The Diversity Landscape of *Plasmodium falciparum var* Genes

Thesis

The Diversity Landscape of *Plasmodium falciparum var* Genes

George Githinji

A thesis submitted for the degree of
Doctor of Philosophy

The Open University (UK)

<u>Affiliated Research Centre</u>

KEMRI-Wellcome Trust Research Programme
Center for Geographic Medicine Research-Coast
Kilifi, Kenya

DATE OF SUBMISSION: 30 SEPTEMBER 2015
DATE OF AWARD: 17 MARCH 2016

Abstract

Plasmodium falciparum Erythrocyte Membrane Protein 1 (PfEMP1) is an important group of cytoadhesive multi-domain *Plasmodium falciparum* surface antigens that are inserted onto the surface of infected erythrocytes and are encoded by the *var* multi-gene family. PfEMP1 antigens are thought to be important targets of naturally acquired immunity to malaria. This thesis describes nucleotide variation in DBLα sequence tags, a semi-conserved region within the DBLα domain that can be PCR amplified using universal primers. PfEMP1 can be classified using a number of approaches some of which are associated with particular expression patterns in severe, non-severe and asymptomatic malaria. This work compared the classification approaches with an aim of understanding the extent of overlap or discordance between them. To explore non-random distribution of variation in respect to *var's* functional specialization, a clustering approach was developed and applied in exploring patterns of nucleotide variation among *var* subsets from different geographical locations in addition to distribution of predicted B and T-cell epitopes among the *var* subsets. In summary the sequences showed distinct patterns of nucleotide substitution that suggests that *var* sequences are under both diversifying and purifying selection.

Dedication

Mum,
Teresiah George

Dad,
Late Jeffrey Kamanu Gathura

Julie and Leon,
Love you

Acknowledgements

I would like to give special thanks to my advisor Dr Peter Bull, for offering me an opportunity to join his research group and for providing supervision and mentorship in the course of my studies. I am extremely grateful to Prof Kevin Marsh and Dr Britta Urban for co-supervising this work. I would also like to thank, Dr Vandana Thathy, Prof James Nokes and Dr Samson Kinyanjui who constituted my student monitoring and appraisal committee (SMAC).

Much gratitude to my colleagues Dr Cheryl Andisi and Dr George Warimwe for allowing me to use the sequence data that they had collected and for useful discussions on malaria immunity. To Dr Carol, Dr Abdi, Faiz, Joshua, Michelle and Esther, I am grateful for our scientific discussions during the lab meetings and a lot of light moments at Pete's "pizza nights". I would like to extend my gratitude to Michael and Jennifer for maintaining the laboratory cultures and for their patience in teaching me to culture malaria parasites when I joined the group.

My sincere appreciation to Dr Vandana and Priscilla for useful discussions and challenging questions which offered good insights on the biology of the malaria parasite. I am very grateful to Dr Greg Fegan for his technical and social support. It was enjoyable talking and sharing insights and opinions on just about everything!

I would like to extend my thank you note to Dr Mathew Berimman for inviting me to the Wellcome Trust Sanger Center in Cambridge, UK and to Prof Chris Newbold for useful discussions on this work. Thank you to Dr Thomas Otto for challenging ideas and for offering a computational platform and support.

A special thank you to Dr Sammy Assefa for making my short stay in Cambridge interesting and memorable and thank you for the exchange of ideas and Dr Sammy Oyola for great friendship.

To my colleagues and office mates Yaw, Victor and James. Thank you for the laughter and useful discussions on immunology and infinite social topics!

Special thanks to my best friend Julie, it is not enough to say thank you for your endless support and to Leon I have reason to begin all over again.

Special thanks to Mrs Elizabeth Murabu, our liaison with the University who tirelessly organized all the official documentation and presentation.

Thank you to the Wellcome-Trust for funding this work through the Strategic Award.

Last and not the least, I am very grateful to the malaria study participants who consented to the malaria studies and provided samples.

Statement of Originality

I hereby declare that this thesis is my own work and contains nothing that is an outcome of collaboration with others unless where stated and that it has not been submitted for award of any degree or diploma.

Sample collection and consenting was conducted as part of a larger integrated study on the development of naturally acquired immunity to malaria in children in Kilifi district.

Samples were sequenced at the Beijing Genomics Institute and the Wellcome-Trust Sanger Institute. These samples were part of a larger study on expression profiling of PfEMP1 sequences.

George Githinji

September 2015

Contents

1 Introduction **1**

1.1 Background 1

1.2 Geographic distribution and life-cycle of malaria parasites 2

1.2.1 Liver stage infection 4

1.2.2 Blood stage infection 5

1.2.3 Sexual stage in the mosquito 7

1.3 Immunity to malaria 8

1.4 The molecular structure of *var* genes 11

1.5 Expression of *var* genes 14

1.6 The role of PfEMP1 in sequestration 15

1.6.1 CD36 16

1.6.2 ICAM-1 16

1.6.3 Chondroitin sulfate-A 17

1.6.4 EPCR 18

1.6.5 Complement receptor 1 18

1.7 Functional classification of *var* genes 18

1.7.1 Working with DBLα sequence tags 21

1.8 Diversity in *var* genes 24

1.8.1 Role of recombination in generation of *var* diversity 25

1.8.2 The role of point mutations in generation of var diversity 26

1.9 Evidence for immunological specialization 26

1.10 Thesis overview 28

1.10.1 Molecular diversity of *var* genes 29

1.10.2 Immunological diversity in *var* genes 29

1.11 Summary of research questions 30

2 Materials and methods 31

2.1 Study location 31

2.2 Sample collection and processing 32

2.3 Sequence data from Kilifi 34

2.4 Sequence data from laboratory isolates 34

2.5 The global data set 34

2.6 DNA amplification 35

2.7 Scripting and code management 35

2.8 Scripting languages and statistical tools 36

3 Development of a sequence clustering and epitope prediction framework 37

3.1 Background 37

3.2 Generating random substitutions 38

3.3 General approach for computer programs 39

3.4 Assembling high quality DBLα sequences 41

3.4.1 Assembling DBLα sequences with phrap 41

3.4.2 CAP3 approach 43

3.5 Defining clusters of similar high quality sequences 48

3.5.1 Reducing noise due to potential PCR errors 48

3.5.2 CD-HIT clustering 49

3.5.3 Vsearch clustering 50

3.6 Aligning and finding mismatches 52

3.7 Computer programs 53

3.7.1 DBLα finder 54

3.7.2 DBLα-Classifier 55

3.7.3 Bio-CD-HIT-Report 55

3.7.4 Block-sharing networks 56

3.7.5 Find-DBLα-mutations 56

3.7.6 DBLα-assembly 57

3.7.7 DBLα CAF Parser 57

3.8 Epitope prediction 57

3.8.1 B-cell epitope prediction 58

3.8.2 Shannon diversity index 59

3.8.3 Hydrophobicity profile 59

3.8.4 Prediction of T-cell epitopes 59

3.8.5 Predivac-util 62

4 Evaluation of classification approaches used to measure expression of var genes in clinical isolates 64

4.1 Background 64

4.2 Methods 66

4.2.1 Data collection and sequence classification 66

4.2.2 Sequence alignment 67

4.2.3 Defining block-sharing groups of sequences using a network approach 67

4.3 Results and discussion 67

4.3.1 Distinct Cys/PoLV groups are associated with *upsA* 70

4.3.2 DC8 and DC13 sequences exhibit distinct recombining patterns 74

4.3.3 Newly described DBLα domains could consist of genetically heterogeneous sequences. 76

4.3.4 Cys/PoLV groups cannot predict DC8 sequences given a global collection of sequences but may do so within a restricted geographical location 77

4.3.5 Other significant DC groups 81

4.4 Conclusion 81

5 The role of point mutations in generating *var* diversity 83

5.1 Background 83

5.2 Methods 84

5.3 Results 85

5.3.1 DBLα sequences are diverse 85

5.3.2 Pattern of random substitutions in DBLα sequences 88

5.3.3 The correlation between frequency of substitutions in random model and in actual sequences 95

5.3.4 Distribution of variation at each codon 97

5.3.5 Density of substitutions in conserved and polymorphic regions of the DBLα sequence tag 98

5.3.6 Variation in homology block D and H 100

5.3.7 Overall distribution of point mutations in DBLα sequence tags . 102

5.3.8 Distribution of point mutations in DBLα sequences from other geographical regions 107

5.3.9 Variation in group A and non-group A sequences 111

5.3.10 Sequence conservation in multiple isolates 113

5.3.11 Substitution patterns in *var1* sequences 114

5.3.12 Insertion/deletions in *var1* 117

5.3.13 *var1* sequences may contain two co-evolving sites 118

5.4 Discussion 119

5.4.1 Limitations of this study 123

5.4.2 Future work 124

6 Identifying potential epitopes 125

6.1 Background 125

6.2 Results 126

6.2.1 Linear B-cell epitopes are correlated with sequence diversity . . 126

6.2.2 The DBLα sequence region contains potential MHC class II T-cell epitopes 129

6.2.3 Predicting peptides with stable peptide-MHC complexes in DBLα sequences 132

6.2.4 Overlapping peptides predicted by both NetMHCII and Predivac were few 134

6.2.5 The frequency of predicted promiscuous epitopes is limited . . 135

6.2.6 Promiscuous peptides from group A-like sequences are conserved137

6.2.7 Predicted promiscuous peptides were expressed largely in group-A sequences 139

6.2.8 Relationship with clinical classification 141

6.2.9 Relationship between peptide expression and age of the host . . 141

6.3 Discussion 142

6.3.1 Linear B-cell epitopes are prevalent in polymorphic regions . . 142

6.3.2 Promiscuous T-cell epitopes are conserved in sequences associated with expression in severe malaria 143

6.4 Conclusion 145

7 Conclusion 146

7.1 Molecular diversity 146

7.2 A recap of the aims of the study 149

7.3 Summary of methods 149

7.4 Understanding relationships between sequences and classification approaches . . . 150
7.5 The role of mutations in generating diversity . . . 150
7.6 Prediction of epitopes and relationship with age and expression . . . 152
7.7 Further work and conclusion . . . 153

References **155**

A Source code **175**

A.1 A script to search for DBLα sequence tags . . . 175
A.2 A Ruby program to classify DBLα sequence tags . . . 179
A.3 A Ruby program to extract DBLα sequence tags from chromatograms . 191
A.4 A bash pipeline to cluster DBLα sequences using an identity cutoff . . . 194
A.5 A program to cluster sequences within isolates . . . 196
A.6 A program to parse cd-hit formatted cluster files . . . 196
A.7 A Ruby program to translate DNA sequences to amino acid sequences . 199
A.8 A Ruby program to find pairwise nucleotide differences from a sequence alignment . . . 200
A.9 A bash program to cluster sequences using uclust . . . 200
A.10 A bash program to generate random substitutions based on a reference sequence . . . 201
A.11 A Perl program to generate random substitutions . . . 204
A.12 An R program to simulate random substitutions . . . 204
A.13 A script to extract regions from a DBLα sequence . . . 205
A.14 A bash program to find variation context . . . 206
A.15 A Python programm for submitting epitope prediction requests to the Predivac server . . . 207
A.16 A program to predict netMHCII epitopes using a parallel approach . . 210

A.17 A Ruby program to print block-sharing networks in csv format 211

A.18 A Ruby program to print homology blocks and hypervarible blocks from sequence tags 213

A.19 A Ruby program to generate random nucleotide substitutions 214

A.20 A Ruby program to parse and work with DBLα sequences from CAF formatted files 216

B Miscellaneous 226

B.1 Regions of diversity in the DBLα sequence tags 226

B.2 Length polymorphism in DBLα sequence tags 228

B.3 Dinucleotide frequency 229

B.4 Relative synonymous codon usage(RSCU) in DBLα sequences 231

B.5 Pattern of variation in var2csa in isolates collected from around the world 234

B.6 Frequency of substitutions in sequences collected from different parts of the world 236

B.7 Clustering ratio 237

B.8 Differences in the distance between polymorphic sites 239

B.9 Homology block F contained relatively more mismatches 240

B.9.1 Within-isolate variation in homology block D and H 241

B.10 HLA prevalence 243

List of Figures

1.1 The life-cycle of *Plasmodium falciparum* parasites 4

1.2 A venn diagram depicting the overlap in severe malaria syndromes . . 7

1.3 The general structure of the *var* gene molecule 12

1.4 Chromosomal arrangement of *var* genes 14

1.5 The general structure of a DBLα sequence tag 22

1.6 A summary of Cys/PoLV classification approach 23

1.7 A block sharing layout of sequences from Kilifi 24

2.1 A map of Kilifi county 32

2.2 Malaria admissions in Kilifi Hospital in the period 1989 - 2012 33

3.1 A figure showing the approach for generating random and uniform substitutions 40

3.2 The DBLα sequence assembly workflow 44

3.3 CAP3 consensus building errors 45

3.4 Examples of contigs with multiple DBLα sequence tags 46

3.5 A summary of DBLα sequence tags and contigs in sequences collected between 2008 and 2010 assembly 47

3.6 The sequence clustering workflow 53

4.1 DBLα sub-domains (x-axis) correspondence with other classification var approaches 69

4.2 Predicting upsA,DC8 and DC13 sequences from Cys/PoLV groups . . . 73

4.3 DBLα block-sharing network 75
4.4 The correspondence between domain cassettes and the Cys/PoLV groups 78
4.5 Bar plots showing the constitution of block sharing groups among the 22 domain cassettes. 79
4.6 A sequence alignment of DC8 sequences from Tanzania and sig2-like sequences from Kilifi 80
4.7 An neighbour-joining tree of DC8 sequences from Tanzania and sig2-like sequences from Kilifi 81

5.1 DBLα sequence clusters generated using Vsearch at 98% similarity . . 86
5.2 DBLα sequence clusters generated using Vsearch at 88% similarity . . 87
5.3 Frequency of random substitutions 91
5.4 Average substitution frequencies in random and observed sequences . 93
5.5 Frequency of random substitutions scaled by the ratio of transitions to transversions 94
5.6 Spearman correlations between frequency of random substitutions, random substitutions scaled with TI:TV ratio and substitutions from observed sequences 96
5.7 Codon base composition statistics for DBLα and AMA1 sequences . . . 98
5.8 The nucleotide diversity between polymorphic(hp) and homology (hb) blocks at 98% and 96% similarity 99
5.9 Distribution of nucleotide substitution in homology block D and homology block H 101
5.10 The distribution of random substitutions and substitutions from DBLα sequences from Kilifi 103
5.11 Nucleotide content in DBLα sequences from Kilifi, Kenya 105
5.12 Distribution of nucleotide substitutions in DBLα sequences collected between 2008 and 2010 106

5.13 Distribution of variation in sequences from Amele in Papua New Guinea108
5.14 Distribution of substitutions in sequences from Ariquemes and Porto Velho in Brazil 110
5.15 Frequency of substitutions in group A-like and non group-A sequences 112
5.16 Frequency of identical sequences in a population of var sequences . . . 114
5.17 The nucleotide content in var1 sequences 116
5.18 The substitution profile of 103 var1 DBLα sequences 116
5.19 Insertion/deletion event in var1 sequences 117
5.20 Neighbour joining tree of var1 DBLα sequences 118

6.1 Correlation between Bepipred scores and sequence diversity 128
6.2 The frequency of predicted binders per HLA class II allele 131
6.3 The binding frequency of predicted immunodominant peptides using Predivac server 133
6.4 A Venn diagram of the overlapping core peptides predicted by NetMHCII and Predivac server 134
6.5 Proportion of Cys/POLV groups among sequences containing promiscuous peptides 138

7.1 Random mutations contribute to new antigenic types 148

B.1 Diversity in DBLα sequences 227
B.2 Distribution of clusters generated with CD-HIT and UCLUST algorithms 228
B.3 Length distribution in DBLα sequence tags from Kilifi 229
B.4 A plot of dinucleotide frequencies in DBLα sequences 231
B.5 Distribution of substitutions in sequences from 19,578 DBLα sequences tags 234
B.6 Distribution of polymorphisms in sequences from 69 var2csa sequences 235

B.7 Frequency of substitutions from 367 sequences collected from different geographical regions 237

B.8 Distribution of substitutions in 162 sequences from Madang in Papua New Guinea 239

B.9 Dotplots showing the nucleotide distance between adjacent substitutions in sequences clustered at different thresholds 240

B.10 Within Isolate distribution of substitutions in homology block F 241

B.11 Frequency of substitutions in homology block D sequences of the same isolate 242

B.12 Within isolate distribution of substitutions in homology block H 243

B.13 HLA alleles in a sample of individuals from Kilifi 244

List of Tables

2.1 This table shows the number of additional sequences from different geographical regions around the world that were collected and used in the study. 34

4.1 A table of different classifications applied to PfEMP1 sequences 65

4.2 The odds of predicting upsA sequences using the cys-POLV definition of group A sequences (cys2-bs1). This group showed significant odds of association with upsA sequences. 71

4.3 The odds of predicting upsA sequences from the CP1 cys-POLV sequences. CP1 sequences are often associated with severe malaria. They had significant odds of predicting upsA sequences. 72

4.4 A logistic regression model showing the odds of predicting upsA sequences using a combination of CP1 and cys2-bs1 (group A-like) sequences. This is a broad definition of sequences that are associated with severe malaria. This group had the highest odds of predicting upsA sequences compaired to CP1 or cys2-bs1 alone. 72

5.1 A table of the number of randomly generated sequences from a DBLα reference sequence 90

5.2 The ratio of synonymous to non-synonymous change (Ka/Ks) was determined in alignment with 57 var1 sequence tags sampled between 2003-2007. Codons 68 and 82 had kaks scores greater than 1 which suggested that they could be under selection pressure. 119

5.3 The ratio of synonymous to non-synonymous change (Ka/Ks) for an alignment of 24 var1 sequence tags that were collected between 2008 and 2010 period. 119

6.1 The frequency of overlapping strong, weak and non binders predicted with NetMHCII from 365 DBLα sequences. There were no strong binders from short peptides with a length of 9 amino acids. The most strong binders were from peptides with 15 amino acids. Peptides with 18 amino acids had fewer strong binders byt the majority of weak binders. 130

6.2 The conservation profile of 34 predicted promiscuous epitopes among from a dataset of sequences from Kenya, Papua New Guinea and South America. 136

6.3 Correlation in the expression profile of 34 promiscuous peptides 140

6.4 A summary of correlations between host age and expression of sequences containing promiscuous peptides 141

B.1 Summary of dinucleotide frequency 230

B.2 Relative synonymous codon usage (RSCU) indices for DBLα sequences in Kilifi. The RSCU is an indicator of codons that are used more frequently than expected. 233

B.3 Clustering ratio for sequences from different geographical regions . . . 238

Chapter 1
Introduction

1.1 Background

Malaria is a major public health and social-economic burden in resource poor settings (Sachs and Malaney 2002). An approximate 1.3% reduction in economic growth in countries with the highest burden of disease is attributed to malaria (Greenwood et al. 2005). Despite a 42% decline in mortality rate in the last decade (World Health Organization 2013), the incidence of malaria in most parts of Africa remains high (Noor et al. 2014). In 2012, the World Health Organization (WHO) reported nearly 207 million cases of malaria, of which an estimated 627 thousand resulted in death (World Health Organization 2013).

Global efforts in eradication and control of malaria indicate a decline in malaria cases in Africa (O'Meara et al. 2008; Okiro et al. 2009a; WHO 2010). Reports of pockets of high malaria prevalence in several regions (Noor et al. 2014) highlight the challenges in malaria eradication efforts.

Malaria is caused by plasmodium parasites. Species of parasites in this genus infect a wide array of vertebrates including humans, reptiles, birds, rodents and monkeys. There are five species of Plasmodium that infect humans; *P. falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi*. *P. knowlesi* was recently described to infect humans (Cox-Singh et al. 2008). The focus of this thesis is *Plasmodium falciparum* the causative agent of the most lethal form of human malaria. *P. falciparum* is transmitted by infected female *Anopheles* mosquitoes.

1.2 Geographic distribution and life-cycle of malaria parasites

The malaria parasite life-cycle involves asexual reproduction in the vertebrate host and sexual reproduction in the insect vector. The life-cycle is characterized by several morphological stages as illustrated in figure 1.1.

The *Plasmodium* species that infect humans are remotely related to each other which suggests that adaptation to humans may have occurred independently. The true origin of these species is a subject of debate (Ollomo et al. 2009; Prugnolle et al. 2010). Duval and colleagues study showed that the great apes are reservoirs for several Plasmodium species (Duval et al. 2010) that are closely related to *P. falciparum*. However, there is still little agreement on which of these closely related plasmodium species is a progenitor to *P. falciparum*. Liu and colleagues suggested a reassessment of the origin of *P. falciparum* based on the available sequence data obtained from plasmodium parasites infecting the great apes (Liu et al. 2010).

P. falciparum is more prevalent in Africa than in other parts of the world (World Health Organization 2013). *P. malariae* is prevalent in sub-saharan Africa, south-east Asia, Indonesia, the Pacific islands and the Amazon basin. *P. ovale* is prevalent in sub-saharan Africa and Asia (Collins and Jeffery 2005). An accurate diagnosis for *P. ovale* is difficult because it presents with non-specific symptoms, for example, digestive and respiratory symptoms which are also characteristic of viral infections.

Most rapid diagnostic and detection kits are not sensitive enough for diagnosis. Unlike in *P. falciparum* infections, *P. ovale* infections are rarely characterized with anaemia.

Plasmodium vivax is predominant in Asia partly because of the wide-spread prevalence of the duffy antigen in the population. The duffy antigen is the receptor for *P. vivax* malaria parasites (Mason et al. 1977; Miller et al. 1976). A case of *P. vivax* like infection has been reported in duffy negative individuals in Kenya (Ryan et al. 2006). Occasionally, a patient may harbour another *Plasmodium* infection in addition to *P. falciparum* (Genton et al. 2008; Douglas et al. 2011). *P. vivax* diagnosis is a challenge because of the low number of parasites in peripheral circulation. Most *P. vivax* infected patients often harbour a dormant liver stage infection (see section 1.2.1) that is only detectable upon relapse (Mueller et al. 2009).

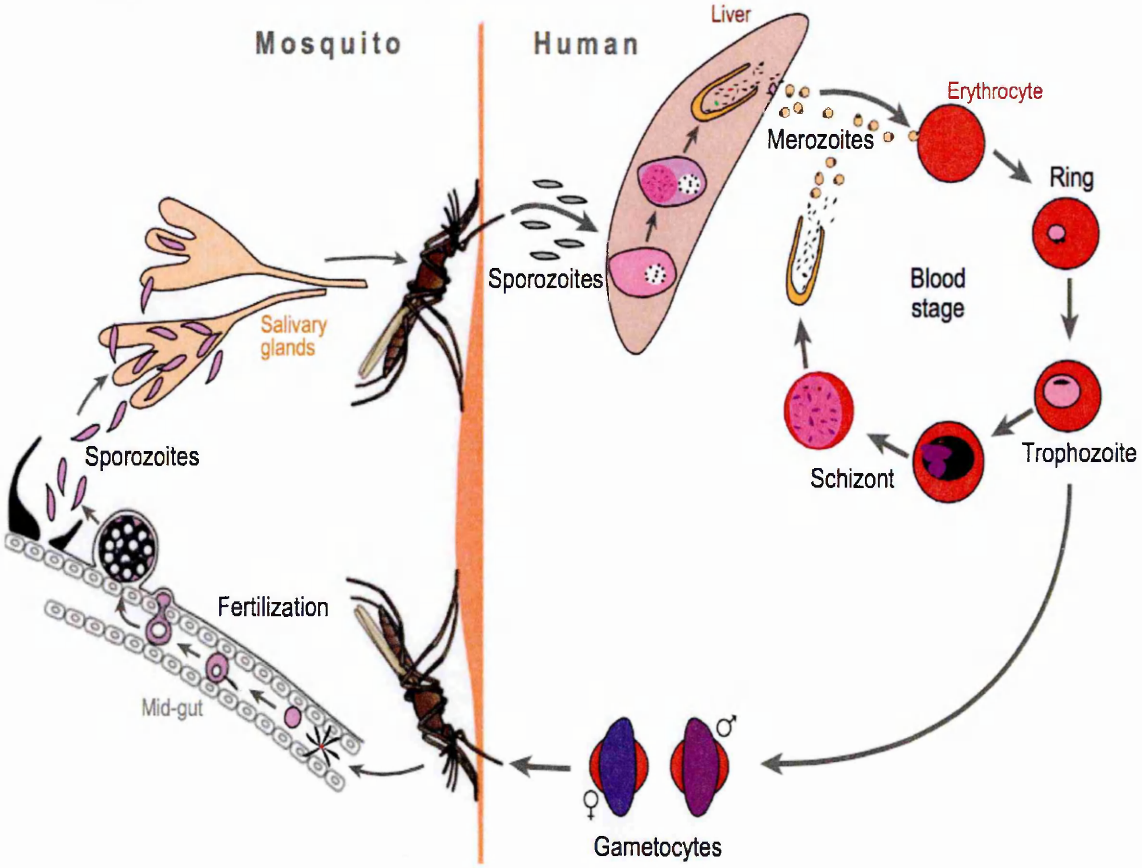


Figure 1.1: *The life-cycle of the malaria parasite Plasmodium falciparum. Sporozoites are injected in to the human host by a female anopheles mosquito during a blood meal. Sporozoites migrate to the liver and invade the liver cells where they mature and form merozoites which are released into the bloodstream. Merozoites invade red blood cells and initiate an asexual multiplication cycle. Some merozoites develop into gametocytes which are the transmissible parasite forms. Gametocytes are ingested by a mosquito during a blood meal where they develop through a sexual cycle and into sporozoites.*

1.2.1 Liver stage infection

An infected female *Anopheles* mosquito can inject between 1 and 100 *Plasmodium falciparum* sporozoites with a median of 15 sporozoites as observed on mineral oil from restrained *Anopheles stephensi* mosquitoes (Rosenberg et al. 1990), and a median of 22 sporozoites from another study that involved mice (Ponnudurai et al. 1991).

The studies also showed that the number of sporozoites on the skin correlated with the size of the innoculum but not with the number of sporozoites in the salivary glads.

Upon successful invasion, the sporozoites lounge in the hepatocytes and within 2-10 days, depending on the parasite species, they asexually replicate during exo-erythrocytic schizogony culminating in the production of merozoites. *P. vivax* and *P. ovale* parasites can remain dormant in the liver for a period of time (Krotoski et al. 1980; Markus 2011b) and reactivate weeks to months after the primary infection. This phenomena is attributed to relapses that are observed in *vivax* and *ovale* malaria infections although some reports have refuted the idea that relapse emanate from the so called "hypnozoite" dormant stages (Markus 2015; Markus 2011a).

1.2.2 Blood stage infection

After maturation, merozoites exit the liver cells and invade the red blood cells. This marks the beginning of the disease development stage and the associated symptoms are usually observed at this stage. Inside the red blood cells, they divide asexually for a period of 1-3 days during which time the parasite goes through a 48-hour morphological transformation; from a ring form, to a trophozoite and a schizont form. *P. malariae* exhibits a 72-hour cycle unlike the other *Plasmodium* species.

In *Plasmodium falciparum*, red blood cells containing rings circulate within the peripheral blood, whereas red blood cells containing trophozoite and schizont parasite forms, may adhere to andothelial cells (Gitau et al. 2012a; Miller 1969). This phenomenon is called sequestration and is involved in pathogenesis and parasite survival. These stages are key to the special pathology observed in *P. falciparum* compared to other parasites. Sequestration is aided by variant surface molecules that also mediate antigenic switching. Sequestration assists the parasite in escaping from splenic clearance (Wyler et al. 1979). Additionally, antigenic switching helps *Plasmodium falciparum* in establishing a chronic blood stage infection in humans. The exact mechanisms of establishing chronic infections have not been fully elucidated.

During a chronic infection, parasites invade and cause the lysis of red blood cells. Parasite-derived molecules produced by the parasite lead to overproduction of serum-bound factors some of which result in inflammatory responses by host-derived cytokines and subsequently lead to the observed intermittent fever episodes given the synchronous lysis of red blood cells. Severe red blood cell lysis leads to anaemia. Furthermore, sequestration and adhering to vascular endothelial surfaces is associated with development of malaria pathogenesis.

In general symptomatic malaria is often defined based on presence of fever and clinical manifestations observed at the time of presentation. Since clinical manifestations and parameters vary with age and geographical location and are often very similar to those observed in other diseases, it is difficult to accurately define severe and non-severe malaria, therefore, severe malaria has been defined using a parasitemia cutoff and in terms of symptoms that are most strongly associated with death. In 1994, Marsh and colleagues, showed that severe malaria is composed of different overlapping clinical syndromes (figure 1.2), i.e. severe anaemia, respiratory distress, and impaired consciousness (Marsh et al. 1995; Berkley et al. 2009).

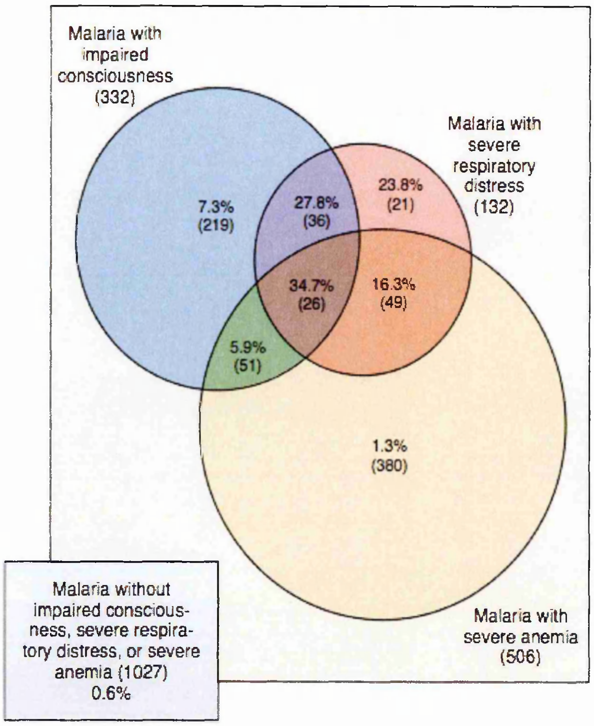

Malaria with impaired consciousness (332)
Malaria with severe respiratory distress (132)
7.3% (219)
27.8% (36)
23.8% (21)
34.7% (26)
16.3% (49)
5.9% (51)
1.3% (380)
Malaria without impaired consciousness, severe respiratory distress, or severe anemia (1027) 0.6%
Malaria with severe anemia (506)

Figure 1.2: *A venn diagram showing the overlap in clinical malaria syndromes. The associated prevalence and mortality are also shown. (Marsh et al. 1995)*

1.2.3 Sexual stage in the mosquito

A few of the merozoites differentiate into gametocytes which are the transmissible forms of the parasite. The gametocytes are picked by a mosquito during a blood meal and form male and female gametes within the mosquito. They fuse and form a diploid zygote within the intestinal wall and ultimately differentiate into oocysts. Inside the oocysts, repeated mitotic divisions form thousands of active sporozoites. Eventually the oocysts burst and release the sporozoites into the body cavity and they migrate to the mosquito's salivary glands where they are injected into a human host during a blood meal.

1.3 Immunity to malaria

The risk and burden of severe malaria in children under five years of age is not well understood and little is known about malaria in infancy (children < 6 months). It is thought that passive transfer of maternal antibodies (Snow et al. 1998; Kitua et al. 1999; Hviid and Staalsoe 2004) and fetal hemoglobin (Amaratunga et al. 2011) could play a role in conferring resistance to severe malaria among these very young children.

In malaria endemic regions, susceptibility to severe life-threatening malaria increases in the first year of life but after the age of five years susceptibility to life threatening disease decreases but sterile immunity to severe malaria is often never achieved. Immunity to non-severe malaria is established slowly in individuals from malaria endemic regions (Gupta et al. 1999). Clinical attacks are rarely encountered in adults although adults often carry asymptomatic infections as reviewed by Marsh and Kinyanjui (2006) and Baird et al. (1998).

Some studies have suggested that immunity to severe non-cerebral malaria is acquired after only a few infections (Gupta et al. 1999) and that protection is often lost if individuals migrate to areas of low transmission (Jelinek et al. 2002) which could suggest that intense and sustained transmission is necessary in order to maintain naturally acquired immunity. In a recent study from Tanzania, Gobcalves and colleagues observed that resistance to severe malaria was not acquired after one or two infections (Gonçalves et al. 2014) as has been proposed by earlier studies. It is likely that lack of exposure and thereby failure to develop naturally acquired immunity may lead to large-scale epidemics if interventions are not put in place. An instance of this phenomenon was described in areas where malaria was re-introduced after successful eradication measures (Mouchet et al. 1997).

Evidence that antibodies are important in control of malaria parasites came from work by Cohen and colleagues who successfully treated children with γ-globulin purified from adult individuals who were exposed to malaria in the Gambia. Cohen showed that treatment with purified γ-globulin from individuals in malaria endemic region in the Gambia, cleared parasites and resolved malaria symptoms in young children (Cohen et al. 1961). This experiment was built upon work by Coggeshall and Kumm who in 1937 demonstrated passive transfer of malaria immunity to normal rhesus monkeys from rhesus monkeys with chronic *P. knowlesi* and *P. inui* parasites, which suggested the presence of protective antibodies to malaria parasites (Coggeshall and Kumm 1937). In addition McGregor found that Gambian children who used weekly doses of chloroquine had significantly lower levels of γ-globulin than children that did not take the weekly dose, suggesting that parasites were important in generation of antibodies (McGregor et al. 1956). These studies did not establish whether the antibodies were protective. Edozien attempted to establish the relationship between serum γ-globulin in protected and unprotected children (Edozien et al. 1960). Their observations were reproduced by later studies using γ-globulin from adults in Cote d'Ivoire which showed reduction in parasite density in Thai patients (Sabchareon et al. 1991). From these studies it was evident that antibodies are important in control of malaria parasites. To link antibodies and protective immunity, it was important to establish the antigenic source that elicited the antibodies.

In 1938, Eaton demonstrated that the serum of eight repeatedly infected monkeys agglutinated parasites more readily and better at lower dilutions than that of two chronically infected monkeys (Eaton 1938). Building on Eaton's agglutination experiments, Brown used *P. knowlesi* parasite in agglutination experiments to demonstrate antigenic variation in Rhesus macaques (Brown and Brown 1965). Later studies that explored agglutination of malaria parasites used homologous and non-homologous serum from children and adults and with human malaria parasites (Marsh and Howard 1986). They proposed that *P. falciparum* protein antigens that are present on the surface of the infected erythrocyte are important targets of protective immunity (**Marsh:1989uy**; Bull et al. 1998).

Of the several parasite proteins on the surface of infected red blood cells, *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) are the most studied. Evidence that PfEMP1 are important targets from protective immunity comes indirectly from observations that antibody levels are correlated with protection from clinical malaria in individuals from West Africa (Ofori et al. 2002) and East Africa (Kinyanjui et al. 2004). Furthermore serological experiments have showed that children are more likely to be recognized by pre-existing homologous antibodies (antibodies elicited by the infecting parasite) more than the heterologous antibodies (Bull et al. 1998) and that agglutination by diverse plasma is associated with severe disease (Bull et al. 2000). Evidence that PfEMP1 are targets of protective antibodies came from gene knockout studies by Chan and colleagues in an experiment where they used gene knockout to modify *P. falciparum* parasite lines to silence the expression of PfEMP1 molecules by using a drug selectable marker under the control of a promoter from a gene encoding native PfEMP1 that was transfected into the parasites. The parasites were grown under drug pressure and expression of the marker allowed the transfectants to grow and silence endogenous PfEMP1 transcription. A significant reduction in the recognition of PfEMP1-silenced infected red blood cells by sera from immune adults was observed. The PfEMP1-silenced parasites were characterized by a decline in opsonic phagocytosis and in-vitro binding to host receptors (Chan et al. 2012). This experiment confirmed that in laboratory isolates, PfEMP1 are major targets of antibodies.

This experiment, together with evidence that PfEMP1 are the targets of antibodies (Leech et al. 1984) as inferred from serological studies (Bull et al. 1998; Piper et al. 1999) support a role of PfEMP1 as targets of naturally acquired immunity.

1.4 The molecular structure of *var* genes

Var genes are 6-13kb two exon molecules (Gardner et al. 2002; Su et al. 1995) that encode the PfEMP1 molecule. Exon 1 encodes an extra-cellular region which has a modular structure and exon 2 encodes an intra-cellular region that anchors the molecule on the surface of the infected erythrocyte (Gardner et al. 2002). *P. falciparum's* haploid genome contains approximately 60 *var* genes (ibid.). A schematic diagram of a *var* gene is shown in figure 1.3.

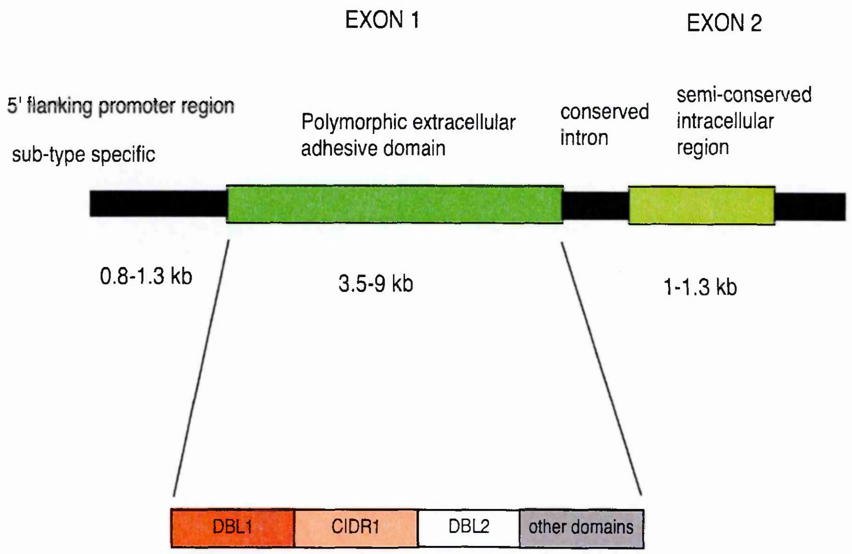

EXON 1
EXON 2
5' flanking promoter region
sub-type specific
Polymorphic extracellular adhesive domain
conserved intron
semi-conserved intracellular region
0.8-1.3 kb
3.5-9 kb
1-1.3 kb
DBL1
CIDR1
DBL2
other domains

Figure 1.3: *Var gene architectures are very diverse. This diagram depicts one of the many typical domain architectures. The molecule comprises of an semi-conserved intra-cellular exon that codes for the acidic terminal segment and an extra-cellular exon that codes for the polymorphic cytoadhesive domain. Exon 1 and 2 are interspersed by a conserved intron. The semi-conserved 5' flanking promoter region has functional relevance and is used for grouping var gene sequences into ups-types. Specifically ups-A and ups-B are associated with parasites that are expressed in severe malaria cases, whereas ups-C are generally expressed in asymptomatic cases.*

The *var* genes encode the high molecular weight PfEMP1 molecules which are made of a combination of two to nine domains organized in a modular architecture comprising an N-terminal segment (NTS), Duffy Binding-like (DBL) which are homologous to adhesive domains in *P. falciparum* EBA-175 and the Duffy-binding proteins of *Plasmodium vivax* and *Plasmodium knowlesi* (Peterson et al., 1995), cysteine rich inter-domain region (CIDR) and the cytoplasmic acidic terminal segment (ATS) (Smith et al. 2000b; Gardner et al. 2002; Lavstsen et al. 2003).

Based on sequence homology, the DBL domains are divided into six sub-groups (α, β, γ, δ, ϵ and ζ), and the CIDR domains are divided into four sub-groups (α, β, δ and

γ) (Su et al. 1995; Lavstsen et al. 2003; Kraemer et al. 2007; Smith et al. 2000a). The DBL domains are further divided into homology blocks A-J that are interspersed by hyper-variable regions (Smith et al. 2000a) and also by three structural sub-domains S1-S3 (Higgins 2008). A study by Rask and colleagues has suggested further divisions of the DBL and CIDR domains based on full length sequence data of 7 laboratory parasite lines (Rask et al. 2010) (Discussed in section 1.7.)

Var genes are found in all fourteen chromosomes of the parasite genome. In the 3D7 genome, the majority are encoded in the subtelomeric and the rest in the centromeric regions (Gardner et al. 2002). Telomeric genes are transcribed in a tail-tail, head-to-tail or head-to-head orientation in respect to each other. Occasionally a member of the *repetitive interspersed families of polypeptides (rifin)* is found between the *var* genes. Centromeric *var* genes are found in tandem in a head to tail orientation of 3-7 genes (Kyes et al. 2007).

The var gene orientation is often associated with the ups type of the 5'-flanking regions (Gardner et al. 2002; Lavstsen et al. 2003). *UpsA* genes are transcribed towards the telomere while upsB are often transcribed towards the centromere. *upsC var* genes are located at the centromeres. The structural arrangement of *var* genes is depicted on figure 1.4. This structural arrangement is thought to be functionally important and to influence recombination patterns (Kraemer et al. 2007) and regulation of gene expression (Ralph et al. 2005; Howitt et al. 2009).

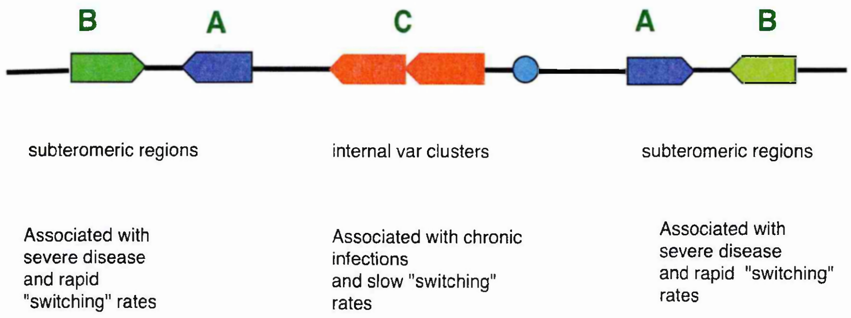

B
A
C
A
B
subteromeric regions
internal var clusters
subteromeric regions
Associated with severe disease and rapid "switching" rates
Associated with chronic infections and slow "switching" rates
Associated with severe disease and rapid "switching" rates

Figure 1.4: *Chromosomal arrangement of var genes. The upstream promoter classification is denoted by letters A-C. Group A genes are transcribed towards the telomere and group B are transcribed away from the telomere. The internal var clusters are found on chromosome 4,6,7,8,and 12 wheres the subteromeric vars are on most chromosomes. Preferential recombination occurs among the group A genes and similarly among the group B var genes.*

Unusually conserved *var* genes across most isolates include the *var1*, the *var2csa* and the *type 3 var* genes. *Var1* and *var2csa* have distinct upstream promoters regions that were described as ups-D and ups-E respectively by Lavstsen and colleagues (Lavstsen et al. 2003).

1.5 Expression of *var* genes

Regulation of expression in *var* genes is controlled at the transcription initiation, translation and epigenetic levels for example through nuclear localization. *Var* genes are expressed in a mutually exclusive manner (Scherf et al. 1998; Kyes et al. 2007) 16-18 hours post invasion. Transcription control involves interaction of the 5' *var* promoter situated upstream of the first exon and the var intron promoter. Epigenetic control of *var* gene expression involves histone acytylation and methylation (Lopez-Rubio et al. 2007; Freitas-Junior et al. 2005).

Silenced *var* genes tend to localize within regions of the nucleus that contain condensed heterochromatin whereas active transcription occurs in regions where chromatin is open for transcription. It has been shown that *var* genes localize mostly at the nuclear periphery regardless of chromosomal location or their activation state they appear to move upon changes in transcription activity (Duraisingh et al. 2005). Telomeric clusters are located within heterochromatin region of the nuclear periphery whereas upon activation it moves to another location of the nuclear periphery where the chromatin is open for transcription. This suggests that sub-nuclear organization also plays a role in regulation of *var* gene expression.

1.6 The role of PfEMP1 in sequestration

While the early stages of the blood stage malaria parasite can be observed under the microscope in a peripheral blood smear, the trophozoite and schizont stages are found in parasites that are sequestered in endothelial tissues. Sequestration is mediated by an interaction between the host endothelial receptors mainly through cytoadhesive domains of the PfEMP1 surface proteins. Parasite infected erythrocytes are able to bind to a wide range of endothelial receptors that include ICAM-1 (Berendt et al. 1989) PECAM, CD36 (Barnwell et al. 1989; Newbold et al. 1997), ELAM-1, VCAM-1, chondroitin sulfate (CSA) (Rogerson et al. 1995; Reeder et al. 1999) and endothelial protein-C (EPCR) (Turner et al. 2013). Infected erythrocytes can bind to uninfected erythrocytes to form rosettes (Udomsangpetch et al. 1989). Interactions between infected erythrocytes and platelets in-vitro result in platelet mediated clumping (Pain et al. 2001).

Several studies have shown that sequestration of infected erythrocytes to the endothelial receptors contributes to development of malaria pathology using different host ligands such as CD36, ICAM-1, CSA, CR1,and EPCR as outlined below.

1.6.1 CD36

CD36 is expressed by several types of cells that include endothelial and epithelial cells, macrophages, monocytes, platelets, erythrocyte precursors and adipocytes (Udomsangpetch et al. 1997). Most parasite infected erythrocytes bind to CD36 (Newbold et al. 1997). Most CD36 studies are small and related to a single functional investigation nonetheless most studies have shown that the CIDR domain of PfEMP1 mediate binding to CD36 (Baruch et al. 1997; Smith et al. 1998). Recombinant PfEMP1 fragments can bind to CD36 and antibodies raised against these fragments can block adherence of infected erythrocytes from different isolates to CD36 (Baruch et al. 1997). Parasites that express *var* group B and C, are important in CD36 binding (Robinson et al. 2003).

1.6.2 ICAM-1

Inter-cellular adhesion molecule 1 also known as CD56 is 90-115kDA transmembrane glyco-protein and a member of the immunoglobulin super family that is expressed on a variety of cell types. It is involved in signal transduction and is a receptor for several ligands. Binding to ICAM-1 by infected erythrocytes may play a role in development of cerebral malaria syndrome pathogenesis (Ochola et al. 2011). Ochola et al. (ibid.) analyzed the role of ICAM-1 binding variants under static and flow adhesion assays to show that parasites from children with cerebral malaria bound more to ICAM-1 under flow conditions. A subset of PfEMP1 molecules with the DBLβ domain pair bind to ICAM1. ICAM1-binding molecules are encoded mainly by group B and group C PfEMP1 proteins (Howell et al. 2008), although some group A PfEMP1 have been shown to encode ICAM-1 binders (Oleinikov et al. 2009). The strength and the level of binding varies. One study reported that infected erythrocytes from children with cerebral malaria tend to bind more to ICAM-1 than from children with asymptomatic malaria (Newbold et al. 1997). In another study there was no difference in binding in children with severe malaria compared to those with non-severe cases (Rogerson et al. 1999). A frequent polymorphism in the ICAM-1 gene in African populations is not associated with protection against severe malaria (Craig et al. 2000) but is reported to be protective against febrile illness in infants (Jenkins et al. 2005). It has been suggested that ICAM-1 could work in a synergistic way with CD36 to enhance static adhesion (McCormick et al. 1997; Yipp et al. 2007).

1.6.3 Chondroitin sulfate-A

P. falciparum parasites adhere to CSA (Rogerson et al. 1995) and a distinct PfEMP1 variant protein, *var2csa*, binds to chondroitin sulphate-A (Salanti et al. 2004) which results in parasite sequestration in the placenta of pregnant and infected women (**Fried:1996ws**). Pregnancy associated malaria is a major cause of poor birth outcomes that include abortion, still-birth and low-birth weight mostly in first-time mothers in malaria endemic regions. Most women develop antibodies that inhibit binding of parasites to the placenta and are less affected in the second, third and later pregnancies (Fried et al. 1998). This observation forms a basis for developing protective

PfEMP1 based vaccine and serves as a model for PfEMP1 based immunity to malaria.

1.6.4 EPCR

Endothelial protein-C receptor (EPCR) is a receptor for activated protein-C, a serine protease involved in the blood coagulation pathway. In 2013 Turner and colleagues reported that parasites causing severe malaria in a group of children expressed EPCR-binding PfEMP1. They showed that EPCR binding was associated with PfEMP1 molecules carrying domain cassette 8 (DC8) and 13 (DC13) (Turner et al. 2013). Domain cassettes are discussed in section 1.7. Importantly they showed that parasite binding to human brain micro-vascular endothelial cells through EPCR was significantly higher in isolates from patients with severe malaria than in children with non-severe malaria.

1.6.5 Complement receptor 1

The complement receptor 1 (CR1) is a complement regulatory protein that is expressed by erythrocytes, leukocytes and dendritic cells. CR1 is implicated in rosetting (Rowe et al. 1997), a parasite phenotype that is associated with severe malaria (**Carlson:1990fq**; Warimwe et al. 2012).

1.7 Functional classification of *var* genes

Var genes exhibit extreme molecular diversity. Only a limited set of functional phenotypes are associated with parasites. Var classification is important to distinguish var types that may be associated with particular virulent phenotypes. Methods used to classify *var* genes are based on homology in coding, non-coding and upstream sequence regions (Voss et al. 2003; Voss et al. 2000; Vázquez-Macías et al. 2002; Gardner et al. 2002) as well as conservation of domain architecture across multiple isolates. On the other hand, the Cys/PoLV DBLα classification method is based on sequence properties of a short PCR amplified region of the DBL1α (see figure 1.3) domain rather than sequence homology.

Based on the conservation of the 5'-flanking region var genes were classified into three major groups, *upsA, upsB* and *upsC* (Gardner et al. 2002). A number of studies have associated upsA *var* sequence expression with severe and life-threatening disease outcomes. First, Rottmann et al. (2006) using real time PCR primers showed that var group A and B transcripts were more abundant in severe malaria patients than in patients with uncomplicated malaria. A study from Kenya associated group A genes with rosetting (Warimwe et al. 2012), a phenotype that has been associated with disease severity in African children (Doumbo et al. 2009). Studies by Warimwe et al. (2009) and Kyriacou et al. (2006) provided indirect association using var DBLα sequence tags. Both studies provided evidence to support that a subset of tag sequences with 2 cysteines tend to be associated with group A *var* genes.

Expression of *upsB* sequences is associated with severe and mild malaria (Rottmann et al. 2006; Kaestli et al. 2006) while a large proportion of *upsC* sequences are associated with expression in asymptomatic cases (Rottmann et al. 2006; Falk et al. 2009). However, significant expression of upsC sequences in severe malaria cases has been reported (Kalmbach et al. 2010)

These observations suggest that the *ups* classification is very broad and does not always discriminate between disease phenotypes. Although there are clear associations between function and features associated with upsA, lack of a distinct sequence signature within *var* genes means that the relationship between sequence and functional role or phenotype cannot be direct.

In 2010, Rask and colleagues described twenty-three non-random domain arrangements that were commonly found in full-length sequences of seven parasite genomes (Rask et al. 2010). These non-random domain arrangements were designated as domain cassettes (DC). *Var* gene segments can be classified based on conservation of these domain structures or cassettes.

A couple of studies have shown that *var* sequences that are characterized with DC8 and DC13 domain cassettes are associated with expression in severe malaria (Lavstsen et al. 2012; Bertin et al. 2013). DC8 and DC13 are important in binding to the endothelial protein-C receptor (EPCR) (Turner et al. 2013). DC8 *var*-encoded genes from the IT4 line were shown to bind to endothelial cells from various organs and notably from the brain endothelial cells (Avril et al. 2012; Claessens et al. 2012). The major drawback with this approach is that the domains were defined based on homology and most variation is not in the parts of the sequences that are directly linked with adhesive function. and the nature of molecular sequence diversity is not correlated with the extent of functional diversity. Therefore, sequencing large numbers of clinical parasites coupled with functional assays is important in order to establish the relationship between sequence conservation and cytoadhesive properties.

1.7.1 Working with DBLα sequence tags

Given the challenges of sequencing full-length *var* genes and classification, a lot of clinical studies have sequenced the DBLα region of in the DBL1 domain because it can be amplified using universal primers.

The DBLα sequence tag is a short 300-500 nucleotide segment found in the DBL1α domain of the *var* gene. DBL1α domains are found in nearly all var genes. While the DBL1 domain is not very conserved in sequence, it is characterized by the presence of semi-conserved homology blocks that can be targeted by universal primers (Taylor et al. 2000a). It is likely that the tag region is in linkage disequilibrium with genetic or structural features upstream or downstream of the molecule. Figure 1.5 shows a schematic diagram of the DBLα sequence tag highlighting the major characteristic features. The `MFK` and `REY` motifs are not present at the same time in a given tag (Bull et al. 2007).

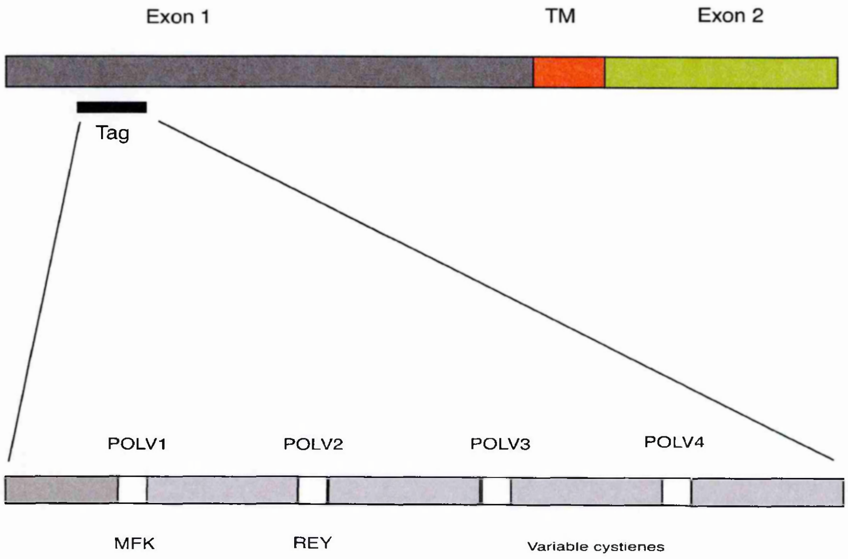


Figure 1.5: *A diagram depicting the size of the DBLα sequence tag relative to the full length var gene and the features that are characteristic of the tag regions. Relatively conserved regions within the tag are shown and they are interspersed with polymorphic regions. The relatively location of two important motifs MFK and REY is shown (Bull et al. 2007)*

DBLα sequence tags are characterized by relatively conserved N-terminal and C-terminal regions, distinct number of cysteine residues, two mutually exclusive motifs and a distinct length distribution (Bull et al. 2005). DBLα sequence tags can be classified into six groups depending on the combination of the its features. The six groups are CP1 (cys2-MFK^{+ve}), CP2 (cys2-REY^{+ve}), CP3 (cys2 REY^{-ve} MFK^{-ve}), CP4 (cys4-REY^{-ve} MFK^{-ve}), CP5 (Cys4-REY^{+ve}) and CP6. CP6 is the default group for sequences that do not fall into any of the previous five groups. Figure 1.6 shows a summary of Cys/PoLV classification system.

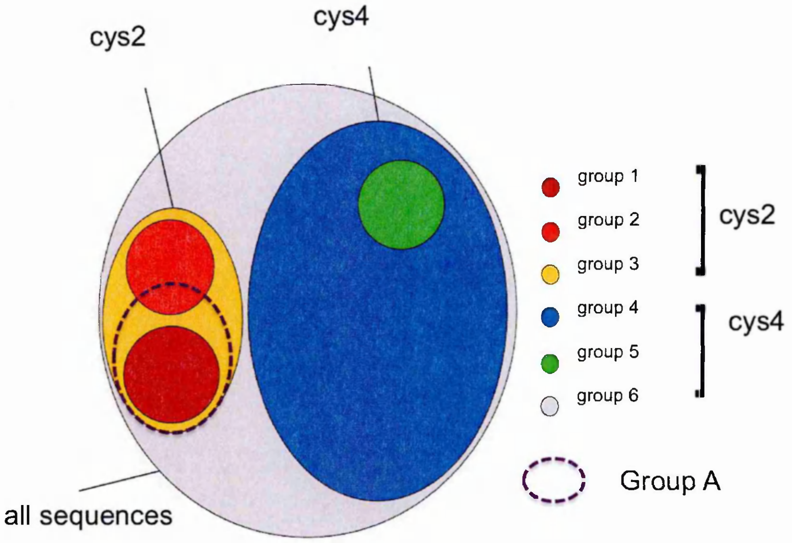


Figure 1.6: *DBLα sequence tags can be classified into six sequence groups. Group 1 (CP1) sequences contain 2 cysteines and an* `MFK` *motif, group 2 (CP2) contain 2 cysteines and an* `REY` *motif, Group 3 (CP3) contain 2 cysteines without any of the motifs. Group 4 (CP4) contain 4 cysteines and Group 5 (CP5) contain 4 cysteines and an* `REY` *motif. Sequences that do not fall into any of the previous groups are classified as group 6 (CP6). Some sequences share short sequences blocks with each other such that they form a network of block sharing sequences. The largest components of this network were termed as Block-Sharing group 1 and Block-sharing group 2. The group-A sequences contain 2 cysteines and tend to fall on block-sharing group 1.*

A study by Bull et al. (2008) showed that sequences that shared 14 amino-acid sequence blocks with each other formed an unbroken network of sequences that comprised of a major and a minor component. Sequences that were connected in the major component were termed as the block-sharing group 2 and those that fell on the minor component were termed as block-sharing group 1. The majority of the sequences did not fall in these components. Sequences in the block-sharing group 1 corresponded well with group A sequences (ibid.).

Not all sequences fitted into these networks and therefore a few sequences were connected to both bs1 and bs2 networks. Figure 1.7 shows a layout of sequences on a block sharing network.

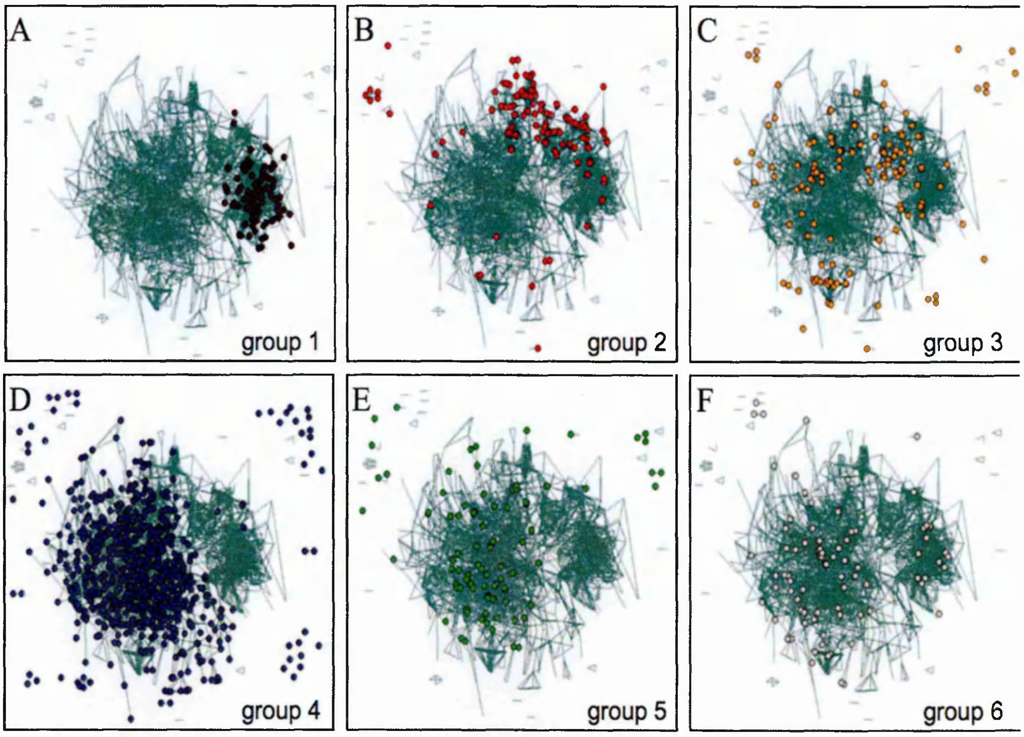


Figure 1.7: *A layout of block sharing networks from a collection of DBLα sequences samples from around the world. DBLα sequences from Kilifi mapped on the block-sharing network. Each point represents a sequence. Sequence belonging to block-sharing group 1 are shown on the network coloured by the respective cys/PoLV group. The majority of these DBLα sequence tags contain two cysteines. (Courtesy of (Bull et al. 2008)).*

1.8 Diversity in *var* genes

In malaria endemic regions an individual can suffer from multiple *P. falciparum* infection and re-infections which suggests that sterilizing immunity against malaria is difficult to establish. Through antigenic variation, the parasite is able to avoid immune clearance and therefore perpetuate a persistent infection by exploiting 'holes' in the immunity of the host as reviewed by Bull and Marsh (Bull and Marsh 2002).

Several studies (Falk et al. 2009; Freitas-Junior et al. 2000; Kraemer et al. 2007; Taylor et al. 2000a) suggest that *var* recombination occurs in coding sequences; and that it generates rapid variation in the encoded proteins (Taylor et al. 2000a; Ward et al. 1999).

1.8.1 Role of recombination in generation of *var* diversity

Recombination in *var* genes is generated during meiosis (Taylor et al. 2000b; Freitas-Junior et al. 2000) and mitosis (Claessens et al. 2014; Bopp et al. 2013). Deitsch and colleagues showed evidence for intra-cluster recombination of *var* genes in chromosome 12 that was characterized with spontaneous switches in the transcription of genes (Deitsch et al. 1999). Using a HB3xDd2 genetic cross in hybridization study, Freitas-Junior showed that parasites from genetically diverse backgrounds can share sequenced blocks with other *var* genes (Freitas-Junior et al. 2000). Zilversmit and colleagues defined large and continuous blocks of homologous sequences among *P. falciparum* and *P. reichenowi* DBLα domains (Zilversmit et al. 2013). This suggests that recombination accounts for diversity within and between species.

Variation in the level of recombination between different var gene subsets appears to play a role in structuring var genes in the populations. For example group A *var* genes tend to recombine with fellow group A and less with non-group A (Kraemer and Smith 2003; Kraemer et al. 2007). This could be important in maintaining overall *var* function. Despite this, the architectures of var genes are poorly maintained between different parasite genomes (Kraemer et al. 2007).

