Antigenic variation and its evolution in *P. falciparum* malaria

Robert Noble

Linacre College

Trinity 2013

A dissertation submitted to the Department of Zoology and the Committee on Graduate Studies of Oxford University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, D.Phil

Abstract

Antigenic variation and its evolution in *P. falciparum* malaria

Robert Noble

D.Phil

Linacre College, University of Oxford, Trinity 2013

This thesis investigates antigenic variation and its evolution in *Plasmodium falciparum*, the cause of the most deadly form of human malaria. Antigenic variation is a strategy for evading immunity by switching between antigenic variants during infection. In *P. falciparum*, such variable antigens confer different binding phenotypes that may affect parasite survival and have also been linked to pathology. Here, a new statistical method is described for determining the switching patterns that underlie antigenic variation. This method is then applied to experimental data to yield a full description of an antigenic switching network in *P. falciparum*. In light of the findings, theoretical modelling is used to show how immune selection and binding phenotypes may have contributed to the evolution of antigenic repertoire structure, expression order and virulence. Related models are also used to investigate parasite population diversity, providing possible explanations for observations reported here and elsewhere, with implications for vaccine design. Together, these chapters advance understanding of *P. falciparum* immune evasion and how it relates to pathology. This work further reinforces the role of host immunity in shaping pathogen population diversity at multiple levels.

Acknowledgements

I am grateful to many friends, family, colleagues and the community of Linacre Common Room for moral support and for countless helpful discussions. Special thanks are due to Mario Recker for his excellent and dedicated supervision, to Sunetra Gupta for her expert co-supervision, and to Zóe Christodoulou, Bob Pinches, Sue Kyes and Chris Newbold, not least for their patience in helping teach me biology.

Contents

1 Introduction **1**

1.1 Biological background 1

1.1.1 Naturally acquired immunity to *P. falciparum* 1

1.1.2 *Var* groups and domain types 4

1.1.3 PfEMP1 and severe disease 6

1.1.4 Sequence diversity 9

1.1.5 Antigenic variation 12

1.2 Mathematical approaches 14

1.2.1 Models of antigenic variation 14

1.2.2 Estimating switching parameters 17

1.2.3 Models of diversity 22

1.3 Types of mathematical model 26

1.4 Thesis outline 28

2 A statistically rigorous method for determining antigenic switching networks **31**

2.1 Introduction 31

2.2 Methods 35

2.2.1 Dynamical model 35
2.2.2 Simulated annealing 36
2.2.3 Experimental error 37
2.2.4 MCMC 38
2.3 Testing the method 40
2.3.1 Resolving switch networks from a single transcription time series 41
2.3.2 Resolving switch networks from multiple transcription time series 49
2.3.3 MCMC 53
2.3.4 Dimension reduction 57
2.3.5 Hypothesis testing 60
2.3.6 Testing with real data 63
2.4 Discussion 63

3 The antigenic switching network of *Plasmodium falciparum* and its implications for the immuno-epidemiology of malaria 68
3.1 Introduction 68
3.2 Methods of modelling and analysis 71
3.3 Laboratory materials and methods 73
3.3.1 Parasite culture and cDNA preparation 73
3.3.2 Primer design and initial testing 73
3.3.3 qRT-PCR 74
3.4 Analysis of the HB3 data set 75
3.4.1 Overview of the data 75
3.4.2 *Var* gene transcription dynamics 76

3.4.3 Parameter estimation 79
3.4.4 Off-rate estimates suggest a role of chromosomal location, not promoter type 81
3.4.5 *Var* gene switching is governed by an activation hierarchy . . 84
3.4.6 *Var* gene switching is influenced by the deactivating gene . . . 88
3.4.7 *Var* gene sequences suggest a role of activation bias in antigenic evolution 90
3.5 Validating the results 94
3.5.1 Initial conditions of the model 94
3.5.2 Cross-validation of parameter estimates 98
3.5.3 Validation of model assumptions 100
3.5.4 Growth rates 101
3.6 Discussion 102

4 Modelling the evolution of structure and expression order in the *Plasmodium falciparum* antigenic repertoire 107
4.1 Introduction 107
4.2 Models of repertoire structure evolution 111
4.2.1 Host births and deaths 112
4.2.2 Assessment of repertoire structures 112
4.2.3 Model comparisons 114
4.3 Expression order model 116
4.4 Results of repertoire structure models 118
4.4.1 Repertoire structure 118
4.4.2 Sub-repertoire sizes 126

4.4.3 Multiple epitope models 130
4.5 Results of expression order model 131
4.6 Discussion 134

5 Modelling the evolution of genetic diversity and strain structure in *P. falciparum* antigenic repertoires 138

5.1 Introduction 138
5.2 Evolutionary model 142
5.2.1 Sequence representation 142
5.2.2 Recombination and immune selection of modified repertoires . 145
5.2.3 Measures of diversity and overlap 148
5.2.4 Model parameters 150
5.3 Analysis of genetic diversity 151
5.3.1 Effects of expression order 151
5.3.2 Effects of gene length 153
5.3.3 Effects of functional constraints 154
5.3.4 Effects of epidemiology 157
5.4 Analysis of repertoire diversity 162
5.4.1 Results of a simplified model 162
5.4.2 Results with sequence recombination 163
5.4.3 Results with sequence recombination and unequal functional constraints 168
5.5 Discussion 171

6 Summary and Conclusion 174

Appendices **183**

Bibliography **229**

Chapter 1

Introduction

1.1 Biological background

1.1.1 Naturally acquired immunity to *P. falciparum*

The mosquito-borne apicomplexan parasite *Plasmodium falciparum* causes the deadliest form of human malaria, accounting for hundreds of thousands of deaths per year, mostly among young children in sub-Saharan Africa [Hay *et al.*, 2010, WHO, 2012]. Having two very different hosts, this single-cell pathogen has evolved a complex, multi-stage life cycle (Figure 1.1). Feeding anopheline mosquitoes inject humans with parasites in the form of sporozoites that are swiftly carried by the bloodstream to the liver. Inside liver cells, sporozoites mature into multinucleated schizonts, which, after a few days, rupture to release into the bloodstream tens of thousands of mononucleated merozoites. The parasites then invade red blood cells, where they absorb nutrients and grow in size. In most cases, development continues with the formation of schizonts, each of which releases multiple new merozoites in an asexual replicative cycle that repeats approximately every 48 hours. A minority of

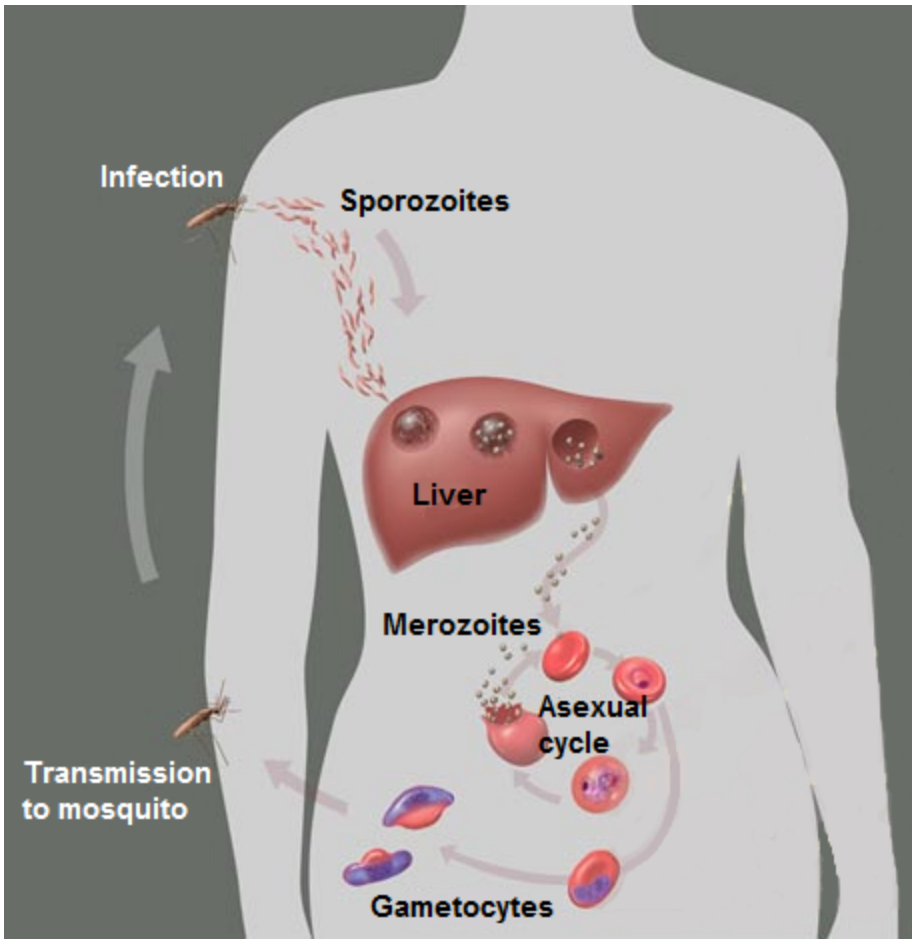


Figure 1.1. *P. falciparum* life cycle. Adapted from an image created by the National Institutes of Health (history.nih.gov/exhibits/bowman/SSmalaria.htm).

merozoites do not undergo schizogony but instead differentiate into male or female gametocytes. When ingested by a mosquito, gametocytes sexually reproduce to form sporozoites and thus begin a new transmission cycle.

Humans infected with *P. falciparum* for the first time almost always show symptoms and are at substantial risk of severe disease in the form of disturbed consciousness, respiratory distress or severe anaemia [Marsh *et al.*, 1995]. Where the parasite is endemic, the risk of morbidity falls markedly with age and children may acquire

immunity to non-cerebral severe malaria after only one or two infections [Gupta *et al.*, 1999, Marsh & Snow, 1997]. In contrast, immunity to infection is acquired slowly [Langhorne *et al.*, 2008]. Therefore in highly endemic regions, most older children and adults are more or less continuously infected with low level parasitemia that is asymptomatic or causes only mild illness, each infecting strain typically remaining in the host for many months [Bretscher *et al.*, 2011, Sama *et al.*, 2004]. The persistence of naturally acquired immunity appears to depend on frequent exposure and in low transmission areas individuals of all ages are at substantial risk of severe disease [Doolan *et al.*, 2009, Langhorne *et al.*, 2008].

The main targets of acquired immunity are variant antigens expressed on the surface of infected red blood cells. Infected hosts readily produce antibody responses that lead to the destruction of cells expressing these antigens. However, the variant surface antigens are enormously variable and the immune responses are generally variant-specific; there is no evidence that hosts ever acquire strain-transcendent sterilising immunity [Tran *et al.*, 2013]. It appears that parasites evade immune responses by expressing variants that correspond to gaps in the host's acquired antibody repertoire [Giha *et al.*, 1999, Marsh & Howard, 1986, Marsh *et al.*, 1989, Ofori *et al.*, 2002]. The most important variant surface antigen is generally thought to be *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) [Bull *et al.*, 1998, Chan *et al.*, 2012], which has the essential function of mediating adherence to other host cells so that infected erythrocytes escape clearance by the spleen [Boone & Watters, 1995, Buffet *et al.*, 2009].

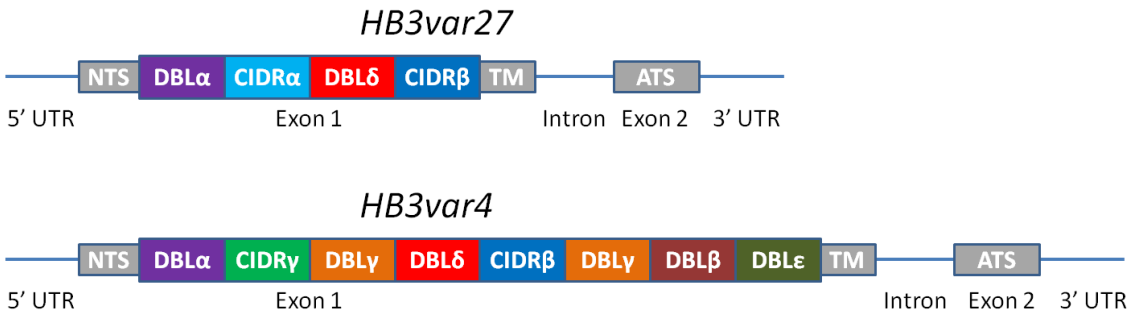


Figure 1.2. Domain structures of two *var* genes from the HB3 isolate. *HB3var27* has a four-domain extracellular structure typical of many UpsB and UpsC genes; *HB3var4* is an UpsA gene with a less common structure containing additional DBL domains [Kraemer *et al.*, 2007]. NTS = N-terminal segment, TM = transmembrane region, ATS = acidic terminal segment.

1.1.2 *Var* groups and domain types

PfEMP1 is encoded by the *var* gene family [Baruch *et al.*, 1995, Smith *et al.*, 1995, Su *et al.*, 1995a]. Each *P. falciparum* genome contains a repertoire of approximately 60 of these genes encoding distinct antigenic variants, only one of which is expressed at a time [Chen *et al.*, 1998, Scherf *et al.*, 1998]. Within and between repertoires, *var* genes have extremely high sequence diversity, yet they can be divided into groups A, B, C and E on the basis of a conserved upstream (Ups) flanking sequence, chromosomal location, and orientation [Kraemer *et al.*, 2007, Lavstsen *et al.*, 2003]. Group A (UpsA) genes are located close to the telomere; Group B includes both subtelomeric and centrally located genes; Group C genes are centrally located; and Group E comprises conserved subtelomeric *var2csa* genes linked to placental malaria. Some subtelomeric genes with UpsB flanking sequences have been described as belonging to an intermediate Group B/A.

Each *var* gene has two exons, the larger of which encodes multiple adhesion do-

mains known as Duffy binding-like (DBL) domains and cysteine-rich inter-domain regions (CIDR) [Smith *et al.*, 2000]. CIDR domains have been grouped according to sequence homologies into five subclasses ($\alpha, \beta, \gamma, \delta$ and pam) and DBL domains have been grouped into six major subclasses ($\alpha, \beta, \gamma, \delta, \epsilon, \zeta$) and five smaller subclasses [Rask *et al.*, 2010, Smith *et al.*, 2000]. Certain domain types and configurations of domain types are associated with particular Ups groups. Most UpsC and many UpsB genes include only the four binding domains DBLα, CIDRα, DBLδ and CIDRβ/γ, whereas typical UpsA genes and some UpsB genes can contain as many as five additional DBL domains [Rask *et al.*, 2010]. Two contrasting gene structures are shown as examples in Figure 1.2. There is evidence that distinct *var* architectures are maintained by a recombination hierarchy, such that recombination is most frequent between similarly structured genes [Bull *et al.*, 2008, Kraemer & Smith, 2003, Kraemer *et al.*, 2007, Lavstsen *et al.*, 2003].

The different *var* structures confer different binding affinities. The domains in UpsE genes – also known as *var2csa* genes – are specialised for binding to placental cells [Fried & Duffy, 1996]. The ability to bind to the human receptor ICAM1 is restricted to a subset of PfEMP1 variants containing DBLβ and encoded by UpsB and UpsC *var* genes [Howell *et al.*, 2008], whereas binding to CD36 depends on sequences within certain CIDRα domains encoded by a large proportion of UpsB and UpsC genes [Baruch *et al.*, 1997, Robinson *et al.*, 2003]. Certain UpsA and UpsB domain types mediate binding to uninfected erythrocytes in a process known as rosetting, which is implicated in some, though by no means all, severe malaria cases [Rowe *et al.*, 2009].

1.1.3 PfEMP1 and severe disease

A link between variant surface antigens and pathology is perhaps clearest in placental malaria, a disease that is life-threatening for both pregnant women and developing foetuses. Placental malaria is characterised by the accumulation of infected red blood cells in placental blood vessels, dependent on binding to chondroitin sulfate A (CSA) [Beeson *et al.*, 2000, Fried & Duffy, 1996]. It is thought that a specialised type of *var* gene, *var2csa*, is unique in mediating this binding [Duffy *et al.*, 2006, Viebig *et al.*, 2005], and placental malaria has been linked to the expression of *var2csa* [Salanti *et al.*, 2004, 2003]. One or two *var2csa* genes are found in almost all parasite repertoires, yet among *var* genes they have unusually low sequence diversity. Women typically acquire antibodies that protect them against placental malaria after their first or second infection during pregnancy [Fried *et al.*, 1998, Maubert *et al.*, 1999, Ricke *et al.*, 2000]. Therefore the extracellular domains encoded by *var2csa* are attractive targets for a vaccine against this particular form of severe malaria.

Partly inspired by the example of placental malaria, extensive efforts have been made to identify other types of variant surface antigen associated with severe disease. Bull *et al.* [2000] tested the agglutination frequency of *P. falciparum* isolates with a panel of heterologous plasma from Kenyan children. They showed that isolates from children with severe malaria were more commonly recognised than those from children with mild malaria, and that older children tended to have less commonly recognised antigens. Studies in Ghana [Nielsen *et al.*, 2002] and Gabon [Lindenthal *et al.*, 2003] confirmed these findings, and Jensen *et al.* [2004] went on to show that selected expression of the commonly recognised antigens was accompanied by up-regulation of Group A *var* genes in a cultured parasite isolate.

Analysis of the *var* genes expressed *in vivo* is hampered by genomic diversity.

Therefore researchers have focused on short sequence tags from the DBLα domain found in most PfEMP1 variants. Studies in various populations (Brazilian adults, Papua New Guinean children, and children in various parts of sub-Saharan Africa) between 2002 and 2010 found that variants containing certain DBLα sequence tags are more commonly expressed in severe malaria than in milder cases [Blomqvist *et al.*, 2010, Bull *et al.*, 2005, Falk *et al.*, 2009, Kaestli *et al.*, 2006, Kalmbach *et al.*, 2010, Kirchgatter & del Portillo, 2002, Kyriacou *et al.*, 2006, Rottmann *et al.*, 2006]. Although not entirely consistent – perhaps because of differing methods, subjects and *P. falciparum* prevalence – these results generally suggested that parasite populations in children with more severe forms of malaria predominantly transcribe *var* genes with sequence tags characteristic of Group A and B/A, whereas asymptomatic infection and mild malaria are associated with tags from Groups B and C.

Another series of studies in Tanzania showed that children acquire PfEMP1 antibodies faster in areas with higher malaria transmission intensity and that some of these antibodies protect against severe malaria [Joergensen *et al.*, 2006, Lusingu *et al.*, 2006, Magistrado *et al.*, 2007, Vestergaard *et al.*, 2008]. Moreover, independently of transmission intensity, antibodies are first acquired against PfEMP1 variants encoded by Group A and B/A *var* genes in an ordered sequence [Cham *et al.*, 2010, 2009]. This finding is consistent with studies of experimentally infected, non-immune volunteers, whose early infections are typically dominated by parasites expressing a subset of UpsA and UpsB *var* genes [Lavstsen *et al.*, 2005, Zhang *et al.*, 2011].

From all of these observations emerged a general consensus that a set of PfEMP1 variants with restricted antigenic diversity are more likely to cause severe malaria and dominate infections until hosts have been sufficiently exposed to acquire pro-

tection against each of them. Warimwe *et al.* [2009] strengthened evidence for this hypothesis. They showed that expression of particular *var* genes was strongly associated with severe malaria and, independently, with young host age. A weaker, negative correlation was found between expression of these genes and the level of the host's IgG antibody response to parasite isolates. The relevant genes were termed cys2 genes because they have DBLα domains containing two cysteine residues, a feature common to all of Group A and some of Group B/A but otherwise highly unusual.

Although the studies described so far found *var* sequence features correlated with severe malaria, they fell short of identifying the cause. A major advance came in 2012 when three research groups published their findings of associations between severe malaria and two specific combinations of binding domains, known as domain cassettes. Parasites selected for binding to human brain endothelial cells were found to express *var* genes containing either domain cassette 8 (DC8) or domain cassette 13 (DC13) and this binding could be disrupted by the serum of children who had experienced severe malaria [Avril *et al.*, 2012, Claessens *et al.*, 2012]. Further, genes containing DC8 or DC13 (always UpsA or UpsB) were shown to be disproportionately expressed in children with severe disease, compared to those with uncomplicated malaria [Lavstsen *et al.*, 2012]. Genes containing DC8 or DC13 were also shown to bind cells from other human tissues [Avril *et al.*, 2012, Claessens *et al.*, 2012] and a subsequent study demonstrated that they could be selected for binding to heart, lung and bone marrow as well as brain cells [Avril *et al.*, 2013]. In 2013, the relevant host receptor in all these tissues was identified as endothelial protein C receptor (EPCR) [Turner *et al.*, 2013]. Another important recent development was the first successful induction of strain-transcending surface-reactive antibodies to a

different subset of PfEMP1 variants associated with rosetting, a binding phenotype that has also been linked to severe disease [Ghumra *et al.*, 2012].

Ascertaining conserved features of the antigens linked to severe malaria gives a major boost to vaccine and drug design. These discoveries also further understanding of *var* gene evolution. Since the late 1990s, it has been hypothesised that the variant surface antigens predominantly expressed in early infections confer a within-host survival advantage, possibly linked to binding affinity and/or avidity [Bull *et al.*, 2000, 1999]. In light of recent findings, a more specific understanding is that *var* genes have evolved into two distinct subsets: one type that encodes CD36-binding domains and another, more antigenically restricted type that instead binds EPCR or binds to uninfected erythrocytes to form rosettes [Smith *et al.*, 2013]. The former type is associated with mild malaria and the latter type is associated with severe disease. The ability of DC8 and DC13 variants to bind avidly to diverse tissues supports the hypothesis that they confer a within-host survival advantage. This argument is further strengthened by evidence that these variants may directly outcompete CD36-binding variants in some sequestration sites [Avril *et al.*, 2013].

1.1.4 Sequence diversity

To maintain antigenic diversity, *var* genes undergo mitotic recombination far more frequently than the majority of the parasite genome [Bopp *et al.*, 2013, Duffy *et al.*, 2009, Freitas-Junior *et al.*, 2000]. It is thought that recombination results from DNA breaks repaired either by the deletion of a sequence fragment or the copying of a sequence fragment from one gene to another (gene conversion). The recombination rate is too low to cause much diversification in parasite cultures but is sufficient to generate new variants within the duration of a typical infection [Bopp *et al.*,

2013]. Immune selection presumably determines which new variants proliferate and are transmitted to other hosts. Sequence recombination has also been observed during meiosis [Kerr *et al.*, 1994] but it is unclear to what extent this contributes to population genetic diversity.

Although very little is known about the extent of *var* sequence homology necessary to generate cross-reactive immune responses to the various PfEMP1 domains, it has often been assumed that sequence diversity and antigenic diversity are correlated. The high degree of recombination among *var* sequences makes standard alignment-based comparisons inappropriate. Instead, one recent analysis by Bull *et al.* [2008] mapped DBLα sequence tags from Kenyan isolates onto a network by connecting sequences that shared one or more polymorphic sequence blocks. The resulting network was highly structured and contained distinct lobes corresponding to Group A and rosetting *var* genes, implying that these types have relatively low sequence diversity and appear to be evolving separately from other types. These findings are consistent with sequence diversity analyses of isolates from diverse epidemiological settings [Barry *et al.*, 2007, Chen *et al.*, 2011, Falk *et al.*, 2009, Mugasa *et al.*, 2012, Trimnell *et al.*, 2006, Ward *et al.*, 1999]

Whereas most research has focussed exclusively on the DBLα domain, a recent analysis of all binding domains in seven sequence genomes from around the world yielded consistent results with respect to sequence diversity within *var* groups [Buckee & Recker, 2012]. As shown in Figure 1.3, a negative correlation was found between sequence conservation of domain subclasses and their frequency within these seven genomes. Further, the study identified a positive correlation between the number of domains in a *var* gene and its sequence conservation. In light of mathematical modelling results (as discussed below), the authors suggested that gene length might

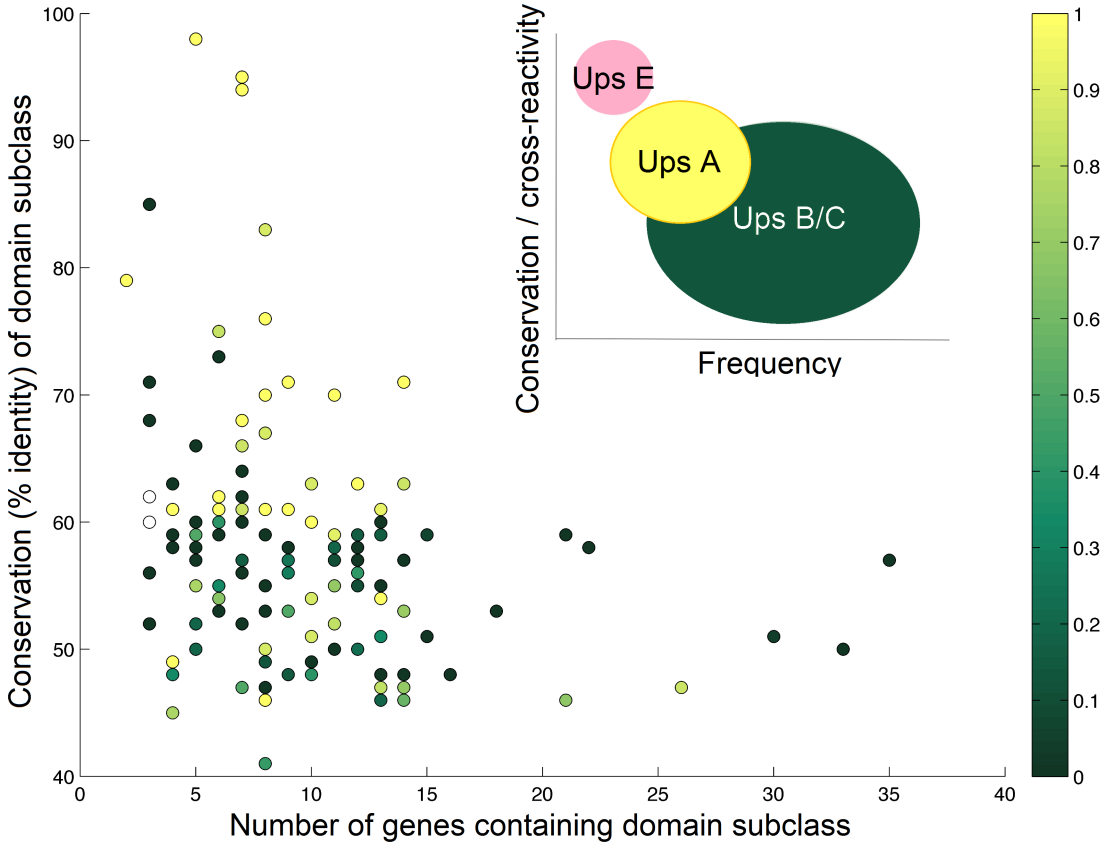


Figure 1.3. Relationship between sequence conservation, Ups group, and frequency of domain subclasses. In the main chart, points correspond to domain subclasses. Yellow represents domains found only in UpsA genes, dark green shows domains only found in UpsB or UpsC genes, and intermediate colours are used for domains found in all Ups groups. The inset chart shows the hypothetical population distribution of *var* genes according to Ups type. Figure adapted from Buckee & Recker [2012].

be a factor in the association between UpsA genes and severe disease.

1.1.5 Antigenic variation

While recombination maintains population diversity, over the course of an infection PfEMP1 population diversity is primarily generated by changing antigens in a systematic and readily reversible manner – a strategy known as antigenic variation [Deitsch *et al.*, 2009]. Antigenic variation of PfEMP1 is achieved by transcriptional switching between *var* genes [Kyes *et al.*, 2007]. Each parasite expresses only one *var* gene at a time and in most cases offspring express the same *var* as their parent. Occasionally, however, a parasite will switch to transcribe a different gene [Roberts *et al.*, 1992].

Recent studies have implicated multiple mechanisms in regulating antigenic variation in *P. falciparum*. Different histone marks are enriched in the upstream flanking regions of active, silent, and "bookmarked" *var* genes. The transcriptionally active gene is associated with di- and trimethylation of the histone H3K4, whereas silent genes are associated with trimethylation of H3K9 [Chookajorn *et al.*, 2007, Flueck *et al.*, 2009, Lopez-Rubio *et al.*, 2007, Perez-Toledo *et al.*, 2009] and H3K36 [Jiang *et al.*, 2013]. When parasites are modified so that they cannot apply the latter mark, almost all of the *var* repertoire is expressed simultaneously [Jiang *et al.*, 2013]. Transcription of the active *var* gene diminishes in later developmental stages but the histone modification H3K4me3 is retained, suggesting that it might bookmark the gene for transcription in the next cell cycle [Lopez-Rubio *et al.*, 2007]. The transcription process itself has been shown to be necessary for a gene to remain active over multiple generations, perhaps by causing the recruitment of histone modifiers [Dzikowski & Deitsch, 2008]. Silencing of *UpsA* and *UpsE* genes appears to depend

on modification of the chromatin environment by the telomere binding protein PfSir2A [Duraisingh *et al.*, 2005, Freitas-Junior *et al.*, 2000]. It has been suggested that the effect of PfSir2A propagates from the telomere and does not extend to centrally located genes, which may be controlled by a different mechanism.

A second layer of regulation is related to subnuclear positioning. Several studies indicate that the active gene relocates to a unique site in the nuclear periphery that contains transcriptionally active non-condensed chromatin [Duraisingh *et al.*, 2005, Marty *et al.*, 2006, Ralph *et al.*, 2005, Voss *et al.*, 2007]. This expression site can however accommodate more than one active *var* gene, so it cannot be solely responsible for exclusive transcription [Dzikowski *et al.*, 2007]. Finally, there is evidence that a *var* intron sequence may be involved in silencing and in controlling mutually exclusive expression [Calderwood *et al.*, 2003, Deitsch *et al.*, 2001, Dzikowski *et al.*, 2007, Frank *et al.*, 2006, Swamy *et al.*, 2011], though this factor does not appear to be absolutely necessary and its significance has yet to be established [Ralph *et al.*, 2005, Voss *et al.*, 2006].

Major questions about this system remain unanswered. Most importantly, it is unknown how the parasite maintains exclusive transcription of exactly one *var* gene, and what factors control transcriptional switching from one gene to another. A recent study found that mutually exclusive transcription does not appear to involve long-range interaction with a unique enhancer element [Lemieux *et al.*, 2013]. One alternative, untested theory is that a necessary transcription factor is expressed at very low levels and accumulates around a single gene due to a positive feedback mechanism [Guizetti & Scherf, 2013].

1.2 Mathematical approaches

Mathematical modelling in malaria has an unusually long history. The epidemiological models created by Ronald Ross in the early twentieth century, and developed by George MacDonald, proved invaluable to the design of effective campaigns targeting mosquito vectors [Smith *et al.*, 2012]. This framework continues to provide the basis of many modern models of the transmission of *P. falciparum* and other vector-borne pathogens. In recent years mathematical modelling has also been used to test and generate hypotheses related to antigenic diversity and immune evasion. The growth in biological understanding of these phenomena has coincided with an increase in computer processing power, enabling mathematical modellers to investigate many-dimensional systems and the interactions of multiple processes. Computational approaches have also facilitated parameter estimation in these large and complex systems. What follows is an overview of some of the more recent developments in mathematical approaches to understanding antigenic variation and diversity of PfEMP1.

1.2.1 Models of antigenic variation

Antigenic variation is subject to opposing evolutionary forces. To prevent early exhaustion of the repertoire during an infection, at any one time only a small minority of antigenic variants should be abundant enough to trigger immune responses [Krause *et al.*, 2007]. On the other hand, the process must generate antigenic diversity fast enough to maintain a transmissible population as the host acquires immunity, or to establish infection of a partially immune host.

Before antigenic variation was described in *P. falciparum*, mathematical mod-

elling was used to investigate a similar system in the protozoan *Trypanosoma brucei*. Early results showed that growth rates alone were unable to generate the observed ordered appearance of antigenic variants in the infecting population [Kosinski, 1980]. An alternative model, which invoked a switching mechanism resulting in brief expression of two antigens per pathogen and differential susceptibility of these "double expressors" to immune responses, was able to mimic ordered expression but failed to find experimental support [Agur *et al.*, 1989]. More recently, Lythgoe *et al.* [2007] and Gjini *et al.* [2010] have shown that intrinsic gene activation rates, independent of the previously active gene, are sufficient to explain the ordered antigenic variation observed in acute *T. brucei* infection. However, the *T. brucei* antigenic repertoire is much larger than that of *P. falciparum*, and in the later stages of infection *T. brucei* relies instead on DNA rearrangement to create novel variants [Aitcheson *et al.*, 2005, Hall *et al.*, 2013, Marcello & Barry, 2007]. Moreover, trypanosomes only infrequently achieve reinfection. It is difficult to conceive of a *var* gene switching network that by itself could ensure the sequential dominance of 60 PfEMP1 variants during an infection, especially in a host likely to have had previous exposure to parts of the repertoire. Stochastic models incorporating multiple factors such as differential growth rates and double expressors have been made to simulate chronic *P. falciparum* infections with dynamics resembling clinical data [Molineaux *et al.*, 2001, Paget-McNicol *et al.*, 2002] but these models relied on questionable assumptions and yielded little biological insight.

Recker *et al.* [2004] proposed an alternative, deterministic model. It was assumed that each PfEMP1 variant contains a unique major epitope that elicits a long-lived immune response and several minor epitopes that trigger transient, cross-reactive immune responses. These assumptions proved sufficient to generate a sustained

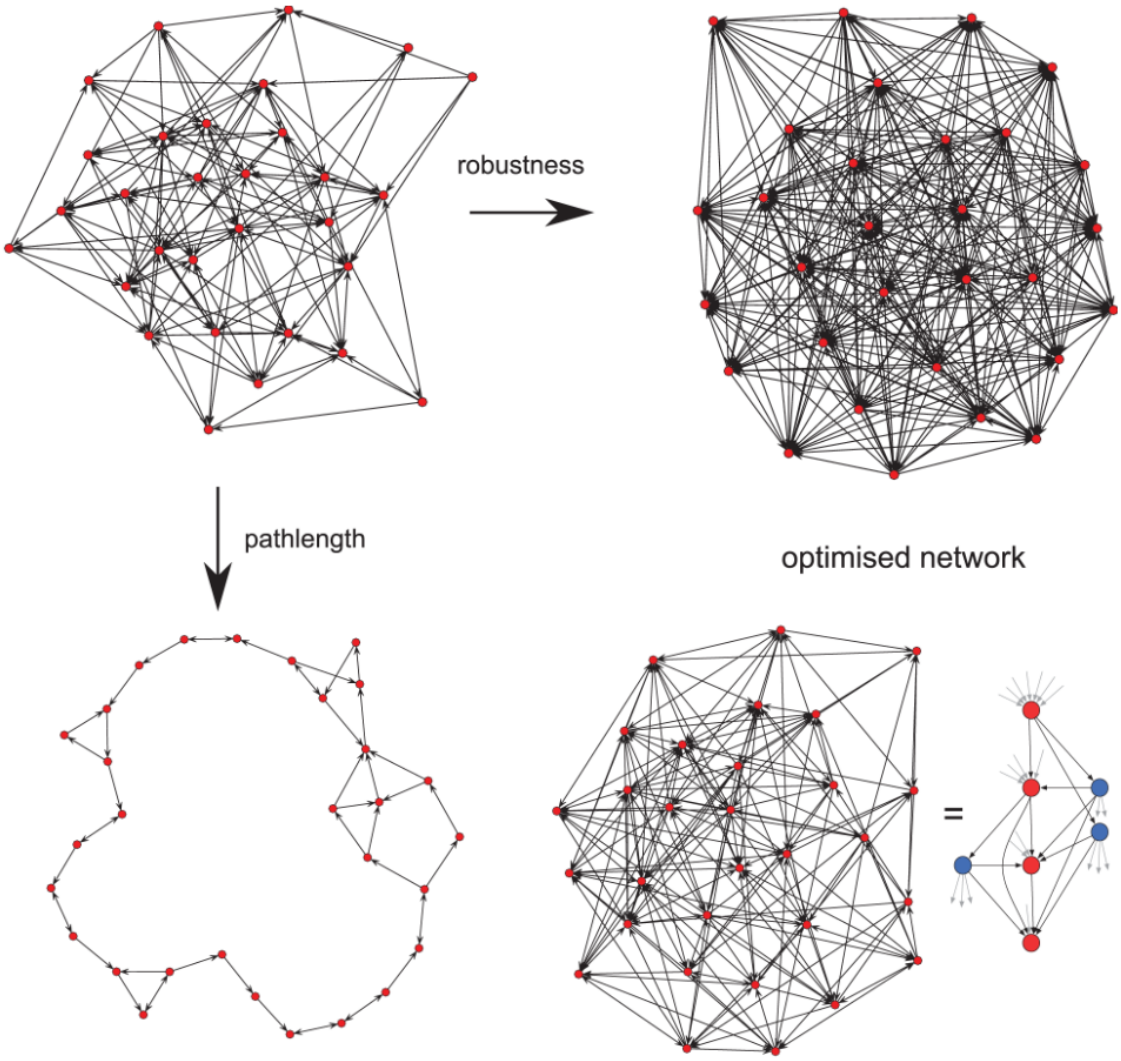


Figure 1.4. Antigenic switching networks derived from optimisation over evolutionary traits. This figure from Recker *et al.* [2011] shows networks describing switching among genes in an antigenic repertoire. An initially random network (top left) was evolved to optimise different evolutionary traits. In the first case (top right), selection depended on the robustness, defined as the ability to establish infection despite host immune responses. Alternatively, the network was optimised for pathlength, which minimised exposure of the repertoire to host immune responses (bottom left). Optimising for both traits led to a lattice-like network (bottom-right). Source nodes (shown in blue in the magnified network fragment on the right) switch to many other variants whereas sink nodes (red) are commonly switched to.

pattern of sequential dominance of variants, each antigenically dissimilar to its immediate predecessors. Nevertheless, the model could not reproduce early infection dynamics because of the time taken to establish the cross-reactive immune responses. The authors suggested that antigenic variation provides the necessary initial conditions by generating the ordered expression of a subset of variants. Recker *et al.* [2011] went on to consider what type of *var* switching pattern might satisfy the trade-off between prolonging infection and achieving reinfection. They showed that, according to the cross-immunity model, the optimal antigenic switching network comprised "source" nodes, which switch frequently, and "sink" nodes, which are frequently switched to. An example of such a network is illustrated in Figure 1.4.

1.2.2 Estimating switching parameters

Although the mechanism of transcriptional switching has not been identified, current understanding is consistent with *var* genes having different rates of activation and inactivation dependent on factors such as chromosomal location and Ups type [Guizetti & Scherf, 2013]. Determining these parameters and their correlates is important for understanding how infections are sustained for long periods, and how parasites establish infections in partially immune hosts. Because the distribution of PfEMP1 variants *in vivo* is determined by immune pressure and other host factors as well as antigenic variation, most studies of *var* gene switching dynamics have looked at parasite cultures. Roberts *et al.* [1992] made the first such investigation and observed an initially isogenic clone switching to various other antigenic types at a rate of approximately 2.4% per parasite generation. Twelve years later Horrocks *et al.* [2004] showed that this rate of switching from active to inactive – known as the "off-rate" – is intrinsic to each *var* gene, and rates within a repertoire are highly

dissimilar. This study also provided evidence of gene activation biases. Among 38 isogenic clones isolated from the same parent strain, 22 predominantly transcribed the same gene, *var* type 6. Six clones, each mainly transcribing one of four other genes, were cultured for 40 parasite generations. Transcription of *var* type 6 grew at a similar rate in all but two clones, both of which had initially transcribed *var* type 41. Flow cytometry analysis confirmed that the corresponding PfEMP1 variant accumulated at the same rate as the transcript. The authors concluded that *var* genes have diverse, intrinsic activation rates, but some switches may be disallowed, such as from type 41 to type 6.

Using a different parasite isolate and an alternative assay (quantitative PCR instead of northern blotting and flow cytometry), Frank *et al.* [2007] looked for correlates of the switching parameters. They investigated a sample of ten *var* genes and concluded that genes positioned in the central regions of chromosomes tend to have lower off-rates than those occupying subtelomeric loci. They proposed that this could explain why transcription of Group A genes declines rapidly during adaptation to laboratory culture [Gatton *et al.*, 2003, Peters *et al.*, 2002, 2007, Zhang *et al.*, 2011]. The study did not include a detailed analysis of the direction of switching but, based on a small number of observations, concluded that this depended on which gene was switching off.

These three studies primarily focused on determining off-rates. They estimated each off-rate by obtaining an isogenic clone and measuring the logarithmic decline of the initially dominant *var* gene transcript or PfEMP1 variant.[1] Horrocks *et al.* [2004] and Frank *et al.* [2007] also calculated so-called "on-rates" by dividing the

[1]Frank and colleagues measured the linear gradient instead of the logarithmic gradient. If they had used the standard formula then their estimates would have been higher, especially for the larger off-rates, but the correlation with chromosomal position would still have held.

increase in the proportion of parasites transcribing a gene by the period between measurements (i.e. the gradient). This measurement has two potential problems. Firstly, since the proportions are initially increasing and are bounded, their gradients must vary over time. Secondly, the on-rate at a point in time depends on the relative transcription levels and off-rates of all the other genes in the repertoire, as well as the off-rate of the gene in question. Recker *et al.* [2011] (following Frank [1999]) proposed instead defining two independent sets of parameters:

- **Off-rate:** the rate at which an active gene becomes inactive.
- **Switch bias:** the switch bias from gene i to gene j is the probability that when i becomes inactive, j becomes active.

Assuming that these are the most important parameters and that they are constant, and noting that parasite generations are synchronised during *in vitro* experiments, antigenic variation can be described by a recurrence relation:

$$m_{i,c,t+1} = (1 - \omega_i)m_{i,c,t} + \sum_{i \neq j} \omega_j \beta_{j,i} m_{j,c,t} \; \forall i \in \{1, \ldots, n\}, \tag{1.1}$$

where $m_{i,c,t}$ is the proportion of parasites in culture c transcribing gene i at generation t, ω_i is the off-rate of gene i, $\beta_{j,i}$ is the switch bias from gene j to gene i, and n is the number of transcripts.

These parameters provide a much clearer insight into the processes controlling antigenic variation. Determining their values should reveal whether, for example, switching is more likely between genes with similar features such as chromosomal location or upstream sequence type. The drawback is that these values are difficult to estimate. One approach is to run simulations using Equation 1.1, testing numerous

combinations of off-rates and switch biases to find which gives the best fit to a data set. A major challenge to this enterprise is the large sample space which makes simple methods unreliable or impractical. For instance, a system of ten genes has 90 parameters; a grid search testing five values for each parameter would require 8.1×10^{62} simulations and would likely miss the globally optimal solution by a wide margin.

Recker *et al.* [2011] addressed this problem using an iterative algorithm. Starting with a random parameter set they applied small perturbations until they found a set that gave a better fit to the data; this set then replaced the original. Repeating the process led them to a locally optimal solution. This "downhill" optimisation method was applied to new data from several cultures and to data from the previously discussed study by Frank *et al.* [2007], which used a different parasite genotype. Because of difficulties in measuring very low transcription levels, the analysis was restricted to the twelve most transcribed genes in each data set. The results indicated that *var* genes have very different sets of switch biases. Subsets of genes were found to have high switch biases toward the same gene, contributing to a highly structured switching network as illustrated in Figure 1.5. This network has similar properties to the network shown in Figure 1.4 that was derived from theoretical modelling. A limitation of the method used in this analysis is that there is no guarantee that a locally optimal parameter set will resemble the true parameter values. Nevertheless, it was shown that the solutions were largely independent of the optimisation algorithm's random starting point.