Few studies have looked at sequence diversity in *var* genes due to lack of epidemiological sampling frameworks given the challenges of obtaining, sequencing and classifying *var* genes from clinical isolates. In a study by Barry and colleagues, DBLα sequences were grouped into "sequence types". By sampling from global population of sequences, a "type" plateau was never arrived at because of the immense number of sequence "types" suggesting extreme diversity within the DBLα region of the PfEMP1 (Barry et al. 2007).

1.8.2 The role of point mutations in generation of var diversity

The role of point mutations and nucleotide substitution in generation of sequence diversity in *var* genes is still poorly understood. While it is known that point mutation occurs in DNA sequences it is difficult to ascertain to what extent point mutation is important in generation of diversity in *var* sequences especially with the long history of recombination. It is important to distinguish the biological processes that generate specific types of point mutations. Relating mutation to specific selective pressure could provide insight on processes maintaining diversity but the challenge is on approaches to quantify and separate diversity that is generated through recombination from diversity occurring through random mutation processes.

1.9 Evidence for immunological specialization

Building on the idea of functional specialization discussed in section 1.7, several studies suggest that distinct *var* groups tend to associate with particular adhesive properties provides evidence for functional specialization. The idea of functional specialization is that if the cost of optimal adhesive properties was conserved antigenic types, then, antigenic types would be rapidly eliminated leading to selection of variants with novel antigenic types (Bull et al. 1999). Evidence in support of immunological specialization comes from an experiment with parasites from different geographical regions. Bull and colleagues observed that parasites associated with severe malaria in non-immune individuals express a subset of variant surface antigens that was commonly recognized by immune serum compared to those that were associated with uncomplicated malaria (Bull et al. 2000). This experiment was confirmed from the observation that antigens expressed by parasites from young Ghanaian children with malaria were commonly and strongly recognized by plasma from healthy children in the same locality, whereas recognition of antigens expressed by parasites from older children was less frequent (Nielsen et al. 2002).

Furthermore the observation that expression of commonly recognised variants was accompanied by up-regulation of group A *var* genes (Jensen et al. 2004) in 3D7 isolates suggests that different subsets of PfEMP1 have different levels of antigenic conservation. Jensen and colleagues showed that group A *var* genes were up-regulated in 3D7 parasites that were commonly recognised. The commonly recognized variants were selected on the basis of the observation that they are recognized by plasma IgG of semi-immune children as previously described by Bull and colleagues.

The above observations have been corroborated by studies using the DBLα tags where clinical expression studies on *var* have shown that CP1, CP2 and CP3 sequences are often expressed in children with severe malaria while CP4 are often expressed in children with non-severe malaria (Kyriacou et al. 2006; Warimwe et al. 2009;

Kirchgatter and Portillo 2002). Additional evidence comes from studies which show that group A *var* genes are often expressed in samples collected from young individual that are less immune (Kyriacou et al. 2006; Warimwe et al. 2009).

Using data from full length *var* genes, Buckee and Recker reported an association between the number of domains in a *var* gene and sequence conservation (Buckee and Recker 2012). From the same group of sequences, they also showed that *upsA* genes were more conserved relative to *upsB* and *upsC*, suggesting some level of structural specialization that is consistent with known recombination patterns (Kraemer et al. 2007) and correlates with functional specialization as discussed in section 1.7.

Taken together, these studies suggest that *var* gene subsets have distinct antigenic properties and that some subsets are more antigenically conserved than others.

1.10 Thesis overview

This thesis work explores molecular and predicted epitope diversity in *var* sequences by analyzing the DBLα sequence tags. The DBLα sequence tags used in this study were largely from samples collected in Kilifi during two study periods 2003-2007 and 2008-2010 and also from published sequences from laboratory parasites and clinical isolates from Papua New Guinea and South America. The sequences from Kilifi were sequenced from samples that were collected from sick children who presented at the County hospital. One of the advantages of using the DBLα sequence tags is that they can be PCR amplified from most *var* genes that contain a DBL1α domain. The thesis work focused on the molecular and immunological diversity of DBLα sequences as explained in the next two sections.

1.10.1 Molecular diversity of *var* genes

The first and the second research questions were investigated by exploring the nature of molecular diversity in the DBLα sequences. The molecular diversity of *var* genes is poorly described given the extensive diversity and recombination that is present in var sequences. Recombination presents a challenge in describing the nature of variation and the role of variation in shaping immunological differentiation and influences sensitivity and specificity in methods that are used to measuring var expression. Chapter 3 of this thesis explores this aspect of diversity. Chapter 4 explores the nature of diversity in DBLα nucleotide sequences of known similarity with an aim of understanding and describing the extent to which processes other than recombination are involved in generation of diversity.

1.10.2 Immunological diversity in *var* genes

Majority of *var* genes contain a DBLα domain but raising cross-reactive and protective antibodies against the DBLα domain is difficult. In a recent study, Blomqvist and colleagues reported success in ability to raise cross-reactive antibodies against conserved peptides in the DBLα domain (Blomqvist et al. 2013). In their experiment antibodies were raised against conserved motifs that were identified in sequences associated with severe malaria from a study in Uganda (Normark et al. 2007). In another study 15-20 mer-long peptides with high binding activity and therefore referred to as high activity binding peptides (HABPs) were used to create highly immunogenic and protection inducing modified HAPBs (Patarroyo et al. 2014).

In chapter 5 of the thesis putative novel epitopes are described based on B-cell and T-cell epitope predictions using the DBLα sequences. The epitopes are analyzed in the context of the expression profile of the respective var sequences and the corresponding relationship with age of the host and data on their molecular diversity. Part of the aim of this work was to test the feasibility of raising cross-reactive antibodies against short linear immunodominant epitopes.

1.11 Summary of research questions

1. How does the *var* classification approaches used to measure expression in clinical isolates compare?

2. Is there evidence for non-random distribution of variation in different subsets of *var* sequences given that PfEMP1 functional and antigenic properties potentially impose different selection pressures?

3. Are predicted linear epitopes associated with sequences expressed in severe malaria and in young individuals? How do they differ between previously defined *var* sequence groups?

Chapter 2
Materials and methods

This chapter provides an overview of the study site, the demography and general methods that were used to process and analyze data.

2.1 Study location

This work was carried out using sequence data that was generated from two immunology studies (Warimwe 2010; Andisi 2014) at the KEMRI-Wellcome Trust Research Programme in Kilifi County, published sequences collected from a studies in Brazil (Albrecht et al. 2010) and Papua New Guinea (Barry et al. 2007) and sequences that were generated from sequenced laboratory isolates (Rask et al. 2010).

The KEMRI-Wellcome Trust Research Programme is located at the Kilifi County Hospital (KCH) situated along the Kenyan coast, 56 kilometers northeast of Mombasa by road and lies on the Kilifi creek as shown in figure 2.1 (Scott et al. 2012). Kilifi county has a population of more than 122,889 people based on data from the 2009 National Demographic Survey and Census.

Kilifi region experiences two rainy seasons in a year; the long rains (April-July) and the short rains (October-November). Most malaria transmission and infection occur during these wet seasons. The estimated entomological inoculation rate is about 21.7 bites per person per year (Midega et al. 2012). The northern part of the Kilifi County has a relatively lower transmission rate compared to the south (Snow et al. 1997). The region has observed a decline in malaria transmission over the last 20 years (Okiro et al. 2009b; O'Meara et al. 2008; Hay et al. 2010). Malaria decline has been attributed to multiple interventions that include widespread use of bed-nets, a change in weather patterns, vector densities and a shift in the feeding patterns of the vectors (Mwangangi et al. 2013) among other social-economic factors.

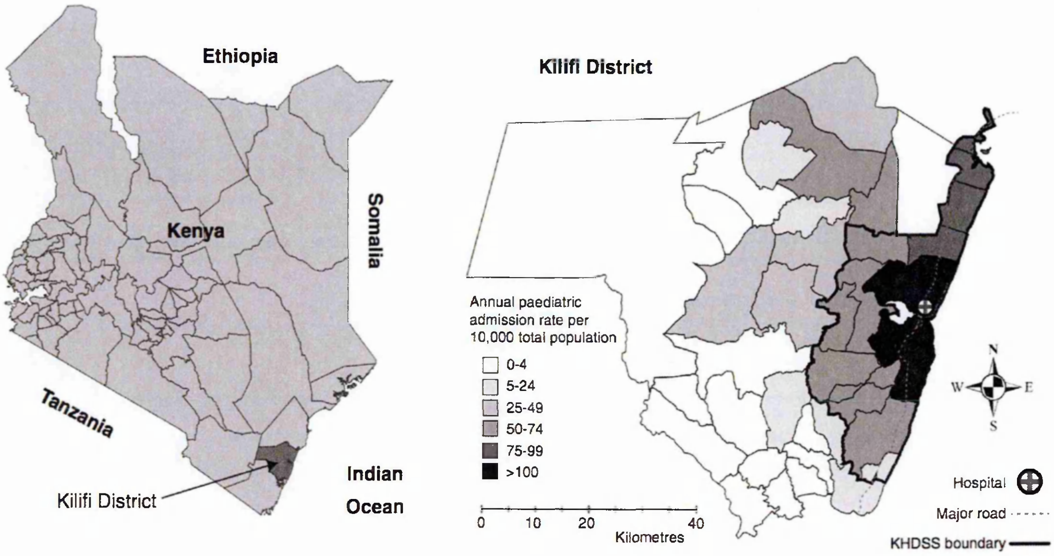


Figure 2.1: *A map of Kilifi district now the Kilifi county. The map on the left shows the location of the county in relation to the rest of the country and the one on the right shows the study area. The rate of pediatric admissions to Kilifi County Hospital by administrative sub-location are colored based on the legend. The dark bold line represents the boundary of the Kilifi Health and Demographic Surveillance System (KHDSS). Courtesy of Scott et al. (2012)*

2.2 Sample collection and processing

The study samples were collected at the Kilifi County Hospital during 2003-2007 and 2008-2010 periods. The period between 2003 and 2007 was a period of rapid decline in malaria admissions in relation to previous years. Malaria admissions were at the lowest during the 2008 to 2010 period as shown in figure 2.2.

An informed consent was obtained at the time of admission. Consent from minors was obtained from the parent or guardian. Parasite samples were classified as severe or non-severe based on clinical information at the time of admission. Severe malaria cases were defined based on Marsh and colleagues case definitions (Marsh et al. 1995).

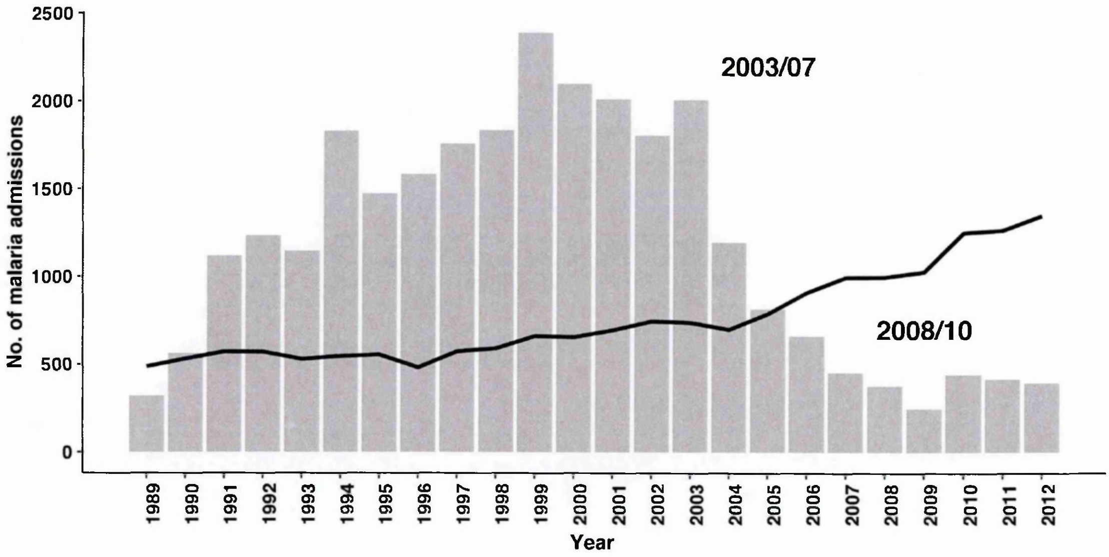


Figure 2.2: *A plot showing the sample collection periods in the background of malaria admission at the Kilifi County Hospital. Sequence data were obtained during the 2003-07 period and 2008-10 period. The 2003-2007 was a period of rapid decline in malaria admissions and during the 2008-10 period malaria admissions were at the lowest. There was a trend towards increasing age for patients who were admitted at the hospital as depicted by the black line (data not shown). Adapted from Andisi (2014)*

The collection and processing of parasites was done prior to sequence analysis using standard and published methods as described in Bull et al. (2005), Warimwe (2010) and Andisi (2014). Sample preparation and parasite culture constituted a separate piece of work which is not included this thesis.

2.3 Sequence data from Kilifi

A total of 8,118 sequences from Kilifi was generated from 450 isolates that were collected in Kilifi in the period 2003-2007. Another dataset was generated from isolates collected in 2008-2010. These sequences were generated using the methods described by Andisi (2014).

2.4 Sequence data from laboratory isolates

Full-length PfEMP1 sequences from seven laboratory isolates were downloaded from sources as described by Rask and colleagues (Rask et al. 2010). These isolates were Dd2 (n=47), 3D7 (n=60), HB3 (n=43), IT4 (n=55), PFCLIN (n=58), Raj116 (n=37) and Preich (n=16) (Otto et al. 2014). DBLα sequence tags were extracted from the full length sequences using the DBLα-finder program.

2.5 The global data set

A total of 354 sequences from 10 geographical regions were downloaded from Genbank and constituted the global dataset as shown in table 2.1.

India	Sudan	Kenya	Cape Verde	Solomon Isl	Phillipines	Vanuaitu	PNG	Thailand	Brazil	Africa
25	22	47	8	90	55	7	22	22	33	23

Table 2.1: *This table shows the number of additional sequences from different geographical regions around the world that were collected and used in the study.*

Data from Papua New Guinea comprised of 460 sequences from Amele and Madange regions that was downloaded from Genbank based on the accession numbers provided by Barry and colleagues (Barry et al. 2007).

An extra dataset of 980 *var2CSA* sequences were also downloaded from Genbank. Only 68 sequence tags met the tag extraction criteria as set out in the DBL-alpha-finder program (A.1). This is not surprising because *var2CSA* sequences cannot be amplified using the DBLα (section A.1) primer target regions which also form definition of the tag.

2.6 DNA amplification

Samples that were collected before 2008 were amplified using Amplitaq Gold polymerase and those that were collected between 2007-2010 were amplified using the Bio-X-ACT DNA polymerase, a high fidelity polymerase with proof-reading activity. Low fidelity enzymes can result in PCR errors that are difficult to account for and bias the distribution of mutations. The effect of PCR errors were minimized using the sequence clustering and selection methods described in later sections. The tag region was amplified in both directions on an ABI Gene Amp PCR system 9700 thermocycler using the DBLα-AF GCACG(A/C)AGTTT(C*/T)GC and DBLα-BR GCCCATTC(G/C)TCGAACCA degenerate primers (Bull et al. 2005).

2.7 Scripting and code management

This thesis was prepared and typeset using Latex and Knitr. Coding and file management was done using the git revision control (`https://git-scm.com/`). Source code was committed to Github repository under a 5 year generous private repository student plan provided by Github team. All source code is provided and released under the open source public plan.

2.8 Scripting languages and statistical tools

Perl, Ruby, Python, Awk and R were used to in the data processing, management and analysis. Notable programs and scripts are listed in appendix A and they are also available from the Github repository (`http://github.com/biorelated`).

Statistical analysis was done using RStudio (`http://rstudio.org`) and the R statistical language (`https://www.r-project.org/`). Majority of the plots were created based on the grammar of graphics as implemented in the ggplot2 package (Wickham 2009). The dplyr package was used to manage data transformation and cleaning. The phylosim and Ape packages were used work with alignment and phylogenetic trees. Layouts were prepared with the gridExtra and Cowplot packages. The Seqinr package and Bioconductor were used in analysis of nucleotide and protein sequences. The Ape package was used for reading and manipulating multiple sequence alignments. Xtable and stargazer packages were used to format tabular data into presentable tables. The work-flows were written using multiple programs written in any of the languages mentioned earlier and they were "glued" together using the Bash programming language and interpreter.

Chapter 3
Development of a sequence clustering and epitope prediction framework

3.1 Background

This chapter describes the methods and approaches that were developed to support the work described in chapter 5 and 6. It describes approaches and tools that were developed to identify high quality DBLα sequence tags, a sequence clustering approach for identifying nucleotide substitutions in groups of similar sequences and methods and tools for predicting both B-cell and T-cell epitopes in DBLα sequences.

Following from chapter 2, there were two main sources of data in this study; immunological studies that were carried out between 2003 and 2007 (Warimwe 2010) and between 2008 and 2010 (Andisi 2014) in Kilifi and sequences that were collected from published sources (Barry et al. 2007; Albrecht et al. 2010; Rask et al. 2010). The sequences from Kilifi formed the bulk of sequences used in the analysis. Data from 2003 - 2007 was downloaded as compressed archives of trace files while the data from 2008 - 2010 was downloaded from the Beijing Genomics Insitute as compressed archive of ABI files and as described in chapter 2. The computer programs described in section 3.7 were developed to unpack the compressed files, process and assemble the raw reads into contigs and then to identify "high quality DBLα sequences" (see section 3.4). The high quality sequences were then clustered, classified using the tag analysis approach described in chapter 1 section 1.7.1 and as described in later sections of this chapter.

3.2 Generating random substitutions

Random substitutions were generated for comparison with substitutions that were observed from clinical sequences. Two approaches were used to generate random substitutions in DBLα sequences. The first approach inserted random substitutions at random nucleotide positions in a randomly selected DBLα reference sequence. A total of 5001 sequences were generated at each identity threshold. An identity threshold was a percent global similarity value that was used to group sequences such that sequences within a cluster were similar at greater than or equal to a specified threshold (see section 3.5). DBLα sequences without stop codons were extracted from the simulated sequences using DBLFinder and clustered at 98%-92% identity using Vsearch. Table 5.1 shows a summary of the DBLα tags at each identity threshold. Lower identities resulted in more substitutions and fewer DBLα sequence tags without stop codons.

In the second approach, substitutions were generated based on a random model and the number of sample sequence alignments was equivalent to the number of clusters at each percent identity. A seed sequence was randomly selected from each actual cluster that contained two or more members to generate random substitutions using the phylosim R package. A neighbour-joining tree from the actual tag sequence data was used as input to the simulation method. The tree from actual sequences ensured that an equivalent number of substitutions were generated in respect to each actual cluster and that the tree topology was maintained. This approach is summarized in figure 3.1.

The main differences in the two approaches was that the first approach used information derived from a single sequence whereas the second approach utilised information from a wide range of sequences and therefore accounted for differences in nucleotide proportions. Therefore the second model was preferred and used in later analysis.

3.3 General approach for computer programs

The computer programs developed in this thesis can be grouped into 1) utility programs, 2) specific programs and 3) work-flows. Utility programs provided fundamental functions such as DNA translation, classification of DBLα sequence tags, slicing and extraction of homology and hyper-variable blocks. Specific programs were developed to provide features that were not available from existing tools or involved novel ways of interacting with the tools for example parsing clustering results, assembling DBLα tags, building networks, mutation analysis as well as plotting results. Work-flows automated repetitive tasks that involved two or more programs to accomplish a task.

Specific programs were packaged for local installation and also released as open source tools to the wider bioinformatics community. The source code for core programs is outlined in the appendix A section and also is available online under open source licenses.

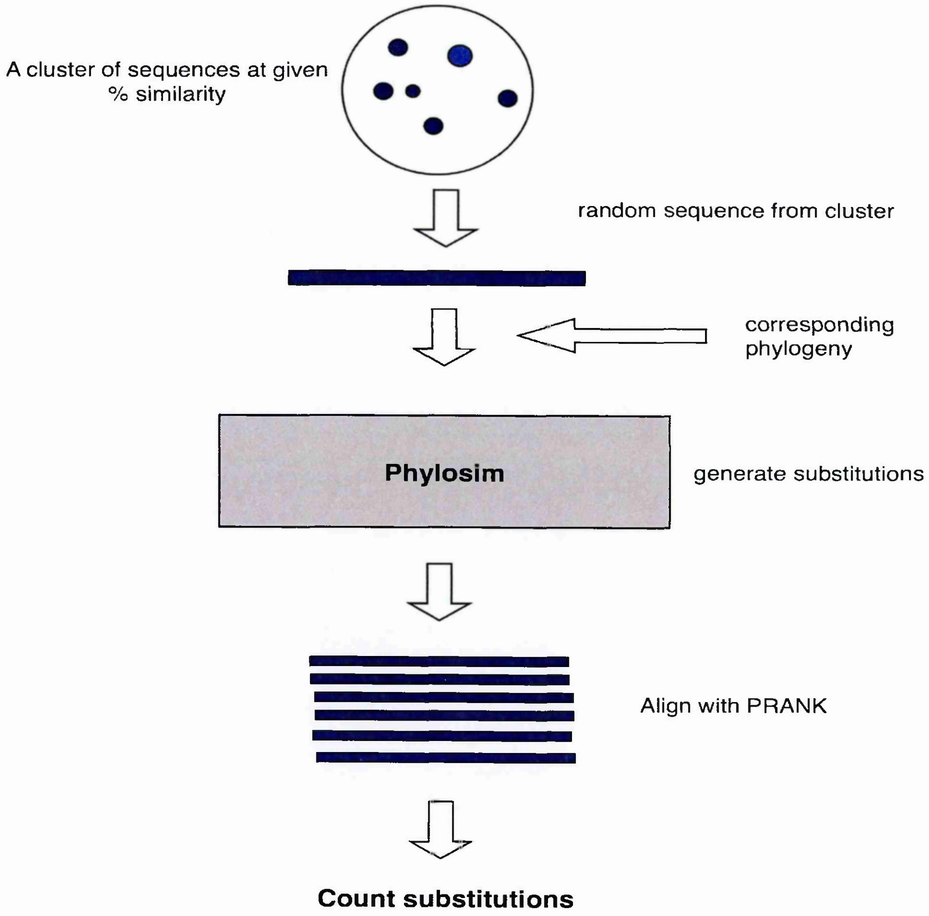


Figure 3.1: *A figure to illustrate the process tha was used to generate random and uniform substitutions using a randomly selected sequence from each cluster that contained more than two members and the respective neighbour-joining tree as input to Phylosim package.*

3.4 Assembling high quality DBLα sequences

A total of 96 colonies per sample were sequenced from data collected between 2003 and 2007 and a total of 32 colonies per sample was sequenced from samples that were collected between 2008 and 2010. Each colony was sequenced in both forward and reverse directions. Two assembly programs were used to assemble the reads, phrap and CAP3.

3.4.1 Assembling DBLα sequences with phrap

Phrap is a program for assembling shot-gun DNA sequences. It was adapted by Bull and colleagues to assemble DBLα sequences based on Sanger capillary sequencing (Bull et al. 2007; Warimwe et al. 2009) and also by other research groups (Blomqvist et al. 2010; Normark et al. 2007) to assemble DBLα sequences in experiments aimed at quantifying the expression of *var* sequences. The assembly process included base quality trimming, vector and primer sequences removal and generation of consensus sequences from reads in each isolate.

A fasta file containing forward and reverse reads with the respective base quality information in a separate file is provided as phrap input. The reads are aligned to produce a collection of high scoring local alignments (Altschul et al. 1990) between pairs of reads, based on a filter for determining which pairs of reads are likely to yield high scoring alignments. Finding every single high scoring alignment is not guaranteed. If two reads have a high-scoring alignment with a region of identity but one read's quality scores are low over this region while in the other they are high, quality scores in the first read are boosted, while in a region of the alignment where there is disagreement, the quality scores of one or both reads are reduced. The reads are then assembled into contigs.

A contig was defined as a set of reads that appear to form an uninterrupted stretch of contiguous bases. First each read is assigned to a contig by itself. Then they are aligned and the alignments are considered from the highest to the lowest scoring one. The contigs containing two aligned reads are merged into a single contig unless the algorithm fails to meet the requirements imposed by the set parameters. The lists of reads and alignments assigned to each contig are used to construct a sequence with quality annotations. After grouping together positions on the individual reads that the alignments suggest have a common origin on the target, a dynamic programming method is used to find the best sequence for each group.

Phrap utilizes local alignment to align and group sequence reads. A possible shortcoming with local alignment is that it can assign tags from the same isolate into a consensus if they share substantial local identities and create contigs from sequences that are globally different. One study showed that phrap produces more errors in consensus sequences compared to CAP3 (Huang and Madan 1999). The phrap-based pipeline as was implemented by Bull and Keane, was more suitable for expression analysis and yielded the protein sequences. It was important to assemble tag sequences based on global similarities and to yield nucleotide sequences as output.

3.4.2 CAP3 approach

CAP3 was used to assemble 192 reads per isolate from the dataset that was collected in 2003-2007 period and 64 reads per isolate for the 2008-2010 dataset. Data from each read was in the form of an Ab1 chromatogram file. The reads were screened to remove read-pairs that did not have a corresponding forward or reverse mate. Each chromatogram was converted into a fastq file which was later split into a fasta file and a phred formatted quality files. The respective fasta and quality file were then provided as CAP3 input files.

CAP3 assembled each read-pair into a consensus contig. For each CAP3 read-pair assembly an ace, contig and a quality file were generated. The particulars of each successive CAP3 execution were saved to a log file. The contig file was then renamed for consistency with the assigned colony and isolate name as was provided in each plate. The respective contig and base quality files were converted into fastq format. All the contigs belonging to each isolate were saved in a file. The DBLα sequence tags were extracted using the DBLα-finder (3.7.1).

The final output was a DBLα sequence tag for each sequenced colony per isolate. All the tag sequences in each isolate were clustered at 96% identity using the Vsearch version v1.1.1_osx_x86_64 clustering program (`https://github.com/torognes/vsearch`). Vsearch is an open source 64-bit drop-in replacement for Usearch. Figure 3.2 shows a summary of this work-flow.

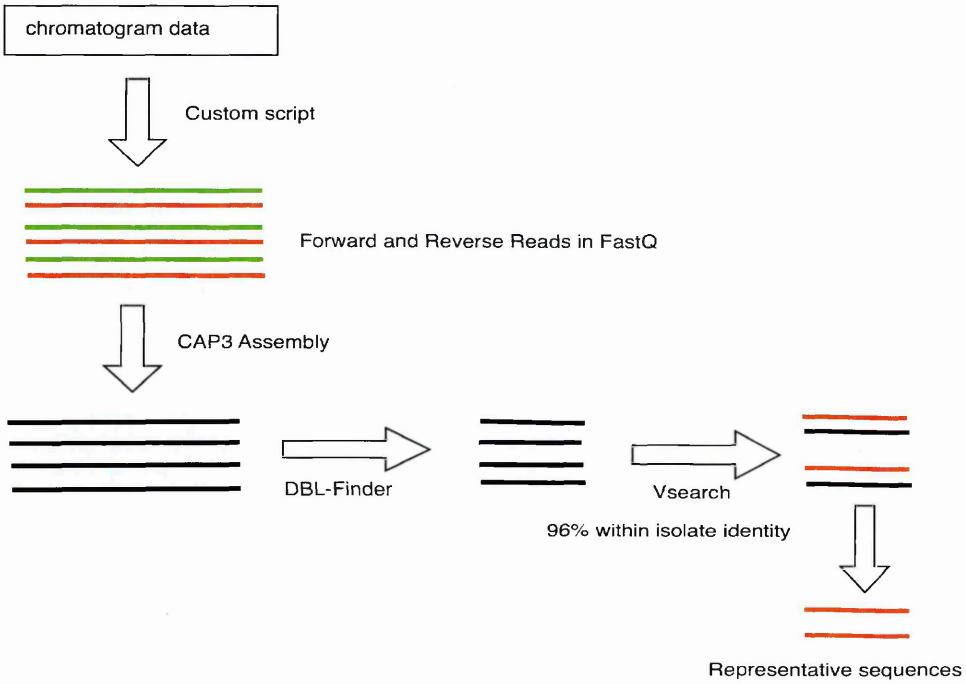


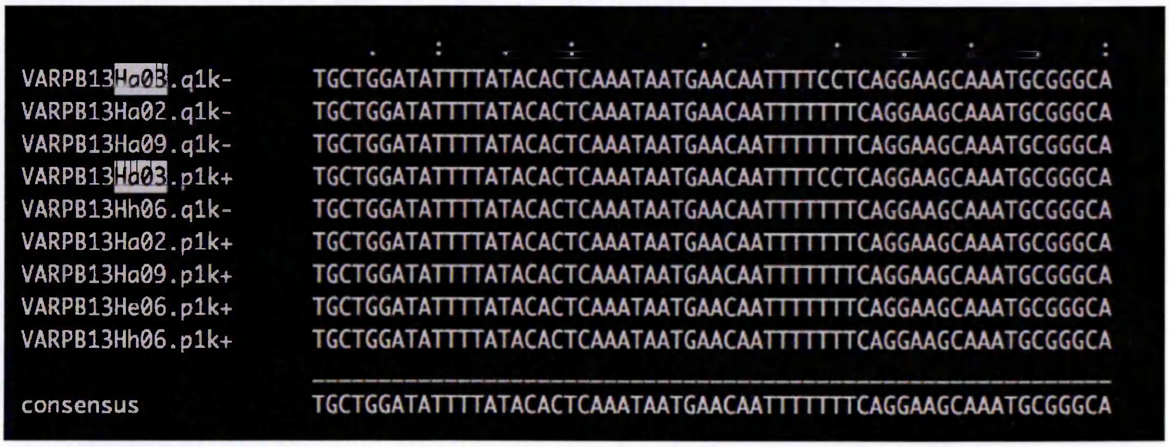
Figure 3.2: *A flow diagram showing the DBLα sequence assembly workflow. The chromatogram data files were converted to raw fastq reads. The reads from each colony were collapsed into "contigs" using CAP3. DBLα sequence tags were extracted from the contigs using the DBL-Finder utility and sequences from each isolate were clustered at 96% identity using Vsearch. Representative sequences from each isolate were written to file for further processing.*

Given that an isolate refers to a lab adapted clinical sample collected from an individual patient at particular time point and 1) individuals can be infected by more than one parasite genotype and 2) each genotype can express more than one variant at any one time because of antigenic variation, a clinical sample can contain RNA from multiple distinct *var* gene variants.

Initially all the reads from a given isolate were provided as input. CAP3 generates consensus sequences from the input reads based on an alignment whose accuracy is dependent on the performance of the multiple alignment method. If the bases of overlapping reads are improperly aligned, conflicting bases may occur in some columns and an inaccurate consensus sequence is generated (Huang and Madan 1999).

Figure 3.3 shows an alignment of high quality reads from isolate 13H. Read-pair Ha03 contained two consecutive cytosine residues that are not reflected on the consensus contig. In this case, majority of the reads contain thiamine residues at the respective column and the sum of quality scores in the majority of the reads is much higher than from the two colonies. In this case, providing all the reads from an isolate in a single step as input to CAP3 could mask actual variants because the reads from a sample may contain multiple variants as explained earlier.

To resolve this, CAP3 was used to generate consensus sequences from each read-pair and the corresponding consensus sequences from each isolate were pooled together and clustered using an arbitrary 96% identity cutoff to remove errors as described in section 3.5.



```
                        .    :         :         :         :         :         :
VARPB13Ha03.q1k-     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTCCTCAGGAAGCAAATGCGGGCA
VARPB13Ha02.q1k-     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
VARPB13Ha09.q1k-     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
VARPB13Ha03.p1k+     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTCCTCAGGAAGCAAATGCGGGCA
VARPB13Hh06.q1k-     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
VARPB13Ha02.p1k+     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
VARPB13Ha09.p1k+     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
VARPB13He06.p1k+     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
VARPB13Hh06.p1k+     TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
                     ____________________________________________________________
consensus            TGCTGGATATTTTATACACTCAAATAATGAACAATTTTTTTCAGGAAGCAAATGCGGGCA
```

Figure 3.3: *Errors in CAP3 consensus sequences. In this figure, the consensus region containing the* `T` *homo-polymer is assigned a* `T` *at the 5th and 6th position despite the fact that the reverse and forward reads in colony Ha03 contains a* `C` *at these positions. To mitigate against such errors, read-pairs from each isolate were provided as input to CAP3. The output contigs were clustered at within isolate level using a 96% global similarity cutoff.*

An assembly of reads from isolate 14G, generated two long contigs, contig 3 and contig 6. Both contigs contained more than one DBLα tag as shown in figure 3.4.

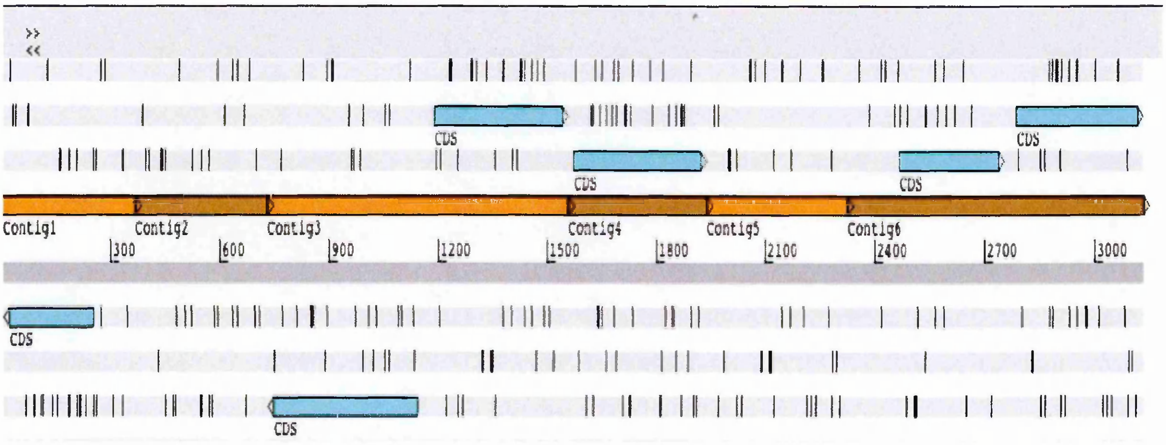


Figure 3.4: *An example of contigs containing multiple DBLα tags. Contig3 contained two DBLα tags in opposite orientation and contig 6 contained two DBLα tags in the same orientation.*

Most of the reads were assembled into contigs and high quality DBLα sequences were extracted as shown in figure 3.5. The black bars represent the consensus sequences and the red bars represent the number of high quality sequence tags from the respective contigs in each sample. High quality sequence tags were defined as sequence tags that did not contain stop codons, had a length of between 300 and 500 bases long, and passed the conservative quality trimming criteria that discarded any read if a nucleotide had a base quality less than 20.

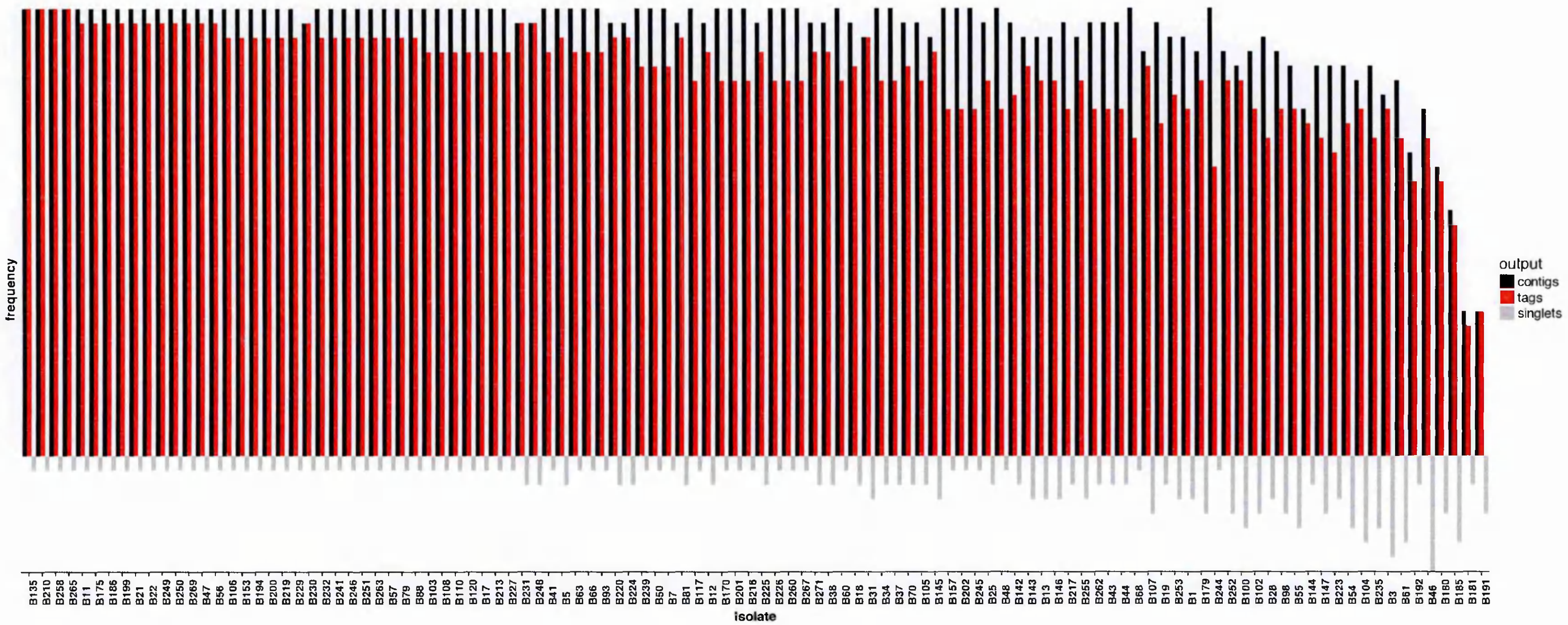


Figure 3.5: *A summary of "assembled" DBLα sequence tags using CAP3 that were collected in Kilifi between 2008 and 2010. The isolate name is indicated on the x-axis and the frequency of the contigs, the tags and singlets are represented by the y-axis. The black bars represents the number of the consensus contigs in each isolate. The red bars represent the number of DBLα sequence tags that did not contain a stop codon and has a phred score >20. The grey bars indicate singlets. A singlet was defined as a sequence that did not have a corresponding forward or reverse read, or where one of the reads was significantly different from the corresponding forward or reverse read. Singlets were discarded from the final list of high quality sequences.*

3.5 Defining clusters of similar high quality sequences

CAP3 only provided an approach for merging forward and reverse reads and taking the base quality information into consideration. It provided a robust approach for defining DBLα sequence contigs per colony per isolate. It did not however take care of PCR errors or minimize their effect. The next sections describe the approaches that were adopted to define high quality tag sequences and to minimize the effect of sequencing and PCR errors in the final batch of tags.

3.5.1 Reducing noise due to potential PCR errors

Base mis-incorporation during PCR leads to substitution bias (Acinas et al. 2005). PCR amplifications in the 2007-2010 dataset used the Bio-X-ACT high fidelity polymerase for amplifications. To reduce the number of unspecific substitutions that could have arisen due to PCR errors, all the sequences were clustered using a global identity cut-off of 96% within each isolate. Only the representative sequences from each cluster in each isolate were considered in between isolate comparisons.

Sequence identity depends largely on how gaps are treated given a local or global alignment. Having established a set of high quality assembled sequences, sequence identity was defined as the proportion of matching residues for a given set of sequences based on global pairwise comparison. Letters were identical if they matched at the same position in an alignment. The two widely used clustering tools CD-HIT (Li and Godzik 2006; Fu et al. 2012) and Usearch (Edgar 2010) as implemented in Vsearch were evaluated.

3.5.2 CD-HIT clustering

CD-HIT is a greedy incremental fast sequence clustering algorithm using global pairwise comparisons. It uses a word filtering algorithm that reduces the number of comparisons in each character along a pair of sequences.

The sequences are first sorted based on decreasing length. The longest sequence is considered as the representative sequence in the first cluster, each remaining sequence is compared to the representative sequence in the existing clusters. If similarity with any representative sequence is above a given threshold, that sequence is grouped into that cluster, otherwise a new cluster is created and the particular sequence is designated as the representative sequence.

The word filtering approach is based on the idea that two sequences with a certain sequence identity must have at least a specific number of identical di-peptides and tri-peptides. For example, for two sequences to have 85% identity over a 100-residue window they have to have at least 70 identical di-peptides, 55 identical tri-peptides, and 25 identical penta-peptides. Based on the short word requirement, CD-HIT skips most pairwise alignments because it assumes that the similarity of two sequences is below certain threshold based on the simple word counts.

Unfortunately, short word filtering is limited to specific cluster identity thresholds. If the mismatches are evenly distributed along an alignment, the numbers of common short words are minimal which results in poor clusters.

The greedy incremental algorithm has a known issue. This can be summarized as follows: if there are two clusters, say the first cluster has 3 sequences A, X and

Y where A is the representative sequence, and the second cluster has 2 sequences B and Z where B is the representative, if sequence Y is more similar to sequence B than to A, it is placed in the first cluster, simply because Y found A first during clustering process.

Occasionally, CD-HIT gives lower rather than higher identities due to misalignment caused by banding errors. The banding errors occur if the difference in length in a segment between two conserved regions exceeds the band width (as defined by CD-HIT) and therefore one or both of the conserved regions are misaligned. Furthermore, CD-HIT gap open penalty is lower than that of BLAST or Usearch. Therefore, CD-HIT alignments are more unreliable because they contain more gaps than BLAST and Usearch.

CD-hit defines the percent identity based on equation 3.1

$$\frac{number\ of\ matches}{length\ of\ shorter\ sequence} \tag{3.1}$$

With CD-HIT gaps in the longer sequence can result in bias in percent identity but gaps in the shorter sequence do not. The percent identity as defined by CD-HIT does not correspond with evolutionary distance of the sequences. CD-HIT clustering was used to cluster sequences in the earlier stages of this work.

3.5.3 Vsearch clustering

The bulk of sequence clustering was performed using Vsearch, an open source implementation of the Usearch algorithm and tool. Usearch is a fast and sensitive sequence clustering tool (Edgar 2010) that can cluster sequences at low percent identity. Given a percent identity threshold t, Usearch attempts to find sequence clusters that fulfill these two criteria,

i. All centroids have similarity $< t$ to each other and

ii. All member sequences have similarity $>= t$ to a centroid.

Most clusters satisfy the two criteria but the first criteria is not always guaranteed. If a sequence matches two different centroids with identity $> t$, it is assigned to the closest centroid. But if there are two or more centroids at the same distance,an arbitrary cluster assignment choice is made for the particular sequence.

The DBLα tag sequences were sorted by length and saved into a separate file. The sorted file was provided as input to the Vsearch program and thereafter sequences were processed in the order they appear in the input file using the `cluster-smallmem` command. Sorting sequences by length ensured that appropriate centroids were designated to the longer sequences. If the next sequence in the file matched an existing centroid, it was assigned to that cluster, otherwise it became the centroid for a new cluster. Clustering was done at using 98 - 88% identity cutoffs.

Sequence identity was defined as the number of matching residues in a sequence divided by the number of columns. This definition is based on the BLAST definition of sequence identity and is calculated as shown equation 3.2.

```
A T C G A A A A
| * * * | | | |
A C T A A A A A
```

For example, the two sequences shown above have a total of 8 bases of which 5 are identical. From equation 3.2 the percent identity is 62.5%.

Identity is defined as shown in equation 3.2

$$identity = \frac{number\ of\ matches}{number\ of\ columns} \tag{3.2}$$

Figure 3.6 shows a summary of the clustering work-flow. The input was a set of reads and the output was a set of clusters at each identity threshold. Each cluster comprised of an alignment from reads that were similar at the respective identity. These alignments were used to determine the location and frequency of mismatches.

3.6 Aligning and finding mismatches

Sequences in each cluster were aligned using PRANK (Löytynoja and Goldman 2005), a sequence alignment tool that aims to produce correct and accurate alignments relative to other alignment tools. The number and location of mismatches was obtained from PRANK alignments using the DBLMutationFinder (`https://github.com/biorelated/find-dbla-mutations`) described in section 3.7.5.

A mismatch was defined as a base inconsistency between a sequence and the most common residue in a particular position in an alignment of sequences. The mismatch finding script produced a tab delimited file that contained, the mismatch position, the type of mismatch, the codon position and the context of the mismatch which on a linear sequence was defined as 5 bases in the 5' direction and 5 bases in the 3' direction relative to the mismatch position. All mismatches at each identity threshold were pooled together into a single file.

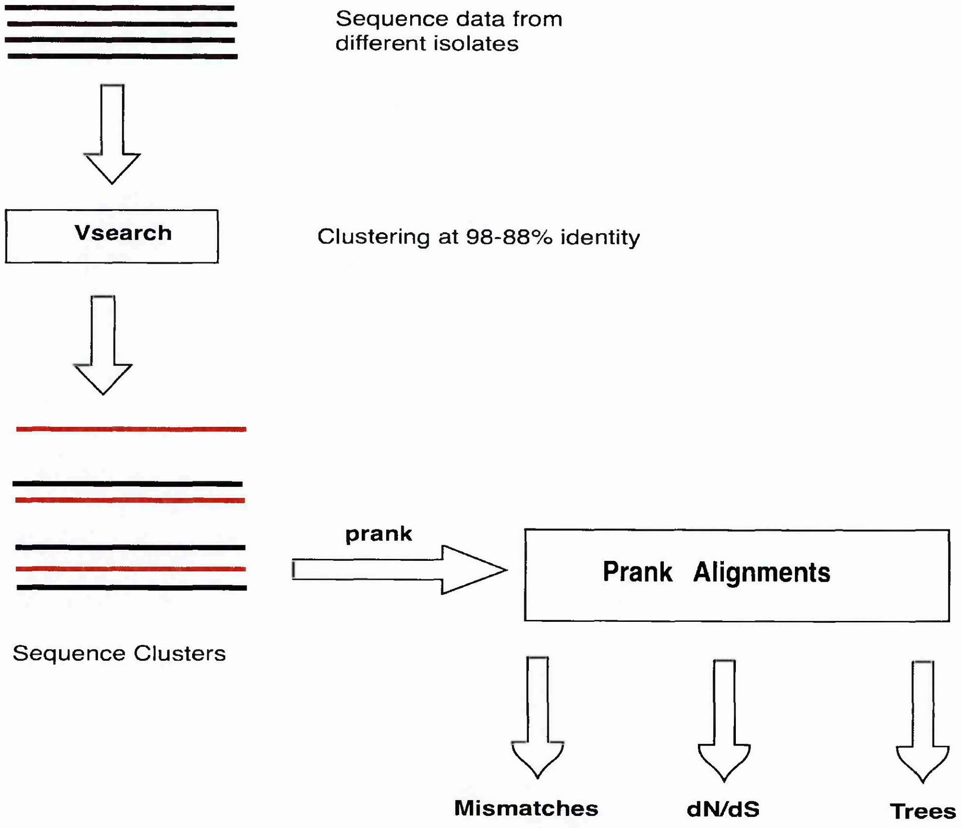

Sequence data from different isolates
Vsearch
Clustering at 98-88% identity
Sequence Clusters
prank
Prank Alignments
Mismatches
dN/dS
Trees

Figure 3.6: *Sequences from different isolates were clustered at different percent identities. Sequences from each cluster were aligned using PRANK and mismatches identified in each alignment using the polymorphisms finder script. The tree output files were later used to generate random sequences using phylosim package. Mismatches were tabulated based on the frequency at different codon positions.*

3.7 Computer programs

The computer programs described in this section were developed to support the analysis work discussed in the previous sections and also in chapter 4, 5 and 6. The full source code is listed in appendix (A) under their respective subsections. The source code was also released online at Github (`www.github.com/biorelated`).

3.7.1 DBLα finder

Online: https://github.com/georgeG/bioruby-dbla-finder

The DBLα-finder (see listing A.1) was developed as a Bioruby plugin (Bonnal et al. 2012) that extends the `Bio::Sequence::NA` class and takes advantage of existing sequence manipulation methods available in the Bioruby sequence class. The main aim of this program is to extract DBLα-like sequences from DBLα assembled contigs sequences. It can also extract the tags from full-length var sequences.

The program takes sequences in fasta and fastq formats as input. The input sequence is translated in all reading frames. The frames containing the least number of stop codons are flagged. For each open-reading-frame, the program searches for the **DIGDI** two amino-acid degenerate motif at the N-terminal and the **PQYLR** degenerate motif at the C-terminal. This step uses a hamming distance metric to generate the degenerate motifs. The hamming distance was preferred because it measures the minimum number substitutions required to change one string into the other and therefore it captures majority of permutations of these motifs. In the strict mode no stop codons in the tag sequences are allowed and the relaxed mode can allow a specified number of stop codons. The extracted sequence tags is orientated in the first reading frame.

The program uses a brute force approach and evaluates all the reading-frames even when they do not contain DBLα sequences. This has impact on the run-time and memory particularly when working with long contigs. For the purpose of this work this was not an issue since the var contigs under consideration were relatively short.

3.7.2 DBLα-Classifier

Availability: https://github.com/georgeG/bioruby-dbla-classifier

DBLα-classifier is a Cys/POLV classification program that was developed based on a Perl classification script by Bull et al. (2007). The program took advantage of existing sequence manipulation methods in Bioruby to add DBLα specific functionality to the `Bio::Sequence::NA` and `Bio::Sequence::AA` classes. It implements more features than the Perl script some of which includes,

- Classify DBLα sequence tags
- Identify the position specific polymorphic blocks (PSPB)
- Identify and print positions of limited variability (POLV) regions
- Identify *var1* and *sig2* like sequences based on sequence motifs that define these sequences.
- Identify group A-like sequence tags
- Identify homology blocks D, F and H given the location of PSPBs, PoLVs, WW and VW motifs.

3.7.3 Bio-CD-HIT-Report

Availability: https://github.com/georgeG/bioruby-cd-hit-report

Bio-CD-HIT-Report was developed to parse CD-Hit cluster files using a convenient and consistent interface. It was developed as a library with a standard interface to call methods for processing clusters and to plug in to the sequence clustering workflow. Given an input cluster file, Bio-CD-HIT-Report can generate a report containing the following information,

- The number of sequences in each cluster.
- A list of sequence names in each cluster.
- A list of representative sequences in each cluster.

Bio-CD-HIT-Report made it easier to process multiple cluster files by embedding its method calls in existing scripts.

3.7.4 Block-sharing networks

Availability: `https://github.com/georgeG/block-sharing-networks`

This is a command-line application that depends on the DBLα-classifier to generate group-sharing networks for DBLα sequences. It returns a list of nodes if sequences share PSPB blocks. The source code is available on listing A.24 of the appendix.

3.7.5 Find-DBLα-mutations

Availability: `https://github.com/biorelated/find-dbla-mutations`

This library was developed to explore polymorphisms in the DBL tag sequences. It contains two main classes, the `Mutation` class and the `Selection` class. The mutation class provides methods for locating and describing polymorphisms given a fasta formatted file of DNA sequences. The selection class provides methods for working with multiple sequence alignments in fasta format. This class is dependent on the Bio-alignment library . Bio-alignment implements a functional approach in parsing alignment files.

3.7.6 DBLα-assembly

DBLα-assembly was initially developed to assemble DBL sequence tags. It implemented functions to 1) read and parse DBL sequences from CAF files and 2) to extract and cluster sequences based on sequence length and sequence identity.

It groups the sequences by length, then it creates an identity matrix from pairwise comparisons. Sequences are clustered based on the pairwise comparisons. The clustering does not account for insertions or deletions and therefore it is only suitable for clustering uniform length sequences.

Given that there are more optimized and published tools for performing sequence clustering, the clustering module was deprecated in favour of CD-HIT and Vsearch clustering approaches.

3.7.7 DBLα CAF Parser

Availability: `https://github.com/biorelated/dbla-caf-parser`

DBLα CAF Parser is a library that was written to provide a functionality for extracting DBLα sequences and sequencing information stored as CAF files.

3.8 Epitope prediction

B-cell and MHC class-II T cell epitopes prediction methods were used as described in section 3.8.1 and section 3.8.4. A total of 100 randomly selected sequences from

Kenya, Papua New Guinea, Brazil were used in B-cell epitope prediction using Bepipred and a total of 365 non-redundant sequences were used for MHC-class II epitope predictions using NetMHCII-pan and the predivac server.

3.8.1 B-cell epitope prediction

Bepipred (Larsen et al. 2006), a B-cell epitope prediction method was used to explore potential B-cell epitopes in DBLα sequences. Bepipred is a combination of Parker propensity scale (Parker et al. 1986) and a hidden markov model that was trained using the Pellequer (Pellequer et al. 1993) and AntiJen (McSparron et al. 2003) epitope data set. It predicts linear B-cell epitopes from amino acid sequences.

A propensity scale method assigns a score to each amino acid in a poly-peptide and applies a running window average to the raw scores. Other notable propensity scale methods other than Parker's scale are Chou and Fasman, Levitt's secondary structure scale (Levitt 1978) and Emini, an accessibility scale (Emini et al. 1985). Many propensity scale methods have poor prediction and only marginally better than a random model according to a study by Blythe (Blythe and Flower 2005). Although bepipred has a low sensitivity, it is more accurate than the propensity methods and it is an improvement to the random model.

Bepipred assigned a prediction score to each amino acid in each input sequence. The default cutoff score was 0.35 which corresponded to a sensitivity of 0.49 and a specificity of 0.75. A mean of the raw score was calculated using a 10 amino acid sliding window approach over the prediction space.

3.8.2 Shannon diversity index

In this work, Shannon's diversity is an index of residue conservation and it was calculated with equation 3.3. The values were calculated in each alignment column and averaged over a running window of 10 amino acids.

$$H_i = -\sum p_{a,i} log_2 p_{a,i} \tag{3.3}$$

where $p_{a,i}$ is the frequency of the amino-acid a at column i.

3.8.3 Hydrophobicity profile

Kyte-Doolittle (Kyte and Doolittle 1982) hydropathy scores were calculated in specified sequences. The scores was averaged using a 10 amino-acid window in the respective sequence.

3.8.4 Prediction of T-cell epitopes

Studies using recombinant peptides from laboratory lines (Allsopp et al. 2002; Sanni et al. 2002) suggest that individuals in endemic regions elicit interferon-γ and interleukin-10. An analysis of CD4$^+$ T-cell responses to DBLα tags to which a child is exposed to was reported to be associated with protection from future malaria episodes suggesting a role for CD4$^+$ T-cells in protection (Gitau et al. 2012b; Gitau et al. 2014). Given the extreme sequence diversity in the DBLα tag and based on the idea that CD4$^+$ T-cells are important for a functional B-cell response.

Only a few studies have found associations between the MHC class II immune responses and *P. falciparum* surface antigens (Troye-Blomberg et al. 1991). The failure to detect associations can be attributed to contribution of non-MHC induced factors and the lack of epitope specificity given the sequence diversity in surface antigens. One study explored HLA class II restriction among a group of 2000 individuals in Gambia and reported that DRB1*1302 - DQB1*0501 are associated with reduced susceptibility to severe malaria. This suggested that MHC polymorphism is maintained by altering susceptibility to infectious pathogens(Hill et al. 1991). Nonetheless, more work is require to explore the possibility of MHC restriction against immuno-dominant CD4$^+$ epitopes in PfEMP1.

The major histo-compatibility molecules (MHC) bind short peptide molecules. The MHC-peptide is presented to the T-cell receptors each of which can only bind particular MHC-peptide combinations. The first studies to report MHC class II association with immune responses used small panels of antigens derived from three *P. falciparum* surface proteins, namely the merozoite surface protein, the circumsporozoite proteins and Pf155/RESA proteins (Riley et al. 1992).

T-cell epitopes were predicted using NetMHCIIpan-3 (Karosiene et al. 2013; Nielsen et al. 2010) and Predivac (Oyarzún et al. 2013). NetMHCII-pan is a pan-specific epitope prediction tool that uses hidden markov models and machine learning to predict MHC class-II T-cell epitopes from amino acid sequences. NetMHCIIpan-3 is based on artificial neural networks that were trained on more than 50,000 quantitative peptide-binding measurements covering HLA-DR, HLA-DP, HLA-DQ and two murine molecules (Karosiene et al. 2013).

Predivac is based on an algorithm that identifies specificity determining residues

(SDRs) (Oyarzún et al. 2013) that identifies the residues involved in forming a stable peptide-MHC complex. Specifity Determining Residues (SDRs) are a small set of structurally conserved positions in the peptide-binding interaction interface that are responsible for specific recognition events. The approach was first described for substrate specificity prediction of protein kinases (Ellis and Kobe 2011; Kobe and Bodén 2012; Saunders et al. 2008; Saunders and Kobe 2008). Predivac works by establishing a correlation between the SDRs in the HLA class II query protein and the SDRs associated with HLA proteins of known specificity. The process involves the following steps:

i. SDRs for each binding position are identified in the query HLA class II protein sequence.

ii. PredivacDB, a purposed-built database of SDRs and high affinity binding data (IC_{50} < 50nM, which is a measure of the effectiveness of binding), is queried and amino acid frequencies and weights are calculated for peptide sequences associated with allotypes sharing similar SDRs as the query protein at each binding position.

iii. A position-specific scoring matrix (PSSM) is built based on the binding data. T-cell epitope mapping is carried out by parsing query protein sequences into overlapping peptides, each of which is assigned a binding score using the PSSM.

The outcome of Predivac is a relative score between 0 and 100. For a given HLA class II protein, a peptide is a "better" candidate to be a MHC class II high-affinity binder, and therefore a CD4+ T cell epitope, if it scores higher than another peptide from a given protein. Given that, most of the immune response in protein-based vaccination is mounted against a few dominant epitopes, despite the presence of many potential epitopes within a given antigen. Vaccine formulations built on epitopes that do not dominate the immune response do not induce effective protection in the vaccinated organism, therefore; it is important to identify immunodominant epitopes.

The peptide-MHC kinetic stability is important in controlling MHC class II peptide's immunogenicity. Predivac was developed on the assumption that positive bias in capturing peptide features correlates with promiscuity and immunodominance (Oyarzún et al. 2013).

Predivac web-server could accept one sequence per MHC allele at the time of analysis. Predivac-util (listing A.22), a python library was developed to submit multiple sequences for multiple MHC class II alleles. Promiscuous epitopes were defined as peptides that bound strongly or weakly to multiple MHC class II alleles.

3.8.5 Predivac-util

The Predivac online server allows a user to submit one sequence and one allele per request. Given hundreds of sequences and over 20 HLA alleles, it would have taken a lot of repetitive and time consuming effort to get prediction results using the Predivac web interface.

Predivac-util (listing A.22) was developed to automate bulk sequence submission to Predivac server. This command line utility was written in Python and is run with two arguments. The first argument is a list of sequences in fasta format and the second argument is a list of HLA DR alleles, one allele per line. A three seconds delay between submissions is enforced to avoid overwhelming the server with multiple requests. This delay can be adjusted upwards or eliminated. Upon a successful run, Predivac-util produces a tab delimited file containing the predictions. Predivac-util relies on an uninterrupted Internet connection. It does not have the capacity to resume jobs if the connection fails during a run. Future versions may fix the issue.

Chapter 4
Evaluation of classification approaches used to measure expression of *var* genes in clinical isolates

4.1 Background

As discussed in chapter 1 section 1.7, *var* classification approaches depend on the conserved upstream promoter region, the coding regions (Lavstsen et al. 2003; Smith et al. 2000a), the semi-conserved DBLα region of the DBL1α domain (Bull et al. 2007) and the organization of the domain architecture in full-length sequences (Rask et al. 2010) to capture putative functional attributes used in grouping *var* genes.

The classification approaches are evaluated on ability to discriminate between severe and non-severe malaria and ability to discriminate severe malaria syndromes such as impaired consciousness, respiratory distress, anaemia and cerebral malaria. The extent to which the *var* sequence grouping approaches are consistent with each other and whether they measure the same functional properties has not been adequately addressed.

This chapter compares the Cys/POLV (Bull et al. 2007), domain cassettes (Rask et al. 2010) and the classification based on the 5' upstream promoter in classifying *var* genes. The results suggest that, functional correspondence between the DBLα domains, the domain cassettes (DC) and the Cys/PoLV (CP) tag classification is not consistent. This may partly be explained by the fact that each classification method captures limited information about a *var* sequence. Nonetheless and despite a long history of recombination DBLα tag sequences provide information on features of full-length *var* genes. This may be particularly the case within specific geographical regions.

Recently a lot of work has focused on the domain classification partly because these classifications are derived from full length sequences and more importantly, specific domain arrangements for example DC8 and DC13, have been shown to be associated with severe malaria (Lavstsen et al. 2012). Furthermore there is evidence that DC8 and DC13 domains are associated with specific binding phenotypes (Turner et al. 2013). (see chapter 1 section 1.7 for details). For a recap, the main PfEMP1 classification approaches are summarized in table 4.1

Name	Description	Reference
Ups classification	Based on the conservation of upstream promoter regions	Smith 2000,Lavstsen 2003
Cys/POLV	Based on the number of cysteines and amino acid motifs in the DBLαsequence tag	Bull 2007
Domain Cassete	Based on domain arrangment of full length PfEMP1 sequences	Rask 2010

Table 4.1: *A summary of the main PfEMP1 classification approaches.*

4.2 Methods

4.2.1 Data collection and sequence classification

DBLα sequence tags were extracted from 403 full-length *var* sequences in seven genomes from a study of *var* sequence diversity and classification of PfEMP1 sequences (Rask et al. 2010). The dataset comprised sequences from the 3D7, IT4, HB3, DD2, RAJ116, IGH-CR14, and the PFCLIN isolates. The sequence tags were classified based on the Cys/POLV approach (Bull et al. 2007) and information on each upstream promoter region and domain cassette was derived from the data set.

A total of 314 sequences were used in the analysis after removing *var2csa* sequences and those that did not have upsA classification data available. The *var2csa* sequences were excluded because they are associated with expression in pregnant women and they cannot be amplified using the DBLα universal primers. Sequences whose 5' upstream promoter regions were missing were removed from the analysis. The Cys/POLV classification used sequences from 3D7 and IT4 to define the block-sharing groups 1 and 2 (Bull et al. 2008) (see chapter 1 section 1.7.1). The block sharing groups were defined based on a global collection of sequences that included sequences from 3D7 and IT4 laboratory isolates. In this analysis, sequences from 3D7 and IT4 isolates were excluded because they comprised the blocks-sharing groups definition and they create bias in the analysis.