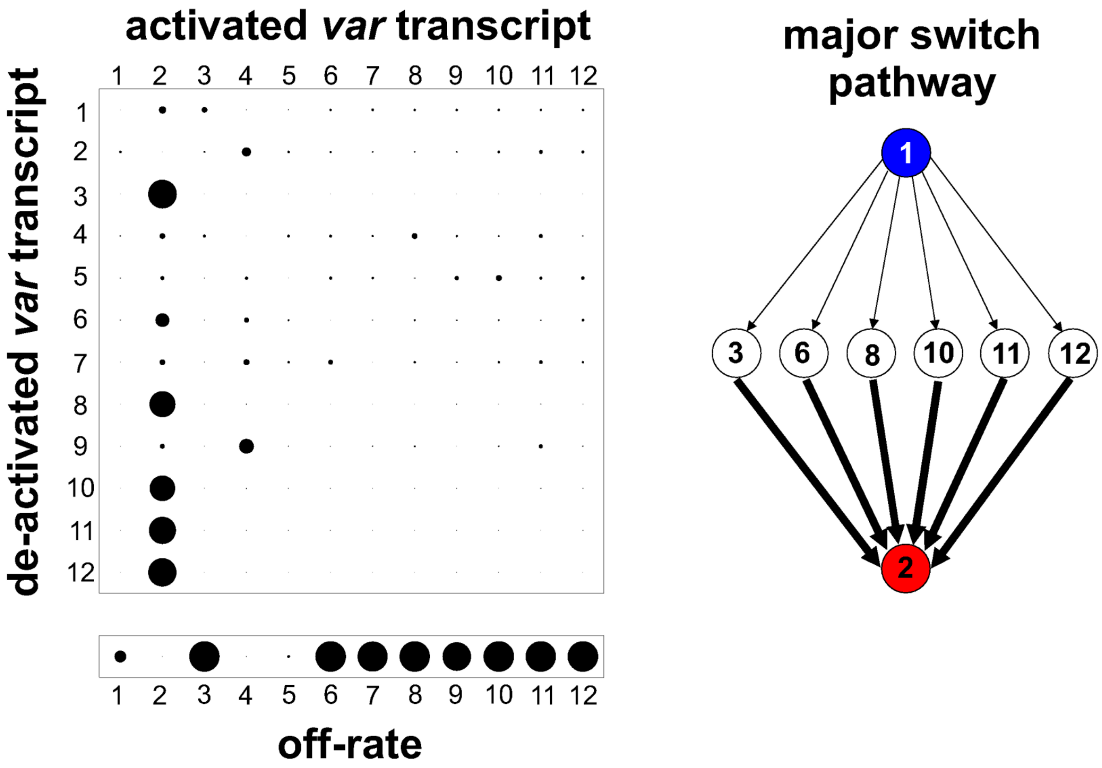


Figure 1.5. A *var* gene switching network derived from transcription data. Left: An example of a set of *var* gene switching parameter estimates produced by the method of Recker *et al.* [2011]. The size of the switch bias from gene i to gene j is indicated by the diameter of the disc in matrix row i and column j. The switch bias is the probability that when gene i is deactivated, gene j will be activated. Off-rates (the rates at which genes deactivate) are shown similarly, with the largest disc representing an off-rate of 6% per generation. Right: Part of the switching network for the same parameter set. In this case the network has a "single-many-single" structure, meaning that genes fall into two distinct sets. Genes in the first set (source nodes, such as genes 3, 6 and 8 in this example) switch at a high rate towards one or a small number of genes in the second set (sink nodes, such as gene 2). Genes in the second set switch much less frequently.

1.2.3 Models of diversity

Several recent modelling studies have focussed on understanding diversity within and between *var* repertoires, which is particularly important for predicting the impact of future vaccines targeting particular antigenic types. This research has been informed by what is known as strain theory. In the first formulation of strain theory [Gupta *et al.*, 1998, 1996], strains were defined by a small number of immunologically dominant loci. For example, in a pathogen with two loci, each having two alleles, there will be four possible strains ax, ay, bx and by, where a and b are alleles at the first locus, and x and y are alleles at the second locus. The model assumed that infection with a strain, such as ax, conferred complete immunity to that specific strain and at least partial immunity to all strains with shared alleles (in this example, ay and bx). It was shown analytically that host immunity will then structure the pathogen population into strains with non-overlapping combinations of alleles (such as ax and by, or ay and bx) to minimise the effects of cross-immunity [Gupta *et al.*, 1998, 1996].

Strain theory emphasises the role of shared antigenic types. Recker *et al.* [2008] investigated how such antigenic overlap might affect the acquisition of immunity to severe malaria. Mathematical analysis and simple stochastic simulations were used to test the consequences of *var* repertoire partitioning, such that each partition comprises genes from only one of several non-recombining groups. It was shown that repertoire partitioning may limit total antigenic diversity and reduce the probability that a host will be infected by a strain with an entirely unfamiliar set of antigens. Next, it was assumed that only one variant group causes severe malaria, and that the risk of developing severe disease depends on infection by a strain containing only severe disease variants to which the host has had no previous exposure. Under this

assumption, it was demonstrated that the development of resistance to severe disease could be very rapid due to overlap between repertoires. Structuring the repertoire into non-overlapping strains maximised the number of infections needed to acquire immunity to severe malaria. The authors speculated that the relative sizes of the *var* groups satisfy evolutionary trade-offs between disease severity, infection length and host adaptation.

To further understanding of *var* repertoire diversity, Buckee *et al.* [2009] simulated the acquisition of immunity in a host population according to competing hypotheses. The model results were compared with serological data from children in a Kenyan district with high malaria transmission intensity. It was found that the data were most consistent with a model in which each *var* repertoire was partitioned into two sets of genes with different strain structures. The first set was selected, as a unit, from a small number of non-overlapping strains, while the other set was randomly generated from a global pool containing hundreds of genes. The model fitted the data best when genes were expressed in order and those belonging to the more restricted set – which would seem to correspond to Group *var* A genes – were expressed earlier in infection. In accordance with strain theory, the authors suggested that in high transmission regions, strain structuring of the Group A repertoire population could result from immune selection since these genes are more frequently expressed and are highly immunogenic.

As explained above, strain theory was originally formulated for pathogens with a small number of immunodominant loci. Artzy-Randrup *et al.* [2012] used an agent-based mathematical model to investigate whether, as suggested by Buckee *et al.* [2009], immune selection might similarly shape population structure in *P. falciparum*, which has a large number of antigenic variants not all expressed at the

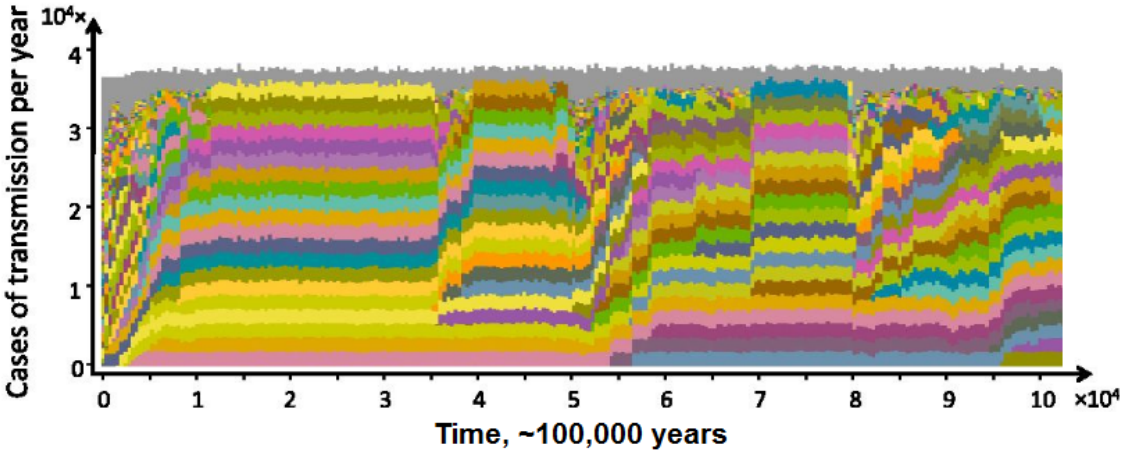


Figure 1.6. Strain structure in a *P. falciparum* model. This figure adapted from Artzy-Randrup *et al.* [2012] shows the frequencies with which pathogen strains were transmitted among hosts over a long time period in an agent-based mathematical model. Coloured bands represent individual strains with distinct antigenic repertoires. All repertoires sampled fewer than 50 times are shown together in grey. Each repertoire comprised 6 genes from a pool of 120 variants. Host immunity structured most of the pathogen population into a small number of strains with very little overlap between their repertoires despite frequent mutation and repertoire recombination.

same time. The model defined pathogen strains on the basis of distinct antigenic repertoires. The entire repertoire was sequentially expressed during infection and hosts acquired specific, lifelong immunity to each expressed variant. It was shown that, despite frequent mutation and repertoire recombination, the pathogen population became structured into a small number of strains with minimal overlap between repertoires (Figure 1.6), as predicted by strain theory.

Strain theory further predicts that pathogens will evolve to reduce cross immunity by minimising the number of immunogenic loci exposed to the immune system. This is apparently inconsistent with the large number of binding domains in many *var* genes, each of which provides additional targets. Buckee & Recker [2012] hypothesised that the accumulation of many domains into single genes could be advantageous if immune responses are partly domain-specific. They created a mathematical model of *var* gene evolution assuming that multi-domain genes retain some binding ability despite host immunity to some (but not all) of their constituent domains. The proportions of single- and multi-domain genes in the evolved population were found to depend strongly on the relative strengths of gene-specific and domain-specific immune responses. Buckee & Recker [2012] went on to consider two models in which *var* gene evolution is subject to a trade-off between maximising within-host fitness and minimising between-host immune pressure. The first model included two types of binding domain, one of which conferred longer infection lengths but was also subject to a higher level of type-specific cross-immunity. It was assumed that domains within a gene act synergistically in determining infection length. These conditions led to the evolution of genes comprising only diverse domains or only conserved domains, with very few mixed architectures. The second model considered the evolution of repertoire structure, assuming that genes are

exclusively expressed in order of within-host fitness. In this case, most repertoires evolved mixed structures containing both conserved and diverse genes. Both results are consistent with what has been observed in *var* repertoires.

1.3 Types of mathematical model

As in the physical sciences, mathematical models in biology are frequently formulated as sets of deterministic ordinary differential equations or recurrence relations. These state variable models assume that population dynamics can be described by assigning individuals (e.g. host organisms in an epidemiological model, or pathogens in a within-host model) to groups with shared characteristics, ignoring individual variation within groups. For more than a century, models of this type have proven useful in understanding diverse epidemiological, evolutionary and ecological systems. They are generally simple to describe, straightforward to replicate, and in some cases can lead to useful analytical solutions. For example, the Ross-McDonald model of malaria transmission, comprising two differential equations, leads to an algebraic expression for the basic reproduction number that clearly shows why some types of intervention are predicted to be more effective than others at reducing malaria prevalence [Smith *et al.*, 2012]. Applied to low-dimensional multi-locus systems, differential equation models can yield precise descriptions of the conditions under which stable or chaotic dynamics are expected [Gupta *et al.*, 1998]. In terms of implementation, state variable models are relatively efficient when the population is large and the number of possible states is small. With the exception of Paget-McNicol *et al.* [2002], all of the within-host modelling studies described in this chapter employ differential equations or recurrence relations, and this is the approach taken to

modelling large pathogen populations in Chapters 2 and 3 of this thesis.

An alternative approach known as individual-based modelling (or agent-based modelling) has rapidly expanded in popularity since the mid-1980s, facilitated by the growth of computer processing power. In an individual-based model (IBM), each member of the population is represented as a discrete entity with characteristics that may change through (usually stochastic) interactions with other individuals or the environment [DeAngelis & Mooij, 2005]. A researcher may choose to use an IBM either because they suspect that a state variable model would fail to capture important dynamics of a system, or for pragmatic reasons [Grimm, 1999]. In the study of *var* gene diversity, for example, the large dimensionality of the system presents a challenge to state variable approaches. When Buckee & Recker [2012] used a differential equation model to investigate the evolution of multi-domain *var* genes, they drastically reduced dimensionality by assuming only gene per parasite repertoire. Later in the same study when they turned to modelling the evolution of multi-gene repertoires, the authors instead used an IBM to allow them to represent wider antigenic and phenotypic diversity. Differential equations have been used to model *var* repertoire evolution in a multi-level (i.e. within-host and between-host) model but only by assuming that each repertoire comprises a small number of blocks of genes expressed in a fixed order, ignoring any variation within these blocks [van Noort *et al.*, 2010]. Models with larger repertoires and/or more complex within-host dynamics have necessarily been IBMs [Artzy-Randrup *et al.*, 2012, Buckee *et al.*, 2009, Severins *et al.*, 2012].

To illustrate a pragmatic choice of IBM, consider the previously described model of Artzy-Randrup *et al.* [2012], which includes between 10 and 500 antigenic variants and 2-30 variants per parasite repertoire. Even for the smallest parameter values

there are $10!/(10-2)! = 90$ possible repertoires, so that a state variable model would require 90 equations to keep track of the classes of infected hosts (ignoring superinfection). Moreover, to achieve its aims the model must also track host immune memories, which have an additional $\sum_{r=0}^{10} 10!/r!(10-r)! = 1,024$ possible states. There are then more than 90,000 classes of hosts – a number that is similar to the size of the host population – and more than 90,000 linked differential equations to solve. A state variable approach is clearly highly inefficient in this case. Instead, the authors opted for an IBM, storing each host's immune memory and the antigenic repertoires of their parasitic infections, and updating these characteristics only as a result of discrete mutation, transmission, recovery, birth and death events. A further benefit of the IBM approach illustrated by this example is that additional details – such as profiling of hosts by age, or the incorporation of various forms of superinfection – can be implemented without greatly increasing model size and complexity. The models in Chapter 4 of this thesis have similarly large dimensions and are therefore also formulated as IBMs for pragmatic reasons. The model of Chapter 5, which includes a further layer of genetic sequence variability, would be all but impossible to implement except as an IBM.

1.4 Thesis outline

It has been explained how the *P. falciparum var* gene family has important roles in immune evasion, within-host survival, and virulence. Therefore none of these aspects can be fully understood in isolation. The aim of this thesis is to provide insight into the evolution of *var* genes and repertoires in response to multiple, possibly conflicting selection pressures.

An obstacle to this endeavour is that the pattern of antigenic variation has not been fully described, partly because it is challenging to estimate very many parameters from necessarily limited data. To address this issue, Chapter 2 describes a new statistical method for determining antigenic switching networks from *in vitro* transcription measurements. The reliability of the method is demonstrated using diverse artificial data sets, designed to resemble experimental data including measurement errors. The method's applicability is further shown using real data from a previous study of *var* switching. Consequent recommendations for experimental design are discussed.

Chapter 3 applies the method developed in Chapter 2 to a set of *var* transcription measurements from the HB3 *P. falciparum* isolate. This yields the first full description of an antigenic switching network in *P. falciparum*, revealing that switching is mainly governed by a hierarchy of gene activation biases. A link is found between activation bias and gene length, which is related to binding phenotype. This is consistent with the hypothesis that binding phenotype, as well as *var* switching, determines the order in which antigenic variants are expressed during infection. Further analysis identifies a correlation between *var* transcription and genetic sequence diversity, which helps motivate the theoretical model of Chapter 5.

Chapter 4 uses theoretical modelling to explore possible consequences of evolutionary trade-offs involving PfEMP1 binding phenotype and antigenic diversity. First, models are used to test competing hypotheses of why *var* repertoires appear to be consistently partitioned according to binding phenotype. Here it is found that cross immunity may be the dominant factor. Next, a similar model is used to investigate the evolution of *var* expression order. The results suggest that transmission fitness, linked to PfEMP1 binding phenotype, might select for a particular expres-

sion order. This finding leads to speculation about the evolution of virulence in *P. falciparum*.

In Chapter 5, theoretical modelling is further applied to investigate population diversity of *var* genes and repertoires. Unlike in previous models, a distinction is made between antigenic diversity and genetic diversity, and a role of immune selection in genetic sequence evolution is explicitly considered. These additional factors are found to have important effects on both sequence diversity and strain structure. The model provides possible explanations for empirical results reported in Chapter 3 and elsewhere.

Versions of Chapters 2 and 3 have been published as papers and the method described in Chapter 2 has been used in a third paper. These publications are included as appendices.

Chapter 2

A statistically rigorous method for determining antigenic switching networks

2.1 Introduction

For pathogenic organisms whose opportunities for transmission are episodic, rather than continuous, long infectious periods are crucial for successful transmission. Most notably vector-borne and sexually transmitted pathogens have evolved various strategies to increase their transmission potential by evading their hosts' immune responses. One of the most sophisticated strategies is antigenic variation whereby the pathogen repeatedly changes its antigens over the course of a single infection. Although the underlying mechanisms differ between pathogens, many involve a polymorphic family of genes encoding functionally similar but antigenically diverse variants (reviewed by Deitsch *et al.* [2009]). Of these genes, only one is

actively transcribed while all other genes are transcriptionally silent. During infection, the active gene can 'switch off', leading to the activation, or 'switching on', of a previously silent gene. It is believed that this mono-allelic gene expression pattern helps the pathogen to guard its available antigenic repertoire from the immune system. Furthermore, gene expression has to be coordinated across the whole parasite population, such that the majority of the population expresses the same gene at the same time, as the host would otherwise build up immunity to all variants early on and clear the infection prematurely. As a result, infections with antigenically variable pathogens are often characterised by successive waves of parasitemia that are sequentially dominated by one or only a few antigenic variants.

Some of the best studied organisms employing antigenic variation are African trypanosomes and *Plasmodium falciparum*. Various theoretical studies have concentrated on determining the underlying mechanisms responsible for the observed coordination in antigen presentation during infection by these pathogens. In particular, gene activation hierarchies or differences in growth rates have been put forward as potential drivers behind their characteristic infection dynamics [Agur *et al.*, 1989, Frank, 1999, Kosinski, 1980, Lythgoe *et al.*, 2007, Molineaux *et al.*, 2001, Paget-McNicol *et al.*, 2002]. Although it was found that parasite intrinsic factors could orchestrate the parasite population in the initial phases of infection, they are insufficient for maintaining sequential dominance of antigenic variants during the later, chronic stages of infections. Instead, immune selection via cross-reactive immune responses has been proposed to offer a more parsimonious solution to this problem, even in the absence of structured differences in switch or growth rates [Recker *et al.*, 2004]. Nevertheless, in vitro studies of malaria parasites have since shown that variant switching is non-random and partly gene specific. For example, Horrocks *et al.*

[2004] demonstrated that *var* genes, which encode the surface-expressed virulence factor PfEMP1 (*P. falciparum* erythrocyte membrane protein 1), switch on and off at different but constant rates during long term culture. Frank *et al.* [2007] also found that these rates differ widely between different genes and that centrally located genes appear to have lower switch rates than those in subtelomeric loci. This has recently been confirmed by Enderes *et al.* [2011] who also showed that these switch preferences appear to be independent of genetic background. Similarly, in *Trypanosoma brucei* it has been shown that switching between *vsg* genes, which encode the pathogen surface coat, is determined by a fixed hierarchy of activation probabilities that appear to depend on features of the genes' loci [Aitcheson *et al.*, 2005, Lythgoe *et al.*, 2007, Morrison *et al.*, 2005]. An expression hierarchy dependent on two particular extragenic elements has also been found to underlie antigenic variation in the spirochete *Borrelia hermsii* [Barbour *et al.*, 2006]. Such observations have led to the hypothesis that structured switching or switch hierarchies might be important for structuring the parasite population during the early stages of infection. On the one hand this would make it easier for the adaptive immune response to desynchronise the parasite population and set up a pattern whereby single variants can successively dominate the infection [Recker *et al.*, 2004]. On the other hand it would also help the parasite to establish an infection in individuals with pre-existing immune responses, as recently shown by Recker *et al.* [2011].

To distinguish between parasite-intrinsic switching and host mediated selection during infection it is important to have a thorough understanding of inherent switch patterns. One approach to investigate these patterns is to analyse longitudinal gene transcription data from *in vitro* cultured parasites in the absence of selection [Horrocks *et al.*, 2004, Roberts *et al.*, 1992]. Given sufficient transcription data, it

should theoretically be possible to determine the switching network of a complete gene repertoire. However, a major challenge is the large sample space which makes simple methods unreliable or impractical. Recker *et al.* [2011] addressed this problem using an iterative algorithm to determine the switching pattern of *P. falciparum var* genes, using data from clonal parasite populations followed over 60-80 generations, with *var* transcription levels measured at regular intervals using quantitative real-time PCR (qrt-PCR). Although this method could robustly determine major switch pathways, it was able to explore only a small subset of the many-dimensional parameter space. Furthermore, qrt-PCR data are subject to measurement error, especially when transcription levels are very low, and the algorithm could not provide an indication of the uncertainty of its parameter estimates.

Here is presented a statistically rigorous solution to this problem in the form of simulated annealing and Markov Chain Monte Carlo (MCMC) algorithms. Using diverse sets of artificially generated gene transcription time courses, it is shown that highly structured and complex switch pathways can be resolved reliably from relatively limited data. Further, it is demonstrated that although experimental noise can have a major effect on estimates, using transcription measurements from several populations of the same pathogen strain can significantly improve accuracy. The method described here can thus be used to resolve complex and high-dimensional switch patterns with high reliability and accuracy from limited and noisy data.

A modified version of this chapter, published as a paper in June 2012, is appended to the dissertation.

2.2 Methods

2.2.1 Dynamical model

The general aim of the method is to find a set of parameter values relating to antigenic switching, i.e. the switch rates and activation biases, that best fit temporal variant transcript distributions. The method of inferring antigenic switch pathways from gene transcription data is based on the approach of Recker *et al.* [2011]. It is assumed that clonal parasite populations are followed over an extended period of *in vitro* culture with relative transcript levels of all genes measured at various time points. Under the assumption that a variant's switch rate and biases are constant over time, the resulting transcription profiles can be described by the following time-discrete model:

$$v_{i,c}^{t+1} = (1 - \omega_i)\, v_{i,c}^{t} + \sum_{j \neq i} \omega_j \beta_{ji} v_{j,c}^{t},\ \forall i \in \{1, \ldots, n\},$$

where $v_{i,c}^{t}$ is the relative transcript level of variant i in culture c at time t, ω_i is the variant specific off-rate, and β_{ji} is the probability of a switch from variant j to variant i. The switch network or pathway underlying a measured change in transcription levels can thus be described by the combination of a switch matrix $\boldsymbol{\beta} = (\beta_{ij})$ and off-rate vector $\boldsymbol{\omega} = \omega_1, \ldots, \omega_n$. The task is then to find the parameter values for $\boldsymbol{\omega} = \omega_i$ and $\boldsymbol{\beta} = \beta_{ij}$ that minimise an error function defined by the deviation between the measured (or simulated) transcript level $v_{i,c}^{t}$ and the model outcome $m_{i,c}^{t}$:

$$E = \sqrt{\frac{1}{qpn} \sum_{c=1}^{q} \sum_{t=1}^{p} \sum_{i=1}^{n} \left(\log_2(v_{i,c}^{t}) - \log_2(m_{i,c}^{t})\right)^2},$$

where p is the number of time points where transcript levels were measured and q is the number of cultures.

It is assumed that off-rates do not exceed 6% per generation, which is consistent with experimental measurements [Aitcheson *et al.*, 2005, Barbour *et al.*, 2006, Frank *et al.*, 2007, Horrocks *et al.*, 2004, Morrison *et al.*, 2005, Roberts *et al.*, 1992].

2.2.2 Simulated annealing

Because of the large dimension of the system, iterative approaches are used to find solutions that minimise the deviation between the model output and the data. The first such approach is simulated annealing, a probabilistic optimisation algorithm that identifies locally optimal parameter sets [Kirkpatrick *et al.*, 1983]. The procedure starts with a random parameter set $B = (\beta, \omega)$ containing the switch biases and off-rates, and applies a random perturbation to produce B'. The move from B to B' is accepted if it reduces the error; otherwise the move is accepted with a probability dependent on how much it increases the error and on a "temperature" parameter. The perturbation and acceptance steps are repeated for many iterations while the temperature is gradually lowered so that moves that increase the error become less likely to be accepted. When the temperature becomes sufficiently low the parameter set converges to a local minimum.

Perturbation of a switch bias parameter β_{ji} is achieved by drawing a random variable β'_{ji} from a Cauchy distribution centred at β_{ji}. The distribution is restricted to the permitted range [0,1] by mapping

$$B_i \mapsto \begin{cases} \text{frac}(\beta_{ji}) & \text{if floor}(\beta_{ji}) \text{ is even} \\ 1-\text{frac}(\beta_{ji}) & \text{if floor}(\beta_{ji}) \text{ is odd,} \end{cases}$$

where floor(β_{ji}) is the largest integer not greater than β_{ji} and frac(β_{ji}) is the fractional part. Off-rates are similarly perturbed after rescaling in proportion to the parameter range [0,0.06]. In each iteration the perturbation procedure is applied to one switch bias per matrix row and the other entries are then rescaled to normalise the row. All off-rates are perturbed in each iteration. The Cauchy distribution is used here because it has a fatter tail than the Gaussian distribution and may therefore be more likely to generate "basin hopping" moves between local optima.

This perturbation procedure is symmetrical, meaning that the probability of moving from B to B' is always equal to the probability of moving from B' to B. However, normalising each row of switch biases causes the algorithm disproportionately to favour small parameter values. Such non-uniformity is permitted in simulated annealing and should not affect the results. The perturbation size is determined by the scale parameter of the Cauchy distribution, which is made constant.

The acceptance threshold uses the Fermi distribution; that is, proposed moves are accepted with probability $1/(1 + \exp\left(\frac{E'-E}{T}\right))$, where E and E' are the errors associated with the current and proposed parameter sets respectively, and T is the temperature. The algorithm follows a geometric cooling schedule and is stopped after 10 million iterations.

2.2.3 Experimental error

Gene transcript levels are calculated from qrt-PCR output using an exponential formula in which the base is the amplification efficiency of the gene transcript (which should be approximately 2) and the exponent is the number of amplification cycles required for the transcript to reach a threshold abundance [Pfaffl, 2001]. It is therefore assumed that error-prone measurements will follow a log-normal distribution

and that, at least for the most abundant transcripts, the standard deviation will be proportional to the mean. Accordingly, measurement errors are simulated by applying a noise function $\phi(v_{i,c,t}) = 2^R v_{i,c,t}$, with $R \sim N(0, \sigma^2)$ to the simulated transcription profiles and then renormalising each set $\{v_{1,c,t}, \ldots, v_{n,c,t}\}$. In most of the trials reported here, $\sigma = 1$, which assumes that transcript levels are typically measured to be two times larger or smaller than their true values.

2.2.4 MCMC

A different optimisation algorithm is used to obtain distributions of likely values for the switch biases and off-rates, instead of only the "best-fit" point estimates. This is a Markov chain Monte Carlo (MCMC) method known as the Metropolis-Hastings algorithm [Hastings, 1970, Metropolis *et al.*, 1953]. The algorithm accepts proposed moves with probability $L(B'|v)/L(B|v)$, which is the likelihood ratio of the proposed parameter set B' and the current parameter set B, given the observed data v. To convert the simulated annealing algorithm to instead perform MCMC, the acceptance threshold is replaced by the exact likelihood ratio, as derived below, and the simulated annealing error is used to estimate the noise parameter σ. The algorithm is run for 10 million iterations and every accepted solution is recorded after an initial "burn-in" period of 1 million iterations.

To simplify the notation, the following derivation of the likelihood function considers the case with only one culture and one time point, so $v = \{v_1, \ldots, v_n\}$ is the observed data and $m = \{m_1, \ldots, m_n\}$ is the output of the model with parameter set B. Assuming as before that each $v_i = 2^{r_i} m_i / \sum_{j=1}^{n} 2^{r_j} m_j$, where each r_i is an

instance of a random variable $R_i \sim N(0, \sigma^2)$, then

$$r_i - r_1 = \log_2(v_i m_1/(m_i v_1)),\ \forall i > 1$$

and

$$\begin{aligned}
L(B|v) &= P(V = v|B) \\
&= \int_{-\infty}^{\infty} P(R_1 = x) \prod_{i=1}^{n-1} P(R_i - R_1 = r_i - r_1 | R_1 = x)\, \mathrm{d}x \\
&= \int_{-\infty}^{\infty} P(R_1 = x) \prod_{i=1}^{n-1} P\left(R_i = \log_2\left(\frac{v_i m_1}{v_1 m_i}\right) + x\right) \mathrm{d}x \\
&\propto \int_{-\infty}^{\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right) \prod_{i=1}^{n-1} \exp\left(\frac{-(\log_2(v_i m_1/(m_i v_1)) + x)^2}{2\sigma^2}\right) \mathrm{d}x.
\end{aligned}$$

Changing the variable to $y = 2^x m_1/v_1$ gives

$$\begin{aligned}
L(B|v) &= \int_0^{\infty} \frac{1}{y} \prod_{i=1}^{n} \exp\left(\frac{-\log_2^2(y v_i/m_i)}{2\sigma^2}\right) \mathrm{d}y \\
&\propto \exp\left[\frac{1}{2\sigma^2}\left(-\sum_{i=1}^{n} \log_2^2\left(\frac{v_i}{m_i}\right) + \frac{1}{n}\log_2^2\left(\prod_{i=1}^{n}\frac{v_i}{m_i}\right)\right)\right].
\end{aligned}$$

It follows that the likelihood ratio for q cultures and p time points is

$$\begin{aligned}
\frac{L(B'|v)}{L(B|v)} &= \prod_{c=1}^{q}\prod_{t=1}^{p} \exp\left[\frac{1}{2\sigma^2}\left(-\sum_{i=1}^{n}\log_2^2\left(\frac{v_{i,c,t}}{m'_{i,c,t}}\right) + \sum_{i=1}^{n}\log_2^2\left(\frac{v_{i,c,t}}{m_{i,c,t}}\right)\right.\right. \\
&\qquad \left.\left. + \frac{1}{n}\log_2^2\left(\prod_{i=1}^{n}\frac{v_{i,c,t}}{m'_{i,c,t}}\right) - \frac{1}{n}\log_2^2\left(\prod_{i=1}^{n}\frac{v_{i,c,t}}{m_{i,c,t}}\right)\right)\right] \\
&= \exp\left[\frac{npq}{2\sigma^2}\left(\mathrm{Var}\left(\log_2\left(\frac{v}{m}\right)\right) - \mathrm{Var}\left(\log_2\left(\frac{v}{m'}\right)\right)\right)\right].
\end{aligned}$$

2.3 Testing the method

The method is first tested for accuracy and reliability in resolving antigenic switch patterns from gene transcription time courses using various test parameter sets representing a wide spectrum of possible switch pathways. That is, a number of artificial switch networks are constructed with different degrees of complexity, in terms of the number of genes that each gene switches to, and the number of consecutive switch events, and these are used to generate temporal gene transcription data. The optimisation methods are then applied to reconstruct the most likely switch pathways underlying the data.

In line with previous *var* gene transcription studies, it is assumed that the parasite population is initially clonal, with every parasite transcribing the same gene, and that relative transcript levels are measured at several time points during *in vitro* culture. A common feature of the previously described transcription profiles is that only a fraction of transcripts reach significant levels during *in vitro* culture [Enderes *et al.*, 2011, Frank *et al.*, 2007, Horrocks *et al.*, 2004, Recker *et al.*, 2011]. It has been argued that less transcribed genes are unlikely to play a dominant role in the switching network in the neighbourhood of the starting gene. Therefore it is assumed that removing data for genes generally transcribed at very low levels will have little effect on the parameter estimates for the interactions between more dominant genes, and that these interactions will be sufficient to determine the most important properties of the switching network. Accordingly, the initial analysis is restricted to a set of 10-16 genes, which could represent the subset of the most dominantly transcribed genes. A numerical justification for using such a reduced system to determine major transcription pathways is presented later in this chapter.

2.3.1 Resolving switch networks from a single transcription time series

Initial testing focuses on four switching networks of different complexities that comprise 10 genes. The first network describes a situation in which each gene simply switches to only one other gene with very high bias, and is referred to as a one-to-one (1:1) pathway. The second network, referred to as single-many-single (SMS), is similar to that proposed to underlie *var* gene transcription data in a previous study [Recker *et al.*, 2011]. It describes switching from the initial gene to a group of genes, which all switch with high bias to another single gene. The latter gene in turn switches to a different group of genes, which switch back to the first gene. The third network has a lattice-like structure containing block-diagonal switch biases, and the fourth is a uniform network, where each gene has identical switch biases. The four networks are illustrated in Figures 2.1 and 2.2 together with the matrix and vector representation of switch biases and off-rates, $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ respectively, and the simulated transcription time courses resulting from these networks of up to 60 generations post cloning.

In the first test, simulated annealing is used to try to recover the original switching parameters from the transcript levels measured without error at generations 20, 40 and 60. Figures 2.3 and 2.4 show that in each case the method finds a good fit between the target and simulated transcript distributions, and it reliably recovers the first four consecutive switch events of the 1:1, SMS and uniform target networks. The switch bias matrix and off-rate vector are less well resolved for the later stages of the 1:1 network (Figure 2.3A), which comprises ten consecutive switches, and for the lattice network (Figure 2.4A), which has an especially complex structure. Note

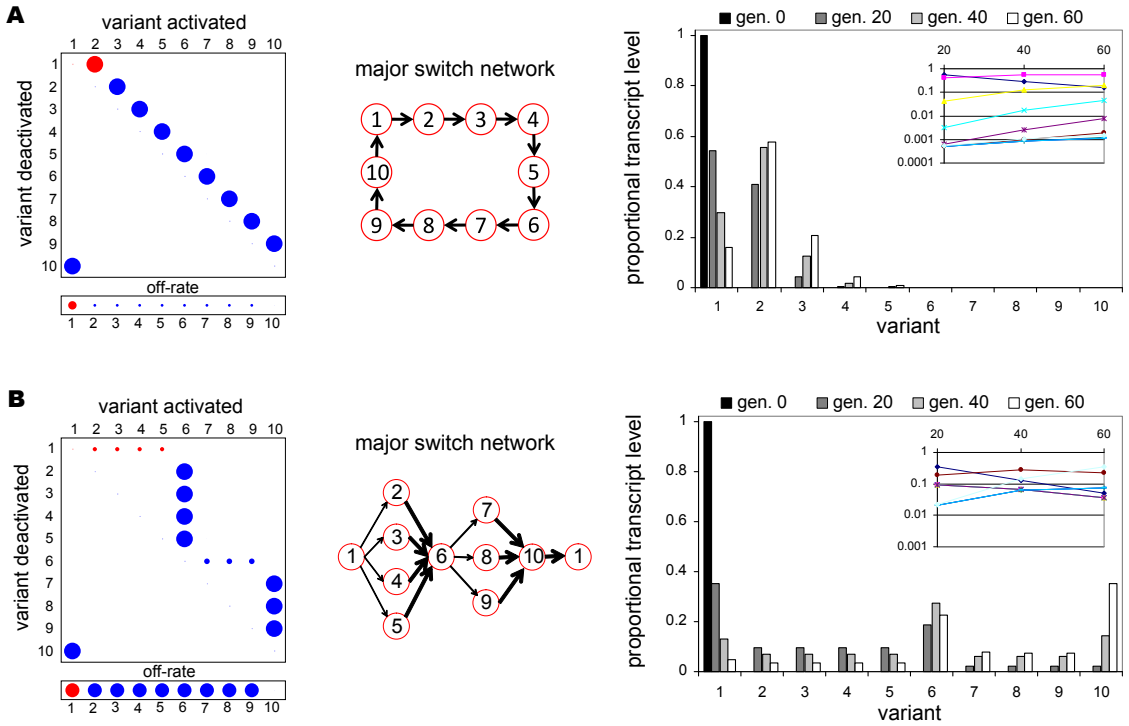


Figure 2.1. Target network and transcript data for initial testing: 1:1 and SMS networks. Here are shown two of the four switch networks of different complexities considered for the initial testing and method calibration: (**A**) one-to-one (1:1), (**B**) single-many-single (SMS). The sizes of the discs in the switch matrices correspond to the transition biases from variant i to variant j, β_{ij}, and the sizes of the discs in the off-rate vectors are proportional to the per-generation de-activation rates, ω_i. The major switch pathways described by these matrices are highlighted in the middle column and the right column shows the proportional transcription levels of all 10 variants from point of cloning until 60 generations post cloning, taken at 20 generation intervals; the insets depict the proportional transcript levels of the 10 variants on a log scale at generations 20, 40 and 60, with each colour representing a different variant.

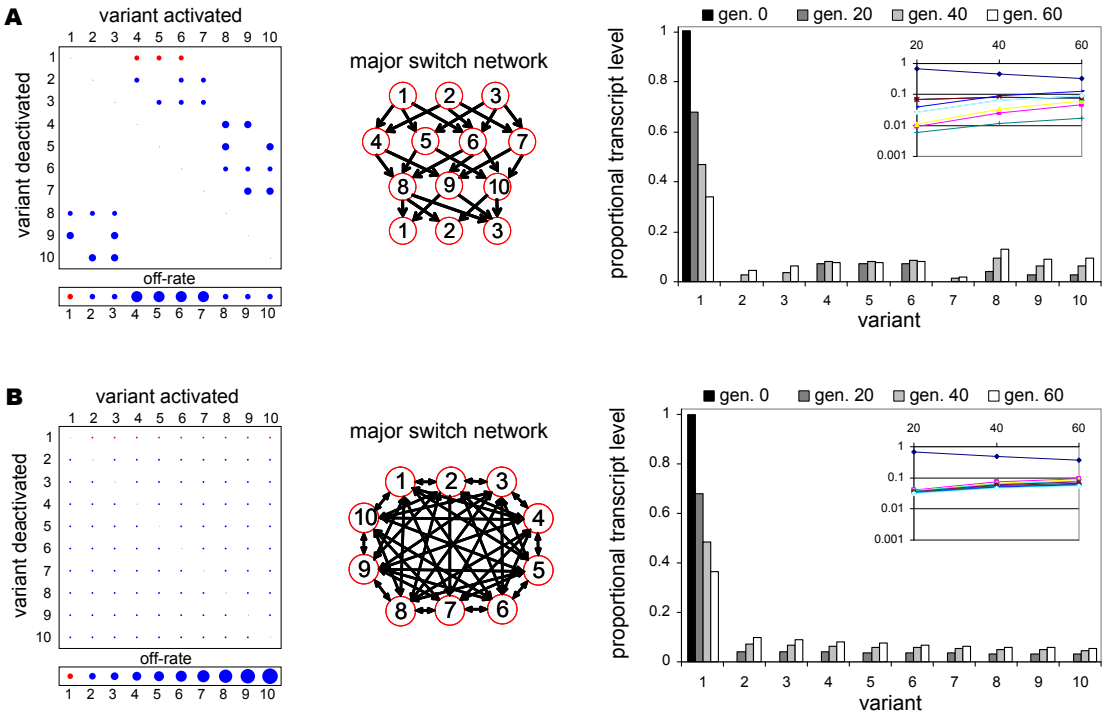


Figure 2.2. Target network and transcript data for initial testing: lattice and uniform networks. Here are shown another two of the four switch networks of different complexities considered for the initial testing and method calibration: (**A**) lattice, and (**B**) uniform. The sizes of the discs in the switch matrices correspond to the transition biases from variant i to variant j, β_{ij}, and the sizes of the discs in the off-rate vectors are proportional to the per-generation de-activation rates, ω_i. The major switch pathways described by these matrices are highlighted in the middle column and the right column shows the proportional transcription levels of all 10 variants from point of cloning until 60 generations post cloning, taken at 20 generation intervals; the insets depict the proportional transcript levels of the 10 variants on a log scale at generations 20, 40 and 60, with each colour representing a different variant.

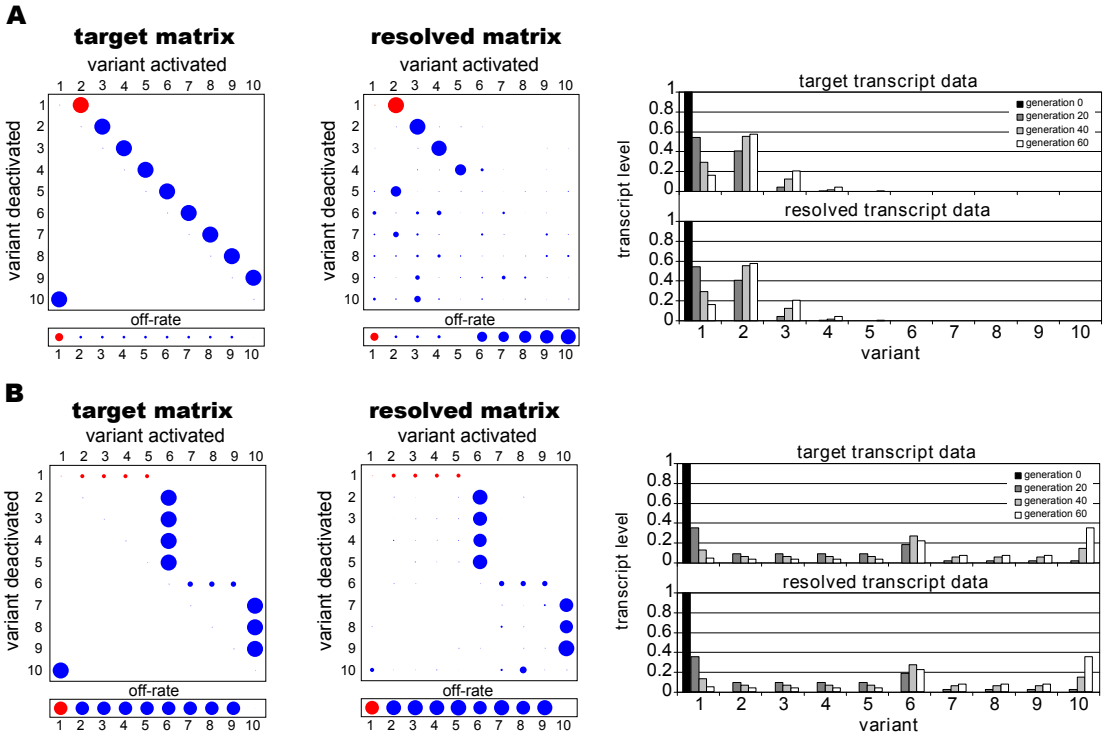


Figure 2.3. Resolved networks from a single time series: 1:1 and SMS networks. Using simulated annealing the (**A**) 1:1 and (**B**) SMS networks are resolved to up to four consecutive switches. The starter genes, which are assumed to be clonally transcribed at generation 0, are highlighted in red. The simulated time courses (right column) show a very good agreement between the target data and the transcription data resulting from the determined network.

that parameters associated with the later network stages, i.e. of those genes that are activated only after several switch events, have relatively minor effects on transcript levels and the simulated annealing algorithm is less sensitive to their variation. That is, even when the model output fits the transcript time courses of these variants very tightly (for all networks the average overall error, or deviation between the input data and the model output, was less than 0.1%), the particular pathways between rarely transcribed genes are less accurately described.

The next test examines the effect of noise, in terms of experimental error, on

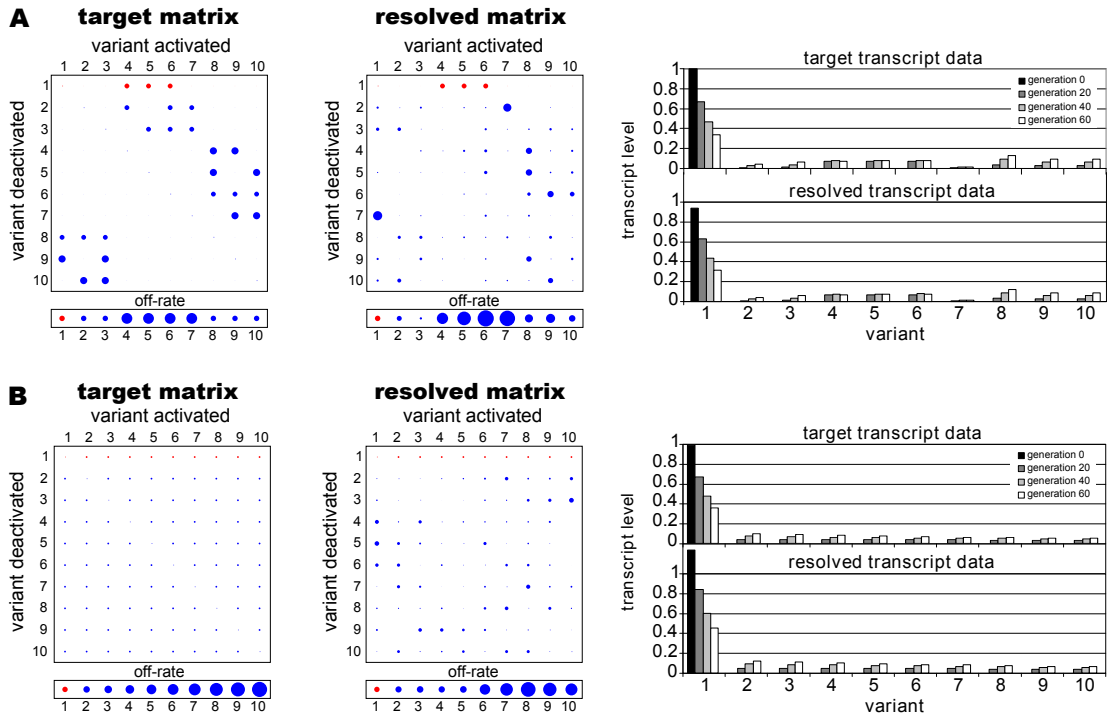


Figure 2.4. Resolved networks from a single time series: lattice and uniform networks. Using simulated annealing the (**B**) uniform network is resolved well, although the complex lattice network (**A**) is less well described. The starter genes, which are assumed to be clonally transcribed at generation 0, are highlighted in red. The simulated time courses (right column) show a very good agreement between the target data and the transcription data resulting from the determined network.

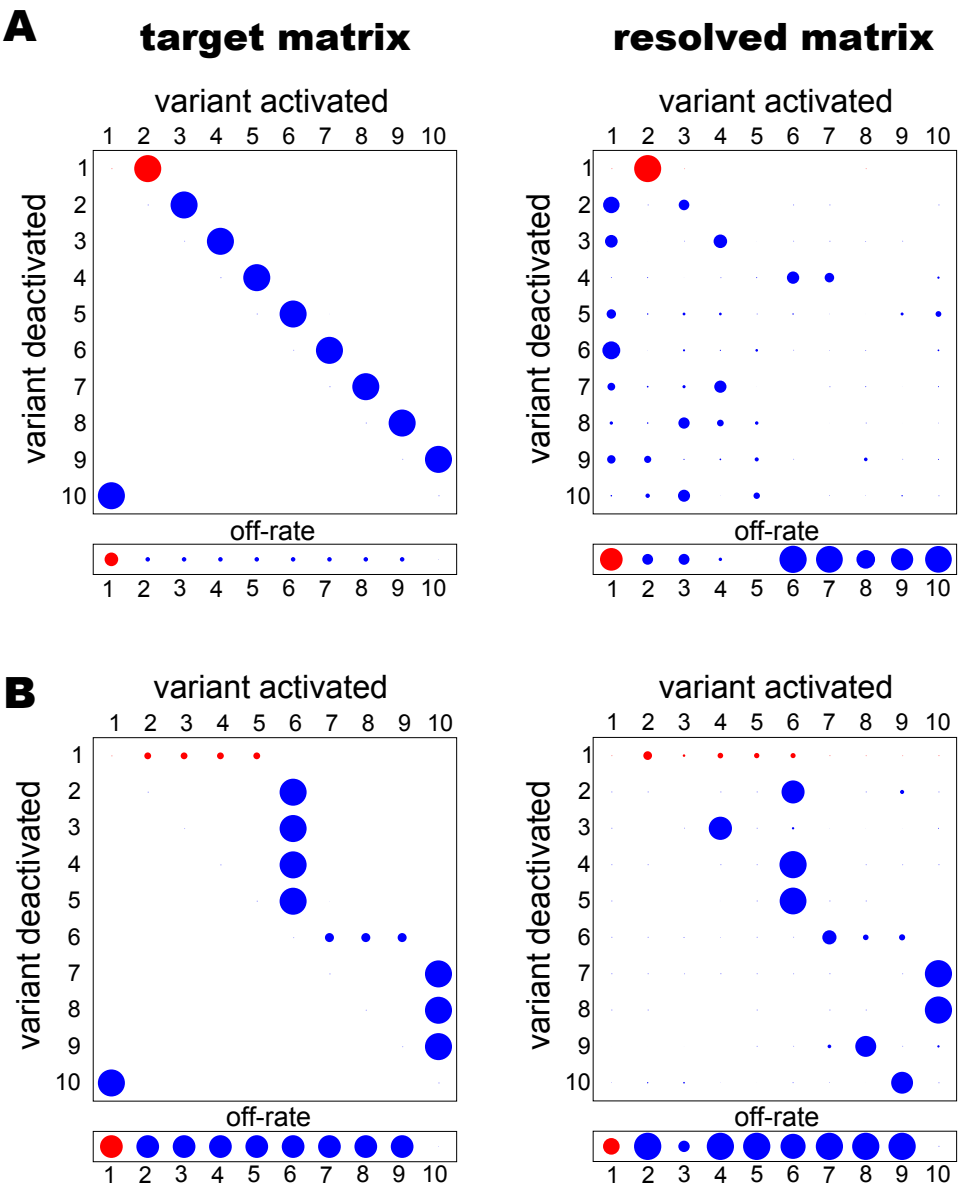


Figure 2.5. Resolved networks from noisy data: 1:1 and SMS networks. The accuracy in determining switch networks from single transcription time courses is significantly affected by the level of noise in the data (here $\sigma = 0.6$) and can lead to poorly resolved networks, as shown for both the (**A**) 1:1 and (**B**) SMS pathways.

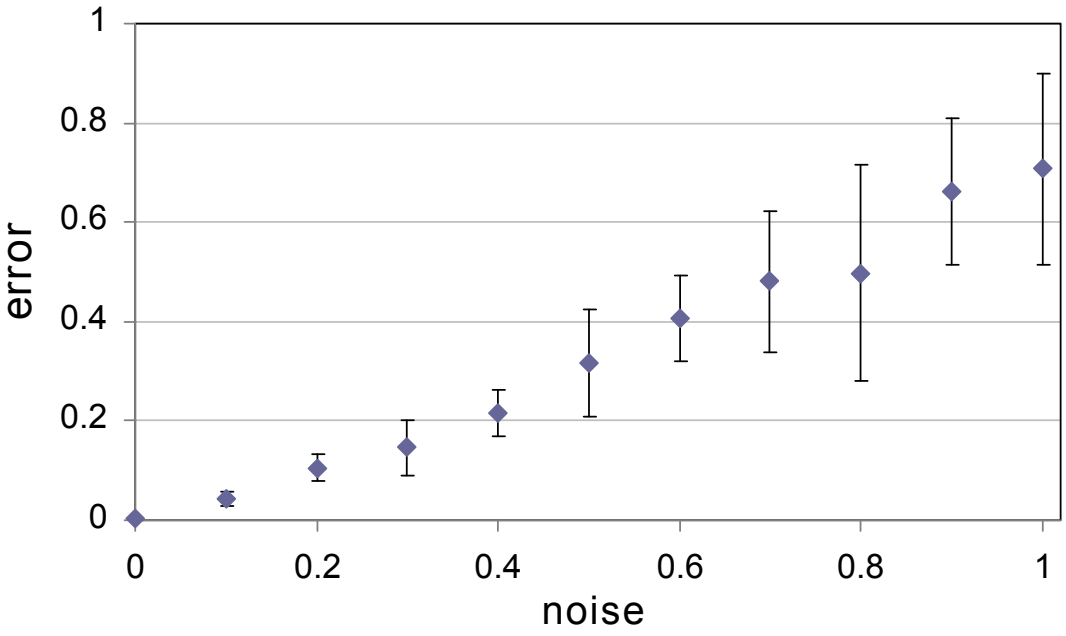


Figure 2.6. Linear correlation between the model-data error and the level of noise.

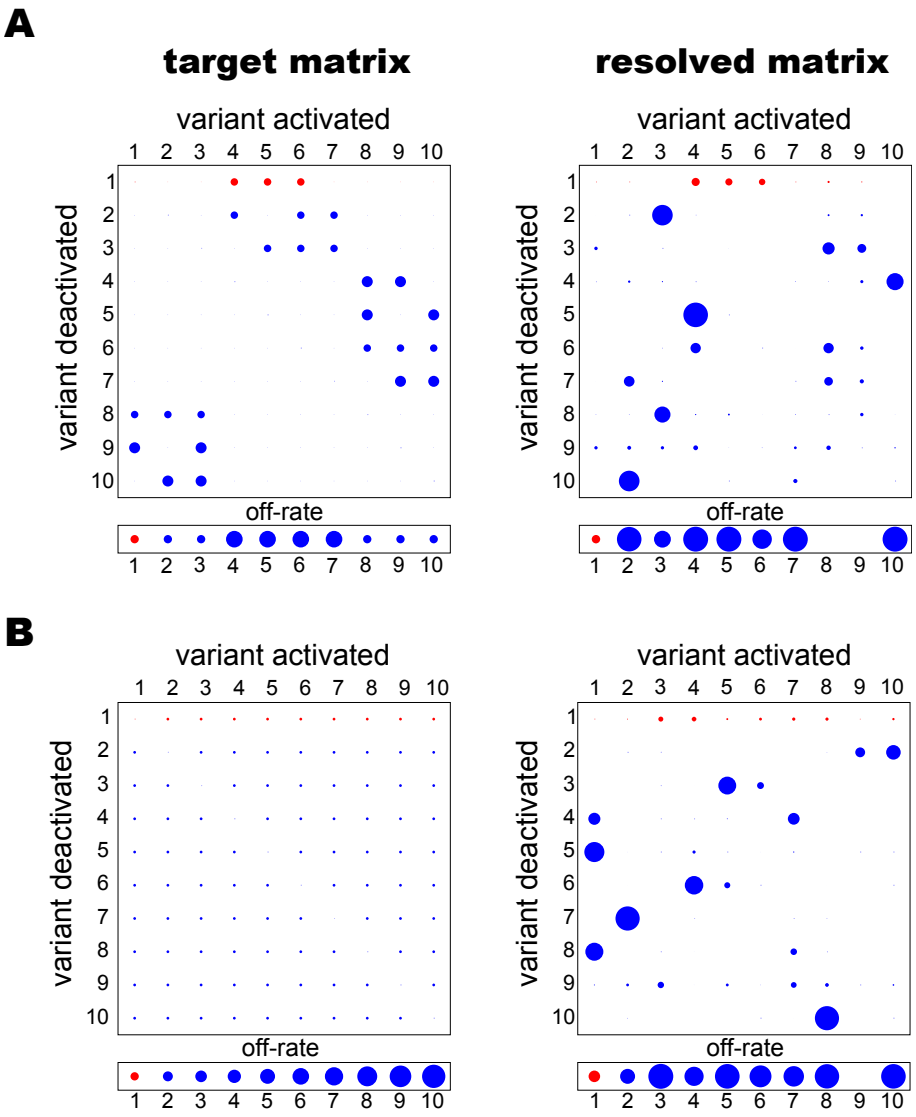


Figure 2.7. Resolved networks from noisy data: lattice and uniform networks. The accuracy in determining switch networks from single transcription time courses is significantly affected by the level of noise in the data (here $\sigma = 0.6$) and can lead to poorly resolved networks, as shown for both the (**A**) lattice and (**B**) uniform pathways.

the method's ability to recover underlying switch patterns. As shown in Figures 2.5 and 2.7, accuracy drops markedly when noise is added to the data. However, under moderate levels of noise, major switch pathways can still be determined with reasonable accuracy, dependent on the particular network structure. The mean error increases linearly with the noise level (Figure 2.6) but there is only a weak correlation between these two variables. It would therefore be impossible to use this error to estimate precisely the degree of noise in an experimental data set. Furthermore, there is a very weak correlation between the noise level and the error in the switch bias estimates (not shown), indicating that there would be wide variation in the accuracy of parameter estimates derived from experimental data of this type.

One possible way to improve the method's accuracy is to increase the number of time points at which transcript levels are measured. For the 1:1 network with noise added, changing from three to seven time points results in a 22% decrease in the root mean square deviation between the output switch biases and the target parameters for the first four network stages. More than doubling the number of measurements to 15 yields only a 16% additional improvement. This non-linear relationship between the number of time points and the deviation of the parameter estimates indicates that the method's accuracy in describing the entire switch pathways of a parasite's gene repertoire is not simply limited by the frequency of measurements.

2.3.2 Resolving switch networks from multiple transcription time series

Results so far show that the simulated annealing method resolves the underlying switch networks with limited resolution for those genes whose transcript levels re-

main low over the whole time course. This issue may be lessened by using several target data sets for each network. That is, the same global switch pathway is used to generate transcription time courses from different clonal starter populations, and the model is fitted to all of the data simultaneously. In tests of this approach, the size of the parasite's repertoire is increased to consist of 16 genes (requiring 240 parameters). Eight of these genes are randomly selected to be the clonally expressed starter genes for separate data sets, corresponding to eight clones taken from the same parent culture. Transcription levels are recorded for each data set at five time points at generations 20, 30, 40, 60 and 80 post cloning.

As shown in Figures 2.8 and 2.9 (middle columns), despite the high complexity of the underlying networks and size of the parameter space all four switch pathways are consistently recovered with high accuracy and resolution by using multiple time series simultaneously. Even when there is a high degree of experimental noise, the switch bias and off-rates are accurately predicted for the starter genes (highlighted in red). It should be noted that the level of noise considered here is considerably higher than what would be expected in experimental data and corresponds to transcription level measurements typically differing from the true values by a factor of two. Although the parameter estimates for non-starter genes (blue discs) are less reliable at moderate and high noise levels, they are considerably more accurate than would be expected by chance. Switch biases of non-starter genes are recovered most accurately for the 1:1 network and less accurately for the SMS and lattice network. This is not surprising, however, given the high degree of complexity of these particular switching networks. Furthermore, the noisy data are fitted more closely by the model predictions than by the true, i.e. noise-less, transcription levels (the error is on average 15% smaller, independent of network type or noise level), which confirms

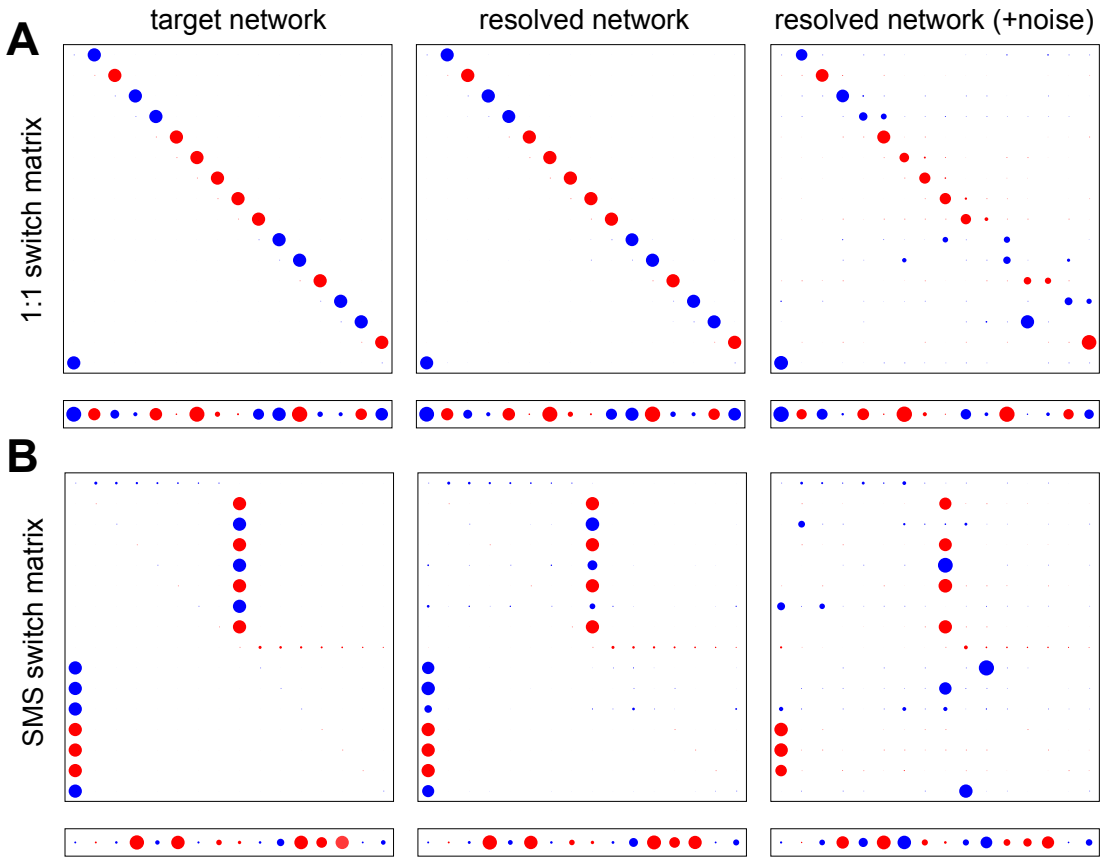


Figure 2.8. Resolved networks from multiple time series: 1:1 and SMS networks. Transcription histories from eight different clones, each defined by a different starting gene (highlighted in red), are used to resolve two different switch pathways describing 1:1 and SMS networks (**A** and **B**, respectively). Without any noise, both matrices and off-rate vectors can be resolved to a high degree of accuracy, even for the non-starter genes (middle column). The use of multiple data sets also yields better estimates when significant levels of noise are added to the data ($\sigma = 1$, right column).

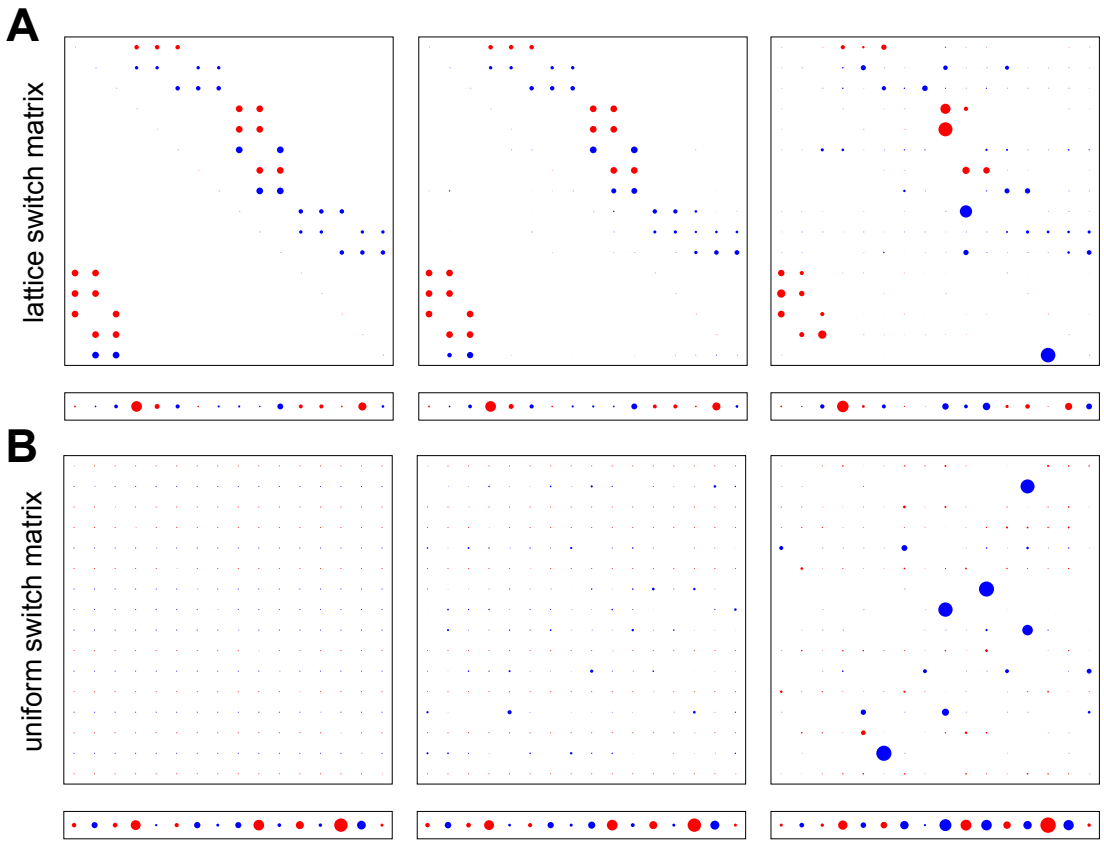


Figure 2.9. Resolved networks from multiple time series: lattice and uniform networks. Transcription histories from eight different clones, each defined by a different starting gene (highlighted in red), are used to resolve two different switch pathways describing lattice and uniform networks (**A** and **B**, respectively). Without any noise, both matrices and off-rate vectors can be resolved to a high degree of accuracy, even for the non-starter genes (middle column). The use of multiple data sets also yields better estimates when significant levels of noise are added to the data ($\sigma = 1$, right column).

that inaccuracies in the parameter estimates are mostly due to the noise and not to a failure of the simulated annealing method to locate good optima. Figures 2.10 and 2.11 show two examples of the transcription time courses of the 10 most highly transcribed genes from two different starter genes (2 and 6) of an underlying 1:1 switch pathway, with and without added noise, together with the transcript distributions resulting from the estimated switching networks. This again highlights the robustness of this method in determining genetic switch patterns from limited and noisy data.

Similarly to the case with only one transcript time series, the error increases linearly with the noise level but this time the two variables are strongly correlated (data not shown). The error value obtained from a sufficiently large set of experimental data would therefore be a good estimator of the degree of experimental noise in the data and could be used as an input parameter for determining the uncertainty of the resolved switching network using a Markov Chain Monte Carlo (MCMC) approach as described below.

2.3.3 MCMC

One explanation for the inaccuracy of the simulated annealing method in estimating the switch parameters for non-starter genes, and generally genes that remain at low transcript levels during the whole experiment, is that the error is less sensitive to their variation. This implies the existence of alternative switch networks that fit the target data almost as well as the predicted best-fit solution. To investigate the level of uncertainty and obtain a probability distribution for each parameter, a Markov Chain Monte Carlo (MCMC) method can be used instead of simulated annealing.

As demonstrated in Figure 2.12, using the same noisy transcription time series as

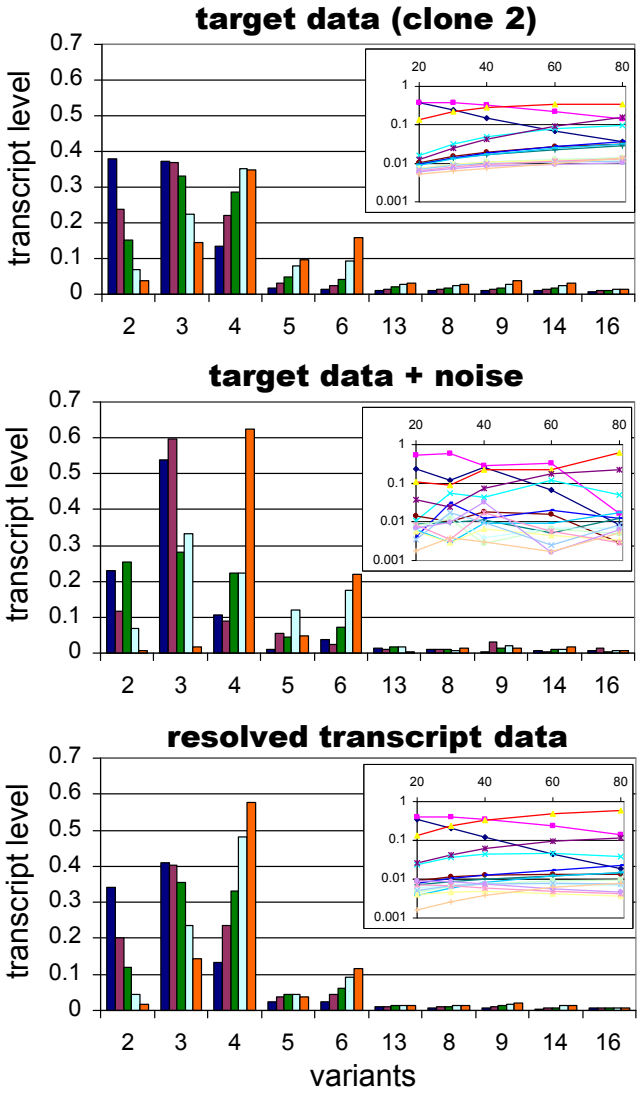


Figure 2.10. Resolved transcription time courses from noisy data. When the model is fitted to multiple data sets with different starting genes, the transcript levels of the predicted network (bottom) are more similar to the noiseless data of the underlying network (top left) than to the noisy data used as input (top right). Results are shown for one different clones of a 1:1 switch network with transcript levels measured at generation 20 (blue), 30 (purple), 40 (green), 60 (light blue) and 80 (orange). The insets depict the proportional transcript levels of the 16 variants on a log scale at generations 20, 30, 40, 60 and 80, with each colour representing a different variant.

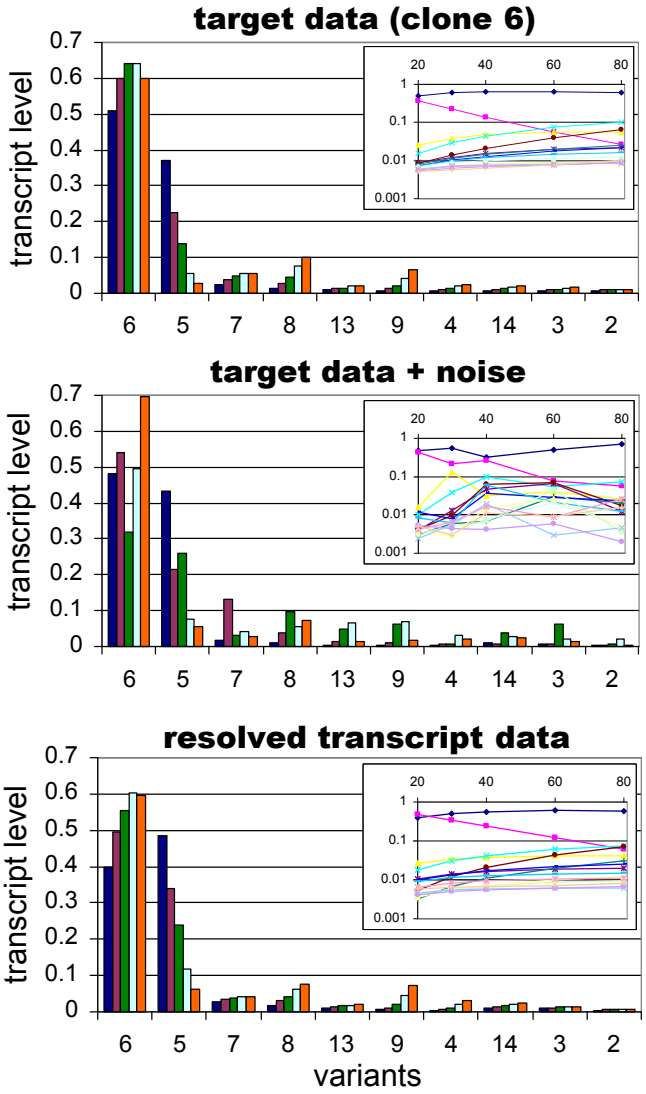
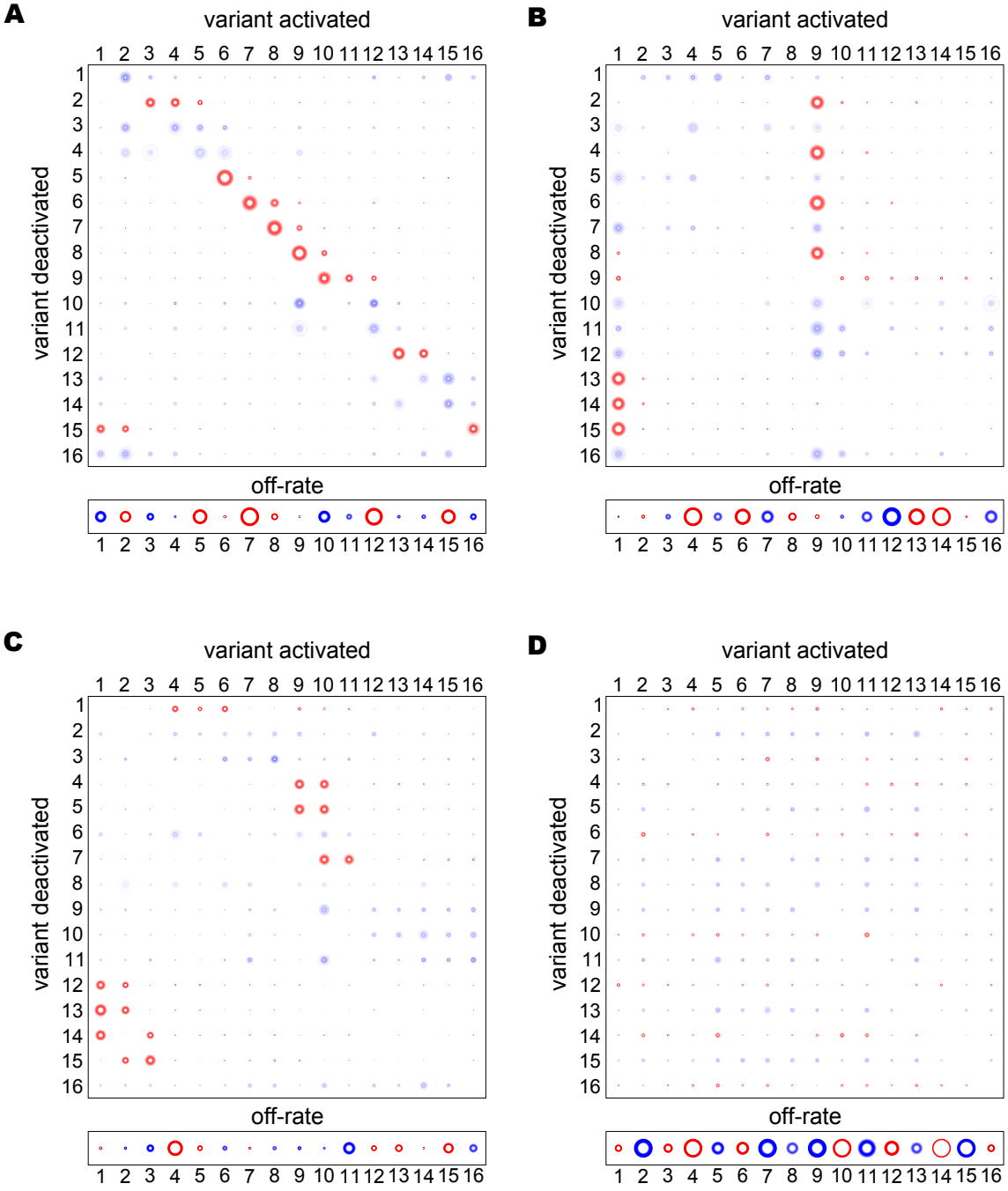


Figure 2.11. Resolved transcription time courses from noisy data (continued). Results are shown for another clone of the 1:1 switch network analysed in Figure 2.10. As in that figure, the transcript levels of the predicted network (bottom) are more similar to the noiseless data of the underlying network (top left) than to the noisy data used as input (top right). Transcript levels were measured at generation 20 (blue), 30 (purple), 40 (green), 60 (light blue) and 80 (orange). The insets depict the proportional transcript levels of the 16 variants on a log scale at generations 20, 30, 40, 60 and 80, with each colour representing a different variant.


A
B
C
D
variant activated
variant deactivated
off-rate
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

in Figures 2.8 and 2.9, the MCMC method produces generally accurate and precise estimates of starter gene parameters (indicated by less fuzzy rings) in all four network types. Moreover, off-rates for non-starter genes are estimated with a good degree of accuracy and precision, except those for the uniform network which are consistently overestimated. Although estimates of the switch rates of non-starter genes are less precise they are still informative. As is evident from comparing the MCMC output in Figure 2.12 to the output from simulated annealing (right column of Figures 2.8 and 2.9), the MCMC method provides a much more reliable indication of the true switching structure. However, the simulated annealing algorithm is still required to estimate the level of noise in the data used to calibrate the MCMC method.

2.3.4 Dimension reduction

It was initially argued that using data from only the most dominant genes within the repertoire is justifiable as long as the majority of gene transcripts remain at very low levels throughout the experiment. This assumption can be verified by creating 60-dimensional versions of the four test networks. Starting genes are selected by first simulating transcription histories for many generations until the transcript levels are close to equilibrium and thus resemble a parent population kept in long-term

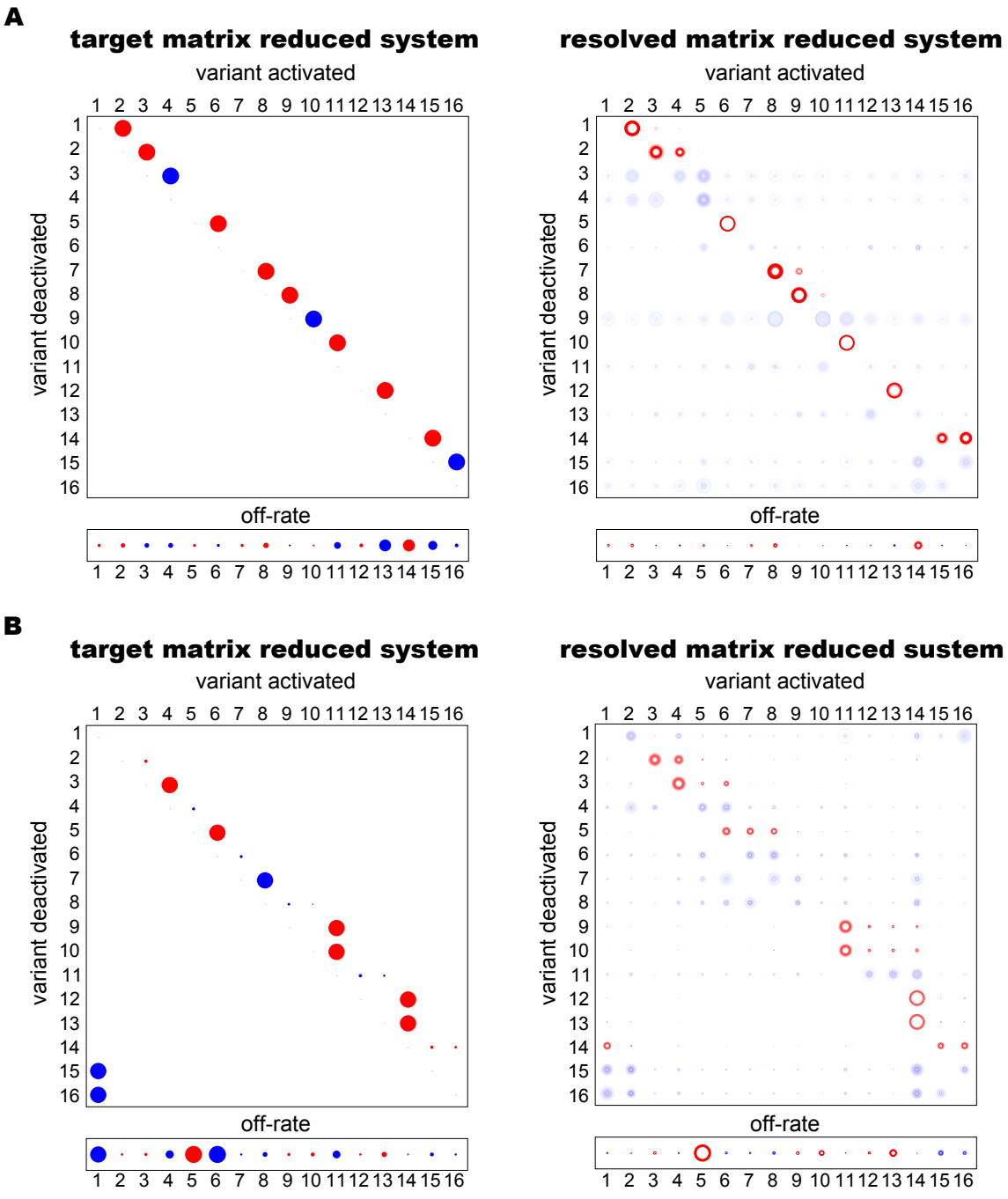
---

Figure 2.12 ***(preceding page).*** **Using MCMC on noisy transcription data.** In the MCMC output for the four different switch pathways: 1:1, SMS, lattice and uniform, (**A-D**, respectively), the parameter range for each switch bias and off-rate is divided into bins, represented by rings. If a large proportion of recorded solutions contain similar values for a parameter then the corresponding ring is coloured dark, indicating a high likelihood that the true parameter value lies within that range. The proportions are measured relative to a null distribution H_0, which assumes that all solutions are equally likely to be accepted. The darkest colour corresponds to proportions that differ from the H_0 mean by at least 25 standard deviations.

culture. Eight distinct starting genes are then randomly selected, such that the probability of selecting each gene is proportional to its level of transcription in the parent culture, thus mimicking experimental selection of starting genes by limiting dilution.

For each network, the MCMC algorithm is applied to transcription data for only the 16 most transcribed genes. To perform the reduction, genes are ranked by their average transcription levels across all time points and all cultures in the data generated by the 60-dimensional matrix (after adding noise). Data for the 16 most highly ranked genes are then selected, renormalised, and used as input. Note that genes chosen in this way may come from diverse regions of the switching network, so that the reduced matrix, which represents disjoint network elements, may look different from the original, full matrix. For example, the reduced SMS and 1:1 matrices (Figures 2.13A and 2.13B) are somewhat similar to each other. It is later shown that hypothesis testing may be used to determine the most likely network type, even when the MCMC output appears ambiguous.

For each of the reduced test networks, the MCMC parameter estimates gives a good indication of the network structure (Figures 2.13 and 2.14). When estimates vary from the true parameter values, this is often because genes included in the reduced system take on roles in the network vacated by omitted genes, so that the general network structure is preserved. These results confirm that accurate estimates of the overall switch pathways can be obtained by using only a subset of highly transcribed genes.


A
target matrix reduced system
variant activated
variant deactivated
off-rate
resolved matrix reduced system
variant activated
variant deactivated
off-rate
B
target matrix reduced system
variant activated
variant deactivated
off-rate
resolved matrix reduced sustem
variant activated
variant deactivated
off-rate

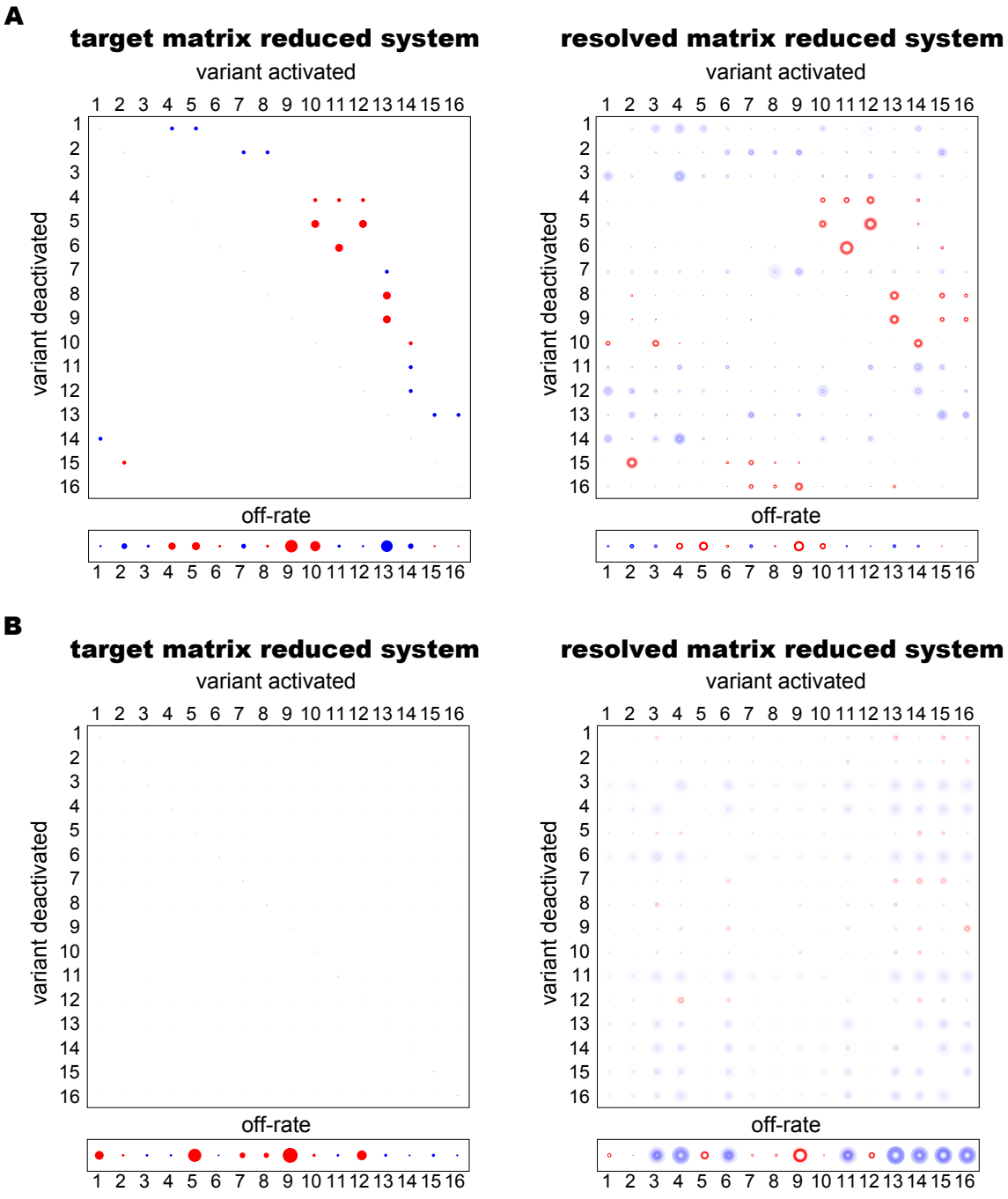
2.3.5 Hypothesis testing

As demonstrated in the previous section, optimisation methods can reliably determine genetic switch pathways from relatively limited data. Nevertheless, parameter estimates can be affected by experimental noise or by sparseness of the available data to be fitted. Although the MCMC approach provides a good indication of the uncertainty in parameter values and thus the underlying switching pathway, the question remains how to proceed if the method results in ambiguous outputs. One potential solution is hypothesis testing, which allows the direct comparison of the likelihoods of alternative switch pathways.

A strength of the likelihood-based approach is that it allows testing of hypotheses regarding the network structure where the predicted switch pathway is not immediately obvious or where there is a high degree of uncertainty around the estimated parameter values comprising this pathway. That is, if the outcome of the MCMC method is ambiguous in terms of the most likely switching network underlying the observed change in transcription levels, this approach can be used to specifically test different hypotheses. To demonstrate how this technique might be useful in practice, likelihood ratio tests are used to compare the power of two alternative switching models to explain the transcription data generated by the networks considered in the previous section (1:1, SMS, lattice and uniform networks). In the first model,

---

Figure 2.13 *(preceding page).* Results following dimension reduction: 1:1 and SMS networks. Left column: target parameters for the (**A**) 1:1 and (**B**) SMS networks reduced from 60 to 16 genes. Right column: MCMC output after adding noise with $\sigma = 1$. To perform the reduction, genes are ranked by their average transcription levels across all time points and all cultures in the data generated by the 60-dimensional matrix (after adding noise). The 16 most highly ranked genes are then selected and their data renormalised. The starter gene parameters are shown in red.


A
target matrix reduced system
resolved matrix reduced system
variant activated
variant deactivated
off-rate
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
B
target matrix reduced system
resolved matrix reduced system
variant activated
variant deactivated
off-rate
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

all genes are assumed to have identical, uniform switch biases, corresponding to a situation with completely unbiased switching; the second model allows all genes to have different sets of switch biases, as in the structured switch pathways considered previously. Simulated annealing is used to find the maximum likelihoods of each model using the noisy, dimension-reduced data sets for each network.

For the 1:1, SMS and lattice networks, the likelihood ratios correctly indicate that, in each case, the switch biases are unlikely to be uniform ($\chi^2_{224} \geq 2,350$, $p < 0.0001$), implying local switch biases and thus rejecting the hypothesis that a change in transcript levels would simply be due to differences in off-rates. Conversely, for the uniform network the test outcome is consistent with uniform switch biases ($\chi^2_{224} = 119$, $p = 1.00$), thus providing no support for the notion of variant-specific switch biases in this case. It is also possible to test for differences between the parameters of specific genes. For example, likelihood ratio tests indicate that in the SMS matrix, the switch biases of gene 10 are likely to differ from those of gene 12 ($\chi^2_{13} = 282$, $p < 0.0001$) but not from those of gene 9 ($\chi^2_{13} = 10.5$, $p = 0.65$). The latter result implies, correctly, that the network is unlikely to have a 1:1 structure.

In the above examples, results of F-tests are similar to those of the likelihood ratio tests. Other model selection statistics such as Bayes factors or information criteria are also compatible with this method. Thus, given sufficient experimental

---

Figure 2.14 ***(preceding page).*** **Results following dimension reduction: lattice and uniform networks.** Left column: target parameters for the (**A**) lattice and (**B**) uniform networks reduced from 60 to 16 genes. Right column: MCMC output after adding noise with $\sigma = 1$. To perform the reduction, genes are ranked by their average transcription levels across all time points and all cultures in the data generated by the 60-dimensional matrix (after adding noise). The 16 most highly ranked genes are then selected and their data renormalised. The starter gene parameters are shown in red.

data, it should be possible to determine with a high degree of confidence whether a hypothesised switching pattern is likely to be correct.

2.3.6 Testing with real data

In a final test, the method is applied to experimental data, comprising a subset of the *P. falciparum var* gene transcription measurements of three different clones (3D7_AS2, IT4_2B2 and IT4_2F6) previously analysed by Recker *et al.* [Recker *et al.*, 2011]. In agreement with the proposed switch pathways, for each data set the simulated annealing algorithm finds that an SMS (single-many-single) network structure gives the best fit to the data. The MCMC results are, however, more ambiguous. Although the results confirm that the switching network is unlikely to have a 1:1 or uniform structure, they also reveal that other parameter sets representing different network structures are equally likely, as indicated by the blurry matrix entries in Figure 2.15. Unfortunately the data are insufficient to carry out the more detailed hypothesis testing outlined above. Similar results can be obtained from a single, artificial data set generated by an SMS matrix (Figure 2.16), which strongly indicates that larger data sets are needed to determine the precise pattern of antigenic switching among *var* genes.