4.2.2 Sequence alignment

Sequence alignments were performed using Muscle, a fast and accurate protein alignment program (Edgar 2004).

4.2.3 Defining block-sharing groups of sequences using a network approach

A total of 1,317 published DBLα sequences was obtained from Kilifi (Bull et al. 2008) and from published parasite genomes (Rask et al. 2010) together with three DC8 sequences from a study that was conducted in Tanzania (Lavstsen et al. 2012).

Sequences that shared 10 amino acid sequence blocks were identified and used to draw a network of sequences that shared common sequences herein referred to as a block-sharing network from the laboratory isolates. The block-sharing networks were visualized using Cytoscape (Shannon et al. 2003) an open source platform for visualizing molecular interaction networks. Similar to the Kilifi network (Bull et al. 2008) the network from the laboratory isolates formed two large connected components, a major component and a minor component as shown in figure 4.3.

4.3 Results and discussion

Figure 4.1 shows the association between *ups* classification and Cys/PoLV groups based on 33 DBLα sub-domains from seven laboratory isolates. This figure was consistent with previous comparisons of Cys/POLV and *ups* classifications (Bull et al. 2007; Bull et al. 2005). While the majority of sequences containing two cysteines were confined within sequences containing the *upsA* promoter, a number of them was also found in sequences that are associated with *upsB* and *upsC* promoters. Sequences with the DBLα-0.3 domain were exclusively from the *upsB* group and contained proportionately higher number of sequences with two cysteines. The DBLα-0.3 is associated with sequences with the DC13 domain architecture.

Sequences with DBLα-2 domain were exclusively from *upsB* and contained both 2 and 4 cysteines. They also appear to contain sequences from both group sharing block 1 and 2. DBLα-2 sequences tend to belong to domain cassette 8 (DC8) (explained further in chapter 1 section 1.7). DBLα-2 domain is associated with sequences with a DC8 domain architecture.

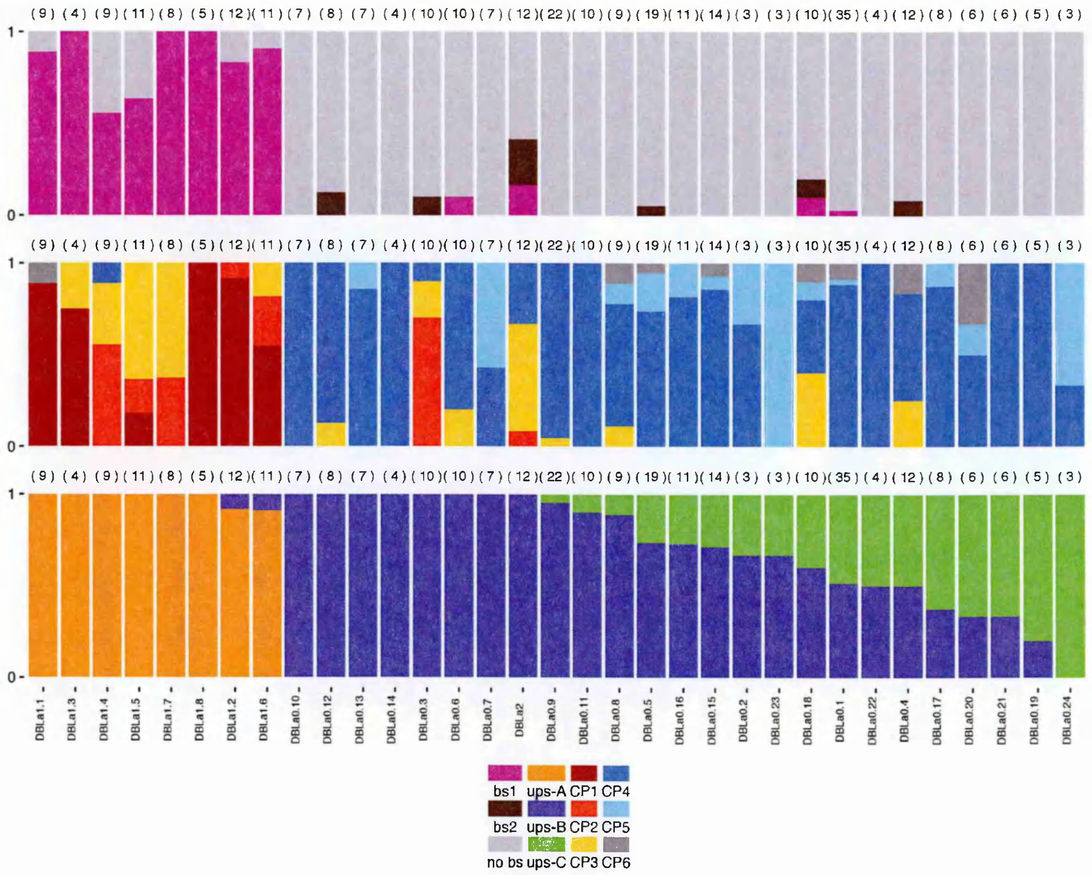


Figure 4.1: *The correspondence between ups and Cys/POLV classification based on the DBLα sub-domains. The upstream promoter region classification is on the bottom panel, Cys/PoLV (CP) in the middle panel and block-sharing groups on the upper panel. The domains are arranged from left to right in order of decreasing proportion of sequences with upstream promoter A to C. The total number of sequences from each domain is shown at the top of the bars. Sequences with 2 cysteines and block sharing group 1 are largely upsA and contain distinct DBLα sub-domains.*

The domain classification that were suggested by Rask et al. (2010) study were partly based on local sequence alignments. Applying sequence alignment to a global collection of recombining *var* sequences is challenging because the alignment process does not take into account the recombination history. It increases the likelihood of bringing together genetically distinct sequences with short local sequence similarities and depending on the length of the sequences relative to the full length molecule, it is also possible to define sequences as distinct while they are actually part of the same group of recombining sequences.

Cys/POLV classification studies suggest that the **MFK** and the **REY** motifs are mutually exclusive, which implies that sequences containing MFK or REY motifs are genetically distinct. Though this genetic distinctness may only apply to a specific homology block within the sequence rather than the full-length sequence. From the analysis of sequence data in the seven genomes, a few sequences with DBLα1.5, 1.2 and 1.6 domains, were identified to belong to CP1-MFK^{+ve} and CP2-REY^{+ve} (figure 4.1) suggesting that the domain classification approach brought together what are otherwise functionally distinct sequences. This has two implications, first, the tag region may not always contain enough information to enable molecular typing of the entire domain. Second, the Rask domain classification may contain little distinctive nucleotide features that are functionally informative given the amount of genetic diversity in *var* genes.

4.3.1 Distinct Cys/PoLV groups are associated with *upsA*

There has been much interest in the idea that specific *var* genes are associated with severe forms of malaria. This is based on the idea that specific cytoadhesive functions may be associated with pathogenic patterns of sequestration. *Var* gene sequences that are found to be associated with severe disease would be of potential importance and good candidates for studies on antibodies that are associated with naturally acquired immunity. As was discussed in section 1.7 in chapter 1, sequences from the *upsA* group and those that are characterized by domain cassette 8 and 13 have been shown to be associated with severe disease (Lavstsen et al. 2012; Turner et al. 2013).

A sensitivity-specificity analysis was used to assess the ability of Cys/POLV tag classification to predict *upsA* sequences and sequences containing DC8 and DC13 cassettes within this *var* sequence dataset. Sequences with two cysteines and belonging to block-sharing group 1 (group A-like) (Warimwe et al. 2009; Bull et al. 2008) were good in predicting *var* sequences with an upsA promoter. Sequences with a combination of three semi-overlapping features, that is two cysteines, block sharing group 1 and and CP1 membership improved upsA prediction as shown in figure 4.2 B. Sequences with DC13 were associated with sequence tags containing two cysteines and belonging to group sharing block 1. The same group of sequences were not associated with *var* genes containing DC8 cassettes as shown in figure 4.2 C.

In addition, logistic regression models were used to calculate the odds of predicting *upsA* from tag classsification methods and the results are shown below.

Table 4.2: *The odds of predicting upsA sequences using the cys-POLV definition of group A sequences (cys2-bs1). This group showed significant odds of association with upsA sequences.*

	Dependent variable:
	upsA
group_A_like	112.000*** (19.882, 2,134.776)
Constant	0.250*** (0.113, 0.496)
Observations	74
Log Likelihood	−26.868
Akaike Inf. Crit.	57.736
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4.3: *The odds of predicting upsA sequences from the CP1 cys-POLV sequences. CP1 sequences are often associated with severe malaria. They had significant odds of predicting upsA sequences.*

	Dependent variable:
	upsA
CP1	38.000*** (7.024, 710.078)
Constant	0.500** (0.278, 0.868)
Observations	74
Log Likelihood	−38.342
Akaike Inf. Crit.	80.684
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4.4: *A logistic regression model showing the odds of predicting upsA sequences using a combination of CP1 and cys2-bs1 (group A-like) sequences. This is a broad definition of sequences that are associated with severe malaria. This group had the highest odds of predicting upsA sequences compaired to CP1 or cys2-bs1 alone.*

	Dependent variable:
	upsA
CP1_plus_groupA	130.500*** (22.856, 2,503.821)
Constant	0.222*** (0.096, 0.454)
Observations	74
Log Likelihood	−25.246
Akaike Inf. Crit.	54.493
Note:	*p<0.1; **p<0.05; ***p<0.01

There was not enough observations to precisely predict DC13 and DC8 sequences from the cyspolv groups and therefore this analysis should be interpreted with care.

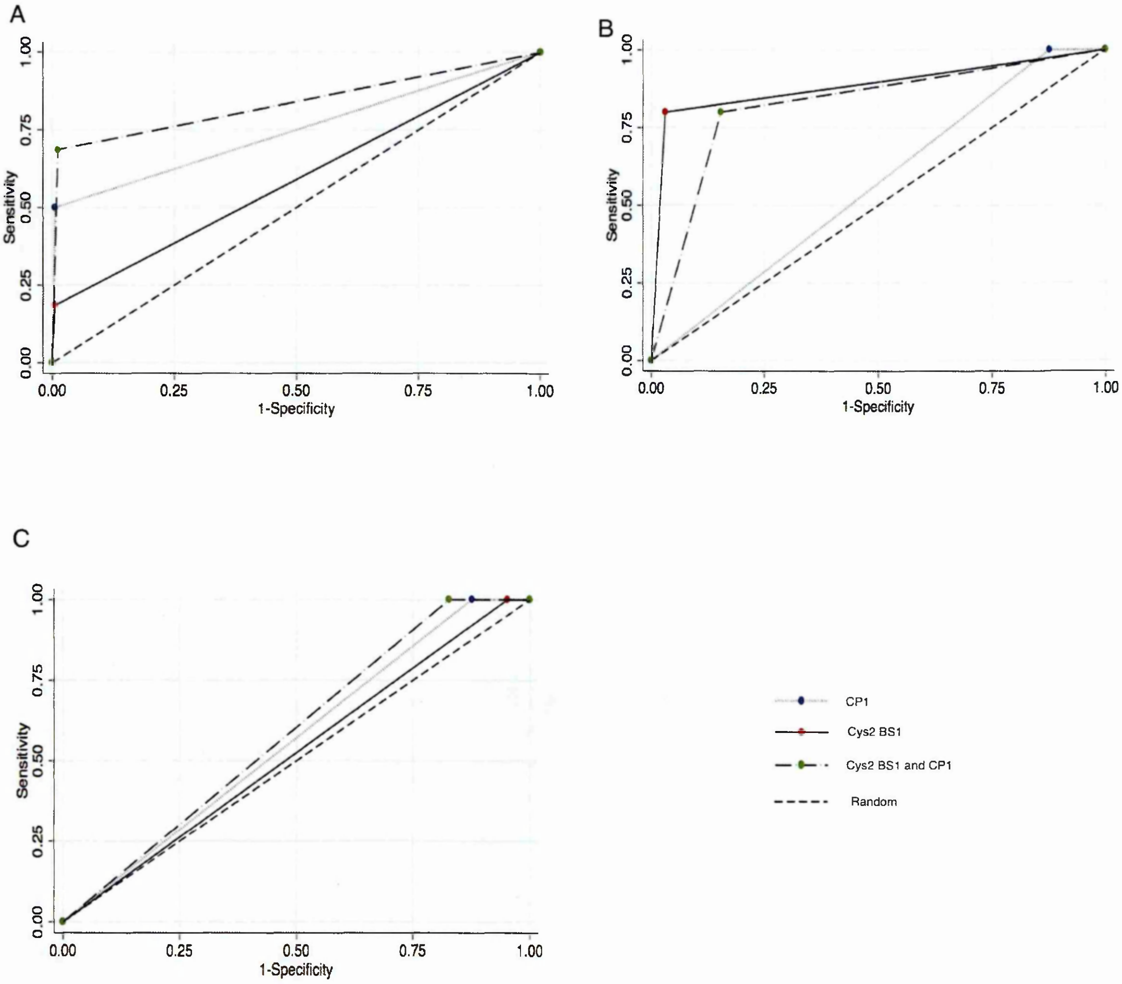


Figure 4.2: *Receiver-operator characteristic curves (ROC) for assessing the prediction of upsA, DC13 and DC8 sequences. (**A.**) CP1 sequences and sequences with two cysteines and belonging to block-sharing group1 (cys2-bs1) were good predictors of upsA sequences. Prediction of sequences with upsA promoter was improved by using tag sequences that classified as two cysteines and block-sharing group 1 together with CP1 group. (**B.**) DC13 sequences were well predicted by sequence tags with two cysteines and block-sharing group 1 but not CP1 group. DC13 appear to be a subset of group-A like var sequences. (**C.**) Predicting DC8 sequences from tag sequences with two cysteines and belonging to block-sharing group 1 together with CP1 group. DC8 cassettes were poorly predicted by cys2-bs1 (group A-like) sequences.*

4.3.2 DC8 and DC13 sequences exhibit distinct recombining patterns

Recombination networks and structuring can be visualized based on sequence blocks that are shared between sequences. Recombining segments can vary in length and therefore several recombination structures can be defined depending on the length of shared blocks. To address this, Bull et al. (2008) used amino acid blocks with a variable number of amino acids to draw a network of shared sequence blocks. Using 14 amino acids, the network segregated into two main components, which were referred to as block-sharing group 1 and 2. DC8 sequences were restricted to block-sharing group 2 component while DC13 were restricted to block-sharing group 1. Sequences from Lavstsen et al. (2012) study in Tanzania belonging to block-sharing group 2 shared sequence blocks and were connected to a component of Kilifi sequences as shown in figure 4.3.

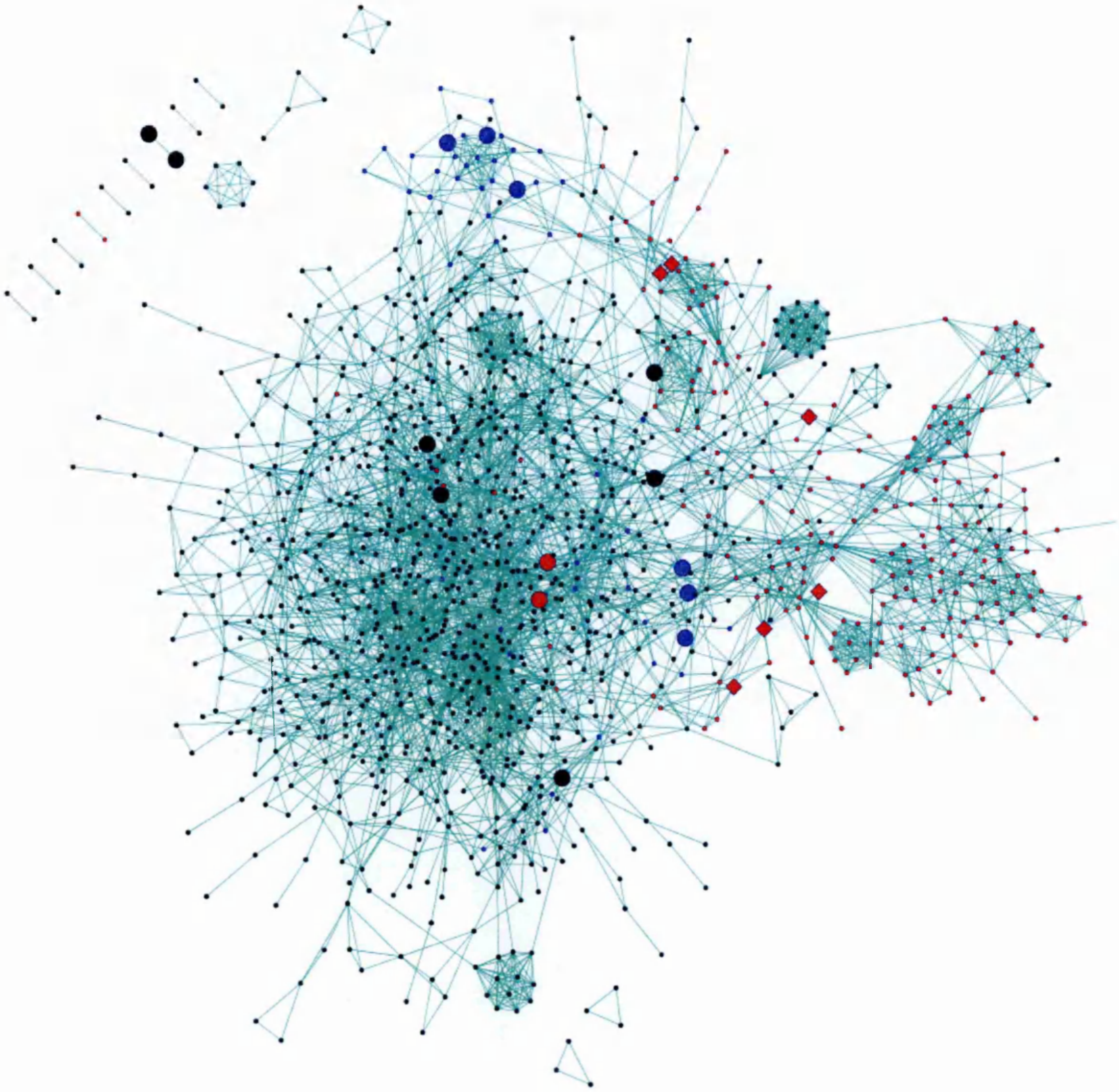


Figure 4.3: *A network of block-sharing sequences collected from Kilifi(Bull et al. 2008), laboratory (Rask et al. 2010) and Tanzanian isolates (Lavstsen et al. 2012). The diamond shaped enlarged nodes represent DC13 sequences. The circular-shaped and enlarged nodes represent DC8 sequences. Red nodes represents block sharing group 1 and the blue nodes represent block-sharing group 2 based on the Cys/POLV classification. All DC13 sequences (diamond shape) appear on the right-most part of the network. DC8 sequences (circular nodes) appear on the larger component but also in between the smaller and the larger component. Two of the DC8 sequences classified as block-sharing group 1 are overlaid on the larger component on this network.*

4.3.3 Newly described DBLα domains could consist of genetically heterogeneous sequences.

The Rask et al. (2010) study (see chapter 1 section 1.7) used sequence alignments to stratify sequences into different sub-domains. Sequence alignments of diverse and recombining global collection of *var* sequences require careful interpretation because multiple alignments ignore the recombination history. Sequences that are otherwise functionally distinct could be brought together by local alignments of recombining segments as discussed in 4.1.

Examination of DBLα tags shows that MFK and REY motifs are never found on the same sequence, suggesting that they contain genetically distinct regions. The DBLα1.5, DBLα1.2 and DBLα1.6 groups defined by Rask and colleagues comprise of a mixture of sequences which contain both MFK and REY motifs. This suggests that the newly defined sub-domains do not always classify sequences into genetically distinct groups. This point is supported by a study by Larremore et al. (2013) that explored community structuring in *var* gene sequences. Larremore defined distinct highly variable sequence blocks from a collection of *var* sequences and showed that recombination constraints shaped the network structures in distinctive ways that agreed with known clinical phenotypes (ibid.). Importantly, the study showed that highly variable regions tend to differ from each other but their recombination communities correspond to known ups classification. They did not explore the *var* sub-domains as defined by Rask. This is the discordance that we present here.

4.3.4 Cys/PoLV groups cannot predict DC8 sequences given a global collection of sequences but may do so within a restricted geographical location

Similar to group A-like sequences, DC8 sequences are associated with severe malaria(Lavstsen et al. 2012; Rask et al. 2010; Bengtsson et al. 2013; Bertin et al. 2013) and associate with a specific class of DBLα2 sequences that could result from a recombination event at a recombination hotspot situated 3' of the DBLα tag region (Lavstsen et al. 2012).

This suggests that the DBLα tag provides no information on this group of var genes. This is consistent with the observation that DC8 sequences also correspond to CP2, CP3 and CP4 Cys/PoLV groups as shown in figure 4.4 and figure 4.5 and in block-sharing group 1 and 2. This suggests that DC8 cassette comprises of a genetically diverse group of sequences and that it is not possible to predict DC8 sequences using the Cys/PoLV groups. In other words DBLα tag analysis is limited in making predictions about important full length molecules.

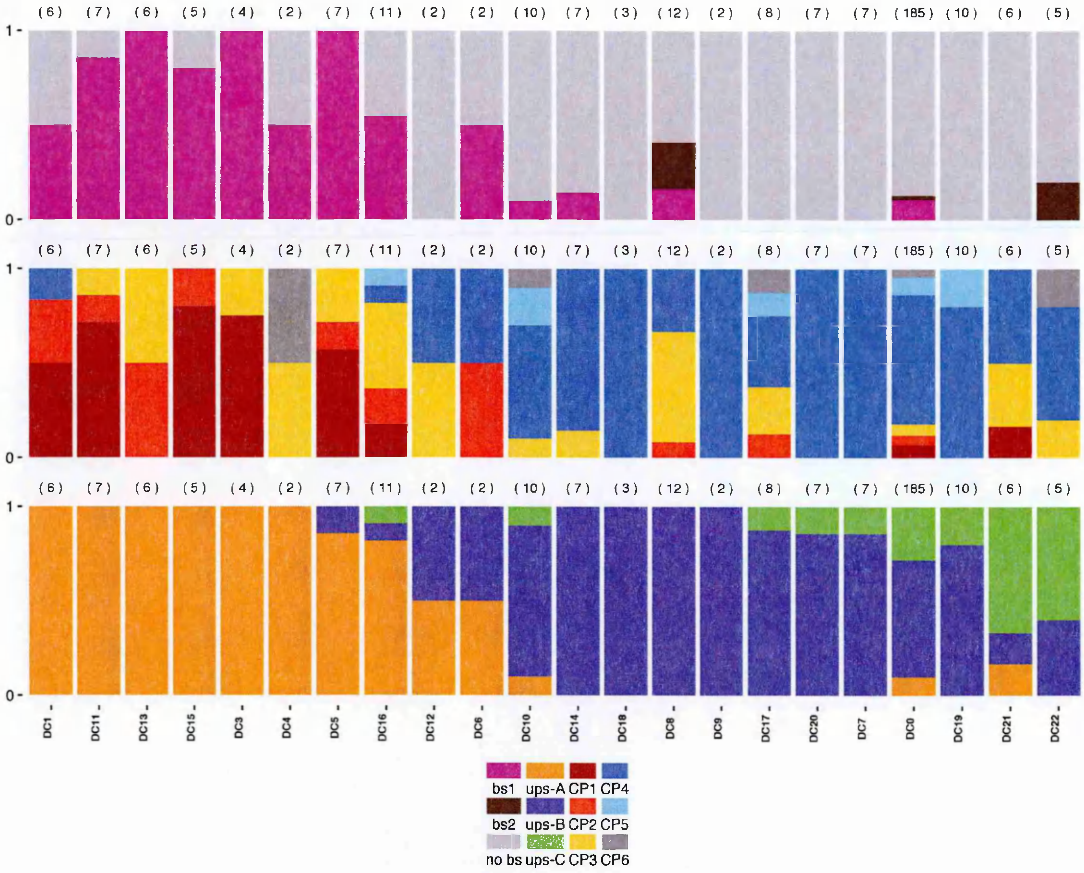


Figure 4.4: *The relationship between domain cassettes (lower panel), the Cys/PoLV (middle panel) and block sharing groups (upper panel). The cassettes sorted from left to right such that the leftmost sequences contain the largest proportion of upsA while sequences to the right contain the largest proportion of upsC. The number of sequences from each domain cassette is shown at the top of each bar. DC0 denotes sequences that were not assigned to a domain cassette according to the original study by Rask and colleagues.*

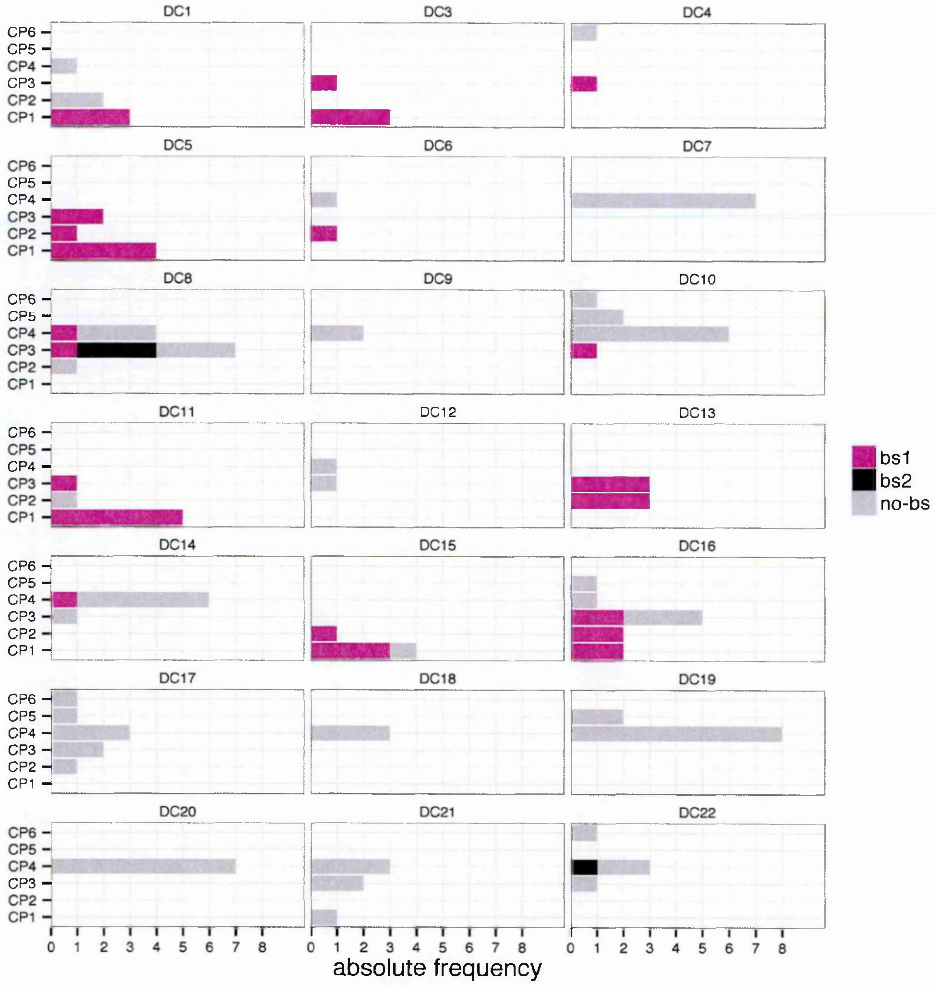


Figure 4.5: *Bar plots showing the distribution of block sharing groups among 23 domain cassettes. All the sequences with DC13 cassettes were block-sharing group 1. Sequences with DC8 cassettes were block-sharing group 1 or 2. Some of them did not belong to either block-sharing group 1 or 2. Sequences with DC5 cassettes were from different Cys/PoLV groups all of which belonged to block-sharing group 1. A CP4 sequence from block-sharing group 2 was found to have a DC22 cassette.*

The highest proportion of observed block-sharing group 2 sequences was found in sequences with DC8 cassettes as illustrated in figure 4.5. In addition the DC8 sequences identified in Tanzania (Lavstsen et al. 2012) were similar to "sig2" sequences as show in the alignment in figure 4.6. "sig2" sequences have a LYLD-VREY-KAIT-PTNL distinct sequence identifier. Sequences from the Lavstsen et al. (ibid.) study in Tanzania those from Bull et al. (2008) study in Kilifi corresponded with CP2 sequences in block-sharing group 2. This suggests that DC8 sequences are associated with specific tag types within a particular region during a specific time of sampling. These "sig2" sequences were found in two severe malaria cases by Bull and colleagues (Bull et al. 2005)

```
1983_3      DIGDIIRGKDLYLDHEPGKQHLEERLERIFANIQN     35
1983_1      DIGDIIRGKDLYLDHEPGKQHLEERLEQMFENIKN     35
4187.dom2   DIGDIIRGKDLYLDHEPGKQHLEERLERIFANIQK     35
1965_1      DIGDIVRGKDLHLRHEPGIQHLEKRLESMFEKIQK     35
4140.dom    DIGDIIRGKDLYLDHEPGKQHLEERLEQMFENIKN     35
4187.dom1   DIGDIIRGKDLYLDHEPGKQHLEERLEQMFENIKN     35
4187.dom3   DIGDIIRGKDLYLDHEPGKQHLEERLEQMFENIKN     35
4187B       DIGDIIRGKDLYLDHEPGKQHLEERLEQMFENIKN     35

1983_3      KN-EKLKDLPLDEVREYWWELNRDQVWKAITCGAT     69
1983_1      NN-EKLKDLPLDEVREYWWALNRVQVWKAITCNAE     69
4187.dom2   EN-GDINTLKPEEVREYWWALNRVQVWKAITCRAE     69
1965_1      NNNNKLSNLSTKEVREYWWALNRVQVWKAITCKAK     70
4140.dom    NNAAKLSELSTAQVREYWWALNRVQVWKAITCRAE     70
4187.dom1   NNAAKLSELSTAQVREYWWALNRVQVWKAITCKAK     70
4187.dom3   NNAAKLSELSTAQVREYWWALNRVQVWKAITCKAK     70
4187B       NNAAKLSELSTAQVREYWWALNRVQVWKAITCKAK     70

1983_3      MKDISSKNIRDAKMILFHYNCGHH-NDKAQTYLDY    103
1983_1      GTDKYFKKSSGGDYLFSGGKCGRN-EEKVPTYLDY    103
4187.dom2   EKDIYSKTTDNGKLILWNYNCGHHVNQDVPTNLDY    104
1965_1      EGDIYSKT-ANGNTTLWNDNCGHHVNQDVPTNLDY    104
4140.dom    EKDTYFKNRENGKLLLWNYKCGHHVNQDVPTNLDY    105
4187.dom1   EGDIYSKTMNNGNMVFWYPKCGHHVKQDVPTNLDY    105
4187.dom3   EGDIYSKT-ANGNTTLWNYNCGHHVNQDVPTNLDY    104
4187B       EGDIYSKT-ANGSTTLWNYNCGHHVNQDVPTNLDY    104

1983_3      VPQYLR   109
1983_1      VPQFLR   109
4187.dom2   VPQFLR   110
1965_1      VPQFLR   110
4140.dom    VPQFLR   111
4187.dom1   VPQFLR   111
4187.dom3   VPQFLR   110
4187B       VPQFLR   110
```

Figure 4.6: *An amino acid sequence alignment of DC8 sequences from Lavstsen et al. (2012) study in Tanzania and sig2-like sequences from Kilifi collected by Bull et al. (2005). Although these sequences were collected at different times and from two different locations within East-Africa, it is interesting that they were similar based on the amino acid sequence and contained a signature* HEPG *motif.*

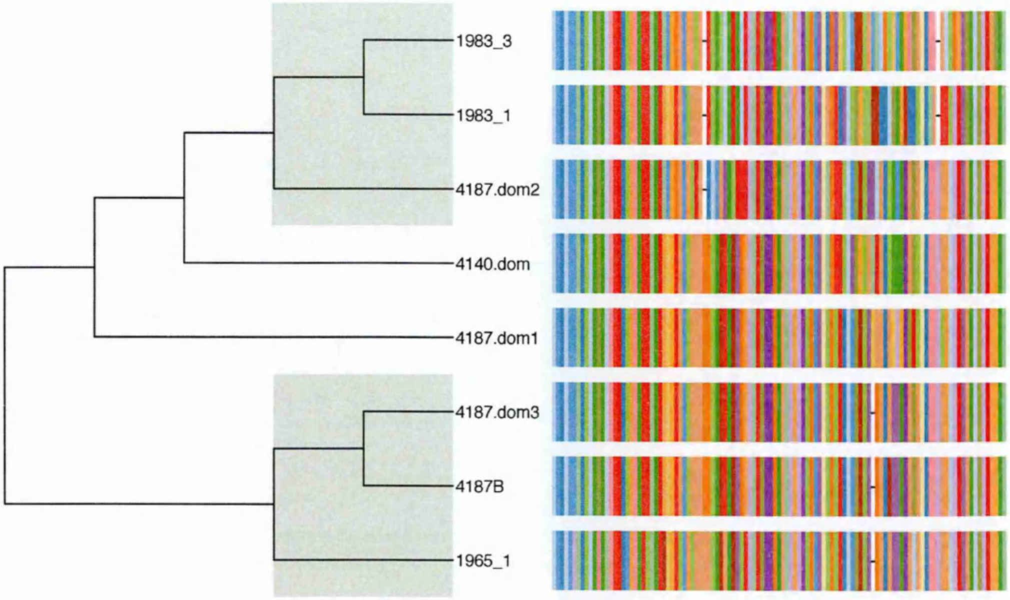


Figure 4.7: *A neighbour-joining tree of DC8 and "sig-2" sequences from East-Africa. Sequence 1983_3, 1983_1 and 1965_1 was were from Lavstsen et al. (2012) study conducted in Tanzania and the rest of the sequences were from Bull et al. (2005) study in Kilifi . Although the sequences were collected at different time points and in two different geographical locations within East-Africa, the similarities between them is remarkable and importantly both studies associated these sequences with severe malaria.*

4.3.5 Other significant DC groups

Sequences with DC4 cassettes are reported to associate with binding to ICAM1(Bengtsson et al. 2013) In this data set, there were only 2 sequences with DC4 cassettes; one sequence correspond with CP3 and the other one with CP6. A larger sample size would more appropriate in order to provide a plausible observation.

4.4 Conclusion

This analysis shows that Cys/POLV analysis is inconsistent with the domain cassette classification and performs poorly in identification of important domain cassettes.

The Rask et al. (2010) DBLα domains are not associated with particular functional groups and that functional and distinct groups of sequences may have geographical restriction.

An important consideration when interpreting these results is that the sequences that were analyzed were from laboratory isolates and therefore they are not representative of *var* sequences from clinical isolates. Second, it is possible for *var* sequences to belong to multiple domain cassettes and presumably the domain cassettes described by Rask et al. (ibid.) could have missed novel cassettes if more clinical parasites were sampled.

The observations show that Cys/PoLV groups containing two cysteines and block-sharing group 1 can predict *upsA* sequences with high sensitivity and specificity and that they can distinguish *var* genes with DC8 domain architecture, at least within a restricted geographical area and at specific time points.

A distinct group of sequence tags contain a HEPG motif correspond to *var* genes with a DC8 architecture (figure 4.6). DC8 sequences also contain a known recombination break-point and carry both CP4 and CP2 DBLα sequence features suggest that when they arise, DC8 sequences could spread rapidly within a population.

The Cys/POLV classification corresponds well with the *ups* classification but is poorly informative on domain architectures. Therefore it is a poor method of assessing the molecular diversity of full-length *var* genes. As highlighted earlier, more sequence data from clinical isolates are required to understand the correspondence between the various expression classification approaches and to provide appropriate data for a detailed thorough analysis.

Chapter 5
The role of point mutations in generating *var* diversity

5.1 Background

Diversity in *var* sequences is generated through recombination (Freitas-Junior et al. 2000; Taylor et al. 2000a; Kraemer et al. 2007; Falk et al. 2009). Recent studies have shown evidence for mitotic recombination (Bopp et al. 2013; Claessens et al. 2014) as discussed in chapter 1 section 1.8). Mitotic recombination adds another level of diversity and is also associated with DNA and chromosomal structural changes (Sander et al. 2013; Claessens et al. 2014). Despite immense molecular diversity, *var* genes are able to maintain discrete functional properties as described in chapter 1 section 1.7.

Among other approaches such as recombination, nucleotide substitution contributes to genetic variation upon which selection acts. It is difficult to estimate substitution patterns in var sequences for several reasons. First, *var* genes are long and diverse molecules that undergo extensive recombination events. Secondly, *var* genes are characterized by length differences which hinders direct comparison of sequences without taking into account how insertion and deletions are treated. Thirdly, the presence of PCR and sequencing errors can add significant noise when profiling mutation patterns. These factors provide significant challenges of estimating selection. Despite these challenges, it is important to understand how the pattern of nucleotide variation is structured to gain an understanding of how it is shaped by immune selection

Molecular diversity studies make use of multiple sequence alignments, distance metrics, and tree building methods to infer sequence relatedness and selection pressure. Regrettably, in presence of recombination, the basic assumptions of many phylogenetic methods are violated (Schierup and Hein 2000; Anisimova et al. 2003).

Zilversmit et al. (2013) developed a statistical method based on a Hidden Markov Model to show the evolutionary history of *genes* that are suggested to overcome some of these challenges. In order to describe patterns of single nucleotide variation in DBLα, sequences were grouped into clusters of known similarity and therefore sequences in the same cluster were assumed to be closely related and thereby comparable. The pattern of variation was obtained from a profile of substitutions and described based on nucleotide type and codon position within isolates and between isolates as well as in hyper-variable and conserved regions of the DBLα.

5.2 Methods

The full methods and approach are described in chapter 2 and chapter 3 of this thesis. In summary, input sequences were clustered at predefined global identity thresholds ranging from 98 to 88 percent with a 2% step to yield six identity thresholds. At each level of identity, sequences were grouped into clusters and nucleotide differences between pairs of sequences from each cluster were identified, annotated and pooled together. For comparison purposes, random substitutions were generated from or based on the respective input files as discussed in chapter 3 section 3.1.

To minimize the effect of PCR generated errors, sequences from each isolate were clustered at 96% identity and only representative sequences from each cluster were considered for analysis. While this approach was aimed at ensuring that the overall quality of sequences was high and that the number of substitutions that were attributable to PCR misincorporations were minimized and resulted in decrease number of sequences.

5.3 Results

5.3.1 DBLα sequences are diverse

Figure 5.1 and figure 5.2 shows an illustration of how the sequences clustered at 98% and at 88% global identities. Each dot represents a sequence and the colors denote group A (red) and non-group A (blue) sequences. The largest clusters corresponded with sequences that had the least amount of intra cluster variation.

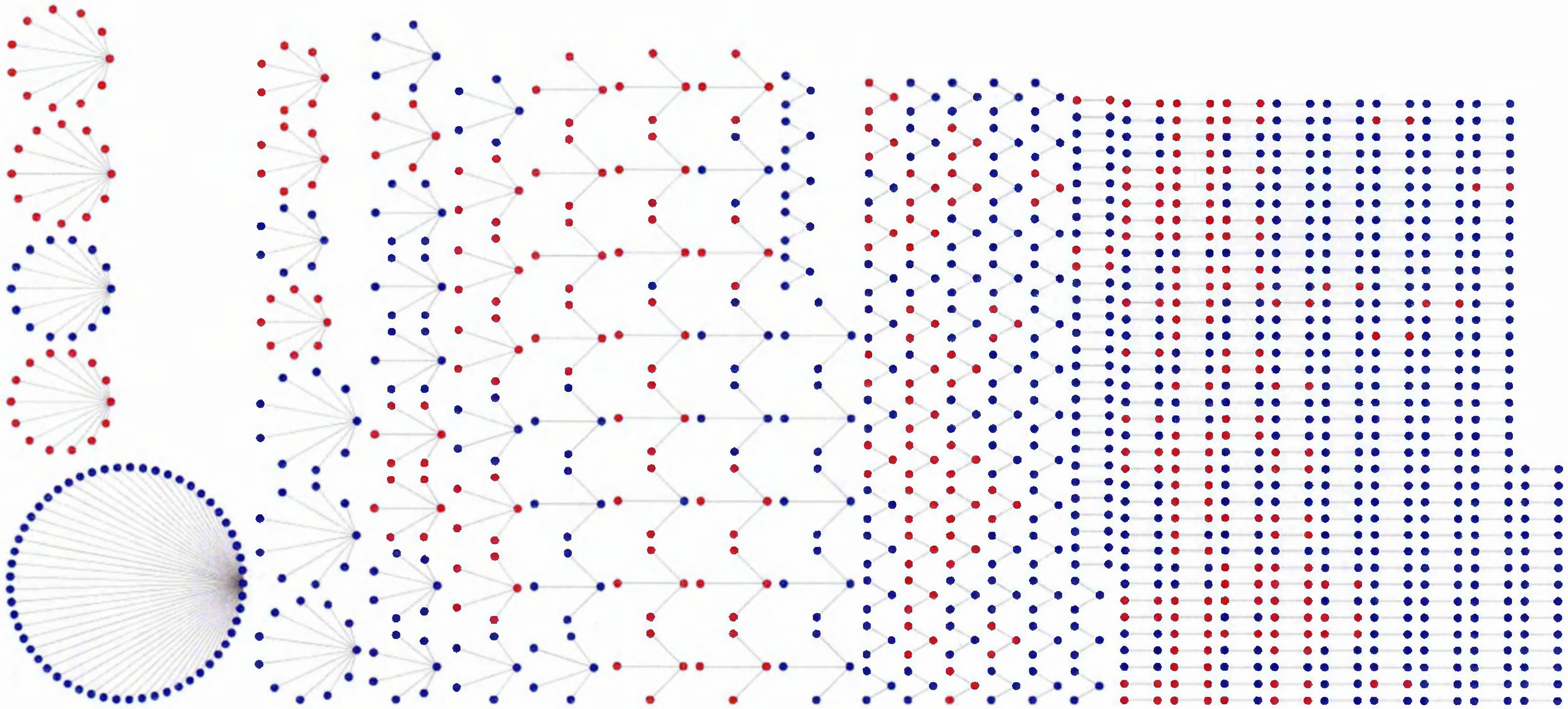


Figure 5.1: *A degree sorted circular layout of clusters of var tag sequences generated using Vsearch (chapter3 section3.5.3) at 98% sequence identity. Input sequences were clustered into groups bases on sequence comparisons to seed sequence defined as the representative sequence. In this figure, the representative sequence is the node with the most out-degrees against which all the sequences in each cluster were compared against. Red nodes represent group-A-like sequences and blue nodes represent non-group A sequences. The clusters are sorted from left to right in the order of size. Only clusters with at least two members are shown.*

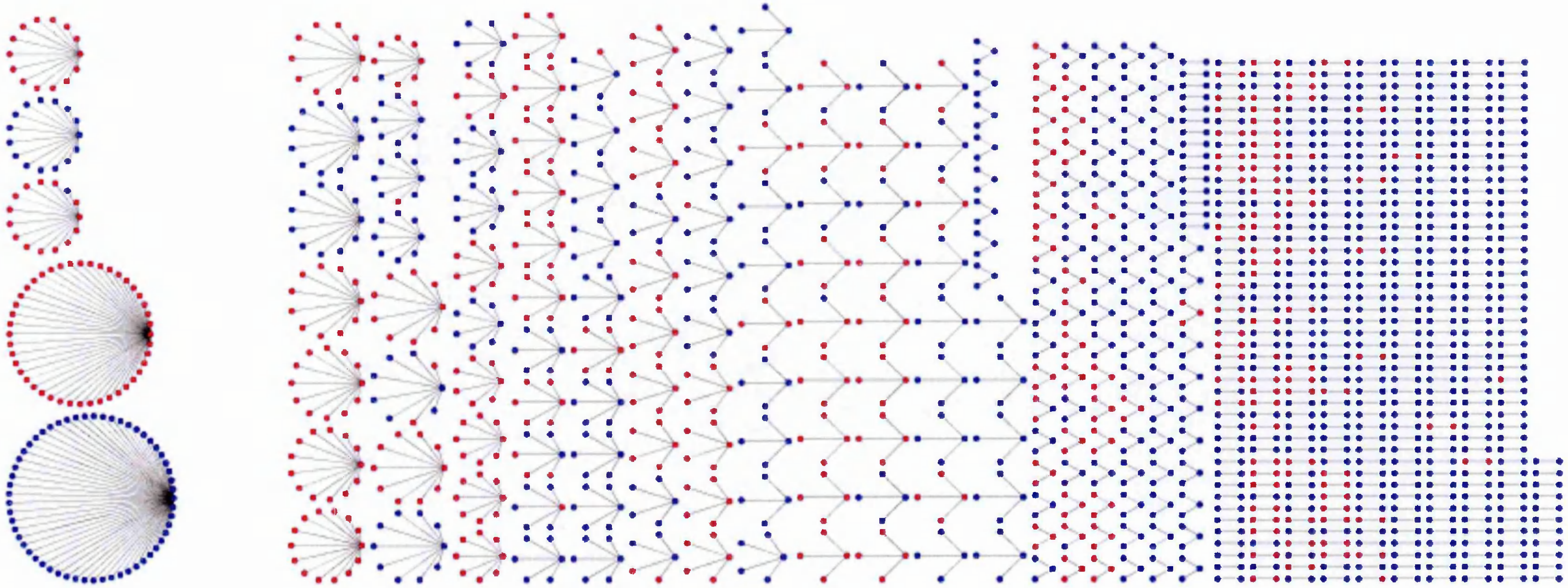


Figure 5.2: *A degree sorted circular layout of clusters generated using Vsearch at 88% sequence identity. The colors in the sequences denoted group-A (red) and non-group-A (blue). At low sequence similarity, some group A and non-group A sequences were clustered together.*

At higher sequence similarities, sequence clusters comprised of either group A or non-group A members while at lower identities some clusters contained a mixture of both group A and non-group A sequences.

5.3.2 Pattern of random substitutions in DBLα sequences

Two substitution profiles were created based on random substitutions of nucleotides in the original sequence data from Kilifi. This was to test the idea that substitution could not be fully accounted by random errors. The first model (hereafter referred to as model A) of substitution utilized data from Kilifi to generate random substitutions based on equal probability of transitions and transversions assuming a uniform model of evolution and no bias for any nucleotide substitution. A neighbour-joining tree that was constructed from a PRANK alignment in each cluster and a randomly selected sequence from each cluster were used as input to the Phylosim R package to generate random substitutions as described in chapter 3.

The second model of random substitutions (hereafter referred to as model B) generated random substitutions in a randomly selected sequence. The sequence was "mutated" 5000 times while maintaining the overall sequence similarity at 98% in the first instance and then another batch was created using the same approach at 96% and so on.

There were no differences in the profile of substitutions between the models at 98% similarity (Kruskal-Wallis rank sum test p-value=0.68) but there were differences between the models in similarities lower than 98%. (Kruskal-Wallis rank sum test p-value < 0.001).

Model A was preferred for comparing profiles with actual tag sequences because it constitutes a pool of changes derived froma larger number of seed sequences with different lengths and variable nucleotide composition.

Identity (%)	sequences	Total DBL tags	Clusters	Distinct types	Diversity ratio
98	5001	3364	578	2257	0.2560
96	5001	2205	127	2011	0.0631
94	5001	1592	70	1500	0.0466
92	5001	1184	61	1100	0.0554

Table 5.1: *A table showing the number of randomly generated sequences using a single template sequence. The clusters column refers to the number of groups of sequences that was generated and excludes 'one-member' clusters whereas the distinct types were the total number of groups including one-member clusters. The diversity ratio refers to the fraction of clusters that were used to generate the pairwise comparisons.*

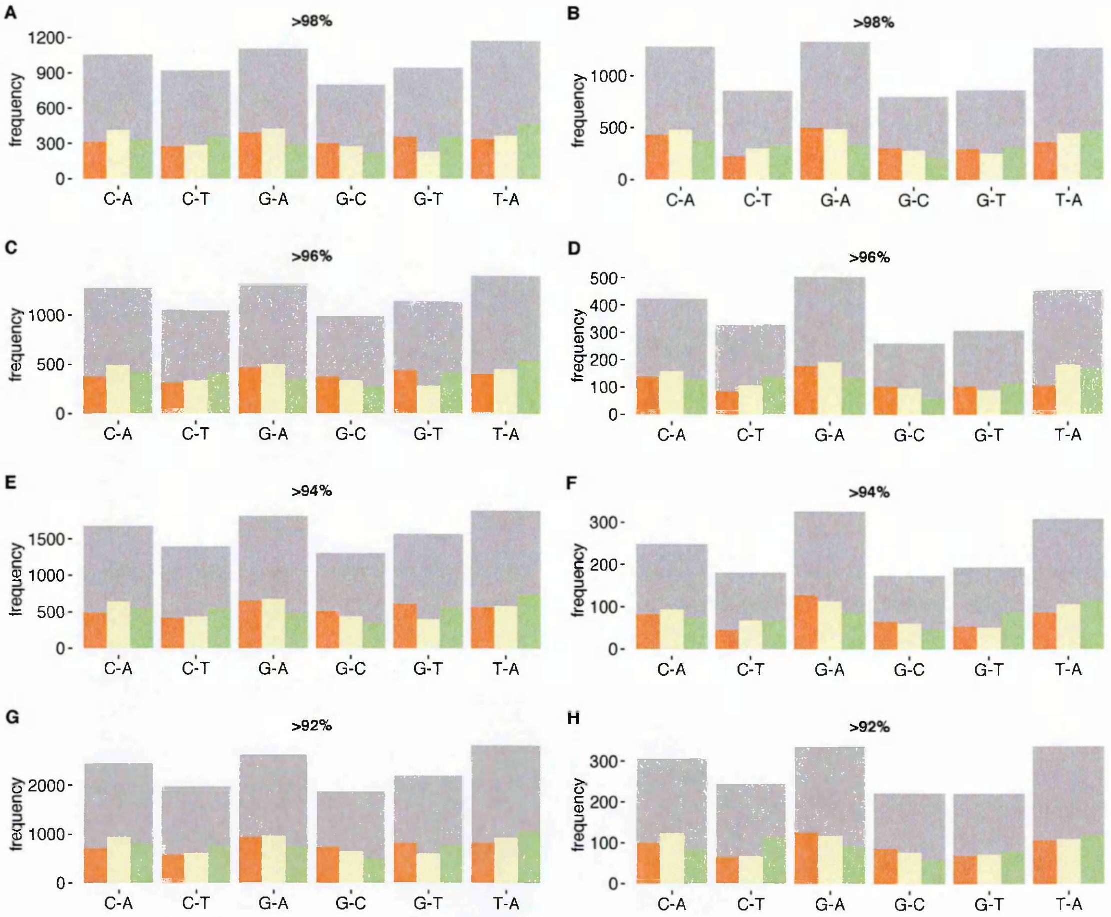


Figure 5.3: *The frequency of random substitutions in DBLα sequence tags generated with two random models. Figure A,C,E,G represent the model A and the rest represent model B's substitution profile. The type of nucleotide substitutions are shown along the x-axis and the inner bars represent the frequency of a substitution at codon position 1 (red), position 2 (yellow) and position 3 (green). The gray bar represents the sum of substitutions for each type of substitution.*

Figure 5.3 shows the distribution of substitutions at each identity threshold in each random model. In model A , the frequency of each type of change was proportionately higher than in model B.

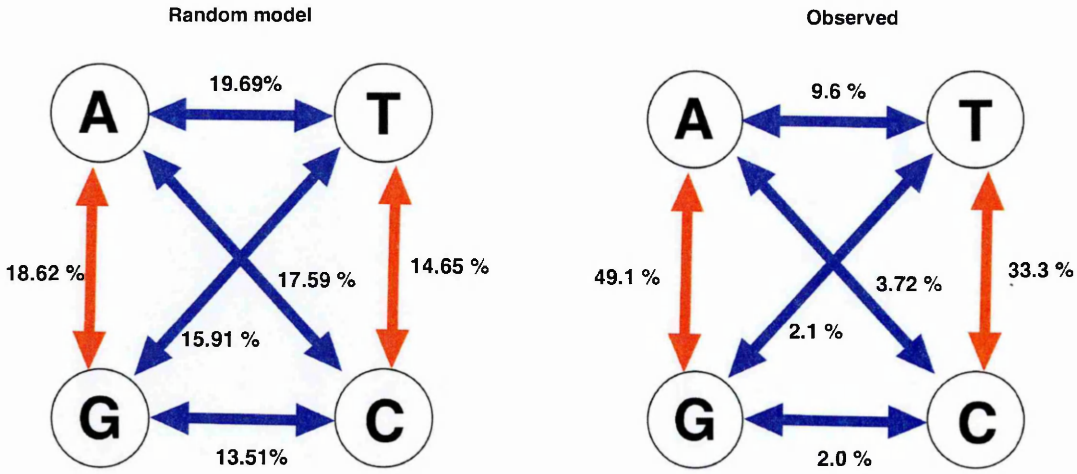


Figure 5.4: *The average frequency of different substitution in percent (%) in based on random model B and sequences collected in 2008-2010 period. The orange arrows indicate transitions and blue arrows indicate transversions. On average G-A transitions accounted for nearly half of all the observed substitutions in sequences from Kilifi.*

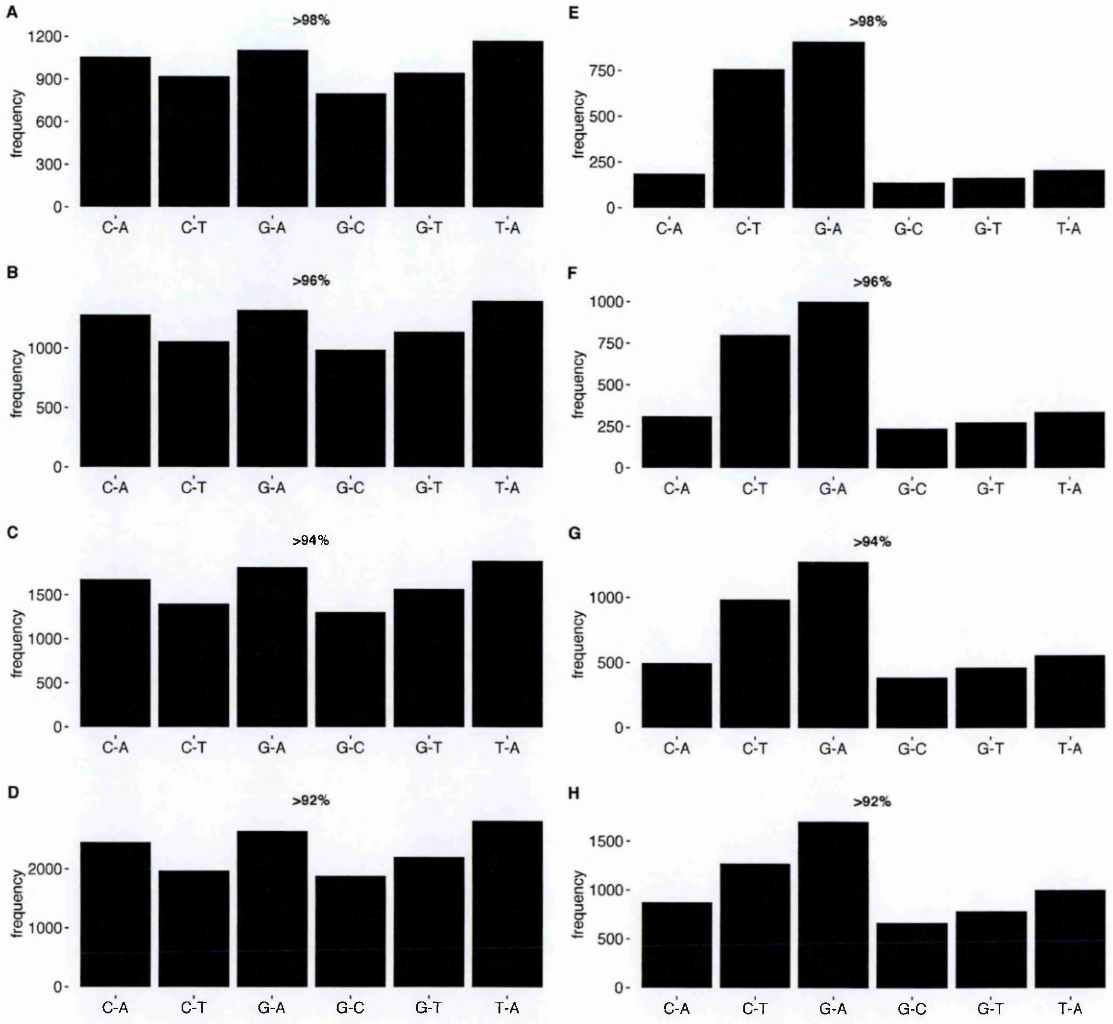


Figure 5.5: *The frequency of random substitutions. Figure A-D shows the relative frequency of random substitutions that were generated based on model B. Figure E-H shows the frequency of the same random substitutions scaled by the ratio of transitions to transversions from actual DBLα sequences collected between 2008-2010 as described in chapter 3 section 3.1. The codon position frequencies were omitted to avoid over adjusting.*

The simplest models of evolution assume that nucleotide changes arises independently and that the transition to transversion ratio (TI:TV) are identical across sites. The second codon position is often constrained. A substitution at the second codon position results in a change in the encoded amino acid. The third codon position is the least constrained and a change in the third codon position does not always result in a change in the encoded amino acid.

The tag sequences showed a strong bias for transition substitutions. The average frequency of transitions and transversions in the random models versus from the observed sequences from Kilifi was calculated and is shown in figure 5.4. The purpose of this calculation was to show the relative differences in the number of substitutions between randomly generated substitutions and from actual sequences.

5.3.3 The correlation between frequency of substitutions in random model and in actual sequences

To assess the distribution of substitutions between the random unscaled profiles and that from actual sequences, a scatter plot of the correlation between the frequency of substitutions in a random model and from sequences in Kilifi sequences is shown in figure 5.6. In summary there was no evidence for a positive correlation between the observed substitutions and the random unscaled substitutions.

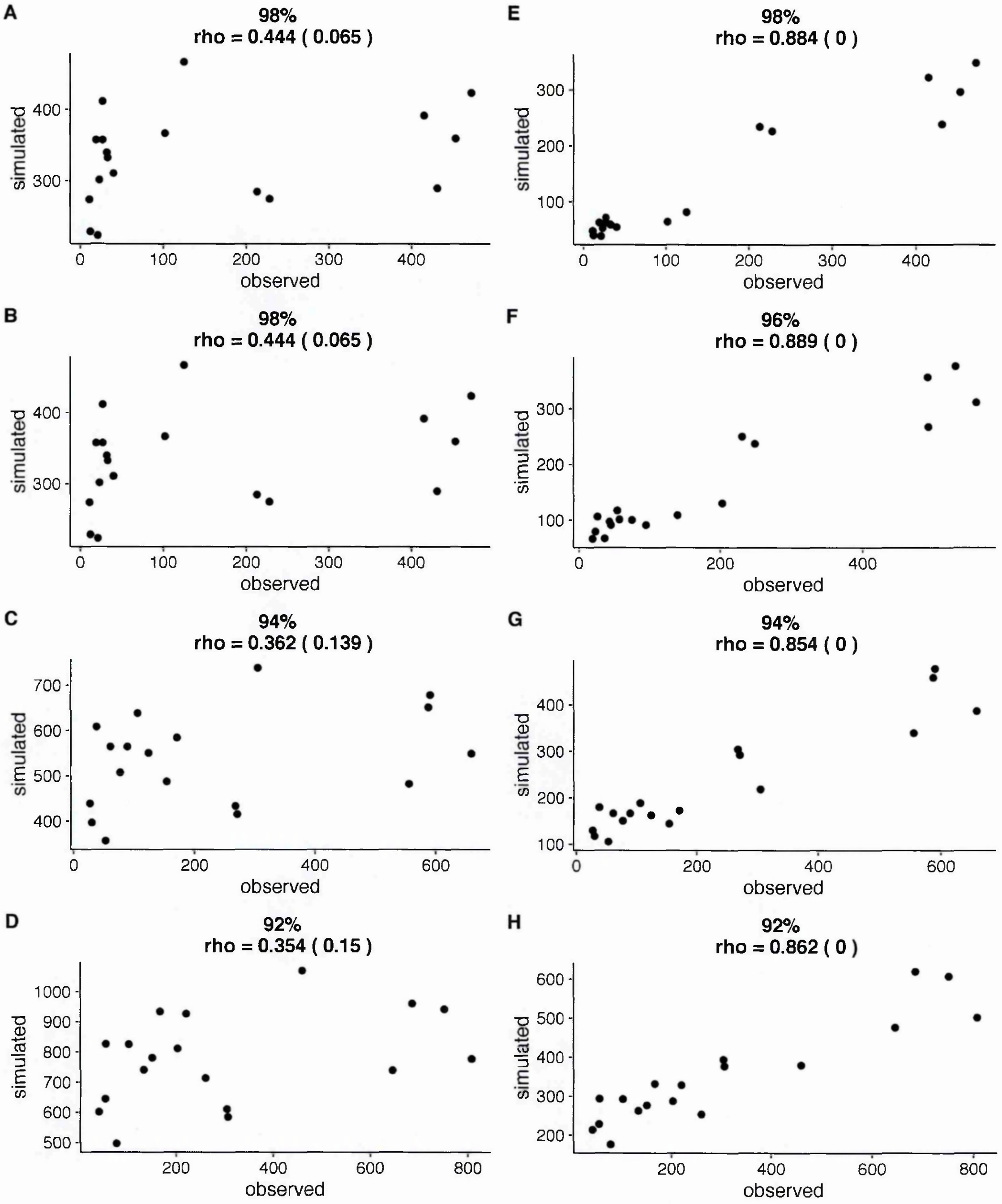


Figure 5.6: *Spearman correlation between frequency of substitutions from randomly generated substitutions and substitutions derived from DBLα sequences collected between 2008 and 2010. Figure A-D shows the random substitutions against the actuals sequences. Figure E-H shows the correlation between the scaled random substitutions and the actual substitutions from Kilifi sequences. The scaling was based on the transition:transversion ratio derived from DBLα sequences. Each point represents the frequency of each type of substitution at the respective codon. Rho indicated above each plot refers to the overall correlation and the p-value is indicated in brackets.*

Figure 5.6 shows that the ratio of transistion and transversion was important in describing the susbstitution models. Without scaling or accounting for this ratio the random and uniform model could not fully account for differences in the substitution profiles. Another way to put it is that that random based on Jukes-Cantor model of substitutions could not fully account for the observed profile of subsititutions. A more through analysis would be important to fully account for this observation.