2.4 Discussion

A problem in determining switch pathways with direct methods is the wealth of data required to estimate gene intrinsic parameters, such as activation biases and off-rates, especially when the system under consideration consists of more than just a couple of genes. Most antigenically variable organisms have gene repertoires that

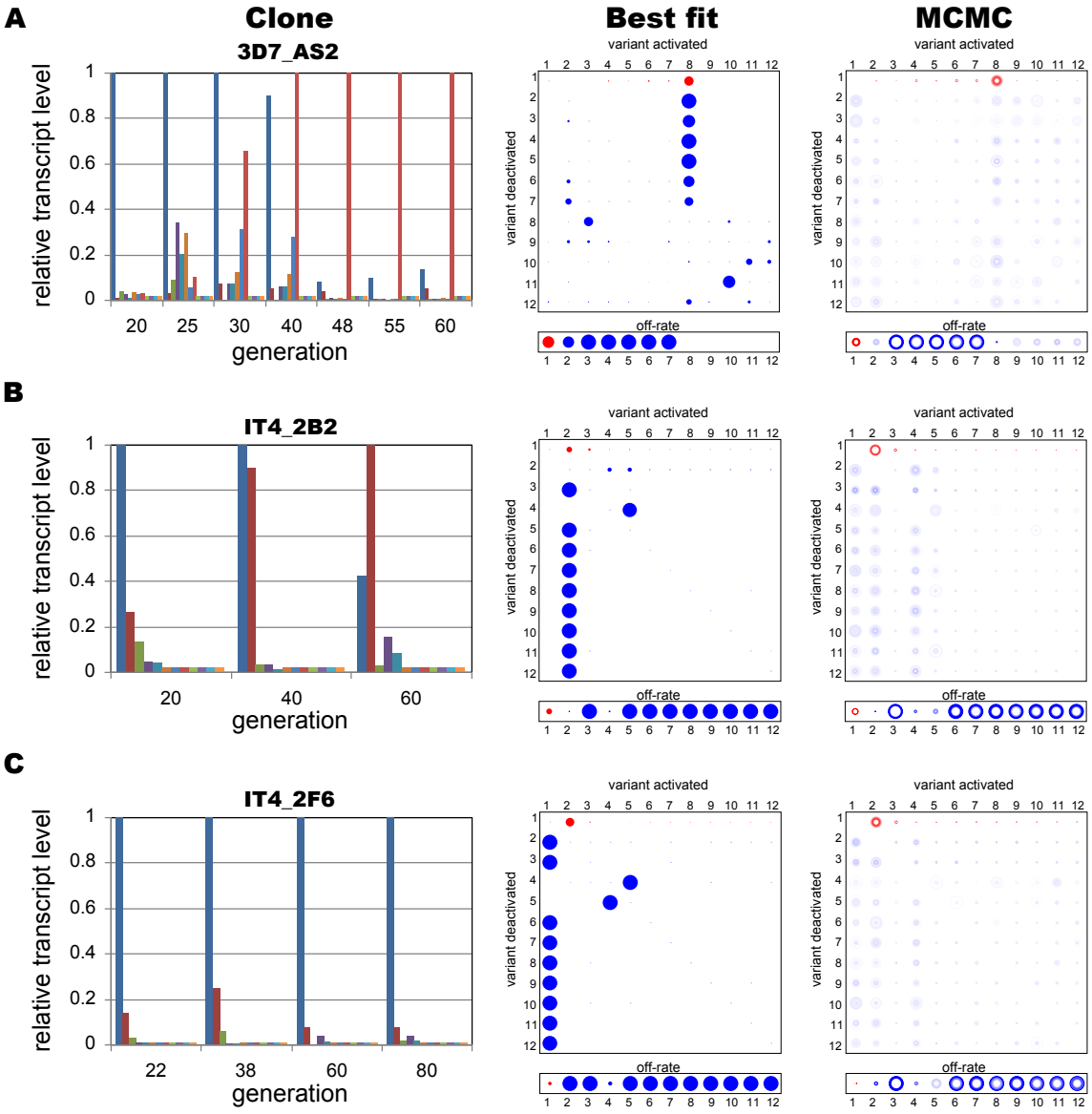


Figure 2.15. Results from experimental data. Best-fit parameter estimates derived by simulated annealing (top row) and MCMC parameter distributions (bottom row) are shown for three sets of *P. falciparum var* gene transcription data previously analysed by Recker *et al.* [Recker *et al.*, 2011]. Each data set comprises a single time series of measurements from an initially clonal culture. The results are consistent with an SMS network structure although the MCMC output (right column) also indicates the likelihood of alternative pathways.

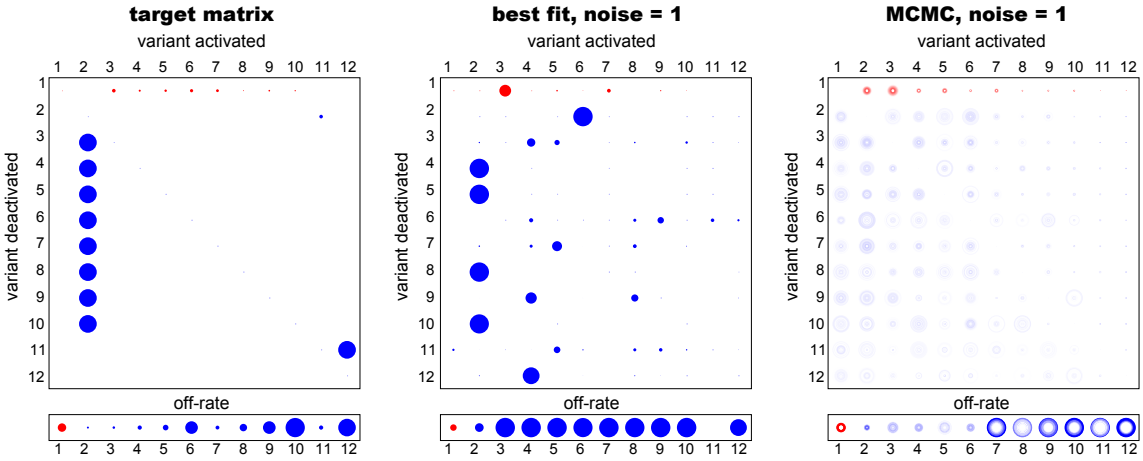


Figure 2.16. Results from simulating hypothesised *var* gene system. In this case the 60-gene SMS network is reduced to 12 genes by selecting only one starter gene, instead of the eight starter genes used in Figures 2.13 and 2.14. Simulated transcription levels are recorded at only three time points and a high level of noise is added to the data. These conditions mimic those of the experimental data sets IT4_2B2 (three time points recorded) and IT4_2F6 (four time points) analysed by Recker *et al.* [Recker *et al.*, 2011] and the results are similar to those found by analysing the original data (Figure 2.15). The best-fit parameter estimates suggest an SMS network but the MCMC output shows that this result is uncertain because of the relatively small data set.

are orders of magnitude larger. To address this problem, a statistically rigorous method for determining genetic switch pathways from quantitative gene transcription data has been described. This method has been shown to be able to resolve complex switch patterns from minimal data to a high degree of accuracy, even when allowing for a reasonable degree of experimental error or noise. Importantly, it was found that using multiple sets of data simultaneously increases accuracy such that even highly complex networks can be determined. This is of particular interest as it suggests that the most reliable information about a gene's or gene repertoire's switch behaviour can be extracted not necessarily from long-term cultures but from analysing transcription profiles which are relatively short but obtained from different clones of the same isolate. In general, the results suggest a strategy for determining genetic switching networks in three stages. The first step is to estimate the amount of noise in the data, using simulated annealing. This is followed by employing the MCMC method – calibrated to the estimated noise level – to return reliable parameter distributions for the parameters comprising the underlying network, i.e. the genes' off-rates and switch biases. Lastly, and strongly dependent on the amount and quality of the available data, hypothesised network structures can be tested by comparing the likelihoods of restricted models.

For structured networks, some variant transcripts remain at very low levels over the whole transcription time courses considered here. This is consistent with experimental studies of longitudinal *var* gene transcription, which also found that some variants are rarely activated, even after long-term *in vitro* culture. Although it was shown that data for only the most dominantly transcribed genes are sufficient for determining the overall switch pattern, uncertainty in parameter estimates for genes with low activation probabilities remains a problem for describing the switch

behaviour of an entire gene repertoire. However, once identified it may be possible to specifically select these genes for cloning and then apply the method described here. Knowledge about genes with low activation probabilities *in vitro* could yield important information about *in vivo* selection processes and should help to distinguish between parasite-intrinsic switching and immune- or receptor-mediated selection underlying the within-host infection dynamics of antigenically variable organisms.

Although the method used here was described with reference to a particular *in vitro* experimental set-up in which *P. falciparum* parasites are followed from an initially clonal population over time, there is no restriction on how the data are generated. In fact, provided that adequate temporal gene transcription data can be obtained, either from *in vitro* culture or *in vivo* infection dynamics, this method can be used to reliably determine the parameters that define the switching network between genes in any antigenically variable organism.

Chapter 3

The antigenic switching network of *Plasmodium falciparum* and its implications for the immuno-epidemiology of malaria

3.1 Introduction

At the core of the variable outcome of *P. falciparum* infection and poor development of immunity lies the major parasite virulence factor and variant antigen, *P. falciparum* erythrocyte membrane protein 1 (PfEMP1). Members of this family of proteins are expressed on the surface of infected red blood cells and are involved in the binding of parasitized cells to other host cells and tissues [Kraemer & Smith, 2006]. The conferred binding phenotypes are believed to enhance parasite survival by avoiding splenic clearance [Boone & Watters, 1995, Buffet *et al.*, 2009] and also

contribute to malaria pathology through parasite sequestration in the deep vasculature [Ockenhouse *et al.*, 1991, Pongponratn *et al.*, 1991]. Mutually exclusive transcriptional switching between PfEMP1 variants during the course of an infection thus affords the parasite an immune evasion strategy and a means to exploit different host tissues [Roberts *et al.*, 1992, Smith *et al.*, 1995].

Each haploid parasite genome contains approximately 60 members of the *var* gene family [Gardner *et al.*, 2002] encoding distinct PfEMP1 variants, with a high degree of divergence between the antigenic repertoires of any two parasites. The high diversity of *var* genes and *var* gene repertoires is predominantly generated and maintained by frequent recombination and gene conversion events [Deitsch *et al.*, 1999, Frank *et al.*, 2008, Freitas-Junior *et al.*, 2000, Kraemer *et al.*, 2007, Taylor *et al.*, 2000, Ward *et al.*, 1999], which has also resulted in a high degree of mosaicism within these genes [Bull *et al.*, 2008, Taylor *et al.*, 2000, Ward *et al.*, 1999]. The modular and highly polymorphic, extracellular portion of PfEMP1 is composed of a variable number of Duffy-binding-like domains (DBL; at least two and up to seven per gene) and cysteine-rich interdomain regions (CIDR; up to two per gene) [Kraemer *et al.*, 2007]. Despite the overall diversity of *var* genes at the sequence level, distinct associations with severe malaria have been identified for several such genes [Jensen *et al.*, 2004, Salanti *et al.*, 2003], with the involvement of *var2csa* in pregnancy-associated malaria being the clearest example to date [Salanti *et al.*, 2003], as well as associations with particular DBL domains [Warimwe *et al.*, 2009] and combinations of DBL and CIDR [Avril *et al.*, 2012, Claessens *et al.*, 2012, Lavstsen *et al.*, 2012].

Var genes can also be classified according to a conserved upstream promoter sequence into four types termed UpsA, UpsB, UpsC and UpsE [Lavstsen *et al.*, 2003].

Of particular interest are the UpsA-type genes, which appear to have diverged from the other groups [Bull *et al.*, 2008, Kraemer *et al.*, 2007] and have been shown to be upregulated during severe infections in young hosts [Bull *et al.*, 2005, Jensen *et al.*, 2004, Kaestli *et al.*, 2006, Kyriacou *et al.*, 2006, Rottmann *et al.*, 2006, Warimwe *et al.*, 2009]. As a possible consequence, acquisition of anti-PfEMP1 immunity appears to follow a particular order, with UpsA variants generally being the first to be broadly recognised in older children [Cham *et al.*, 2010, Warimwe *et al.*, 2009]. Nevertheless, clear associations between the expression of particular genes or Ups-groups, host age and severity of disease are still missing and could in fact be strain- or location-specific. Understanding the underlying pattern of antigenic switching between *var* genes is therefore important to explain not only the mechanisms and dynamics of persistent infections but also the relationship between immune-mediated expression of particular PfEMP1 sub-types and infection outcomes.

Various studies have investigated *var* gene expression during natural or experimental infections (see e.g. [Blomqvist *et al.*, 2010, Lavstsen *et al.*, 2005, Peters *et al.*, 2002, Wunderlich *et al.*, 2005]). However, as the diversity and order of PfEMP1 variants expressed in the human host is influenced by immune responses and other host factors, attempts to elucidate the patterns of antigenic change and estimate related switching parameters are commonly based on *in vitro* cultured parasites in the absence of selection pressure. In this setting, Horrocks *et al.* [2004] found that the rates at which *var* genes activate or deactivate are non-random, gene-specific and highly dissimilar. It has furthermore been suggested that *var* genes occupying subtelomeric loci tend to switch off faster than those positioned in central chromosomal regions [Frank *et al.*, 2007], offering a possible explanation for the rapid decline in the transcription of subtelomeric genes in parasites during culture adaptation [Pe-

ters *et al.*, 2007, Zhang *et al.*, 2011]. More recently, Recker *et al.* [2011] proposed that *var* gene transcriptional change involves a highly structured pathway that has evolved in response to a trade-off between the within-host and the between-host level fitness. However, most of these previous studies were limited in analytical depth due to restrictions in the number of variants considered and/or time points at which gene transcription levels were measured, and could therefore only describe small fragments of the whole switching network.

Chapter 2 described a novel statistical method that can determine antigenic switching networks with much improved accuracy. Here, this new method is applied to produce the first detailed characterisation of antigenic switching in the HB3 isolate of *P. falciparum*, using *var* gene transcription data collected by researchers at the Weatherall Institute of Molecular Medicine. The analysis reveals a global hierarchy in *var* gene activation favouring a highly genetically diverse set of genes in central chromosomal locations. In addition, evidence is found for a role of active gene transcription in the generation of antigenic diversity. These results provide the foundation for the theoretical models of Chapters 4 and 5.

A modified version of this chapter, published as a paper in September 2013, is appended to the dissertation.

3.2 Methods of modelling and analysis

As in Chapter 2, it is assumed that *var* transcription profiles can be described by the time-discrete model of Recker *et al.* [2011]:

$$v_{i,c,t+1} = (1 - \omega_i)\, v_{i,c,t} + \sum_{j \neq i} \omega_j \beta_{ji} v_{j,c,t},$$

where $v_{i,c,t}$ is the relative transcript level of variant i in culture c at time t, ω_i is the variant specific off-rate, and β_{ji} is the switch bias (probability of a switch) from variant j to variant i. Also as in Chapter 2, it is assumed that off-rates are below 6% per generation, which is consistent with previous studies and theory [Frank *et al.*, 2007, Horrocks *et al.*, 2004, Recker *et al.*, 2011, Roberts *et al.*, 1992].

The Markov chain Monte Carlo (MCMC) method described in Chapter 2 is used to obtain posterior distributions of likely values for the switch biases and off-rates, using the pooled data from all observed cultures as input. For model comparisons, the simulated annealing algorithm described in Chapter 2 is used to find maximum likelihood parameter sets. To estimate activation rates, the model is modified so that the direction of switching is independent of the previously active gene (that is, $\beta_{ji} = \beta_{ki}$ for all i, j and k).

Before conducting statistical tests, switch biases are transformed using the logit function $\text{logit}(x) = \log(x/(1-x))$ to map from the range $(0, 1)$ to the range $(-\infty, \infty)$. Additionally, to account for uncertainty in the parameter estimates, the same statistical tests are performed on the collected output of the MCMC method. This method samples from the posterior distribution, which means it selects numerous parameter sets such that the probability of each being chosen depends on how likely it is to explain the data. Statistical tests are performed on all the selected parameter sets and the mean p-values are reported. The same parameter sets are used to obtain credible intervals for estimates.

3.3 Laboratory materials and methods

All laboratory work described in this section was conducted by collaborators at the Weatherall Institute of Molecular Medicine.

3.3.1 Parasite culture and cDNA preparation

Parasites were cultured and sorbitol-synchronized using standard techniques [Lambros & Vanderberg, 1979, Trager & Jensen, 1976] and RNA was extracted from saponin lysed, mid-to-late ring stage parasites with Trizol as previously described [Kyes *et al.*, 2000]. For cDNA preparation, 5 μg RNA was treated in 20 μl reactions with TurboFree DNase (Ambion; following manufacturer's recommendations except incubation was at 16°C for 40 minutes). All DNase treated RNA samples were tested by PCR with primers to fructose-bisphosphate aldolase (PF14_0425) [Salanti *et al.*, 2003] to ensure complete removal of gDNA. Duplicate 20 μl reactions consisting of 6 μl RNA were reversed transcribed with 100 ng random hexamers (Invitrogen) and Bioscript reverse transcriptase (Bioline; following manufacturer's instructions) at 40°C for 40 minutes. The resulting cDNA was stored at -20°C in single-use aliquots and diluted 1:6 for real time PCR analysis. cDNA was synthesised in triplicate (84% of all data points) or duplicate, and real time PCR was done in duplicate for each primer pair on each cDNA.

3.3.2 Primer design and initial testing

Primers specific for the HB3 *var* exon 1 repertoire were optimised to meet several basic criteria: that they distinguished between the known *var* gene sequences within the HB3 genome; that they were specific for HB3 *var* genes with no cross-reactivity

to 3D7 *var* genes (for future application to studies on 3D7xHB3 cross progeny); and that they reflected measurement of full-length *var* transcript. Primer pairs were designed using Primer3plus [Kraemer *et al.*, 2007, Untergasser *et al.*, 2007], with parameters manually set for product size range (100-150nt), primer length (20nt), Tm (60°C), GC clamp (1nt), salt correction formula and thermodynamic parameters (SantaLucia 1998); all other parameters were default settings. To avoid general cross-detection of all *var* types, the first kilobase of exon 1 containing the conserved DBLα domain was avoided. The final kilobase was also excluded to avoid sterile transcripts [Kyes *et al.*, 2007, Su *et al.*, 1995b]. Each primer pair was specific for a single gene except, unavoidably, the *var3* primer pair and the two pairs designed to the highly similar copies of *var2csa*. Each HB3 primer pair was tested five times on gDNA; pairs that varied more than +/- 1.5 threshold cycles (Ct) values from the median or had efficiency value (Amplification) less than 1.85 were redesigned. Primer pairs that amplified 3D7 genomic DNA were discarded and redesigned so that they were specific for HB3 *var* genes only.

3.3.3 qRT-PCR

Quantitative real time PCR (qRT-PCR) reactions were prepared with a Corbett CAS1200 liquid handling system and run on a Corbett Rotor-Gene 6000. Each 10 μl reaction contained 1 μl template, 5 μl SensiMix SYBR (Bioline), 3 μl water and 1 μl primer mix (1.6 μM of each primer). The cycling conditions were as follows: Hold, 95°C/10 minutes; 45 cycles (95°C/25 seconds, 58°C/25 seconds, 68°C/30 seconds); Melt, 55°C-99°C. Amplicons were checked for a single product of the expected size by agarose gel electrophoresis. Melt analysis confirmed single products.

Comparative quantification results were given as expression levels relative to a

control (housekeeping) gene, seryl-tRNA synthetase (Seryl-tRNAF agctacctcagaa-caaccattatgtgc, Seryl-tRNAR atcctttccatgtgcccctgc), which was arbitrarily set at 1. Minor differences in Take Off values between primer pairs were accounted for using a correction factor derived from 5 replicates on gDNA.

Dominant *var* gene expression detected by qRT-PCR was verified as both full-length and dominant by northern blot analysis, comparing specific probes for each dominant gene vs a generic exon 2 probe. The primer pairs for *var* genes 5, 11, 13, 27, 28, 29, 31 and 35 were thus verified as specific for detecting full-length transcripts. The *var3* primer pair, which was known to also detect *var5* and pseudogene *var46*, indicated dominant *var3* alongside *var5* in culture #4. However, it was not possible to detect *var3* expression by northern blot. As this sample was positive by northern blot probed with *var5*, and it was the only one with 'dominant' *var3* expression by qRT-PCR, all *var3* qRT-PCR data were discarded as ambiguous. Data from one of the two *var2csaA*/*var2csaB* primers pairs were also discarded, as the data sets obtained from these primers were extremely similar.

3.4 Analysis of the HB3 data set

3.4.1 Overview of the data

Laboratory experiments and data collection were carried out by scientists at the Weatherall Institute of Molecular Medicine. Eleven subclones of the lab-adapted HB3 isolate of *P. falciparum* were obtained by serial dilution of the parent culture. The clones were cultivated and transcript levels of 38 *var* genes measured by quantitative PCR, first at 29-33 parasite generations (8-9 weeks) after cloning and then at several time points over the following 60-79 generations in each culture (Table

1). These 38 genes comprised almost all known functional HB3 *var* genes [Kraemer *et al.*, 2007, Rowe *et al.*, 2009] and one partial gene. The HB3 parent culture was also observed at seven time points. At the first observation, each of the eleven clones predominantly transcribed one of nine different *var* genes, which were most likely the genes transcribed at the time of cloning. Importantly, these nine genes – hereafter referred to as "starter genes" – were representative of the *var* repertoire diversity in terms of chromosomal location and Ups type.

Table 3.1. Parasite culture and RNA sampling information.

Culture	Starter gene	Observed generations
1	*var11*	29, 48, 59, 69, 89
2	*var13*	29, 48, 69, 89, 99, 108
3	*var13*	30, 49, 70, 90, 100, 109
4	*var5*	30, 39, 49, 60, 70, 90
5	*var27*	30, 49, 70, 90
6	*var27*	29, 38, 48, 69, 89
7	*var28*	33, 52, 63, 73, 93
8	*var29*	30, 49, 70, 80
9	*var30*	29, 48, 59, 69, 89
10	*var31*	29, 48, 59, 69, 89, 99, 108
11	*var35*	32, 51, 65, 72, 92
Parent	-	0, 9, 19, 40, 60, 70, 79

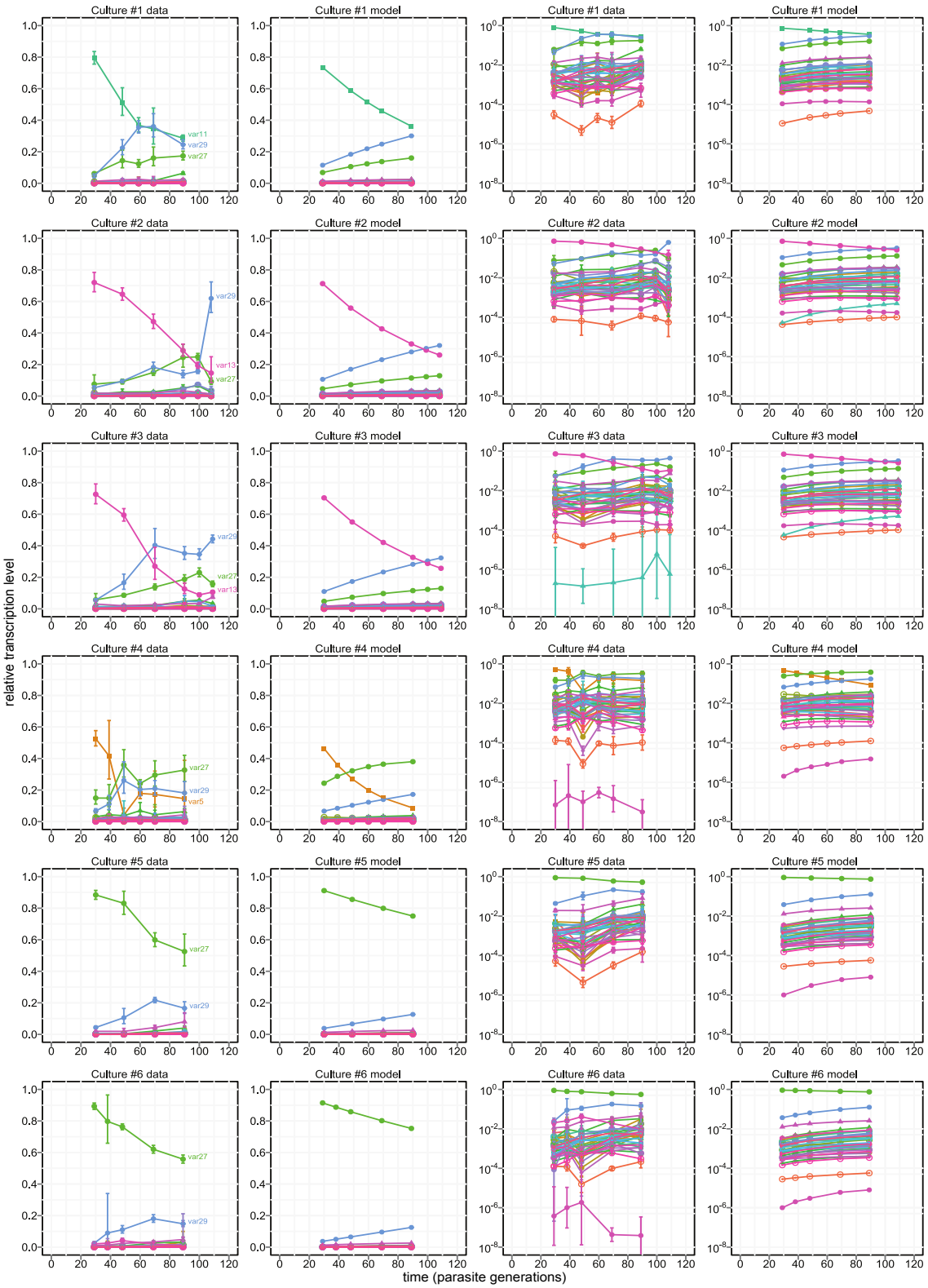
3.4.2 *Var* gene transcription dynamics

The first step of data analysis is to convert the raw quantitative real time PCR (qRT-PCR) measurements into transcription levels to be used as model input. This is done using the comparative quantification method of Pfaffl [Pfaffl, 2001]. That is, the level for each replicate measurement is determined using $A^{T.O.-CF}$, where A is the amplification rate, $T.O.$ is the number of replication cycles taken to reach a cut-off threshold, and CF is a correction factor. Expression levels are converted to

proportions and the geometric mean is calculated for each set of replicates. Replicates are excluded as outliers if they deviate from the mean by a factor of more than $10^{1.5}$ or have an amplification efficiency below 1.5. Thus, 1.1% of replicates from the HB3 data set are excluded, most of which meet both criteria.

As shown in Figures 3.1 and 3.2, the proportional transcript levels of the different starter genes declined at notably different rates over the course of the experiment. For example, the centrally located, UpsC starter gene *var29* in culture #8 accounted for more than 90% of the total transcript level at the first measurement and remained the dominant variant over the following 60 generations. In contrast, the transcript level of the *var35* starter gene of culture #11 (subtelomeric, UpsA promoter type) was already below 50% of the total transcript level at the point of the first observation (approximately 30 generations post cloning) and continued to decline during the time course of culture, to be eventually replaced by other genes as dominant transcripts. Similar patterns are seen in cultures #3 and #7, which both had centrally located, UpsC starter genes, whereas the remaining five starter genes show intermediate rates of decline.

Of particular interest is the apparent convergence of most cultures towards high transcription of *var27* and/or *var29* (Figures 3.1, 3.2), which are both situated in central chromosomal location but have different promoter types (UpsB and UpsC, respectively). In light of the very different initial conditions, in terms of *var* gene transcription levels at which these cultures were established, this indicates that antigenic switching in *P. falciparum* might be mainly governed by an activation hierarchy.

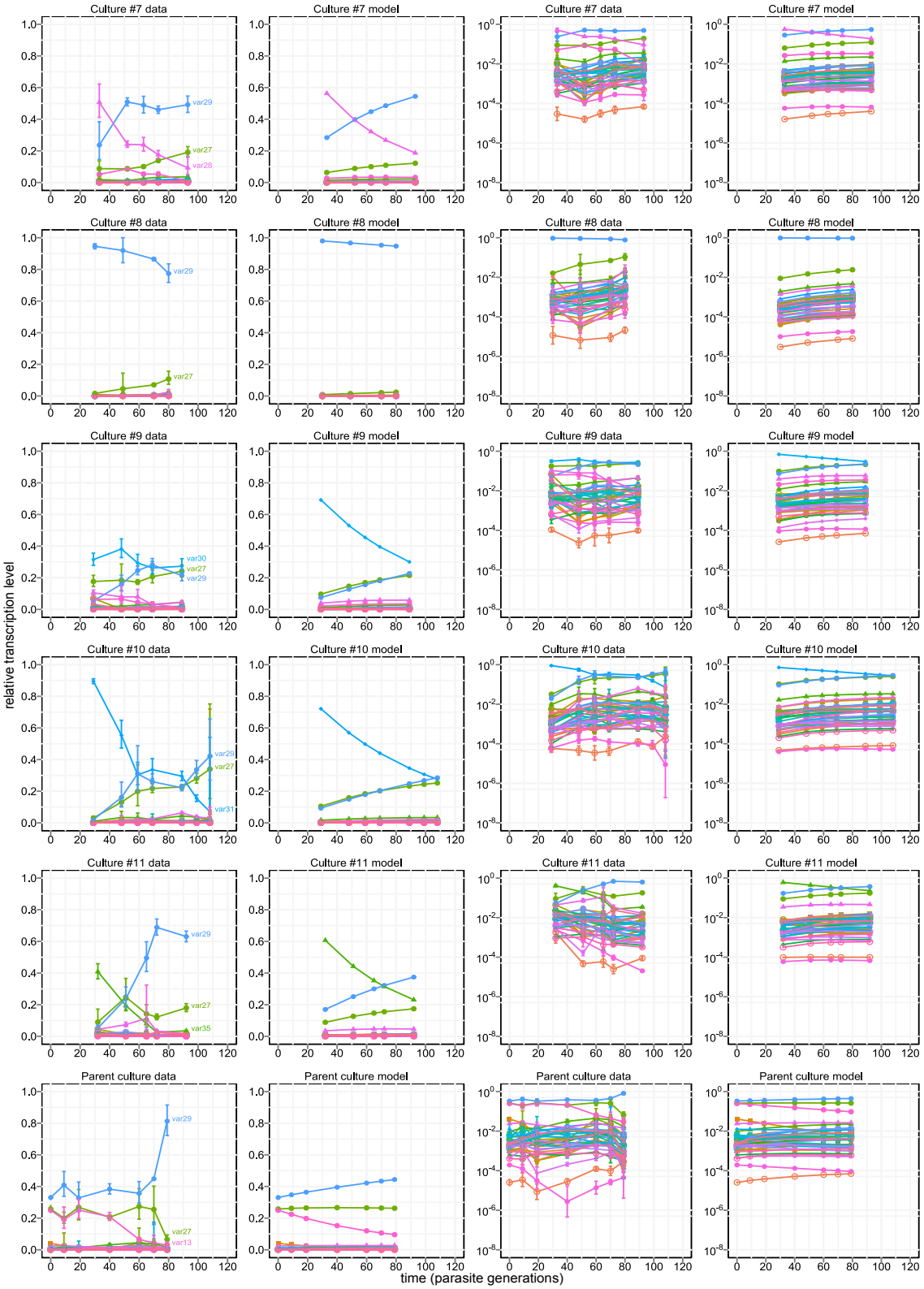

Culture #1 data
Culture #1 model
Culture #2 data
Culture #2 model
Culture #3 data
Culture #3 model
Culture #4 data
Culture #4 model
Culture #5 data
Culture #5 model
Culture #6 data
Culture #6 model
relative transcription level
time (parasite generations)
var11
var29
var27
var13
var5

3.4.3 Parameter estimation

To fully understand the patterns of antigenic switching underlying the observed transcriptional changes in the parasite cultures, the relevant switch parameters are estimated using the Markov chain Monte Carlo (MCMC) method described and tested in the preceding chapter. It has been proposed that *var* gene switching can be fully described dynamically by two sets of parameters [Recker *et al.*, 2011]. First, each gene i has an off-rate, ω_i, which is the per generation probability of switching from active to silenced. Second, each gene has a switch bias, β_{ij}, which is the probability that when gene i switches off, gene j becomes activated.

In Chapter 2, it was shown that reliable parameter estimates can be obtained from data for the 16 most highly transcribed *var* genes, which in the HB3 data set comprises the nine starter genes and the seven other genes with highest average transcript levels. These 16 genes together account, on average, for more than 92% of the sum of transcript levels in each culture (excluding the culture's starter gene), whereas none of the excluded genes account for more than 1%. Therefore the algorithm is applied to the full 38-gene data set and separately to a reduced set of the 16 most transcribed genes. Both models generally give good and similar fits between the data and predicted transcript levels (Figures 3.1, 3.2), with deviations resembling stochastic measurement errors.

Figure 3.1 *(preceding page).* Proportional *var* transcript levels and model output for cultures 1-6. The parasite cultures initially expressed a variety of dominant "starter genes", which switched off at notably different rates. Nevertheless, most cultures converged towards high level transcription of two centrally located genes *var27* and *var29*. Columns 1 and 2 show the measured transcript levels at linear and logarithmic scale, respectively; columns 3 and 4 show the model fit. Error bars show approximate 95% confidence intervals estimated from qRT-PCR replicates.


Culture #7 data
Culture #7 model
Culture #8 data
Culture #8 model
Culture #9 data
Culture #9 model
Culture #10 data
Culture #10 model
Culture #11 data
Culture #11 model
Parent culture data
Parent culture model
relative transcription level
time (parasite generations)
var29
var27
var28
var30
var31
var35
var13

The MCMC method provides estimated switch parameters for the entire *var* repertoire, which are illustrated in Figure 3.3 for the 16 most highly transcribed genes. Due to higher signal to noise ratios, parameter estimates of the starter genes (indicated in red) are more accurate and result in much narrower credible intervals, as indicated by less fuzzy rings that represent respective switch probabilities. As clearly demonstrated in Figure 3.3, there is wide distribution in switch biases and off-rates, and the next step is to analyse for specific associations between these parameters and other gene characteristics.

3.4.4 Off-rate estimates suggest a role of chromosomal location, not promoter type

In the HB3 results, estimated off-rates range from 0.3% to more than 5% per generation, with a mean of approximately 2.9% (Table 3.2). Notably, the two most transcribed genes, *var29* and *var27* (both centrally located on chromosome 4), are found to have exceptionally low off-rates of approximately 0.3% and 0.8% respectively, whereas the other starter genes switch off at rates between 1.4% and 3.5% per generation. In general off-rates of centrally located genes are found to be significantly lower than of genes in subtelomeric regions (Welch's $t = 3.3, p = 0.002$),

Figure 3.2 *(preceding page)*. Proportional *var* transcript levels and model output for cultures 7-11. The parasite cultures initially expressed a variety of dominant "starter genes", which switched off at notably different rates. Nevertheless, most cultures converged towards high level transcription of two centrally located genes *var27* and *var29*. Columns 1 and 2 show the measured transcript levels at linear and logarithmic scale, respectively; columns 3 and 4 show the model fit. Error bars show approximate 95% confidence intervals estimated from qRT-PCR replicates.

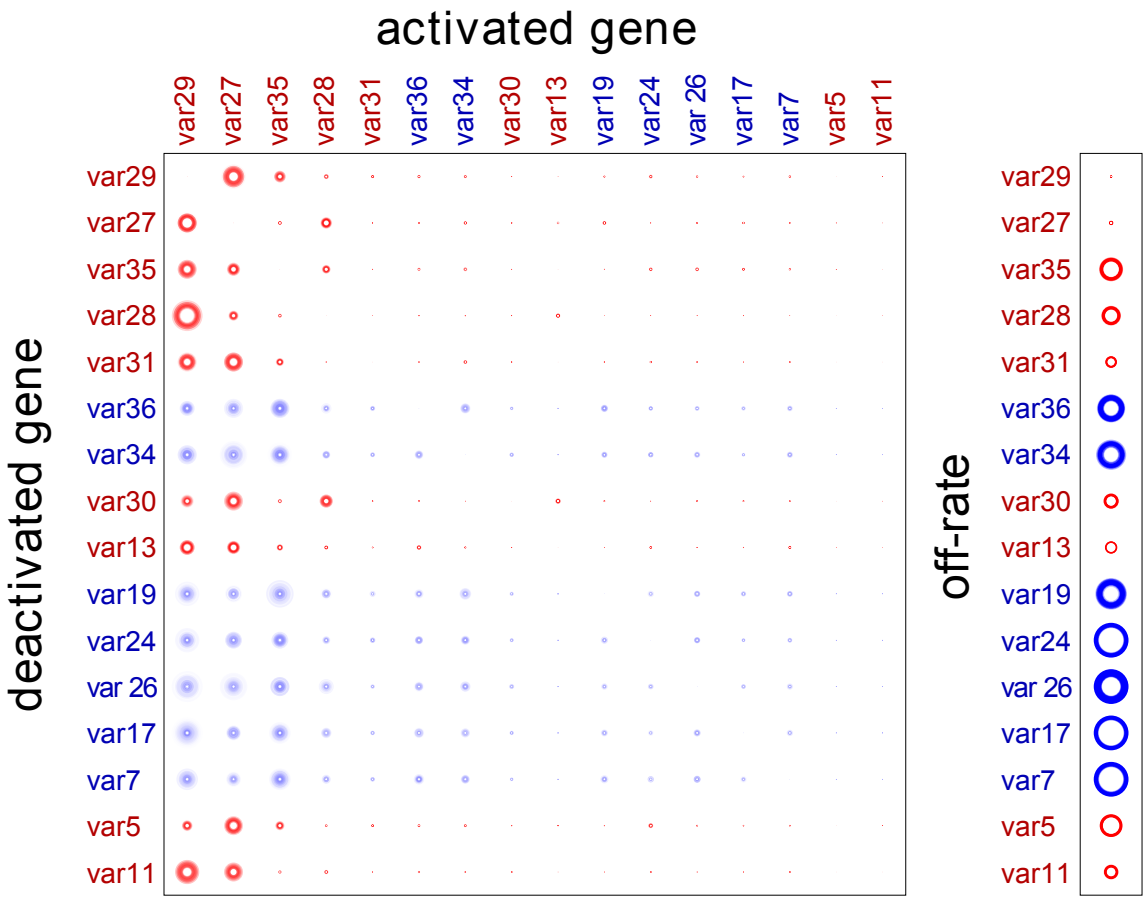


Figure 3.3. Estimation of switch parameters of highly transcribed genes. Parameter estimations for the 16 most transcribed *var* genes are represented as a switch matrix and an off-rate vector. The diameter of a circle in the ith row and jth column of the matrix is proportional to the switch bias β_{ij} from gene i to gene j, and the diameter of a circle in the off-rate vector is proportional to the rate ω_i at which gene i becomes silenced. The fuzziness indicates uncertainty in the estimate, such that the darkness of each concentric ring is proportional to the likelihood that the parameter is within the corresponding range (see preceding chapter). Starter gene parameters, which are more precisely estimated, are in red, and genes are arranged from left to right, and top to bottom, in order of average transcript level.

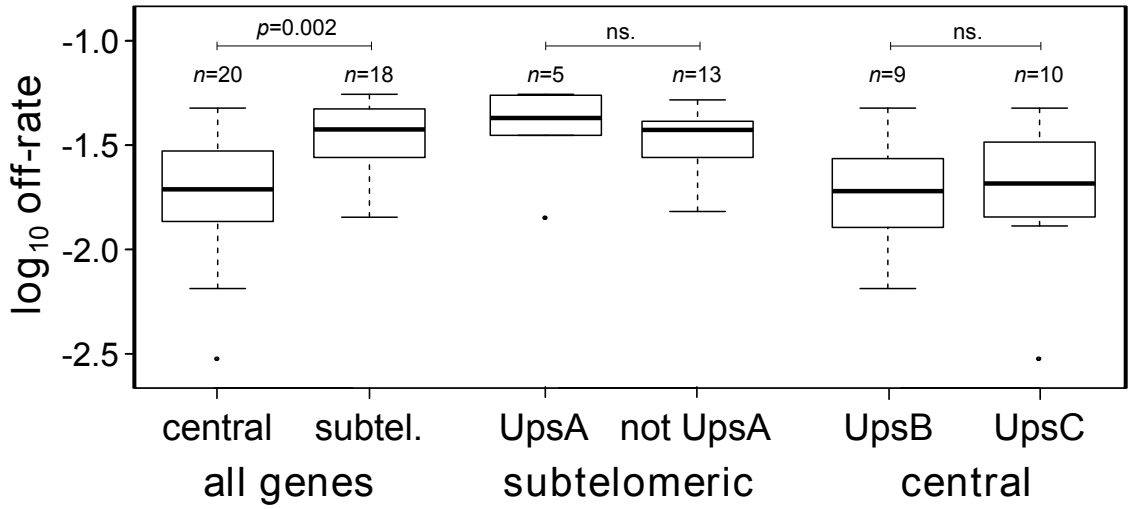


Figure 3.4. Off-rate estimates show strong dependency on chromosomal location. The estimated rates at which active *var* genes become silenced are significantly lower for centrally located *var* genes than for those in subtelomeric locations. No significant effect of upstream promoter type is found by differentially testing off-rates of genes in subtelomeric location (UpsA vs. non-UpsA) or in central location (UpsB vs. UpsC).

as shown in Figure 3.4. This result remains significant ($p = 0.035$) after accounting for uncertainty in the estimates, as described in Section 3.2.

Previous studies reported that UpsA *var* genes from patient isolates were down-regulated faster than other *var* genes during adaptation to culture [Peters *et al.*, 2007, Zhang *et al.*, 2011]. The HB3 repertoire contains seven UpsA genes, all but one of which are subtelomeric. A comparison of the off-rates of the subtelomeric UpsA genes with those of other subtelomeric genes reveals no significant difference (Welch's $t = 0.65, p = 0.54$; accounting for uncertainty $p = 0.55$), as shown in Figure 3.4. Among centrally located genes, comparison of UpsB with UpsC again reveals no significant difference (Figure 3.4, Welch's $t = 0.44, p = 0.66$; accounting for uncertainty $p = 0.57$). This lends support to a previous study suggesting that the

rate at which transcriptionally active *var* genes become silenced is less dependent on promoter type than on chromosomal location [Frank *et al.*, 2007].

3.4.5 *Var* gene switching is governed by an activation hierarchy

In each of the HB3 cultures, gene transcription levels spanned several orders of magnitude and only a minority of genes were ever highly transcribed during the experiment. As mentioned earlier, despite the diversity of the starter genes, the relative transcript levels of the activated genes appear to converge towards similar levels across all cultures (Figures 3.1, 3.2). The degree of similarity in gene activation is tested by excluding the nine starter genes and ranking the rest of the repertoire according to average transcript levels at later time points (more than 65 generations) for each culture. Pair-wise comparison then shows that the ranking to be highly correlated between the cultures (Spearman's $\rho > 0.8, p < 10^{-5}$ for all pairs), suggesting that the outcome of antigenic switching in *P. falciparum*, or at least within the HB3 isolate, is largely governed by an activation hierarchy that is independent of the currently activated gene.

Consistent with this hypothesis, the estimated switch biases between genes, β_{ij}, are remarkably similar, so that $\beta_{ij} \approx \beta_{kj}$ for each i, j and k. (Figure 3.3). All genes are found to have high switch biases towards the same few genes, which explains the observed convergence in *var* transcription in the cultures. As expected, the highest biases across the repertoire are found towards *var29* and *var27*, with some genes also frequently switching towards *var35* and/or *var28* (both UpsC-types in central chromosomal location).

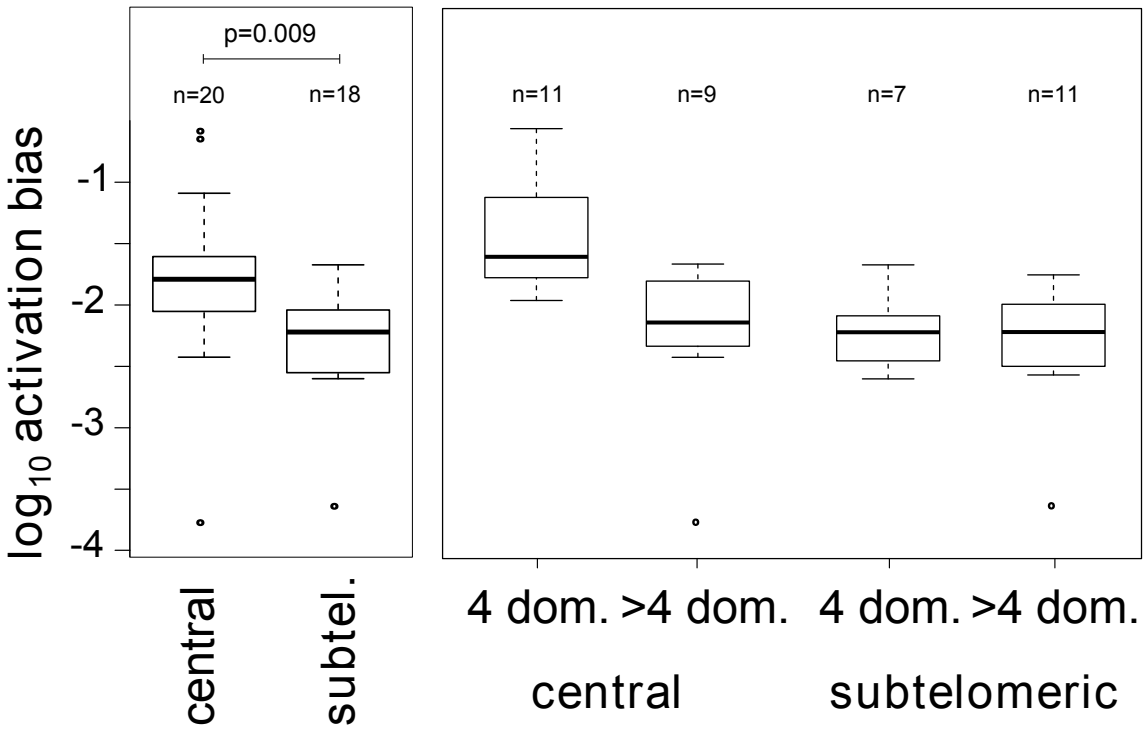


Figure 3.5. Associations between activation biases and genetic attributes. The mean activation bias is higher for centrally located *var* genes. Within the set of centrally located genes, those encoding only four binding domains have higher activation biases than longer genes. Together, short, central genes have activation biases ≈10 times higher than the rest of the repertoire.

To better describe the observed activation hierarchy, the MCMC method is used to estimate the average switch bias towards each gene, which is hereafter referred to as the gene's activation bias. That is, the activation bias of gene j is defined as the average of the switch biases β_{ij} for all deactivating genes i. It can be understood as a gene specific, *per switch* activation probability and is therefore distinct from the previously defined *on-rate* [Horrocks *et al.*, 2004], in that it is invariant to transcript levels. As shown in Table 2, there is a clear hierarchy in gene activation biases spanning over three orders of magnitude. The ranking also appears to be non-random so that, for example, most genes within central chromosomal location are situated in the top half of the table. Analysing for possible associations with genetic attributes confirms that activation biases of centrally located *var* genes are significantly higher than those of subtelomeric genes (Figure 3.5, Welch's $t = 2.8, p = 0.009$; accounting for uncertainty $p = 0.011$). In fact, of the highest four activation biases, three belong to neighbouring genes on chromosome 4: *var29*, *var27* and *var28*. Importantly, no significant and independent effect of promoter type is found, which again points towards a strong influence of chromosomal location on *var* gene activation.

Previous research has suggested that gene length, in terms of the number of binding domains, is another important *var* gene characteristic besides chromosomal location and promoter type [Buckee & Recker, 2012]. In the HB3 data, grouping genes depending on whether they encode only four domains (short genes) or more than four domains (long genes) shows that activation biases differ significantly depending on gene length (Figure 3.5, $F = 9.7, p = 0.004$) as well as chromosomal location ($F = 8.1, p = 0.007$) and the interaction of these two factors ($F = 5.0, p = 0.032$). As such, central *var* genes encoding only four binding domains (also known as Type 1 *var* genes [Gardner *et al.*, 2002]) have on average a 10 times higher activation bias

Table 3.2. Parameter estimates for the HB3 *var* repertoire.

Gene	Location	Domains*	Ups	Activation rate	Off-rate
var29	Central	4	C	29%	0.3%
var27	Central	4	B	25%	0.7%
var35	Central	4	C	8.0%	3.2%
var28	Central	4	C	6.6%	2.1%
var32	Central	4	C	3.1%	4.6%
var31	Central	4	C	2.2%	1.4%
var13	Subtelomeric	4	B	2.1%	1.9%
var36	Central	4	C	1.9%	0.9%
var30	Central	4	C	1.8%	1.9%
var34	Central	7	C	1.7%	0.8%
var10	Subtelomeric	6	B	1.5%	2.6%
var17	Central	7	B	1.4%	2.5%
var1csa	Subtelomeric	9	A	1.3%	5.4%
var26	Central	4	B	1.2%	0.9%
var24	Central	7	B	1.1%	1.1%
var7	Subtelomeric	8	B	1.1%	1.6%
var19	Central	4	B	1.1%	0.6%
var25	Central	7	B	0.9%	1.9%
var4	Subtelomeric	9	A	0.8%	4.0%
var33	Central	4	C	0.8%	1.9%
var21	Central	8	B	0.8%	4.4%
var16	Subtelomeric	4	B	0.7%	2.3%
var14	Subtelomeric	4	B	0.7%	4.0%
var18	Subtelomeric	4	B	0.7%	4.9%
var11	Subtelomeric	5	B	0.6%	1.5%
var39p†	Subtelomeric	0	B	0.6%	4.6%
var22	Central	7	B	0.6%	1.3%
var8	Subtelomeric	8	B	0.6%	3.6%
var5	Subtelomeric	7	A	0.5%	3.4%
var50‡	Central	7	C	0.4%	3.9%
var2csaA	Subtelomeric	6	E	0.3%	4.5%
var23	Central	7	B	0.3%	0.3%
var9	Subtelomeric	7	B	0.3%	2.4%
var20	Subtelomeric	4	B	0.2%	3.7%
var12	Subtelomeric	4	B	0.2%	3.4%
var2	Subtelomeric	7	A	0.2%	0.8%
var1	Subtelomeric	8	A	0.03%	5.4%
var6	Central	9	A	0.01%	2.1%

*Number of encoded DBL (Duffy binding-like) and CIDR (Cys rich inter-domain region). †Partial gene. ‡Psuedo gene

than the rest of the repertoire. Results are similar after accounting for uncertainty in the estimates ($p = 0.005, p = 0.0096, p = 0.037$, respectively).

3.4.6 *Var* gene switching is influenced by the deactivating gene

In the method of determining switching networks described in Chapter 2, the third and final stage is hypothesis testing. An important question arising from the HB3 results is whether the currently active gene has an influence on the direction of switching, i.e. if in general $\beta_{ij} \neq \beta_{kj}$. Therefore the simulated annealing algorithm described in Chapter 2 is used to fit two alternative models to the data for the 16 most highly transcribed *var* genes: (i) a general model including off-rates and individual switch biases (240 independent parameters), and (ii) a simpler model, in which all sets of switch biases are required to be identical, i.e. where $\beta_{ij} = \beta_j$ for all i and j (31 independent parameters). A likelihood ratio test indicates that the model with individual switch biases is significantly more likely than the second model ($F = 1.84, p < 0.0001$), confirming that antigenic switching in *P. falciparum* is not only controlled by an activation hierarchy but is also influenced by the currently active gene.

The individual variations can be estimated from a subset of the MCMC switch bias estimates. Each set $\{\beta_{1j}, \beta_{2j}, \ldots, \beta_{nj}\}$ is standardised by subtracting the mean and dividing by the standard deviation (after rescaling to compensate for $\beta_{jj} = 0$). Negative values then correspond to switch biases that are smaller than average, and positive values to those that are larger. For HB3, a suitable subset (balancing sample size and uncertainty of estimates) includes the switch biases from the nine

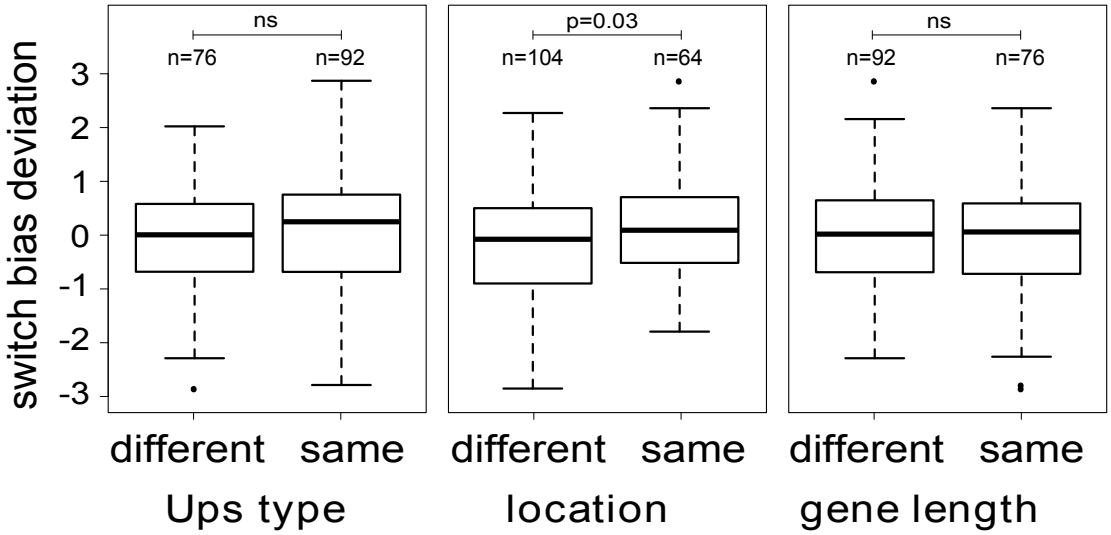


Figure 3.6. Variation in switch biases (deviation from the mean). Antigenic switching between *var* genes is more frequent between genes with matching chromosomal locations, e.g. from central to central, or from subtelomeric to subtelomeric. No significant associations between switch bias and matching Ups type or gene length are found.

starter genes to the 25 most transcribed genes, omitting the diagonal terms. Testing for associations with chromosomal location and promoter type then indicates that switch biases are higher between *var* genes with matching chromosomal locations than between differently located genes (Figure 3.6, Welch's $t = 2.2, p = 0.029$), whereas no significant association is found between switch biases and matching Ups types or gene lengths. This result needs to be confirmed with larger data sets, however, as statistical significance is lost after accounting for uncertainty in the parameter estimates.

3.4.7 *Var* gene sequences suggest a role of activation bias in antigenic evolution

The final stage of analysis focuses on testing for relationships between antigenic switching parameters and genetic diversity. *Var* gene sequences can be described using a set of 628 conserved homology blocks with an average length of 19 amino acids [Rask *et al.*, 2010]. Homology blocks found only within the DBL1 and CIDR1 domains, which are present in all HB3 *var* genes except *var1csa*, *var2csaA*, *var2csaB* and the partial gene *var39*, are used to construct a *var* gene homology network, in which edges between genes are defined and weighted by the number of shared homology blocks (Figure 3.8). Then, the centrality of each gene within this network is calculated using a weighted closeness measure [Newman, 2001] implemented by the tnet package for R [Opsahl, 2009]. This is a distance measure between all pairs of nodes (i.e. genes) and can therefore be used as a proxy for the relatedness of a gene to the rest of the repertoire. The inclusion of network centrality as an explanatory variable in the previous ANOVA model confirms that the degree by which *var* genes share sequence blocks is positively correlated with activation bias, independently of chromosomal location and gene length (Figure 3.9, $F = 10.5, p = 0.003$ for the model comparison; accounting for uncertainty $p = 0.006$).

It should be noted that prior to this analysis, transformations are applied to the data to better conform to the assumptions of the statistical model. First, the relatedness values are rescaled and then both they and the activation biases are logit transformed (this is a standard method to map proportional data onto an unbounded domain). When the linear model ($relatedness \sim bias + location \times domain$) is applied to these data there is weak heteroscedasticity, meaning that the magnitude of the

residuals is weakly correlated with the fitted values (Figure 3.7A, second plot). The residuals distribution is also slightly skewed (Figure 3.7A, third plot) but this cannot be corrected with a power transformation (Figure 3.7A, fourth plot). Applying a square root transformation to the activation bias data, centred on the midpoint of the data, spreads the data more evenly along the bias axis and results in a more linear relationship between residuals and fitted values (Figure 3.7B, first and second plots, non-constant variance score test: $\chi^2 = 0.24, p = 0.62$). The recommended power transformation to correct for non-normality is still approximately 1 (Figure 3.7B, fourth plot). Two outliers (points 7 and 23) have a disproportionate influence on the model fit, as shown by large Cook's distances (Figure 3.7B, fifth plot), but if these points are removed then the result of the model comparison remains significant ($F = 21.8, p = 0.00007$).

The correlation between relatedness and activation bias leads to a hypothesis that activation bias, or gene transcription, has a role in *var* gene sequence evolution. Under the assumption that gene transcription facilitates recombination, more frequently activated genes would be expected to share more sequence blocks with other genes, simply because they more often act as recombination sites. This idea is illustrated by a simple model that explicitly simulates gene conversion by choosing pairs of genes at random, weighted by their activation bias, and copying homology blocks between them. Specifically, gene conversion is simulated in a repertoire of 30 genes, each comprising 40 homology blocks, with 30 possible variants for each homology block. Genes are initially constructed of randomly chosen homology blocks and are assigned activation rates according to a geometric sequence. For 30,000 iterations, pairs of genes are selected at random, with the probability of selection being proportional to activation rate. A randomly chosen homology block is copied

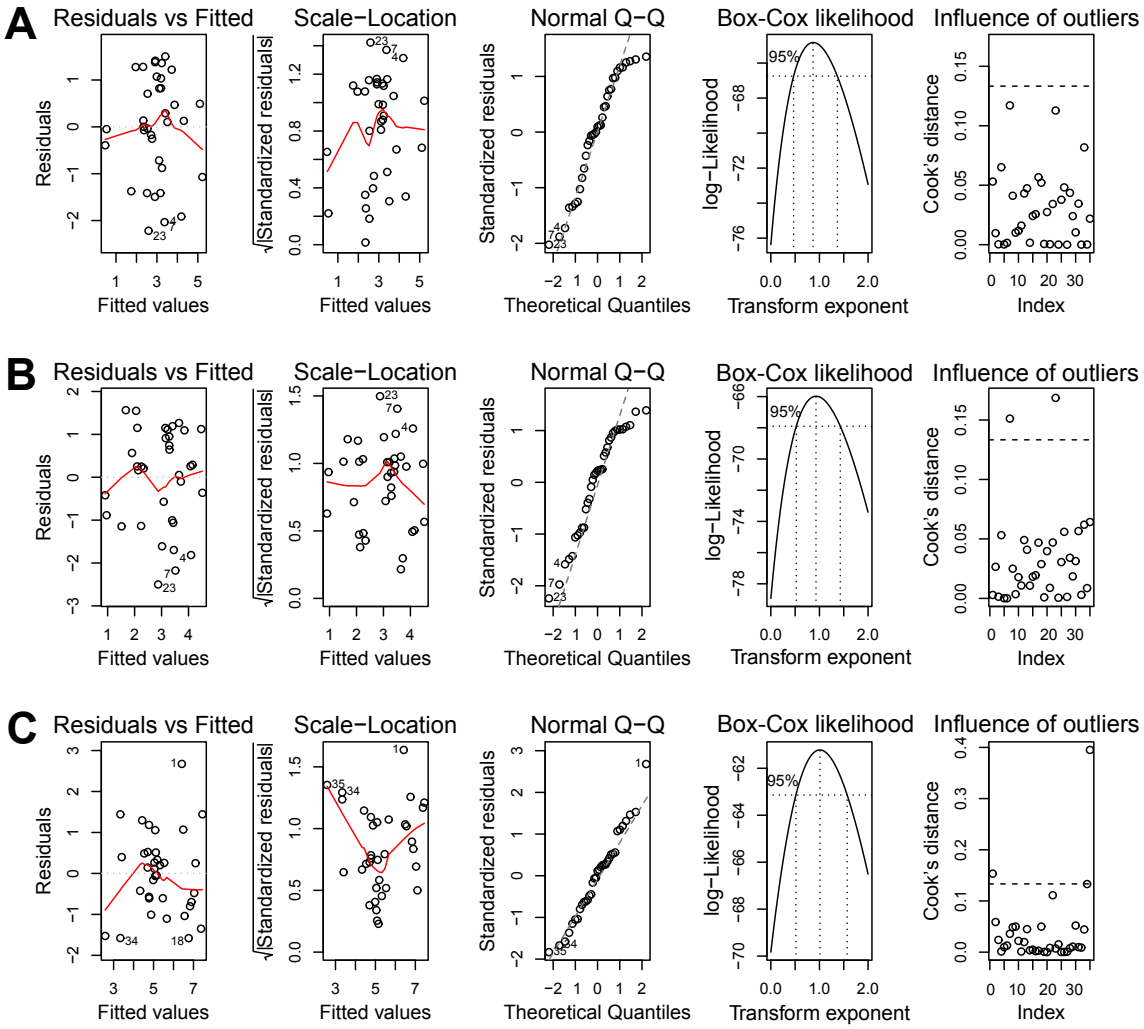


Figure 3.7. Diagnostic plots for linear models. First column: residuals versus fitted values; the trend line should be horizontal. **Second column:** absolute residuals (standardised and square-rooted) versus fitted values; the trend line should be horizontal. **Third column:** quantile-quantile plot of standardised residuals versus normal residuals; the points should lie on the dashed line. **Fourth column:** likelihood distribution of the parameter of a power transformation to correct for a skewed residuals distribution; if the most likely value is 1, no transformation is suggested. **Fifth column:** Cook's distance, which measures the influence of data points; those with high values are influential outliers. **A.** The model $relatedness \sim bias + location \times domain$. **B.** Results for the same model after transforming activation biases by adding 5 and then square-rooting the magnitude. **C.** The model $conservation \sim bias + location \times domain$.

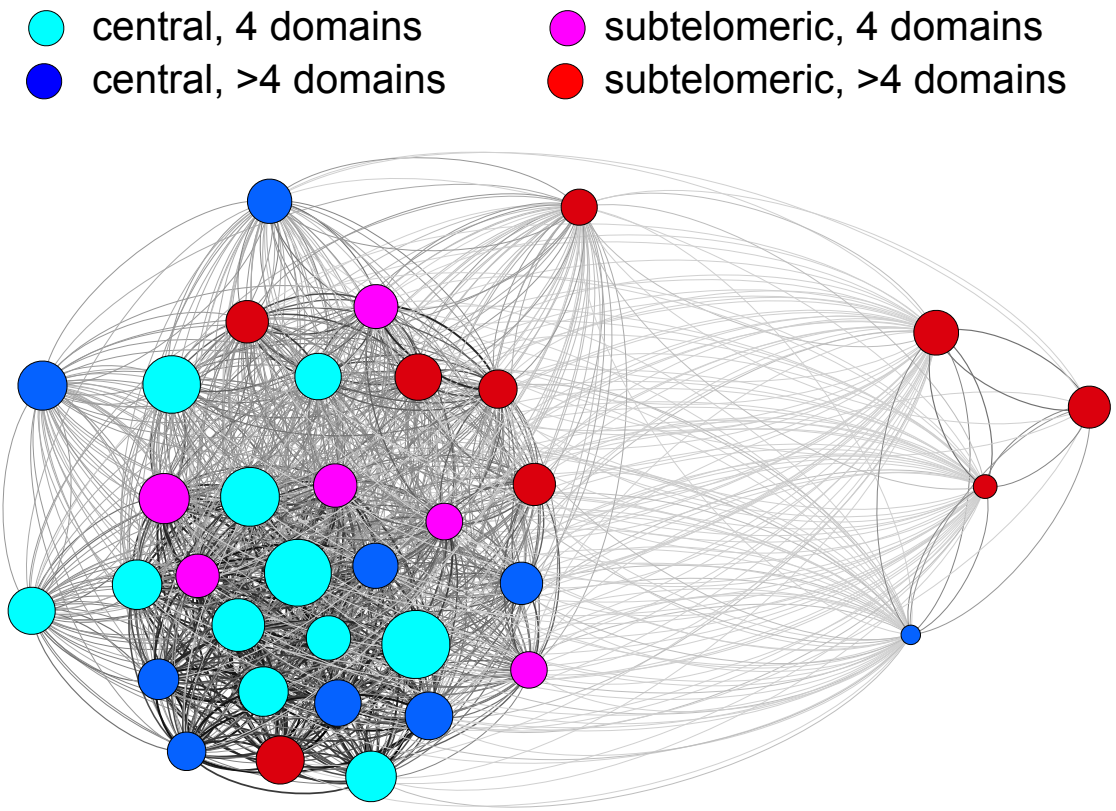


Figure 3.8. Relatedness network of the HB3 repertoire. Each node represents a *var* gene, colour coded according to chromosomal location and number of binding domains. Node size indicates estimated activation bias. The strength of a connection between two nodes is proportional to the number of shared homology blocks in the DBL1 and CIDR1 domains. Stronger connections are shown thicker and darker and the layout is force-directed.

from the first of the chosen genes to replace the corresponding block in the second gene. Homology blocks in all genes also undergo random mutation at a rate ten times lower than the gene conversion rate. The relatedness network is constructed and centrality scores are calculated in the same way as for the HB3 data. As expected, this simple model generates a strong positive correlation between activation bias and relatedness in terms of network centrality (Figure 3.10). A more sophisticated model incorporating epidemiology, host immunity and antigenic variation is considered in Chapter 5.

Under the same assumption, one would also expect genes with higher activation biases to be overall more diverse. Indeed, using the average sequence conservation of all domains as a measure of population level diversity [Buckee & Recker, 2012], this is again found to be significantly and positively correlated, independently of chromosomal location and gene length (Figure 3.11, $F = 15.8, p = 0.0004$ for the model comparison; accounting for uncertainty $p = 0.0006$), suggesting a role of active gene transcription in generating antigenic diversity in *P. falciparum* malaria. In this model there are two data points (indices 1 and 35) that especially skew the residuals distribution (Figure 3.7C, third plot) and disproportionately influence the model fit (Figure 3.7C, fifth plot). If these points are removed then the model comparison is still significant ($F = 9.4, p = 0.005$).

3.5 Validating the results

3.5.1 Initial conditions of the model

The model of switching dynamics requires a set of initial conditions. Assuming that each culture, excluding the parent, began with a single parasite and that populations

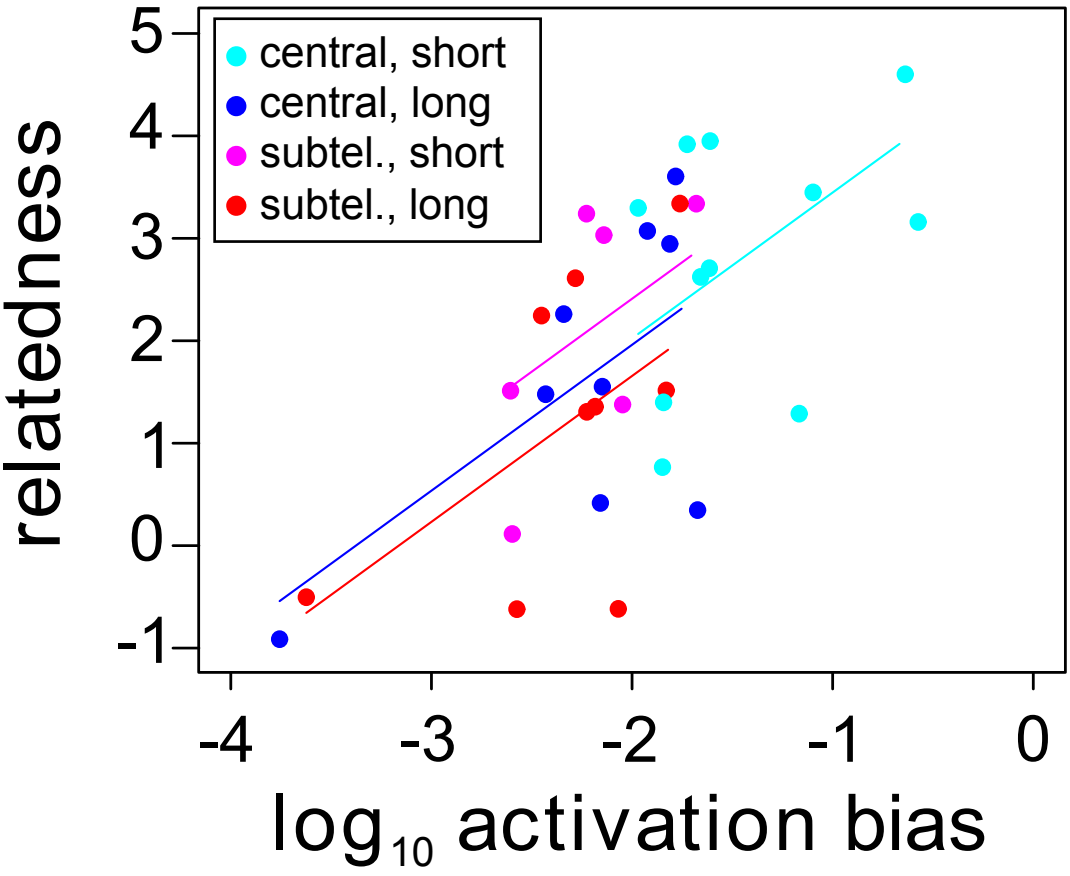


Figure 3.9. Observed effect of *var* gene activation on relatedness. *Var* gene activation biases are significantly correlated with each gene's relatedness to the rest of the HB3 repertoire, as measured by the gene's centrality within a shared homology block network, independently of chromosomal location or gene length. Trend lines show the fit of a linear regression model.

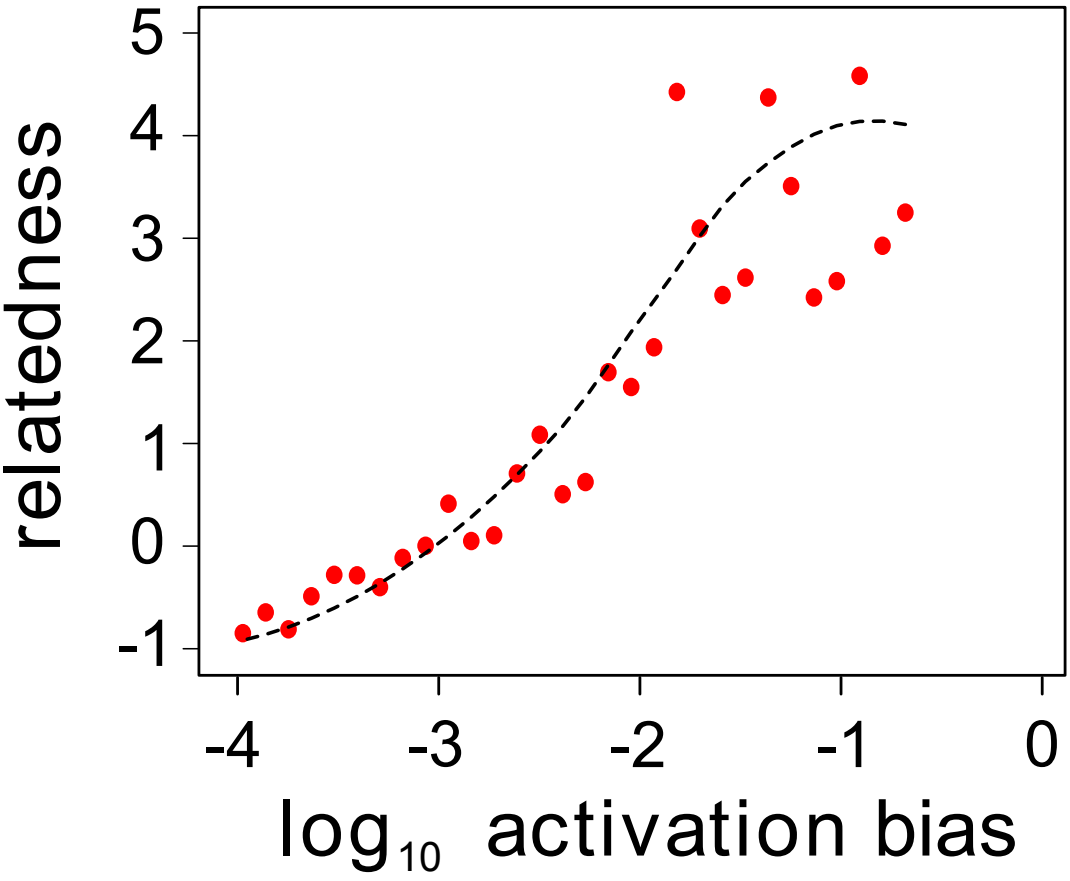


Figure 3.10. Simulated effect of *var* gene activation on sequence evolution. A simple mathematical model was used to simulate *var* gene evolution by gene conversion, whereby homology blocks are swapped among pairs of genes chosen at random according to their activation biases, shows a similar correlation between the genes' activation biases and their relatedness to the rest of the repertoire. The red points show the outcome of one simulation and the dashed line is the smoothed average of 50 runs.

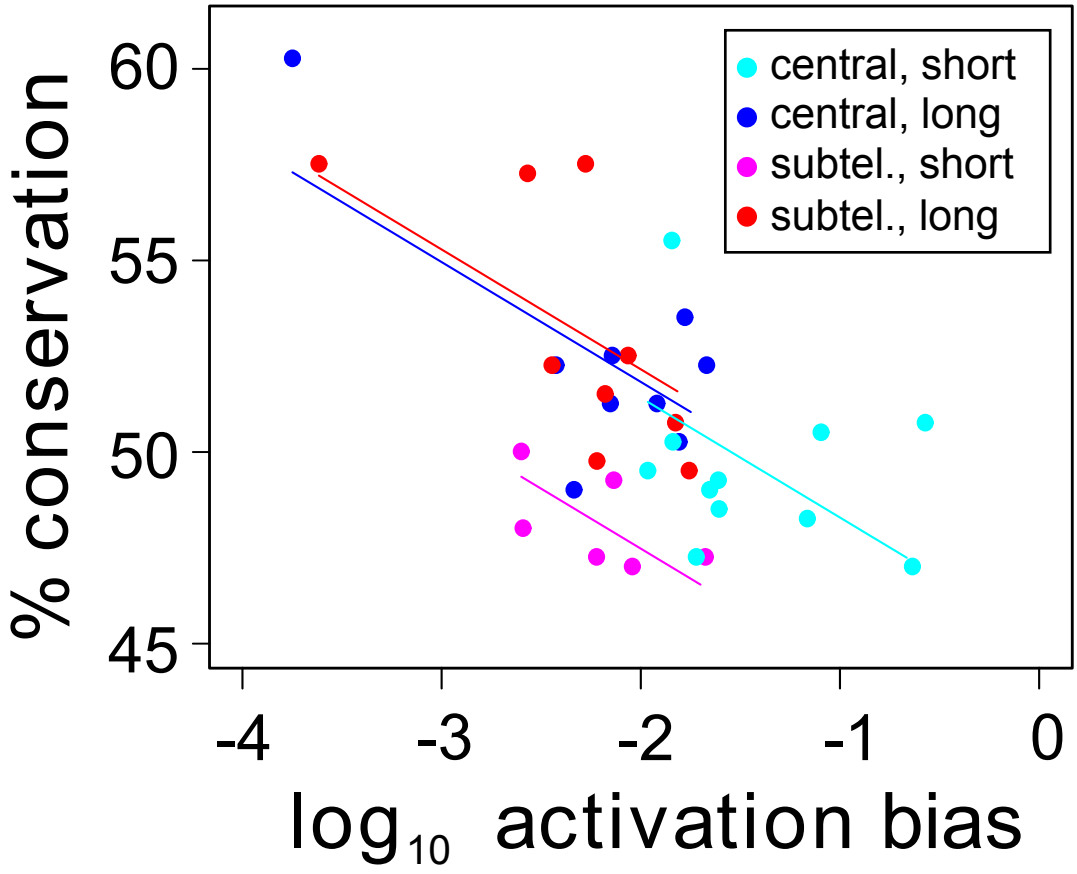


Figure 3.11. Observed effect of *var* gene activation on sequence diversity. Activation biases are negatively correlated with the gene's average domain sequence conservation, as a measure of population-level diversity, independently of chromosomal location and gene length. Trend lines show the fit of a linear regression model.

expanded approximately fivefold per generation, it follows that the first transcriptional switch most likely occurred, on average, two or three generations after cloning. Therefore in the analyses reported above it is assumed that each of these cultures remained clonal for the first three generations. For the parent culture, the first observation is used as the initial condition. If a switch had occurred in one or more of the initially clonal cultures within the first few generations, when the population was very small, then the assumed initial conditions would be inaccurate. However, this issue is found not to affect the results. When the algorithms are instead run using the first observations as the initial conditions for all cultures, the parameter estimates are highly consistent with those obtained using clonal initial conditions.

3.5.2 Cross-validation of parameter estimates

The simplest way to estimate switch biases and off-rates for a single *var* gene is to obtain a clone transcribing the gene, grow it in culture and analyse changes in transcript levels over time. A particular strength of the method used here is that it also estimates parameters for genes not initially transcribed by clones. To assess the accuracy of results obtained in this way, cross-validation is performed by sequentially excluding data sets. For example, to derive alternative estimates for *var29* the MCMC algorithm is run without the data for Culture 8 (which initially expressed *var29*), and to derive alternative estimates for *var27*, data for Cultures 5 and 6 are excluded. The results, shown in Figure 3.12, confirm my method's ability to estimate switch biases of non-starter genes with sufficient accuracy to determine the general switching pattern. Off-rate estimates are however not always accurate, especially for genes with very low transcript levels.

A similar method is used to test the accuracy of the method for estimating

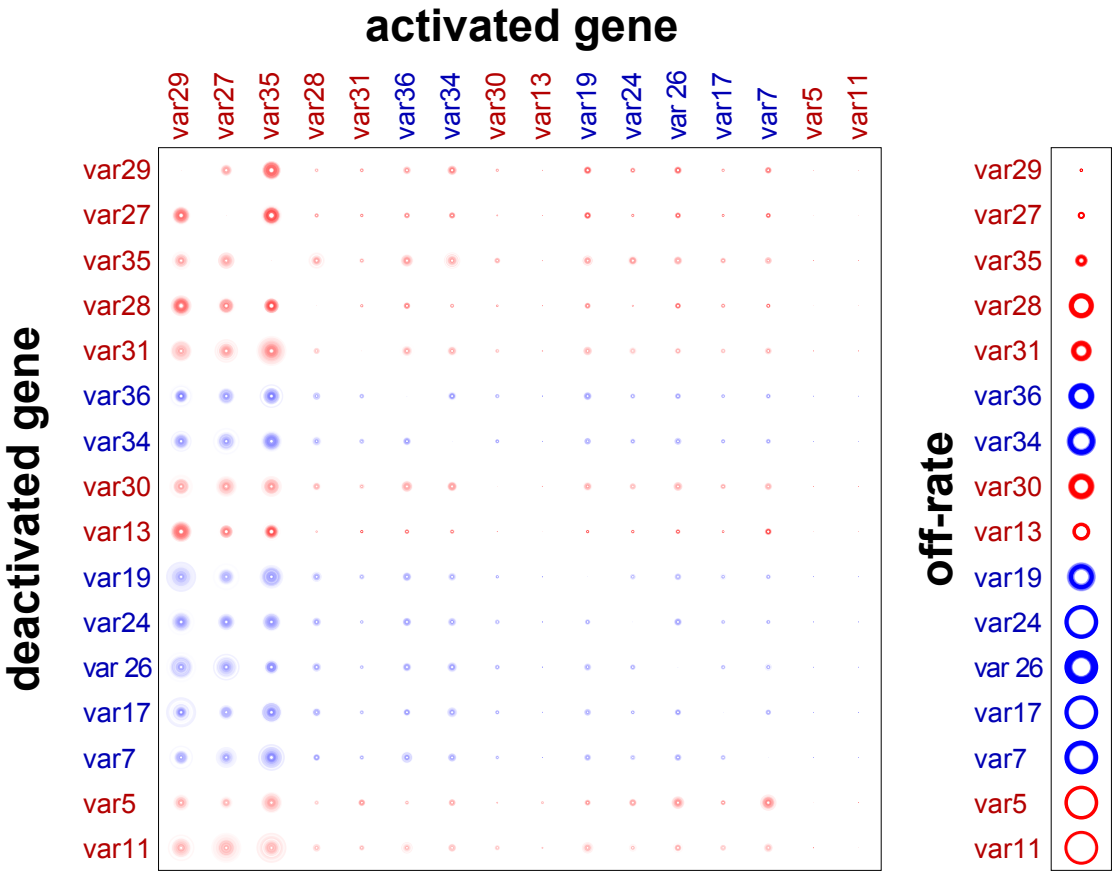


Figure 3.12. Validation of estimated switching parameters. The method is able to estimate switching parameters not only for starter genes, which were initially transcribed in the cultures, but also for other *var* genes. The accuracy of the latter estimates is tested using a cross-validation technique. For example, to generate an alternative set of parameter estimates for *var29*, the MCMC algorithm is run without the data for culture #8, which began as a clone transcribing *var29*. The matrix shown here is a composite of estimates derived in this way. The similarity to Figure 3.3 confirms that the results are reliable.

variations in switch biases. Cultures 2 and 3 initially expressed the same starter gene *var13*, and Cultures 5 and 6 both initially expressed *var27*. By excluding the data from Culture 2, estimates are derived for the *var13* switch biases mostly dependent on the Culture 3 data, and vice versa. Similarly, data from Culture 5 and Culture 6 are alternately excluded to obtain two sets of estimates for the *var27* switch biases. For each of the two different data sets, the deviations from the mean switch biases are found these to be positively correlated (Pearson's $\rho = 0.43$, $p = 0.015$), confirming the validity of the method.

3.5.3 Validation of model assumptions

The likelihood function derived in Chapter 2, used here for parameter estimation and model comparisons, assumes that input data are subject to measurement errors following a log-normal distribution. In Chapter 2 it was shown that the MCMC algorithm reliably determines network structures when the measurement error distribution is log-normal (base 2) with standard deviation $\sigma \leq 1$, and that σ can be reliably estimated using the simulated annealing algorithm. In the 16-gene model fitted to the HB3 data, $\sigma \approx 0.73$, which is well within the required range.

To test whether the measurement errors obeyed a log-normal distribution, the distribution of the deviations between the data and the simulated annealing output are analysed, after applying log transformations. The distribution for the 16-gene model is found to be unimodal and only slightly asymmetrical (skewness 0.4) but is more peaked than a normal distribution (excess kurtosis 2.4), so that it somewhat resembles a Laplace distribution. Running the simulated annealing algorithm with an error function that assumes log-Laplace-distributed noise makes very little difference to the results. The deviations for the constrained 38-gene model are similarly

distributed, excluding a very small minority of outliers.

Between the model output and the HB3 transcription data, unusually large deviations occur at 13 of the 64 observed time points, eight of which are between generations 48 and 52 and therefore belong to the same round of measurements. Transcript levels are found to deviate more for longer genes than for shorter genes. When the team that collected the data conducted northern blotting of the samples taken at the affected time points, they found unusually low levels of long, e.g. *var* length, RNA, confirming that the cause was not PCR error but most likely an as yet unidentified artefact of RNA preparation. Excluding the 13 data points affected by this phenomenon makes very little difference to the parameter estimates and statistical test results.

Analysis of the qRT-PCR replicates reveals that measurements of lower transcript levels have higher standard errors but the average standard error increases only threefold for each 1,000-fold difference in transcript level.

3.5.4 Growth rates

It has been suggested that *var* gene transcript levels in culture may be affected by small differences in parasite growth rates [Horrocks *et al.*, 2004], although no evidence has been found to support this hypothesis. To test whether such differences might affect estimates, the model is extended to include an excess growth rate parameter ϕ_i for each gene i:

$$v_{i,c,t+1} = (1-\omega_i)(1+\phi_i)v_{i,c,t} + \sum_{j \neq i} \omega_j \beta_{ji} (1+\phi_j) v_{j,c,t}.$$

When growth rates are allowed to vary by up to 0.5% per generation (equivalent

to approximately 50% over the duration of the experiment), the extended model is not significantly more likely than the model with only switch biases and off-rates. When the the upper bound on the excess growth rates is increased to 1% per generation, the extended model becomes significantly more likely. However, such large differences in growth rates *in vitro* are biologically implausible and have never been observed. Importantly, the inclusion of growth rate differences has very little effect on the switch bias and off-rate estimates.

3.6 Discussion

Several studies have noted that *var* genes located near the centres of chromosomes tend to be more highly transcribed *in vitro* than those in subtelomeric location [Enderes *et al.*, 2011, Fastman *et al.*, 2012, Frank *et al.*, 2007, Peters *et al.*, 2007, Zhang *et al.*, 2011], and it has been suggested that this might be due to low deactivation rates [Frank *et al.*, 2007]. Here it was found that high transcription levels are confined to a subset of central genes that encode only four PfEMP1 binding domains (termed Type 1 *var* genes [Gardner *et al.*, 2002]) and, importantly, are the result of high activation biases rather than low off-rates. That is, even though some of the most highly transcribed genes had very low off-rates, the total sum of other genes switching off with high bias towards these genes significantly outweighs the effect of off-rates. This is consistent with the results of analysing artificial transcription data in Chapter 2, which also showed that off-rates, at least within a biologically plausible range, can have only a modest effect on the total transcript levels within parasite populations.

Reanalysis of data from a recent study on the NF54 parasite strain by Fast-

man *et al.* [2012] supports these findings. Consistent with the results for HB3, centrally located *var* genes encoding only four binding domains are found to be more highly transcribed than the rest of the repertoire (Welch's $t = 2.3, p = 0.02$), with PF3D7_0809100, PF3D7_0421300 and PF3D7_0800100 being the most highly transcribed genes. According to the transcriptional switching model, these genes are predicted to have high activation biases, and indeed, PF3D7_0809100 has previously been reported as the most transcribed *var* gene in NF54 bulk culture whereas PF3D7_0421300 and PF3D7_0800100 were found to be the two genes most commonly activated by *var* gene switching in previous studies [Dzikowski *et al.*, 2006, Enderes *et al.*, 2011, Frank *et al.*, 2007]. Also in agreement with the HB3 results, NF54 gene transcription levels are found to be positively correlated with relatedness to the rest of the repertoire in terms of shared DBL1 and CIDR1 homology blocks (Spearman's $\rho = 0.3, p = 0.044$).

Here it was also found that the direction of transcriptional switching between *var* genes depends not only on intrinsic activation biases but also on the particular preferences of each deactivating gene. Even though these variations were found to have only a modest effect on *in vitro* switching, they are expected to play a much greater role during infection, where factors such as immune responses and binding affinities can significantly alter the *var* gene expression profiles. Furthermore, the combination of an activation hierarchy and these specific variations could explain a previously proposed source-sink model underlying antigenic switching in *P. falciparum* [Recker *et al.*, 2011], in which genes with high in-degrees, i.e. those most commonly switched to, correspond to the genes here identified as having high activation biases. Intriguingly, also, the results of the HB3 analysis suggest that switch biases may be stronger between genes in similar chromosomal locations: cen-

tral genes prefer to switch to other central genes, and subtelomeric genes prefer to switch to other subtelomeric genes. More data is needed to fully establish the factors influencing variation in switch biases, however, especially for less activated genes, where parameter estimations are more prone to uncertainties.

Despite recent advances in elucidating the genetics and epigenetics of *var* gene activation and silencing, not much is known about how mutually exclusive switching between the 60 *var* genes is controlled at the molecular level. The discovery of intrinsic *var* gene activation biases and their associations with chromosomal locations and other genetic features could therefore have important implications for future studies. These factors are clearly correlated, however, and it is uncertain how they are linked by cause and effect. For example, in a recent study it was found that gene length is positively correlated with sequence conservation [Buckee & Recker, 2012], and long *var* genes are more often positioned in subtelomeric regions. Nevertheless, the correlation between activation bias and relatedness, which was found to be independent of chromosomal location, gene length and upstream promoter, suggests a role of gene transcription in facilitating recombination and therefore in generating and maintaining antigenic diversity in *P. falciparum*. The nature of this role is unclear. One possibility is a molecular mechanism: transcription-associated recombination has been observed in other eukaryotes and may result from increased accessibility to recombination proteins or from stalled replication forks [Gottipati & Helleday, 2009, Prado & Aguilera, 2005]. An alternative explanation involving immune selection is considered in Chapter 5.

It is further possible that some *var* genes are under balancing selection, whereby the relative sequence conservation of longer genes could be due to optimized functionality, for example in terms of binding avidity and/or affinity. This assumption

would also help to reconcile apparent contradictions between observations reported here and *var* gene transcription *in vivo*. That is, various *in vivo* studies have reported UpsA *var* genes to be over-expressed during infections in individuals with little previous exposure [Bull *et al.*, 2005, Jensen *et al.*, 2004, Kyriacou *et al.*, 2006, Lavstsen *et al.*, 2005, Warimwe *et al.*, 2009], despite their low activation rates *in vitro*. Each *var* repertoire consists of a set of relatively long and conserved genes, such as those predominantly found with in the UpsA group, and another, short and diverse set of genes, such as those predominantly found within the UpsB and UpsC groups [Buckee & Recker, 2012]. Under the assumption that more conserved genes have an *in vivo* survival advantage, possibly due to higher binding affinities [Avril *et al.*, 2012, Claessens *et al.*, 2012, Lavstsen *et al.*, 2012], these variants will dominate early infections in naive hosts, regardless of activation rates. Their lower sequence diversity, however, will then lead to a more rapid acquisition of protective immunity, causing a transition towards the expression of more diverse variants, facilitated by their inherent activation biases. In light of this it can be speculated that the proposed order in which individuals acquire immunity (anti-UpsA followed by anti-UpsB/C), as well as the proposed associations between the expression of UpsA *var* genes and severe disease outcome, especially in young children, is less determined by the Ups group itself and is more a consequence of within-host selection of long, conserved genes, which are often but not exclusively found within the UpsA group. This hypothesis and its possible implications are explored theoretically in Chapters 4 and 5.