5.3.4 Distribution of variation at each codon

Nucleotide composition varies widely among genomes and the G + C content differs between genes and organisms. The mechanisms of codon variation are related to codon and amino acid compositions. In most genes composition bias reflects the action of mutation and selection and can be linked to polymerase and transcription efficiency. The most likely factor that determines codon or amino acid usage is mutational bias that shapes GC composition constantly when genomes are either replicated or repaired by DNA polymerases

Figure 5.7 shows the G + C content and the proportion of purines at each codon for a group DBLα sequences from Kilifi that were collected between 2008 and 2010. The purines were more frequent at codon position 1 and 2 when compared to pyrimidines. The 3rd codon had the least G+C content and comprised of about an equal proportion of purines and pyrimidines.

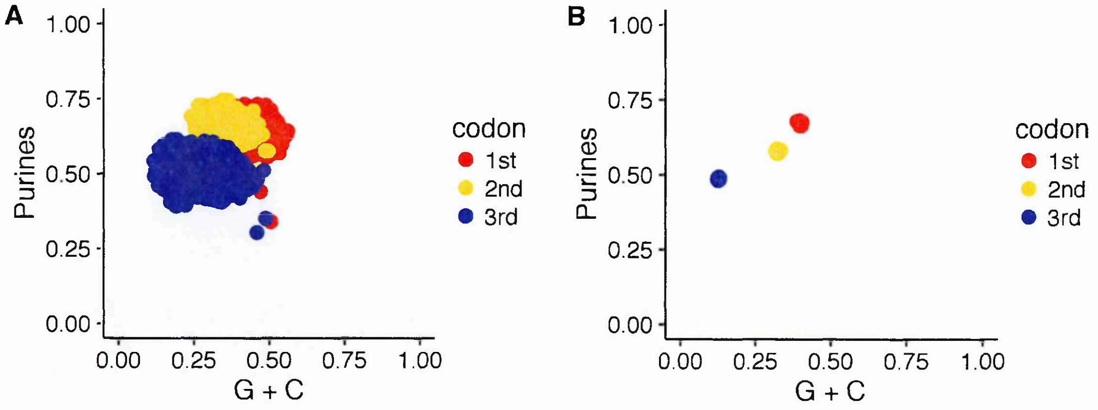


Figure 5.7: *A scatter plot showing the percent GC content (x-axis) and the average proportion of purines at each of the codon positions in DBLα and AMA1 sequences. DBLα sequence tag regions are purine rich. codon position 2 is more GC-rich than codon position 1 and 3. The plot was generated by calculating the average purine content and the average G+C content of the bases at each codon. The figure implies that variation is likely to be influenced by base composition.*

5.3.5 Density of substitutions in conserved and polymorphic regions of the DBLα sequence tag

Based on sequence alignment of PfEMP1 sequences, the DBLα sequence region falls within the homology block D to H (Smith et al. 2000a). The homology blocks are the most relatively conserved regions of the sequence and are interspersed by the polymorphic blocks which are denoted with abbreviation HP in this thesis. Nucleotide diversity (π) scores were calculated to explore differences in density of substitutions between the homology and the polymorphic regions. Nucleotide diversity scores in homology block D and H and the polymorphic regions HP1 and HP2 in each cluster were calculated as shown in figure 5.8. In the figure, the density of substitutions was plotted in sequences containing randomly substitutions (simulated) and the actual sequences collected from Kilifi (observed). Figure 5.8 shows the substitution profile of sequences clustered at 98% similarity.

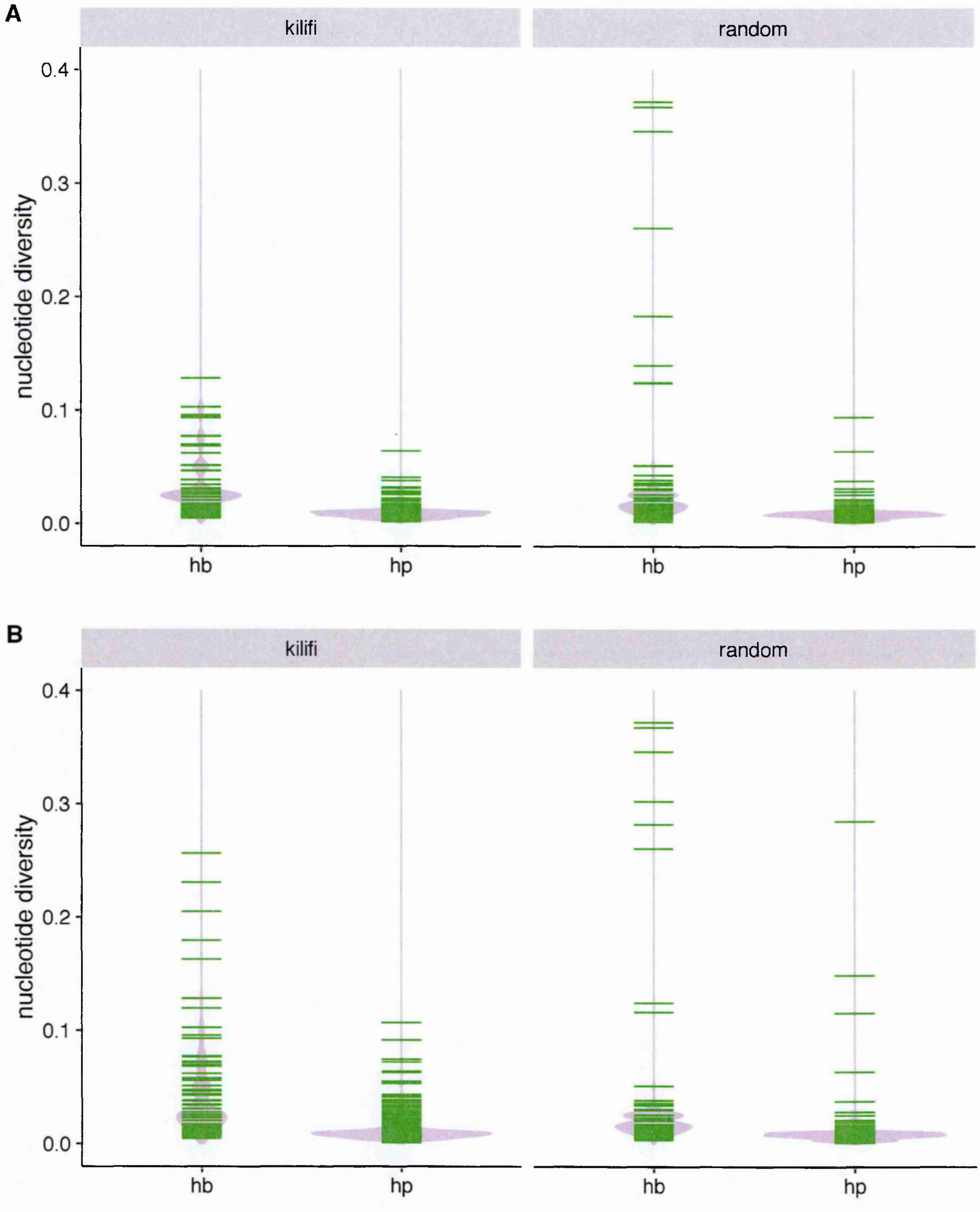


Figure 5.8: *The figure shows the nucleotide diversity (π) from polymorphic blocks (hp) and homology blocks (hb) in alignments of sequences that were similar at 98% (A) and 96% (B) that were collected. A similar profile was calculated in sequences that contained random substitutions. Each line is a diversity score from a single alignment and the grey shade shows the distribution. The homology blocks from Kilifi sequences showed a higher proportion of substitutions than the polymorphic blocks.*

5.3.6 Variation in homology block D and H

The homology blocks are relatively conserved regions of the PfEMP1 molecule (Smith et al. 2000a). Here we considered the distribution of variation in homology block D and block H which are at the 5' and 3' ends of the tag region. Figure 5.9 shows the distribution of point mutations in homology block D and in homology block H. Clusters of homology blocks for comparison were extracted from alignments of full-length sequence tags. The complete DBLα sequences were first aligned using different cutoffs and then the homology blocks were extracted from full length alignments of the tag region. This ensured that the substitutions were from sequences that were similar at the global level.

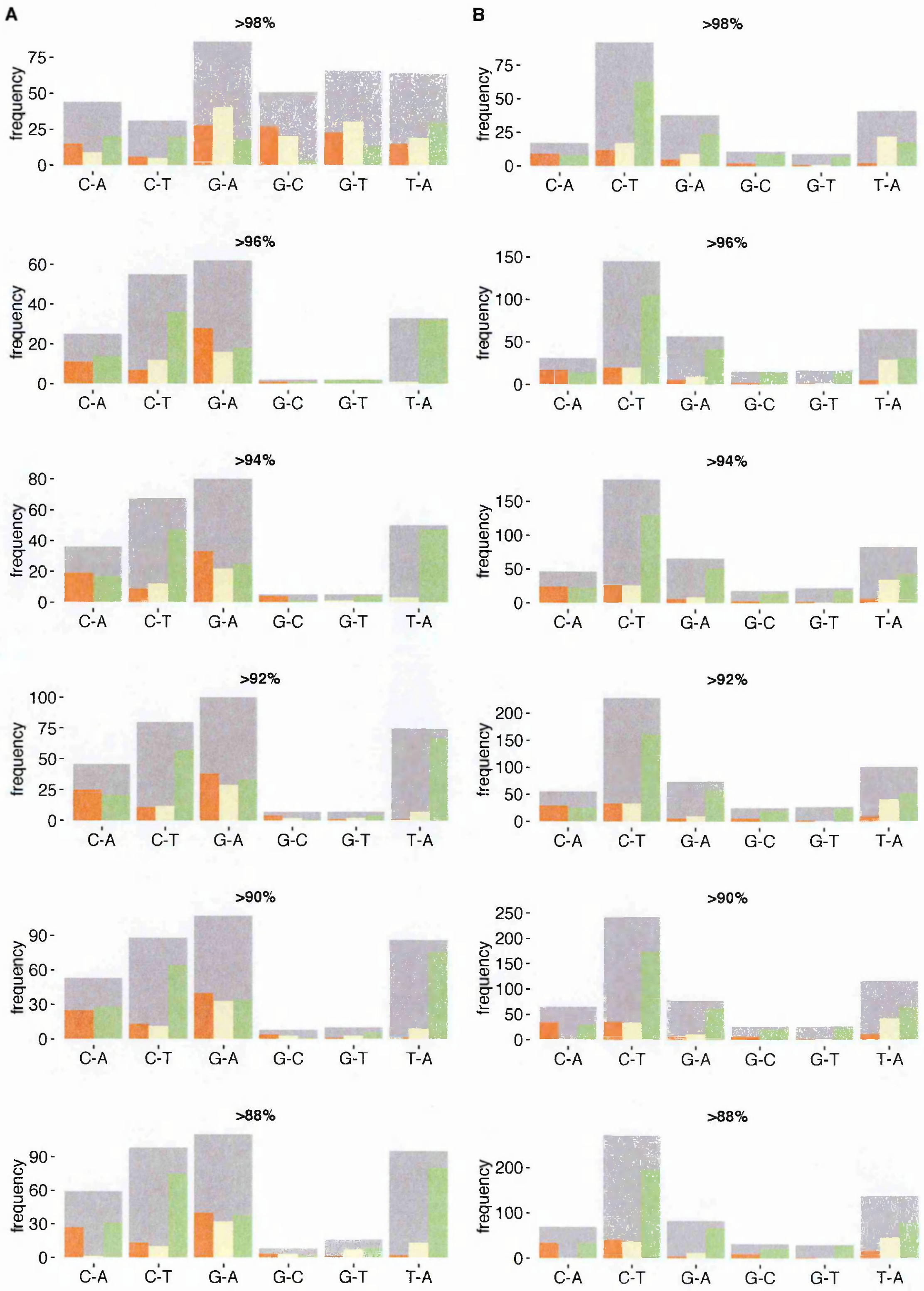


Figure 5.9: *Distribution of nucleotide substitutions in homology block D (A) and block H (B). The homology blocks were extracted from full-length alignments of DBLα sequences and therefore these alignments reflect substitution in very closely related sequences. The red bars represent subsitutions in the first codon position and the yellow and green bars represent substitutions in the second and third codon positions respectively. The gray bar represent the overall sum of substitutions for each type of substitution.*

The profile of substitutions within the conserved regions of the DBLα tag was different from that observed from full-length sequence tags. The homology blocks showed a remarkable bias in substitutions at the third codon compared to the full tag sequences.

5.3.7 Overall distribution of point mutations in DBLα sequence tags

The DBLα sequence tags are A-T rich as illustrated in figure 5.11. Nucleotide composition varied across the codon positions as illustrated in the figure 5.11. The proportion of T residues at the first codon was higher compared with that at the second and the third codon position. The proportion of **C** was the lowest at all the codons positions. The proportion of **T** at the third codon position was relatively higher than that of **C** and **G** and almost equal to that of **A**.

The most frequent type of substitutions in DBLα sequences were the **G-A/C-T** transitions. Generally, the frequency of **G-A** transitions was higher than **C-T** in most sequence groups and as illustrated in figure 5.10 and figure 5.12. The **G-A** transitions were more frequent at the first and the second codons positions. The proportion of Guanine (G) and Adenine (A) was proportionately higher at the first and the second codon position compared to the third codon position (shown in figure 5.7).

There was relatively higher number of **C-T** transitions at the third codon position. The most common transversions in sequences from Kilifi were T-A and C-A nucleotide changes. The T-A transversions were more common at the third codon position and the C-A transversions were proportionately higher at the first and second codon positions.

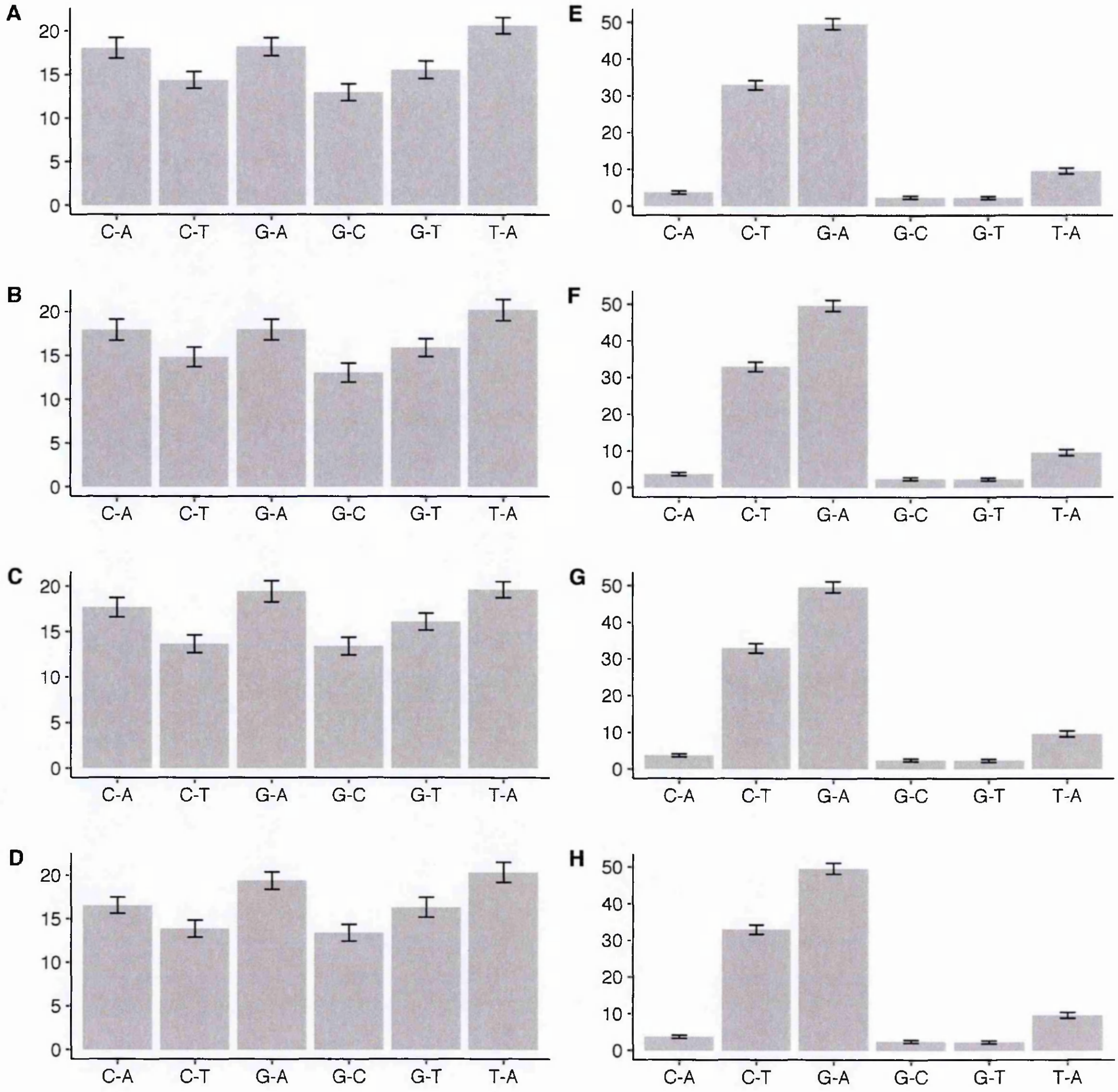


Figure 5.10: *The overall distribution of random substitutions (A-D) and from actual sequences tags from Kilifi collected between 2008 and 2010 (E-H). This plot shows the mean frequency of each type of substitution that were selected at random. In each figure, 50 random substitution were selected from the pool of all substitutions with replacement and the process was repeated 100 times. The mean number of each type of substitution is shown, the error bars indicate the 95% confidence interval. There was significant difference between the distribution of substitutions (Kruskal-Wallis rank test p-value < 0.001).*

Figure 5.10 suggests that substitutions in the DBLα sequences are not just random but are also shaped by the evolutionary pressures acting on the sequences. The figure also provides evidence that the subsitutions cannot be fully accounted by PCR errors. More analysis of codon usage and selection pressures pressures is required to fully account for these substitution profiles.

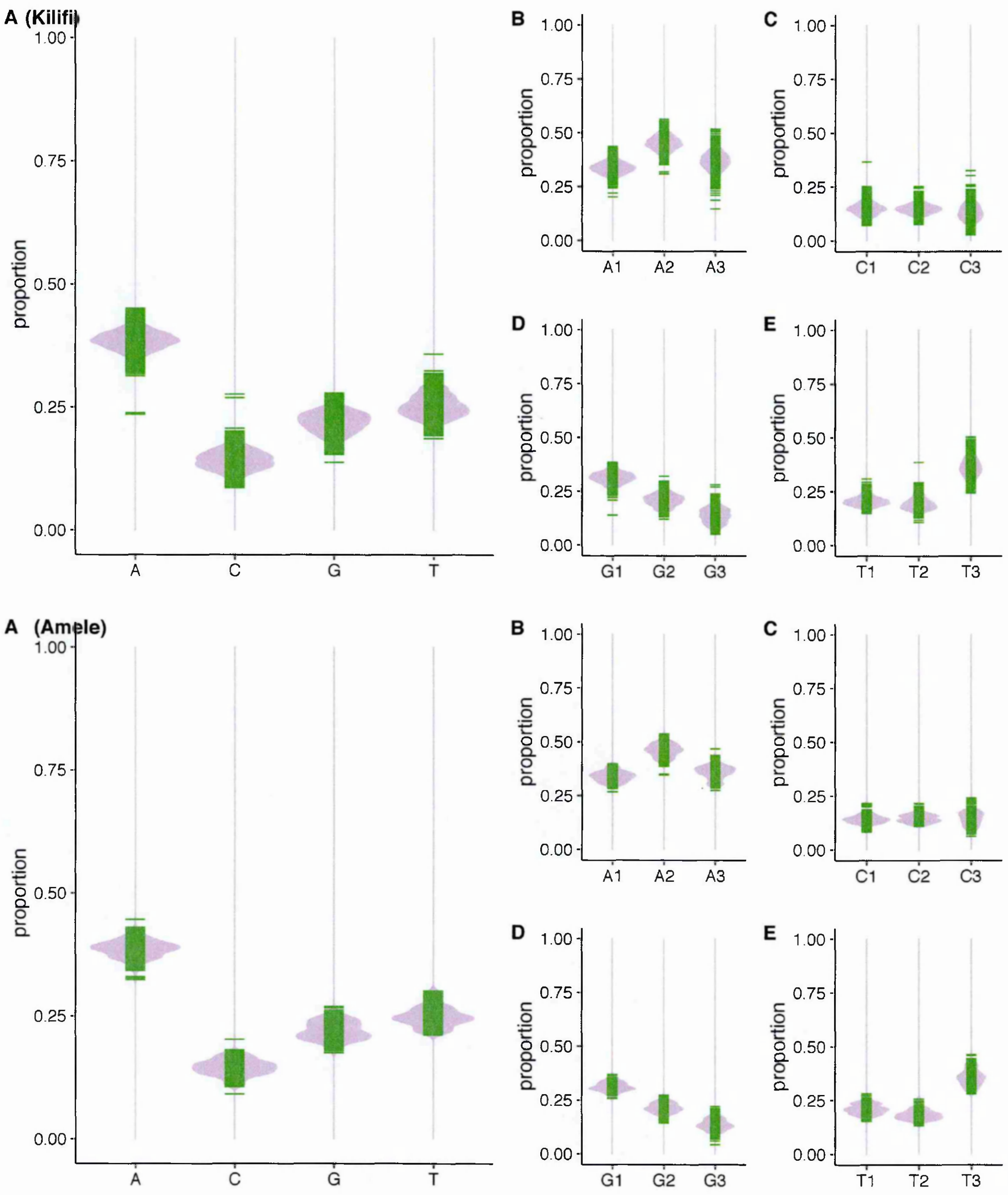


Figure 5.11: *The distribution of nucleotide content in DBLα sequences tags from Kilifi and Amele region of Papua New Guinea is shown in the figure. Panel A shows the overall distribution of A,C,T and G nucleotides while panels B-E shows the distribution of each type of nucleotide at each respective codon for both Kilifi and Amele. The green lines represent individual observations and the grey area shows the respective distribution.*

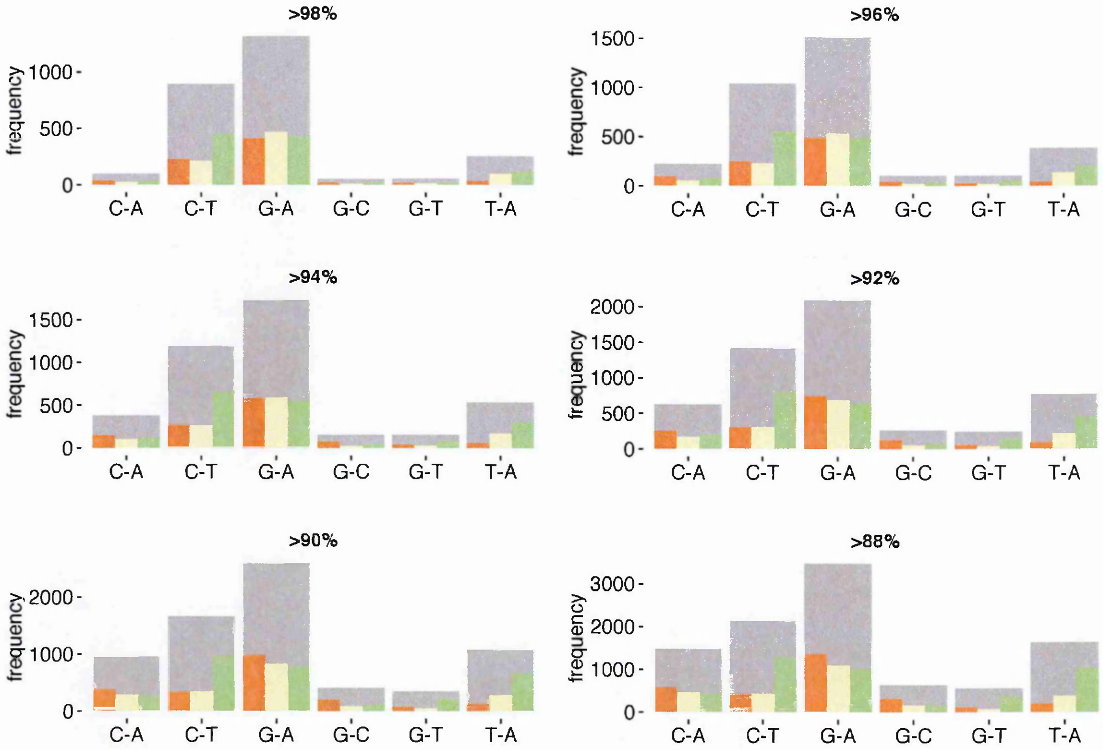


Figure 5.12: *The per codon nucleotide substitution distribution from 4817 DBLα sequence tags collected in Kilifi. Each plot shows the absolute frequency of each type of substitution. In comparison to the random substitution model, there was significant different in the pattern of changes (Kruskal-Wallis test p-value < 0.001). C/T and G/A transitions were the dominant type of substitution changes.*

5.3.8 Distribution of point mutations in DBLα sequences from other geographical regions

Transmission and host immunity varies between geographical regions which in turn can impose different selection pressures on the expressed *var* repertoire. In this section the distribution of nucleotide substitutions was explored in sequences from two geographic locations other than Kilifi. Sequences from Papua New Guinea were obtained from a published genomic epidemiology study (Barry et al. 2007), and sequences from South America were obtained from a *var* diversity study (Albrecht et al. 2010).

The distribution of the proportion of nucleotide content in sequences from Papua New Guinea was similar compared to sequences from Kilifi (figure 5.11). The G-A/C-T transitions were the most frequent transitions. The frequency of G-A was relatively higher than that of C-T. In both cases there were more variation at the third codon position as shown in figure 5.13.

C-A changes were the most frequent types of transversion substitutions and unlike the Kilifi sequences the number of T-A changes was relatively low. The transversions occurred more regularly at the third codon. The distribution of substitutions was a remarkable departure from the random model and also from what was observed in Kilifi sequences.

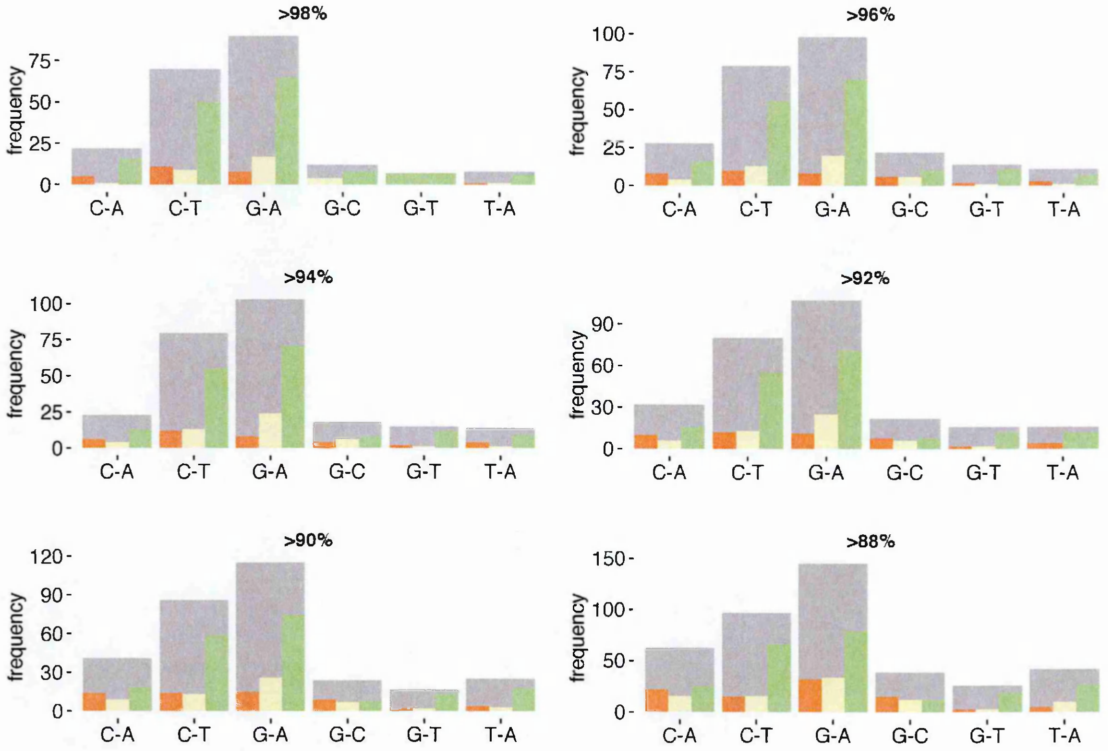


Figure 5.13: *A summary of the type and codon location of substitutions in 890 DBLα tags from Amele in Papua new Guinea. Transitions and the* `T-A` *transversion were the most frequent type of substitutions. The third codon position contributed most of the transitions and transversions.*

The *var* repertoire from parasites in the amazon basin has been described as relatively conserved (Albrecht et al. 2010; Bopp et al. 2013). Following from the observations in Amele region, it was interesting to describe the substitution profile of sequences from this region. There were no differences in the distribution of nucleotide content in sequences from Brazil compared to sequences from Kilifi or Amele. Overall, the frequency of transitions was higher than that of the transversions. T-A and C-A transvertions were the most dominant across all the similarity thresholds and similar to what was observed in sequences from Kilifi and Amele. C-T transitions were more frequent at the third codon. G-A transitions occurred proportionately frequent across all the codons as shown in figure 5.14.

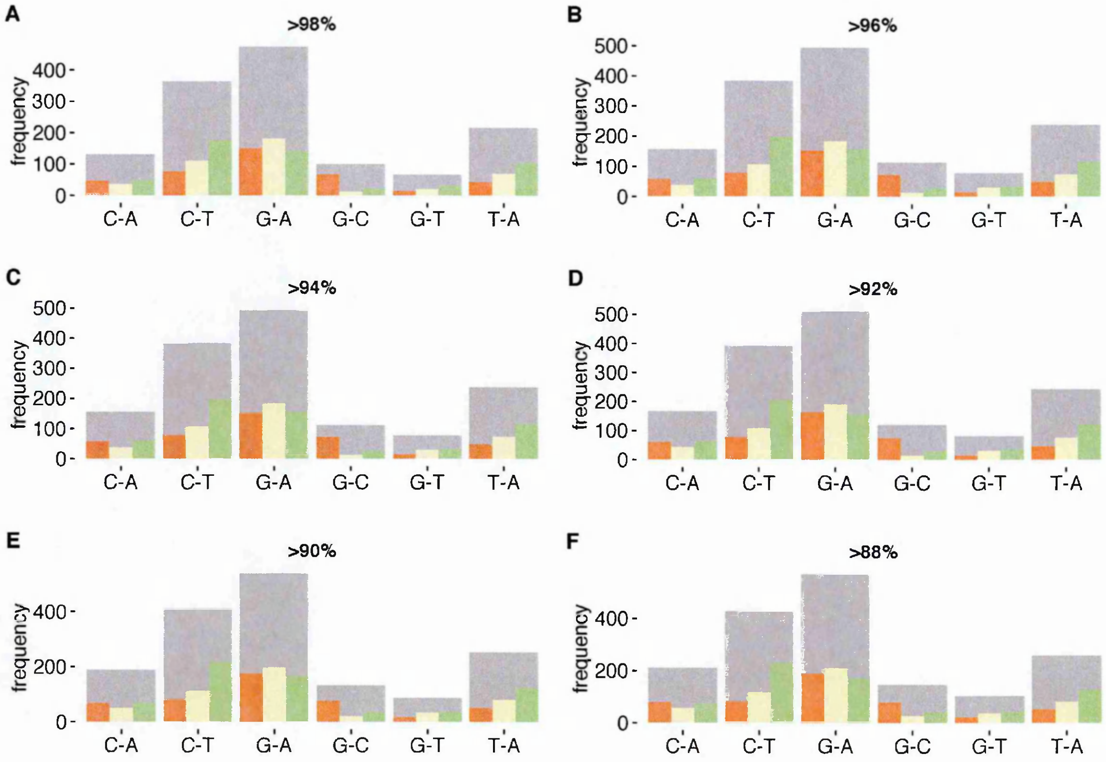


Figure 5.14: *The frequency of substitutions in 1,993 DBLα sequence tags from Ariquemes and Porto Velho regions of Brazil. Each plot shows the frequency of each type of substitution at each codon.*

5.3.9 Variation in group A and non-group A sequences

Group-A *var* genes are relatively conserved and are associated with expression in severe malaria (Warimwe et al. 2009) and are of considerable importance as discussed in chapter 1 section 1.7. Group-A like DBLα sequences were defined as sequences characterized by two cysteine residues and belong to group-sharing block 1 as discussed in chapter 1 section 1.7.1 and elaborated further in chapter 4. Given the role of group A var in immunity, it was hypothesized that they could be under different selection pressure from non-group A sequences.

A total of 1000 sequences from each group were randomly selected and clustered as described in chapter 3. Figure 5.15 shows the distribution of substitutions using an equal number of randomly selected group A and non-group A sequence tags from Kilifi and that were collected in the 2008 - 2010 time period.

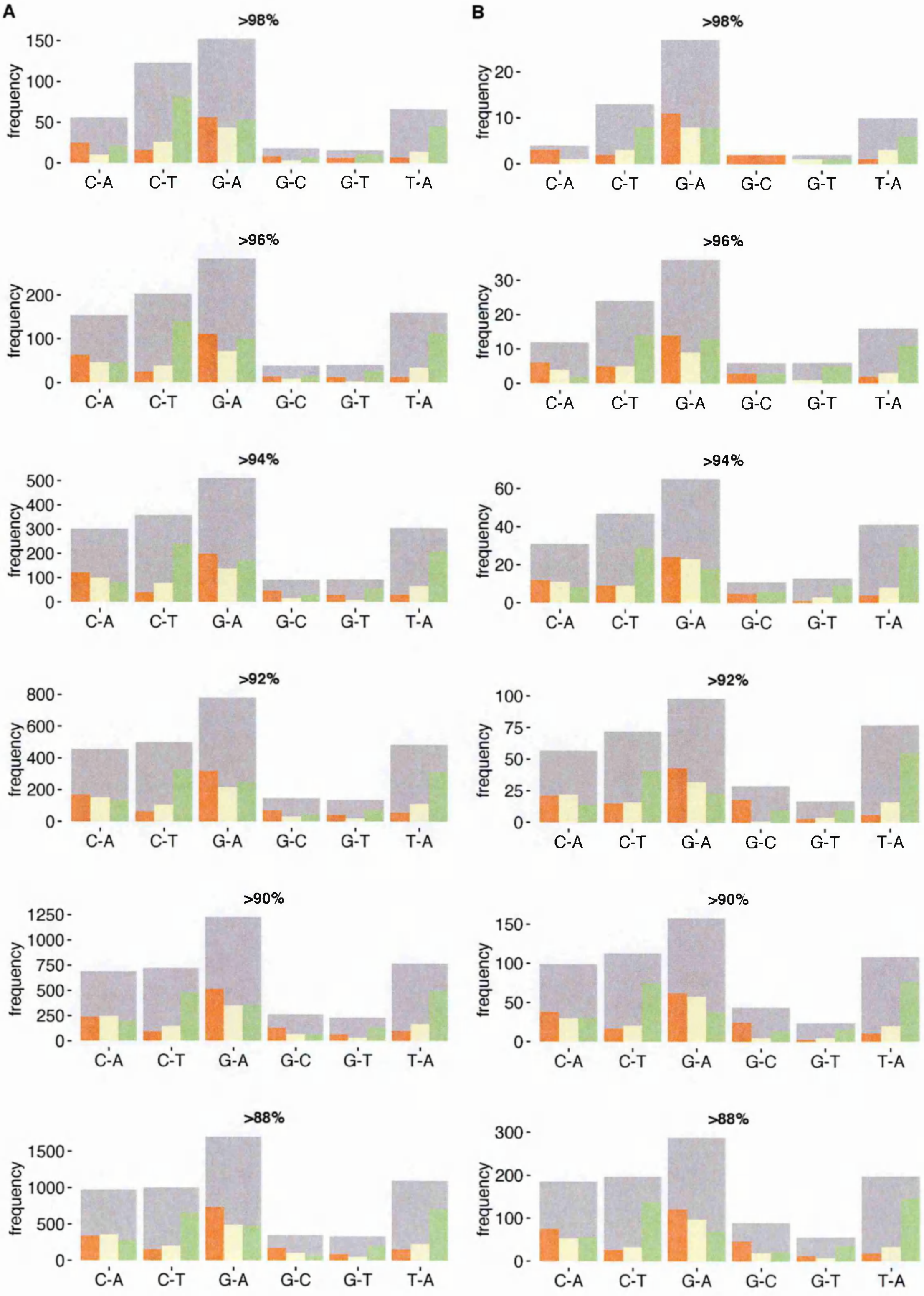


Figure 5.15: *The frequency of substitutions in group A-like and non group A sequences (column B). Group A-like sequences were defined as sequences with 2 cysteines and belonging to block-sharing group 1. In each case 1000 randomly selected sequences were clustered and the number of substitutions in each cluster at each identity threshold is shown..*

From visual inspection of the substitution profile, there were no obvious differences between group A and non group A sequences. Non-group A var tags had a proportionately higher absolute count in the number of substitutions compared to group A as shown in figure 5.15. The overall distribution of changes was also statistically significant at each percent similarity threshold based on Kruskal-Wallis rank test.

The non-group A sequences were diverse and comprised of fewer clusters compared to the group A-like sequences. This explains the difference in absolute number of substitutions that was observed between the two groups. The mechanism that drives the substitution in group A and non-group A sequences could be the same regardless of the expression profile in severe and non severe malaria.

5.3.10 Sequence conservation in multiple isolates

To assess whether there were sequences tags that were common in multiple isolates, all the sequences from Kilifi were clustered at 100% identity and the representative sequences from each cluster written to a file. A representative sequence was the index sequence against which all the sequences in a given cluster were compared to such that they had similar or greater identity with the representative sequence. Therefore representative sequences at 100% identity represent the non-redundant set of all the sequences under comparison. The representative sequences were clustered between isolates at 98% identity. Each sequence in each cluster was annotated with the respective frequency at 100% identity and number of distinct isolates as shown figure 5.16.

There was little 100% conservation among the sequences. The most conserved sequences were the *var1* as depicted by the sequences from cluster #2799 in figure 5.16.

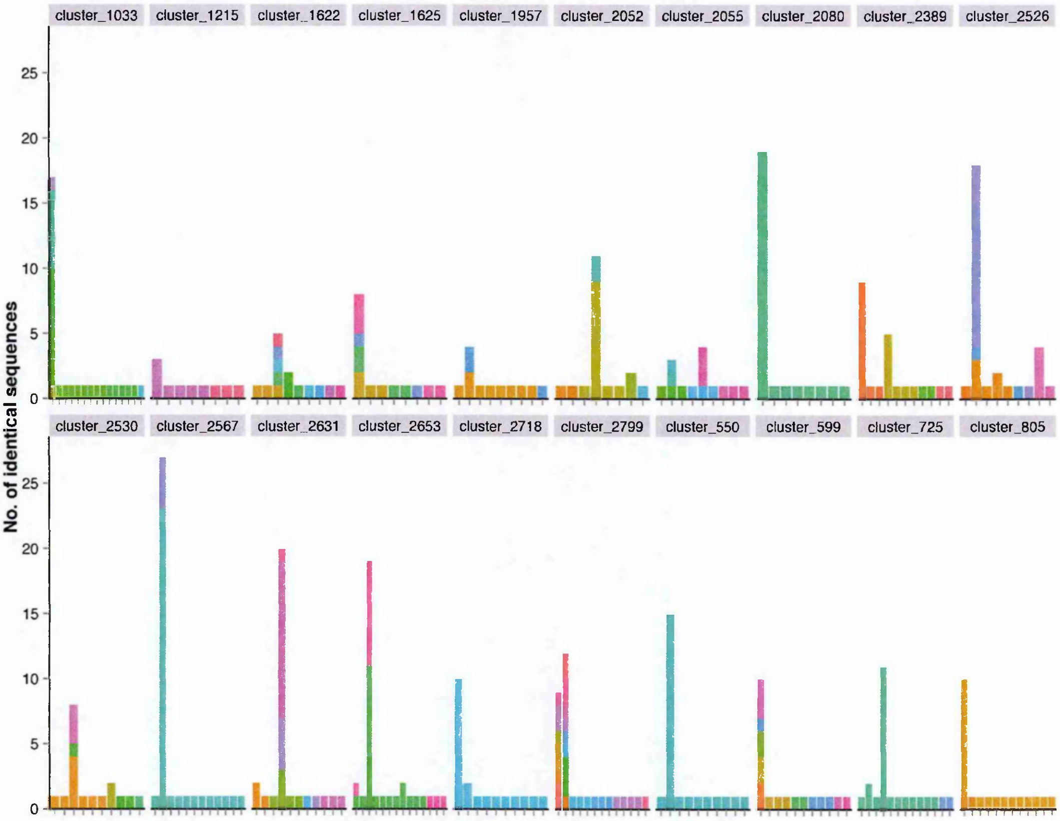


Figure 5.16: *The frequency (y-axis) of identical sequences (x-axis) from a population of var sequences. The frequency of each representative sequence is annotated with the number of isolates that contained an exact 100% identical sequence. Isolates are depicted by the different colours. Sharing of identical sequences was limited to only a few isolates.*

5.3.11 Substitution patterns in *var1* sequences

The *var1* sequences were analyzed in more detail following the observation that they were the least diverse group and that they formed clusters with the most number of sequences across all the identity thresholds that were considered as shown in figure 5.1, figure 5.2 and figure 5.16. Furthermore, *var1* sequences are thought not to be expressed on the surface of the parasite (Winter et al. 2003) and presumably they are not under the same selection pressure as other var genes and in particular group A var genes.

Var1 had a similar profile of nucleotide content as non var1 sequences as shown in figure 5.17. The `G-A/C-T` transitions were over represented in the *var1* sequences. There was little variation at 98% identity compared to 88% identity. The overall number of substitutions at the various identity thresholds were relatively low compared to what was observed in other DBLα sequences.

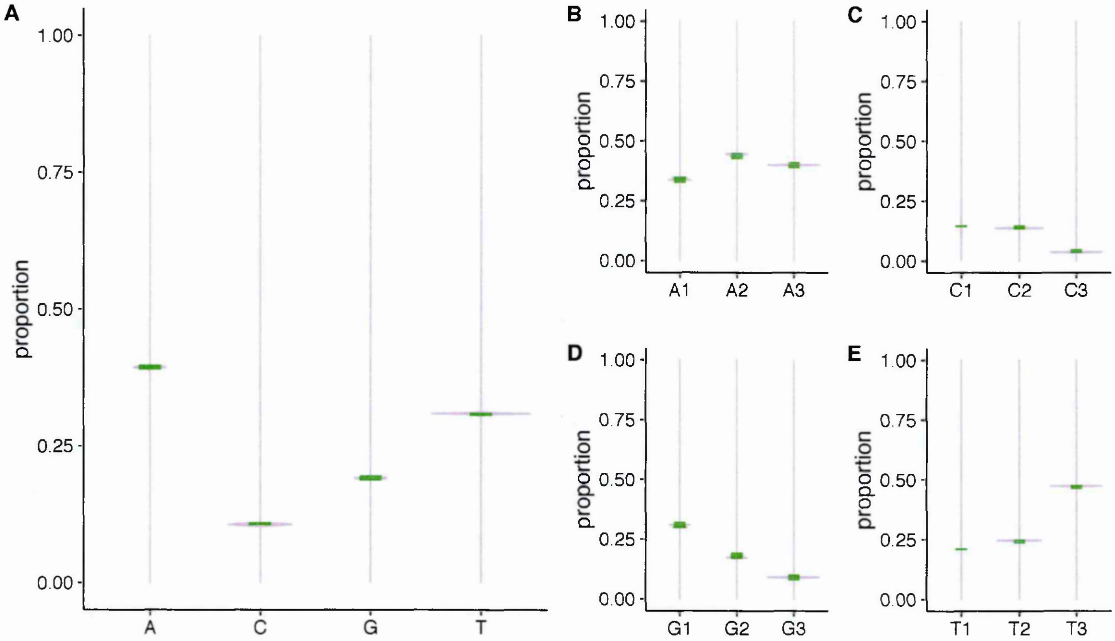


Figure 5.17: *The nucleotide content of var1 sequences is shown below for a total of 103 var1 sequences that were collected from Kilifi in the period 2003-2007. This dataset contained the largest repertouire of var1 sequences from a single location. Figure A shows the overall distribution of nucleotides and figure B-E shows the distribution of each nucleotide at each codon.*

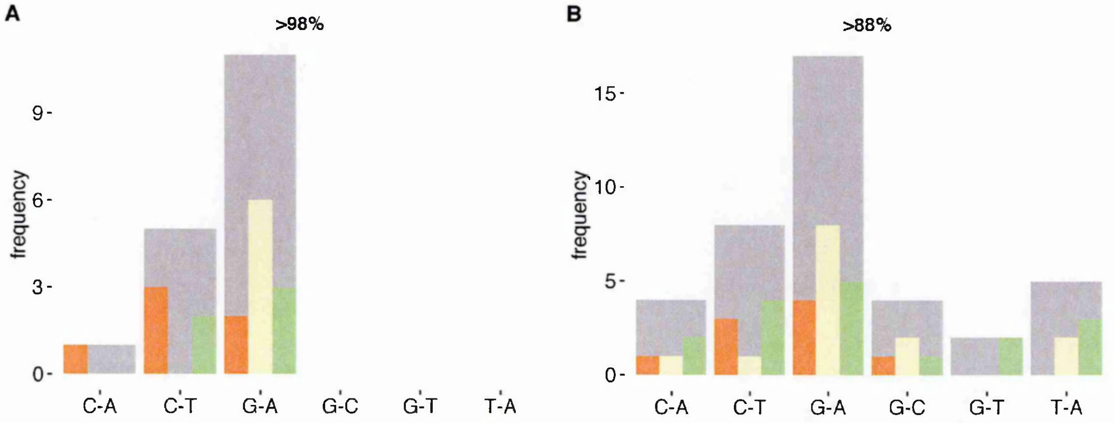
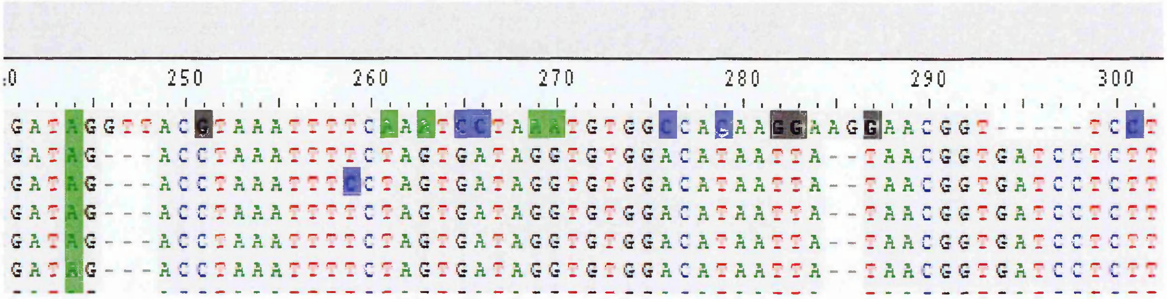


Figure 5.18: *The distribution of substitutions in 103 DBLα var1 sequences at 98% and 88% sequence identity. Var1 were conserved between parasite genomes and appeared to have a constant sequence length of 330 nt. The majority of changes were G-A and C-T transitions. G-C, G-T and T-A changes were not observed at 98% sequence identity and they were relatively few at 88%.*

5.3.12 Insertion/deletions in *var1*

Var1 DBLα sequences contained several substitutions as profiled in figure 5.18. Majority of *var1* substitutions were confined to a region that was between codon 68 and 109.

The sequences from the *var1* alignment had deletions/insertions at two locations relative to the *var1* sequence `60A-1c08`. The first was a codon deletion at position 246-248 and the second was a two nucleotide deletion at position 281-282. Interestingly, sequence `60A-1c08` contained a 5 nucleotide deletion from position 294-298. This somewhat appeared to compensate for the length of the molecule.



```
GATAGGTTACGTAAATTTTCAAATCCTAAATGTGGCCACAAGGAAGGAACGGT-----TCCT
GATAG---ACCTAAATTTTCTAGTGATAGGTGTGGACATAATTA--TAACGGTGATCCTCTT
GATAG---ACCTAAATTTCCTAGTGATAGGTGTGGACATAATTA--TAACGGTGATCCTCTT
GATAG---ACCTAAATTTTCTAGTGATAGGTGTGGACATAATTA--TAACGGTGATCCTCTT
GATAG---ACCTAAATTTTCTAGTGATAGGTGTGGACATAATTA--TAACGGTGATCCTCTT
GATAG---ACCTAAATTTTCTAGTGATAGGTGTGGACATAATTA--TAACGGTGATCCTCTT
```

Figure 5.19: *A truncated image showing an unusual insertion/deletion events in var1 sequences. Deletions at position 246-248 and 281-282 were observed in 102 sequences relative to the first sequence which contained a five nucleotide deletion at position 294-298 thereby maintaining the overall length of the molecule at 330 nucleotides. The above alignment was generated as a codon alignment using the PRANK alignment tool.*

Although this observation is interesting and raises questions on role of var1 sequences, it is also difficult to interprate given that it was only observed once. It could be that the var1 sequence is a more recent recombinant and that the alignment is largely made of ancient sequences. It would be important to explore more sequences from multiple isolates.

5.3.13 *var1* sequences may contain two co-evolving sites

Var1 constituted of clusters with the highest number of members across the similarity thresholds that were considered (98-88%). A phylogenetic tree of var1 sequences is shown in Figure 5.20. Members of *var1* were divided into two groups based on mutations at nucleotide 204 and 246 (codon 68 and at codon 82 respectively).

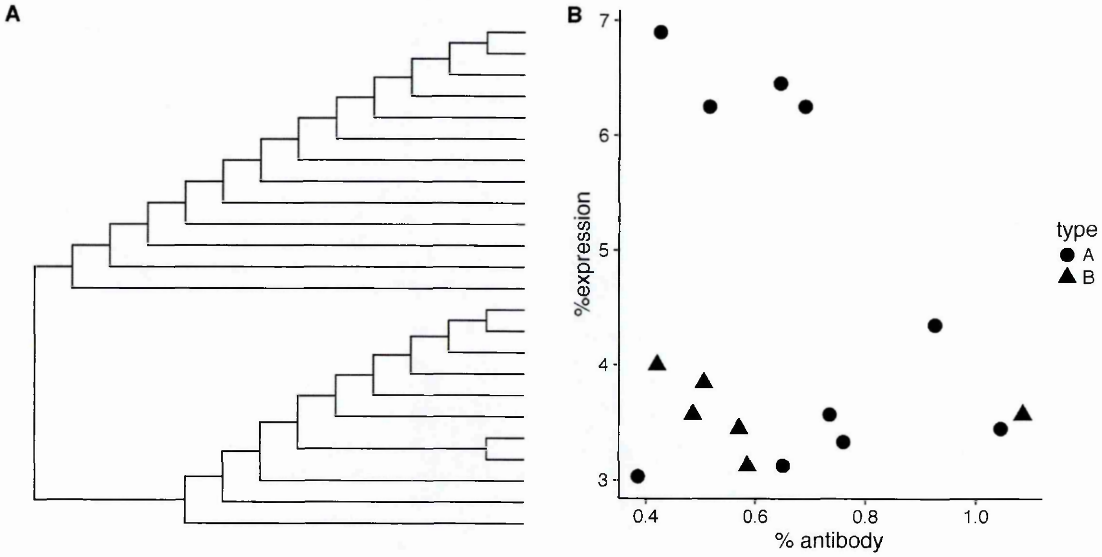


Figure 5.20: ***A.*** *A neighbour joining tree of var1 sequences that were similar at 90% and were collected in 2008-2010. Based on the NJ tree, the sequences could be grouped into two major clades based on mutations at nucleotides 204 and 246.*
B. *A scatter plot of the correlation between the proportion of expression (in percentages) of var1 DBLα sequences collected in 2008-2010 (same sequences in panel A) in each isolate and their respective antibody reactivity.*

To explore the functional relevance of the sequences in each of the two clades, the expression profile for each of the var1 sequences was determined.

	mutation	kaks	lod	freq
1	68	16.50	5.81	0.59
2	82	16.50	5.81	0.59

Table 5.2: *The ratio of synonymous to non-synonymous change (Ka/Ks) was determined in alignment with 57 var1 sequence tags sampled between 2003-2007. Codons 68 and 82 had kaks scores greater than 1 which suggested that they could be under selection pressure.*

	mutation	kaks	lod	freq
1	68	7.00	2.47	0.61
2	82	6.50	2.29	0.57

Table 5.3: *The ratio of synonymous to non-synonymous change (Ka/Ks) for an alignment of 24 var1 sequence tags that were collected between 2008 and 2010 period.*

The explore if variation in *var1* sequences was of functional importance, the Ka/Ks ratio was determined in *var1* sequences from the two study periods. The first dataset comprised of 53 *var1* sequences from samples that were collected in 2003-2007 and the second dataset comprised of 23 *var1* sequences from samples that were collected in 2008-2010. Table 5.2 and 5.3.13 shows the Ka/Ks values in each respective dataset. The table shows two distinct mutations within the var1 sequences and that could be important in structuring this conserved repertoire.

5.4 Discussion

These results show that nucleotide variation is an important mechanism in generation of variation in *var* sequences. Transition changes formed the bulk of all observed changes other than T-A transvertions relative to a random substitutions model. Nucleotide changes at the third codon was pronounced in C-T and T-A substitutions in sequences from Kilifi, PNG and Brazil. Sequences from Amele had a striking bias for changes at the third codon position. The pattern of nucleotide content was consistent across all the sequences from different geographical regions although there was variation in the actual distributions. Interestingly diversity in homology blocks was relatively higher than in the previously described polymorphic blocks.

There was little variation in the pattern of changes between group A and non-group A sequences but there were differences in the absolute number of changes in which case, non group A sequence tags were more diverse than the group A var sequences. This independently confirmed other studies that suggest that group A var genes are more conserved. *Var1* sequences were very conserved and they split into two groups based on two mutations. A thorough analysis with a larger data set is required to confirm these observations.

The distribution of transition and transversions was consistent across sequences from different geographical regions although the profile of substitutions by codon position was remarkably different between sequences from Kilifi, Amele and Brazil. In conserved sequences the majority of substitutions occurred at the third codon position compared to substitutions from random substitution profiles.

The samples that were collected between 2003-2007 had been amplified using the Amplitaq Gold polymerase. Amplitaq has an error rate of 2.6 x 10^{-5} and like most Taq polymerases it lacks the 3'-5' proof-reading exonuclease activity. Samples that were collected in 2008-2010 period were amplified with Bio-X-ACT, a proprietary mix of high fidelity enzymes that contain 5'-3' DNA polymerase activity and 3'-5' proof-reading activity to prevent nucleotide misincorporation during a reaction. The profile of substitutions between sequences that were amplified with Bio-X-ACT and those that were amplified with Amplitaq gold suggested that the choice of polymerase had little effect on distribution of substitutions and presumably PCR errors.

Sequences from Kilifi exhibited a higher density of substitutions in conserved regions that in polymorphic regions in sequences that were very similar to each other. The distribution of changes in these sequences was not contiguous but as the percent threshold was relaxed from 98% to 88%, the number of contiguous changes increased.

Sequences from Amele region that were collected in 2006, showed an interesting profile in the distribution of substitutions (figure 5.13). There was a strong bias for transition changes accompanied by a strong bias for substitutions at the 3rd codon. These sequences were identified as relatively conserved at the amino acid sequence level. The nucleotide content was similar to sequences in Kilifi. This might be explained by a strong purifying selection on these sequences.

DBLα sequences from Brazil are less diverse compared to sequences from African parasites (Albrecht et al. 2010). It has been reported that Brazilian sequences are characterized by high type sharing (Barry et al. 2007). These sequences formed larger clusters at 98% similarity but the distribution of substitutions was not different from sequence in Kilifi and was unlike that which was observed in PNG isolates.

Network-based non-alignment approaches (Bull et al. 2007; Bull et al. 2008; Larremore et al. 2013) have been used to study DBLα sequence tag expression in clinical cases. These studies have reported differences in expression profile of sequences that are expressed in younger children compared to older individuals (Bull et al. 2008; Warimwe et al. 2009). They have also shown that the *var* expression profile is modified by the host immunity (Warimwe et al. 2009), which could suggest that immunity exerts selection pressure on *var* genes.

Var1 (Kyes et al. 2003) is an unusual conserved molecule in *P. falciparum* parasite genomes (Lavstsen et al. 2003; Kraemer et al. 2007) with a distinct 5' promoter region (Vázquez-Macías et al. 2002). *Var1* is constitutively transcribed throughout the intra-erythrocytic stages (Kyes et al. 2003; Lavstsen et al. 2005) and expressed by parasites in nearly all disease phenotypes (Rottmann et al. 2006). These observations suggest that *var1* does not have a specialized function,nonetheless, var1 is actively expressed and maintained in the parasite *var* repertoire.

The function of *var1* sequences in *P. falciparum* genomes is not well understood. Initially it was thought that *var1* could be as important as the pregnancy associated and geographically conserved *var2csa*, but the fact that *var1* is expressed in all disease phenotypes and across the erythrocytic cycle is enigmatic. Figure 5.18 shows the distribution of point mutations in *var1* sequences. Transition changes were the primary sources of single point variation in generating nucleotide diversity in *var1*. Interesting mutations at position 204 and 246 suggested that *var1* can be divided into two separate groups. It would be interesting to explore the functional or structural relevance of such mutations.

In one group of *var1* sequences most of the substitutions were confined between nucleotide position 68 and 109. From table 5.2 and table 5.3.13 an estimate of the Ka/Ks ratio suggests evidence of diversifying selection on the respective codon sites. And raises the question of whether this region is an active recombination site given that the changes appear in a non-random manner and only within this region. The observation that all the changes were confined to these contiguous positions would support this idea. The *var1* orthologue in *Plasmodium reichenowi* is reported to be

1,584 nucleotides longer than in the 3D7 parasite line (Otto et al. 2014) suggesting that *var1* may have lost a functional component.

Var1 sequences are largely conserved in multiple parasites. Point mutations at two distinct positions partition the *var1* sequences into two groups and both position could be under diversifying selection. A thorough examination using *var1* full-length sequences from different geographical locations across different time points is required to ascertain and confirm these observations.

In summary, there were interesting differences in the distribution of substitutions at the third codon in sequences from Amele in Papua New Guinea compared to sequences from Kilifi and from Ariquemes in Brazil. Although *var1* are conserved, a number of substitutions were observed at different similarity thresholds. It would be important to investigate these in relation to functional relevance and structure. A more rigorous population genetic analysis would be required to understand the nature of substitution distribution within the tag sequences and if there is a relation with immune selection. In addition is would be important to re-sequence some of the isolates to confirm the observations.

5.4.1 Limitations of this study

The total number of sequences in most clusters was relatively small and therefore these sequences or clusters could not be considered for population genetic analysis. Secondly, it was not possible to fully account for PCR errors and to derive a quantifiable error rate partly due to lack of appropriate reference sequences. Thirdly, it was not possible to assign the direction of a substitution and therefore the total number of substitutions from one nucleotide to another were pooled together.

5.4.2 Future work

This study developed a simple approach to explore substitution patterns in highly diverse and recombining sequences. Some of the methods are adhoc and could benefit if more sophisticated approaches of molecular evolution are applied.

The study of selection and the question of validity of observed subsitutions could easily be done with deep sequencing. This would allow more sophisticated and state of the art approaches to be applied in the study. By setting coverage thresholds for known mutations, it would be possible to analyze structural changes, SNPs as well as minority variants in multiple clinical samples.

Although, mosaicism is a challenge in assembly of full length *var* genes, this study would have benefited from looking at how changes are distributed along the entire length of PfEMP1 sequence in parasite isolates collected from areas that are under different transmission intensities. It would also be important to compare the substitution bias across the genome relative to the telomeric genes.

One of the interesting observations was that *var1* sequences had a similar distribution of substitutions compared to other *var* sequences. This raises an important question on the role of *var1*. On the other hand *var1* sequences had an unsual pattern of variation which raises the question on whether *var1* contain a recombination break-point?

Further investigation should focus on comparing the distribution of changes in recent *var* sequences relative to ancient sequences.

Chapter 6
Identifying potential epitopes using a predictive approach

6.1 Background

In this chapter, a computational approach is described which was used to explore the epitope diversity in DBLα sequences and to identify commonly occurring epitopes and epitope positions. Epitope predictions were performed against linear B-cell epitopes and T-cell epitopes using a limited group of MHC class II T-cell alleles. Two main approaches were used to investigate immune selection on predicted epitopes as outlined below,

1. Explored whether the regions associated with these epitopes were associated with specific patterns of nucleotide substitution using methods described in chapter 5.

2. By using an existing data set of *var* expression data to test whether expression of commonly occurring predicted MHC-class II PfEMP1 epitopes were correlated with host age, to explore whether expression of common epitopes within these key parasite antigens are selected against as children develop naturally acquired immunity.

6.2 Results

B and T-cell epitopes were identified using an in-silico epitope prediction approach as described in the methods development chapter 3 section 3.8.4 and in section 3.8.1.

For Bepipred based B-cell epitope predictions the cutoff threshold for epitope was the default score of 0.35 as described in chapter 3. NetMHC-II T-cell epitopes were identified as overlapping sets of peptides. They were categorized as strong binders, weak binders (IC_{50} >50nM and IC_{50} < 500nM) and non-binders (IC_{50} > 500nM) based on the predicted IC_{50}. The predicted IC_{50} was a surrogate measure of the strength of binding.