Together, these results demonstrate how selection pressures operating on multiple ecological scales have shaped the phenotypic plasticity embodied within the antigenic repertoire of *P. falciparum*. Within this setup, *var* genes and the molecular

mechanisms that underlie their sequential activation and silencing have co-evolved to allow the parasite to express the most advantageous phenotype in response to its current environment.

Chapter 4

Modelling the evolution of structure and expression order in the *Plasmodium falciparum* antigenic repertoire

4.1 Introduction

Plasmodium falciparum var genes can be categorised by various features such as upstream flanking sequence (Ups type), chromosomal location and number of binding domains [Kraemer *et al.*, 2007, Lavstsen *et al.*, 2003]. These classifications are not independent; for example, UpsA genes are typically subtelomeric and have five or more binding domains, whereas UpsC genes are typically centrally located and have only four binding domains. Understanding the diversity of *var* types is especially important because particular types have been associated with severe malaria

[Blomqvist *et al.*, 2010, Bull *et al.*, 2005, Falk *et al.*, 2009, Kaestli *et al.*, 2006, Kalmbach *et al.*, 2010, Kirchgatter & del Portillo, 2002, Kyriacou *et al.*, 2006, Rottmann *et al.*, 2006, Warimwe *et al.*, 2009]. Recent studies suggest that this association is largely due to extracellular domains encoded by a subset of UpsA and UpsB genes that mediate binding between infected red blood cells and the host receptor EPCR [Avril *et al.*, 2012, Claessens *et al.*, 2012, Lavstsen *et al.*, 2012, Turner *et al.*, 2013]. Another subset of UpsA genes cause infected cells to bind to uninfected red blood cells in a process known as rosetting, which is also linked to severe disease [Ghumra *et al.*, 2012, Rowe *et al.*, 2009].

There is evidence that recombination is more frequent between genes with similar characteristics and that UpsA and rosetting genes in particular are evolving separately [Bull *et al.*, 2008, Kraemer *et al.*, 2007]. UpsA genes also appear to encode fewer antigenically distinct variants [Rottmann *et al.*, 2006, Warimwe *et al.*, 2009]. Very little is known about the limits of antigenic diversity but one analysis of serological data from a Kenyan district hospital suggested that, in this particular population, there were in the order of 75 circulating antigenic variants of a more antigenically restricted type (which was supposed to comprise mainly UpsA genes) and 260-600 other variants [Buckee *et al.*, 2009]. The ratio of gene types within parasite repertoires is similarly unequal and, remarkably, is also highly consistent among sequenced repertoires from diverse geographical and epidemiological settings [Bull *et al.*, 2008, Rask *et al.*, 2010], as illustrated in Figure 4.1. Since repertoires frequently undergo gene duplication events and meiosis, this observation suggests that repertoire structure is subject to selection pressures.

One hypothesis is that different gene types are adapted for different ecological niches and that repertoires containing a mix of types therefore have a fitness advan-

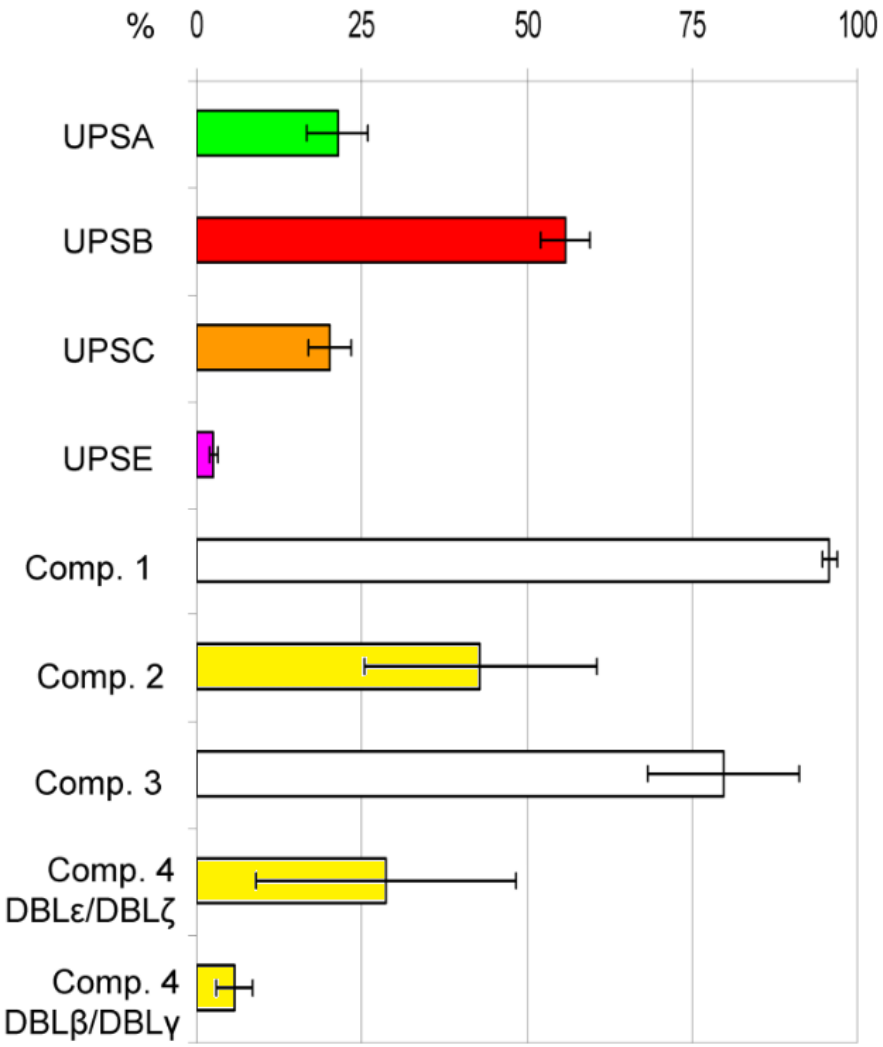


Figure 4.1. Composition of the *var* repertoires of seven *P. falciparum* isolates. The first four bars show the mean proportion of *var* genes per repertoire that have each upstream promoter type. The next five bars show the mean proportion of *var* genes per repertoire that contain each component. Here "components" are sets of DBL and CIDR extracellular domains. Components 2 and 4 (in yellow) are characteristic of genes with more than four domains. Error bars represent 95% confidence intervals. Figure adapted from Rask *et al.* [2010] using data from the seven sequenced genomes 3D7, HB3, DD2, IT4, PFCLIN, RAJ116 and IGH.

tage resulting from phenotypic flexibility. Buckee & Recker [2012] tested this idea using an evolutionary mathematical model with two gene types. Parasites expressing one gene type were assumed to have a transmission advantage (due to a longer infection period) but this type was also subject to higher cross immunity. That is, a more conserved gene type was optimised for between-host transmission success and a more diverse type for within-host survival. These assumptions were shown to result in the evolution of repertoires containing both gene types, in a system with two genes per repertoire [Buckee & Recker, 2012]. It may further be hypothesised that the trade-off between antigenic diversity and within-host fitness might cause parasite repertoires to evolve a particular, optimal ratio of gene types.

An alternative hypothesis is that host immunity is the single dominant factor that shapes *var* repertoire structures. Since UpsA genes have different binding properties from UpsB and UpsC genes [Smith *et al.*, 2013], it may be speculated that immune responses are more strongly cross-reactive within *var* types than between types. The possible effects of cross immunity on repertoire structure have not been investigated.

A related question concerns the order of *var* gene expression. In endemic areas, hosts first acquire antibody responses to UpsA and a subset of UpsB *vars* [Cham *et al.*, 2010, 2009, Warimwe *et al.*, 2009]. One hypothesis is that these genes are predominantly expressed in early infections because they confer a within-host survival advantage, perhaps because of their different binding phenotypes [Bull *et al.*, 2000, 1999, Smith *et al.*, 2013]. The same factor might help explain why such genes are associated with higher risk of severe disease.

This chapter introduces a suite of mathematical models of the evolution of *var* repertoires in *P. falciparum*, integrating antigenic variation, adaptive immunity, and transmission between hosts. The models are first used to test competing hypotheses

regarding the selection forces that shape antigenic repertoire structure. The results indicate that differential expression periods and levels of antigenic diversity are insufficient to drive the evolution of consistent structure in large repertoires. Conversely, it is shown that host immune responses can select for consistent repertoire structure.

Next, similar models are used to investigate the evolution of expression order. In this case, in contrast to the previous results, it is found that the different attributes of the gene types are critical. Assuming that one gene type confers a higher transmission rate, parasites will evolve to express this type first, even at the cost of increased host mortality. This leads to a speculative hypothesis regarding the evolution of virulence in *P. falciparum*.

4.2 Models of repertoire structure evolution

A stochastic, individual-based model is used to describe antigenic variation and acquisition of immunity within hosts, and transmission of strains between hosts. The structure of the algorithm is presented as pseudocode in Algorithm 1. In the model, hosts may be infected by parasites with diverse repertoires of antigen-encoding genes. Each of these genes belongs to one of two types labelled α and β. Thus, each antigenic repertoire comprises an α sub-repertoire and a β sub-repertoire. It is assumed that there are limited numbers of possible genes of each type and that different genes encode distinct antigenic types. During infection, genes from either sub-repertoire are exclusively expressed within the host in a random order. Each gene has a limited expression period, representing the time taken for the host to develop effective adaptive immune responses to the encoded antigen. Immunity is assumed to be long lasting so the same gene cannot be expressed more than once

in the lifetime of a host. If a host becomes immune to the whole of the repertoire of an infecting strain then the strain is no longer transmissible. All infections have a maximum length, which is generally less than the time needed to express the whole of the repertoire. A limit on infection length may be assumed to result from non-variant-specific immune responses. Nevertheless, the limit was conservatively set towards the upper end of plausible values so that it would not unduly affect the results.

Transmission is modelled with random mixing of hosts, ignoring vectors, since vector dynamics are not expected to have much influence on the evolutionary processes considered here. Each host can harbour multiple simultaneous infections with distinct strains, up to a certain number, but these strains do not directly interact. Since strains can become extinct, random parasite strains are continually introduced into the population at a low rate to maintain diversity.

4.2.1 Host births and deaths

The host mortality rate depends on age following a Gompertz function Ae^{Bx}, where x represents host age and A and B are constants. Figure 4.2A shows how mortality rate changes with age, and Figure 4.2B shows the resulting distribution of lifespans. To maintain a constant population size, each death is immediately followed by the birth of a fully susceptible host.

4.2.2 Assessment of repertoire structures

The first aim is to determine which factors can lead to the evolution of a parasite population that is homogeneous with respect to α and β sub-repertoire sizes. This

```
for each parameter set:
    for each generation:
        for each host and each of their infections:
            Decrement infection length countdown;
            if infection length countdown is zero:
                Clear the infection;
                Make host immune to all genes in list of expressed genes;
        for i = 0 to number_of_hosts × transmission_rate:
            Select a random pair host A and host B;
            if host A is infected:
                select random infections to transmit;
            for each infection to be transmitted:
                Apply mutation;
                Transmit infection to host B;
                Set infection length countdown to zero;
                for each gene in the repertoire, in order of expression:
                    if host B is not immune to the gene:
                        Add gene to list of expressed genes;
                        Add gene's expression period to infection length
                        countdown;
                        if infection length countdown has reached maximum:
                            Exit for loop;
        for i = 0 to introduction_rate:
            Select a random host;
            Generate a random repertoire;
            Transmit infection to host;
            Set infection length countdown to zero;
            for each gene in the repertoire, in order of expression:
                if host is not immune to the gene:
                    Add gene to list of expressed genes;
                    Add gene's expression period to infection length countdown;
                    if infection length countdown has reached maximum:
                        Exit for loop;
        for each host:
            Increment host age;
            Dependent on host age, implement death and birth;
```

Algorithm 1: Pseudocode version of the algorithm used to implement the evolutionary model.

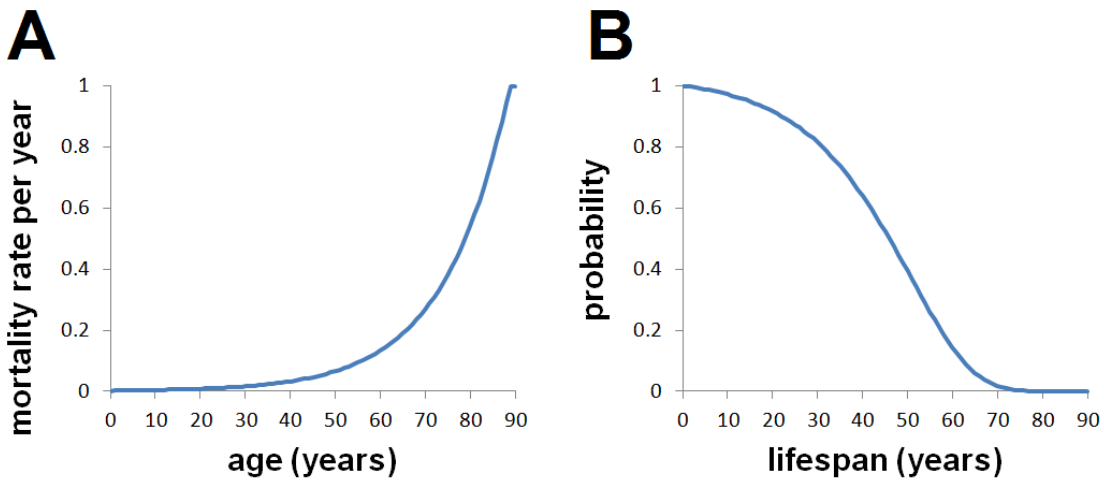


Figure 4.2. Host mortality. A. Mortality rate versus host age. Mortality rate was Ae^{Bx}, where x = host age in years, $A = 0.0002$ and $B = 0.07$. **B.** The resulting lifespan probability distribution.

is equivalent to the population distribution of α or β sub-repertoire sizes having a single, sharp peak. Distributions resulting from the models are assessed using four measures of shape: variance, kurtosis, Pearsons chi-squared statistic, and Hartigan's dip test statistic for unimodality (implemented by the diptest package for R). Assessment criteria using these four measures were selected empirically by testing on subsets of distributions.

4.2.3 Model comparisons

Where possible, parameter values are informed by empirical studies. One hypothesis to be tested invokes differential transmission rates, depending on the type of expressed gene. In the models considered here, as in a previous study [Buckee & Recker, 2012], transmission rates are assumed to correlate with expression periods. Therefore α and β expression periods are treated as independent variables, and sepa-

Table 4.1. Variable and parameter values of repertoire structure models

Independent variable	Values	Refs
Expression period per α gene (parasite generations)	6, 8, 12, 16	[1,2]
Expression period per β gene (parasite generations)	6, 8, 12, 16	[1,2]
Varied parameter	**Values**	**Refs**
Numbers of possible variants N_α, N_β	{50,300}, {50:600}, {75, 450}, {100, 300}, {100, 600}	[3]
Contact rate per host per generation	0.01, 0.1	-
Constant parameter	**Value**	**Refs**
Repertoire size	60	-
Host population	10000	-
Maximum multiplicity of infection	2	[4,5]
Maximum infection length (parasite generations)	288	[1,2,6]
Introduction of new strains (per parasite generation)	0.01	-

[1] = Collins & Jeffery [1999b]; [2] = Collins & Jeffery [1999a]; [3] = Buckee *et al.* [2009]; [4] = Mueller *et al.* [2012]; [5] = Koepfli *et al.* [2011], [6] = Bretscher *et al.* [2011].

rate models are run for multiple combinations of expression periods within plausible ranges. Two other factors are treated as parameters with vague priors. These are host contact rate, and numbers of possible α and β variants (N_α and N_β), which are not precisely known. Each model is run repeatedly for different combinations of values of these two parameters. Thus, for each pair of expression periods, there are multiple sets of results. It is assumed that, a priori, all sets of parameter values are equally probable. Therefore, from a Bayesian point of view, the most likely model is the one that produces the desired outcome for the highest proportion of parameter sets. All values of independent variables and parameters are listed in Table 4.1. Selected results from testing wider ranges of variable and parameter values are also reported.

Initially, models are seeded with 1,000 parasite strains with random sub-repertoire

sizes and random genes within each sub-repertoire. The models are stopped after 50,000 parasite generations (approximately 274 years, assuming 48 hours per generation). On a standard desktop computer (quad core, 3 GHz, 4 GB RAM), each model can be run for all parameter sets within 24 hours.

4.3 Expression order model

The model described above is modified to investigate the evolution of sub-repertoire expression order. The numbers of α and β genes in repertoires are not allowed to vary; instead, their ratio is always the same as the ratio of α and β variants in the population of possible variants (N_α:N_β), consistent with observations of PfEMP1 antigenic diversity [Buckee *et al.*, 2009] and *var* sequence diversity [Barry *et al.*, 2007, Chen *et al.*, 2011, Falk *et al.*, 2009, Mugasa *et al.*, 2012, Trimnell *et al.*, 2006, Ward *et al.*, 1999]. Diversity is maintained by occasional introduction of novel strains and also by mutation, such that each gene has an equal probability of mutating into any other possible gene of the same type. That is, an α gene can mutate into any other α gene but cannot mutate into a β gene, and vice versa. Mutation is modelled at the time of transmission so has no effect on infection dynamics within the transmitting host.

Expression order is used to mean the order in which genes are predominantly expressed in the infecting parasite population. It is assumed that this order may depend on within-host survival rates as well as activation biases. Each strain has one of three possible orders of expression:

- α genes always expressed before β genes,
- β genes always expressed before α genes,

- equal probabilities of an expressed gene being selected from the α sub-repertoire or the β sub-repertoire.

Rarely, a strain is mutated to have a different ordering. Thus this model structure can be used to explore evolution of the order of expression. In accordance with the findings of Chapter 3, there is an expression hierarchy within each sub-repertoire.

Initially, models are seeded with 500 parasite strains with random repertoires and orders. After 5,000 parasite generations, all infections are cleared and the model is seeded again with 500 strains to lessen founder effects. With the above parameters, infection prevalence stabilises between 35% and 40%. Models are stopped after 50,000 parasite generations (approximately 274 years, assuming 48 hours per generation). Additional parameter values are listed in Table 4.2.

Table 4.2. Parameter values of expression order model

Parameter	Value(s)	Refs
Expression period per gene (parasite generations)	10	[1,2]
Number of possible α gene variants, N_α	64, 256	[3]
Number of possible β gene variants, N_α	256	[3]
Contact rate per host per parasite generation	0.01	-
Repertoire size	64	-
Host population	10,000	-
Maximum multiplicity of infection	2	[4,5]
Maximum infection length (parasite generations)	160, 320	[1,2,6]
Introduction of new strains (per parasite generation)	0.1	-
Gene mutation rate (per gene per transmission event)	0.0001	-
Order mutation rate (per transmission event)	0.001	-

[1] = Collins & Jeffery [1999b]; [2] = Collins & Jeffery [1999a]; [3] = Buckee *et al.* [2009]; [4] = Mueller *et al.* [2012]; [5] = Koepfli *et al.* [2011], [6] = Bretscher *et al.* [2011].

4.4 Results of repertoire structure models

4.4.1 Repertoire structure

First to be considered are the simplest conceivable models, in which parasites express genes from their repertoires in random order, regardless of type. It is assumed that α and β genes may have different expression periods and that there are fewer possible α variants than β variants. For each combination of expression periods within plausible ranges, these models rarely result in the evolution of consistent repertoire structures (Figure 4.3A). Instead of being concentrated around a peak, distributions are generally broad and highly skewed (Figure 4.4). Consistent repertoire structures emerge only when the α expression period is at least three times longer than the β expression period (24 versus 8 or fewer parasite generations, not shown).

In light of this result, it may be supposed that each gene can be expressed for a long period within each host, but that this expression period is spread over several infections. In other words, hosts acquire lifelong immunity to each variant only after multiple exposures. This assumption only marginally improves model outcomes (Figure 4.3B).

Alternatively (or in addition), it may be supposed that expression order matters, since in *P. falciparum*, genes of the more restricted type appear to be expressed first in non-immune hosts. Note that expression order here means the order in which genes are predominantly expressed in the infecting parasite population, which may depend on within-host fitness as well as the activation biases considered in Chapter 3. Also, expression order here is regarded as fixed for the entire parasite population (the evolution of expression order is considered separately, later in this chapter). Therefore models may be modified so that α genes are expressed before β genes.

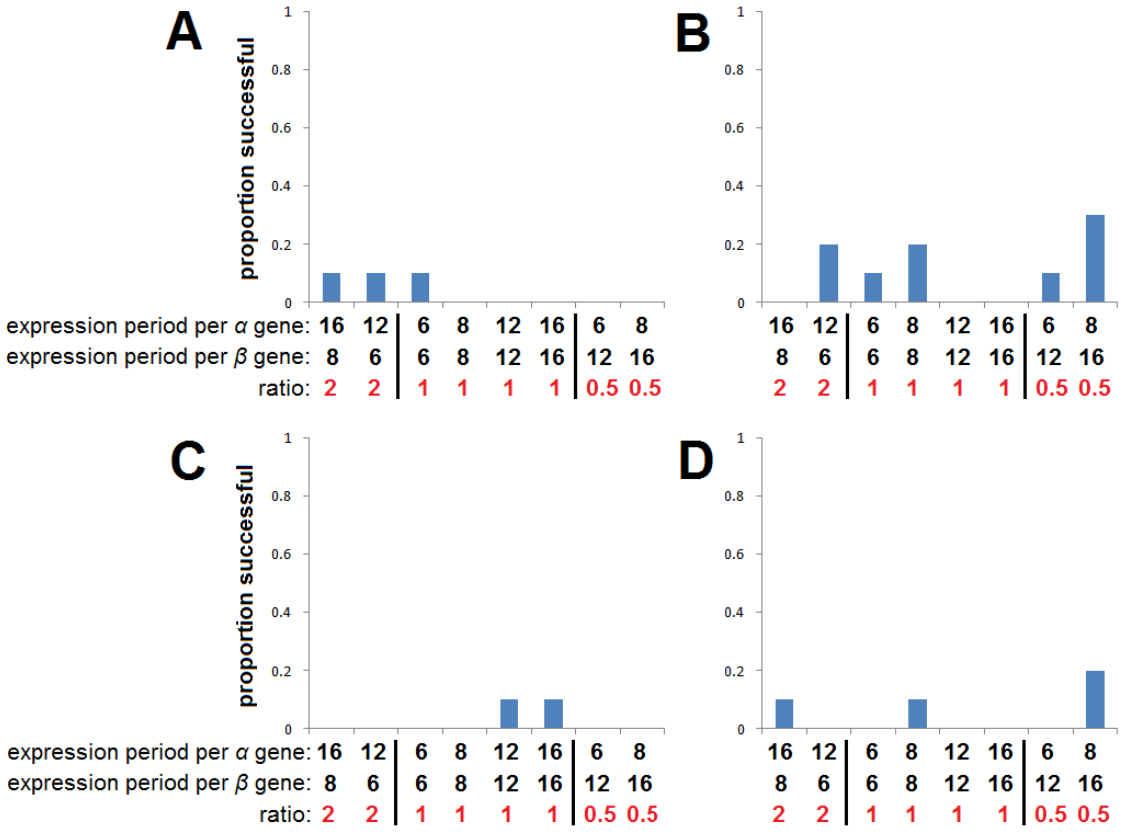


Figure 4.3. Model results without cross immunity. Models are categorised according to assumptions regarding expression order and acquisition of immunity. Within each category there are 8 models with different combinations of α and β expression periods (measured in parasite generations). Each of these models is run for multiple combinations of parameter values and the proportion of "successful" runs resulting in consistently structured repertoires is calculated. Models with lower scores are judged less likely to be correct. Parameter values are as listed in Table 4.1. **A.** Random expression order. **B.** Random expression order and immunity acquired after four exposures to a variant. **C.** α genes expressed before β genes. **D.** α genes expressed before β genes according to a global expression hierarchy.



Figure 4.4. A sample of sub-repertoire size distributions resulting from models without cross-immunity. Distributions in black satisfy the criteria for being concentrated around a single peak, indicating consistent repertoire structure, whereas those in grey do not. In this category of models, even the few distributions that meet the criteria are highly skewed and the mode number of α genes is typically close to zero.

That is, an infecting strain will never predominantly express β genes unless the host is immune to all α genes in its repertoire. As shown in Figure 4.3C, such models are unsuccessful in generating consistent repertoire structure.

Serological analyses [Buckee *et al.*, 2009, Cham *et al.*, 2010, 2009] indicate that immunity is acquired to the most frequently recognised variant surface antigens in a predictable order. This pattern may lead to speculation about a universal *var* expression hierarchy, perhaps related to within-host fitness, such that the order of expression is the same for all repertoires (that is, if variant i is above variant j in the hierarchy then i will be expressed before j in all repertoires containing both variants). Models incorporating a universal expression hierarchy also generally fail to generate consistent repertoire structure for plausible parameter sets (Figure 4.3D).

Decreasing the maximum infection length or increasing the maximum multiplicity of infection in the above models does not improve the results (not shown).

Finally, models are extended to include cross immunity. The simplest, most general way to do this is to make each expressed gene have an expression period shorter than the previously expressed gene of the same type, within the same infection, according to a scaling factor. For a range of scaling factors, a majority of plausible parameter sets result in consistently structured repertoires for several combinations of expression periods (Figure 4.5). Similar results are seen when α genes are expressed before β genes (Figure 4.6). Compared to the results without cross immunity, the distributions of sub-repertoire sizes are generally narrower and more symmetrical (Figure 4.7). All results are summarised in Figure 4.8.

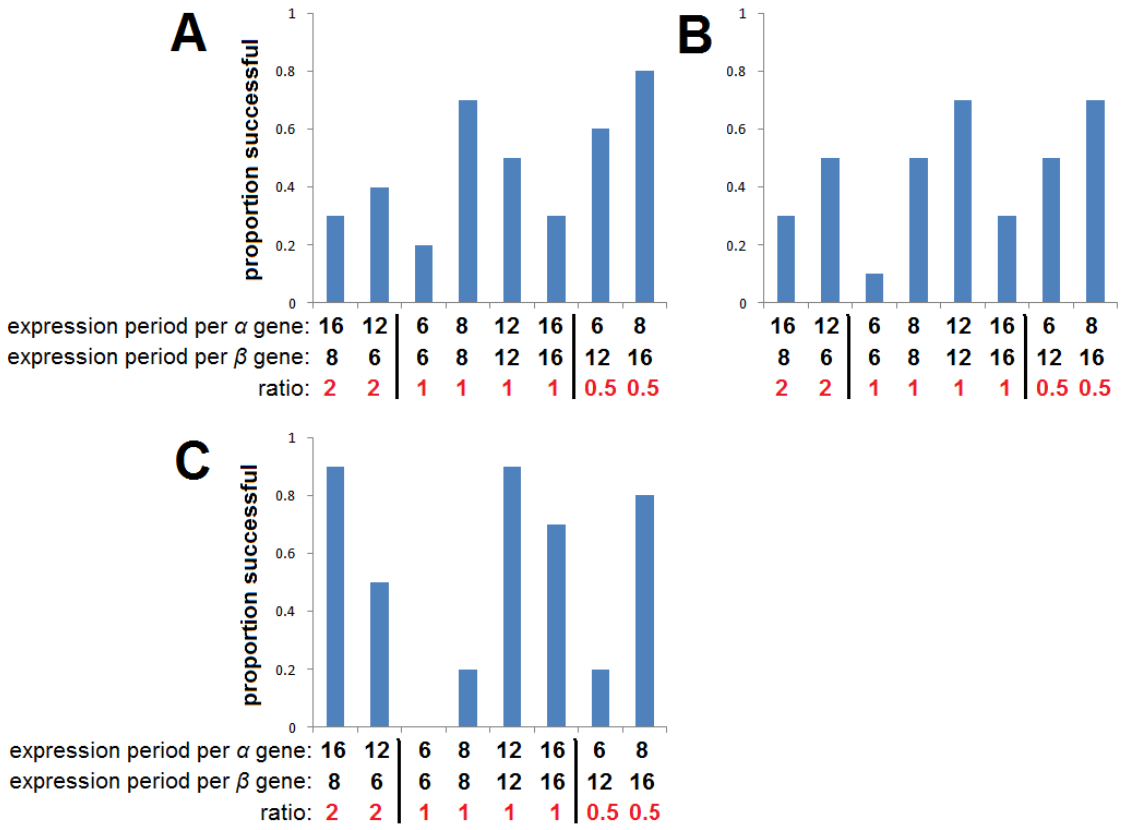


Figure 4.5. Model results with cross immunity. Models are categorised according to the strength of cross immunity, determined by a scaling factor. Smaller scaling factors correspond to stronger cross immunity. Within each category there are 8 models with different combinations of α and β expression periods (measured in parasite generations). Each of these models is run for multiple combinations of parameter values and the proportion of "successful" runs resulting in consistently structured repertoires is calculated. Models with lower scores are judged less likely to be correct. Parameter values are as listed in Table 4.1. **A.** Scaling factor 0.95. **B.** Scaling factor 0.92. **C.** Scaling factor 0.87.

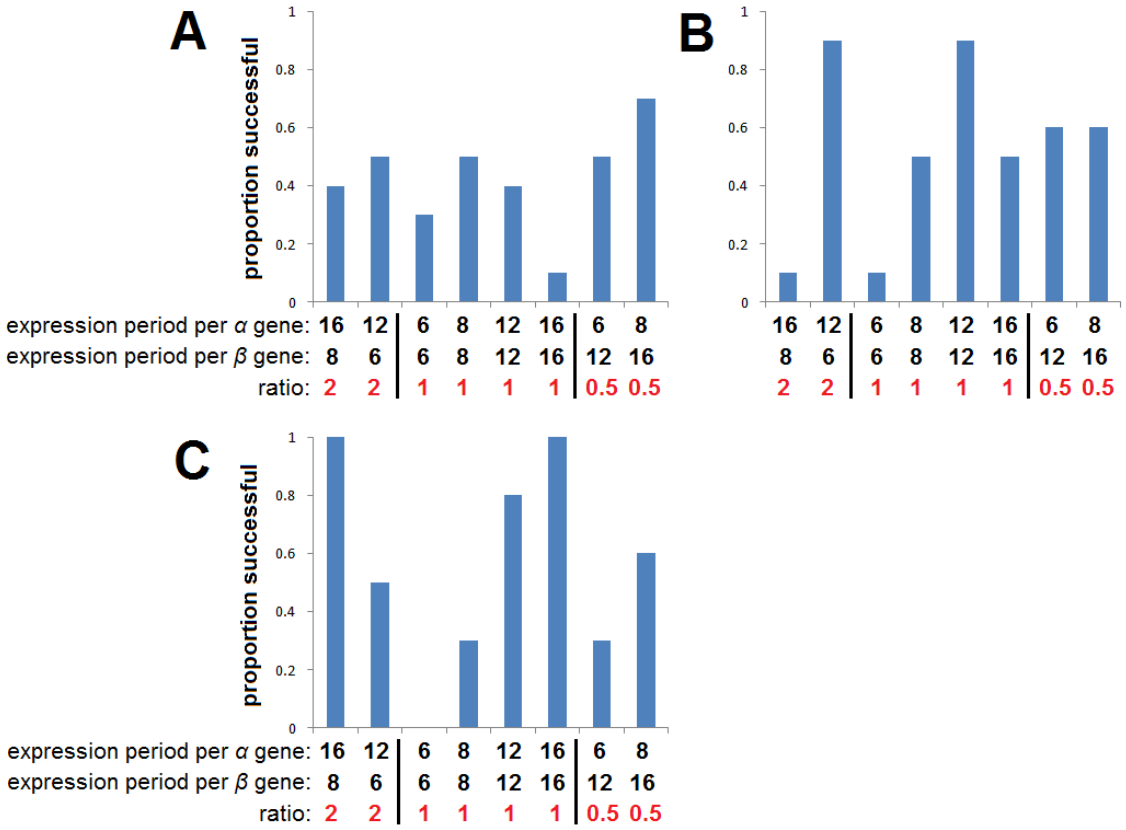


Figure 4.6. Model results with cross immunity and α genes expressed before β genes. MModels are categorised according to the strength of cross immunity, determined by a scaling factor. Smaller scaling factors correspond to stronger cross immunity. Within each category there are 8 models with different combinations of α and β expression periods (measured in parasite generations). Each of these models is run for multiple combinations of parameter values and the proportion of "successful" runs resulting in consistently structured repertoires is calculated. Models with lower scores are judged less likely to be correct. Parameter values are as listed in Table 4.1. **A.** Scaling factor 0.95. **B.** Scaling factor 0.92. **C.** Scaling factor 0.87.

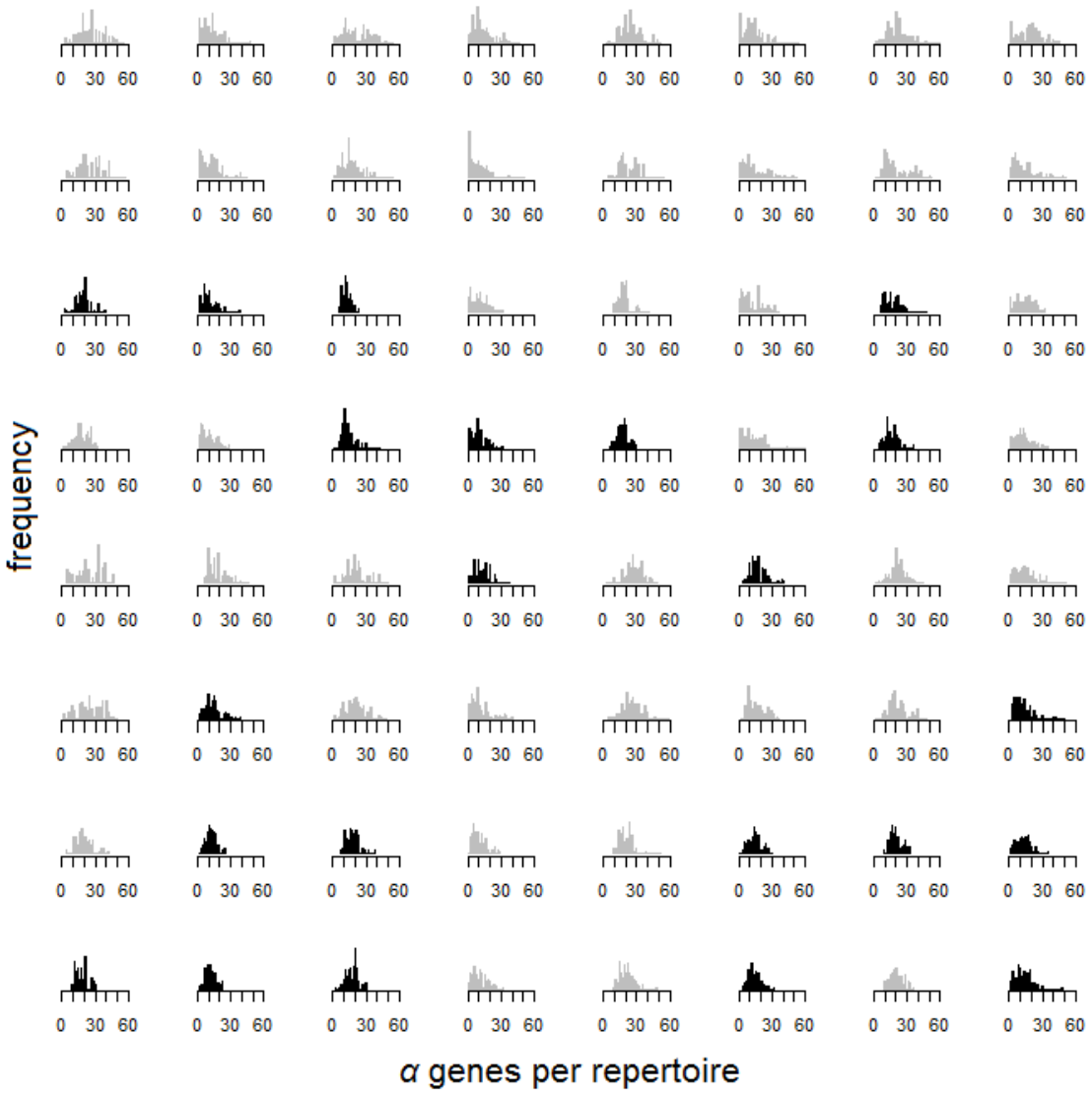


Figure 4.7. A sample of sub-repertoire size distributions resulting from models with cross-immunity. Distributions in black satisfy the criteria for being concentrated around a single peak, indicating consistent repertoire structure, whereas those in grey do not. The distributions are generally narrower and more symmetrical than those resulting from models without cross-immunity.

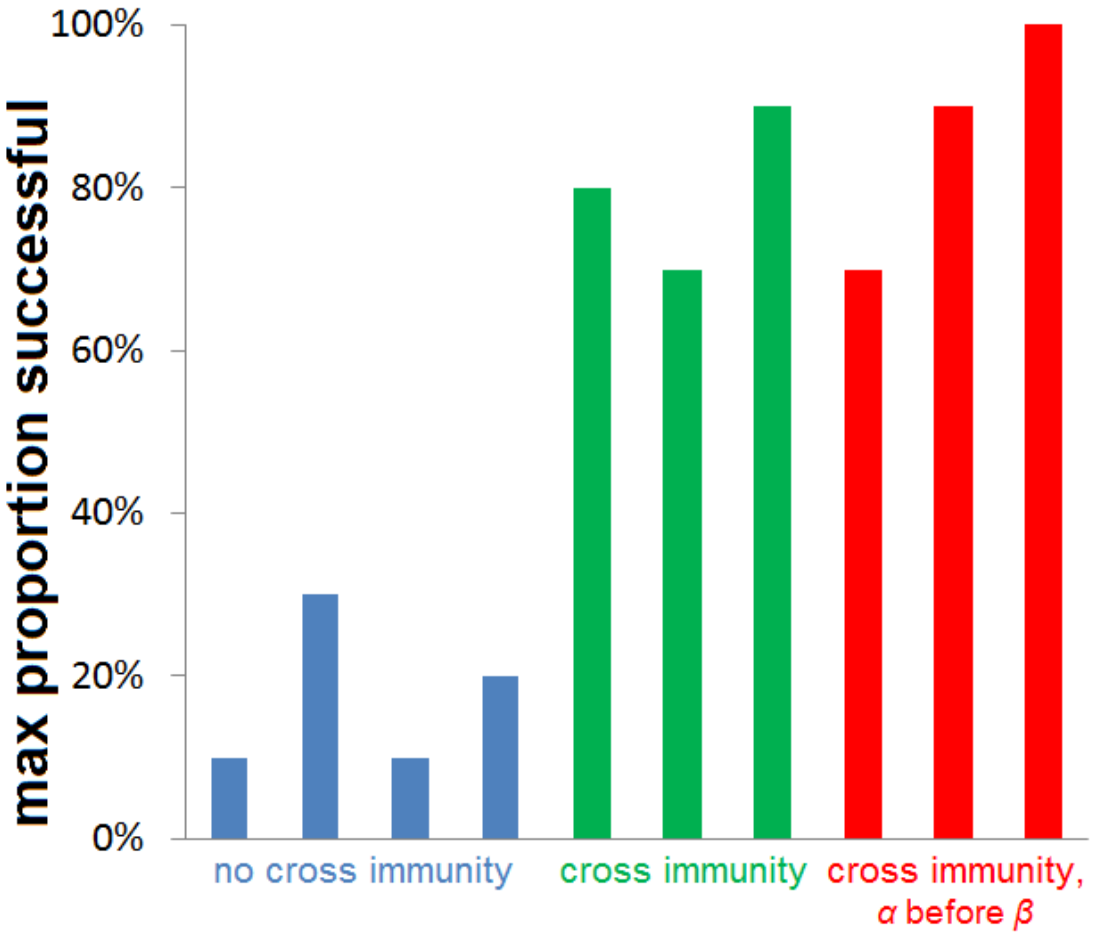


Figure 4.8. Summary of repertoire structure results. This chart summarises the results shown in Figures 3, 5 and 6 by showing the maximum proportions of "successful" outcomes for models of each category. Here, successful models are those that generate consistent repertoire structure. Thus, the first column is the same as the tallest column in Figure 4.3A, the second column is the same as the tallest column in Figure 4.3B, and so on. Models with lower scores are judged less likely to be correct. Blue columns from left to right: random expression order; random expression order and immunity acquired after four exposures to a variant; α genes expressed before β genes; α genes expressed before β genes with global expression hierarchy. Green columns: random expression order with cross reactive immunity scale factor 0.95, 0.92 and 0.87, respectively. Red columns: same as second set but with α genes expressed before β genes. Parameter values are as listed in Table 4.1.

4.4.2 Sub-repertoire sizes

So far, parameter sets have been judged on whether they result in consistent numbers of α and β genes in repertoires, no matter what those numbers might be. To fit what is observed in *P. falciparum* (Figure 4.1), the mode number of α genes per repertoire should be smaller than the number of β genes and not too close to zero. This corresponds to a mode between approximately 8 and 20 α genes per repertoire of 60 genes.

Without cross immunity, model results show that the mean proportion of α genes per repertoire is affected by the proportion of α genes in the population of possible variants, N_α / (N_α + N_β), and by the ratio of the α and β infection periods. This mean frequently falls between 8 and 20 α genes per repertoire. However, the distributions of sub-repertoire sizes are generally highly skewed, so the mode proportion of α genes per repertoire is substantially smaller than the mean (Figure 4.4). This means that, even among distributions narrow and peaked enough to meet the shape criteria, the mode number of α genes is generally close to zero (Figure 4.9). When α genes are expressed before β genes, repertoires contain even fewer α variants. This is because selection pressure applies only to the fraction of the repertoire typically expressed within the lifetime of a host, and the unexpressed portion of the repertoire must necessarily consist entirely of β genes. Therefore, when distributions are filtered by mode number of α genes as well as by distribution shape, the models without cross immunity become even less likely (Figure 4.10).

In the models with cross immunity, the distributions of sub-repertoire sizes are narrower and more symmetrical, and cross immunity makes it advantageous for parasites to split their repertoires more evenly (Figure 4.7). Therefore the mode number of α genes more frequently falls between 8 and 20 per repertoire (Figures 4.10

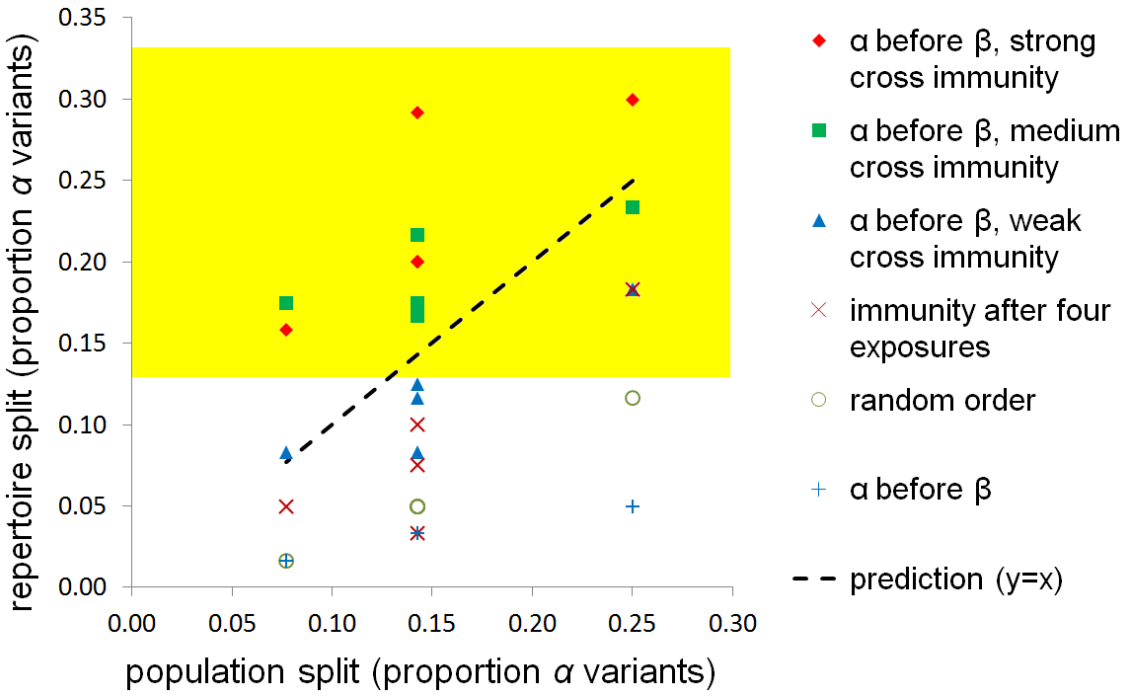


Figure 4.9. Ratio of α and β genes per repertoire versus ratio of possible α and β variants. Shown here are the medians for all combinations of expression periods and parameter values. For each model category there are five (possibly overlapping) points corresponding to five combinations of α and β population sizes (that is, numbers of possible variants, N_α and N_β). The highlighted region corresponds to between 8 and 20 α genes per repertoire. Parameter values are as listed in Table 4.1.