The Predivac server was used to predict and score peptides based on their ability to form stable MHC-peptide complexes. Predivac assigns values between 1 and 100 to the predictions. Peptides with a prediction value of >60 are assigned as epitopes. In this study, a conservative approach was adopted and only predicted peptides that were within an arbitrary 95th percentile of the scores were considered as immunodominant.

The T-cell prediction results were presented based on the predicted affinity scores and the predicted ability to bind to five or more HLA alleles. A conservative approach was adopted such that only predicted strong binders were considered for further investigation.

6.2.1 Linear B-cell epitopes are correlated with sequence diversity

Linear B-cell epitopes were predicted with Bepipred (Larsen et al. 2006) (`http://www.cbs.dtu.dk/services/BepiPred/`) as described in chapter 3 section 3.8.1.

The mean Bepipred scores for an alignment of sequences are shown in figure 6.1. The predicted linear B-cell epitope scores were significantly correlated with hydrophobicity (Kyte and Doolittle 1982) and sequence diversity, calculated using Shannon's diversity index (Shannon 1948). Regions with high Bepipred epitope scores were the most diverse as shown in figure 6.1. Group-A and non-group-A sequences had a slightly different binding profile which was more detectable in the 3' region of the molecule.

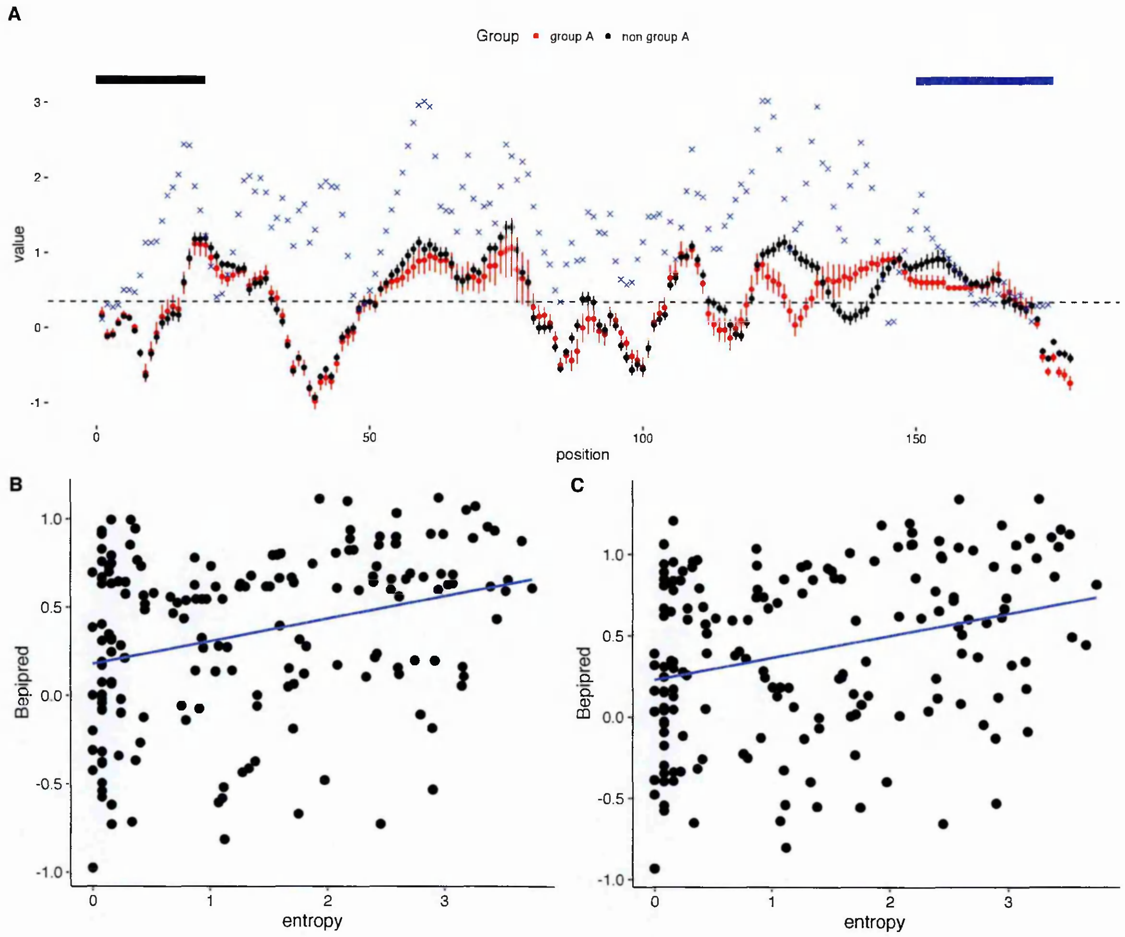


Figure 6.1: *Figure A shows a plot of the mean Bepipred scores and sequence diversity in aligned and randomly selected DBLα sequences. The red dots represents Bepipred scores at each position in group-A sequences and the black are the Bepipred scores from non-group-A sequences. The blue crosses represent the mean Shannon diversity index calculated using a window size of 5 amino acids along the alignment columns. There was a significant positive correlation(rho=0.3,p-value < 0.001) between the mean Bepipred scores and the sequence diversity. Differences in the Bepipred scores between group-A and non-group-A sequences were more pronounced towards C-terminus. The black and blue lines on figure A represent the homology blocks D and H respectively. Figure B and C shows the correlation between diversity and Bepipred scores in both group A and non-group A respectively.*

6.2.2 The DBLα sequence region contains potential MHC class II T-cell epitopes

There are several approaches and tools for predicting T-cell epitopes. NetMHCIIpan (Nielsen et al. 2010; Karosiene et al. 2013) is a pan-specific approach to epitope prediction and was shown to outperform several other prediction tools and therefore it was adopted for the purpose of T-cell epitope prediction in this analysis and as described in chapter 3 section 3.8.4. Because it is known that the length of a peptide often correlates with binding strength (O'Brien et al. 2008), the length of the peptides were arbitrary chosen as 9 amino-acids 15 and 18 amino acids to represent short medium and long peptides respectively. A total of 39 arbitrarily chosen HLA alleles were used in the predictio; 13 were HLA-DP-DQ and 26 were HLA-DR alleles. A total of 365 non-redundant DBLα sequence tags from Kilifi collected between 2008 and 2010 were used as input. It was not computationally possible to use all the HLA alleles for prediction given the limited resources. Secondly, fi there was an important signal, the rationalle should have picked it with predictions from the limited HLA predictions.

The binding profiles for predicted strong binders are shown in figure 6.2. The upper and lower panels represents peptides with 15 and 18 amino-acids respectively. There were no predicted strong binders for peptides with 9 amino-acids presumably because they were too short for MHC class II binding.

A peptide was considered a strong binder if the predicted binding affinity in IC_{50} was less than 50nM and a weak binder if the binding affinity was between 50nM and 500nM. Peptides with IC_{50} above 500nM were considered as a non binders. These are the defined cutoff thresholds for binders and non-binders. The total number of predicted binders and non-binders for each peptide group is summarized in table 6.1.

	strong_binders	weak_binders	non_binders
9 amino acids	0	334	1339961
15 amino acids	11956	122173	1135752
18 amino acids	9376	157241	1068251

Table 6.1: *The frequency of overlapping strong, weak and non binders predicted with NetMHCII from 365 DBLα sequences. There were no strong binders from short peptides with a length of 9 amino acids. The most strong binders were from peptides with 15 amino acids. Peptides with 18 amino acids had fewer strong binders byt the majority of weak binders.*

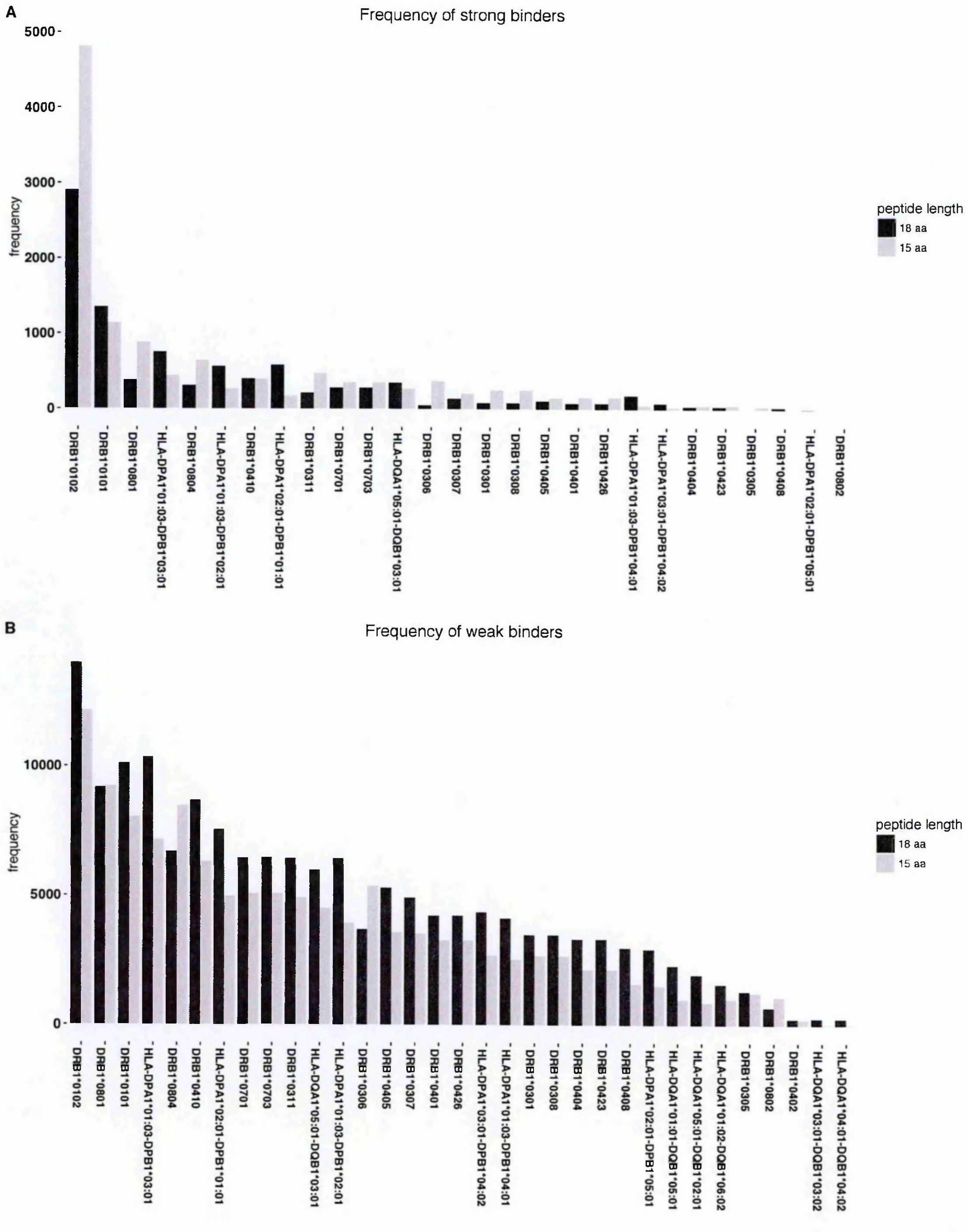


Figure 6.2: *The frequency (y-axis) of predicted NetMHCII peptide cores per HLA allele (x-axis). Figure A shows the frequency of strong binders. The frequency of 18-mers and 15-mers is shown in black and grey respectively. Within the strong binders, there was a lot of variation in the frequency of predicted binders for each HLA allele under consideration. The peptides with 18 amino acids were largely weak binders.*

6.2.3 Predicting peptides with stable peptide-MHC complexes in DBLα sequences

Formation of stable peptide-MHC (p-MHC) complexes is important for immunodominant epitopes. Mutations in a peptide alter the strength of binding and it is important to identify the key residues that are important for a peptide-MHC interaction. The most immunodominant epitopes could not be identified based on NetMHCIIpan affinity values. The Predivac server (Oyarzún et al. 2013) was used to predict core epitopes that form stable peptide-MHC complexes. The pMHC kinetic stability is important in controlling the MHC class II peptide immunogenicity.

Immunodominant epitopes were defined as peptides that are predicted to induce the most potent immune responses from a set of immunogenic peptides competing that bind to the same MHC molecule. The Predivac server allowed only one sequence and one HLA-DR allele input at a time. A Python script, predivac-util (listing A.22, described in chapter 3 section 3.8.5), was developed to automate sequence submission and retrieval from predivac server.

Predivac was developed using high-affinity binding data in the belief that there are peptide features that correlate with promiscuity and immunodominance. Based on the prediction criteria, peptides are assigned a normalized binding score between 1 and 100 where 100 are the strongest binder. These scores are obtained by establishing a correlation between the specificity determining residues (SDRs) in the HLA query protein and the SDRs associated with HLA proteins of known specificity as described in chapter 3 section 3.8.4. According to Predivac developers a predicted peptide score >60 was used to discriminate potential epitopes from non-epitopes.

For the purpose of this analysis, an arbitrary conservative definition of a Predivac epitope was defined to only consider peptides that were in the top 5% of the scores. Promiscuous binders were defined as peptides that were within the top 5% of the predicted binders and could bind to an arbitrary chosen minimum of 5 of the 27 number of HLA alleles that were tested.

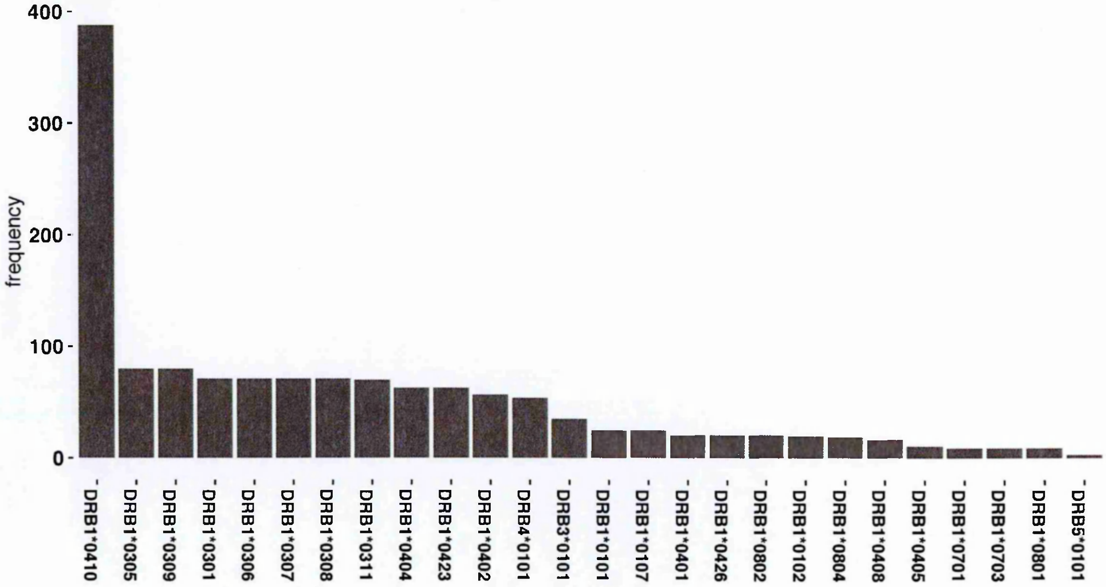


Figure 6.3: *A bar plot of the frequency of predicted immunodominant HLA class II peptides using the Predivac server. Predivac scores predicted core peptides on a scale of 1 - 100. Peptides with scores >60 are defined as epitopes. Only the 95th percentile of epitopes are shown. HLA allele DRB1*0410 had the highest frequency of predicted peptides.*

In relation to NetMHCII predictions, Predivac predictions shown in figure 6.3 were considerably fewer. This would be explained by the fact that only a few epitopes are likely to be immunodominant. It is also important to note that a very conservative threshold was considered and that could partly explain the fewer numbers.

6.2.4 Overlapping peptides predicted by both NetMHCII and Predivac were few

Using Predivac, a total of 1,157 core peptides were associated with strong binders and 329 core peptides were predicted to form strong peptide-MHC complexes. A total of 98 core peptides were predicted by both NetMHCII and Predivac. A total of 34 out of the 98 peptides were predicted to bind to 5 or more HLA alleles. Figure 6.4 illustrates the overlap between the predicted strong binders (netMHCII), strong pMHC complexes (predivac) and binding to multiple alleles (promiscuous).

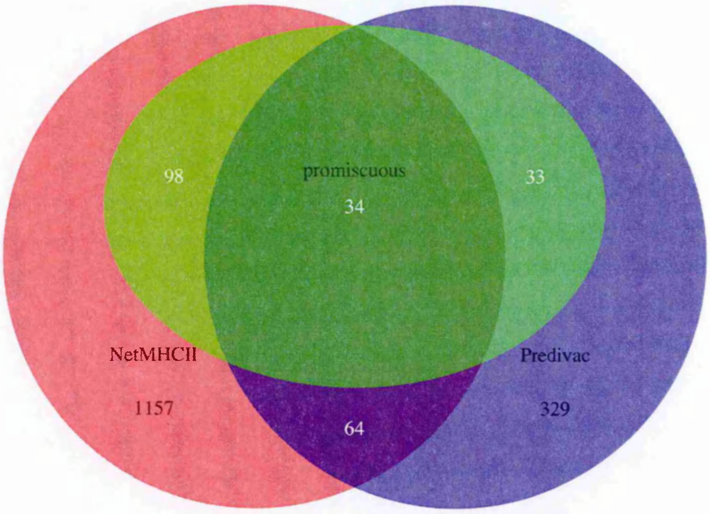


Figure 6.4: *A venn diagram showing the total number of overlapping peptide cores that were predicted by both netMHCIIpan as strong binders and by Predivac server as immunodominant. NetMHCII peptide cores were based on the 15 amino-acid long peptides. Promiscous peptides were defined as core peptides that were predicted by both NetMHCII and predivac and were also predicted to bind to five or more HLA molecules respectively.*

6.2.5 The frequency of predicted promiscuous epitopes is limited

Conservation of the 34 promiscuous epitopes among DBLα sequence tags from Kilifi, Amele and Brazil was calculated using the epitope conservation tool (Bui et al. 2007) (`http://tools.immuneepitope.org/tools/conservancy/iedb_input`). Peptide conservation was defined as the best local alignment of a peptide in a DBLα sequence. The percentage identity was calculated from the number of peptides that matched the aligned sequences above 90%. Table 6.2 shows the name of the peptide, the number of identical hits, the sequence search space, and the percent conservation in each of the 34 promiscuous epitopes and listed from the most to the least conserved. The sequence search space comprised of sequences from Kilifi, Papua New Guinea and South America.

Peptide	Hits	Sequences	Conservation
ISGDTKVFT	59	13974	0.42
IYKDLKDLH	45	13974	0.32
YEGLKNNGA	42	13974	0.30
LTCDARNNA	35	13974	0.25
YQKDAPNYY	25	13974	0.18
YEGLSNNGA	21	13974	0.15
IWRALTCHA	22	13974	0.16
IRNDDRTLK	15	13974	0.11
LYFDGRCGR	14	13974	0.10
IQLEERLEQ	12	13974	0.09
YLGDVRTTL	13	13974	0.09
FKKIYNKLI	11	13974	0.08
IYKSLTPEA	9	13974	0.06
FRNTCSSKS	10	13974	0.07
IGADGRVTE	11	13974	0.08
VKLSNNLRA	7	13974	0.05
FGKIYNKLI	8	13974	0.06
LYFDDRCGR	9	13974	0.06
YNGLSNNGV	6	13974	0.04
YKNLKNPAQ	4	13974	0.03
FLNIQNDNS	5	13974	0.04
LKLEEKLKQ	4	13974	0.03
ITCDNRLRG	5	13974	0.04
YGSDTRNYY	5	13974	0.04
FRFTCSKGV	2	13974	0.01
IYEDLKDLH	2	13974	0.01
IYKDLKDGV	2	13974	0.01
IYKDLKDAK	3	13974	0.02
LKGDARTHY	2	13974	0.01
IHKDVTNRK	1	13974	0.01
LMEDLKNDR	2	13974	0.01
YKNDTKNYY	2	13974	0.01
ITCDARHDA	3	13974	0.02
LTCDDKLSK	1	13974	0.01

Table 6.2: *The conservation profile of 34 predicted promiscuous epitopes among from a dataset of sequences from Kenya, Papua New Guinea and South America.*

6.2.6 Promiscuous peptides from group A-like sequences are conserved

Sequence hits from table 6.2 were classified based on the Cys-POLV classification as shown in figure 6.5. The most conserved peptides were from group A-like sequences with the exception of peptide **LTCDARNNA** which is conserved largely in CP4 sequences.

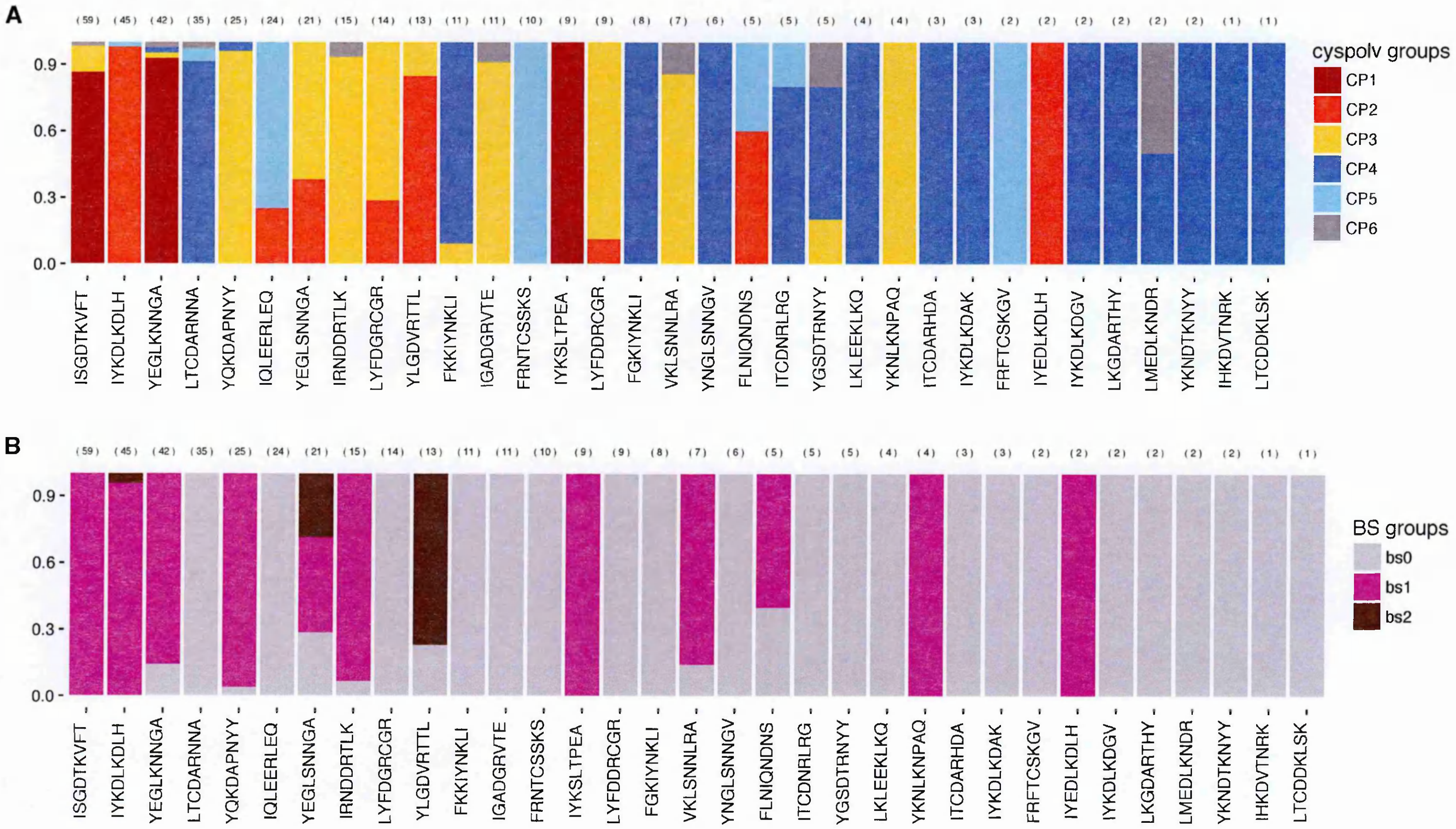


Figure 6.5: *The Cys/POLV classification of sequences containing each of the predicted promiscuous peptide. The frequency of the sequences is shown on top of each bar in parentheses and the sequences are ordered from high to low frequency with the most frequent sequences on the left. Most of the sequences containing the peptides were from group-A like sequences. The lower panel shows the distribution of block-sharing groups among sequences containing each peptide.*

From figure 6.5 (A), the least conserved peptides were from non group A like sequences. The most conserved peptides came from sequences belonging to block-sharing group 1. Three of the peptides were from sequences containing block-sharing group 2 as shown in figure 6.5 (B).

6.2.7 Predicted promiscuous peptides were expressed largely in group-A sequences

The relation between the 34 promiscuous peptides and *var* expression was evaluated using data from clinical isolates collected in Kilifi. Another evaluation was carried out using a randomly chosen set of 34 sequences that had been predicted as strong binders. The random peptides were selected from a collection of strong binders. The aim was to determine if the peptides were expressed in a subset of PfEMP1 proteins from commonly recognized parasites. Table 6.3 shows the correlation between the 34 peptides and expression based on Cys-POLV groups.

	CP1	CP2	CP3	CP4	CP5	CP6	BS1_cys2	BS1_cp2
Random peptides	-0.22(0.000)	-0.02(0.741)	-0.22(0.000)	0.23(0.000)	0.17(0.002)	0.00(0.934)	-0.23(0.000)	-0.03(0.588)
Promiscuous peptides	0.12(0.028)	0.14(0.013)	0.25(0.000)	-0.04(0.435)	-0.02(0.775)	0.08(0.143)	0.22(0.000)	0.13(0.017)

Table 6.3: *A table of spearman correlation in the expression profile of 34 promiscuous binders that were predicted by both NetMHCII and Predivac and could bind to 5 or more HLA alleles. The random binders were selected from a list of strong binders that were predicted by NetMHCII*

There was a positive correlation between the sequences and group A-like sequences (BS1-cys2) and CP3 sequences. A positive correlation was also observed with sequences from CP1 and CP2. There was no correlation between the promiscuous and CP4 sequences. The random peptides had a negative correlation with group A-like (BS1-cys2). These random peptides showed a negative correlation with CP1, amd no correlation with CP2. A negative correlation with CP3 and a positive correlation with CP4 was observed.

6.2.8 Relationship with clinical classification

Clinical classification was determined in parasite isolates expressing sequences containing one or more of the 34 promiscuous peptides. A total of 178 isolates had clinical information. The majority of the isolates (n=78) were from individuals with severe disease and the rest (n=55) were from individual with non-severe disease. A few (n=6) isolates were from individuals with asymptomatic disease.

6.2.9 Relationship between peptide expression and age of the host

Table 6.4 shows the that there was no correlation observed between sequences containing the promiscuous epitopes and age of the host although. No correlation was observed with the random peptides either.

	age
promiscuous peptides	-0.02(0.720)
random peptides	0.02(0.737)

Table 6.4: *The correlation between host age and the expression of sequences containing the 34 promiscuous peptides shows that there was no evidence of an association.*

6.3 Discussion

The results show that linear B-cell epitopes are associated with diverse regions and that a few MHC-class II epitope peptides that were widely recognized by HLA alleles are conserved in sequences that were expressed by individuals with severe disease and were associated with young host age. The most conserved promiscous T-cell epitopes were identified in a subset of DBLα sequences that were associated with severe disease and young host age as shown in figure 6.5.

6.3.1 Linear B-cell epitopes are prevalent in polymorphic regions

Predicted linear B-cell epitope scores were higher in amino acid residues that were found in polymorphic and hydrophobic regions of the DBLα sequence tag as shown in figure 6.1. One explanation for this is that Bepipred takes into account Parker's hydrophilicity scale to predict B-cell epitopes, therefore it is not surprising that hydrophobic residues provided high Bepipred scores. What is interesting is that, variation in these regions appears to converge to residues with specific physico-chemical properties which may suggest surface exposure and potentially an interaction with antibodies. Using a Markov model derived from known antibody-antigen binding data is an improvement to the prediction accuracy (Larsen et al. 2006).

Evidence to support the importance of short linear epitopes comes from laboratory studies using peptide arrays and predicted linear epitopes. A predicted linear peptide in AMA1 was shown to be recognized by serum samples that also recognized domain II of Plasmodium vivax AMA-1 (Bueno et al. 2010). Ditlev et al. (2011) used Bepipred to identify B-cell epitopes in DBL4ϵ of var2CSA in the FCR3 isolate (Ditlev et al. 2011). These studies showed that ELISA derived peptides could be predicted in-silico and that the peptides showed high reactivity to sera from multi-gravid women. There is evidence to suggest that rosetting is mediated by the variable regions of the PfEMP1 protein and that short linear-peptides in sub-domain 1 and 2 of the DBL1alpha domain contain anti-rosetting features (Albrecht et al. 2013).

One of the main drawbacks of linear B-cell epitopes is that B-cell cross-reactive epitopes are likely to be discontinuous and that the prediction accuracy of linear B-cell epitopes is poor (Blythe and Flower 2005). Discontinuous epitopes rely on prediction of accurate antigen-antibody structures. In relation to this work, regions that interact with antibodies would be under diversifying selection. There was not enough evidence to suggest that diversifying selection was acting on regions that were identified with high B-cell epitopes. Moreover the DBLα sequence tag is short relative to the full-length molecule and may not be key to B-cell immunity, at least not in a direct way.

6.3.2 Promiscuous T-cell epitopes are conserved in sequences associated with expression in severe malaria

To identify potential important regions of the DBLα molecule, two pan-specific methods were used to predict MHC class II T-cell epitopes from 39 arbitrary selected HLA-alleles. A total of 34 promiscuous peptides were identified based on a conservative filtering criteria. The peptides were not very conserved in a majority of the sequences. The most conserved peptide was only present in a subset of sequences containing the MFK motif. Although the MFK motif is often associated with cys2 sequences, this motif was found in a cys4 sequence that contained an MFK motif and a few sequences in CP3 (figure 6.5). This observation may be a sequencing error. It is difficult to ascertain or drawn any conclusion from the single observation.

Peptide elongation has been reported to increase the MHC class II binding affinity (O'Brien et al. 2008). From table 6.1 the frequency of strong binders was high in 15 amino acid peptides suggesting that this was the preferred optimal length for a peptide. Nonetheless, it is still difficult to discern the optimal length of the binders without experimentation. There were no strong binders with 9 amino acid peptides, which suggests the importance of residues other than the core peptides.

Although pan-specific methods are developed to cope with HLA class II allele diversity (Nielsen et al. 2010), most pan-specific tools predict epitopes for only a single locus (mostly the HLA DR). HLA restriction was a major challenge in T-cell epitope prediction given that HLA alleles identify peptides with varying affinities. If an allele recognizes a peptide, it is not guaranteed that a different HLA allotype will recognize and bind to the same or similar peptide. Another challenge in MHC class-II T-cell epitope prediction is the breadth of available tools and the respective accuracy. An important epitope must be recognized by a sufficiently large number of MHC class II alleles and at the same time it should be sufficiently conserved among the PfEMP1 molecules.

6.4 Conclusion

Based on the limited data set, sequences containing the predicted and promiscuous epitopes were found to be expressed in individuals with severe disease. It would be important to confirm the ability to raise T-cell immune responses against the predicted binders and seroprevalence studies on recombinant tags would be potentially useful.

Chapter 7
Conclusion

7.1 Molecular diversity

The immense molecular diversity in *var* genes (Barry et al. 2007) is generated largely through recombination (Kirkman et al. 2014; Sander et al. 2013; Zilversmit et al. 2013; Duffy et al. 2009; Freitas-Junior et al. 2000; Kerr et al. 1994). Mitotic recombination could account for majority of the sequence variation that is observed within an isolate (Bopp et al. 2013; Claessens et al. 2014) and meiotic recombination accounts for the majority of observed variation between different isolates and between populations. Lack of a suitable or defined *var* gene epidemiological sampling framework is challenge for exploring both sequence and molecular diversity including the role of single nucleotide substitutions. Due to recombination, sequences from diverse isolates tend to contain sequence blocks with disparate "ages" which complicates measures of evolution and tests of selection.

Figure 7.1 shows how new antigenic determinants can be generated through random mutations and selected by antibodies and thereby shape the *var* gene repertoire. Evidence for geographical conservation of antigenic determinants came from the Aguiar study that showed that despite presence of antigenic diversity, there was considerable pan-agglutination of parasite samples by serum collected from diverse malaria endemic geographical regions (Aguiar et al. 1992). This suggested that there are shared surface antigen epitopes between isolates from different geographic regions.

A functional role may constrain serological diversity and is particularly important to identify such proteins because they are potential vaccine candidates and they can protect against specific forms of malaria. On the other hand, and as illustrated in figure 7.1 serological diversity may be restricted by functional constraints that restrict diversity.

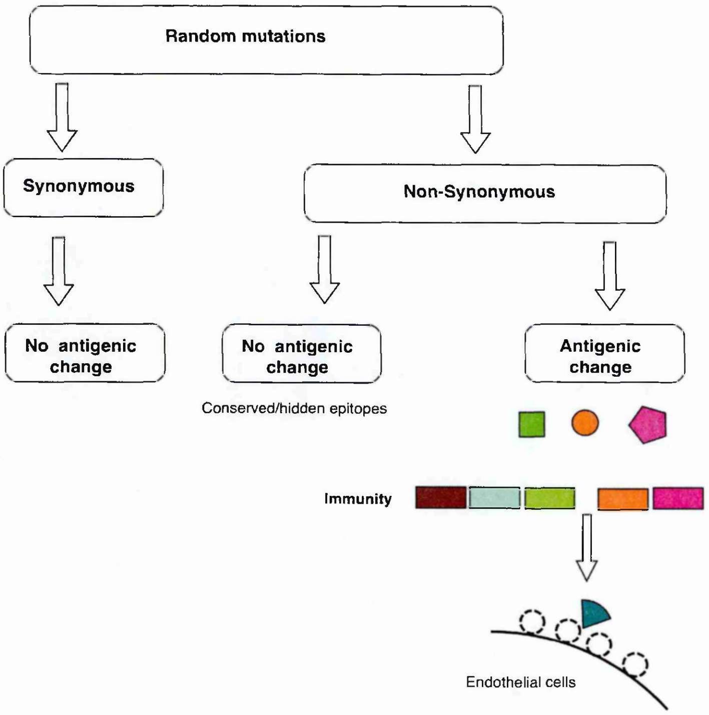


Figure 7.1: *Non-synonymous mutations random mutations contribute to new antigenic types. The host immunity structures antigenic repertoire. A successful antigenic variant is recquired to escape the host immunity and maintain binding to one or more receptors. The balance between novel antigenic diversity and maintenance of a functional role for example binding to endothelium, is thought to be important in structuring the var gene repertoire.*

7.2 A recap of the aims of the study

This study attempted to understand the relationship between sampled sets of *var* sequences from clinical parasite isolates by comparing approaches for classifying *var* sequences. The study was extended to explore the pattern of sequence diversity in DBLα sequences by comparing pairwise differences between groups of similar sequences at different sequence identity thresholds and later comparing patterns of changes in relation to predicted epitopes regions.

An unexpected observation was that at 98% and 96% similarities, the frequency of substitutions within DBLα homology blocks between isolates was relatively higher compared to substitutions within the polymorphic blocks.

7.3 Summary of methods

Sequences data in this study was prepared using approach than utilized a consensus reads read alignment using CAP3, a sequence assembly tool (version date 12-21-07) (Huang and Madan 1999) and Vsearch (version 1.0.16), a tool for processing metagenomic sequences, including searching, clustering, sorting, masking and shuffling. Sequences were clustered using a global alignment procedure and then grouped into clusters based on similarity at a given thresholds. An immunoinformatic approach was used to identify potentially immunogenic peptides in the DBLα tag. Predicted epitopes were evaluated based on the expression profile in samples that were collected in Kilifi and the age of the host.

7.4 Understanding relationships between sequences and classification approaches

In chapter four, approaches for measuring the expression of *var* genes from clinical studies were compared with Cys/PoLV classification using 306 published sequences from 3D7, DD2, HB3, IT4, RAJ116, PFCLIN and IGH3 laboratory isolates (Rask et al. 2010). Results from chapter four showed that;

- Distinct Cys/PoLV groups are associated with *upsA* genes. Here we developed an approach to rapidly identify group-A sequences which highlights the importance of the Cys/PoLV classification method in predicting features of the full-length molecule.

- DBLα domains represent very heterogeneous sequences. Therefore using trees to define these sequences maybe misleading. Mutually exclusive sequences can be found among collections of sequences that have been placed in different subgroups of the Cys/POLV classification.

- Cys/PoLV groups do not predict sequences with DC8 cassettes given global collection of sequences but may do so within a restricted geographical location such as East Africa.

7.5 The role of mutations in generating diversity

Data from chapter five showed that point mutations contribute to the overall *var* diversity. The analysis approach attempted to minimize the effect of PCR errors, but without a quantifiable approach of accounting for PCR errors, it is difficult to make a firm statement on the role of these substitutions. In summary,

- In similar sequences, diversity in the homology blocks was relatively higher than in the polymorphic blocks in sequences. The known homology blocks were defined based on collection of DBL domains from diverse sequences and presumably they contain recombining sequence blocks of different ages. By looking at sequences collected from the same geographic area and that were clustered at high similarity, the observed differences can be attributed to real substitutions.

- The T-A substitution was the most frequent transversion and had higher frequency at the third codon compared to the first and the second codon. From the random model of substitutions, there was no difference in the number of T-A substitutions. Even upon accounting for the ratio of transitions to transversions, the relative frequency of T-A substitution did not differ in a proportionally. Sequences from Kilifi, differed in a statistically significant way in the distribution of substitutions. Therefore the differences in distribution of T-A changes cannot be fully accounted by T-A proportion in the sequences.

- Sequences from Kilifi isolates showed a different substitution pattern compared to sequences from Papua New Guinea. Sequences from the Amele region in Papua New Guinea had a striking substitution bias at the third codon position across all types of substitutions. We considered the possibility that this difference might merely be due to composition bias at the third codon within these sequences. Interestingly there were no differences in the substitution profile compared to sequences from Kilifi.

- Substitutions within the homology blocks were less frequent and they were predominant at the third codon. `C-T` transversion was more prevalent in homology block H than homology block D.
- Substitution bias at the third codon was observed in homology blocks D and H from distantly related sequences.

Previously we have identified *var1* DBLα sequences using sequence motifs. The clustering approach was successful in independently identifying *var1* sequences based sequence similarity. Chapter 5 showed that *var1* genes are characterized by a mutation hot-spot that may be under diversifying selection or active recombination. *var1* showed a different pattern of substitutions in that **GT** changes were rare at 98% and 88% similarities. Overall the proportion of **GA** and **CT** changes was similar to what was observed in other *var* sequences. Unlike other *var* sequences, *var1* sequences that were sampled in this study had a distinct uniform length of 330 nucleotides.

7.6 Prediction of epitopes and relationship with age and expression

Chapter 7 explored predicted linear B-cell epitopes and MHC class II T-cell epitopes in relation to expression in *var* sequences and host age. Predicting B-cell and T-cell epitopes is not trivial given the sequence diversity and MHC allele diversity. The predictions were limited to a few arbitrary selected MHC class II alleles. Linear B-cell epitopes scores were correlated with sequence diversity and the hydrophobicity profile which could be due to the fact that Bepipred linear B-cell epitope prediction partly relies on propensity scales. The most diverse regions were potent B-cell epitopes and there was a difference in B-cell epitope predictions between group-A and non-group-A DBLα sequences towards the N-terminal as shown in figure 6.1.

In this study a conservative approach was adopted and only 34 promiscuous peptides were identified using two MHC class II T-cell epitope prediction tools. This is a very limited set of peptides to make general conclusions about predicted epitopes. Furthermore, the number of HLA alleles that were considered were also a limited set. Nonetheless, it is important to appreciate that these limited set of epitopes were associated with expression in group-A *var* genes, secondly a large proportion of the sequences containing these peptides were from severe malaria individuals and the rest from non-severe cases. Only a tiny proportion (n=6) was from asymptomatic individuals but that may be because there was only a few asymptomatic individuals.

7.7 Further work and conclusion

This work was an exploratory description of variation in *var* sequences based on the 300-500 base-pair DBLα sequence tag. For a more comprehensive analysis focus should shift to full length sequences. This could be better achieved by making use of parasite cloning and sequencing of *var* genes to confirm the variation bias in C-T/G-A transitions as well as collection of newly available sequences from around the world.

It would be interesting to investigate role of potential mutagenic events in plasmodium parasites. An interesting focus is the role of cytosine deamination during transcription. It has been shown that B-cell-specific-activation-induced cytidine deaminase (AID) works directly on DNA to deaminate deoxycytidine (Petersen-Mahrt et al.

2002). AID is thought to be responsible for antibody gene diversification through gene conversion and class-switch recombination although the exact mechanism is not well understood.

Studies in *E. coli* have shown that deamination of cytosine leads to uracil U, uracil simulates thymine (T) in DNA and pairs with adenine (A), and after two rounds of replication the substitution from C/T is generated (Francino and Ochman 2001). In the human genome, CpG dinucleotides have been shown to mutate at a higher rate because of cytosine deamination. Methylated CpG dinucleotides go through deamination of 5-methylcytosine to produce thymidine (Coulondre et al. 1978). Given these accounts, it would be interesting to investigate the role of mutagenic events and in particular, DNA polymerases and mismatch repair in *var* sequences. This study showed an relatively higher number of G-A and C-T substitutions although it was difficult to assign direction to an observed substitution. It is possible that some substitutions maybe generated during replication or transcription. Whether this is a deliberate effort by the parasite to allow errors to persist remains to be investigated.

B and T-cell epitopes should be confirmed by peptide synthesis and testing for live cell surface reactivity and induction of T-cell responses respectively. Furthermore the predictions should be based on a wider panel of HLA alleles that comprise of alleles from the local population from where the acute samples were collected. Potential source of this data is whole genome sequences. A combination of prediction and empirical epitope discovery may be more useful in exploring PfEMP1 based epitopes. Examples include the approaches that were considered by Patarroyo et al. (2014) and Albrecht et al. (2013) and discussed in chapter 1 section 1.10.2

References

Acinas, Silvia G, Ramahi Sarma-Rupavtarm, Vanja Klepac-Ceraj and Martin F Polz (2005). 'PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample.' *Applied and environmental microbiology* 71.12, pp. 8966–8969.

Aguiar, J C, G R Albrecht, P Cegielski, Brian M Greenwood, J B Jensen, G Lallinger, A Martinez, I A McGregor, J N Minjas, J Neequaye, M E Patarroyo, J A Sherwood and R J Howard (1992). 'Agglutination of Plasmodium falciparum-infected erythrocytes from east and west African isolates by human sera from distant geographic regions.' *The American journal of tropical medicine and hygiene* 47.5, pp. 621–632.

Albrecht, Letusa, Catarina Castiñeiras, Bruna O Carvalho, Simone Ladeia-Andrade, Natal Santos da Silva, Erika H E Hoffmann, Rosimeire C dalla Martha, Fabio T M Costa and Gerhard Wunderlich (2010). 'The South American Plasmodium falciparum var gene repertoire is limited, highly shared and possibly lacks several antigenic types.' *Gene* 453.1-2, pp. 37–44.

Albrecht, Letusa, Davide Angeletti, Kirsten Moll, Karin Blomqvist, Davide Valentini, Fabio Luiz D'Alexandri, Markus Maurer and Mats Wahlgren (2013). 'B-cell epitopes in NTS-DBL1α of PfEMP1 recognized by human antibodies in Rosetting Plasmodium falciparum.' *PLoS ONE* 9.12, e113248–e113248.

Allsopp, Catherine E M, Latifu A Sanni, Lieke Reubsaet, Francis Ndungu, Christopher I Newbold, Tabitha Mwangi, Kevin Marsh and Jean Langhorne (2002). 'CD4 T cell responses to a variant antigen of the malaria parasite Plasmodium falciparum, erythrocyte membrane protein-1, in individuals living in malaria-endemic areas.' *The Journal of infectious diseases* 185.6, pp. 812–819.

Altschul, S F, W Gish, W Miller, E W Myers and D J Lipman (1990). 'Basic local alignment search tool.' *Journal of Molecular Biology* 215.3, pp. 403–410.

Amaratunga, Chanaki, Tatiana M Lopera-Mesa, Nathaniel J Brittain, Rushina Cholera, Takayuki Arie, Hisashi Fujioka, Jeffrey R Keefer and Rick M Fairhurst (2011). 'A role for fetal hemoglobin and maternal immune IgG in infant resistance to Plasmodium falciparum malaria.' *PLoS ONE* 6.4, e14798.

Andisi, Cheryl (2014). 'Profiling PfEMP1 Variants Associated With Severe Malaria And Low Host Immunity Over Time'. PhD thesis. Milton-Keynes.

Anisimova, Maria, Rasmus Nielsen and Ziheng Yang (2003). 'Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.' *Genetics* 164.3, pp. 1229–1236.

Avril, Marion, Abhai K Tripathi, Andrew J Brazier, Cheryl Andisi, Joel H Janes, Vijaya L Soma, David J Sullivan, Peter C Bull, Monique F Stins and Joseph D Smith (2012). 'A restricted subset of var genes mediates adherence of Plasmodium falciparum-infected erythrocytes to brain endothelial cells.' *Proceedings of the National Academy of Sciences of the United States of America* 109.26, E1782–90.

Baird, J K JK, S S Masbar, H H Basri, S S Tirtokusumo, B B Subianto and S L SL Hoffman (1998). 'Age-dependent susceptibility to severe disease with primary exposure to Plasmodium falciparum.' *The Journal of infectious diseases* 178.2, pp. 592–595.

Barnwell, J W, A S Asch, R L Nachman, M Yamaya, M Aikawa and P Ingravallo (1989). 'A human 88-kD membrane glycoprotein (CD36) functions in vitro as a receptor for a cytoadherence ligand on Plasmodium falciparum-infected erythrocytes.' *Journal of Clinical Investigation* 84.3, pp. 765–772.

Barry, Alyssa E, Aleksandra Leliwa-Sytek, Livingston Tavul, Heather Imrie, Florence Migot-Nabias, Stuart M Brown, Gilean A V McVean and Karen P Day (2007). 'Population genomics of the immune evasion (var) genes of Plasmodium falciparum.' *PLoS Pathogens* 3.3, e34.

Baruch, D I, X C Ma, H B Singh, X Bi, B L Pasloske and R J Howard (1997). 'Identification of a region of PfEMP1 that mediates adherence of Plasmodium falciparum infected erythrocytes to CD36: conserved function with variant sequence.' *Blood* 90.9, pp. 3766–3775.

Bengtsson, Anja, Louise Joergensen, Thomas S Rask, Rebecca W Olsen, Marianne A Andersen, Louise Turner, Thor G Theander, Lars Hviid, Matthew K Higgins, Alister Craig, Alan Brown and Anja T R Jensen (2013). 'A novel domain cassette identifies Plasmodium falciparum PfEMP1 proteins binding ICAM-1 and is a target of cross-reactive, adhesion-inhibitory antibodies.' *The Journal of Immunology* 190.1, pp. 240–249.

Berendt, A R, D L Simmons, J Tansey, Christopher I Newbold and Kevin Marsh (1989). 'Intercellular adhesion molecule-1 is an endothelial cell adhesion receptor for Plasmodium falciparum.' *Nature* 341.6237, pp. 57–59.

Berkley, James A, Philip Bejon, Tabitha Mwangi, Samson Gwer, Kathryn Maitland, Thomas N Williams, Shebe Mohammed, Faith Osier, Samson Kinyanjui, Greg Fegan, Brett S Lowe, Mike English, Norbert Peshu, Kevin Marsh and Charles R J C Newton (2009). 'HIV infection, malnutrition, and invasive bacterial infection among children with severe malaria.' *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 49.3, pp. 336–343.

Bertin, Gwladys I, Thomas Lavstsen, François Guillonneau, Justin Doritchamou, Christian W Wang, Jakob S Jespersen, Sem Ezimegnon, Nadine Fievet, Maroufou J Alao, Francis Lalya, Achille Massougbodji, Nicaise Tuikue Ndam, Thor G Theander and Philippe Deloron (2013). 'Expression of the domain cassette 8 Plasmodium falciparum erythrocyte membrane protein 1 is associated with cerebral malaria in Benin.' *PLoS ONE* 8.7, e68368.

Blomqvist, Karin, Johan Normark, Daniel Nilsson, Ulf Ribacke, Judy Orikiriza, Petter Trillkott, Justus Byarugaba, Thomas G Egwang, Fred Kironde, Björn Andersson and Mats Wahlgren (2010). 'var gene transcription dynamics in Plasmodium falciparum patient isolates'. *Molecular and biochemical parasitology* 170.2, pp. 74–83.

Blomqvist, Karin, Letusa Albrecht, Maria del Pilar Quintana, Davide Angeletti, Nicolas Joannin, Arnaud Chêne, Kirsten Moll and Mats Wahlgren (2013). 'A sequence in subdomain 2 of DBL1α of Plasmodium falciparum erythrocyte membrane protein 1 induces strain transcending antibodies.' *PLoS ONE* 8.1, e52679.

Blythe, Martin J and Darren R Flower (2005). 'Benchmarking B cell epitope prediction: underperformance of existing methods.' *Protein science : a publication of the Protein Society* 14.1, pp. 246–248.

Bonnal, R J P, J Aerts, George Githinji, N Goto, D MacLean, C A Miller, H Mishima, M Pagani, R Ramirez-Gonzalez, G Smant, F Strozzi, R Syme, R Vos, T J Wennblom, B J Woodcroft, T Katayama and P Prins (2012). 'Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics'. 28.7, pp. 1035–1037.

Bopp, Selina E R, Micah J Manary, A Taylor Bright, Geoffrey L Johnston, Neekesh V Dharia, Fabio L Luna, Susan McCormack, David Plouffe, Case W McNamara, John R Walker, David A Fidock, Eros Lazzerini Denchi and Elizabeth A Winzeler (2013). 'Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families.' *PLoS Genetics* 9.2, e1003293.

Brown, K N and I N Brown (1965). 'Immunity to malaria: antigenic variation in chronic infections of Plasmodium knowlesi.' *Nature* 208.5017, pp. 1286–1288.

Buckee, Caroline O and Mario Recker (2012). 'Evolution of the multi-domain structures of virulence genes in the human malaria parasite, Plasmodium falciparum.' *PLoS Computational Biology* 8.4, e1002451.

Bueno, Lilian Lacerda, Francisco Pereira Lobo, Cristiane Guimarães Morais, Luíza Carvalho Mourão, Ricardo Andrez Machado de Ávila, Irene Silva Soares, Cor Jesus Fontes, Marcus Vinícius Lacerda, Carlos Chavez Olórtegui, Daniella Castanheira Bartholomeu, Ricardo Toshio Fujiwara and Erika Martins Braga (2010). 'Identification of a highly antigenic linear B cell epitope within Plasmodium vivax apical membrane antigen 1 (AMA-1).' *PLoS ONE* 6.6, e21289–e21289.

Bui, Huynh-Hoa, John Sidney, Wei Li, Nicolas Fusseder and Alessandro Sette (2007). 'Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines.' *BMC Bioinformatics* 8.1, p. 361.

Bull, P C, B S Lowe, M Kortok, C S Molyneux, Christopher I Newbold and Kevin Marsh (1998). 'Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria.' *Nature Medicine* 4.3, pp. 358–360.

Bull, P C, B S Lowe, M Kortok and Kevin Marsh (1999). 'Antibody recognition of Plasmodium falciparum erythrocyte surface antigens in Kenya: evidence for rare and prevalent variants.' *Infection and Immunity* 67.2, pp. 733–739.

Bull, P C, M Kortok, O Kai, F Ndungu, A Ross, B S Lowe, Christopher I Newbold and Kevin Marsh (2000). 'Plasmodium falciparum-infected erythrocytes: agglutination by diverse Kenyan plasma is associated with severe disease and young host age.' *The Journal of infectious diseases* 182.1, pp. 252–259.

Bull, Peter C and Kevin Marsh (2002). 'The role of antibodies to Plasmodium falciparum-infected-erythrocyte surface antigens in naturally acquired immunity to malaria.' *Trends in microbiology* 10.2, pp. 55–58.

Bull, Peter C, Matthew Berriman, Sue A Kyes, Michael A Quail, Neil Hall, Moses M Kortok, Kevin Marsh and Christopher I Newbold (2005). 'Plasmodium falciparum variant surface antigen expression patterns during malaria.' *PLoS Pathogens* 1.3, e26.

Bull, Peter C, Sue A Kyes, Caroline O Buckee, Jacqui Montgomery, Moses M Kortok, Christopher I Newbold and Kevin Marsh (2007). 'An approach to classifying sequence tags sampled from Plasmodium falciparum var genes.' *Molecular and biochemical parasitology* 154.1, pp. 98–102.

Bull, Peter C, Caroline O Buckee, Sue A Kyes, Moses M Kortok, Vandana Thathy, Bernard Guyah, José A Stoute, Christopher I Newbold and Kevin Marsh (2008). 'Plasmodium falciparum antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks.' *Molecular microbiology* 68.6, pp. 1519–1534.

Chan, Jo-Anne, Katherine B Howell, Linda Reiling, Ricardo Ataide, Claire L Mackintosh, Freya J I Fowkes, Michaela Petter, Joanne M Chesson, Christine Langer, George M Warimwe, Michael F Duffy, Stephen J Rogerson, Peter C Bull, Alan F Cowman, Kevin Marsh and James G Beeson (2012). 'Targets of antibodies against Plasmodium falciparum-infected erythrocytes in malaria immunity.' *Journal of Clinical Investigation* 122.9, pp. 3227–3238.

Claessens, Antoine, Yvonne Adams, Ashfaq Ghumra, Gabriella Lindergard, Caitlin C Buchan, Cheryl Andisi, Peter C Bull, Sachel Mok, Archna P Gupta, Christian W Wang, Louise Turner, Monica Arman, Ahmed Raza, Zbynek Bozdech and J Alexandra Rowe (2012). 'A subset of group A-like var genes encodes the malaria parasite ligands for binding to human brain endothelial cells.' *Proceedings of the National Academy of Sciences of the United States of America* 109.26, E1772–81.

Claessens, Antoine, William L Hamilton, Mihir Kekre, Thomas D Otto, Adnan Faizullabhoy, Julian C Rayner and Dominic Kwiatkowski (2014). 'Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis.' *PLoS Genetics* 10.12, e1004812–e1004812.

Coggeshall, L T and H W Kumm (1937). 'Demostration of Passive Immunity in Experimental Monkey Malaria'. *The Journal of experimental medicine* 66.2, pp. 177–190.

Cohen, S, I A McGregor and S CARRINGTON (1961). 'Gamma-globulin and acquired immunity to human malaria.' *Nature* 192, pp. 733–737.

Collins, William E and Geoffrey M Jeffery (2005). 'Plasmodium ovale: parasite and disease.' *Clinical microbiology reviews* 18.3, pp. 570–581.

Coulondre, C, J H Miller, P J Farabaugh and W Gilbert (1978). 'Molecular basis of base substitution hotspots in Escherichia coli.' *Nature* 274.5673, pp. 775–780.

Cox-Singh, J, T M E Davis, K S Lee, S S G Shamsul, A Matusop, S Ratnam, H A Rahman, D J Conway and B Singh (2008). 'Plasmodium knowlesi Malaria in Humans Is Widely Distributed and Potentially Life Threatening'. *Clinical Infectious Diseases* 46.2, pp. 165–171.

Craig, A, D Fernandez-Reyes, M Mesri, A McDowall, D C Altieri, N Hogg and C Newbold (2000). 'A functional analysis of a natural variant of intercellular adhesion molecule-1 (ICAM-1Kilifi).' *Human molecular genetics* 9.4, pp. 525–530.

Deitsch, K W, A del Pinal and T E Wellems (1999). 'Intra-cluster recombination and var transcription switches in the antigenic variation of Plasmodium falciparum.' *Molecular and biochemical parasitology* 101.1-2, pp. 107–116.

Ditlev, Sisse B, Morten A Nielsen, Mafalda Resende, Mette Ø Agerbæk, Vera V Pinto, Pernille H Andersen, Pamela Magistrado, John Lusingu, Madeleine Dahlbäck, Thor G Theander and Ali Salanti (2011). 'Identification and characterization of B-cell epitopes in the DBL4ε domain of VAR2CSA.' *PLoS ONE* 7.9, e43663–e43663.

Douglas, Nicholas M, François Nosten, Elizabeth A Ashley, Lucy Phaiphun, Michèle van Vugt, Pratap Singhasivanon, Nicholas J White and Ric N Price (2011). 'Plasmodium vivax recurrence following falciparum and mixed species malaria: risk factors and effect of antimalarial kinetics.' *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 52.5, pp. 612–620.

Doumbo, Ogobara K, Mahamadou A Thera, Abdoulaye K Koné, Ahmed Raza, Louisa J Tempest, Kirsten E Lyke, Christopher V Plowe and J Alexandra Rowe (2009). 'High levels of Plasmodium falciparum rosetting in all clinical forms of severe malaria in African children.' *The American journal of tropical medicine and hygiene* 81.6, pp. 987–993.

Duffy, Michael F, Timothy J Byrne, Celine Carret, Alasdair Ivens and Graham V Brown (2009). 'Ectopic recombination of a malaria var gene during mitosis associated with an altered var switch rate.' *Journal of Molecular Biology* 389.3, pp. 453–469.

Duraisingh, Manoj T, Till S Voss, Allison J Marty, Michael F Duffy, Robert T Good, Jennifer K Thompson, Lucio H Freitas-Junior, Artur Scherf, Brendan S Crabb and Alan F Cowman (2005). 'Heterochromatin Silencing and Locus Repositioning Linked to Regulation of Virulence Genes in Plasmodium falciparum'. *Cell* 121.1, pp. 13–24.

Duval, Linda, Mathieu Fourment, Eric Nerrienet, Dominique Rousset, Serge A Sadeuh, Steven M Goodman, Nicole V Andriaholinirina, Milijaona Randrianarivelojosia, Richard E Paul, Vincent Robert, Francisco J Ayala and Frédéric Ariey (2010). 'African apes as reservoirs of Plasmodium falciparum and the origin and diversification of the Laverania subgenus.' *Proceedings of the National Academy of Sciences of the United States of America* 107.23, pp. 10561–10566.

Eaton, M D (1938). 'The Agglutination of Plasmodium knowlesi by Immune Serum.' *The Journal of experimental medicine* 67.6, pp. 857–870.

Edgar, R C (2004). 'MUSCLE: multiple sequence alignment with high accuracy and high throughput'. *Nucleic Acids Research* 32.5, pp. 1792–1797.

Edgar, Robert C (2010). 'Search and clustering orders of magnitude faster than BLAST.' 26.19, pp. 2460–2461.

Edozien, J C, A E Boyo and D C Morley (1960). 'The relationship of serum gamma-globulin concentration to malaria and sickling.' *Journal of Clinical Pathology* 13, pp. 118–123.

Ellis, Jonathan J and Boštjan Kobe (2011). 'Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge.' *PLoS ONE* 6.7, e21169.

Emini, E A, J V Hughes, D S Perlow and J Boger (1985). 'Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide.' *Journal of virology* 55.3, pp. 836–839.

Falk, Nicole, Mirjam Kaestli, Weihong Qi, Michael Ott, Kay Baea, Alfred Cortés and Hans-Peter Beck (2009). 'Analysis of Plasmodium falciparum var genes expressed in children from Papua New Guinea.' *The Journal of infectious diseases* 200.3, pp. 347–356.

Francino, M P and H Ochman (2001). 'Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences.' *Molecular biology and evolution* 18.6, pp. 1147–1150.

Freitas-Junior, L H, E Bottius, L A Pirrit, K W Deitsch, C Scheidig, F Guinet, U Nehrbass, T E Wellems and A Scherf (2000). 'Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum.' *Nature* 407.6807, pp. 1018–1022.

Freitas-Junior, Lucio H, Rosaura Hernandez-Rivas, Stuart A Ralph, Dvorak Montiel-Condado, Omar K Ruvalcaba-Salazar, Ana Paola Rojas-Meza, Liliana Mâncio-Silva, Ricardo J Leal-Silvestre, Alisson Marques Gontijo, Spencer Shorte and Artur Scherf (2005). 'Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites.' *Cell* 121.1, pp. 25–36.

Fried, M, F Nosten, A Brockman, B J Brabin and P E Duffy (1998). 'Maternal antibodies block malaria.' *Nature* 395.6705, pp. 851–852.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu and Weizhong Li (2012). 'CD-HIT: accelerated for clustering the next-generation sequencing data.' 28.23, pp. 3150–3152.

Gardner, Malcolm J, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue A Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Christopher I Newbold, Ronald W Davis, Claire M Fraser and Bart Barrell (2002). 'Genome sequence of the human malaria parasite Plasmodium falciparum.' *Nature* 419.6906, pp. 498–511.

Genton, Blaise, Valérie D'Acremont, Lawrence Rare, Kay Baea, John C Reeder, Michael P Alpers and Ivo Müller (2008). 'Plasmodium vivax and mixed infections are associated with severe malaria in children: a prospective cohort study from Papua New Guinea.' *PLoS Medicine* 5.6, e127.

Gitau, Evelyn N, James Tuju, Liz Stevenson, Eva Kimani, Henry Karanja, Kevin Marsh, Peter C Bull and Britta C Urban (2012a). 'T-Cell Responses to the DBLα-Tag, a Short Semi-Conserved Region of the Plasmodium falciparum Membrane Erythrocyte Protein 1'. *PLoS ONE* 7.1, e30095.

— (2012b). 'T-cell responses to the DBLα-tag, a short semi-conserved region of the Plasmodium falciparum membrane erythrocyte protein 1.' *PLoS ONE* 7.1, e30095.

Gitau, Evelyn N, James Tuju, Henry Karanja, Liz Stevenson, Pilar Requena, Eva Kimani, Ally Olotu, Domtila Kimani, Kevin Marsh, Peter Bull and Britta C Urban (2014). 'CD4+ T cell responses to the Plasmodium falciparum erythrocyte membrane protein 1 in children with mild malaria.' *The Journal of Immunology* 192.4, pp. 1753–1761.

Gonçalves, Bronner P, Chiung-Yu Huang, Robert Morrison, Sarah Holte, Edward Kabyemela, D Rebecca Prevots, Michal Fried and Patrick E Duffy (2014). 'Parasite burden and severity of malaria in Tanzanian children.' *The New England journal of medicine* 370.19, pp. 1799–1808.

Greenwood, Brian M, Kalifa Bojang, Christopher J M Whitty and Geoffrey A T Targett (2005). 'Malaria.' *Lancet* 365.9469, pp. 1487–1498.

Gupta, S, R W Snow, C A Donnelly, Kevin Marsh and C Newbold (1999). 'Immunity to non-cerebral severe malaria is acquired after one or two infections.' *Nature Medicine* 5.3, pp. 340–343.

Hay, Simon I, Emelda A Okiro, Peter W Gething, Anand P Patil, Andrew J Tatem, Carlos A Guerra and Robert W Snow (2010). 'Estimating the global clinical burden of Plasmodium falciparum malaria in 2007.' *PLoS Medicine* 7.6, e1000290.

Higgins, Matthew K (2008). 'The structure of a chondroitin sulfate-binding domain important in placental malaria.' *The Journal of biological chemistry* 283.32, pp. 21842–21846.

Hill, A V, C E Allsopp, D Kwiatkowski, N M Anstey, P Twumasi, P A Rowe, S Bennett, D Brewster, A J McMichael and Brian M Greenwood (1991). 'Common west African HLA antigens are associated with protection from severe malaria.' *Nature* 352.6336, pp. 595–600.

Howell, Dasein P-G, Emily A Levin, Amy L Springer, Susan M Kraemer, David J Phippard, William R Schief and Joseph D Smith (2008). 'Mapping a common interaction site used by Plasmodium falciparum Duffy binding-like domains to bind diverse host receptors.' *Molecular microbiology* 67.1, pp. 78–87.

Howitt, Cali A, Daniel Wilinski, Manuel Llinás, Thomas J Templeton, Ron Dzikowski and Kirk W Deitsch (2009). 'Clonally variant gene families in Plasmodium falciparumshare a common activation factor'. *Molecular microbiology* 73.6, pp. 1171–1185.

Huang, X and A Madan (1999). 'CAP3: A DNA sequence assembly program.' *Genome research* 9.9, pp. 868–877.

Hviid, Lars and Trine Staalsoe (2004). 'Malaria immunity in infants: a special case of a general phenomenon?' *Trends in parasitology* 20.2, pp. 66–72.

Jelinek, T, C Schulte, R Behrens, M P Grobusch, J P Coulaud, Z Bisoffi, A Matteelli, J Clerinx, M Corachán, S Puente, I Gjørup, G Harms, H Kollaritsch, A Kotlowski, A Björkmann, J P Delmont, J Knobloch, L N Nielsen, J Cuadros, C Hatz, J Beran, M L Schmid, M Schulze, R Lopez-Velez, K Fleischer, A Kapaun, P McWhinney, P Kern, J Atougia, G Fry, S da Cunha and G Boecken (2002). 'Imported Falciparum malaria in Europe: sentinel surveillance data from the European network on surveillance of imported infectious diseases.' *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 34.5, pp. 572–576.

Jenkins, Neil E, Tabitha W Mwangi, Moses Kortok, Kevin Marsh, Alister G Craig and Thomas N Williams (2005). 'A polymorphism of intercellular adhesion molecule-1 is associated with a reduced incidence of nonmalarial febrile illness in Kenyan children.' *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 41.12, pp. 1817–1819.

Jensen, Anja T R, Pamela Magistrado, Sarah Sharp, Louise Joergensen, Thomas Lavstsen, Antonella Chiucchiuini, Ali Salanti, Lasse S Vestergaard, John P Lusingu, Rob Hermsen, Robert Sauerwein, Jesper Christensen, Morten A Nielsen, Lars Hviid, Colin Sutherland, Trine Staalsoe and Thor G Theander (2004). 'Plasmodium falciparum associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group A var genes.' *The Journal of experimental medicine* 199.9, pp. 1179–1190.

Kaestli, Mirjam, Ian A Cockburn, Alfred Cortés, Kay Baea, J Alexandra Rowe and Hans-Peter Beck (2006). 'Virulence of malaria is associated with differential expression of Plasmodium falciparum var gene subgroups in a case-control study.' *The Journal of infectious diseases* 193.11, pp. 1567–1574.

Kalmbach, Yvonne, Matthias Rottmann, Maryvonne Kombila, Peter G Kremsner, Hans-Peter Beck and Jürgen F J Kun (2010). 'Differential varGene Expression in Children with Malaria and Antidromic Effects on Host Gene Expression'. *Journal of Infectious Diseases* 202.2, pp. 313–317.

Karosiene, Edita, Michael Rasmussen, Thomas Blicher, Ole Lund, Søren Buus and Morten Nielsen (2013). 'NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ.' *Immunogenetics* 65.10, pp. 711–724.

Kerr, P J, L C Ranford-Cartwright and D Walliker (1994). 'Proof of intragenic recombination in Plasmodium falciparum.' *Molecular and biochemical parasitology* 66.2, pp. 241–248.

Kinyanjui, Samson M, Tabitha Mwangi, Peter C Bull, Christopher I Newbold and Kevin Marsh (2004). 'Protection against clinical malaria by heterologous immunoglobulin G antibodies against malaria-infected erythrocyte variant surface antigens requires interaction with asymptomatic infections.' *The Journal of infectious diseases* 190.9, pp. 1527–1533.

Kirchgatter, Karin and Hernando del A Portillo (2002). 'Association of severe noncerebral Plasmodium falciparum malaria in Brazil with expressed PfEMP1 DBL1 alpha sequences lacking cysteine residues.' *Molecular medicine (Cambridge, Mass.)* 8.1, pp. 16–23.

Kirkman, Laura A, Elizabeth A Lawrence and Kirk W Deitsch (2014). 'Malaria parasites utilize both homologous recombination and alternative end joining pathways to maintain genome integrity.' *Nucleic Acids Research* 42.1, pp. 370–379.

Kitua, A Y, H Urassa, M Wechsler, T Smith, P Vounatsou, N A Weiss, P L Alonso and M Tanner (1999). 'Antibodies against Plasmodium falciparum vaccine candidates in infants in an area of intense and perennial transmission: relationships with clinical malaria and with entomological inoculation rates.' *Parasite immunology* 21.6, pp. 307–317.

Kobe, Boštjan and Mikael Bodén (2012). 'Computational modelling of linear motif-mediated protein interactions.' *Current topics in medicinal chemistry* 12.14, pp. 1553–1561.

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

Kraemer, Susan M, Sue A Kyes, Gautam Aggarwal, Amy L Springer, Siri O Nelson, Zóe Christodoulou, Leia M Smith, Wendy Wang, Emily Levin, Christopher I Newbold, Peter J Myler and Joseph D Smith (2007). 'Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates.' *BMC Genomics* 8, p. 45.

Krotoski, W A, D M Krotoski, P C Garnham, R S Bray, R Killick-Kendrick, C C Draper, G A Targett and M W Guy (1980). 'Relapses in primate malaria: discovery of two populations of exoerythrocytic stages. Preliminary note.' *British medical journal* 280.6208, pp. 153–154.

Kyes, Sue A, Zóe Christodoulou, Ahmed Raza, Paul Horrocks, Robert Pinches, J Alexandra Rowe and Christopher I Newbold (2003). 'A well-conserved Plasmodium falciparum var gene shows an unusual stage-specific transcript pattern.' *Molecular microbiology* 48.5, pp. 1339–1348.

Kyes, Sue A, Susan M Kraemer and Joseph D Smith (2007). 'Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family.' *Eukaryotic cell* 6.9, pp. 1511–1520.

Kyriacou, Helen M, Graham N Stone, Richard J Challis, Ahmed Raza, Kirsten E Lyke, Mahamadou A Thera, Abdoulaye K Koné, Ogobara K Doumbo, Christopher V Plowe and J Alexandra Rowe (2006). 'Differential var gene transcription in Plasmodium falciparum isolates from patients with cerebral malaria compared to hyperparasitaemia.' *Molecular and biochemical parasitology* 150.2, pp. 211–218.

Kyte, J and R F Doolittle (1982). 'A simple method for displaying the hydropathic character of a protein'. *Journal of Molecular Biology*.

Larremore, Daniel B, Aaron Clauset and Caroline O Buckee (2013). 'A network approach to analyzing highly recombinant malaria parasite genes.' *PLoS Computational Biology* 9.10, e1003268.

Larsen, Jens Erik Pontoppidan, Ole Lund and Morten Nielsen (2006). 'Improved method for predicting linear B-cell epitopes.' *Immunome research* 2, p. 2.

Lavstsen, Thomas, Ali Salanti, Anja T R Jensen, David E Arnot and Thor G Theander (2003). 'Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions.' *Malaria Journal* 2, p. 27.

Lavstsen, Thomas, Pamela Magistrado, Cornelus C Hermsen, Ali Salanti, Anja T R Jensen, Robert Sauerwein, Lars Hviid, Thor G Theander and Trine Staalsoe (2005).

'Expression of Plasmodium falciparum erythrocyte membrane protein 1 in experimentally infected humans.' *Malaria Journal* 4.1, p. 21.

Lavstsen, Thomas, Louise Turner, Fredy Saguti, Pamela Magistrado, Thomas S Rask, Jakob S Jespersen, Christian W Wang, Sanne S Berger, Vito Baraka, Andrea M Marquard, Andaine Seguin-Orlando, Eske Willerslev, M Thomas P Gilbert, John Lusingu and Thor G Theander (2012). 'Plasmodium falciparum erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children.' *Proceedings of the National Academy of Sciences of the United States of America* 109.26, E1791–800.

Leech, J H, J W Barnwell, Louis H Miller and R J Howard (1984). 'Identification of a strain-specific malarial antigen exposed on the surface of Plasmodium falciparum-infected erythrocytes.' *The Journal of experimental medicine* 159.6, pp. 1567–1575.

Levitt, M (1978). 'Conformational preferences of amino acids in globular proteins.' *Biochemistry* 17.20, pp. 4277–4285.

Li, Weizhong and Adam Godzik (2006). 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.' 22.13, pp. 1658–1659.

Liu, Weimin, Yingying Li, Gerald H Learn, Rebecca S Rudicell, Joel D Robertson, Brandon F Keele, Jean-Bosco N Ndjango, Crickette M Sanz, David B Morgan, Sabrina Locatelli, Mary K Gonder, Philip J Kranzusch, Peter D Walsh, Eric Delaporte, Eitel Mpoudi-Ngole, Alexander V Georgiev, Martin N Muller, George M Shaw, Martine Peeters, Paul M Sharp, Julian C Rayner and Beatrice H Hahn (2010). 'Origin of the human malaria parasite Plasmodium falciparum in gorillas.' *Nature* 467.7314, pp. 420–425.

Lopez-Rubio, Jose Juan, Alisson M Gontijo, Marta C Nunes, Neha Issar, Rosaura Hernandez-Rivas and Artur Scherf (2007). '5′ flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites'. *Molecular microbiology* 66.6, pp. 1296–1305.

Löytynoja, Ari and Nick Goldman (2005). 'An algorithm for progressive multiple alignment of sequences with insertions.' *Proceedings of the National Academy of Sciences of the United States of America* 102.30, pp. 10557–10562.

Markus, Miles B (2011a). 'Malaria: origin of the term "hypnozoite".' *Journal of the history of biology* 44.4, pp. 781–786.

— (2011b). 'The hypnozoite concept, with particular reference to malaria.' *Parasitology research* 108.1, pp. 247–252.

— (2015). 'Do hypnozoites cause relapse in malaria?' *Trends in parasitology*.

Marsh, Kevin and R J Howard (1986). 'Antigens induced on erythrocytes by P. falciparum: expression of diverse and conserved determinants.' *Science* 231.4734, pp. 150–153.

Marsh, Kevin and S Kinyanjui (2006). 'Immune effector mechanisms in malaria.' *Parasite immunology* 28.1-2, pp. 51–60.

Marsh, Kevin, D Forster, C Waruiru, I Mwangi, M Winstanley, V Marsh, C Newton, P Winstanley, P Warn and N Peshu (1995). 'Indicators of life-threatening malaria in African children.' *The New England journal of medicine* 332.21, pp. 1399–1404.

Mason, S J, Louis H Miller, T Shiroishi, J A Dvorak and M H McGinniss (1977). 'The Duffy blood group determinants: their role in the susceptibility of human and animal erythrocytes to Plasmodium knowlesi malaria.' *British journal of haematology* 36.3, pp. 327–335.

McCormick, C J, A Craig, D Roberts, Christopher I Newbold and A R Berendt (1997). 'Intercellular adhesion molecule-1 and CD36 synergize to mediate adherence of Plasmodium falciparum-infected erythrocytes to cultured human microvascular endothelial cells.' *Journal of Clinical Investigation* 100.10, pp. 2521–2529.

McGregor, I A, H M Gilles, J H Walters, A H Davies and F A Pearson (1956). 'Effects of heavy and repeated malarial infections on Gambian infants and children; effects of erythrocytic parasitization.' *British medical journal* 2.4994, pp. 686–692.

McSparron, Helen, Martin J Blythe, Christianna Zygouri, Irini A Doytchinova and Darren R Flower (2003). 'JenPep: a novel computational information resource for immunobiology and vaccinology.' *Journal of chemical information and computer sciences* 43.4, pp. 1276–1287.

Midega, Janet T, Dave L Smith, Ally Olotu, Joseph M Mwangangi, Joseph G Nzovu, Juliana Wambua, George Nyangweso, Charles M Mbogo, George K Christophides, Kevin Marsh and Philip Bejon (2012). 'Wind direction and proximity to larval sites determines malaria risk in Kilifi District in Kenya.' *Nature communications* 3, p. 674.

Miller, Louis H (1969). 'Distribution of mature trophozoites and schizonts of Plasmodium falciparum in the organs of Aotus trivirgatus, the night monkey.' *The American journal of tropical medicine and hygiene* 18.6, pp. 860–865.

Miller, Louis H, S J Mason, D F Clyde and M H McGinniss (1976). 'The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy.' *The New England journal of medicine* 295.6, pp. 302–304.

Mouchet, J, S Laventure, S Blanchy, R Fioramonti, A Rakotonjanabelo, P Rabarison, J Sircoulon and J Roux (1997). 'The reconquest of the Madagascar highlands by malaria.' *Bulletin de la Société de pathologie exotique (1990)* 90.3, pp. 162–168.

Mueller, Ivo, Mary R Galinski, J Kevin Baird, Jane M Carlton, Dhanpat K Kochar, Pedro L Alonso and Hernando A del Portillo (2009). 'Key gaps in the knowledge of Plasmodium vivax, a neglected human malaria parasite.' *The Lancet infectious diseases* 9.9, pp. 555–566.

Mwangangi, Joseph M, Charles M Mbogo, Benedict O Orindi, Ephantus J Muturi, Janet T Midega, Joseph Nzovu, Hellen Gatakaa, John Githure, Christian Borgemeister, Joseph Keating and John C Beier (2013). 'Shifts in malaria vector species composition and transmission dynamics along the Kenyan coast over the past 20 years.' *Malaria Journal* 12.1, p. 13.

Newbold, C, P Warn, G Black, A Berendt, A Craig, B Snow, M Msobo, N Peshu and Kevin Marsh (1997). 'Receptor-specific adhesion and clinical disease in Plasmodium falciparum.' *The American journal of tropical medicine and hygiene* 57.4, pp. 389–398.

Nielsen, Morten, Sune Justesen, Ole Lund, Claus Lundegaard and Søren Buus (2010). 'NetMHCIIpan-2.0 - Improved pan-specific HLA-DRpredictions using a novel concurrent alignmentand weight optimization training procedure'. *Immunome research* 6.1, p. 9.

Nielsen, Morten A, Trine Staalsoe, Jørgen A L Kurtzhals, Bamenla Q Goka, Daniel Dodoo, Michael Alifrangis, Thor G Theander, Bartholomew D Akanmori and Lars Hviid (2002). 'Plasmodium falciparum variant surface antigen expression varies between isolates causing severe and nonsevere malaria and is modified by acquired immunity.' *Journal of immunology (Baltimore, Md. : 1950)* 168.7, pp. 3444–3450.

Noor, Abdisalan M, Damaris K Kinyoki, Clara W Mundia, Caroline W Kabaria, Jonesmus W Mutua, Victor A Alegana, Ibrahima Socé Fall and Robert W Snow (2014). 'The changing risk of Plasmodium falciparum malaria infection in Africa: 2000-10: a spatial and temporal analysis of transmission intensity.' *Lancet* 383.9930, pp. 1739–1747.

Normark, Johan, Daniel Nilsson, Ulf Ribacke, Gerhard Winter, Kirsten Moll, Craig E Wheelock, Justus Bayarugaba, Fred Kironde, Thomas G Egwang, Qijun Chen, Björn Andersson and Mats Wahlgren (2007). 'PfEMP1-DBL1alpha amino acid motifs in severe disease states of Plasmodium falciparum malaria.' *Proceedings of the National Academy of Sciences of the United States of America* 104.40, pp. 15835–15840.

O'Brien, Cathal, Darren R Flower and Conleth Feighery (2008). 'Peptide length significantly influences in vitro affinity for MHC class II molecules.' *Immunome research* 4, p. 6.

Ochola, Lucy B, Bethsheba R Siddondo, Harold Ocholla, Siana Nkya, Eva N Kimani, Thomas N Williams, Johnstone O Makale, Anne Liljander, Britta C Urban, Peter C Bull, Tadge Szestak, Kevin Marsh and Alister G Craig (2011). 'Specific receptor usage in Plasmodium falciparum cytoadherence is associated with disease outcome.' *PLoS ONE* 6.3, e14741.

Ofori, Michael F, Daniel Dodoo, Trine Staalsoe, Jørgen A L Kurtzhals, Kwadwo Koram, Thor G Theander, Bartholomew D Akanmori and Lars Hviid (2002). 'Malaria-induced acquisition of antibodies to Plasmodium falciparum variant surface antigens.' *Infection and Immunity* 70.6, pp. 2982–2988.

Okiro, Emelda A, Abdullah Al-Taiar, Hugh Reyburn, Richard Idro, James A Berkley and Robert W Snow (2009a). ‘Age patterns of severe paediatric malaria and their relationship to Plasmodium falciparum transmission intensity’. *Malaria Journal* 8.1, p. 4.

Okiro, Emelda A, Victor A Alegana, Abdisalan M Noor, Juliette J Mutheu, Elizabeth Juma and Robert W Snow (2009b). ‘Malaria paediatric hospitalization between 1999 and 2008 across Kenya.’ *BMC medicine* 7.1, p. 75.

Oleinikov, Andrew V, Emily Amos, Isaac Tyler Frye, Eddie Rossnagle, Theonest K Mutabingwa, Michal Fried and Patrick E Duffy (2009). ‘High throughput functional assays of the variant antigen PfEMP1 reveal a single domain in the 3D7 Plasmodium falciparum genome that binds ICAM1 with high affinity and is targeted by naturally acquired neutralizing antibodies.’ *PLoS Pathogens* 5.4, e1000386–e1000386.

Ollomo, Benjamin, Patrick Durand, Franck Prugnolle, Emmanuel Douzery, Céline Arnathau, Dieudonné Nkoghe, Eric Leroy and François Renaud (2009). ‘A new malaria agent in African hominids.’ *PLoS Pathogens* 5.5, e1000446.

O’Meara, Wendy P, Phillip Bejon, Tabitha W Mwangi, Emelda A Okiro, Norbert Peshu, Robert W Snow, Charles RJC Newton and Kevin Marsh (2008). ‘Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya’. *The Lancet* 372.9649, pp. 1555–1562.

Otto, Thomas D, Julian C Rayner, Ulrike Böhme, Arnab Pain, Natasha Spottiswoode, Mandy Sanders, Michael Quail, Benjamin Ollomo, François Renaud, Alan W Thomas, Franck Prugnolle, David J Conway, Christopher I Newbold and Matthew Berriman (2014). ‘Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts.’ *Nature communications* 5, p. 4754.

Oyarzún, Patricio, Jonathan J Ellis, Mikael Bodén and Boštjan Kobe (2013). ‘PREDIVAC: CD4+ T-cell epitope prediction for vaccine design that covers 95% of HLA class II DR protein diversity.’ *BMC Bioinformatics* 14, p. 52.

Pain, A, D J Ferguson, O Kai, B C Urban, B Lowe, Kevin Marsh and D J Roberts (2001). ‘Platelet-mediated clumping of Plasmodium falciparum-infected erythrocytes is a common adhesive phenotype and is associated with severe malaria.’ *Proceedings of the National Academy of Sciences of the United States of America* 98.4, pp. 1805–1810.

Parker, J M, D Guo and R S Hodges (1986). ‘New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites.’ *Biochemistry* 25.19, pp. 5425–5432.

Patarroyo, Manuel E, Martha Patricia Alba, Hernando Curtidor, Magnolia Vanegas, Hannia Almonacid and Manuel A Patarroyo (2014). ‘Using the PfEMP1 head structure binding motif to deal a blow at severe malaria.' *PLoS ONE* 9.2, e88420–e88420.

Pellequer, J L, E Westhof and M H Van Regenmortel (1993). 'Correlation between the location of antigenic sites and the prediction of turns in proteins.' *Immunology letters* 36.1, pp. 83–99.

Petersen-Mahrt, Svend K, Reuben S Harris and Michael S Neuberger (2002). 'AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification.' *Nature* 418.6893, pp. 99–103.

Piper, K P, D J Roberts and K P Day (1999). 'Plasmodium falciparum: analysis of the antibody specificity to the surface of the trophozoite-infected erythrocyte.' *Experimental parasitology* 91.2, pp. 161–169.

Ponnudurai, T, A H Lensen, G J van Gemert, M G Bolmer and J H Meuwissen (1991). 'Feeding behaviour and sporozoite ejection by infected Anopheles stephensi.' *Transactions of the Royal Society of Tropical Medicine and Hygiene* 85.2, pp. 175–180.

Prugnolle, Franck, Patrick Durand, Cécile Neel, Benjamin Ollomo, Francisco J Ayala, Céline Arnathau, Lucie Etienne, Eitel Mpoudi-Ngole, Dieudonné Nkoghe, Eric Leroy, Eric Delaporte, Martine Peeters and François Renaud (2010). 'African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum.' *Proceedings of the National Academy of Sciences of the United States of America* 107.4, pp. 1458–1463.

Ralph, Stuart A, Christine Scheidig-Benatar and Artur Scherf (2005). 'Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations.' *Proceedings of the National Academy of Sciences of the United States of America* 102.15, pp. 5414–5419.

Rask, Thomas S, Daniel A Hansen, Thor G Theander, Anders Gorm Pedersen and Thomas Lavstsen (2010). 'Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes–divide and conquer.' *PLoS Computational Biology* 6.9.

Reeder, J C, Alan F Cowman, K M Davern, J G Beeson, J K Thompson, S J Rogerson and G V Brown (1999). 'The adhesion of Plasmodium falciparum-infected erythrocytes to chondroitin sulfate A is mediated by P. falciparum erythrocyte membrane protein 1.' *Proceedings of the National Academy of Sciences of the United States of America* 96.9, pp. 5198–5202.

Riley, E M, O Olerup, S Bennett, P Rowe, S J Allen, M J Blackman, M Troye-Blomberg, A A Holder and Brian M Greenwood (1992). 'MHC and malaria: the relationship between HLA class II alleles and immune responses to Plasmodium falciparum.' *International immunology* 4.9, pp. 1055–1063.

Robinson, Bridget A, Teresa L Welch and Joseph D Smith (2003). 'Widespread functional specialization of Plasmodium falciparum erythrocyte membrane protein 1 family members to bind CD36 analysed across a parasite genome.' *Molecular microbiology* 47.5, pp. 1265–1278.

Rogerson, S J, S C Chaiyaroj, K Ng, J C Reeder and G V Brown (1995). 'Chondroitin sulfate A is a cell surface receptor for Plasmodium falciparum-infected erythrocytes.' *The Journal of experimental medicine* 182.1, pp. 15–20.

Rogerson, S J, R Tembenu, C Dobaño, S Plitt, T E Taylor and M E Molyneux (1999). 'Cytoadherence characteristics of Plasmodium falciparum-infected erythrocytes from Malawian children with severe and uncomplicated malaria.' *The American journal of tropical medicine and hygiene* 61.3, pp. 467–472.

Rosenberg, R, R A Wirtz, I Schneider and R Burge (1990). 'An estimation of the number of malaria sporozoites ejected by a feeding mosquito.' *Transactions of the Royal Society of Tropical Medicine and Hygiene* 84.2, pp. 209–212.

Rottmann, Matthias, Thomas Lavstsen, Joseph Paschal Mugasa, Mirjam Kaestli, Anja T R Jensen, Dania Müller, Thor Theander and Hans-Peter Beck (2006). 'Differential expression of var gene groups is associated with morbidity caused by Plasmodium falciparum infection in Tanzanian children.' *Infection and Immunity* 74.7, pp. 3904–3911.

Rowe, J A, J M Moulds, Christopher I Newbold and Louis H Miller (1997). 'P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1.' *Nature* 388.6639, pp. 292–295.

Ryan, Jeffrey R, José A Stoute, Joseph Amon, Raymond F Dunton, Ramadhan Mtalib, Joseph Koros, Boaz Owour, Shirley Luckhart, Robert A Wirtz, John W Barnwell and Ronald Rosenberg (2006). 'Evidence for transmission of Plasmodium vivax among a duffy antigen negative population in Western Kenya.' *The American journal of tropical medicine and hygiene* 75.4, pp. 575–581.

Sabchareon, A, T Burnouf, D Ouattara, P Attanath, H Bouharoun-Tayoun, P Chantavanich, C Foucault, T Chongsuphajaisiddhi and P Druilhe (1991). 'Parasitologic and clinical human response to immunoglobulin administration in falciparum malaria.' *The American journal of tropical medicine and hygiene* 45.3, pp. 297–308.

Sachs, Jeffrey and Pia Malaney (2002). 'The economic and social burden of malaria.' *Nature* 415.6872, pp. 680–685.

Salanti, Ali, Madeleine Dahlbäck, Louise Turner, Morten A Nielsen, Lea Barfod, Pamela Magistrado, Anja T R Jensen, Thomas Lavstsen, Michael F Ofori, Kevin Marsh, Lars Hviid and Thor G Theander (2004). 'Evidence for the involvement of VAR2CSA in pregnancy-associated malaria.' *The Journal of experimental medicine* 200.9, pp. 1197–1203.

Sander, Adam F, Thomas Lavstsen, Thomas S Rask, Michael Lisby, Ali Salanti, Sarah L Fordyce, Jakob S Jespersen, Richard Carter, Kirk W Deitsch, Thor G Theander, Anders Gorm Pedersen and David E Arnot (2013). 'DNA secondary structures are associated with recombination in major Plasmodium falciparum variable surface antigen gene families.' *Nucleic Acids Research* 42.4, pp. 2270–2281.

Sanni, Latifu A, Catherine E M Allsopp, Lieke Reubsaet, Ambaliou Sanni, Christopher I Newbold, Virander S Chauhan and Jean Langhorne (2002). 'Cellular responses to Plasmodium falciparum erythrocyte membrane protein-1: use of relatively conserved synthetic peptide pools to determine CD4 T cell responses in malaria-exposed individuals in Benin, West Africa.' *Malaria Journal* 1, p. 7.

Saunders, Neil F W and Boštjan Kobe (2008). 'The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information.' *Nucleic Acids Research* 36.Web Server issue, W286–90.

Saunders, Neil F W, Ross I Brinkworth, Thomas Huber, Bruce E Kemp and Boštjan Kobe (2008). 'Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites.' *BMC Bioinformatics* 9.1, p. 245.

Scherf, A, R Hernandez Rivas and P Buffet (1998). 'Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium ...' *The EMBO ...*

Schierup, M H and J Hein (2000). 'Consequences of recombination on traditional phylogenetic analysis.' *Genetics* 156.2, pp. 879–891.

Scott, J Anthony G, Evasius Bauni, Jennifer C Moisi, John Ojal, Hellen Gatakaa, Christopher Nyundo, Catherine S Molyneux, Francis Kombe, Benjamin Tsofa, Kevin Marsh, Norbert Peshu and Thomas N Williams (2012). 'Profile: The Kilifi Health and Demographic Surveillance System (KHDSS).' *International journal of epidemiology* 41.3, pp. 650–657.

Shannon, C E (1948). 'A mathematical theory of communication'. *Bell System Technical Journal, The* 27.3, pp. 379–423.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski and Trey Ideker (2003). 'Cytoscape: a software environment for integrated models of biomolecular interaction networks.' *Genome research* 13.11, pp. 2498–2504.

Smith, J D, S Kyes, A G Craig, T Fagan, D Hudson-Taylor, Louis H Miller, D I Baruch and Christopher I Newbold (1998). 'Analysis of adhesive domains from the A4VAR Plasmodium falciparum erythrocyte membrane protein-1 identifies a CD36 binding domain.' *Molecular and biochemical parasitology* 97.1-2, pp. 133–148.

Smith, J D, G Subramanian, B Gamain, D I Baruch and Louis H Miller (2000a). 'Classification of adhesive domains in the Plasmodium falciparum erythrocyte membrane protein 1 family.' *Molecular and biochemical parasitology* 110.2, pp. 293–310.

Smith, J D, A G Craig, N Kriek, D Hudson-Taylor, S Kyes, T Fagan, T Fagen, R Pinches, D I Baruch, Christopher I Newbold and Louis H Miller (2000b). 'Identification of a Plasmodium falciparum intercellular adhesion molecule-1 binding domain: a parasite adhesion trait implicated in cerebral malaria.' *Proceedings of the National Academy of Sciences of the United States of America* 97.4, pp. 1766–1771.

Snow, R W, J A Omumbo, B Lowe, C S Molyneux, J O Obiero, A Palmer, M W Weber, M Pinder, B Nahlen, C Obonyo, C Newbold, S Gupta and Kevin Marsh (1997). 'Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa.' *The Lancet* 349.9066, pp. 1650–1654.

Snow, R W, B Nahlen, A Palmer, C A Donnelly, S Gupta and Kevin Marsh (1998). 'Risk of severe malaria among African infants: direct evidence of clinical protection during early infancy.' *The Journal of infectious diseases* 177.3, pp. 819–822.

Su, X Z, V M Heatwole, S P Wertheimer, F Guinet, J A Herrfeldt, D S Peterson, J A Ravetch and T E Wellems (1995). 'The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes.' *Cell* 82.1, pp. 89–100.

Taylor, H M, S A Kyes, D Harris, N Kriek and Christopher I Newbold (2000a). 'A study of var gene transcription in vitro using universal var gene primers.' *Molecular and biochemical parasitology* 105.1, pp. 13–23.

Taylor, H M, S A Kyes and Christopher I Newbold (2000b). 'Var gene diversity in Plasmodium falciparum is generated by frequent recombination events.' *Molecular and biochemical parasitology* 110.2, pp. 391–397.

Troye-Blomberg, M, O Olerup, A Larsson, K Sjöberg, H Perlmann, E Riley, J P Lepers and P Perlmann (1991). 'Failure to detect MHC class II associations of the human immune response induced by repeated malaria infections to the Plasmodium falciparum antigen Pf155/RESA.' *International immunology* 3.10, pp. 1043–1051.

Turner, Louise, Thomas Lavstsen, Sanne S Berger, Christian W Wang, Jens E V Petersen, Marion Avril, Andrew J Brazier, Jim Freeth, Jakob S Jespersen, Morten A Nielsen, Pamela Magistrado, John Lusingu, Joseph D Smith, Matthew K Higgins and Thor G Theander (2013). 'Severe malaria is associated with parasite binding to endothelial protein C receptor.' *Nature* 498.7455, pp. 502–505.

Udomsangpetch, R, B Wåhlin, J Carlson, K Berzins, M Torii, M Aikawa, P Perlmann and M Wahlgren (1989). 'Plasmodium falciparum-infected erythrocytes form spontaneous erythrocyte rosettes.' *The Journal of experimental medicine* 169.5, pp. 1835–1840.

Udomsangpetch, R, P H Reinhardt, T Schollaardt, J F Elliott, P Kubes and M Ho (1997). 'Promiscuity of clinical Plasmodium falciparum isolates for multiple adhesion molecules under flow conditions.' *Journal of immunology (Baltimore, Md. : 1950)* 158.9, pp. 4358–4364.

Vázquez-Macías, Aleida, Perla Martínez-Cruz, María Cristina Castañeda-Patlán, Christine Scheidig, Jürg Gysin, Artur Scherf and Rosaura Hernandez-Rivas (2002). 'A distinct 5' flanking var gene region regulates Plasmodium falciparum variant erythrocyte surface antigen expression in placental malaria.' *Molecular microbiology* 45.1, pp. 155–167.

Voss, T S, J K Thompson, J Waterkeyn, I Felger, N Weiss, Alan F Cowman and H P Beck (2000). 'Genomic distribution and functional characterisation of two distinct and conserved Plasmodium falciparum var gene 5' flanking sequences.' *Molecular and biochemical parasitology* 107.1, pp. 103–115.

Voss, Till S, Mirjam Kaestli, Denise Vogel, Selina Bopp and Hans-Peter Beck (2003). 'Identification of nuclear proteins that interact differentially with Plasmodium falciparum var gene promoters.' *Molecular microbiology* 48.6, pp. 1593–1607.

Ward, C P, G T Clottey, M Dorris, D D Ji and D E Arnot (1999). 'Analysis of Plasmodium falciparum PfEMP-1/var genes suggests that recombination rearranges constrained sequences.' *Molecular and biochemical parasitology* 102.1, pp. 167–177.

Warimwe, George M, Thomas M Keane, Gregory Fegan, Jennifer N Musyoki, Charles R J C Newton, Arnab Pain, Matthew Berriman, Kevin Marsh and Peter C Bull (2009). 'Plasmodium falciparum var gene expression is modified by host immunity.' *Proceedings of the National Academy of Sciences of the United States of America* 106.51, pp. 21801–21806.

Warimwe, George M, Gregory Fegan, Jennifer N Musyoki, Charles R J C Newton, Michael Opiyo, George Githinji, Cheryl Andisi, Francis Menza, Barnes Kitsao, Kevin Marsh and Peter C Bull (2012). 'Prognostic indicators of life-threatening malaria are associated with distinct parasite variant antigen profiles.' *Science Translational Medicine* 4.129, 129ra45.

Warimwe, George Mbugua (2010). 'Studies on PfEMP1 Expression in Clinical Isolates From Kenyan Children With Malaria'. PhD thesis. Milton-Keenes: Open University(UK).

WHO (2010). 'World Malaria Report:2010', pp. 1–220.

Wickham, Hadley (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

Winter, Gerhard, Qijun Chen, Kirsten Flick, Peter Kremsner, Victor Fernandez and Mats Wahlgren (2003). 'The 3D7var5.2 (var COMMON) type var gene family is commonly expressed in non-placental Plasmodium falciparum malaria.' *Molecular and biochemical parasitology* 127.2, pp. 179–191.

World Health Organization (2013). 'World Malaria Report:2013'. *World Health Organization*, pp. 1–286.

Wyler, D J, C N Oster and Quinn T C (1979). *The role of the spleen in malaria infections*. Basel, Switzerland.

Yipp, Bryan G, Michael J Hickey, Graciela Andonegui, Allan G Murray, Sornchai Looareesuwan, Paul Kubes and May Ho (2007). 'Differential roles of CD36, ICAM-1, and P-selectin in Plasmodium falciparum cytoadherence in vivo.' *Microcirculation (New York, N.Y. : 1994)* 14.6, pp. 593–602.

Zilversmit, Martine M, Ella K Chase, Donald S Chen, Philip Awadalla, Karen P Day and Gil McVean (2013). 'Hypervariable antigen genes in malaria have ancient roots.' *BMC Evolutionary Biology* 13, p. 110.

Appendix A
Source code

A.1 A script to search for DBLα sequence tags

This program depends on the DBLα classifier library

Listing A.1: A Ruby program to search for DBLα sequence tags

```
#!/usr/bin/env ruby
require 'bio'
require 'optparse'
require 'ostruct'
require 'bio-dbla-finder'

options = OpenStruct.new

OptionParser.new do |opts|
  opts.banner = 'Usage: dbla_finder -i infile -s stopcodons -o outfile'

  opts.on('-h','--help','Display this screen') do
    puts opts
    exit
  end

  opts.on('-i','--infile INPUT','DNA file to search for tags [required]↩
      ') do |infile|
    options.infile = infile
  end

  opts.on('-s','--stopcodon STOPCODON',Integer, 'Number of stop codons ↩
      to accept in the tag') do |stopcodon|
    options.stopcodon = stopcodon
  end

  opts.on('-o','--outfile OUTPUT', 'File to write the tags to') do |↩
      outfile|
    options.outfile = outfile
  end
end.parse!

infile    = options.infile #file_path =  ARGV[0]
stopcodon = options.stopcodon
outfile   = options.outfile
```

```
#create a finder class in this case named robot
robot = Bio::Finder.new(infile)

begin
  #call the find_tags method to list the tags to the stderr
  tags = robot.find_tags

  if outfile.nil?
    $stdout.puts tags
  else
    output = File.open(outfile,'w')
    output << robot.find_tags
    output.close
  end
  #$stderr.puts "#{tags.size} tags found!"

rescue #Bio::FlatFile::UnknownDataFormatError
  $stderr.puts "bad_contigs_format"
end
```

Listing A.2: nucleotide class

```
class Bio::Sequence::NA

  #translate the sequence to protein
  def to_protein(frame)
    seq.translate(frame)
  end

  #get the hamming distance of 2 strings of equal length
  def hamming_dist?(str1,str2,mismatches=1)
    str1.chars.zip(str2.chars).count{|ca,cb| ca!=cb} <= mismatches
  end

  #find strings that are similar based on the hamming distance
  def find_similar(haystack,needle,mismatches)
    haystack.chars.each_cons(needle.length).map(&:join).select{|s| ↩
        hamming_dist?(s,needle,mismatches)}
  end

  #given a tag return the potential motifs in the n-terminus
  def n_term_motifs(protein)
    find_similar(protein,'DIGDI',2)
  end

  #given a tag return the potential motifs in the c-terminus
  def c_term_motifs(protein)
    find_similar(protein,'PQYLR',2)
  end

  #positions for all n_terminus motifs
  def n_term_positions(motifs,protein)
    motifs.map{|motif| protein.match(/#{Regexp.quote(motif)}/).offset↩
        (0)[0]}
  end

  #positions for all c_terminus motifs
  def c_term_positions(motifs,protein)
    motifs.map{|motif| protein.match(/#{Regexp.quote(motif)}/).offset↩
        (0)[1]}
  end

  #pairwise length differences for each n_terminus and c_terminus pair
  def tag_len_diffs(c_term_pos,n_term_pos)
    c_term_pos.map{|i| n_term_pos.map{|j| i - j}}
  end

  #potential tags
  def candidate_tag_lens(tag_len_diffs,tag_range)
    tag_len_diffs.map{|list| list.each_with_index.map{|i,index| index ↩
        if tag_range.include? i}}
  end

  #get the tag region
  def clip_tag(start,stop,step=0,num_sc=0)
    tag = seq.subseq(start + step,stop + step)
    amino_tag = tag.translate
    last_5 = amino_tag[-5..-1] || amino_tag
    stop_codons = amino_tag.scan(/\*/).size
    tag if stop_codons <= num_sc && last_5 != 'PTYLD'
  end
end
```

Listing A.3: finder class

```
module Bio
  class Finder
    attr_accessor :tag_range
    attr_accessor :stop_codons

    def initialize(path)
      @file ||= path
    end

    def find
      warn "[DEPRECATION] `find` is deprecated. use `extract_tags` ↩
          instead"
      extract_tags
    end

    def extract_tags
      ext = File.extname(@file)
      if ext == '.fastq' || ext == '.fq'
        find_tags_fastq(@file)
      elsif ext == '.fasta' || ext == '.fna' || ext == '.fas'
        $stderr.puts 'Only Fastq input is supported!' #find_tags_fasta(↩
            @file)
      else
    $stderr.puts "File Read Error: '#{ext}' is not a supported ↩
       sequence file format"
  end
end

def find_tags_fastq(file)
  fastq = Bio::Faster.new(file)
  tags = []
  fastq.each_record do |seq_header, sequence, quality|
    tags << find_tags(seq_header,sequence,quality)
  end
  tags #.map{|r| r.join("\n")}.flatten.join
end

def get_frames(bioseq)
  orfs = 6.times.map{|frame| bioseq.translate(frame+1)}.map{|trans|↩
      trans.split('*')}
  orfs.each_with_index.map{|orf,frame| frame + 1 unless orf.map{|e|↩
      e if e.seq.size > 99}.compact.empty?}.compact
end

def bioseq_obj(sequence_string)
  Bio::Sequence::NA.new(sequence_string)
end

def to_bioseq_generic(sequence_string)
  Bio::Sequence.new(sequence_string)
end

def to_fastq(name,sequence,qualities)
  output_array = []
  output_array << "@#{name}"
  output_array << sequence
  output_array << "+#{name}"
  unless qualities.empty?
    if qualities.is_a? Array
      output_array << qualities.map{|score| (score + 33).chr}.join
    else
      output_array << qualities.split(/\s+/).map{|score| (score.↩
         to_i + 33).chr}.join
    end
  else
    output_array << 'D' * sequence.length # a default value incase ↩
       no quality data is provided
  end
  output_array
end

def find_tags(seq_header,sequence,quality=[])
  bio_seq = bioseq_obj(sequence)
  frames = get_frames(bio_seq)

  frames.each do |frame|
    protein = bio_seq.to_protein(frame)
```

```
        n_term_motifs = bio_seq.n_term_motifs(protein)
        c_term_motifs = bio_seq.c_term_motifs(protein)

        n_term_positions = bio_seq.n_term_positions(n_term_motifs,↩
           protein)
        c_term_positions = bio_seq.c_term_positions(c_term_motifs,↩
           protein)

        tag_len_diffs = bio_seq.tag_len_diffs(c_term_positions,↩
           n_term_positions)
        candidate_tag_lens = bio_seq.candidate_tag_lens(tag_len_diffs,↩
           tag_range)

        candidate_tag_lens.each_with_index do |list,index|
          unless list.compact.empty?
            aa_start = n_term_positions[list.compact.flatten.pop]
            aa_stop  = c_term_positions[index]

            tag_start = ((aa_start + 1) * 3) - 2
            tag_stop  = (aa_stop * 3)

            frame > 3 ? @bioseq   = bio_seq.reverse_complement : ↩
               @bioseq = bio_seq.seq
            frame > 3 ? qualities = quality.reverse : qualities = ↩
               quality

            tag_quals = qualities[tag_start..tag_stop]

            case frame
            when 1,4
              tag = @bioseq.clip_tag(tag_start,tag_stop,0,stop_codons)
              return to_fastq(seq_header,tag,tag_quals) if tag

            when 2,5
              tag = @bioseq.clip_tag(tag_start,tag_stop,1,stop_codons)
              return to_fastq(seq_header,tag,tag_quals) if tag

            when 3,6
              tag = @bioseq.clip_tag(tag_start,tag_stop,2,stop_codons)
              return to_fastq(seq_header,tag,tag_quals) if tag
            else
            end #case statement
          end #unless statement
        end #candidate_tag_lens
      end #frames loop
    end #find_tags method
  end #class
end #module
```

A.2 A Ruby program to classify DBLα sequence tags

Listing A.4: A program to classify DBLα sequence

```
#This script depends on the bio-dbla-alpha classifier package.

#!/usr/bin/env ruby
require '/Users/george/Code/Ruby/bioruby-dbla-classifier/lib/bio-dbla-↩
    classifier'
seq_file = ARGV[0]

Bio::FlatFile.open(seq_file).each do |entry|
 tag = Bio::Sequence::AA.new(entry.seq)
 puts "#{entry.definition}\t#{tag.cyspolv_group}\t#{tag.dsid}\t#{tag.↩
    bs_group}\t#{tag.cys_count}\t#{tag.length}\t#{tag.var1_status}\t#{↩
    tag.sig2_status}\t#{tag.groupA_status}"
end
```

Listing A.5: The nucleotide class

```
class Bio::Sequence::NA
  attr_accessor :mut_pos

  #position specific polymorphic block 1
  def pspb1_dna(anchor_pos=0,win_len=42)
    self[42 + anchor_pos,win_len]
  end

  #position specific polymorphic block 2
  def pspb2_dna(anchor_pos=0,win_len=42)
    if !ww_missing?
      return self[(ww_pos * 3) - 12 - anchor_pos - win_len, win_len]
    elsif !vw_missing?
      return self[(vw_pos * 3) - 36 - win_len - anchor_pos, win_len]
    else
      return '....'
    end
  end

  #position specific polymorphic block 3
  def pspb3_dna(anchor_pos=0,win_len=42)
    if !ww_missing?
      return self[(ww_pos * 3) + 42 + anchor_pos, win_len]
    elsif !vw_missing?
      return self[(vw_pos * 3) + 8 + anchor_pos, win_len]
    else
      return '....'
    end
  end

  #position specific polymorphic block 4
  def pspb4_dna(anchor_pos=0,win_len=42)
    self[self.length - 36 - win_len - anchor_pos, win_len]
  end
```

```
def polv1_dna_pos
  mut_pos.map{|mut| mut - (aa_seq.polv1_pos * 3)}
end

def polv2_dna_pos
  mut_pos.map{|mut| mut - (aa_seq.polv2_pos * 3)}
end

def polv3_dna_pos
  mut_pos.map{|mut| mut - (aa_seq.polv3_pos * 3)}
end

def polv4_dna_pos
  mut_pos.map{|mut| mut - (aa_seq.polv4_pos * 3)}
end

#return an array of distances from each polv
def dist_from_polvs
  [polv1_dna_pos, polv2_dna_pos, polv3_dna_pos, polv4_dna_pos]
end

#return the homology block D
#def block_D
# self[0,36]
#end

#return the sequences in homology block E
#def block_H
# self[-36..-1]
#end

def block_D
  self[0, ((aa_seq.polv1_pos + 1) * 3) + 9]
end

def block_d_start
  1
end

def block_d_stop
  ((aa_seq.polv1_pos + 1) * 3) + 9
end

#+14 aa at the end of PSPB1
def block_E
  start = index(pspb1_dna) + 42
  myend = start + 42
  self[start..myend]
end

def block_e_start
  index(pspb1_dna) + 42
end
```

```
def block_e_stop
  block_e_start + 42
end

def block_F
  mystart = index(pspb2_dna) + 42
  myend   = index(pspb3_dna)
  self[mystart..myend]
end

def block_f_start
  index(pspb2_dna) + 42
end

def block_f_stop
  index(pspb3_dna)
end

def block_H
  self[(aa_seq.polv4_pos * 3)..-1]
end

def block_h_stop
  self.size
end

def block_h_start
  aa_seq.polv4_pos * 3
end

#catch all methof missing.
#TODO: do i really need it?
def method_missing(m,*args, &block)
  puts 'undefined block'
end

private
def accepted_length
  aa_seq.accepted_length #300..500
end

def aa_seq
  self.translate
end

def ww_pos
  aa_seq.ww_pos #rindex("WW")
end

def vw_pos
  aa_seq.vw_pos #rindex("VW")
end

def ww_missing?
```

```
    aa_seq.ww_missing? #true unless aa_seq =~ /WW/i
  end

  def vw_missing?
    aa_seq.vw_missing? #true unless aa_seq =~ /VW/i
  end

  def vw_ww_missing?
    aa_seq.vw_ww_missing? #true if ww_missing? && vw_missing?
  end
end
```