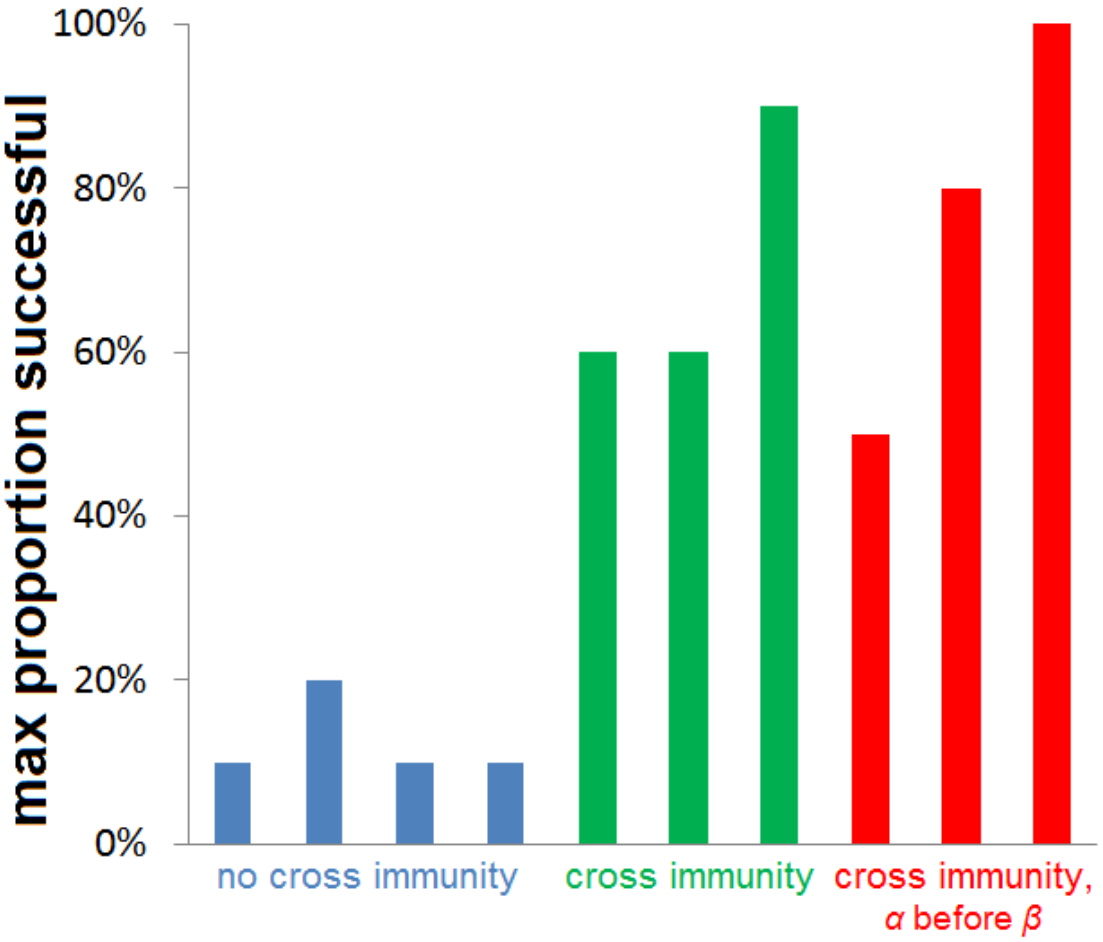


Figure 4.10. Model results after filtering for sub-repertoire size. Column order is the same as in Figure 4.8. Shown here are the maximum numbers of "successful" outcomes for models of each category, where successful models are those that generate consistently structured repertoires containing between 8 and 20 α genes (mode average). Parameter values are as listed in Table 4.1.

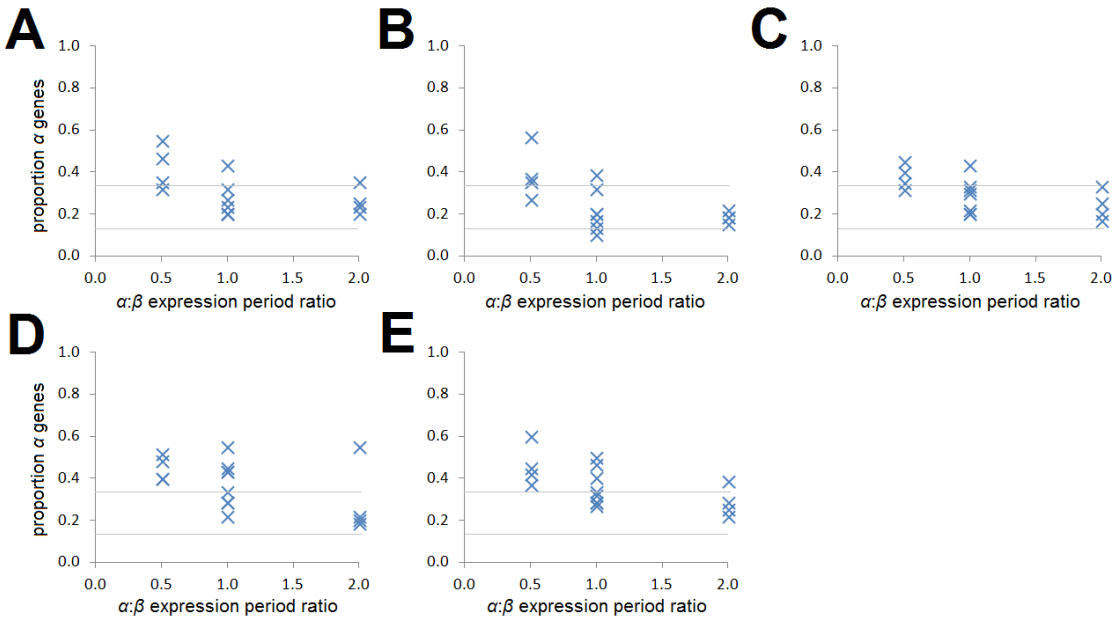


Figure 4.11. Effect of expression periods on sub-repertoire size. Each data point represents the mode number of α genes per repertoire resulting from a model with a unique combination of expression periods and transmission rates. Grey lines indicate the region with between 8 and 20 α genes per repertoire. **A.** $N_\alpha = 50$, $N_\beta = 300$; **B.** $N_\alpha = 50$, $N_\beta = 600$; **C.** $N_\alpha = 75$, $N_\beta = 450$; **D.** $N_\alpha = 100$, $N_\beta = 300$; **E.** $N_\alpha = 100$, $N_\beta = 600$. Other parameter values are as listed in Table 4.1.

and 4.9). The number of α genes per repertoire is negatively correlated with the ratio of α and β expression periods (Figure 4.11). For a given expression period ratio, the number of α genes per repertoire is further negatively correlated with the magnitude of the expression periods. Nevertheless, both these factors have relatively modest effect sizes. As shown in Figure 4.11, the number of α genes per repertoire lies between 8 and 20 genes more frequently when the α expression period is at least as long as the β expression period. Within the parameter ranges considered here, the numbers of possible α and β variants have little effect on the number of α genes per repertoire, as can be seen by comparing Figures 4.11A-E.

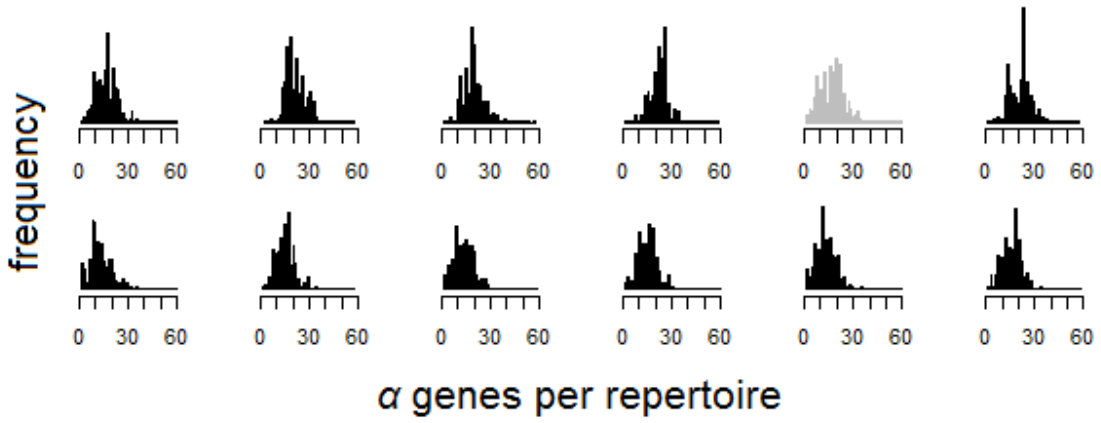


Figure 4.12. A sample of sub-repertoire size distributions resulting from models with cross-immunity defined by shared minor epitopes.
Distributions in black satisfy the criteria for being concentrated around a single peak, indicating consistent repertoire structure, whereas those in grey do not. In this example there are 2 minor epitopes per variant, each of which has 4 possible types. $N_\alpha = 64$, $N_\beta = 448$ and the expression periods are 12 generations for all genes. Immune responses to minor epitopes last for 5 expression periods.

4.4.3 Multiple epitope models

Finally, to confirm that the above results are not uniquely dependent on the particular way of modelling cross immunity, a set of alternative models are considered, based on the idea of short-lived immune responses to shared minor epitopes [Recker *et al.*, 2004]. Each α variant is assigned two minor epitopes, one from each of two small sets, so that pairs of variants frequently have at least one minor epitope in common. Similarly, β variants are assigned minor epitopes but from different sets, so no epitopes are shared between α variants and β variants. Hosts gain long-lasting immunity to each specific expressed variant (as before), but also acquire short-lived immunity to each expressed variants minor epitopes. Like the previous models, these models produce consistent numbers of α genes per repertoire within

the required range (Figure 4.12).

4.5 Results of expression order model

In modelling the evolution of repertoire structure, within-host *var* expression order has been assumed to be a fixed property of all parasite strains. Now, instead, a related model is used to investigate the evolution of expression order, to understand how parasites might benefit from expressing one gene type before another. In this case, expression periods are made constant and each parasite strains has one of three expression orders:

- α genes always expressed before β genes,
- β genes always expressed before α genes,
- equal probabilities of an expressed gene being selected from the α sub-repertoire or the β sub-repertoire.

In the simplest case, when there is no difference between α and β genes, the model does not result in selection of any particular expression order. Restricting the number of possible α variants to be lower than the number of possible β variants does not result in selection so long as the ratio of α and β genes per repertoire remains in proportion (Figure 4.13A).

Next, setting $N_\alpha = N_\beta$, it is assumed that parasites expressing α genes have a higher probability of transmission, relative to those expressing β genes. This causes strong selection of parasites expressing α genes first (Figures 4.13B and 4.14A). The outcome is similar in a model without multiplicity of infection, confirming that

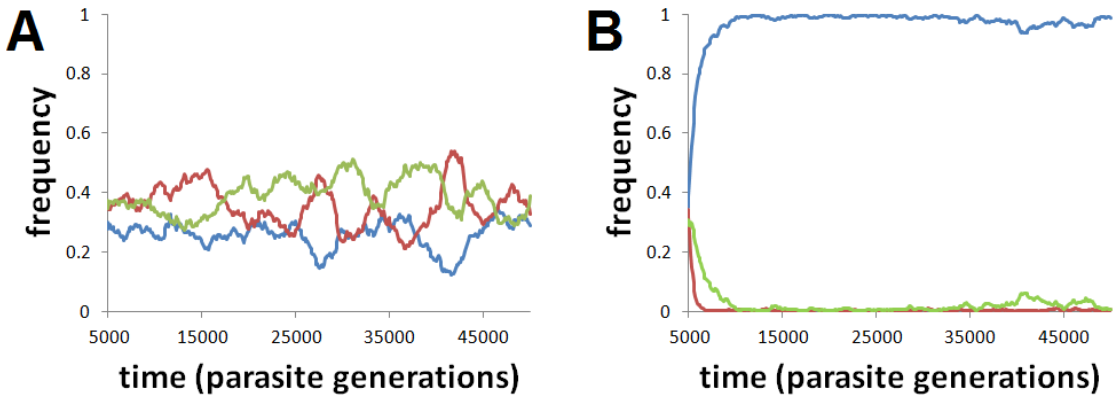


Figure 4.13. Evolution of sub-repertoire expression order. Blue = α genes before β genes; red = β genes before α genes; green = random order. **A.** When expressed genes confer equal transmission rates, all expression orders are equally frequent in the population of circulating strains. In this example the number of possible α variants is lower than the number of possible β variants ($N_\alpha = N_\beta/4$). **B.** When α genes confer a transmission rate advantage (here 50%), there is strong selection for strains expressing α genes before β genes. In this example there are equal numbers of possible variants ($N_\alpha = N_\beta$). In both models, $\leq 25\%$ of the repertoire is expressed per infection and $N_\beta = 256$. Other parameter values are as listed in Table 4.2.

the result depends on expression order between infections and not within infections (Figure 4.14B).

When the maximum infection length is higher, so that a greater proportion of the repertoire may be expressed in each infection, the selection effect is diminished (Figure 4.14C). Similarly, decreasing the number of α genes per repertoire (by setting $N_\alpha = N_\beta/4$) weakens selection for expressing α genes first (Figure 4.14D). In both cases, there remains strong selection against expressing β genes before α genes, but weaker selection against random expression order. This is because, in these scenarios, parasites with random expression order have a high probability of expressing most of their α genes in first or second infections, so the outcome of random expression order is little different from the outcome of strictly expressing α genes before

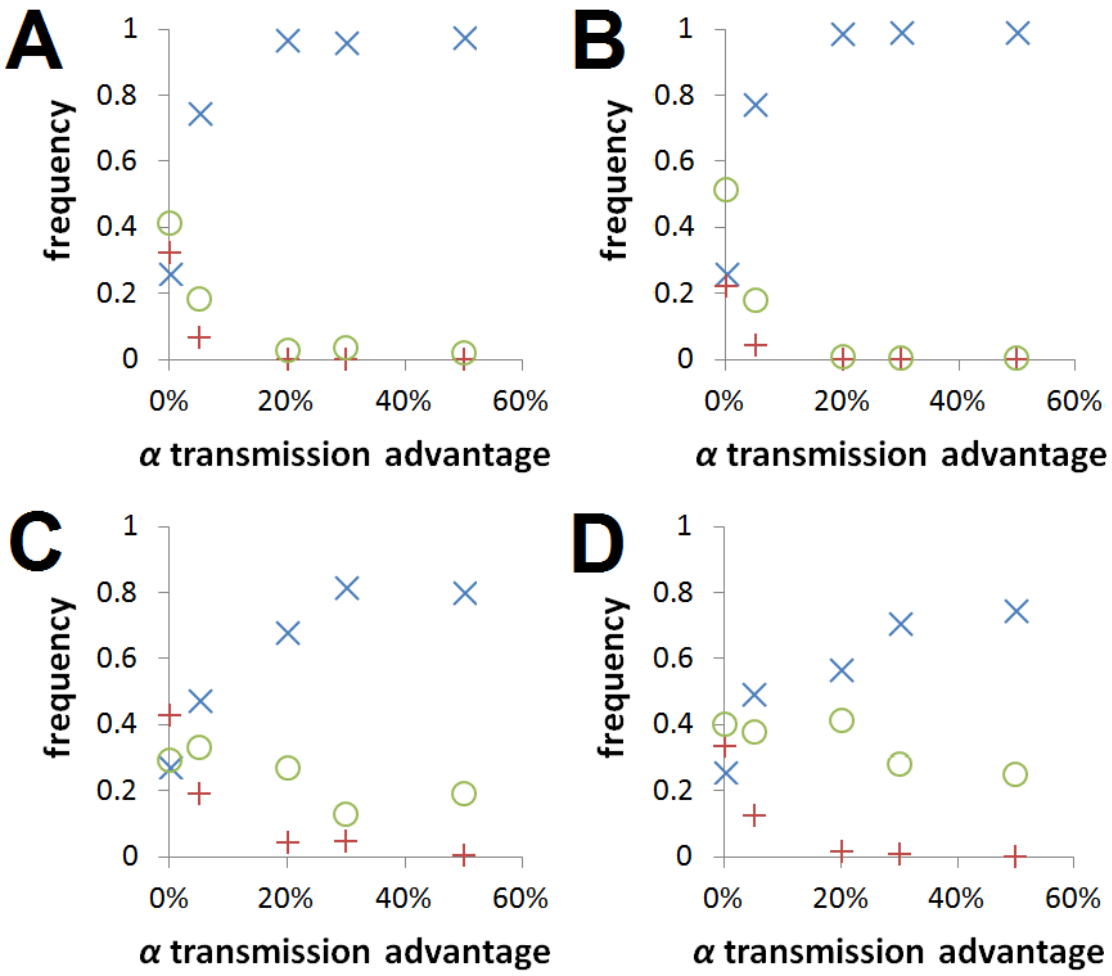


Figure 4.14. Effect of differential transmission rates on evolved expression order. The frequency of parasite strains with each expression order is averaged over the final 20,000 parasite generations of the simulation. Blue $\times = \alpha$ genes before β genes; red $+ = \beta$ genes before α genes; green $\circ$ = random order. **A.** Default parameters ($N_\alpha = N_\beta$, maximum genes expressed per infection = 25% of repertoire, maximum multiplicity of infection = 2). **B.** The same as A, except that maximum multiplicity of infection = 1. **C.** The same as A, except that maximum genes expressed per infection = 50% of repertoire. **D.** The same as A, except with unequal numbers of possible α and β variants ($N_\alpha = N_\beta/4$). Other parameter values are as listed in Table 4.2.

β genes.

An effect of differential transmission rates on the evolved order of expression remains even when there is a moderate cost, in terms of host mortality, to expressing α genes before β genes. This supposes that parasites expressing α genes out-compete those expressing β genes because they have a survival advantage in non-immune hosts, but the survival advantage is associated with higher virulence. In the case $N_\alpha = N_\beta$, there is selection for expressing α genes before β genes when mortality rates are below 1 in 36 per expressed α gene with a transmission rate advantage of 50%, and below 1 in 100 per expressed α gene with a transmission rate advantage of 20%. When mortality rates are higher, random expression order is selected. There is always strong selection against expressing β genes before α genes.

4.6 Discussion

Two aspects of antigenic repertoire evolution were investigated using mathematical models. First, attention was focussed on repertoire structure. It has been observed that *P. falciparum var* repertoires are structured such that each has approximately the same ratio of UpsA genes to UpsB and UpsC genes. Previous theoretical modelling showed that type-dependent factors such as antigenic diversity and transmission probability per expressed gene can cause the evolution of repertoires containing a mix of gene types, at least in a system with only two genes per repertoire [Buckee & Recker, 2012]. This chapter examined whether such factors can explain why there is so little variation in the ratio of gene types per repertoire. The results show that type-dependent factors can influence the mean numbers of genes of each type per repertoire. However, such factors were found to be insufficient to create consistent

repertoire structure; rather, there was wide variation around the mean, and for plausible parameter values a large proportion of repertoires comprised only a single gene type.

An alternative hypothesis that host immune responses alone drive the evolution of consistent repertoire structure was also tested. Specifically, it was assumed that cross immunity acts within gene types but not between gene types to reduce infection length. Modelling results were consistent with this hypothesis. The results were robust to variation of expression periods for each gene type, as well as to variation of the strength of cross immunity, variant population sizes, transmission rate, and expression order. Importantly, this finding is not inconsistent with the hypothesis that gene types have evolved to fill different host niches. In the cross immunity models, expression periods were found to have only a modest effect on repertoire structure. Nevertheless, the results best fitted observations when the more antigenically restricted gene type conferred a modest advantage in terms of expression period.

The second model presented here looked at how such differential transmission rates might shape the order in which antigenic variants are predominantly expressed during infections. Assuming that some genes confer a higher probability of transmission, it was found to be advantageous for a parasite strain to express these genes at the earliest opportunity. This is because the strain might not get a second chance: in the time between infection and reinfection, the host might be infected by other strains and thus acquire immunity to shared variants. A corollary is that if a high proportion of circulating strains have evolved this expression order then there is even greater selection pressure driving the rest of the population to do likewise.

It has been proposed that pathogens evolve to satisfy a trade-off between vir-

ulence and transmissibility [Alizon & van Baalen, 2008, Anderson & May, 1982]. For *P. falciparum* to be transmitted between human hosts, a biting mosquito must ingest mature gametocytes. The probability that this will happen has been found to be only weakly correlated with gametocyte density and even more weakly correlated with asexual parasite density [Babiker *et al.*, 2008, Bousema & Drakeley, 2011]. Instead, it may be hypothesised that the transmission rate depends on the predominantly expressed *var* gene for reasons not directly related to asexual parasite numbers. For example, differences in binding phenotypes might enable asexual parasites to sequester in tissues, such as bone marrow, that provide safer environments for their maturing gametocyte offspring [Smalley *et al.*, 1981]. In this scenario, the results presented here indicate that there is a strong advantage to first expressing the genes that confer this higher transmission fitness, even at the cost of higher virulence. Thus it may be that the benefit to be gained from expression order drives the evolution of very high effective multiplication rates , so that parasites with the highest transmission fitness outcompete the rest of the population during the initial phase of expansion after exiting the liver. In this case, the effective multiplication rate is a function of the asexual reproduction rate and the parasite survival rate, and it is the latter that is assumed to be under the control of the expressed *var* gene.

An additional factor not considered here is the benefit that may result from outcompeting other parasite strains. One study of *Plasmodium chabaudi* in laboratory mice found that strains with higher effective multiplication rates had a competitive advantage in mixed-strain infections, resulting in a transmission advantage [de Roode *et al.*, 2005]. A similar effect in *P. falciparum* might further contribute to the evolution of virulence [Pollitt *et al.*, 2011, van Baalen & Sabelis, 1995].

Why should expression order be orchestrated by a difference in effective multiplication rate as well as by the transcriptional switching hierarchy described in Chapter 3? One possible reason is that the former is more robust to the effects of host diversity and competing infections. Additionally, there are benefits to having two levels of regulation. Parasites with higher effective multiplication rates will predominate even if they are only rarely activated by transcriptional switching. The genes that confer higher effective multiplication rates can therefore be assigned very low activation biases, allowing the higher biases to be assigned to other variants. This gives the parasite greater control over expression order throughout the repertoire and means that fewer genes with lower effective multiplication rates need be assigned low activation biases.

Chapter 5

Modelling the evolution of genetic diversity and strain structure in *P. falciparum* antigenic repertoires

5.1 Introduction

Plasmodium falciparum var genes appear to have diverged into types associated with different binding properties and different levels of antigenic diversity. Chapter 4 looked at how such structuring of the *var* population might lead to structuring of *var* repertoires. This chapter considers how the existence of distinct types might influence patterns of *var* sequence diversity and repertoire population diversity.

On the basis of an upstream flanking sequence (UpsA, B, C or E), gene architecture and chromosomal location, *var* genes can be categorised into Groups A, B, C and E, and an intermediate Group A/B [Kraemer *et al.*, 2007, Lavstsen *et al.*, 2003]. Hosts first acquire immune responses to Group A and A/B *var* genes and it

is thought that parasites predominantly express these genes during early infections, perhaps because they confer a within-host survival advantage [Cham *et al.*, 2010, 2009, Warimwe *et al.*, 2009].

Studies of the DBLα domain, which is found in almost all *vars*, indicate that this domain has lower sequence diversity among UpsA genes than among other genes [Barry *et al.*, 2007, Bull *et al.*, 2008, Chen *et al.*, 2011, Falk *et al.*, 2009, Mugasa *et al.*, 2012, Trimnell *et al.*, 2006, Ward *et al.*, 1999]. Domains found only in UpsA and longer UpsB genes also have relatively high sequence conservation [Buckee & Recker, 2012]. Intriguingly, the same study reported that gene length, in terms of number of encoded binding domains, is correlated with sequence diversity independently of Ups type.

According to results presented in Chapter 3, the highest sequence diversity is seen among shorter genes in central chromosomal location. The same chapter reports that such genes also have high *in vitro* activation biases, which means they are frequently activated by transcriptional switching. It is expected that, *in vivo*, the hierarchy of activation biases generates expression order among otherwise similar genes. All of these findings are summarised in Table 5.1. It should be noted, however, that Table 5.1 is a simplification; in reality categories are less sharply defined and some attributes vary over continuous ranges.

Table 5.1. Typical properties of *var* gene types.

Ups type	A or B	B or C	
Location	Subtelomeric	Subtelomeric	Central
Activation bias	Low	Low	High
Sequence diversity	Low	Low	High
Gene length	Long	Short	
***In vivo* expression**	Early	Late	

The findings of Chapter 3 led to the hypothesis that genes with higher activation biases, which are presumed to be expressed more frequently, undergo more frequent recombination. The cause of this correlation is unknown. One possibility is that transcription facilitates recombination via a molecular mechanism, as has been observed in other eukaryotes [Gottipati & Helleday, 2009, Prado & Aguilera, 2005]. It should be noted, however, that this hypothesis fits the data only if UpsA genes, which have unusually high sequence conservation, are among the least frequently transcribed *vars*.

An alternative explanation may depend on considering how genetic diversity is generated. Each *var* gene comprises several domains that are themselves modular. A set of 628 homology blocks, with an average length of 21 amino acids, can be used to describe, on average, 83.5% of a *var* sequence [Rask *et al.*, 2010]. Frequent recombination events during mitosis result in multiple homology blocks being copied between genes [Bopp *et al.*, 2013, Duffy *et al.*, 2009, Freitas-Junior *et al.*, 2000]. It appears that recombination occurs more often between genes with similar architectures, so that combinations of sequence blocks defining distinct domain classes are maintained [Bull *et al.*, 2008, Kraemer *et al.*, 2007, Kraemer & Smith, 2006, Lavstsen *et al.*, 2003]. Previous theoretical mathematical modelling has assumed a trade-off between antigenic diversity and within-host fitness [Buckee & Recker, 2012] but has not explicitly considered genetic diversity. The potential effects of *var* structure and expression order on the evolution and maintenance of genetic diversity have not been investigated.

Differences between gene types may affect diversity not only among genes, but also at the repertoire level. According to strain theory [Gupta *et al.*, 1998, 1996], host immunity can structure pathogen populations into strains with minimally over-

lapping immunodominant epitopes. This idea has implications for understanding the acquisition of immunity and for predicting the impact of vaccines that target particular epitopes or particular strains. Buckee *et al.* [2009] investigated *P. falciparum* strain structure by analysing patterns of immune recognition among Kenyan children. They found that the data were most consistent with a model in which *var* repertoires comprise two sets of genes. The smaller sub-repertoire is selected, as a block, from a small number of non-overlapping strains, whereas the rest of the repertoire consists of variants chosen more randomly from a large pool. In a first step to towards a theoretical understanding of these patterns, Artzy-Randrup *et al.* [2012] investigated the evolution of antigenic repertoire diversity using a mathematical model with a single gene type and no restriction on mutation (so that every variant was equally likely to mutate into every other variant). In this model, immune selection was found to impose strong strain structure on the repertoire population, meaning that the population was dominated by a small number of strains with very low overlap between their repertoires. However, the study did not consider how strain structure changes when there are multiple gene types with different levels of diversity or other attributes.

To test hypotheses regarding the observed patterns of *var* diversity, this chapter employs mathematical modelling of the evolution of a parasite with two types of antigen-encoding gene. It is found that a model incorporating immune selection and two other essential factors – expression order and gene length – can reproduce all of the major patterns in *var* genetic diversity reported in Chapter 3. The model additionally generates predictions regarding *P. falciparum* strain structure that are consistent with empirical observations.

5.2 Evolutionary model

5.2.1 Sequence representation

The mathematical model structure described in the final part of Chapter 4 is extended to incorporate genetic sequences, to enable investigation of the evolution of genes as well as repertoires. The structure of the algorithm is presented as pseudocode in Algorithm 2. As in the previous model, there are two gene types α and β, which are assumed not to recombine with one another. Little is known about how sequence differences relate to antigenic differences among *var* genes. Therefore, rather than attempting to replicate molecular mechanisms, the structure here focuses on reproducing the most essential features. Thus, the method allows numerous distinct but similar sequences to encode the same antigenic type, so that a change in sequence does not necessarily result in a change of antigenic type. This means that sequences can be shaped by neutral modifications as well as by host immunity.

Specifically, each gene comprises a set of sequence blocks $b_1, \ldots, b_k$. If the gene is of type α then $k = \log_2 N_\alpha$, and otherwise $k = \log_2 N_\beta$, where N_α and N_β are, respectively, the numbers of possible α and β variants. Each sequence block is made up of sequence elements $s_1, \ldots, s_l$ and there are v possible values for each sequence element. The sequence elements together determine the antigenic type of the gene as follows.

1. Within each block, count how many sequence elements have a value greater than $v/2$,

2. If the count is larger than $l/2$ then the value of the block is 1; otherwise the

```
for each generation:
    for each host and each of their infections:
        Decrement infection length countdown;
        if there is an expressed gene:
            Decrement expression period countdown of the expressed gene;
            Make host immune to the expressed gene;
        if infection length countdown is zero:
            Clear the infection;
    for each host and each their infections:
        if there is an expressed gene and expression period countdown is zero:
            for each gene in the repertoire, in order of expression:
                if host is not immune to the gene:
                    Make gene expressed;
                    Initialise expression period countdown;
                    Exit for loop;
    for i = 0 to introduction_rate:
        Select a random host;
        Generate a random repertoire;
        Transmit infection to host;
        Initialise infection length countdown;
    for i = 0 to number_of_hosts × transmission_rate:
        Select a random pair of hosts A and B;
        if A is infected then select random infections to transmit:
            for each infection to be transmitted:
                Apply recombination and mutation;
                Transmit infection to host B;
                Initialise infection length countdown in host B;
    for each host:
        Increment host age;
        Dependent on host age, implement death and birth;
```

Algorithm 2: Pseudocode version of the algorithm used to implement the evolutionary model.

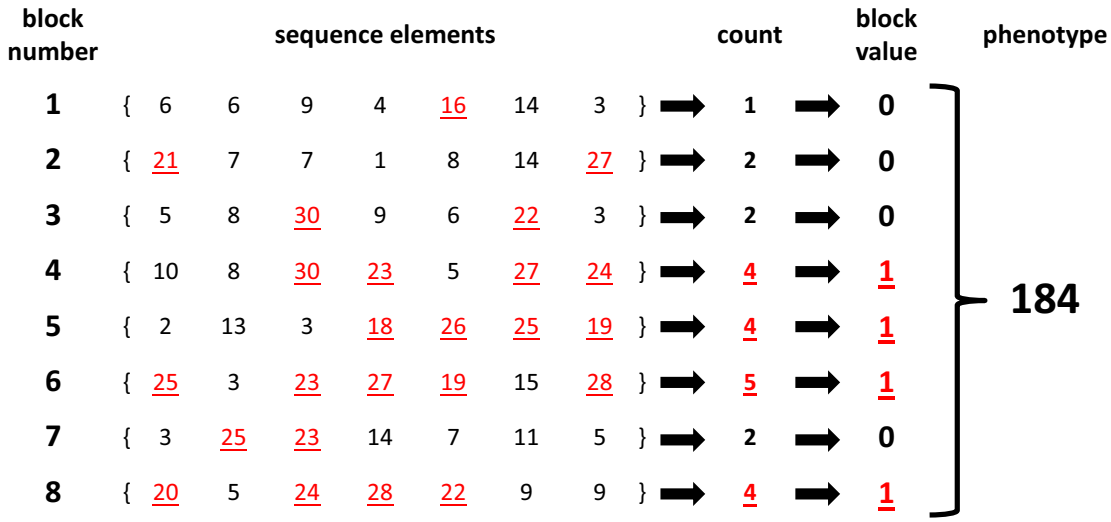


Figure 5.1. Relationship between sequence and antigenic type in the model. Within each block, the count of high-valued sequence elements (underlined and in red) determines the block value. The sequence of block values determines the antigenic type (see main text for a fuller explanation). In this example, $k = 7$, $l = 7$ and $v = 30$.

value of the block is 0,

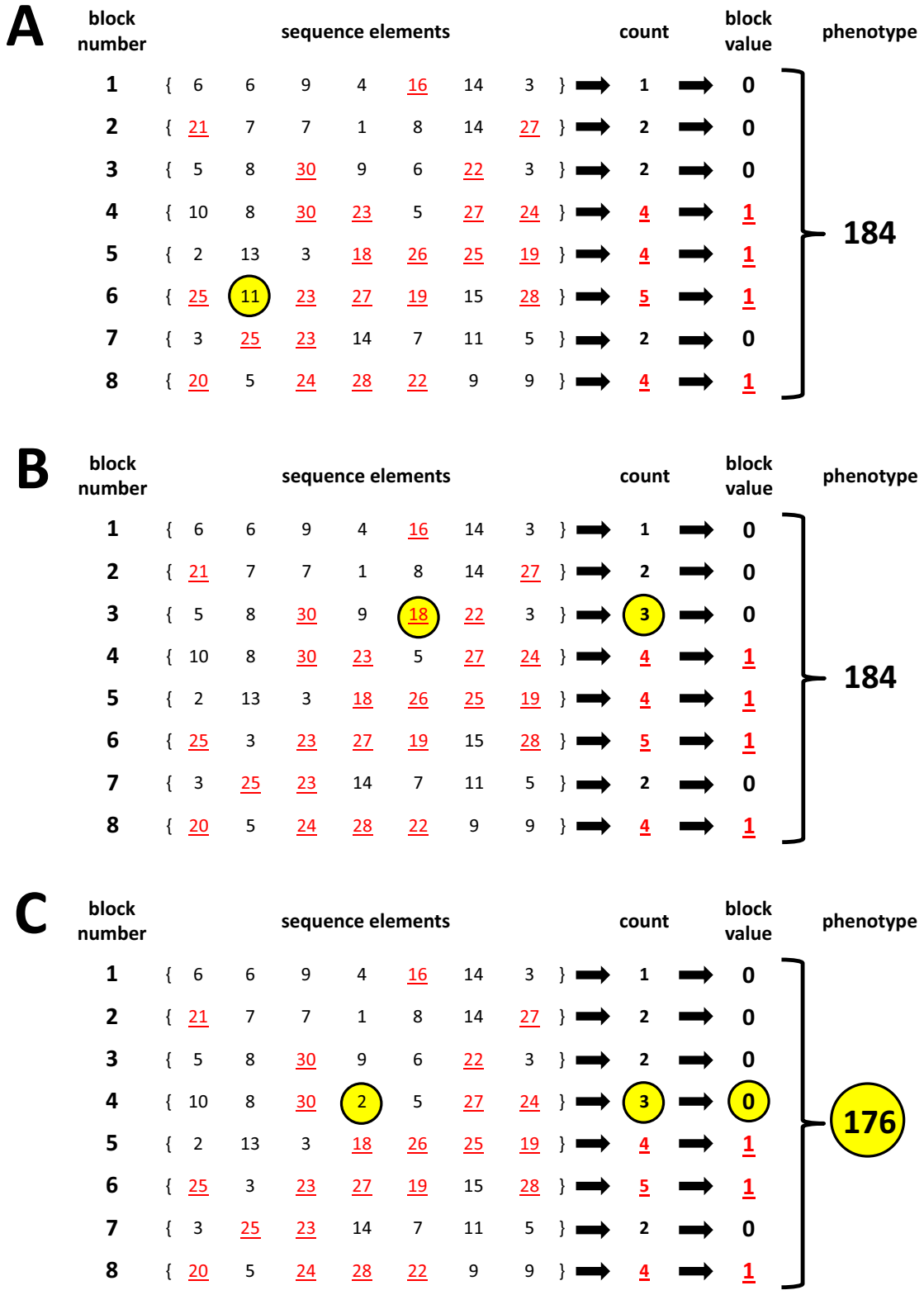
3. The binary block sequence uniquely determines the antigenic type.

This scheme is illustrated by example in Figure 5.1.

This method improves upon the common modelling assumption of unrestricted, instantaneous mutation between antigenic variants. It enables mutation and recombination events to change a sequence without changing the antigenic type. Also, the scale of change matters: a switch of antigenic type is more likely to result from a recombination event affecting multiple sequence elements, or from the cumulative effects of consecutive modifications, than from a single mutation. There are three possible outcomes of a sequence modification. First, a synonymous modification may change sequence elements without having any immediate or cumulative effect on antigenic type (Figure 5.2A). Alternatively, a non-synonymous modification may change sequence elements with no immediate effect but a potential cumulative effect on antigenic type (Figure 5.2B). Lastly, a non-synonymous modification may change sequence elements causing an immediate switch of antigenic type (Figure 5.2C). Importantly, the symmetry of the system ensures that modifications are unbiased. That is, within a random set of genes, the effects of recombination and mutation on diversity are independent of sequence and independent of the sites of the modifications.

5.2.2 Recombination and immune selection of modified repertoires

Unlike the model in Chapter 4, mutation acts on randomly selected sequence elements rather than entire genes. Further, it is assumed that, far more commonly


A
block number
sequence elements
count
block value
phenotype
1 { 6 6 9 4 16 14 3 } 1 0
2 { 21 7 7 1 8 14 27 } 2 0
3 { 5 8 30 9 6 22 3 } 2 0
4 { 10 8 30 23 5 27 24 } 4 1
5 { 2 13 3 18 26 25 19 } 4 1
6 { 25 11 23 27 19 15 28 } 5 1
7 { 3 25 23 14 7 11 5 } 2 0
8 { 20 5 24 28 22 9 9 } 4 1
184
B
block number
sequence elements
count
block value
phenotype
1 { 6 6 9 4 16 14 3 } 1 0
2 { 21 7 7 1 8 14 27 } 2 0
3 { 5 8 30 9 18 22 3 } 3 0
4 { 10 8 30 23 5 27 24 } 4 1
5 { 2 13 3 18 26 25 19 } 4 1
6 { 25 3 23 27 19 15 28 } 5 1
7 { 3 25 23 14 7 11 5 } 2 0
8 { 20 5 24 28 22 9 9 } 4 1
184
C
block number
sequence elements
count
block value
phenotype
1 { 6 6 9 4 16 14 3 } 1 0
2 { 21 7 7 1 8 14 27 } 2 0
3 { 5 8 30 9 6 22 3 } 2 0
4 { 10 8 30 2 5 27 24 } 3 0
5 { 2 13 3 18 26 25 19 } 4 1
6 { 25 3 23 27 19 15 28 } 5 1
7 { 3 25 23 14 7 11 5 } 2 0
8 { 20 5 24 28 22 9 9 } 4 1
176

than mutation, genes undergo recombination. In this process, one or more randomly selected sequence elements are copied from a donor gene and substituted for the corresponding sequence elements in a recipient gene.

Like mutation, recombination is modelled at the time of transmission. Importantly, it is not assumed that transcription is directly linked to recombination. Instead, the assumptions here follow from the biological argument that a modified strain is unlikely to replace its parent strain unless it is selected for by the immune system. Consider four scenarios:

- An unexpressed gene is modified,
- An expressed gene is modified but retains its antigenic type,
- An expressed gene is modified and acquires a new antigenic type, to which the host is already immune,
- An expressed gene is modified and acquires a new antigenic type, to which the host is not immune.

In the first three scenarios, the parasite strain with the modified gene has no within-host fitness advantage and so is unlikely to out-compete the unmodified

---

Figure 5.2 *(preceding page).* Possible effects of sequence modifications on antigenic type. The unmodified sequence is shown in Figure 5.1. Changed values are circled and highlighted in yellow. **A.** A sequence element is changed from one low value to another with no effect on the count of high-valued sequence elements, block value or antigenic type. **B.** A sequence element is changed from a low value to a high value, which increases the count of high-valued sequence elements but has no immediate effect on block value or antigenic type. **C.** A sequence element is changed from a high value to a low value, which decreases the count of high-valued sequence elements and immediately changes the block value and the antigenic type.

strain. It is therefore unlikely that the modified strain will be represented among the relatively small number of parasites ingested by a biting mosquito. In the final scenario, however, the strain with the modified gene might outgrow the unmodified strain and thus has a higher probability of being present in a mosquito blood meal. It is therefore argued that, even if the recombination rate is the same for all genes, host immunity will select for modifications to expressed genes only.

To mimic the effects of immune selection, recombination events are modelled such that the recipient gene is selected from genes that have been expressed during the current infection, whereas the donor is randomly selected from all genes belonging to the same sub-repertoire. Further, a strain with a modified gene is transmitted only if the modified gene encodes an antigenic type that is novel to the transmitting host; otherwise the unmodified version of the strain is transmitted.

In the model, recombination can result in the transfer of multiple, randomly chosen sequence elements according to a Bernoulli process. That is, if the probability of modification of one sequence element is r then the probability of modification of m elements is r^m. The probability r is hereafter referred to as the recombination rate.

For the sake of simplicity, meiotic recombination between gene repertoires is not considered here.

5.2.3 Measures of diversity and overlap

A measure of sequence diversity is derived from Hamming distances. The average sequence conservation of genes at repertoire position i is defined as the average number of sequence elements shared by pairs of genes at repertoire position i among circulating repertoires.

Diversity in repertoire populations is measured using the inverse Simpson index, which was previously used in a similar modelling study [Artzy-Randrup *et al.*, 2012]:

$$\text{diversity } D = \frac{A(A-1)}{\sum_{i=1}^{S} a_i(a_i-1)},$$

where S is the number of distinct strains in the population, a_i is the number of infections with strain i, and A is the total number of infections (that is, $A = \sum_{i=1}^{S} a_i$). This diversity measure is 1 when there is only a single strain and is infinite when there is only one infection for every strain.

Repertoire overlap (the frequency of shared genes or domains) is measured using an index similar to the Simpson index:

$$\text{overlap} = \frac{1}{DG} \sum_{i=1}^{G} \left(1 \Big/ \sum_{j=1}^{S} \frac{d_{ij} a_j^2}{A_i^2} \right),$$

where G is the number of antigenic types in the population, A_i is the number of infections with strains containing antigenic type i (that is, $A_i = \sum_{j=1}^{S} d_{ij} a_j$), and

$$d_{ij} = \begin{cases} 1 & \text{if antigenic type } i \text{ is in strain } j \\ 0 & \text{otherwise.} \end{cases}$$

Artzy-Randrup *et al.* [2012] used a similar index but did not divide by diversity, D, which meant that in their results, overlap was highly dependent on diversity. Dividing by diversity ensures that the two measures are independent, and causes the overlap measure to be 0 when no antigenic types are shared by distinct strains, and 1 when there is only a single strain. The overlap measure, thus modified, correlates with the Jaccard index, which is also frequently used to assess overlap

between sets.

5.2.4 Model parameters

Parameter values are generally the same as those used in the second part of Chapter 4 but new parameters are added to implement sequence recombination.

Table 5.2. Parameter values of expression order model

Parameter	Value(s)	Refs
Expression period per gene (parasite generations)	10	[1,2]
Number of possible α gene variants, N_α	64, 256	[3]
Number of possible β gene variants, N_α	256, 1024	[3]
Contact rate per host per parasite generation	0.007, 0.01, 0.02	-
Repertoire size	64	-
Host population	10,000	-
Maximum multiplicity of infection	2	[4,5]
Maximum infection length (parasite generations)	160, 320	[1,2,6]
Introduction of new strains (per parasite generation)	0.1	-
Gene mutation rate per sequence element per transmission event	0.0001	-
Number of sequence elements per sequence block	7	-
Number of possible variants of each element	30	-
Recombination rate	0.05, 0.5	[7]

[1] = Collins & Jeffery [1999b]; [2] = Collins & Jeffery [1999a]; [3] = Buckee *et al.* [2009]; [4] = Mueller *et al.* [2012]; [5] = Koepfli *et al.* [2011], [6] = Bretscher *et al.* [2011], [7] = Bopp *et al.* [2013].

Of particular note, the last of these parameter values is informed by a study by Bopp *et al.* [2013], who found that the *var* recombination rate is high enough to generate new sequences during a typical infection.

Initially, the model is seeded with 500 parasite strains with random repertoires. After 5,000 parasite generations, all infections are cleared and the model is seeded again with 500 parasite strains (this procedure is designed to lessen founder effects).

With the above parameters, infection prevalence stabilises between 35% and 40%. The model is stopped after 50,000 parasite generations (approximately 274 years, assuming 48 hours per generation). On a standard desktop computer (quad core, 3 GHz, 4 GB RAM), each model takes less than two hours to run.

Host mortality is modelled using the Gompertz function described in the Methods of Chapter 4. To maintain a constant population size, each death is immediately followed by the birth of a fully susceptible host.

5.3 Analysis of genetic diversity

5.3.1 Effects of expression order

To test for effects of expression order, the model is set up so that all repertoires express α genes before β genes and, within each sub-repertoire, genes are sequentially expressed according to expression hierarchies. As expected, these expression hierarchies are found to influence genetic diversity, in terms of shared sequence elements. Genes nearer the top of an expression hierarchy are, on average, more diverse than those of lower rank (Figure 5.3). This result is consistent with the finding that within *P. falciparum var* sub-repertoires, the order of gene expression is correlated with genetic diversity, as reported in Chapter 3.