Listing A.6: amino acid class

```
class Bio::Sequence::AA
  BS1    = /DDDVEKGLKIVFEK|DKGEKKKLEKNLKD|DSRTDKLEENLRKI|GGGGRKKLEDNLKE|↩
      GGRGRKKLEDNLIE|GGRGRKKLEDNLKE|GGRGRKQLEENLQK|GKKKEKEKIYGNIE|↩
      GPKQEKKELEENLK|GPSQEKIKLEENLK|GPSQEKKKLEENLK|HEQGNNKLEAILKT|↩
      HEQGNNKLEARLKT|HEQGYNKLEAILKT|HEQGYNKLEAISKT|HNHIKKPLLENLEQ|↩
      HNHKKKPLLDNLEK|HNHKKKPLLENLEQ|HNNKKKALLDNLEK|HNQKKINLEKSLHR|↩
      HQQRKGKLEENLRN|HQQRKRKLEENLRN|HSKEKEKLQTNLKN|KNESEIKRKEKLQR|↩
      KNESEKNTKKKLQG|KNESEKRTKEKLQG|NDADKVEKGLQVVF|NDADKVQKGLQVVF|↩
      NDDVEKGLKIVFEK|NDEDDVEKGLKIVF|NDKDAAQKVLRTVF|NDKDAVQKGLRAVF|↩
      NDKDAVRHGLKVVF|NDKDYVENGLKKVF|NDKEKDQRKKLDEI|NDKEKDQRKKLDEN|↩
      NDKVEKGLQVVFGK|NDKVEKGLREVFKK|NDKVEKGLREVFRK|NDKVENGLKKVFDK|↩
      NDKVENGLREVFKK|NDNVEKGLKAVFRK|NDNVEKGLKKVFDK|NDNVENGLREVFKK|↩
      NDQDDVEKGLKIVF|NDQDEVWNGLRSVF|NEDDKVQKGLQVVF|NEDVEKGLKVVFKK|↩
      NEDVEKGLKVVFQK|NEEDAVQKGLKKVF|NEEDAVQKGLKVVF|NEEDAVQKGLRAVF|↩
      NEKDAVQNGLKKVF|NEKVEIGLKKVFDK|NEKVEIGLKKVFEK|NEKVEIGLKKVFKK|↩
      NEKVEYGLRKLFKK|NEMVEIGLKKVFKK|NENVEKGLKIVFEK|NENVEKGLKKVFDK|↩
      NENVEKGLQVVFGK|NEQDEVWKGLRDVF|NGDYKEKVSNNLKT|NGDYKEKVSNNLRA|↩
      NGDYKKKVSNNLKT|NKDDKIEKSLRAIF|NKDDKVEKGLRAIF|NKDDKVQKGLKAVF|↩
      NKDDKVQKGLQVVF|NKDDKVQKGLRAVF|NKHDNIEKGLREVF|NKNVEIGLKKVFDK|↩
      NKNVEIGLKNVFKN|NKQEKEKREKLDEN|NKQRKKILQEKLEN|NNDDDKIKKGKLRG|↩
      NNDNDKIKKEKLQE|NNDNDKIKKEKLRG|NNDNDKIKKGKLRG|NNDNDKVKKEKLQN|↩
      NNDNDKVKKEKLRG|NNDNDRVKKEKLQN|NNDVEKGLDVVFKK|NNDVEKGLKVVFKK|↩
      NNDVVKGLDVVFKK|NNEKDMREKQKLQS|NNEKDMTEKQKLQS|NNESEIKRKEKLRG|↩
      NNESEKKKREELQG|NNETDKEQKVKLEK|NNHDNVEKGLKAVF|NNHDNVEKGLKKVF|↩
      NNHDNVENGLKAVF|NNHDNVENGLREVF|NNKEKEKIEKSLQN|NNKENEKLQENLKR|↩
      NNVDAVQEGLKVVF|NPEDKVHEGLKVVF|NPEVEKGLKAVFRK|NPEVENGLREVFNK|↩
      NPQDKVQEGLKNVF|NPQDKVQEGLKVVF|NPQDKVQKGLREVF|NQEDKVQEGLKVVF|↩
      NRKEKGKLQTNLKN|NSDDKVEKGLREVF|NSDDKVENGLKKVF|NTVDKIHEGLKVVF|↩
      NTVDKVHEGLKVVF|NVHDKVEKGLQVVF|NVHDKVEKGLQVVL|NVHDKVEKGLREVF|↩
      NVHDKVERGLREVF|NVHDKVETGLREVF|SEKVEYGLRKLFKK|SNKEKEKIENSLQN|↩
      TDKDAVQKGLRAVF|TDKDDVENGLREVF|TDKDEVKEGLKVVF|TDKDEVWKGLRAVF|↩
      TDKDYVENGLKAVF|TDKDYVENGLKKVF|TDKVENGLKEVFDK|TDKVENGLKKVFDK|↩
      TDKVENGLKKVFDN|TDNDAVQKGLRAVF|TDNDEVWKGLGSVF|TDNDEVWKGLRSVF|↩
      TDNDEVWTGLRSVF|TDNVEKGLRAVFGK|TDNVENGLREVFKK|TEKDDVEKGLKIVF|↩
      VNGNDKLESNLKKI|YDEKEKNRRKQLEN|YNERDRAQKKKLQD|YNERDRDKKRKLQE|↩
      YNERDREKKRKLQD|AKNDYTGDHPNYYK|ALKHYKDDTKNYYQ|ARYKKDEEDGNYYK|↩
      ARYKKDEEDGNYYQ|DKNRGKLGALSLDD|DNNSDKLRDLSVDK|EHYEDVDGSGNYLK|↩
      EHYKDVDGSGNYYK|EKNYPDDGSGNYSK|EKNYYNDGTGNYYK|EYYEDKDPDKNYYQ|↩
      EYYNDTNNKINYVK|GIIDYDHDGPHYYK|GINAYNDGSENYYK|GINDCDRDGPEYYK|↩
      GINDYDGDGPEYYK|GINDYDRDGPEHYK|GINDYDRDGPEYYK|GINDYNDGSGNYFK|↩
```

```
GINDYNDISGNYYK|GKKYYNDETGNYYK|GKKYYNDGSGNYYK|GKKYYNDGTGNYVK|↩
GKNYPDDGSGNFYK|GKNYPDDGSGNYYK|GKQYYNDENGNYYK|GQTYPDDGSGNYYK|↩
HHYKDDDISGNYSK|IEARYKKDDDNYYQ|IETRYENDGPNYYQ|IETRYGSDTTNYYQ|↩
IHNYDDNGSGNYYK|IKNDKTLNNLSNGQ|ISYYNADEKGNFYK|KAKERYKDIKNYYQ|↩
KAKYEDLKTLPIDD|KARYKDRKDPNYYK|KDHYKDEKDGNFFQ|KDHYQDDGTGNYYK|↩
KDYYNADEKGNYYK|KEEYGDLKDVPIDD|KEISDYDNDPNYYK|KEKYGDLKDVPIDD|↩
KEYYQDDGTGNYYK|KGINDYDGDPNYYK|KHYADEDGSGNYYK|KHYAHDDGSGNYYK|↩
KHYAHDDGSVNYYK|KHYAHGDGSGNYLK|KHYAHGDGSGNYSK|KHYAHGDGSGNYYK|↩
KHYTDTHGSIDYDK|KIKDYDGDGPEYYK|KINDYDGDGPEYYK|KITHYDDISGNYYK|↩
KKHYENDTDKNYYQ|KKHYKKDEDPNYYK|KKKKKGLSELSTEK|KKVYPEDVTGNYFK|↩
KKVYPEDVTGNYYK|KNAYPDDGFGNYYK|KNAYPDDGSGNYFK|KNAYPDDGSGNYYK|↩
KNDYNPDGSGNYFK|KNDYNPDGSGNYYK|KNENTDLNKLTTEK|KNKNTKLSTLTLEK|↩
KNNDRTLNNLSIGQ|KNYNYDEDGPEYYK|KNYNYDKDGPEYYK|KNYYNPDEAGNYYK|↩
KNYYNPDGAGNYYK|KNYYNPDGSGNYYK|KPHYKDDGFGNYYK|KQHYKDDGSVNYYK|↩
KQHYKEDKDENYYK|KQNNKKLKDLTDKH|KRYYNDDTDDNFYQ|KRYYNDDTDNNFYQ|↩
KRYYNDDTDNNLYQ|KSYYDADEKGNYYK|KSYYNADEKGNYYK|KSYYNADGEGNFYK|↩
KTIYADLKDVEIDD|KTSNSNLKELSLDK|KTSNSNMDTLSLEQ|KTSNTNMNTLSLDK|↩
KVHYKENKDGNYVK|KVHYKENKDGNYYK|KVKYPDLKDLQIDD|KVKYPDLNDIEIDD|↩
KVKYPDLNDVEIDD|KVKYQDLKDVEIDD|KYYNDTNNKINYVK|LKKHYQKDAPNYYK|↩
LKTRYKKDDDNYYQ|LQARYKKDGDDFFK|LQERYNDPKGDFFQ|LQTRYTNDGDNYFK|↩
MESNANLKKHTLER|NEHYKEVKNGNYVK|NHYKDDDGSGNYYK|NHYKDDDISGNYSK|↩
NHYKDDDISGNYYK|NHYKDDNGSENYYK|NHYKDDNGSGNYYK|NKNKPPLDKLSVDK|↩
NKNKSPLDKLSLEQ|NKNNVPLDKLSLDK|NKNNVPLHNLSLDK|NVHYKDDGSENYYK|↩
NVHYKDDGSGNYYK|NVHYKEVKNGNYVK|NYNYDEDGSGNYVK|NYNYDEDGSGNYYK|↩
NYYADGDKSGNYYK|NYYNNTGNNANYAK|NYYNNTGNNVDYVK|NYYNNTGNNVNYAK|↩
PHYTNDRGLADYVK|QGIIDYDNDPNYYK|QISDYDGDGPEYYK|QISDYTGDHPNYYK|↩
QKHYEDDGSGNYYK|QKIYEDINNLPIDD|QKIYKDLNNLPIDD|QKNNSALKKLTDKQ|↩
QKSDSSLQRLSIEK|QNYYADDGSGNYSK|QNYYADDGSGNYVK|QNYYKDDPKKNYYK|↩
QQNNNTLENLTDKQ|REHYKEVKNGNYIK|REHYKEVKNGNYYK|REKYKDLKDLPIDD|↩
REKYKDLKDVEIDD|RERYKDLKDVEIDD|RHYADHDKSGNYYK|RIRHYDDGSGNYSK|↩
RIRHYDDGSGNYYK|RITHYNDGSGNYVK|RITHYNGVSGNCVK|RKNNSSLRKLTNEQ|↩
RNENNNLGKLSNEQ|RVKETYKDDPNYYK|SDYKDDDGSGNYYK|SDYKDDDIDGNYYK|↩
SDYKDDDIDGNYYQ|SHYADHDKSGNYLK|SHYADHDKSGNYYK|SHYEDGDKSGNYYK|↩
SHYEDKDKSGNYIK|SHYTDTHGSIDYDK|SKINDYDGDPNYYK|SKITDYDNDPNYYK|↩
SVQERYGNDPNFFQ|TETLYKDEEGNYLK|THYADEDGSENYYK|THYADEDGSGNYVK|↩
TVKETYKDDPNYYK|TVKGTYKDDPYYYK|VKAHYKKDAPYYYK|VKAHYQKDAPNYYK|↩
CAARGNDLYSKNIG|CAARYHPGYFKKSD|CDAPKDANYFIGSG|CDAPQKVDYFRKGS|↩
CDAPQKVDYFRKIS|CDAPRDADYFKNVA|CDAPRDADYFRKGS|CDAPRDAHYFLKSS|↩
CDAPRDANFFIKNS|CDAPYKSRYFIQSE|CDAPYKSRYFMQSE|CDASYKSGYFMQSE|↩
CDTEESDTYFKQSS|CEAPENAYIIKRRI|CEAPGDAHYFRKGP|CEAPKDANYFIGSG|↩
CEAPQKVDYFRKGL|CEAPQKVDYFRKGS|CEASKNANFFIKDS|CEASKNANFFIKNS|↩
CFADGSEEYFIKSS|CFADGSEEYFIQSE|CFADGSEEYFIQSS|CFAHNTEEYFIKSE|↩
CGAGAKDTYFTYSK|CGAGARDEYFIKPS|CGAGEKDTYFTYSK|CGAGEKDTYFTYSN|↩
CGAGEKDTYFVQLD|CGALPKSAYFLQSE|CGALPKSAYFMQLE|CGALPKSAYVLQSE|↩
CGATMNDIFSKNIG|CGTGENDTYFKNSS|CIAPRDAHYFLKSS|CKAKEGDIYSKTTD|↩
CKAPEDADYFRKGS|CKAPGDVNFFIKNS|CKAPGDVNYFRKIS|CKAPKDADYFRKGS|↩
CKAPKDAHYFLKSS|CKAPKDANFFIKIS|CKAPKDANYFIGSG|CKAPKGANYFRKES|↩
CKAPNGANYFRKKS|CKAPPKVDYSRNIS|CKAPQDANYFRKGL|CKAPQDANYFRKIS|↩
CKAPQDANYFRNIS|CKAPQDANYFRNVS|CKAPQDANYFTKES|CKAPQGANYFRNIS|↩
CKAPQKANYFRKGS|CKAPQKVDYFRKGS|CKAPQSVHYFIKTS|CKAPTGADYFKKKS|↩
CKAPTGADYFVYKP|CKAPTGAHYFLKSS|CKASKNANFFIKNS|CKASRNAHYFLKSS|↩
CKASRNANYFRKAL|CKASRNANYFRKIS|CNAPDKAEYFVYKS|CNAPDNVNYFRKYS|↩
CNAPGDAHYFRKDP|CNAPGDVHYFRKDP|CNAPNISGYFMQSE|CNAPNISGYFMQSG|↩
CNAPYDANYFRKTS|CNAPYDANYVRRKS|CNAPYDANYYRKYS|CNAPYDANYYRQTC|↩
```

```
    CNAPYEAQYFIKPS|CNAPYEAQYFIKSS|CNAPYEAQYYIKSS|CNAPYKAQYYIKSS|↩
    CNAPYKAWYFMHSE|CNAPYKSRYFIQSE|CNAPYKSRYFMHSE|CNAPYKSRYFMQSE|↩
    CRAPKNAHYFIKSS|CRAPNEANYFKNVA|CRAPNGANYFRKGL|CRAPQKANYFKNVA|↩
    CSADDSEDYFIQSE|CSADDSEDYFIRSE|CSADGSEDYFIKSS|CSADGSEEYFIQSE|↩
    CSADGSEEYFKKQS|CSAGPKDTYFIKSG|CSAGQKDTYFIKPN|CSAPDNAKYFKPPK|↩
    CSAPDNAKYVKYFP|CSAPDYAKYFRQTC|CSAPGDAKYVKNFP|CSAPGDAKYVKYFP|↩
    CSAPGDVNYFRKES|CSAPGDVNYFRKFS|CSAPGDVNYFRKGL|CSAPGDVNYFRKIS|↩
    CSAPHNAQYVKYVP|CSAPRDADYFIKNS|CSAPRDAQYFIKSS|CSAPYCADYFKKKS|↩
    CSAPYDANYVRRKS|CSAPYEAQYFIKSS|CSAPYEAYYFTYKS|CSAPYGANYYRKYS|↩
    CSAPYHPGYFRQSK|CSAPYKSQYFIKSS|CSAPYNAHYFIKSS|CSAPYYADYFKKKP|↩
    CSAPYYADYFKKKS|CSAPYYADYFKSVA|CSAPYYADYFRKGS|CSAQNNEVYFINSE|↩
    CSVPYEAYYFTYKS|CTAPDKANYFIYKS|CTAPDNVNYFRKYS|CTAPYGANYYRKYS|↩
    CVAGEGNTYFIQLD|CVAPENAYFRKTEA|CYAPNNANYFIGSG|CYIPYCVNYFKNIS|↩
    CYIPYYVNYFKDIS|CYIPYYVNYFKKKS|CYIPYYVNYFKKTP|CYIPYYVNYFKKTS|↩
    CYIPYYVNYFKNIS|YKAPKDAHYFLKSS|YKAPQDANYFRNVS|YKAPRKADYFRNIS|↩
    YKAPRKANYFIYKS|AVSSNKCGHNDMNV|EFTGGYCGRDETDV|EFTSGYCGRNETNV|↩
    FENAGKCGHNDNRV|FENAGKCRRNDNKV|FLYPKCGHNNKNDL|FSDNGHCGRNETNV|↩
    FSDNGPCGRKELIV|FSNEHCGHHNNDDP|FSNEHCGHYKNGDP|FSNEYCGHKKNEDP|↩
    FSNEYCGHYKNGDP|FSNNGPCGRNETDV|FSNNKCGHSNGGDP|FSNPKCGHSNGGDP|↩
    FSNRGPCGRNETDV|FSNSGKCGGKEAPV|FSNSGPCGRKELTV|FSNSGTRGRKELTV|↩
    FSNSKCGHHNNDGP|FSSDRCGHNNNDGP|FSSEGKCGHKEGTV|FSSEYCGHYKNGDP|↩
    FSSHGKCGHNEGAP|FSSQGQCGHTEGTV|FSSSGPCGRDEAPV|FTDDGKCGHYEGAP|↩
    FTDGHCGRTQEGHV|FTDIGKCGGKEAPV|FTDIGKCGHGDKDV|FTDIGKCGHKQGNV|↩
    FTDIGKCGHNEGAP|FTDIGKCGHNKGSV|FTGGGQCGRNETDV|FTGGGQCRRNDNSV|↩
    FTNDGKCGHTEGTV|FTNDGKCGHYEDAP|FTNDGKCGHYEDNV|FTNDGKCGHYEGAP|↩
    FTNDGKCGHYENNI|FTNDGKCGRYEGAP|FTRQGYCGHSETNV|FTSAGKCRHNDNSV|↩
    FTSAGKCRRNDNSV|FTSEGKCGHNDNRV|FTSEGKCGRNETNV|FTSEGQCGHDENKV|↩
    FTSEGQCGHNDKSV|FTSEGQCGHSETNV|FTSEGQCRRNDNSV|FTSEGQRGHSETNV|↩
    FTSEGRCGHSETNV|FTSHGKCGHSEGAP|FTSHGKCGRNETNV|FTSIGKCGHNKGSV|↩
    FTSQGQCGHKEGTV|FTSQGQCGHSETNV|FTSQGQCGHTEGTV|FTSQGQCGRNERNV|↩
    FTSQGYCGHSETNV|FTSQGYCGRKEAPV|FTSQGYCGRKELTV|FTSVGYCGHNKGIR|↩
    FTSVGYCGHNKGSV|FTTEGYCGRDEGAP|FTTEGYCGRNEGAP|FWDRKCGHSNEGAL|↩
    FWDRKCGHSNEGAP|FWDRKCGHSNGGDP|IFSNEHCGHKQGSV|IVSFDQCGHNDMDV|↩
    IVSFDQCGHNDVDV|KFSERKCGHDENAP|KFSERKCGHNEGSP|KFSSDRCGHNEGDP|↩
    LFDYNCGHHKDNNV|LFSDGHCGNKDGTV|LFSDHKCGHEESRV|LFSDYKCGHYEDAP|↩
    LFSDYKCGHYEGSP|LFSNAYCGHYEGSP|LFSNDYCGHKQGNV|LFSNPKCGHEQGNV|↩
    LFSNPKCGHEQGTV|LFSNPKCGHKQGKV|LFSNRQCGHDENKV|LFSNRQCGHEQGNV|↩
    LFSNRQCGHGEHEV|LFSNRQCGHNEGAP|LFSNSKCGHDENKV|LFSNSKCGHDESKV|↩
    LFSNSKCGHEQGNV|LFSNSKCGHRQGNV|LFSNYKCGHYEDAP|LFWDRKCGHDERNV|↩
    LLFSNYKCGHYEGS|LWNDKCGHHVDKDV|NFSNPKCGHDEGIV|NFSNPKCGHKQGNV|↩
    NLILTHPKCGHDTD|PSYIKCGHNNKDDP|PSYLKCGHNNKDDP|SEGKCGHKETERDL|↩
    SFADAYCGRGDENV|SFSDHKCGHDENAP|SFSDHKCGHGDKDV|SFSDHKCGHYEGAP|↩
    SFSDRKCGHYEGAP|SFSNDYCGHNENKV|SFSNDYCGHRQGSV|SFSNEYCGHRQGSV|↩
    SFSNGQCGHRDENV|SFSNPKCGHEQGNV|SFSNPKCGHGDNEV|SFSNPKCGHGEHEV|↩
    SFSNPKCGHNENKV|SFSNSKCGHGEHEV|SFSSEYCGHEQGNV|SFSSEYCGHGDNEV|↩
    SFSSEYCGHRQGSA|SFSSEYCGHRQGSV|SFTNGQCGHNEENV|SFTNPKCGHDENKV|↩
    SFTNPKCGHGDNEV|SFTNPKCGHGEHEV|SFTNPKCGRGDNEV|SLILPYSKCGRDTD|↩
    SQGQCGRNENNGYP|SSTNTQCRCATNDV|TEGYCGRNENNGYP|TFSGYWCGHYEGAP|↩
    TFSNDYCGHGEHEV|TFTYTKCGHDENKV|TSEGQCGHNDKMRP|TVSFDQCGHNDMDV|↩
    TVSFDQCGHNDMHV|TVSNAKRREGDENP|TVSSNKCGHNDMDV|VFSNRQCGHYEDAP|↩
    VFSNRQCGHYEDVP/
BS2    = /DKGEKKKLEENLKN|DQERKHLLEKRLET|DQERKQHLEKRLET|DQQEKAKLENNLKR|↩
    DQQEKLYLENNLKK|DRKEKVKLEENLKN|GPDQEKKKLEENLR|GPNQEKKLLENKLK|↩
    HEPGIQHLEKRLES|HEPGIQYLEKRLES|HEPGKQHLEERLEQ|HEPGKQHLEERLER|↩
    HEPGKQHLGERLEQ|HEQGINRLEARLKT|HEQGNNRLEARLKT|HEQGYNRLEARLKT|↩
    HNRRKEKLETRLEE|HQQRKHLLEKRLET|KKKLEENLRNIFKN|NDEEKKKRDELEKN|↩
    NTHESAQRKKLEEN|REKGKSRLEARLKT|REPGKQHLEERLER|YNEKDQEEKRKLQE|↩
    YNETDKVQKAILQQ|YSQKYKDEKSKLEE|AKNHYNDTSKNYYK|ARDHYNDTSGNYYQ|↩
    DAKKHYGDDENYYK|IEHYKDDPEENFYE|IKSNYNDSEGNYFK|IKSQYDDNEGNYFK|↩
    KDHYKGDEANNYFQ|KDNNTKLNDLSIQE|KKTNPALKSFTNEE|KNNNNELNNLSLDK|↩
    KNNNNKLSNLSTKE|KSSYNDDGTGNYFK|KSYYKNDNDRNYFK|NKNKSPLDKLSLDK|↩
    NKNNPPLYKLSLEK|NNNAAKLSELSTAQ|NYYEDNDTDKNYYQ|QKENGDINTLKPEE|↩
    QNKNENLKSLSLDK|QNNNTKLQNIPLHE|QRNNIKLQNIPLHE|QRNNIKLQTLTLHQ|↩
    SAKEHYQDTENYYK|TRYKKDDEDGNFFQ|VKDRYQNDGPDFFK|CDAGAADEYFKKSG|↩
    CDAGQKDTYFKQSS|CEAGTSDKYFRKTA|CEAKSDDKYNVIGP|CGAGMKDIYSKTMN|↩
    CGAPKEAKYFRKTA|CGAPSDAQYFRNTC|CGASEDAKYKVIGP|CGATMNDIFSKNIR|↩
    CGATVDDIFSKNIR|CGATVDDISSKNIR|CGTEDKDTYFIKSG|CGVEENAKYFRESS|↩
    CHAPPDAQYTKKGP|CKAEVDDIYSKTAN|CKAKEGDIYSKTAN|CKAKEGDIYSKTMN|↩
    CKAKKGDIYSKTMN|CKAKKGGIYSKTMN|CKANDDAEYFRKKD|CKAPDKANYFEPPK|↩
    CKAPDKANYFKPPK|CKAPEEDHYFKPAQ|CNAPKDANYFEYNS|CNAPTGADYFVYKP|↩
    CNAWGNTYFRKTCS|CRAEEKDIYSKTTD|CRAEEKDTYFKNRE|CRAEEKEIYSKTTD|↩
    CRAEEKGTYFKNRE|EFSGGKCGHKDNNV|EFTDGHCGHNEENV|EFTDGHCGHRQGNV|↩
    EFTGGQCGRDGENV|ESNKGQCRCFSGDP|ESNMGQCRCFSGDP|ESNMVQCRCFSGDP|↩
    ETHGYCRCVNRVDV|FSNDQCGHNNGGAP|FSNDQCGHNNRGDP|FWYPKCGHHVKQDV|↩
    FWYPKCGHHVKQEV|FWYPKCGHHVRQDV|FWYPKCSHHVKQDV|GANAIKAGDNVSIV|↩
    GGHYKNCHCIGGDV|GGTYKNCRCASGNV|ILFDYKCAHDNDKV|KFSNPKCGHNEGSP|↩
    LFYYKCGHYVYKDV|LWNDKCGHHVKQDV|LWNYKCGHHVNQDV|LWNYNCGHHVNQDV|↩
    LWNYNCGHHVNRDV|PCSVQKCTCINGDP|PNKCRCEDANADQV|SFTNGQCGRDGENV|↩
    SWYPKCGHHVKQDV|TLFDYKCGHDENAP|TLWNEKCGHGDYNL|TPTQGKCHCIDGTN|↩
    TPTQGKCHCIDGTV/
SIG2  = /AKLSELSTAQ|CRAEEKDTYF|KCGHHVNQDV|HEPGKQHLEE|EPGKQHLEER|↩
    PGKQHLEERL|GKQHLEERLE|KQHLEERLEQ|QHLEERLEQM|HLEERLEQMF|LEERLEQMFE|↩
    EERLEQMFEN|ERLEQMFENI|RLEQMFENIK|LEQMFENIKN|EQMFENIKNN|QMFENIKNNN|↩
    MFENIKNNNA|FENIKNNNAA|ENIKNNNAAK|NIKNNNAAKL|IKNNNAAKLS|KNNNAAKLSE|↩
    NNNAAKLSEL|NNAAKLSELS|NAAKLSELST|AAKLSELSTA|RAEEKDTYFK|AEEKDTYFKN|↩
    EEKDTYFKNR|EKDTYFKNRE|KDTYFKNREN|DTYFKNRENG|TYFKNRENGK|YFKNRENGKL|↩
    FKNRENGKLL|KNRENGKLLL|NRENGKLLLW|RENGKLLLWN|ENGKLLLWNY|NGKLLLWNYK|↩
    GKLLLWNYKC|KLLLWNYKCG|LLLWNYKCGH|LLWNYKCGHH|LWNYKCGHHV|WNYKCGHHVN|↩
    NYKCGHHVNQ|YKCGHHVNQD|AKLSELSTAQ|CKAKEGDIYS|NCGHHVNQDV|HEPGKQHLEE|↩
    EPGKQHLEER|PGKQHLEERL|GKQHLEERLE|KQHLEERLEQ|QHLEERLEQM|HLEERLEQMF|↩
    LEERLEQMFE|EERLEQMFEN|ERLEQMFENI|RLEQMFENIK|LEQMFENIKN|EQMFENIKNN|↩
    QMFENIKNNN|MFENIKNNNA|FENIKNNNAA|ENIKNNNAAK|NIKNNNAAKL|IKNNNAAKLS|↩
    KNNNAAKLSE|NNNAAKLSEL|NNAAKLSELS|NAAKLSELST|AAKLSELSTA|KAKEGDIYSK|↩
    AKEGDIYSKT|KEGDIYSKTA|EGDIYSKTAN|GDIYSKTANG|DIYSKTANGN|IYSKTANGNT|↩
    YSKTANGNTT|SKTANGNTTL|KTANGNTTLW|TANGNTTLWN|ANGNTTLWNY|NGNTTLWNYN|↩
    GNTTLWNYNC|NTTLWNYNCG|TTLWNYNCGH|TLWNYNCGHH|LWNYNCGHHV|WNYNCGHHVN|↩
    NYNCGHHVNQ|YNCGHHVNQD|GDINTLKPEE|CRAEEKDIYS|NCGHHVNQDV|HEPGKQHLEE|↩
    EPGKQHLEER|PGKQHLEERL|GKQHLEERLE|KQHLEERLER|QHLEERLERI|HLEERLERIF|↩
    LEERLERIFA|EERLERIFAN|ERLERIFANI|RLERIFANIQ|LERIFANIQK|ERIFANIQKE|↩
    RIFANIQKEN|IFANIQKENG|FANIQKENGD|ANIQKENGDI|NIQKENGDIN|IQKENGDINT|↩
    QKENGDINTL|KENGDINTLK|ENGDINTLKP|NGDINTLKPE|RAEEKDIYSK|AEEKDIYSKT|↩
    EEKDIYSKTT|EKDIYSKTTD|KDIYSKTTDN|DIYSKTTDNG|IYSKTTDNGK|YSKTTDNGKL|↩
    SKTTDNGKLI|KTTDNGKLIL|TTDNGKLILW|TDNGKLILWN|DNGKLILWNY|NGKLILWNYN|↩
    GKLILWNYNC|KLILWNYNCG|LILWNYNCGH|ILWNYNCGHH|LWNYNCGHHV|WNYNCGHHVN|↩
    NYNCGHHVNQ|YNCGHHVNQD/

def has_accepted_length?
  true if accepted_length.include? self.length
```

```
end

def start_motif
  #warn "[DEPRECATION] 'start_motif' is deprecated.  Please use `↩
      n_terminal_motif` instead."
  n_terminal_motif
end

def end_motif
  #warn "[DEPRECATION] 'end_motif' is deprecated.  Please use `↩
      c_terminal_motif` instead."
  c_terminal_motif
end

def n_terminal_motif
  self[0,5]
end

def c_terminal_motif
  self[-5,self.length]
end

def ww_pos
  rindex("WW")
end

def vw_pos
  rindex("VW")
end

#number of cysteines
def cys_count
  scan(/C/).size
end

#get the 5' end of the sequence from the 'middle'
def polv1_to_polv2
  slice(rindex(polv1),rindex(polv2) - 6)
end

#get the 3'end of the sequence from the "middle"
def polv3_to_polv4
  slice(rindex(polv3),(rindex(polv4) - rindex(polv3) + 4))
end

#The first position of limited variability(polv1)
def polv1
  self[10,4]
end

#The second position of limited variability(polv2)
def polv2
  if !ww_missing?
    return self[ww_pos - 4,4]
  elsif !vw_missing?
    return self[vw_pos - 12,4]
  else
    return '....'
  end
end

#The third position of limited variability(polv3)
def polv3
  if !ww_missing?
    return self[ww_pos + 10,4]
  elsif !vw_missing?
    return self[vw_pos + 2,4]
  else
    return '....'
  end
end

#The fourth position of limited variability(polv4)
def polv4
  self[self.length - 12,4]
end

#Assigning dsid group based on number of cysteines, presence of REY ↩
    motif in polv2 and MFK in polv1,
def cyspolv_group
  case
  when cys_count > 4 || cys_count == 3 || cys_count < 2
    group = 6
  when cys_count == 4 && polv2 =~ /REY/i
    group = 5
  when cys_count == 4
    group = 4
  when cys_count == 2 && polv1 =~ /MFK/i
    group = 1
  when cys_count == 2 && polv2 =~ /REY/i
    group = 2
  else
    group = 3
  end
  group
end

def polv1_pos
  index(polv1)
end

def polv2_pos
  index(polv2)
end

def polv3_pos
  index(polv3)
end
```

```
def polv4_pos
  index(polv4)
end

#return the block sharing group
def bs_group
  case
  when self =~ BS1
    block_sharing = 1
  when self =~ BS2
    block_sharing = 2
  else
    block_sharing = 0
  end
  block_sharing
end

def is_bs1_and_bs2?
 !!(self =~ BS1) && !!(self =~ BS2)
end

def is_var1_cp1?
  true if (cyspolv_group == 1) && (self =~ /NVHDKVEKGLREVF|↩
      NVHDKVETGLREVF/i)
end

def is_var1_cp2?
  true if (cyspolv_group == 2) && (self =~ /APNKEKIKLEENLKK/i)
end

def is_var1?
  return true if is_var1_cp1? || is_var1_cp2?
end

def var1_status
  if is_var1?
    status = 'var1'
  else
    status = 'not-var-1'
  end
  status
end

# return var group A like tags. Group A like sequences are associated↩
    with disease severity.
def is_groupA_like?
  true if (cys_count == 2 && bs_group == 1) || cyspolv_group == 1 #&&↩
       ( cyspolv_group == 1)
end

def groupA_status
  if is_groupA_like?
    status = 'A'
  else
    status = 'other'
  end
  status
end

#distict sequence identifier(DSID)
def dsid
  "#{polv1}-#{polv2}-#{polv3}-#{cys_count.to_s}-#{polv4}-#{self.↩
      length}"
end

#position specific polymorphic block 1
def pspb1(anchor_pos=0,win_len=14)
  self[14 + anchor_pos,win_len]
end

#position specific polymorphic block 2
def pspb2(anchor_pos=0,win_len=14)
  if !ww_missing?
    return self[ww_pos - 4 - anchor_pos - win_len, win_len]
  elsif !vw_missing?
    return self[vw_pos - 12 - win_len - anchor_pos, win_len]
  else
    return '....'
  end
end

#position specific polymorphic block 3
def pspb3(anchor_pos=0,win_len=14)
  if !ww_missing?
    return self[ww_pos + 14 + anchor_pos, win_len]
  elsif !vw_missing?
    return self[vw_pos + 6 + anchor_pos, win_len]
  else
    return '....'
  end
end

#position specific polymorphic block 4
def pspb4(anchor_pos=0,win_len=14)
  self[self.length - 12 - win_len - anchor_pos, win_len]
end

def accepted_length
  100..168
end

def ww_missing?
  true unless self =~ /WW/i
end

def vw_missing?
  true unless self =~ /VW/i
  end

  def vw_ww_missing?
    true if ww_missing? && vw_missing?
  end

  def group6_bs1
    true if cyspolv_group == 6 && bs_group == 1
  end

  def sig2_like?
    return true if self =~ SIG2
  end

  def sig2_status
    if sig2_like?
      sig2 = 'sig2'
    else
      sig2 = 'non-sig2'
    end
    sig2
  end
end
```

A.3 A Ruby program to extract DBLα sequence tags from chromatograms

Listing A.7: Extracting DBLα sequence tags from chromatograms

```
#!/bin/bash

#This script should be run from the  grouped_by_clones folder. i.e all ↩
    the sequences for each isolate's clones are located
#run the copy_per_colony script first to generate this list of ↩
    sequences

#get the list of isolates from ls the current folder
echo "generating a list of isolate names"

#ls -l | awk '{print $9}' | perl -lne 'print /(^B[0-9]{1,3})/' | sort -↩
    k1.2nr | uniq >isolate_names.txt

find . -type f -name '*.fasta' | grep -Eo 'B[0-9]{1,}' | sort -k1.2n | ↩
    uniq >isolate_names.txt

sed '/^$/d' isolate_names.txt >isolate_names

while read isolate
do
  mkdir -p ../$isolate

  echo "copying data for $isolate"
  cp $isolate\-* ../$isolate

  cd ../$isolate
```

```
#ls | perl -lne 'print /B[0-9]{1,3}-([0-9]{1,2})/' >$isolate\_clones.↩
    txt
ls *.fasta | perl -lne 'print /B[0-9]{1,3}-([0-9]{1,2})/' >$isolate\↩
    _clones.txt

#remove empty lines
sed '/^$/d' $isolate\_clones.txt | uniq >temp.txt
mv temp.txt $isolate\_clones.txt

while read clone
do
  echo "creating clone $clone directory"

  mkdir -p $isolate\_$clone

  echo "copying $isolate-$clone.fasta and $isolate-$clone.fasta.qual ↩
      files to directory $clone"

  mv $isolate\-$clone\_* $isolate\_$clone

  cd $isolate\_$clone

  #change the name of the qual file to *.fasta.qual
  #mv $isolate\_$clone\-M13.qual $isolate\_$clone\-M13.fasta.qual

  #echo "running cap3 for $clone read pairs"
  ~/cap3/cap3 $isolate\-$clone\_M13.fasta $isolate\-$clone\_M13.fasta↩
      .qual >$isolate\_$clone\.cap.txt

  #append the clone names to the contigs rename the contigs
  sed "s/^>/>$isolate\-$clone\_/" $isolate\-$clone\_M13.fasta.cap.↩
      contigs >$isolate\_$clone.renamed.contigs.fasta
  sed "s/^>/>$isolate\-$clone\_/" $isolate\-$clone\_M13.fasta.cap.↩
      contigs.qual >$isolate\_$clone.renamed.contigs.qual

  #create a single line fasta and qual for the assembles and renamed ↩
      contigs
  awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {↩
      printf("\n");}' < $isolate\_$clone.renamed.contigs.fasta >↩
      $isolate\_$clone.SL.renamed.contigs.fasta
  awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {↩
      printf("\n");}' < $isolate\_$clone.renamed.contigs.qual >↩
      $isolate\_$clone.SL.renamed.contigs.qual

  sed '/^$/d' $isolate\_$clone.SL.renamed.contigs.qual >temp.txt
  mv temp.txt $isolate\_$clone.SL.renamed.contigs.qual

  sed '/^$/d' $isolate\_$clone.SL.renamed.contigs.fasta >temp.txt
  mv temp.txt $isolate\_$clone.SL.renamed.contigs.fasta

  #create a fastq file from the assembled contig and the ↩
      corresponding quality file
  echo "creating a fastq file"
```

```
    perl ~/Softwares/makefastq.pl $isolate\_$clone.SL.renamed.contigs.↩
       fasta $isolate\_$clone.SL.renamed.contigs.qual >$isolate\_$clone↩
       .SL.renamed.contigs.fastq

    #ruby ~/Softwares/to_fastq.rb $isolate\_$clone.SL.renamed.contigs.↩
       fasta $isolate\_$clone.SL.renamed.contigs.qual >$isolate\_$clone↩
       .SL.renamed.contigs.fastq

    echo "calling dbla tags for $clone contig"
    ruby ~/Code/Ruby/bioruby-dbla-finder/lib/bio-dbla-finder.rb ↩
       $isolate\_$clone.SL.renamed.contigs.fastq >$isolate\_$clone.SL.↩
       tag.renamed.contigs.fastq

    #echo "appending clone names to the dbla tags"
    #sed "s/^>/>$isolate\_$clone\_/" $isolate\_$clone.tags.contigs >↩
       $isolate\_$clone.renamed.tags.contigs

    #echo "appending clone name to the contigs"
    #sed "s/^>/>$isolate\_$clone\_/" $isolate\_$clone-M13.fasta.cap.↩
       contigs >$isolate\_$clone.renamed.fasta.cap.contigs

    #echo "translating dbla tags for $clone"
    #translate $isolate\_$clone.renamed.tags.contigs >$isolate\_$clone.↩
       translated.dbla.tags.contigs

    cd ..

  done <$isolate\_clones.txt

  cd ../data_per_colony

done <isolate_names

cd ..

echo "concatenate the fastq files"

find . -type f -name '*.SL.tag.renamed.contigs.fastq' -print0 | xargs ↩
   -0 cat >tags.fastq

#print tags.fasta
ruby ~/Softwares/read_fastq.rb tags.fastq seq >tags.fasta

#count the total number of potential tags
countseqs tags.fasta

#sometimes the dblfinder detects multitple tag like regions in longer ↩
   contigs.
#These need to be corrected/removed manually
grep ">" tags.fasta | sort | uniq -c | sort -nk 1 | uniq | awk '{if($1↩
   >1) print}' >duplicated_tags.txt
```

A.4 A bash pipeline to cluster DBLα sequences using an identity cutoff

This bash script uses Vsearch to cluster sequences, Bio-CDHIT-report to parse clustered files.

Listing A.8: A program to cluster sequences using an identity cut-offs

```
#!/bin/bash
########################################################
#This bash script takes a sequence file and clusters the sequences at ↩
   predifined identities (98 - 88)
#For each cluster in each identity, we align the sequences and search ↩
   for changes, positions of the changes and the type of changes.
#We tally all the changes in each identity

#Dependencies
#Bio-cd-hit-report
#Ruby
#AWK
#sed
#translate.rb
#mutation.rb

#Author: George Githinji
#Email: ggithinji@kemri-wellcome.org
#MIT Licence
########################################################

seqfile=$1

[ $# -eq 0 ] && { echo "Usage: $0 nucleotide_file"; exit 1;}

identities=(98 96 94 92 90 88)

seqfilefullpath=$( cd $(dirname $seqfile); pwd)/$(basename $seqfile)

#sort the input sequence with usearch sortbylength command
vsearch -sortbylength $seqfile --output $seqfile.sorted.fasta

sortedfilefullpath=$( cd $(dirname $seqfile); pwd)/$(basename $seqfile.↩
   sorted.fasta)

#create the directories to hold output for each cluster identity
for id in ${identities[@]}; do
  mkdir -p "${id}"
  fr_id=$(bc<<<"scale=2; $id/100")
  echo $fr_id

  #use cd-hit to cluster
  #cd-hit-est -i $seqfile -o $id/cluster.${id} -c $fr_id #retire cd-hit

  #use the uclust algorithm to perform the clustering
  #uclust --input $sortedfilefullpath --uc $id/cluster.${id}.uc --id ${↩
     fr_id} --optimal
```

```
#use vsearch1.1.1
vsearch -cluster_smallmem $sortedfilefullpath -id ${fr_id} -centroids↩
     $id/cluster.${id}.nr.fasta -uc $id/cluster.${id}.uc

#convert the output to cd-hit format
uclust --uc2clstr $id/cluster.${id}.uc --output $id/cluster.${id}.↩
    clstr

cd "${id}"

#Write a file containing the cluster-id and respective list of ↩
    cluster members
#echo "Parsing clusters"

ruby ~/Softwares/parse-cdhit.rb cluster.$id.clstr member >↩
    seqs_per_cluster.txt
ruby ~/Softwares/parse-cdhit.rb cluster.$id.clstr size >↩
    seq_numbers_per_cluster.txt

echo "Sorting cluster numbers"
sort -t$',' -k2 -nr seq_numbers_per_cluster.txt >sorted.↩
    seq_numbers_per_cluster.txt
awk -F, '{if($2 > 1) print $1}'< sorted.seq_numbers_per_cluster.txt >↩
    most_clusters.txt

#create directories for each cluster in most_cluster.txt
while read clustr; do
  echo "processing $clustr"

  mkdir cluster_$clustr
  cd cluster_$clustr

    #write the fasta members for clustr $name
    grep "^${clustr}:" ../seqs_per_cluster.txt | awk -F: '{print $2}'↩
         | tr , '\n' >members.txt
    fastagrep -F -X -f members.txt $sortedfilefullpath >↩
        cluster_$clustr.fasta

    #echo "Translating"
    #ruby ~/Softwares/translate.rb -i cluster_$clustr.fasta -f 1 -o ↩
        cluster_$clustr.aa.fasta

    #echo "Aligning sequences"
    #muscle -in cluster_$clustr.fasta -out cluster_$clustr.aln #-↩
        physout cluster_$clustr.phylip
    prank -d=cluster_$clustr.fasta -o=cluster_$clustr -codon -F -↩
        showtree -quiet

    #replace mutation.rb with polymorphic_sites.rb
    #echo "Searching for polymorphic sites"
    #ruby ~/Code/Ruby/blocks/lib/polymorphic_sites.rb cluster_$clustr↩
        .aln >cluster_$clustr.snps.muscle.txt
    ruby ~/Code/Ruby/blocks/lib/polymorphic_sites.rb cluster_$clustr.↩
        best.fas >cluster_$clustr.snps.prank.txt
```

```
      #echo "searching dN and dS changes"
      #~/Softwares/selection.rb -i cluster_$clustr.aln -S | awk '{print↩
          $4}' | sort | uniq -c | sed -e 's/^[ \t]*//' >cluster_$clustr↩
         .dN_dS.txt
      ~/Softwares/selection.rb -i cluster_$clustr.best.fas -S | awk '{↩
         print $4}' | sort | uniq -c | sed -e 's/^[ \t]*//' >↩
         cluster_$clustr.dN_dS_prank.txt
    cd ..
  done <most_clusters.txt

  echo "collect snps in $id identity threshold"
  #find . -type f -name *.snps.muscle.txt | xargs cat | sed -e 's/^[ \t↩
     ]*//' >all.snps.muscle.txt
  find . -type f -name *.snps.prank.txt | xargs cat | sed -e 's/^[ \t↩
     ]*//' >all.snps.prank.txt

  cd ..
done
```

A.5 A program to cluster sequences within isolates

Listing A.9: A program to cluster sequences within isolates

```
#!/bin/bash

#find the isolate file name in seqs folder
for file in ../seqs/*.fa
do
  fullpath=$(cd $(dirname $file); pwd)/$(basename $file)
  filepath="${file##*/}"
  filename="${filepath%.*}"
  mkdir -p $filename
  cd $filename
  ruby ~/Softwares/get_blocks.rb $fullpath >$filename.block_H.fasta
  clusteringpipeline $filename.block_H.fasta
  cd ..
done

mkdir block_H
mv *.nr block_H
```

A.6 A program to parse cd-hit formatted cluster files

Listing A.10: A program to parse CD-HIT clusters

```
require 'bio-cd-hit-report'

clusterfile=ARGV[0]
out = ARGV[1]

def print_members(report)
```

```ruby
  report.each_cluster{|c| puts "#{c.cluster_id}:#{c.members}"}
end

def print_sizes(report)
  report.each_cluster{|c| puts "#{c.cluster_id},#{c.size}"}
end


report = Bio::CdHitReport.new(clusterfile)

if out == "member"
  print_members(report)
elsif out == "size"
print_sizes(report)
else
  puts "Error"
end
```

Listing A.11: parser class

```ruby
class CdHitParser
  attr_accessor :report_file

  def each
    data,header = nil, nil
    File.open(report_file).each do |line|
      if line[0].chr == '>'
        yield Cluster.new(:name => header,:data => data) if data
        data = ''
        header = line[1..-1].strip
      else
        data << line
      end
    end
    yield Cluster.new(:name => header, :data => data)
  end
end
```

Listing A.12: cluster class

```ruby
class Cluster
  attr_accessor :name, :data

  def initialize(arg={})
    self.name = arg[:name]
    self.data = arg[:data]
  end

  def cluster_id
    name.scan(/Cluster\s(.*)/).join
  end

  def members
    entries.join(',')
  end

  def representative
    data.split("\n").map{|line|line.scan(/>(.+)\.{3}\s\*/)}.join
  end
  alias :rep_seq :representative

  def size
    entries.size
  end
  alias :length :size

  def entries
    data.split("\n").map{|line|line.scan(/>(.+)\.{3}/)}
  end
end
```

Listing A.13: report class

```
module Bio
  require_relative 'cluster'
  require_relative 'parser'

  class CdHitReport
    include Enumerable

    def initialize(file)
      @report  = CdHitParser.new
      @report.report_file = file
    end

    def clusters
      cls = []
      @report.each do |c|
        cls << c
      end
      cls
    end

    def each_cluster(&block)
      clusters.each(&block)
    end

    def total_clusters
      clusters.size
    end

    def get_members(cluster_id)
      clusters.select {|cluster| cluster.cluster_id == cluster_id.to_s↩
          }.map{|c|c.members}
    end
    alias :get_cluster :get_members
  end
end
```

A.7 A Ruby program to translate DNA sequences to amino acid sequences

Listing A.14: A program to translate DNA sequences

```
#!/usr/bin/env ruby

require 'optparse'
require 'ostruct'
require 'bio'

defaults = {frame: 1}
options = OpenStruct.new(defaults)

OptionParser.new do |opts|
  opts.banner = 'Usage: translate -i inputfile -f frame -o outputfile'

  opts.on('-h', 'Translate DNA to Amino Acids') do
    puts opts
    exit
  end

  opts.on('-i', '--infile FILE','DNA input file [required]') do |infile↩
     |
    options.infile = infile
  end

  opts.on('-f', '--frame FRAME',Integer, 'Frame to translate [optional]↩
     ') do |frame|
    options.frame = frame
  end

  opts.on('-o','--outfile FILE','amino acid output file [optional]') do↩
      |outfile|
    options.outfile = outfile
  end
end.parse!


class Utility
  def translate(file,frame)
    Bio::FlatFile.auto(file).map do |entry|
      ">#{entry.definition}\n#{Bio::Sequence::NA.new(entry.seq).↩
         translate(frame)}"
      #puts  Bio::Sequence::NA.new(entry.seq).translate(frame)
    end
  end
end

# read options
file    = options.infile
frame   = options.frame
outfile = options.outfile

util = Utility.new
```

```
begin
  trans_seqs = util.translate(file,frame)

  if outfile.nil?
    #write to the std output stream
    $stdout.puts trans_seqs
  else
    #write to the specified outfile
    output = File.open(outfile,'w')
    output << trans_seqs.join("\n")
    output.close
  end

  $stderr.puts "#{trans_seqs.size} sequences translated!"

rescue TypeError
  $stderr.puts 'Error! Provide an input file. Run translate -h '
end
```

A.8 A Ruby program to find pairwise nucleotide differences from a sequence alignment

Listing A.15: A program to locate pairwise changes from DNA alignments

```
#!/usr/bin/env ruby

#require 'bio'
require File.join("#{ENV['HOME']}/Code/Ruby/find-dbla-mutations/lib","↩
    find-dbla-mutations.rb")
file_path = ARGV[0]

#initilize a class
mutation = Mutation.new

#get the data
data = mutation.fasta_to_hash(file_path)

#locate mismatches
mutation.locate_mismatches(data)

#output a summary of the differences
summary = mutation.change_summary

#which sequences did we analyse?
#puts mutation.subjects

#write the output to the command line. Can be piped to a file
puts summary
```

A.9 A bash program to cluster sequences using uclust

Listing A.16: A bash script to cluster sequences with usearch

```
#!/bin/bash

#Check to see if at least one argument was specified
if [ $# -lt 1 ] ; then
  echo "You must specify the directory -d and identity -i arguments"
  exit 1
fi

#Process the arguments
while getopts d:i: opt
do
  case "$opt" in
    d) dir=$OPTARG;;
    i) identity=${OPTARG};;
esac
done
 echo $dir
 echo $identity

for file in $dir/*.fasta
  do
  filename=$(basename $file)
  filename="${filename%.*}"

  #sort the sequence
  uclust --sort $file --output $filename.sorted.fasta

  #cluster at given identity
  uclust --input $filename.sorted.fasta --uc $filename.uc --id ${↩
      identity}

  #write a fasta output of all the seed sequences as a nr database
  uclust --input $filename.sorted.fasta --uc2fasta $filename.uc --types↩
       S --output $filename.$identity.fasta
done

mkdir -p sorted_fasta
mv *.sorted.fasta sorted_fasta

mkdir -p cluster_files
mv *.uc cluster_files

#sort the cluster file
sort -n -k2 -k4 cluster_files/*.uc >cluster_files/sorted.nr.uc

#from the sorted cluster file write a new file containing the frequency↩
     of members for each cluster
awk '{print $2}' sorted.nr.uc | sort -n|uniq -c | awk '{if($1>2) print↩
    }' | sort -nk1 >clusters.names.sizes.txt
```

A.10 A bash program to generate random substitutions based on a reference sequence

Listing A.17: A script to create random substitutions given a reference sequence in fasta format

```
#!/usr/bin env ruby

require 'optparse'
require 'ostruct'
require 'bio'
require 'securerandom'
require 'pickup'

options = OpenStruct.new

OptionParser.new do |opts|
  opts.on('-h', 'shows this help screen') do
    puts opts
    exit
  end

  opts.on('-i', '--infile FILE','input fasta file') do |infile|
    options.infile = infile
  end

  opts.on('-r','--error_rate Error', Float, 'Error rate per nucleotide'↩
     ) do |error_rate|
    options.error_rate = error_rate
  end

  opts.on('-o','--outfile FILE','Output file') do |outfile|
    options.outfile = outfile
  end

  opts.on('-w' '--weights Weight',Float,'weighting levels') do |weights↩
     |
    options.weights = weights
  end

end.parse!


class Bio::Sequence
  attr_accessor :substitutions
  attr_accessor :weights

  NUCLEOTIDES = ['A','C','T','G']

  # The number of random positions to change
  def positions_to_sub
    (1..self.length).to_a.sample(substitutions)
  end

  # Weighted random nucleotides
  def weighted_random_nucleotides
    weights = []
    weighted_nt = Pickup.new(weights)
    weighted_nt.pick(subsitutions)
  end

  # Returns an array of x random nucleotides where x is the number of ↩
     given substitutions
  def random_nucleotides
    substitutions.times.map{NUCLEOTIDES[rand(NUCLEOTIDES.length)]}
  end

  # Rate based mutations
  def rate_based_mutate
    #substitute nucleotide given positions with random nucleotides
    positions_to_sub.zip(random_nucleotides).map do |position,↩
       nucleotide|
      if self[position] == nucleotide
        nucleotide = (['A','C','G','T'] - [nucleotide]).sample(1)
        self[position] = nucleotide.join
      else
        self[position] = nucleotide.downcase
      end
    end
  end
end


infile = options.infile
error_rate = options.error_rate
weights = options.weights

def biosequences(file)
  Bio::FlatFile.auto(file).map do |entry|
    Bio::Sequence.new(entry.seq.upcase)
  end
end

def calculate_mismatches(error_rate,seq_length)
  (error_rate * seq_length).round
end

def mutate_seqs(bioseq_objects,error_rate)
  bioseq_objects.each do |bioseq|
    seq_len = bioseq.length
    bioseq.substitutions = calculate_mismatches(error_rate,seq_len)
    bioseq.rate_based_mutate
    $stdout.puts ">#{SecureRandom.hex(5)}\n#{bioseq.seq}"
  end
end

begin
sequences = biosequences(infile)
mutate_seqs(sequences,error_rate)

rescue TypeError => type_error
  $stderr.puts type_error.message
```

```
rescue Bio::FlatFile::UnknownDataFormatError =>format_error
  $stderr.puts format_error.message << "provide a fasta format"
end
```

A.11 A Perl program to generate random substitutions

Listing A.18: A perl script to generating random mutations given a sequence

```
#!/usr/bin/perl

use strict;
use warnings;
use Bio::Seq;
use Bio::SeqIO;
use Bio::SeqEvolution::Factory;
use Getopt::Long;

#read command line args
my $seqfile = ''; #= $ARGV[0];

#%identity with the template sequence
my $identity = '';

#transition:transversion ratio
my $rate  = 2;

GetOptions('seqfile=s'=>\$seqfile,'identity=i'=>\$identity,'rate=i'=>\↩
    $rate);

my $inputstream = Bio::SeqIO->new(
  -file   => $seqfile,
  -format => 'Fasta',
);

my $seqobj = $inputstream->next_seq();
my $seq = lc($seqobj->seq);

my $tempseq = Bio::Seq->new(-seq => $seq, -alphabet => 'dna');

my $evolve = Bio::SeqEvolution::Factory->new (-rate => $rate, -seq => ↩
    $tempseq, -identity => $identity);

my @mutated;
for (1..10) { push @mutated, $evolve->next_seq }

foreach(@mutated){print $_-> seq ."\n";}
```

A.12 An R program to simulate random substitutions

Listing A.19: A script to simulate substitutions based on a model

```
#!/bin/bash
```

```
clusterfile=$1

clusterFileFullPath=$( cd $(dirname $clusterfile); pwd)/$(basename ↩
   $clusterfile)

directory=$(dirname ${clusterFileFullPath})

 while read cluster_no; do

   #create a neighbour joining tree
   #muscle -maketree -in $directory/cluster_$cluster_no/↩
      cluster_$cluster_no.fasta -out cluster_$cluster_no.tree -cluster ↩
      neighborjoining

   #awk '{printf("%s",$0)}'  <cluster_$cluster_no.tree >↩
      cluster_$cluster_no_2.tree

   #overwrite the original tree file
   #mv cluster_$cluster_no_2.tree cluster_$cluster_no.tree

   #get a random sequence from this cluster
   #This solution is based on reservoir sampling. It is optimal in ↩
      terms of both time and space complexity and works with an ↩
      arbitrarily large input file.
   #https://www.biostars.org/p/18831/
   bioawk -c fastx -v k=1 '{y=x++<k?x-1:int(rand()*x);if(y<k)a[y]=">"↩
      $name"\n"$seq}END{for(z in a)print a[z]}' $directory/↩
      cluster_$cluster_no/cluster_$cluster_no.fasta >↩
      cluster_$cluster_no.fa

   Rscript ~/Code/R/phylosim_simulation.R cluster_$cluster_no.fa ↩
      $directory/cluster_$cluster_no/cluster_$cluster_no.best.dnd ↩
      cluster_$cluster_no.s.fa

  ruby ~/Code/Ruby/blocks/lib/polymorphic_sites.rb cluster_$cluster_no.↩
     s.fa >cluster_$cluster_no.snps.txt

 done <$clusterfile

  #find . -type f -name *.snps.txt | xargs cat | sed -e 's/^[ \t]*//' >↩
     all.snps
```

A.13 A script to extract regions from a DBLα sequence

Listing A.20: A script to extract sequence regions of interest

```
#!/usr/bin/env ruby
require 'bio'
require 'commander/import'

program :version, '0.0.1'
program :description, 'This program extracts sequence regions given a ↩
   position and a window length'
program :author, 'George Githinji email:ggithinji@kemri-wellcome.org'
```

```
command :regions do |r|
  r.syntax = 'extract region [option]'
  r.summary = 'find and extract a region of sequence given a variant ↩
     position'
  r.description  = 'Extracts overlapping sequence regions from given a ↩
     MSA variant position and a window length.'

  r.example 'extract region', 'extract region --infile FILE --pos ↩
     INTEGER --win INTEGER'

  r.option '--infile FILE', 'Input file'
  r.option '--pos INTEGER', Integer, 'A variant position'
  r.option '--win INTEGER', Integer, 'Window length'

  r.action do |args,options|
    infile  = options.infile
    pos     = options.pos
    win     = options.win

    # Read in a FASTA-formatted multiple sequence alignment
    fasta_seqs = Bio::Alignment::MultiFastaFormat.new(File.open(infile)↩
       .read)

    fasta_seqs.entries.each do |e|
      for i in 0..win
        $stdout.puts e.seq[(pos - 1) - i, win] unless pos - i <= 0
      end
    end
  end
end
```

A.14 A bash program to find variation context

Listing A.21: A simple pipeline to find the frequency of isolates sharing sequence regions around a variant position

```
#!/bin/bash
db="/Users/george/454_data/blast-db/454"
classified="/Users/george/454_data/blast-db/454.classified.txt"

cluster=$1
position=$2
window=$3

clsname=$(basename "$cluster")
clsext="${clsname##*.}"
clsname="${clsname%.*}"

#remove suffix from alignment
~/Code/Ruby/Blocks/lib/extract regions --infile $cluster --pos ↩
   $position --win $window | sort | \
  uniq | awk -v cl=$clsname.$position '{print ">"cl"." FNR "\n" $0}' \
  >"$clsname.$position.$window.fa"
```

```
blastn -query $clsname.$position.$window.fa -db $db -outfmt 6 |\
  awk -v win=$window '{if($3 == 100 && $4 == win) print $2}' | \
  sort -n | uniq >hits.454.$clsname.$position.$window.txt

grep -wFf hits.454.$clsname.$position.$window.txt $classified |\
  awk '{print $1,$7,$2}' | sort -k2 | sed 's/\..*$//g' | sort -n | uniq↩
      -c
```

A.15 A Python programm for submitting epitope prediction requests to the Predivac server

Listing A.22: Automated prediction of T-cell epitopes using the Predivac online server.

```
#!/usr/bin/env python

from bs4 import BeautifulSoup
from urllib2 import urlopen
from ClientForm import ParseResponse
from optparse import OptionParser
from Bio import SeqIO
import sys
import time

#main class
#↩
    ----------------------------------------------------------------------------

class Predivac(object):
    """
    Scrapping the predivac web database using beautifuloup and python
    """

    def __init__(self,Sequence,Allele):
        """
        Initialize class parameters
        """
        self.Sequence = Sequence
        self.Allele = Allele
        self.PredivacLink = 'http://predivac.biosci.uq.edu.au/cgi-bin/↩
            binding.py'

    def FillForm(self):
        """
        Fill the webform
        """
        ProtString = "\n".join(self.Sequence)
        webpage = urlopen(self.PredivacLink)
        forms = ParseResponse(webpage)
        form = forms[0]
        form['subtype'] = ['protein']
        form['sequence'] = ProtString
        form['allele'] = [self.Allele]
        form['threshold'] = ['3']
        response = urlopen(form.click())
        soup = BeautifulSoup(response)
        return soup

    def ParseResults(self):
        """
        Parse the prediction table
        """
        soup = Predivac.FillForm(self)
        table = soup.find("table", id = "prediction-table")
        predictions = []
        try:
            for row in table.findAll('tr'):
                tds = row('td')
                if len(tds) > 1:
                    id = tds[0].text
                    nanomer = tds[1].text
                    score = tds[2].text
                    start = tds[3].text
                    end = tds[4].text
                    full_entry = id + ":" + nanomer + ":" + score + ":"↩
                        + start + ":" + end
                    predictions.append(full_entry)
        except AttributeError:
            return

        return ",".join(predictions)

#splitting the protein sequence to accomodate predivac
#↩
    ------------------------------------------------------------------------------

def chunks(s, n):
    """Produce `n`-character chunks from `s`."""
    for start in range(0, len(s), n):
        yield s[start:start+n]

#Parse fasta entries
#↩
    ------------------------------------------------------------------------------

def ParseFasta(infile):

    handle = open(infile, "rU")
    FastaResult = []
    for record in SeqIO.parse(handle, "fasta"):
        Id = record.id
        Sequence = record.seq
        Entry = (Id,Sequence)
        FastaResult.append(Entry)
    return FastaResult

#Parse allele list
#↩
```

```
    ----------------------------------------------------------------------------

def ParseAllele(infile):
    Alleles = []
    for line in open(infile, "rU"):
            Alleles.append(line.strip())
    return Alleles

#Setting commandline options
#←
    ----------------------------------------------------------------------------

def commandline_options():

    parser = OptionParser(usage="usage: %prog [options]")
    parser.add_option("-p", "--proteins",
                      type="string",
                      dest="protein_file",
                      help="Fasta file with proteins")
    parser.add_option("-a", "--allele",
                      type="string",
                      dest="allele_file",
                      help="File with list of alleles")
    parser.add_option("-d","--delay",
                      type="int",
                      dest="time_delay",
                      default=3,
                      help="Minimum time delay before hitting server (←
                         Default 3s)")
    parser.add_option("-o", "--output",
                      action="store",
                      type="string",
                      dest="result_file",
                      default="results.txt",
                      help="The final output file")

    (options, args) = parser.parse_args()
    options_args_parser = [options,args,parser]

    return options_args_parser


#main
#←
    ----------------------------------------------------------------------------

def main():
    options, args, parser = commandline_options()
    ProteinFile = options.protein_file
    AlleleFile = options.allele_file
    TimeDelay = options.time_delay
    OutputFile = options.result_file
    if ProteinFile is None or AlleleFile is None:
        sys.stderr.write("\nError: A mandatory option is missing!\n")
        parser.print_help()
        sys.exit(-1)
    else:
        FastaFile = ParseFasta(ProteinFile)
        Alleles = ParseAllele(AlleleFile)
        FOut = open(OutputFile,'w')
        for Entry in FastaFile:
            for Allele in Alleles:
                ProtId,ProtSeq = Entry
                ProtId_M = '>' + ProtId
                ProtSeq = str(ProtSeq)
                SeqPieces = [ProtId_M] #divind the sequence into pieces
                for chunk in chunks(ProtSeq, 50):
                    SeqPieces.append(chunk)

                sys.stderr.write("predicting %s with %s\n" % (ProtId,↩
                   Allele))
                soup = Predivac(SeqPieces,Allele)
                soup.FillForm()
                Results = soup.ParseResults()
                FOut.write("%s\t%s\t%s\n" %(ProtId,Allele,Results))
                time.sleep(TimeDelay)
        sys.stderr.write("\nPrediction complete, results written to to ↩
           %s\n" % (OutputFile))


#Run it
#↩
   ----------------------------------------------------------------------------

if __name__ == '__main__':
    main()
```

A.16 A program to predict netMHCII epitopes using a parallel approach

Listing A.23: A script to predict MHC-class II epitopes

```
#!/bin/sh
## This script should be improved with GNU parallel! :)

prediction_method="NetMHCII-pan-3"

fasta_file=$1
allele_file=$2
peptide_len=$3
#output_file=$4

echo "Predicting class II T-cell epitopes" #$file_no" #>> $LOG_FILE

function wait_run_in_parallel()
{
  local number_to_run_concurrently=$1
  if [ `jobs -np | wc -l` -gt $number_to_run_concurrently ]; then
    wait `jobs -np | head -1` # wait for the oldest one to finish
  fi
}

#run the prediction and sanitize the output to csv file
function predict(){
  local allele=$1
  netmhcpan3 -f $fasta_file -a $allele -length $peptide_len |
  sed '/\#/d' | sed '/---/d' |
  sed '/Pos/d' | sed '/Number/d' |
  sed 's/^ *//' | sed 's/[ \t]*$//' |
  sed 's/<=WB//' | sed 's/<=SB//' |
  tr -s '[:blank:]' ',' |
  sed '/^$/d'| sed 's/,$//'
}

# function to run MHCII program for each alleles and send the jobs to ↩
    the background
run_prediction () {
  #Read the allele file
  while read mhc_allele
  do
    predict $mhc_allele >>"$mhc_allele.pred.temp" &
    wait_run_in_parallel 10 #run 10 jobs at a time and wait to execute.
  done <$allele_file
}

#execute the jobs
run_prediction
#wait for all background jobs
wait

#concatenate the files
cat *.pred.temp >>all.hla.preds

#cleanup!
rm *.pred.temp
```

A.17 A Ruby program to print block-sharing networks in csv format

Listing A.24: A script to print block-sharing networks

```
#!/usr/bin/env ruby

require 'bio-dbla-classifier'
require 'commander/import'

program :version, '0.0.1'
program :description, 'Generates a network from DBL-alpha amino acid ↩
    sequence tags'
program :author, 'George Githinji email: biorelated@gmail.com'

command :create do |c|
#A structure to hold the name and pspb types
class Pspb < Struct.new(:id,:name,:pspb1,:pspb2,:pspb3,:pspb4)
    def all_pspbs
        [pspb1,pspb2,pspb3,pspb4]
    end
end

c.syntax = 'network create [options]'
c.summary = 'creates a network given a list of amino acid sequences↩
    '
c.description = 'This command generates a tab-separated list of 2 ↩
   columns where each entry in each column is a node.'
c.example 'network', 'create --infile FILENAME --blocksize INTEGER ↩
    --outfile FILENAME'
c.option '--infile FILE', 'Specify an input file(amino acid only)'
c.option '--blocksize INTEGER', Integer, 'specify the length of the↩
    pspb block. Default is 10'

c.action do |args, options|
    options.default :blocksize => 10
    seq_file    = options.infile
    blocksize = options.blocksize

    pspbs = Bio::FlatFile.open(seq_file).each_with_index.map do |↩
       entry,index|
        pspb1 = Bio::Sequence::AA.new(entry.seq).pspb1(0,blocksize)
        pspb2 = Bio::Sequence::AA.new(entry.seq).pspb2(0,blocksize)
        pspb3 = Bio::Sequence::AA.new(entry.seq).pspb3(0,blocksize)
        pspb4 = Bio::Sequence::AA.new(entry.seq).pspb4(0,blocksize)

        Pspb.new(index + 1,entry.definition,pspb1,pspb2,pspb3,pspb4↩
           )
    end

    puts "node1\tnode2"
    pspbs.combination(2).each do |i|
        puts "#{i[0].name}\t#{i[1].name}" if (i[0].all_pspbs & i↩
           [1].all_pspbs).size > 0
    end
end

command :attribute do |a|
    a.syntax = 'network attribute [options]'
    a.summary = 'generate network attributes file'
    a.description = 'This command generates a network attributes ↩
       file for loading into cystoscape'

    a.example 'network', 'attribute --infile FILENAME'
    a.option '-f','--infile FILE', 'specify the sequence input file↩
       '

    #a structure to hold the cyspolv classification and bs sharing ↩
       groups
    class Attr < Struct.new(:id,:name,:cyspolv,:bsgroup);end
```

```
        a.action do |args, options|
            seqfile = options.infile

            attrs = Bio::FlatFile.open(seqfile).each_with_index.map do ↩
               |entry,index|
                cyspolv = Bio::Sequence::AA.new(entry.seq).↩
                   cyspolv_group
                bsgroup = Bio::Sequence::AA.new(entry.seq).bs_group
                Attr.new(index + 1,entry.definition,cyspolv,bsgroup)
            end

            puts "gene\tcyspolvgroup\tbsgroup"
            attrs.each do |attri|
                puts "#{attri.name}\t#{attri.cyspolv}\t#{attri.bsgroup}↩
                   "
            end
        end
    end
end
```

A.18 A Ruby program to print homology blocks and hypervarible blocks from sequence tags

Listing A.25: A program to print the homology blocks D and H and hypervarible blocks HP1 and HP2

```
#!/usr/bin/env ruby
require 'bio'
require 'bio-alignment'
require 'bio-alignment/bioruby'
include  Bio::BioAlignment

file = ARGV[0]

include Bio::BioAlignment

aln = Alignment.new
Bio::FlatFile.auto(file).map.each_entry do |entry|
  aln << Sequence.new(entry.definition,entry.seq)
end


seq_len = aln[0].to_s.length

block_H_start = seq_len - 39

ww_motif_pos = aln[0].to_s.rindex(/TGGTGG/)

hp2_start = ww_motif_pos + 42

hp1_start = 42
hp1_stop = ww_motif_pos - 21
```

```
#block D
puts aln.columns[0..42].map{|c| c.to_a}.transpose.each_with_index.map{|↩
    s,i| ">seq_#{i+1}\n#{s.join}"}

#before WW hypervariable region
#puts aln.columns[hp1_start..hp1_stop].map{|c| c.to_a}.transpose.↩
    each_with_index.map{|s,i| ">seq_#{i+1}\n#{s.join}"}

#after WW hypervarible block
#puts aln.columns[hp2_start..block_H_start].map{|c| c.to_a}.transpose.↩
    each_with_index.map{|s,i| ">seq_#{i+1}\n#{s.join}"}

#block_H
#puts aln.columns[block_H_start..seq_len].map{|c| c.to_a}.transpose.↩
    each_with_index.map{|s,i| ">seq_#{i+1}\n#{s.join}"}
```

A.19 A Ruby program to generate random nucleotide substitutions

Listing A.26: A Ruby implementation of a program to generate random substitutions

```
#!/usr/bin env ruby

require 'optparse'
require 'ostruct'
require 'bio'
require 'securerandom'
require 'pickup'

options = OpenStruct.new

OptionParser.new do |opts|
  opts.on('-h', 'shows this help screen') do
    puts opts
    exit
  end

  opts.on('-i', '--infile FILE','input fasta file') do |infile|
    options.infile = infile
  end

  opts.on('-r','--error_rate Error', Float, 'Error rate per nucleotide'↩
      ) do |error_rate|
    options.error_rate = error_rate
  end

  opts.on('-o','--outfile FILE','Output file') do |outfile|
    options.outfile = outfile
  end

  opts.on('-w' '--weights Weight',Float,'weighting levels') do |weights↩
      |
    options.weights = weights
  end
```

```
end.parse!


class Bio::Sequence
  attr_accessor :substitutions
  attr_accessor :weights

  NUCLEOTIDES = ['A','C','T','G']

  # The number of random positions to change
  def positions_to_sub
    (1..self.length).to_a.sample(substitutions)
  end

  # Weighted random nucleotides
  def weighted_random_nucleotides
    weights = []
    weighted_nt = Pickup.new(weights)
    weighted_nt.pick(subsitutions)
  end

  # Returns an array of x random nucleotides where x is the number of ↩
      given substitutions
  def random_nucleotides
    substitutions.times.map{NUCLEOTIDES[rand(NUCLEOTIDES.length)]}
  end

  # Rate based mutations
  def rate_based_mutate
    #substitute nucleotide given positions with random nucleotides
    positions_to_sub.zip(random_nucleotides).map do |position,↩
        nucleotide|
      if self[position] == nucleotide
        nucleotide = (['A','C','G','T'] - [nucleotide]).sample(1)
        self[position] = nucleotide.join
      else
        self[position] = nucleotide.downcase
      end
    end
  end
end


infile = options.infile
error_rate = options.error_rate
weights = options.weights

def biosequences(file)
  Bio::FlatFile.auto(file).map do |entry|
    Bio::Sequence.new(entry.seq.upcase)
  end
end

def calculate_mismatches(error_rate,seq_length)
```

```
  (error_rate * seq_length).round
end

def mutate_seqs(bioseq_objects,error_rate)
  bioseq_objects.each do |bioseq|
    seq_len = bioseq.length
    bioseq.substitutions = calculate_mismatches(error_rate,seq_len)
    bioseq.rate_based_mutate
    $stdout.puts ">#{SecureRandom.hex(5)}\n#{bioseq.seq}"
  end
end

begin
sequences = biosequences(infile)
mutate_seqs(sequences,error_rate)

rescue TypeError => type_error
  $stderr.puts type_error.message
rescue Bio::FlatFile::UnknownDataFormatError =>format_error
  $stderr.puts format_error.message << "provide a fasta format"
end
```

A.20 A Ruby program to parse and work with DBLα sequences from CAF formatted files

Listing A.27: assemble tags

```
require 'bio'
['caf', 'cafrecord', 'phrap'].each do |name|
  require File.join(File.expand_path(File.dirname(__FILE__)),"#{name}")
end

puts "Processing ...."
dir_path = "#{ENV['HOME']}/Batch1_files_from_thomas/batch1_cafs"

Dir.chdir(dir_path) do
  p = Caf::Parser.new('batch1.caf')

  puts "Total caf records = #{p.caf_records.size}"

  #rejected_reads = p.rejected_seqs
  #p.to_file('rejected_reads_2.fasta',rejected_reads)

  puts 'removing vector sequences'

  without_vector_seqs = p.no_vector_seqs
  p.to_file('batch1_without_vector_seqs_28_07_2011.fasta',p.↩
      no_vector_seqs)

end

puts "Finished successfuly"
```

Listing A.28: caf class

```
module Caf
 class Parser

  def initialize(file)
    @file = file
  end

  def caf_records
    $/ = "\n\n"
    records = []
      File.open(@file) do |f|
        f.each_line do |line|
          read = CafRecord.new
          case line
            when /^DNA/
              records << CafRecord.new
              records.last.dna_data = line.strip
            when /^BaseQuality/
              records.last.quality_data = line.strip
            when /^Sequence/
              records.last.metadata = line.strip
            else
              puts "Unrecognized line: #{line}"
          end
        end
      end
      records
    end

    def sequence_list
      caf_records.map {|read| read.sequence}
    end

    def name_list
     caf_records.map {|read| read.name}
    end

    def no_vector_seqs
      caf_records.map {|read| read.trimmed_fasta if read.↩
         passes_vector_filter?}.compact       #remove nils
    end

    def rejected_seqs
      caf_records.map {|read| read.raw_fasta unless read.↩
         passes_vector_filter?}.compact       #remove nils
    end

    def no_primer_seqs
      caf_records.map {|read|read.deprimed_fasta if read.passes_filters↩
         ?}.compact          #remove nils
    end

    def no_primer_seqs_quals
```

```
      caf_records.map {|read|read.deprimed_base_qual_fasta if read.↩
         passes_filters?}.compact       #remove nils
    end

    def raw_caf_reads
      caf_records.map {|read| read.raw_fasta}.compact
    end

    #write an output to a physical file
    def to_file(path,contents)
      File.open("#{path}",'w') do |f|
        f.puts contents
      end
    end

  end #class
end #module
```

Listing A.29: cafrecord class

```
class CafRecord
  attr_accessor :dna_data,:quality_data,:metadata

  #return the name
  def name
   #dna_data = dna_data.to_s
   # possible bug for different records names
   # for georges sequences
   #name_reg = create_regexp("(VARPB.*)$")
   name_reg = create_regexp("(VAR[PB|BP].*)$") #allow mistyped read ↩
      names VARPB/VARBP
   # name_reg = create_regexp("(VARPB\\d\\d..\\d.\\d\\d\\..\\d.*)")
   dna_data.to_s.scan(name_reg).join
  end

  #return the raw read dna sequence
  def sequence
   # name_reg = create_regexp("(^DNA\\s:\\sVARPB\\d\\d..\\d.\\d\\d↩
      \\..\\d.*)")
   name_reg = create_regexp("(^DNA : VARPB.*)$")
   #dna_data.to_s[25,dna_data.to_s.size].delete("\n").strip rescue "" #↩
      if dna_data
   dna_data.to_s.gsub(name_reg,'').delete("\n").strip
  end

  #string of base quality scores
  def quality_scores
    #quality_data = quality_data.to_s
    score_regexp = create_regexp("(^\\d.*)")
    quality_data.to_s.scan(score_regexp).to_s
  end

  #get the vector positions xxxxxxxacatatatataxxxxxxx
  def vector_positions
  seqvector = create_regexp("^Seq_vec\\s+SVEC\\s+(\\d+\\s+\\d+)")
  #get the vector positions
  #svec1_start, svec1_stop, svec2_start, svec2_stop
  get_vector_positions(seqvector,metadata.to_s)
  #return svec1_start, svec1_stop, svec2_start, svec2_stop
end


def clip_positions
  clipping = create_regexp("^Clipping\\s+QUAL\\s+(\\d+\\s+\\d+)")
  get_clip_positions(clipping, metadata.to_s)
  #return clip_qual_start, clip_qual_stop
end

#returns a sequence with no vector regions
def trimmed_seq
  begin
  highest_start = vector_clips[0]
  lowest_end    = vector_clips[1]

  clip_len = lowest_end.to_i - highest_start.to_i
  clip_len
  #len_range = (300..500)
  #if len_range.include? clip_len
    trimmed_seq = sequence[highest_start, clip_len]
  #else
    #trimmed_seq = ""
  #end
    #puts "clip_len:#{clip_len}"
    #puts "trimmed_seq_size:#{trimmed_seq.to_s.size}"
 return trimmed_seq if trimmed_seq.to_s.size > 300
  rescue
  end
end

#returns base qualities devoid of vector quality scores.
def trimmed_base_qualities
  highest_start = vector_clips[0]
  lowest_end = vector_clips[1]

  clip_len = lowest_end.to_i - highest_start.to_i
  clip_len
  len_range = (300..500)
  if len_range.include? clip_len
    trimmed_base_q = quality_scores.split(/ /).slice(highest_start, ↩
        clip_len).join(" ")
  else
    trimmed_base_q = ""
  end
  trimmed_base_q
end

#returns highest start and lowest ends for each read
def vector_clips
```

```
  #first step: set highest_start to svec1_stop and lowest_end to 10 ↩
     000
   highest_start = vector_positions[1]
   lowest_end = 10_000

  #second step
  #check whether svec2_start is less than 10 000; if true set ↩
     lowest_end to svect2_start
   unless vector_positions[2] == 0
     lowest_end = vector_positions[2]  if vector_positions[2] < ↩
        lowest_end
   else
     lowest_end = clip_positions[1]  #if we only have 1 svec entry,
   end
    highest_start = clip_positions[0] if highest_start < ↩
       clip_positions[0]
    lowest_end    = clip_positions[1] if lowest_end > clip_positions↩
       [1]
   return highest_start, lowest_end
end

# a good read i.e size > 1 after removing the vector sequences
def passes_vector_filter?
  true if trimmed_seq.to_s.size > 1     # can put a minimun of what ↩
     length of read to accept
end

#return the read's region without primers
def deprimed
  seq = Bio::Sequence::NA.new(trimmed_seq)
  start_pos = primer_boundaries[0]
  end_pos = primer_boundaries[1]

  if start_pos.nil? or end_pos.nil?
    start_pos, end_pos = 0,0
  end

  clip_len  = end_pos - start_pos
  seq.to_s[start_pos, clip_len + 1].upcase
end

#return the base qualities of the read without primers
 def deprimed_base_qual
   start_pos = primer_boundaries[0]
   end_pos   = primer_boundaries[1]

    if start_pos.nil? or end_pos.nil?
      start_pos, end_pos = 0,0
    end

    clip_len  = end_pos - start_pos
    trimmed_base_qualities.split(/ /).slice(start_pos,clip_len + 1).↩
       join(" ")
 end
```

```
   def deprimed_base_qual_fasta
       ">#{name}\n#{deprimed_base_qual}"
   end

  #returns the start and end positions of unmasked region
  def primer_boundaries
      seq = Bio::Sequence::NA.new(trimmed_seq)
      primer_regs.each do |reg|
        seq.gsub!(reg) { |x| "X" * x.length }
      end

      # Replace all 5' bases before "X" with "X".
#        seq.sub!(/\A[^X]+X/) { |x| "X" * x.length }

#       Replace all 3' bases after "X" with "X".
        #seq.sub!(/X[^X]+\z/) { |x| "X" * x.length }

#        last_match_size = seq.scan(/[^(X+)]*$/)[0].size
       #Replace any left end base on the 3' end with a '' only if ↩
          XXXcgctctXXXtc and not XXXcgctctgctc
        seq.sub!(/[^(X+)]*$/,'')  if seq.scan(/[^(X+)]*$/)[0].size < 20

      # Get the start and end positions of the unmasked region.
#        start_pos = seq.index(/[^X]/) #start position
#        end_pos   = seq.rindex(/[^X]/) #end position

      return seq.index(/[^X]/),seq.rindex(/[^X]/)
  end

  #remove any read where stop codons >= 6 or or read length > 500
  def passes_sc_filter?
    #translate read sequence in all 6 frames
    passed = []
    6.times do |frame|
      stop_codons = Bio::Sequence::NA.new(deprimed).translate(frame + ↩
         1).scan(/\*/).size
      passed << deprimed unless stop_codons >= 6  || deprimed.size > ↩
         500 #have a look at this line!!
    end
   true if passed.size > 0
  end

  def passes_filters?
    passes_vector_filter? && passes_sc_filter?
  end

  def raw_fasta
   fasta(sequence)
  end

  def trimmed_fasta
   fasta(trimmed_seq)
  end
```

```
 def deprimed_fasta
  fasta(deprimed)
 end

 def six_frame_traslate_deprimed
  six_frame_translate(deprimed)
 end

private
 #creates a regular expression given a string
 def create_regexp(expression,opts={})
  options = {:case_sensitive=>true}.merge!(opts)    #make them case ↩
     insentive by default
  Regexp.new("#{expression}",options)
 end

 def get_vector_positions(seqvector, value)
  vector_positions =  metadata.to_s.scan(seqvector)
  svec1_startq, svec1_stopq = vector_positions[0].to_s.split(/ /)
  svec2_startq, svec2_stopq = vector_positions[1].to_s.split(/ /)
  return svec1_startq.to_i, svec1_stopq.to_i, svec2_startq.to_i, ↩
     svec2_stopq.to_i
 end

  def get_clip_positions(clipping, value)
  clipping_positions = metadata.to_s.scan(clipping)
  clip_qual_startq, clip_qual_stopq = clipping_positions[0].to_s.↩
     split(/ /)
  return clip_qual_startq.to_i, clip_qual_stopq.to_i
  end

 def get_sequence(sequence,start,clip_len)
  sequence[start + 1, clip_len] if clip_len >= 300
 end

 def get_qualities(qualities,clip_len,start)
   #split the quality data to get an array
   q = qualities.to_s.split(/ /)
   q.slice(start,clip_len).join(" ") if clip_len > 300
 end

 #return the length of the sequence
  def size(seq)
  seq.size
  end

 #return a biosequence object
  def biosequence(seq)
    Bio::Sequence.new(seq)
  end

  #create a fasta format for the read
 def fasta(seq)
```

```
    # puts name.class
    biosequence(seq).output(:fasta,:header =>name.to_s,:width=>500)
  end

  #returns the g + c content for the sequence
  def gc_contect(seq)
   biosequence(seq).gc_content
  end

  #returns a list of six frame translations
  def six_frame_translate(seq)
    content = []
    6.times do |i|
       content << Bio::Sequence::NA.new(seq).translate(i+1)
    end
    content
  end

  #creates a collection of primer sequences to look for
  def primer_regs
   primers = ['G*CACG[A|C]AGTTT[C|T]GC','G*CCCATTC[G|C]TCGAACCA','GC[G|↩
      A]AAACT[T|G]CGTGC','TGGTTCGA[C|G]GAATGGGC']
   primers.collect! { |primer| create_regexp(primer) }
  end

end
```

Listing A.30: Phrap wrapper

```
module Phrap
  class Runner
    #run the phrap program
    def bull_phrap(fasta_file)
    phrap_out = %x(phrap_99 -retain_duplicates -ace -preassemble -↩
       group_delim . -minscore 200 #{fasta_file})
    puts phrap_out
    end

    #run this for reads from the same isolate
    def bloqvist_phrap(fasta_file)
       phrap_out = %x(phrap_99 -retain_duplicates -ace -↩
          repeat_stringency 0.9 -minmatch 20 #{fasta_file})
    end

  end

end
```

Listing A.31: Caf to Phrap

```
require 'bio'
class Caf2phrap
  #install caf2phrap executable
  def run(caf_file,fasta_name)
  caf2phrap_out = %x(caf2phrap -caf #{caf_file} -fasta #{fasta_name})
end

#creates a regular expression given a string
def create_regexp(expression,opts={})
  options = {:case_sensitive=>true}.merge!(opts)    #make them case ↩
      insentive by default
  Regexp.new("#{expression}",options)
end

def primer_regs
  ['G*CACG[A|C]AGTTT[C|T]GC','G*CCCATTC[G|C]TCGAACCA','GC[G|A]AAACT[T↩
      |G]CGTGC','TGGTTCGA[C|G]GAATGGGC'].map! { |primer| create_regexp↩
      (primer) }
end

def trim(filename)
  seqs = []
  Bio::FlatFile.auto(filename){ |f| f.map {|entry| entry} }.each do |↩
      entry|
    #puts entry.seq
    start_pos = entry.seq.index(/[^x]/).to_i
    stop_pos = entry.seq.rindex(/[^x]/).to_i

    clipped = stop_pos - start_pos
    seq = entry.seq[start_pos, clipped + 1]
    seq_len_range = (300..500)
    if seq_len_range.include? seq.size
      #remove the primers
      primer_regs.each do |reg|
        seq.gsub!(reg) { |x| "x" * x.length }
      end

      seq.sub!(/[^(x+)]*$/,'')  if seq.scan(/[^(x+)]*$/)[0].size < 20

      primer_start_pos = seq.index(/[^x]/)
      primer_stop_pos =  seq.rindex(/[^x]/)
      tag = primer_stop_pos - primer_start_pos
      seqs << Bio::Sequence.new(seq[primer_start_pos, tag + 1].upcase↩
          ).output(:fasta,:header => entry.definition,:width=>600)
    else
      puts "#{entry.definition} does not meet the upper(500) or lower↩
          (300) nt thresholds"
    end
  end
  seqs
end

#write data to file
def to_file(path,contents)
  File.open("#{path}",'w') do |f|
    f.puts contents
  end
end
```

```
end

input_file = "#{ENV['HOME']}/Batch2_files_from_thomas/batch2_cafs/↩
   batch2_caftools.fasta"
output_file = "#{ENV['HOME']}/Batch2_files_from_thomas/batch2_cafs/↩
   batch2_caftools_no_primer.fasta" #caftools_no_primers.fasta

puts 'initializing'

caf2 = Caf2phrap.new

puts 'trimming reads'

reads = caf2.trim(input_file)

puts 'writing to file'
caf2.to_file(output_file,reads)

puts 'Done!'
```

Appendix B
Miscellaneous

This chapter includes work that supports specific sections of the main text. Some of this analysis was too short or inconclusive on its own and may require additional data or experiments that could not be conducted given the time and budget restrictions.

B.1 Regions of diversity in the DBLα sequence tags

Figure B.1 shows the codon diversity (upper panel) and pairwise distance (lower panel) for a group of 100 randomly selected DBLα sequence tags from Kilifi. This figure shows that relative sequence conservation is confined within specific regions of homology that are interspersed by very diverse regions. The codon usage is also not uniform even in conserved amino acid columns. The mean pairwise identity along the alignment positions is shown by the black line in the lower panel.

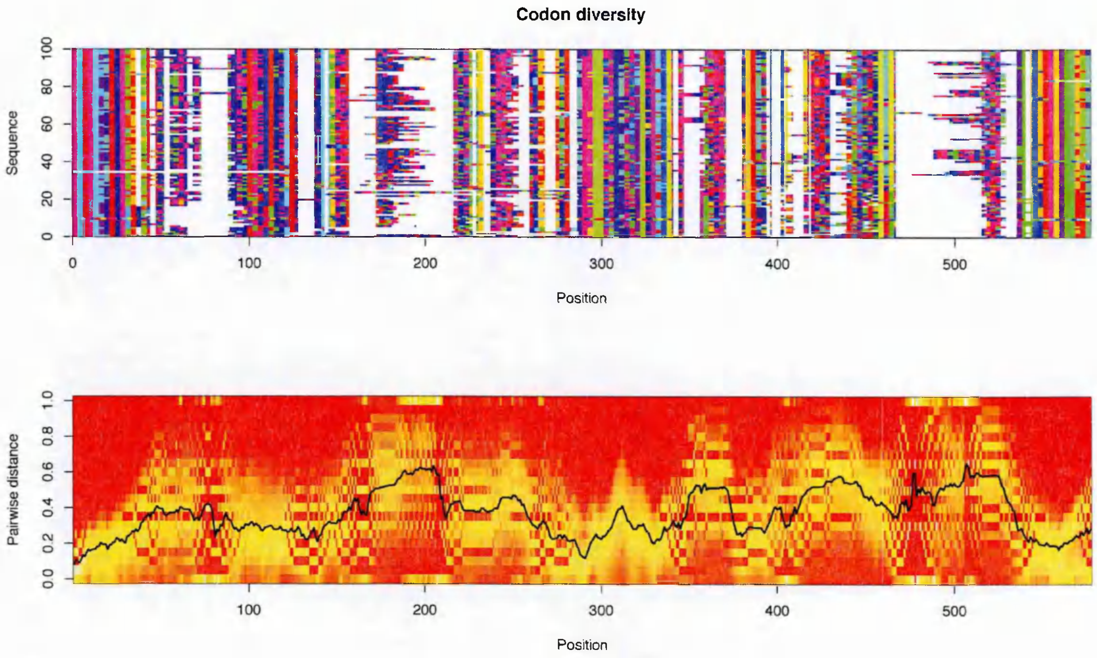


Figure B.1: *A codon diversity plot (upper panel) and a pairwise distance plot (lower panel) summarizing the extent of codon and nucleotide diversity in a group of 100 randomly selected DBLα sequence tags from Kilifi. Each codon is shown by a different colour. The pairwise distances (y-axis) were calculated using a 10 nucleotide sliding-window along the length of the alignment (x-axis). The black line shows the mean pairwise distance for each column of the alignment.*

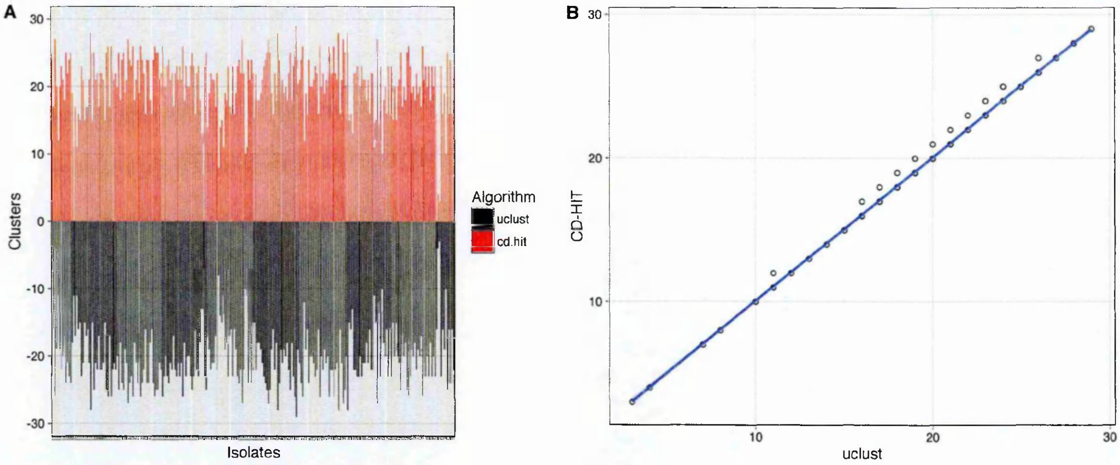


Figure B.2: *Distribution of clusters and non redundant sequences generated with CD-HIT and UCLUST version 5 algorithms. A) The x-axis represents the isolates and the y-axis shows the frequency of representative sequences for each isolate. The bar plots show that at 100% similarity the two algorithms create the same number of clusters per isolate. Diverse isolates tend to contain higher number of clusters. B) A scatter plot showing the linear relationship between the two clustering algorithms at 100% similarity. CD-HIT produced more clusters for some isolates relative to uclust.*

This work uses the Usearch/Vsearch approach in clustering given that the algorithm meets a definition of sequence identity that is applicable to this work. Furthermore it is straightforward to understand and has several advantages over CD-hit as explained in the previous sections.

B.2 Length polymorphism in DBLα sequence tags

DBLα sequences differ in length. Figure B.3 shows a histogram and a density plot of the mean distribution of sequence lengths from 6,200 DBLα sequence tags collected from Kilifi during the 2003 -2007 and 2007 - 2010 periods.

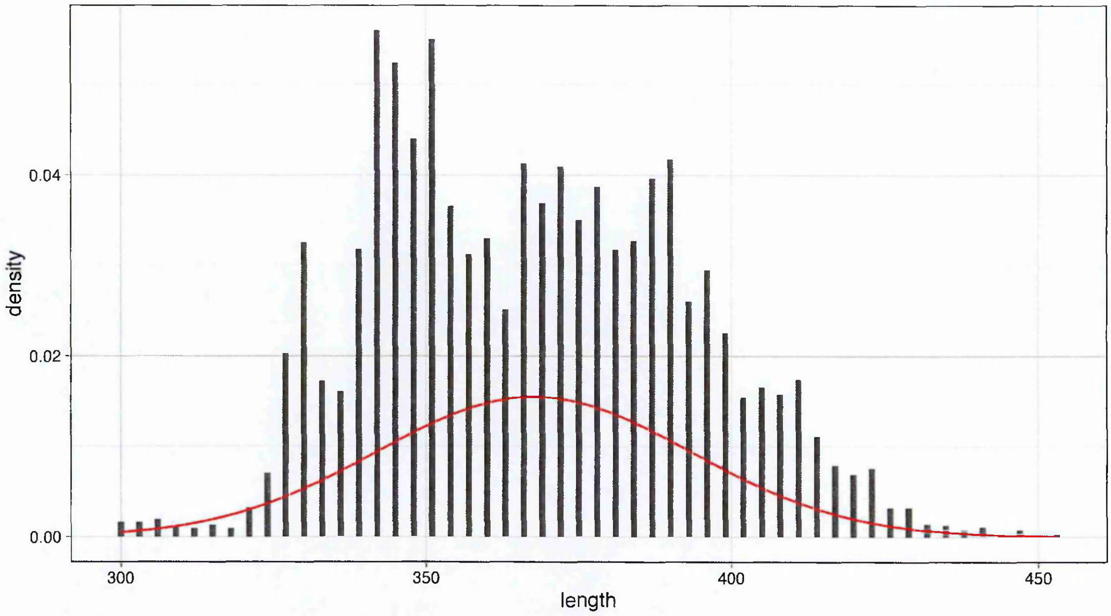


Figure B.3: *A distribution of length in 6,200 DBLα sequences collected between 2003-2007 and 2007-2010 time periods in Kilifi. The red plot shows the mean density in length of the tag sequences. Although the upper limit for DBLα sequences is defined at 500 nucleotides majority of sequences did not surpass 450nt in length. The mean length was normally distributed.*

B.3 Dinucleotide frequency

Because dinucleotide frequency is known to affect codon bias, which in turn can affect the pattern of substitutions, a comparison of the expected and the actual frequency of 16 dinucleotides was assessed. If a particular dinucleotide was over-represented, it meant that it occurred many more times than what was expected by chance and if was under-represented, it meant it occurred fewer times than expected.

The ρ (rho) statistic was calculated for the observed and expected dinucleotides. Rho is a measure of over or under-represented dinucleotides. If a DNA sequence had a frequency f_x of a 1-nucleotide DNA word x, and a frequency f_y of a 1-nucleotide DNA word y, then the frequency of the dinucleotide xy is expected to be the product of f_x and f_y.

Rho (ρ) is defined by equation B.1

$$\rho(xy) = f_{xy}/(f_x * f_y) \tag{B.1}$$

where f_{xy} and f_x are the frequencies of dinucleotide *xy* and *x* in a sequence.

From equation B.1, dinucleotide frequecies in a sequence are expected to be equal to the products of the frequencies of the two nucleotides that compose them and therefore ρ should be equal to 1 if they are not under or over-represented.

Figure B.4 shows the dinucleotide ρ statistic from the 16 dinucleotide pairs.

Table B.1: *Summary of dinucleotide frequency*

Statistic	N	Mean	St. Dev.	Min	Max
AA	5,280	57.077	7.492	21	85
AC	5,280	18.979	4.304	6	33
AG	5,280	26.559	4.437	6	43
AT	5,280	39.386	4.218	23	55
CA	5,280	20.193	4.001	8	34
CC	5,280	8.648	2.708	1	28
CG	5,280	12.716	5.316	1	28
CT	5,280	10.424	2.451	3	27
GA	5,280	35.068	5.971	12	53
GC	5,280	11.930	4.126	3	28
GG	5,280	17.755	3.669	5	34
GT	5,280	16.245	3.286	6	28
TA	5,280	29.842	4.280	14	47
TC	5,280	12.920	2.645	5	37
TG	5,280	23.038	3.852	11	36
TT	5,280	27.093	4.176	14	42

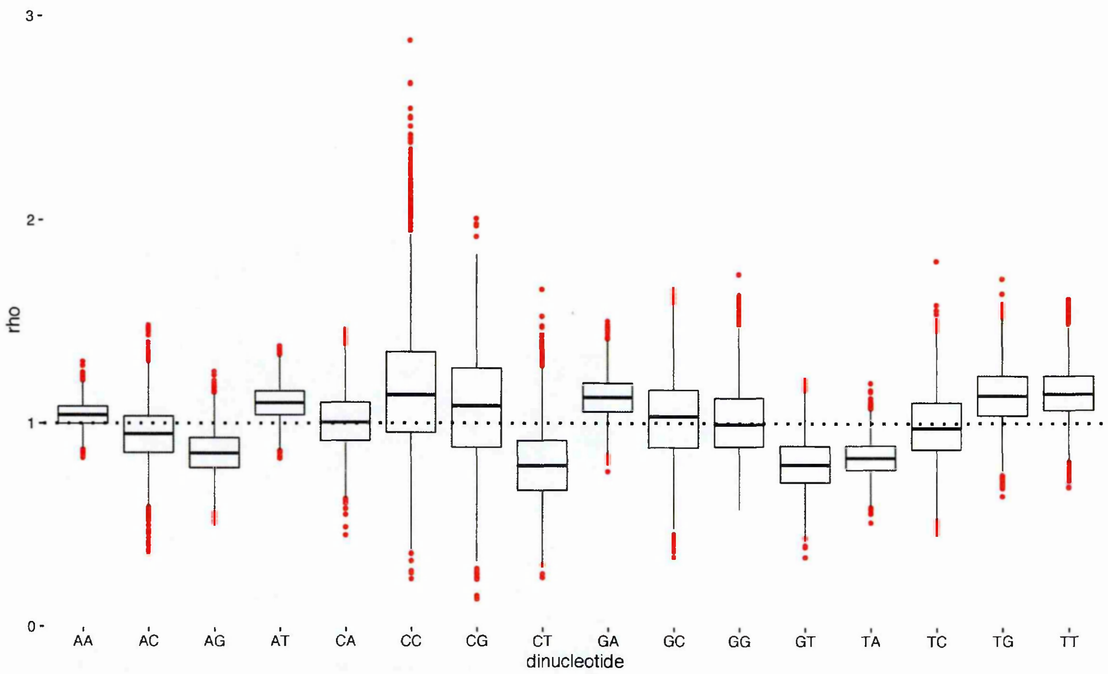


Figure B.4: *A plot of dinucleotide rho statistic in DBLα sequences. The rho-statistic is a measure of how over or under represented a given dinucleotide is. The most over-represented dinucleotide are GA, TT, TG, CC, CG and AT. There was a lot of variation in some dinucleotides like CC where some sequences seem to be very rich in CC and other very under-represented CC. CT,AG, GT and TA dinucleotides are largely under-represented although there is a lot of variation.*

The ρ statistic shows a lot of variation in dinucleotide representation. Generally, GA, TG, TT and CC were over-reprented and CT,AG, GT and TA dinucleotides are largely under-represented.

B.4 Relative synonymous codon usage(RSCU) in DBLα sequences

The relative synonymous codon usage (RSCU) score is a measure of the frequency of a particular codon relative to the frequency that the codon would be observed in the absence of any codon usage bias.

The RSCU score was calculated using equation B.2

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} ni \tag{B.2}$$

where g_{ij} is the observed number of the ith codon for j^{th} amino acid which had type n_i synonymous codons.

A codon that is used less frequently than expected had a value of less than 1 and a value >1.0 indicated positive codon bias meaning that the codon was used more than expected.

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	TTT	14827	1.38	Ser	TCT	3520	1.16
	TTC	6678	0.62		TCC	464	0.15
Leu	TTA	15056	2.46		TCA	4228	1.39
	TTG	6956	1.13		TCG	987	0.32
Tyr	TAT	29303	1.58	Cys	TGT	9848	1.22
	TAC	7701	0.42		TGC	6299	0.78
ter	TAA	0	0	ter	TGA	0	0
ter	TAG	0	0	Trp	TGG	14900	1
Leu	CTT	7829	1.28	Pro	CCT	5097	1.12
	CTC	882	0.14		CCC	3171	0.7
	CTA	4268	0.7		CCA	7169	1.58
	CTG	1795	0.29		CCG	2723	0.6
His	CAT	6182	1.46	Arg	CGT	4223	0.61
	CAC	2258	0.54		CGC	6735	0.98
Gln	CAA	18245	1.52		CGA	8487	1.23
	CAG	5701	0.48		CGG	2259	0.33
Ile	ATT	11607	1.2	Thr	ACT	6555	0.79
	ATC	3883	0.4		ACC	3797	0.46
	ATA	13491	1.4		ACA	16409	1.98
Met	ATG	2341	1		ACG	6401	0.77
Asn	AAT	34498	1.63	Ser	AGT	6993	2.3
	AAC	7748	0.37		AGC	2076	0.68
Lys	AAA	48572	1.53	Arg	AGA	16360	2.38
	AAG	14747	0.47		AGG	3247	0.47
Val	GTT	5169	0.87	Ala	GCT	9208	1.38
	GTC	4247	0.72		GCC	3243	0.49
	GTA	8509	1.43		GCA	9967	1.49
	GTG	5816	0.98		GCG	4309	0.64
Asp	GAT	43675	1.5	Gly	GGT	15898	1.5
	GAC	14389	0.5		GGC	5129	0.48
Glu	GAA	30033	1.7		GGA	17107	1.62
	GAG	5300	0.3		GGG	4206	0.4

Table B.2: *Relative synonymous codon usage (RSCU) indices for DBLα sequences in Kilifi. The RSCU is an indicator of codons that are used more frequently than expected.*

Table B.2 shows a summary of the observed RSCU values for different codons in DBLα sequences collected from Kilifi.

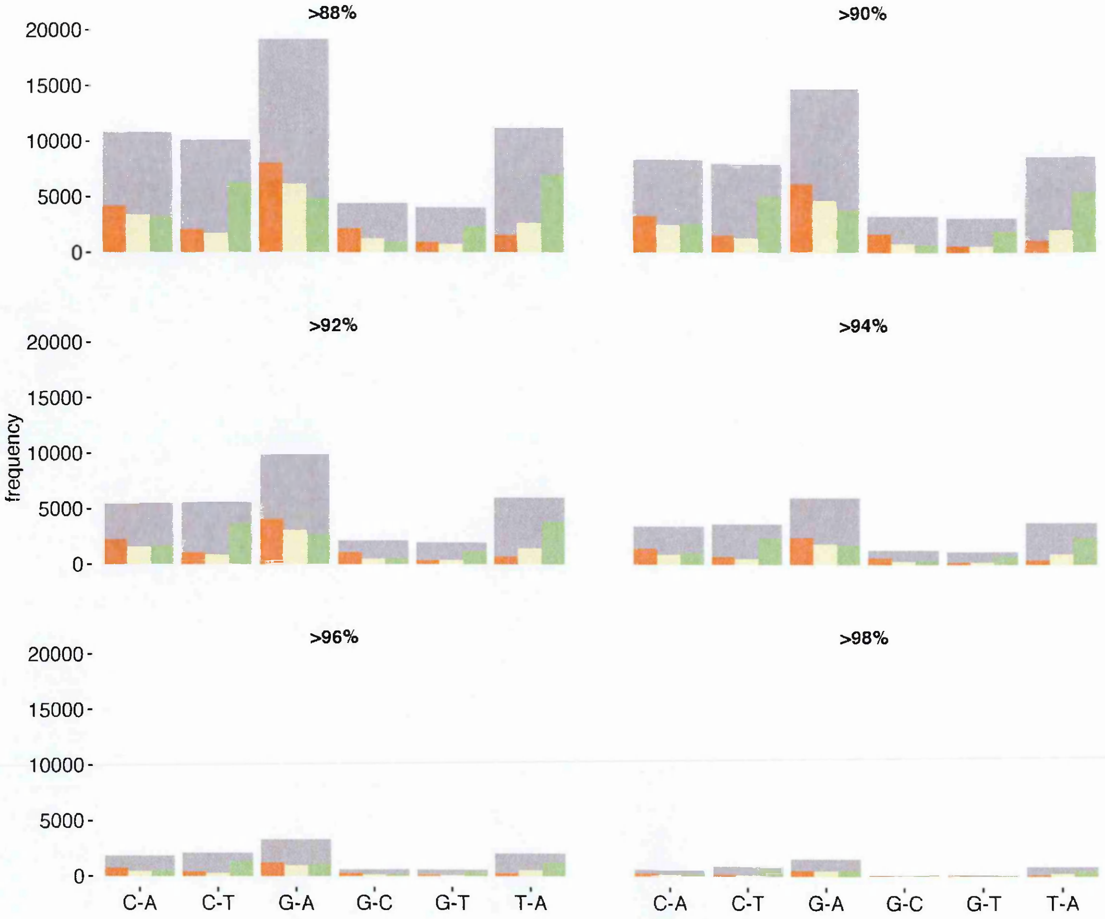


Figure B.5: *A summary of substitutions from DBLα sequence tags that were sequenced using the 454-platform. The y-axis shows the frequency of substitutions and the x-axis shows the substitution type and the frequencies at the respective codon positions.* `G-A, C-T, C-A` *and* `T-A` *substitutions were predominant. Substitutions at the 3rd codon position were preferred by* `C-T, G-T` *and* `T-A` *substitutions.* `C-A, G-A and G-C` *substitutions showed preference for 1st codon positions. The frequency of* `C-A` *substitutions was remarkably higher in this data-set compared to the capillary sequenced data-set.*

B.5 Pattern of variation in var2csa in isolates collected from around the world

A total of 978 full-length and partial *var2csa* sequences were downloaded from Genbank using a keyword search. The DBLα-finder was used to search for published DBLα tags and the 68 sequences that were identified by the program, were clustered as explained previously.

The pattern of substitutions is shown in figure B.6. The most frequent substitutions were G-A, C-T and T-A. The most frequent substitutions at the third codon position were C-T, G-T and T-A. G-A substitutions were more frequent at the 1st codon position in these sequences.

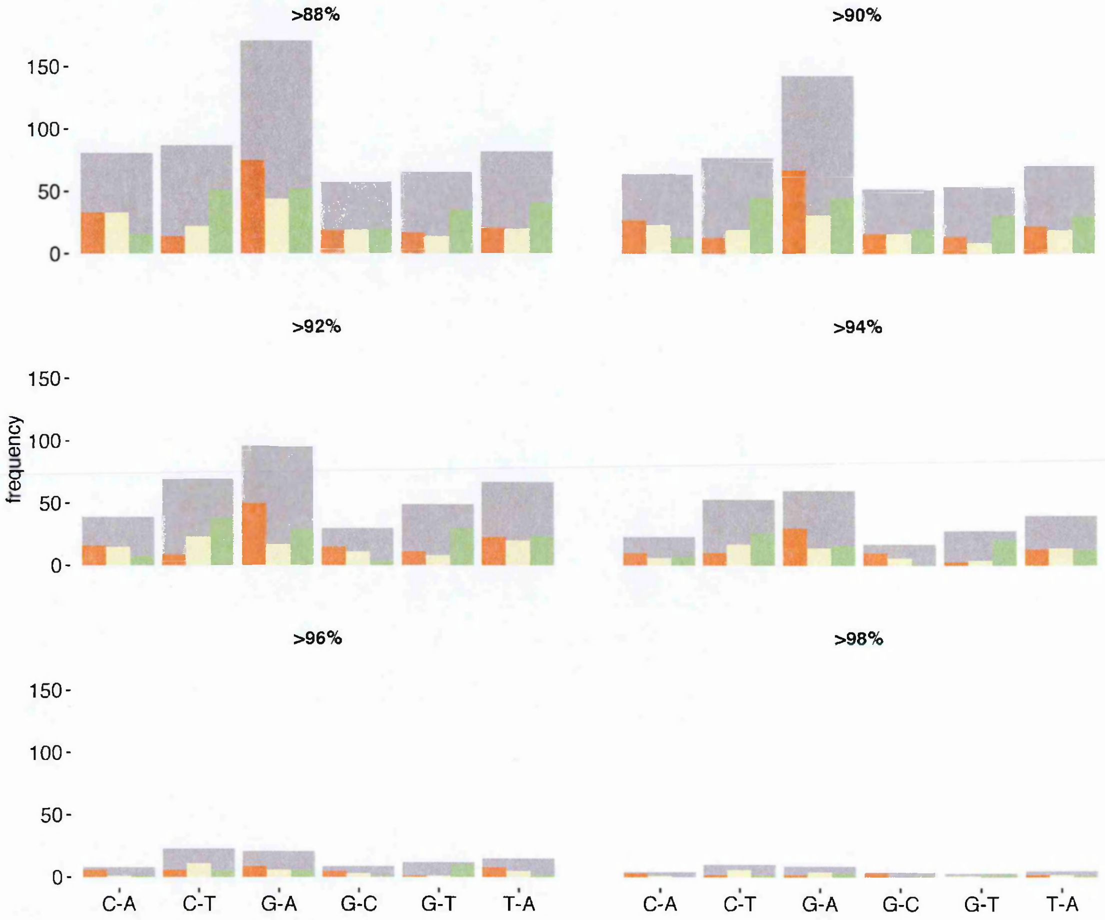


Figure B.6: *A summary of the type and codon location of polymorphisms in 69 DBLα tags from global isolates. The sequences were clustered at between 98% and 88% identity. The y-axis shows the frequency of polymophisms and the x-axis shows the type and codon location of the polymorphisms.* `G-A`*,* `C-T` *and* `T-A` *transversion were the most frequent.*

B.6 Frequency of substitutions in sequences collected from different parts of the world

398 sequences from India, Sudan, Cape Verde, Solomon islands, Philippines, Kenya, Brazil, Thailand, Africa,Vanuatu were retrieved to form a global collection of DBLα sequences. The full list of accession is provided in appendix B.

The sequences were clustered and substitutions were enumerated and plotted. Figure B.7 shows a summary of the distribution of the substitutions.

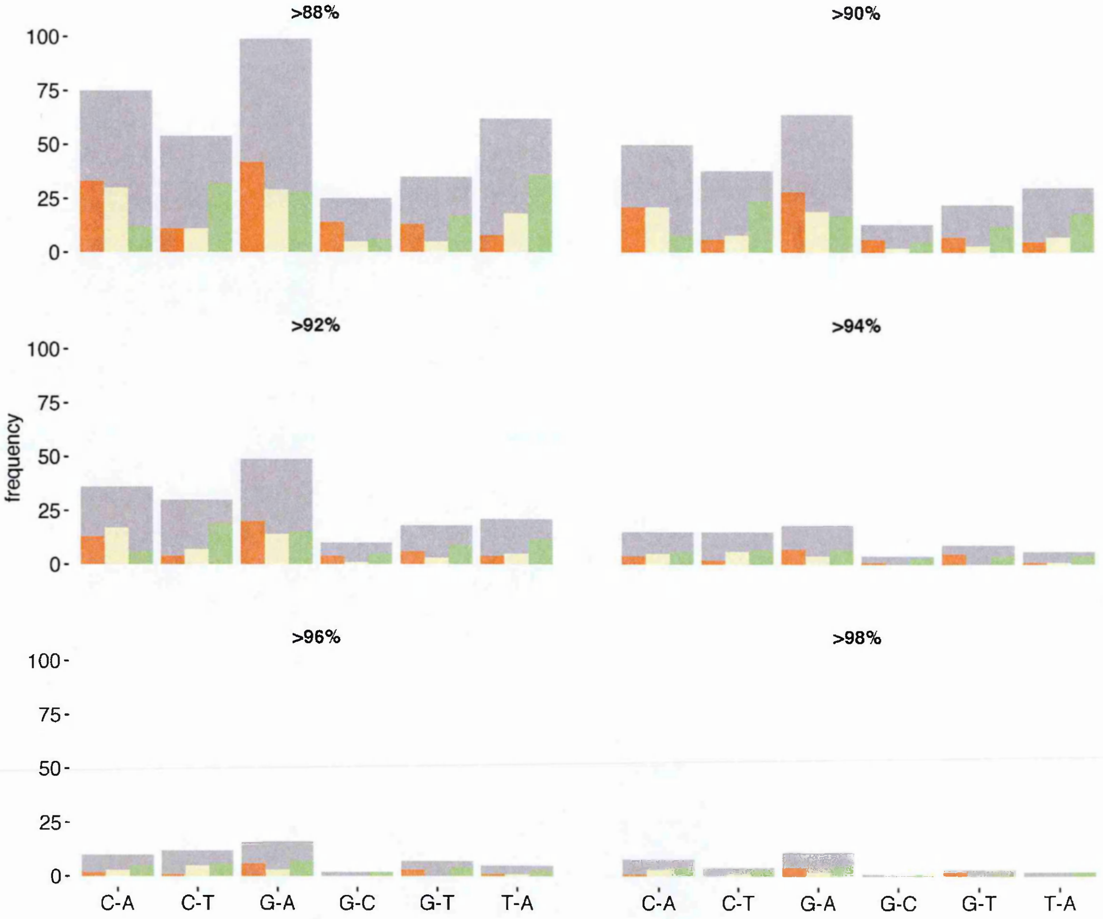


Figure B.7: *The frequency of 367 DBLα tags collected from different geographical regions. The sequences were clustered at between 98 and 88 percent sequence identity. The y-axis represents the frequency of substitutions and the x-axis represents the type and codon position in each of the substitution type.* `G-A` *and* `C-T` *substitutions were more prevalent which is consistent with previous observation from particular geographical regions.*

B.7 Clustering ratio

Table B.3 shows the number of clusters and the fraction of clusters that were used for pairwise comparisons.

Sequences from Kilifi had the lowest clustering ratio compared to sequences from other regions in the world. The sequences from Brazil has relatively higher clustering ratios. This could be explained by the fact that the sequences were collected from

Region	% Identity	Clusters	Total clusters	Ratio
Amele	98	77	180	0.438
	96	76	172	0.441
	94	74	170	0.435
	92	75	168	0.446
	90	75	165	0.455
	88	75	163	0.460
Kilifi	98	465	2725	0.171
	96	479	2665	0.180
	94	496	2617	0.190
	92	526	2553	0.206
	90	543	2497	0.217
	88	571	2422	0.236
Global	98	47	309	0.152
	96	48	306	0.157
	94	48	303	0.158
	92	52	299	0.174
	90	52	295	0.176
	88	53	291	0.182
Brazil	98	160	213	0.751
	96	152	192	0.792
	94	152	192	0.792
	92	152	189	0.804
	90	150	186	0.812
	88	149	182	0.824

Table B.3: *The clustering ratio for sequences collected in various geographical regions are given. The cluster ratio is the proportion of sequences that were used in the pairwise comparisons to the total number of clusters that were created at a given identity threshold. For very diverse sequences the ratio is close to 0 and for more conserved dataset clustering ratio is close to 1. The clustering ratio increased with decrease in sequence identity within each dataset. Sequences from Brazil were more conserved compared to sequences from Kilifi.*

fewer isolates. Nonetheless, sequences from south America have been described as relatively conserved and could miss a lot of virulent sequences compared to sequences from Africa(Albrecht et al. 2010).

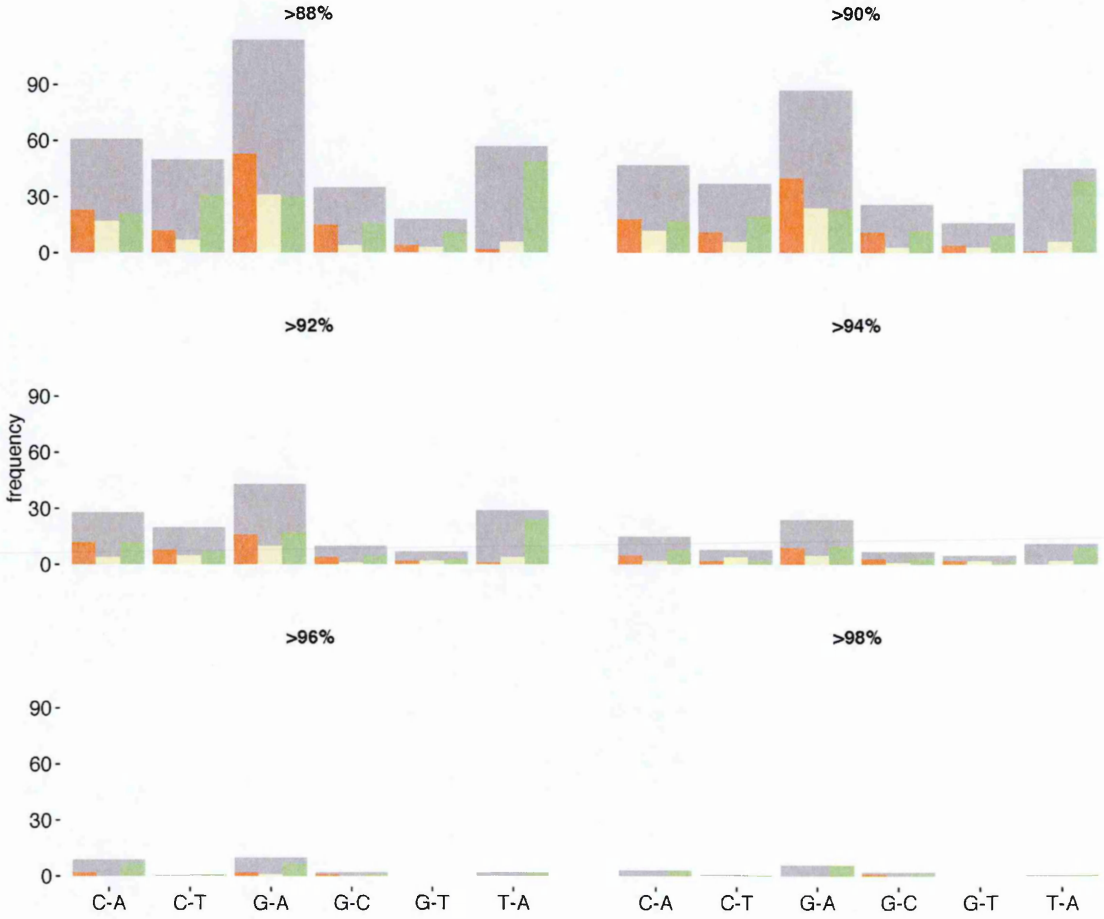


Figure B.8: *Frequency of substitutions in 162 DBLα tags from Madang in Papua new Guinea. The y-axis shows the frequency of substitutions and the x-axis shows the type and codon position for each substitution.*

B.8 Differences in the distance between polymorphic sites

The distance between polymorphic sites in sequence varied based on the clustering threshold. Nucleotide distances were lower in sequences with low identities suggesting that presence of contiguous polymorphic sites. The frequency of nucleotide distance between polymorphic sites is shown in figure B.9.

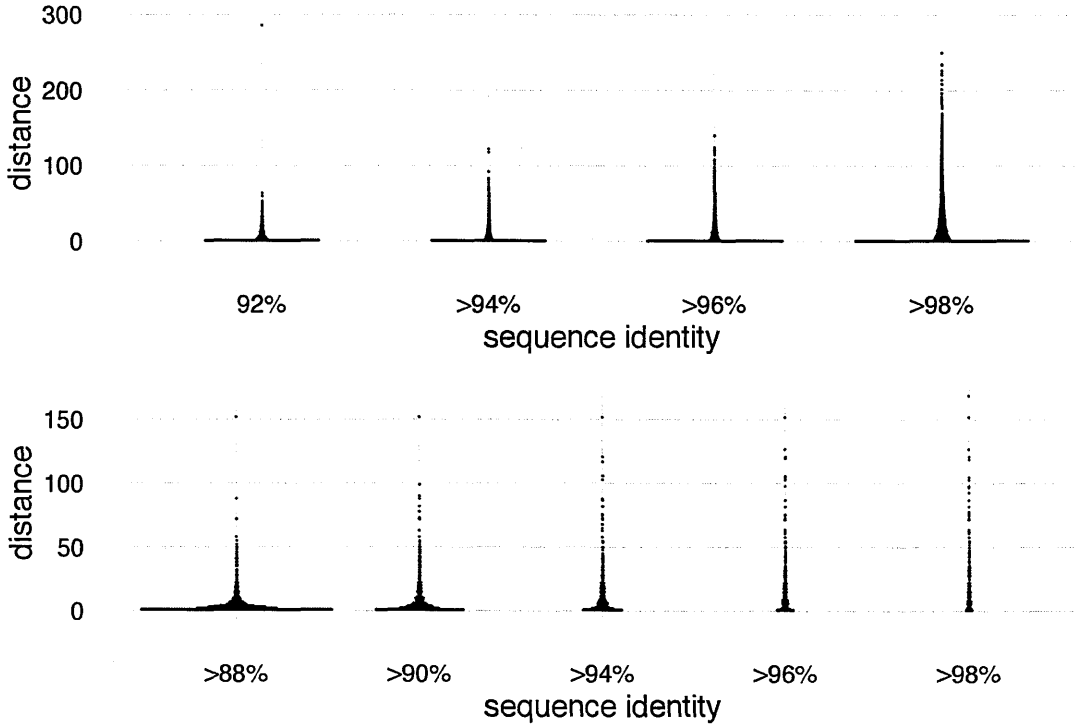


Figure B.9: *A dotplot of nucleotide distances (Y-axis) between adjacent substitution sites in sequences clustered at >98% to 88% sequence identity. The upper panel shows nucleotide distances in sequences where random substitutions were introduced.*

B.9 Homology block F contained relatively more mismatches

There were relatively more substitutions at block F compared to block D. The changes were predominated by G-A and C-T substitutions with most occurred at the 3rd codon except for C-A changes that occured mostly at the 1st codon and the 3rd codon.

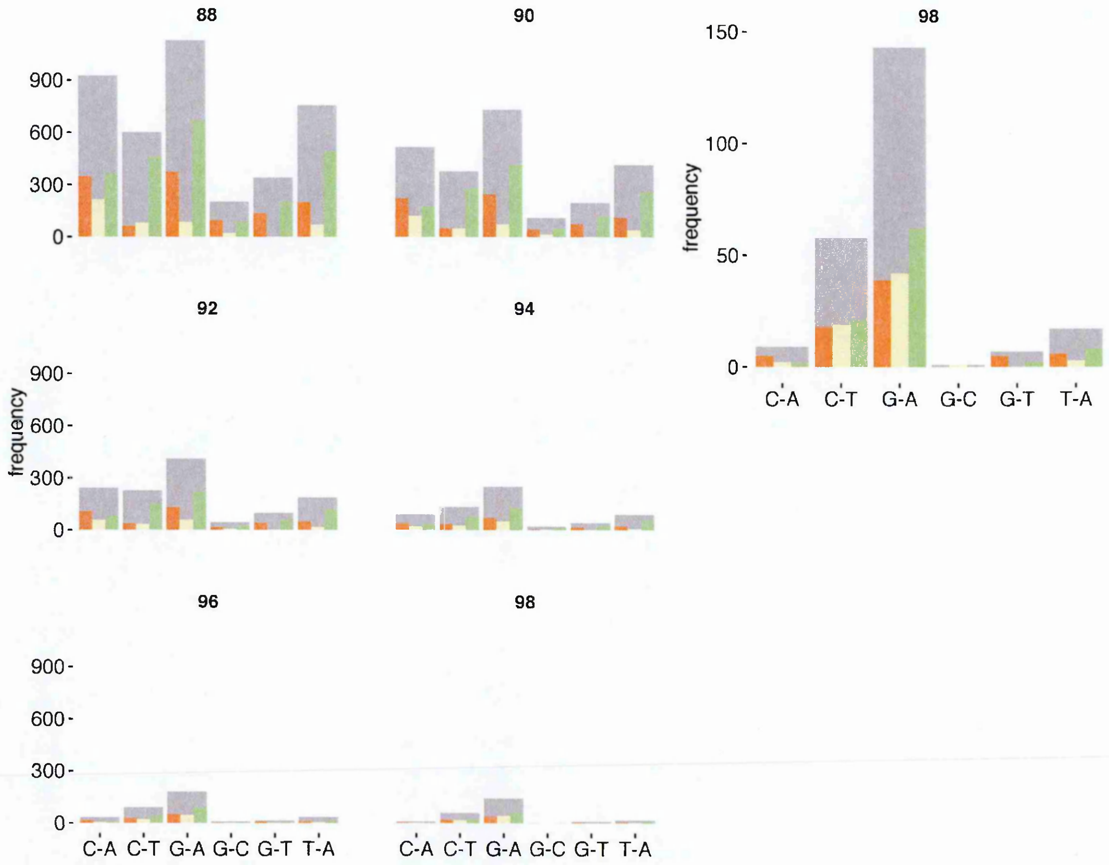


Figure B.10: *A summary of within isolate distribution of polymorphisms in homology block F*

B.9.1 Within-isolate variation in homology block D and H

Figure 4.4 shows the distribution of substitutions in homology blocks D and H in sequences from the same isolate.

Sequences from the same isolate were clustered and aligned and the pairwise differences were tallied across all the isolates. The frequency of substitution was summarized using the frequency plots shown in figure B.11 (block D), figure B.10(block F) and figure B.12 (block H).

Block D region is relatively conserved and very few mutations were counted at

98% sequence identity. The number of observed mutations increase gradually with decrease in sequence identity. Most changes were confined to the third codon. There were very few changes at 98% identity and the profile of changes did not match the one that was observed in lowe percent identities.

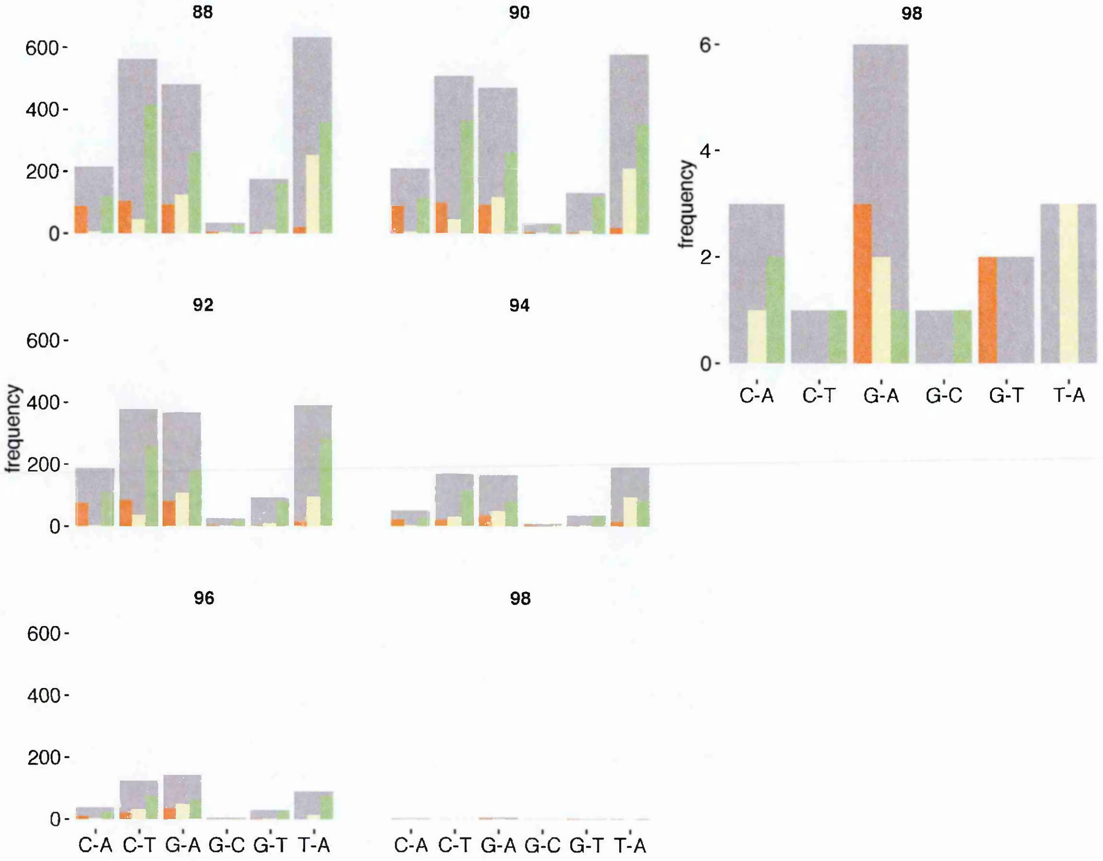


Figure B.11: *A plot of the frequency of substitutions in homology block D region from sequences collected from the same isolate.*

Block H changes were dominated by C-T substitutions with relatively high frequencies at the third codon position. The profile for changes at 98% identity was different from the rest, though the numbers were few and may be due to errors.

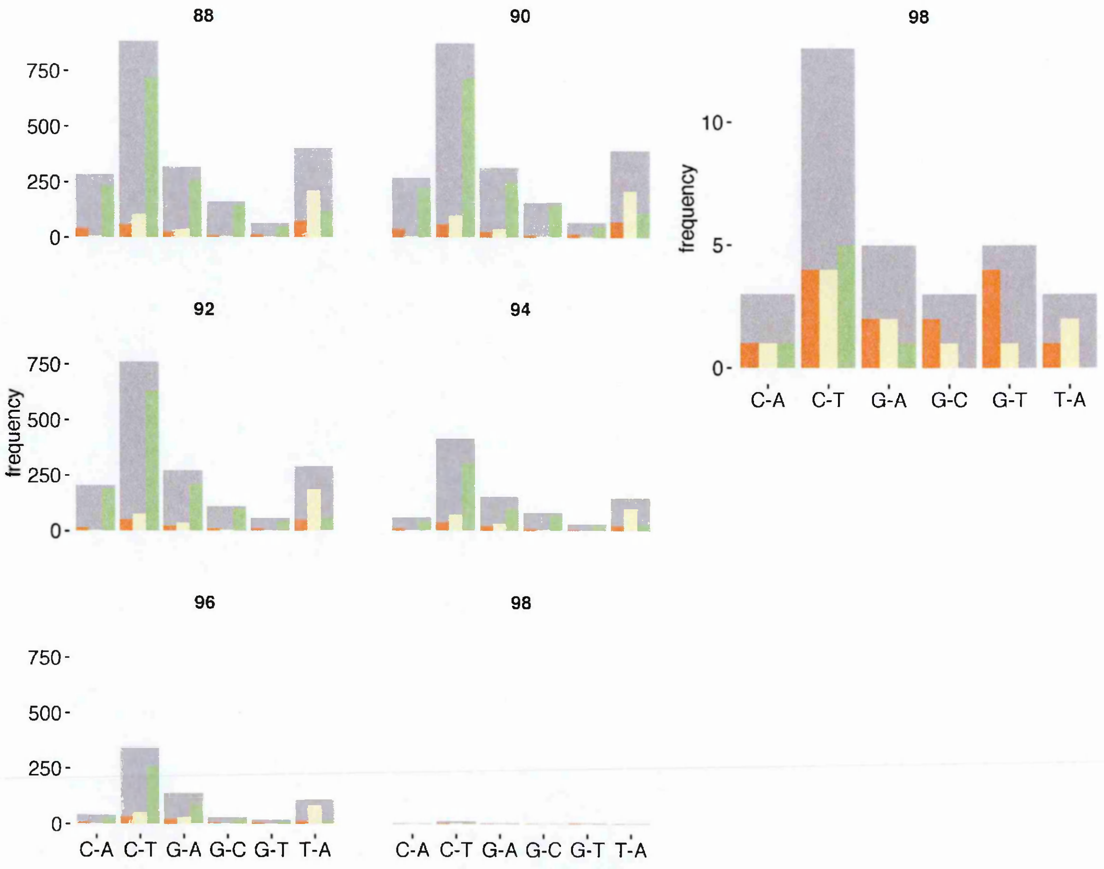


Figure B.12: *Within isolate distribution of substitutions in homology block H. Sequences with equal to or greater than 98% identity were few they have the same profile that was found for full length sequence tags. There was a bias for substitutions at the 3rd codon position except for T-A changes where substitutions at the 2nd codon position were more preferred.*

B.10 HLA prevalence

The frequecies of HLA alleles among individuals in the coastal region is unknown. In the absence of HLA allele frequency, HLA alleles were chosen to cover the known supertypes such that they covered the 9 HLA-DR supertypes, 6 HLA-DQ, and 6 HLA-DP alleles.

A sense of HLA alleles that are prevalent in Kilifi was obtained from limited data aimed at identifying prevalent HLA alleles among children in Kilifi (unpublished).

HLA typing was perfomed on 11 samples that were collected from children in Kilifi (which datasets? How representative are they?). Figure B.13 shows a frequency plot of the HLA alleles from the 11 samples. While this is not conclusive, a sense of

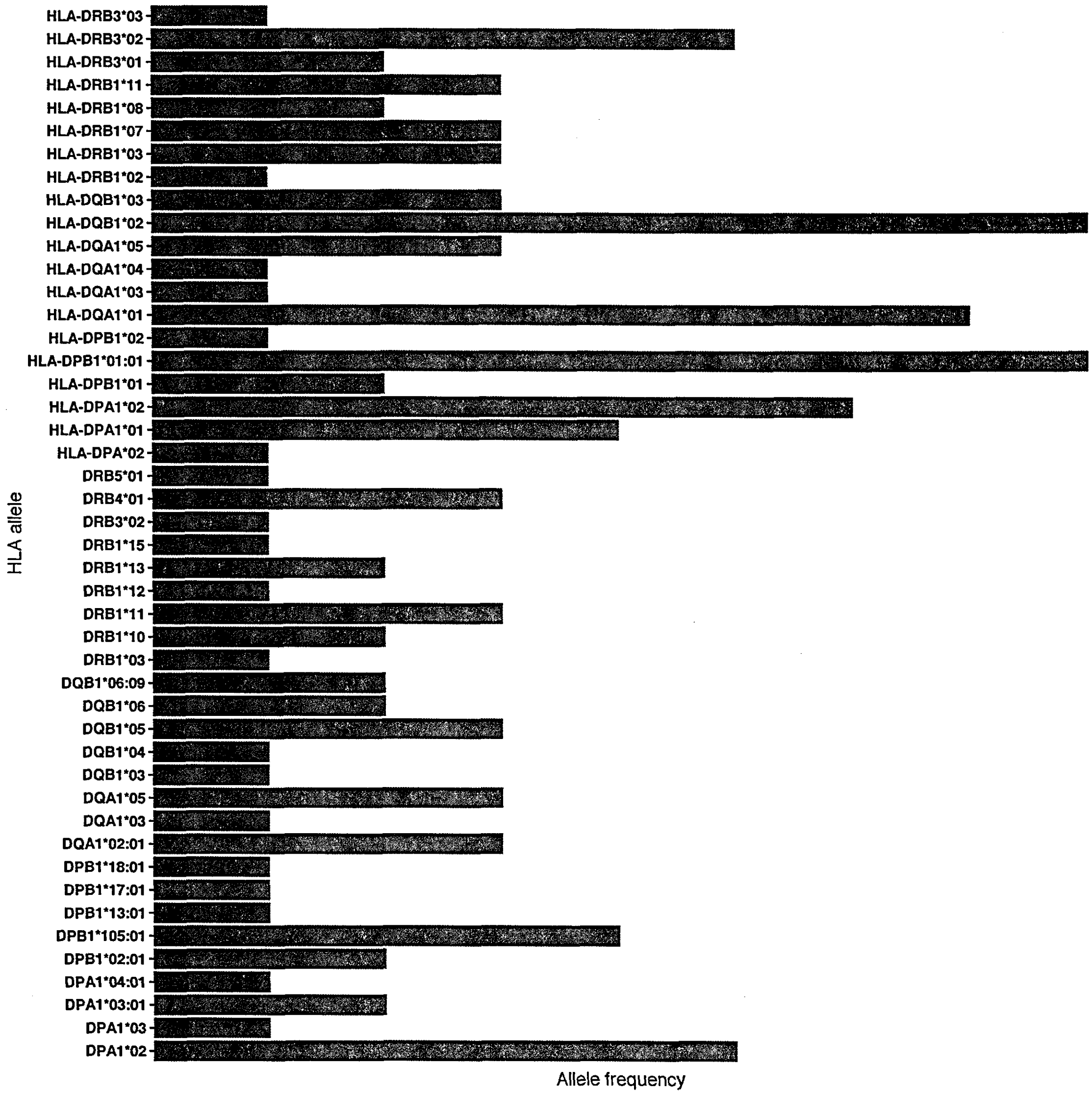


Figure B.13: *A frequency histogram plot of HLA alleles from 11 samples in Kilifi. The T-cell HLA typing was perfomed using samples from Kilifi. This information was not used and did not influenced the MHC T-cell predictions used in this thesis.*