As well as shaping diversity within sub-repertoires, expression order can result in different levels of diversity between sub-repertoires. When α genes are expressed before β genes, they become generally more diverse than β genes, as shown in Figure 5.3. This contrasts with what has been observed in *P. falciparum*. In the domains common to all major *var* types, genes thought to be predominantly expressed during early infections (corresponding to α genes in the model) have relatively high

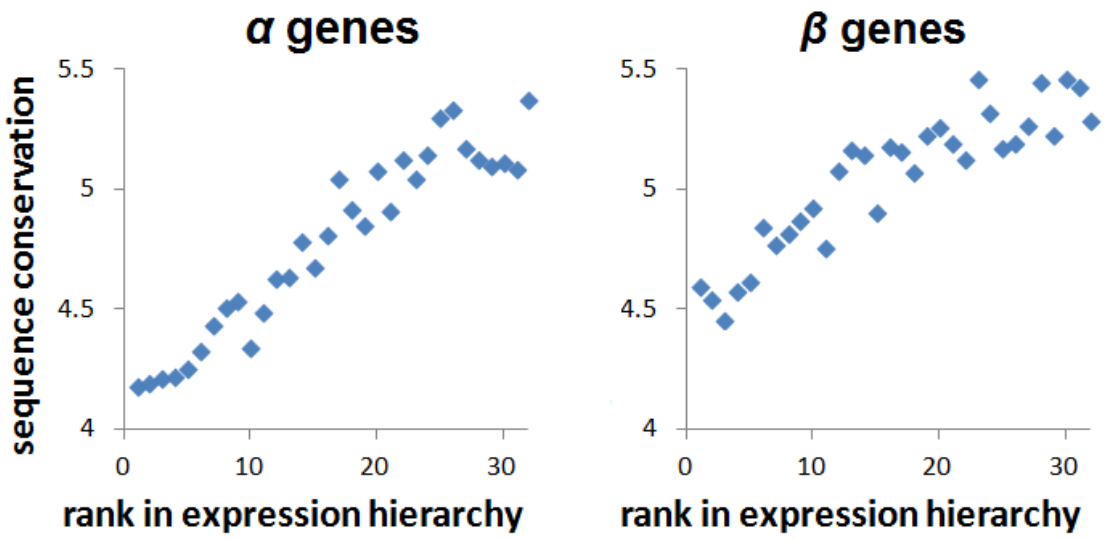


Figure 5.3. Effect of expression hierarchy on genetic diversity in the single-domain model with sequence recombination. Expression order is correlated with sequence conservation among α genes (left) and among β genes (right). In this model the recombination rate is 0.05, there are equal numbers of possible α and β variants ($N_\alpha = N_\beta = 256$), the contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2.

sequence conservation [Barry *et al.*, 2007, Bull *et al.*, 2008, Chen *et al.*, 2011, Falk *et al.*, 2009, Mugasa *et al.*, 2012, Trimnell *et al.*, 2006, Ward *et al.*, 1999]. According to the results of Chapter 3, the greatest diversity is seen in another subset of genes with exceptionally high expression biases, which, it may be conjectured, places these genes near the top of their sub-repertoire expression hierarchy.

5.3.2 Effects of gene length

A review of Table 5.1 suggests a possible reason for the difference between the model results and what is seen in *P. falciparum*. The missing factor is gene length, which corresponds to the number of encoded binding domains. The *var* genes expressed during early infections generally have more binding domains than the genes expressed in subsequent infections and a recent study reported that this factor is independently correlated with sequence conservation [Buckee & Recker, 2012]. Accordingly, the model is modified so that α genes encode two domains, whereas β genes, as before, encode only one domain. It is assumed that following exposure to an antigenic variant with domains d_1 and d_2, a host will be immune to every other variant containing either d_1 or d_2.

The multi-domain model is first run with $N_\alpha = N_\beta$ and α genes expressed before β genes. In contrast to the previous results, the most diverse genes are now the β genes near the top of the β expression hierarchy. Diversity among α genes is similar to diversity among the remainder of the β genes. Therefore the model output now qualitatively resembles the observations reported in Chapter 3. A wide range of recombination rates give similar results (Figure 5.4).

These consistent outcomes reflect a balance of two selection forces. Recall that strains modified by recombination are selected for transmission only if two conditions

are satisfied. First, the modification must affect the expressed gene. Second, the modification must result in a switch to an antigenic type to which the host is not immune. As previously shown, the first condition means that genes higher in a sub-repertoire expression hierarchy undergo more frequent recombination. The switch to an unrecognised antigenic type, however, depends on a gene acquiring beneficial mutations in all of its domains. Over the course of an infection, this is less likely to happen to a two-domain gene than to a single-domain gene. Hence the second selection constraint slows the diversification of genes with a higher number of binding domains.

5.3.3 Effects of functional constraints

A next step is to consider PfEMP1 function. In *P. falciparum*, the additional extracellular domains encoded by longer *var* genes have been shown to have different binding affinities and avidities [Rowe *et al.*, 2009, Turner *et al.*, 2013]. It has been hypothesised that these different binding properties confer higher fitness in non-immune hosts, and it can be speculated that such an advantage imposes greater functional constraints on these additional domains [Bull *et al.*, 2000, 1999, Smith *et al.*, 2013]. This offers a simple explanation for why the domains found only in longer genes have especially high sequence conservation. However, longer genes also have higher sequence conservation in the DBLα domain, which occurs in almost all *var* genes and has not been directly linked to binding phenotype. The relationship between activation bias and sequence diversity reported in Chapter 3 also held for the DBLα domain.

To examine how functional constraints on binding domains might affect results, it is assumed that the second α domain has fewer possible variants than the other

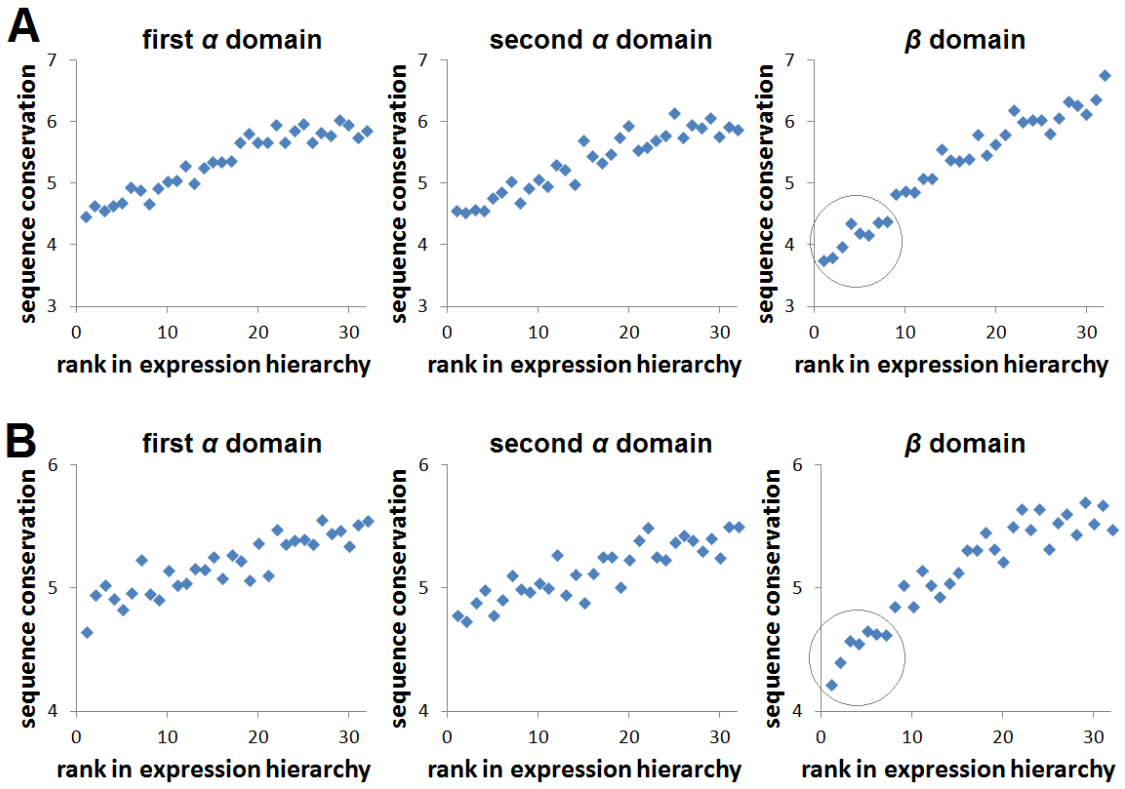


Figure 5.4. Effect of expression hierarchy on genetic diversity in the multi-domain model with equal numbers of possible variants for all domains. The contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. **A.** $N_\alpha = N_\beta = 256$ and recombination rate 0.5. Circled is a subset of β genes with greater diversity than α genes, consistent with observations of *var* repertoires. **B.** $N_\alpha = N_\beta = 256$ and recombination rate 0.05.

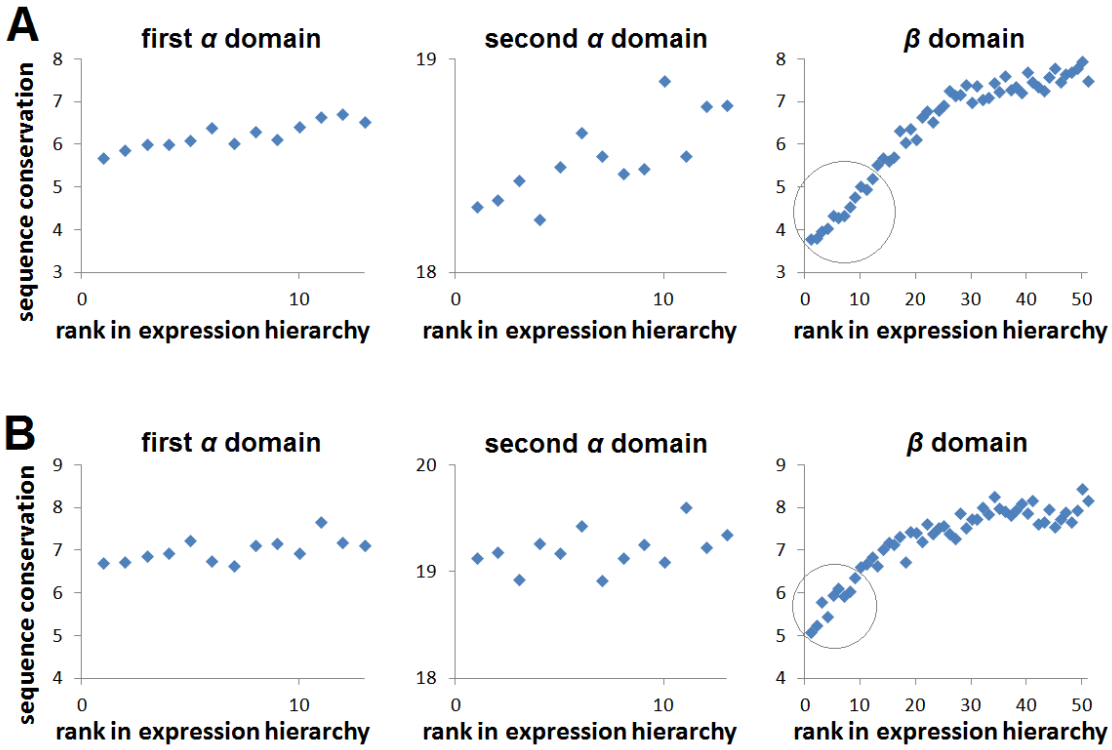


Figure 5.5. Effect of expression hierarchy on genetic diversity in the multi-domain model with fewer possible variants for the second α domain. The contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. **A.** $N_{\alpha 1} = N_\beta = 256$ and $N_{\alpha 2} = N_\beta/4$ and recombination rate 0.5. Circled are β genes with greater diversity than α genes. **B.** $N_{\alpha 1} = N_\beta = 256$ and $N_{\alpha 2} = N_\beta/4$ and recombination rate 0.05. Note changes of scale.

two domains. This is done by setting $N_{\alpha 1} = N_\beta$ and $N_{\alpha 2} = N_\beta/4$, where $N_{\alpha 1}$ and $N_{\alpha 2}$ are the numbers of possible variants of the first and second α domains, respectively. The first α domain and the β domain, which have equal functional constraints, might then correspond to DBLα. As shown in Figure 5.5, in this model the second α domain is, predictably, much more highly conserved. Nevertheless, the diversity patterns in the other two domains continue to be consistent with empirical observations.

5.3.4 Effects of epidemiology

Humans first acquire immune responses to antigens encoded by Group A and A/B *vars*, which correspond to α genes in the model. Where *P. falciparum* is highly prevalent, immunity to these variants is typically gained at a young age, but immunity to all parasite strains is apparently never achieved. Figure 5.6 illustrates how host immune memory grows with age in the multi-domain model considered above, and how this affects gene expression and incidence of infection. Especially when there are fewer possible antigenic variants for the second α domain (Figure 5.6B), the frequency of α gene expression drops steeply over the first few years of life (Figure 5.6, left column), resembling the pattern seen in human populations. This trend results from the rapid acquisition of immunity to the most antigenically restricted α domain (Figure 5.6, middle column). Infection prevalence declines with host age but complete immunity is typically never acquired (Figure 5.6, right column).

In hosts beyond middle age, the model with equal numbers of possible antigenic variants shows an increase in α gene expression (Figure 5.6A, left column). This trend coincides with the gradients of the immune acquisition curves becoming equal for each domain type (Figure 5.6A, middle column). Such a phenomenon has not been observed in nature, perhaps reflecting a limitation of the model in capturing the ways in which immunity is acquired. Alternatively, this trend may occur in reality but has not been readily detectable, especially in the context of hosts whose immune systems have been weakened by old age.

The prevalence of *P. falciparum* infection is highly variable and the rate of acquisition of immune memory in some populations may be very unlike what is shown in Figure 5.6. It is important to test how this difference affects modelling results. As shown in Figure 5.7A, when the transmission rate is made much lower, hosts build

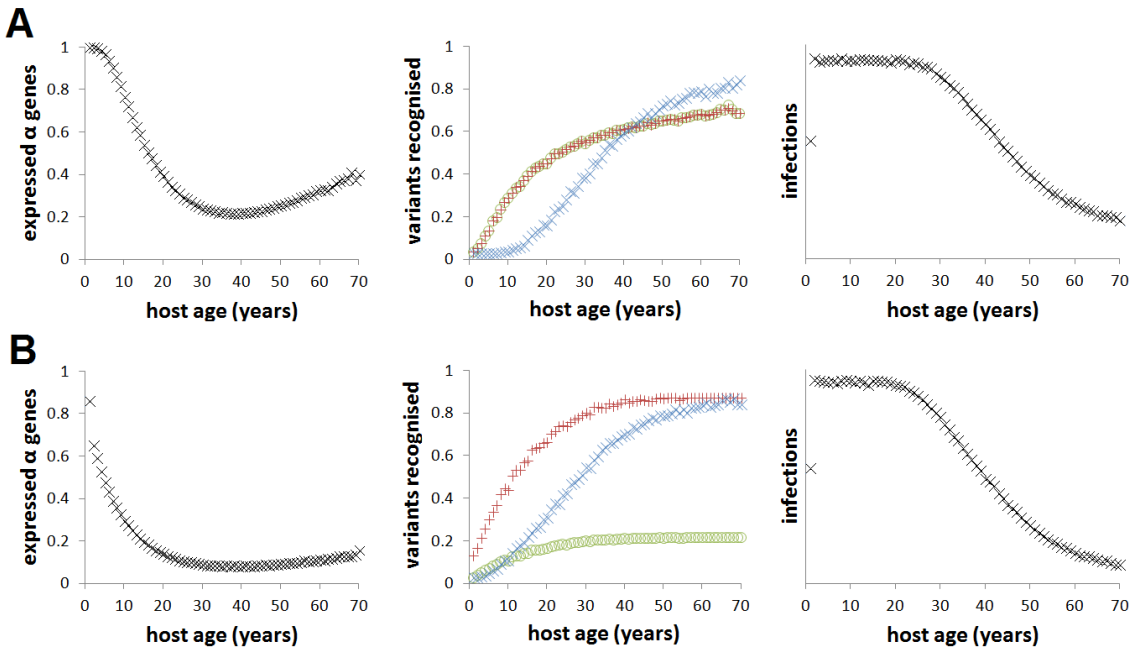


Figure 5.6. Acquisition of immunity to different antigen types. The recombination rate is 0.05 (results with recombination rate 0.5 are highly similar to those shown here), the contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. **Left column:** Mean proportion of α genes among expressed genes as a function of host age. **Middle column:** Mean proportion of all possible antigenic variants recognised as a function of host age. Green $\circ$ = first α domain, red $+$ = second α domain, blue $\times$ = β domain. **Right column:** Total number of infections per year as a function of host age. **A.** Equal numbers of possible variants for all domains ($N_{\alpha 1} = N_{\alpha 2} = N_\beta = 256$). **B.** Fewer possible variants for the second α domain ($N_{\alpha 1} = N_\beta = 256$ and $N_{\alpha 2} = N_\beta/4$).

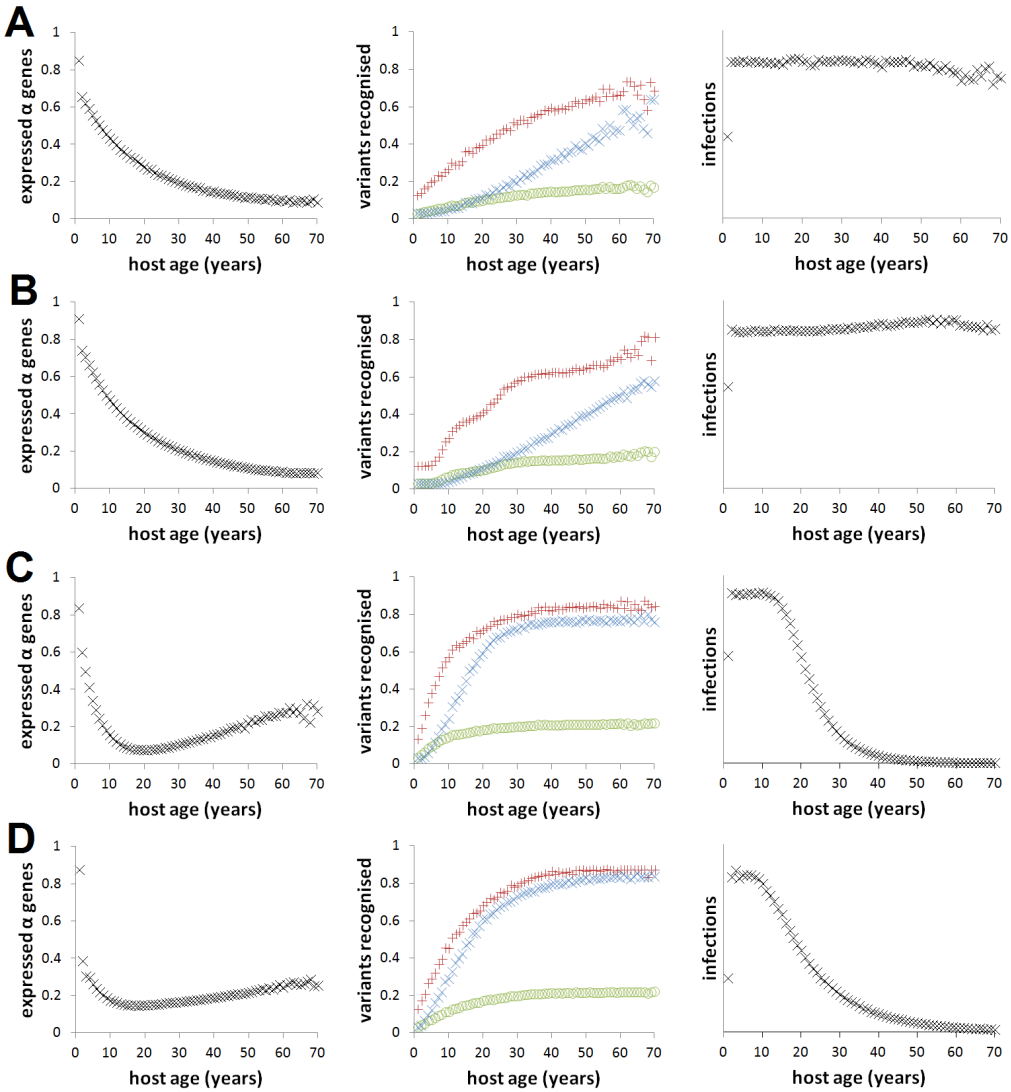
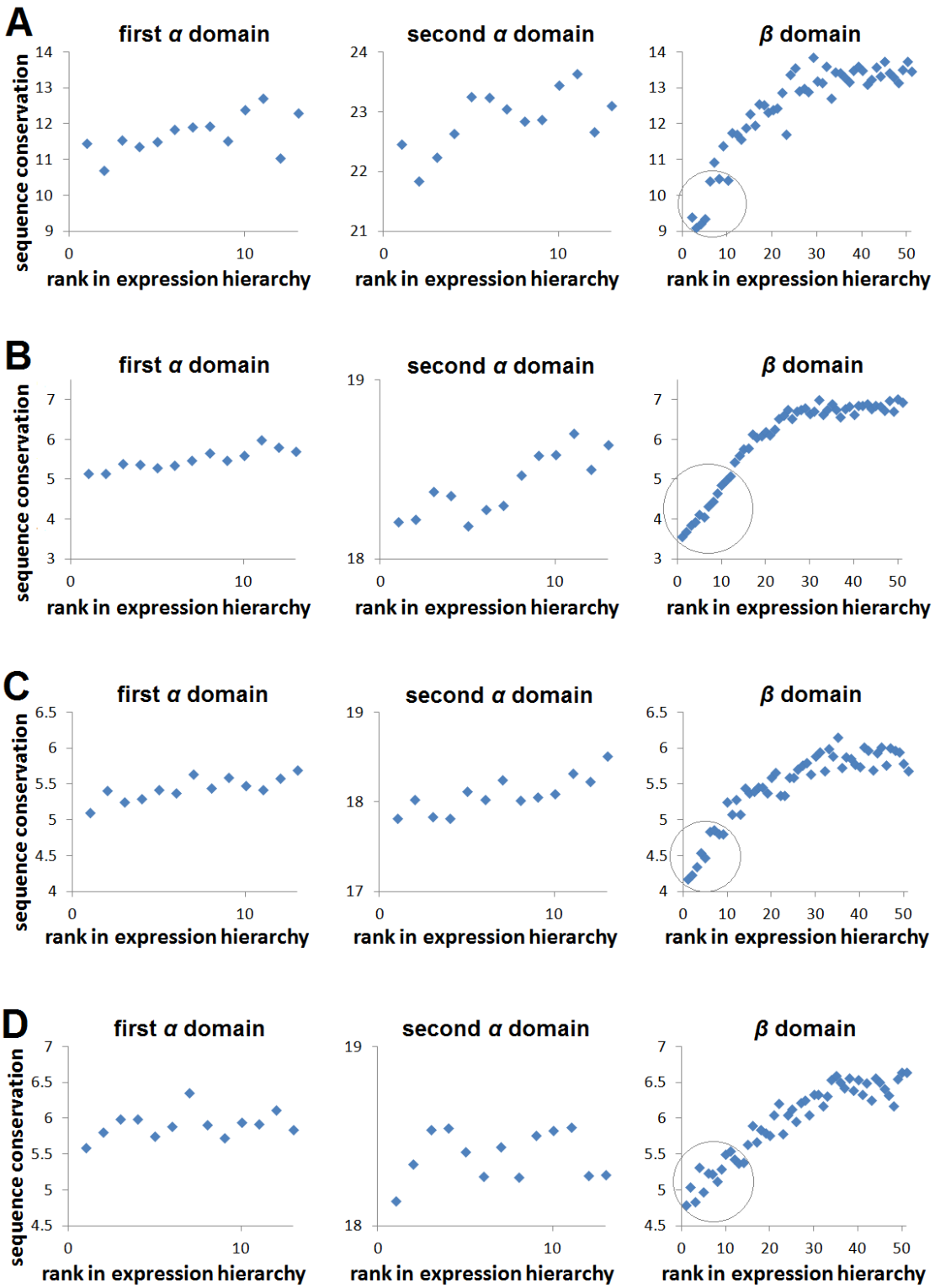


Figure 5.7. Effect of parameters changes on acquisition of immunity. In all cases, $N_{\alpha 1} = N_\beta$ and $N_{\alpha 2} = N_\beta/4$ and recombination rate = 0.05. Parameter values not specified here are as listed in Table 5.2. Left column: Mean proportion of α genes among expressed genes as a function of host age. Middle column: Mean proportion of all possible antigenic variants recognised as a function of host age. Green = first α domain, red + = second α domain, blue = β domain. Right column: Total number of infections per year as a function of host age. **A.** Lower transmission rate (0.007 instead of 0.01). **B.** Larger number of possible variants ($N_{\alpha 1}$ = 1,024 instead of 256). **C.** Higher transmission rate (0.02 instead of 0.01). **D.** Longer maximum infection length (up to 50% of repertoire expressed per infection, instead of 25%).


A
first α domain
second α domain
β domain
sequence conservation
rank in expression hierarchy
B
first α domain
second α domain
β domain
sequence conservation
rank in expression hierarchy
C
first α domain
second α domain
β domain
sequence conservation
rank in expression hierarchy
D
first α domain
second α domain
β domain
sequence conservation
rank in expression hierarchy

immune memory more slowly and the incidence of infection barely declines with age. This change also affects genetic diversity, as illustrated in Figure 5.8A. Compared to the equivalent model with moderate transmission rate (Figure 5.5B), all domains have moderately higher sequence conservation. Smaller populations might be expected to be less diverse, but this outcome might also be influenced by the fact that the model simulates recombination and mutation at the point of transmission. Importantly, however, the qualitative pattern of sequence diversity is unchanged by lowering the transmission rate (Figure 5.8A).

Increasing the total number of possible variants results in age profiles for immune acquisition and incidence of infection (Figure 5.7B) that are very similar to those for the low transmission model. In this case, sequence diversity is slightly higher in all domains than in the model with fewer possible variants. Nevertheless, the qualitative sequence diversity pattern remains (Figure 5.8B).

Conversely, increasing the transmission rate leads to more rapid acquisition of immunity, so that hosts typically become immune to almost all variants many years before death (Figure 5.7C). A similar effect results from increasing the maximum infection length, so that generally more genes are expressed in each infection (Figure 5.7D). In both scenarios, all domains evolve slightly higher sequence conservation but the overall genetic diversity pattern is, once again, unchanged (Figures 5.8C

Figure 5.8 *(preceding page).* Effect of parameters changes on sequence diversity. In all cases, $N_{\alpha 1} = N_{\beta}$ and $N_{\alpha 2} = N_{\beta}/4$ and recombination rate = 0.05. Parameter values not specified here are as listed in Table 5.2. Circled are β genes with greater diversity than α genes. Note changes of scale. **A.** Lower transmission rate (0.007 instead of 0.01). **B.** Larger number of possible variants ($N_{\alpha 1} = 1,024$ instead of 256). **C.** Higher transmission rate (0.02 instead of 0.01). **D.** Longer maximum infection length (up to 50% of repertoire expressed per infection, instead of 25%).

and 5.8D).

5.4 Analysis of repertoire diversity

The factors that shape genetic diversity may also affect diversity at another level, in the repertoire population. In this analysis, the definition of the term "strain" is important. Parasites are considered to be members of the same strain with respect to β genes if they have identical sets of β antigenic types. Similarly, parasites may be separately sorted into strains with respect to α genes or domains. Population structure is analysed within each population of strains by measuring strain diversity and overlap. A low diversity index indicates that the population is evenly divided into a small number of distinct strains, and a low overlap index indicates that, on average, few antigenic types are shared by members of distinct strains. A population that has low values of both indices is considered to have "strain structure" [Gupta *et al.*, 1998, 1996].

5.4.1 Results of a simplified model

[Artzy-Randrup *et al.*, 2012] demonstrated that host immune selection can impose strain structure on *var* repertoires in a model with only one *var* type and no explicit sequence recombination. Consistent with that study, a simplified version of the model described in this chapter, with mutation of entire genes only and one domain per gene, produces strong strain structure. At any one time, both the α and the β sub-repertoire populations are dominated by a small number of strains. This is illustrated in the lower panels of Figure 5.9, where each colour represents a distinct strain and band width is proportional to frequency in the circulating parasite

population. The measure of diversity, shown in the top panel of Figure 5.9, is orders of magnitude lower than would be expected from a random mixture of repertoires. The measure of overlap, also shown in the top panel of Figure 5.9, is much lower than the maximum value of 1.

Modifying this simplified model to have unequal numbers of possible α and β variants makes only a small difference to diversity and overlap in the sub-repertoire populations, as shown in Figure 5.10A (recall that the $\alpha : \beta$ ratio in the repertoire is the same as the ratio N_α:N_β). Adding a transmission rate advantage for one gene type, as considered in Chapter 4, also has very little effect on strain structure (Figure 5.10B).

Previously it was shown that gene length can affect sequence diversity. The effect of gene length is similarly tested in the simplified model, without sequence recombination, by having two domains per α gene and imposing a functional constraint affecting one α domain. As before, the latter is implemented by setting $N_{\alpha 1} = N_\beta$ and $N_{\alpha 2} = N_\beta/4$. As shown in Figure 5.11, repertoire diversity is equal with respect to all domains. However, overlap is higher with respect to the more constrained second α domain, as expected because of the lower number of possible antigenic variants ($N_{\alpha 2}$) of this domain. Overlap is also higher with respect to the β domain, as expected because of the larger number of β genes per repertoire.

5.4.2 Results with sequence recombination

Next, the model is extended to include sequence recombination. In the simplest case of single-domain genes and $N_\alpha = N_\beta$, as shown in Figure 5.12A, diversity is very much higher than in the model with only simple genetic mutation. No strain accounts for more than 1% of the population. The diversity values are around 2,000

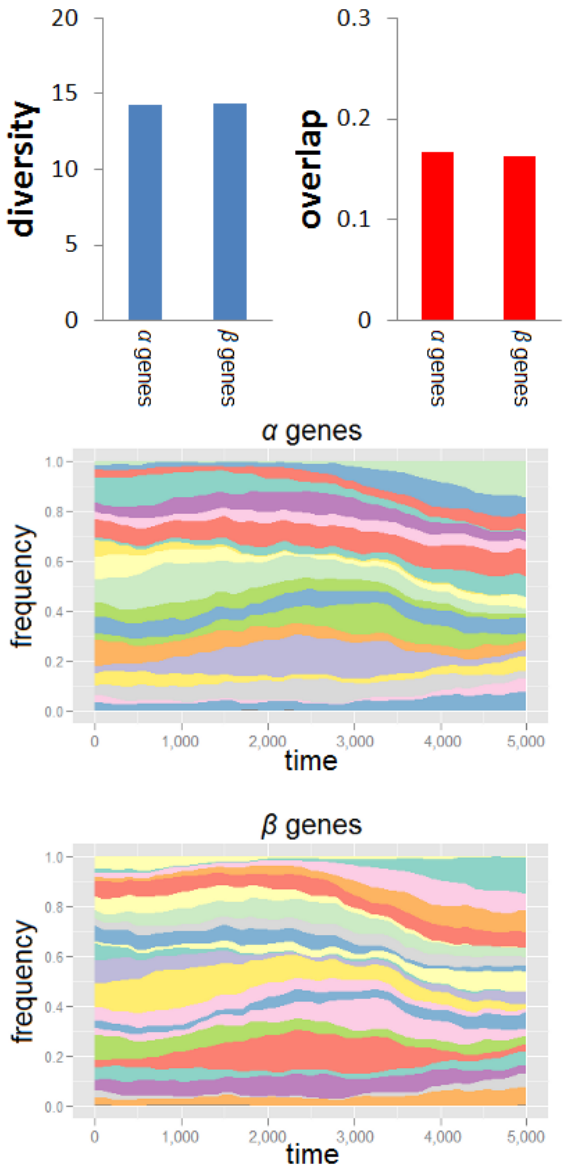


Figure 5.9. Diversity and overlap in sub-repertoire populations in a simplified model without recombination. In this model, instead of sequence recombination, genes undergo only simple mutation from one antigenic type to another. In the lower panels, each coloured band represents circulating strains with identical repertoires and dark grey represents strains that never exceeded 1% of the population. Results are shown for the final 5,000 parasite generations of the simulation. Here there are equal numbers of possible α and β variants ($N_\alpha = N_\beta = 256$). The contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2.

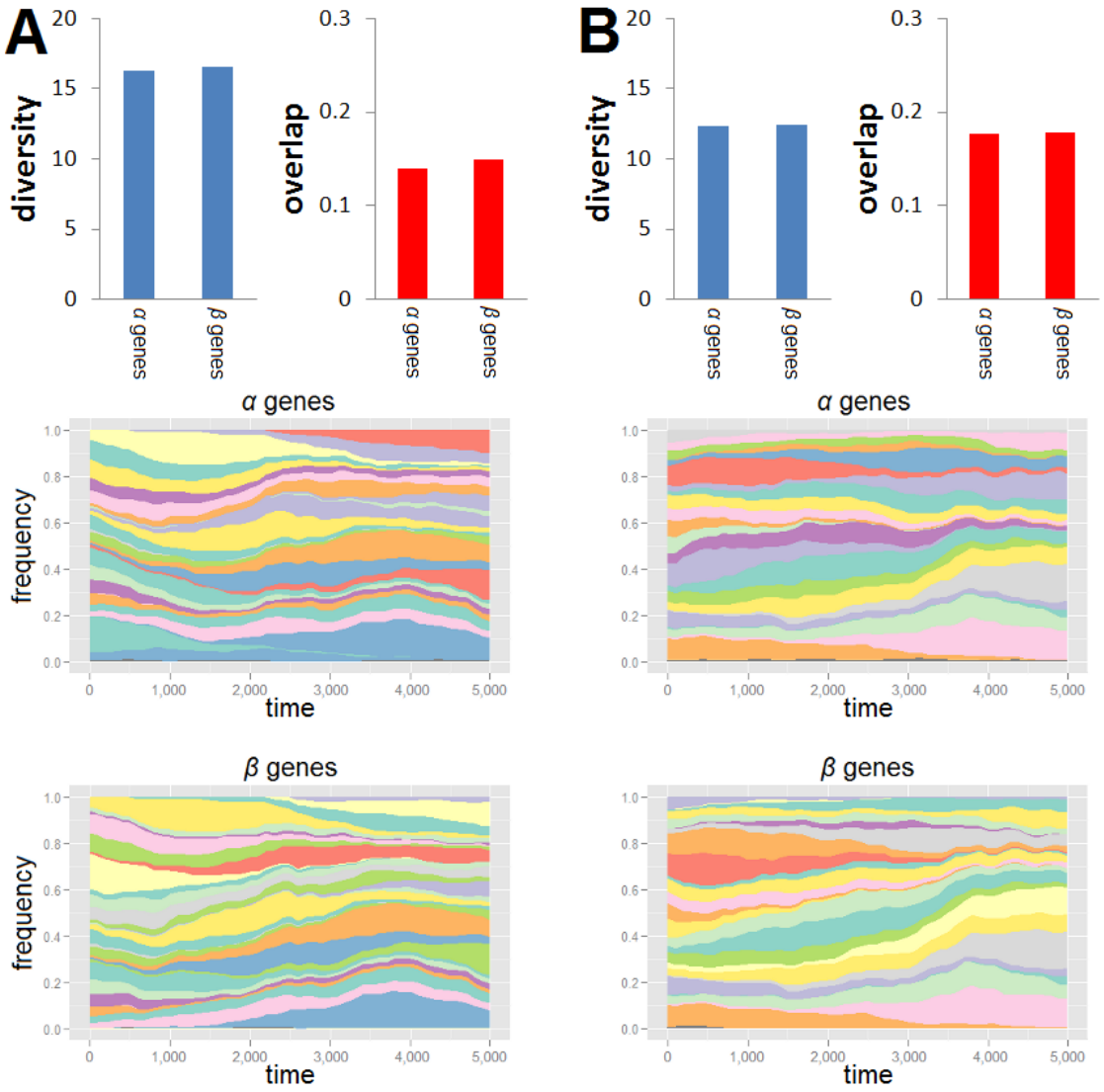


Figure 5.10. Diversity and overlap in sub-repertoire populations in a simplified model without recombination and with changed parameters. In this model, instead of sequence recombination, genes undergo only simple mutation from one antigenic type to another. In the lower panels, each coloured band represents circulating strains with identical repertoires and dark grey represents strains that never exceeded 1% of the population. Results are shown for the final 5,000 parasite generations of the simulation. The contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. **A.** Unequal numbers of possible α and β variants ($N_\alpha = N_\beta/4 = 64$). **B.** Equal numbers of possible α and β variants and α genes confer a 50% transmission rate advantage, as described in Chapter 4.

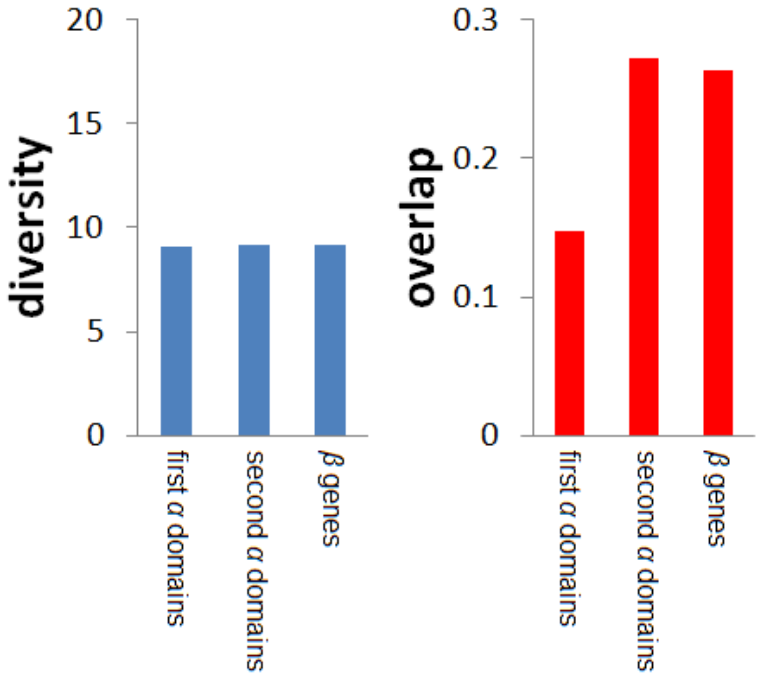


Figure 5.11. Diversity and overlap in a simplified model without recombination and with multiple domains. In this model, instead of sequence recombination, genes undergo only simple mutation from one antigenic type to another. The contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. As in the model of Figure 5.13, there are two domains per α gene and fewer possible variants for the second α domain ($N_{\alpha 1} = N_\beta = 256$ and $N_{\alpha 2} = N_\beta/4$).

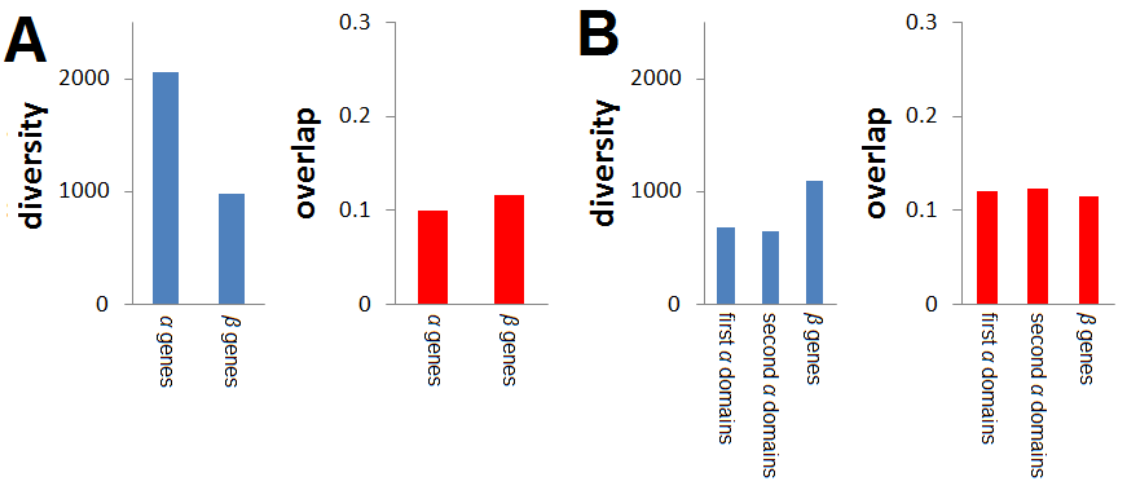


Figure 5.12. Diversity and overlap in the full model with recombination. The recombination rate is 0.05, the contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. **A.** One domain per gene and equal numbers of possible α and β variants ($N_\alpha = N_\beta = 256$). **B.** Two domains per α gene and equal numbers of variants for all domains ($N_{\alpha 1} = N_{\alpha 2} = N_\beta = 256$).

with respect to α sub-repertoires and 1,000 with respect to β sub-repertoires. For comparison, in a model in which all genes are continually randomly mutated, the corresponding diversities are approximately 8,000 and 10,000 (not shown). Overlap between strains, however, is at least as low as in the model without sequence recombination.

Figure 5.12A further shows that when $N_\alpha = N_\beta$ and α genes are expressed before β genes, the α sub-repertoire population is substantially more diverse than the β sub-repertoire population. Therefore in this model, expression order can affect not only sequence diversity but also repertoire population diversity.

Earlier it was shown that when there are two domains in each α gene and $N_\alpha = N_\beta$, sequence diversity changes so that α genes become less diverse than β genes. Similarly, adding a second α domain causes repertoire diversity to be higher with

respect to β genes than with respect to either α domain (Figure 5.12B).

5.4.3 Results with sequence recombination and unequal functional constraints

Finally, the effect of a functional constraint affecting one α domain is considered by setting $N_{\alpha 1} = N_\beta$ and $N_{\alpha 2} = N_\beta/4$. In this model, diversity and overlap in the β sub-repertoire population, as shown in Figure 5.13, are even higher than in the model with equal functional constraints. In sharp contrast, the α sub-repertoire population has much lower diversity with regard to both α domains, as shown in Figure 5.13. Overlap is also very low with respect to the first α domain. As in the model without sequence recombination (Figure 5.11), higher overlap is seen with respect to the second α domain and the β domain.

As shown in Figure 5.14A, no distinct strain, with respect to either α domain, accounts for more than a small fraction of the population. However, grouping strains that differ by 3 or fewer genes reveals much stronger strain structure, as shown in Figure 5.14B. Therefore the population is seen to be structured into a relatively small number of strains with minimal overlap between strains and with very high (although not necessarily complete) overlap within strains.

Importantly, this pattern is seen even though the first α domain is no more functionally constrained than the β domain. Comparison with Figure 5.11, which shows the results of a model with the same functional constraint but without sequence recombination, confirms that sequence recombination has an essential role in generating these results. As in the case of sequence diversity, the fact that multiple domains are combined within genes is seen to have a profound effect on repertoire

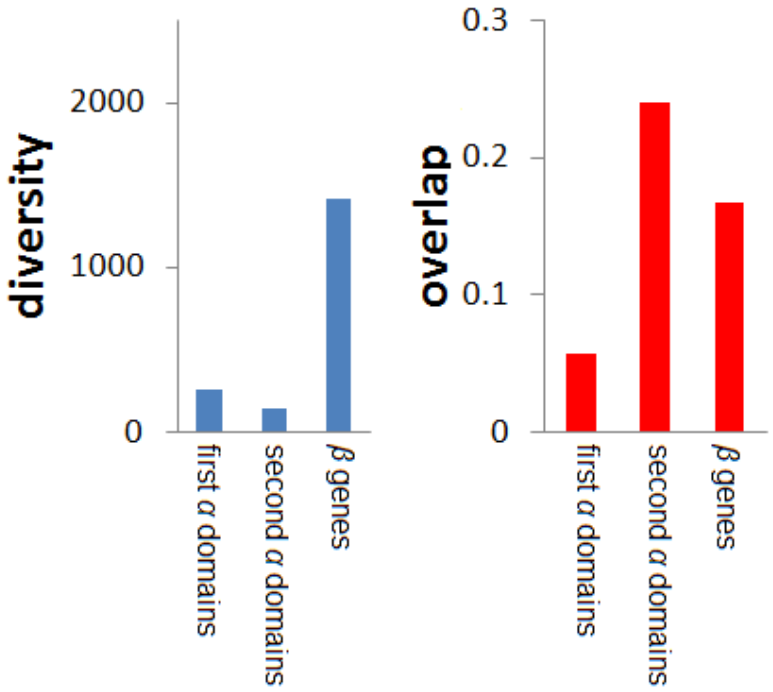


Figure 5.13. Diversity and overlap in the full model with recombination and a functional constraint on one domain. The recombination rate is 0.05, the contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. Here there are two domains per α gene and fewer possible variants for the second α domain ($N_{\alpha 1} = N_\beta = 256$ and $N_{\alpha 2} = N_\beta/4$).

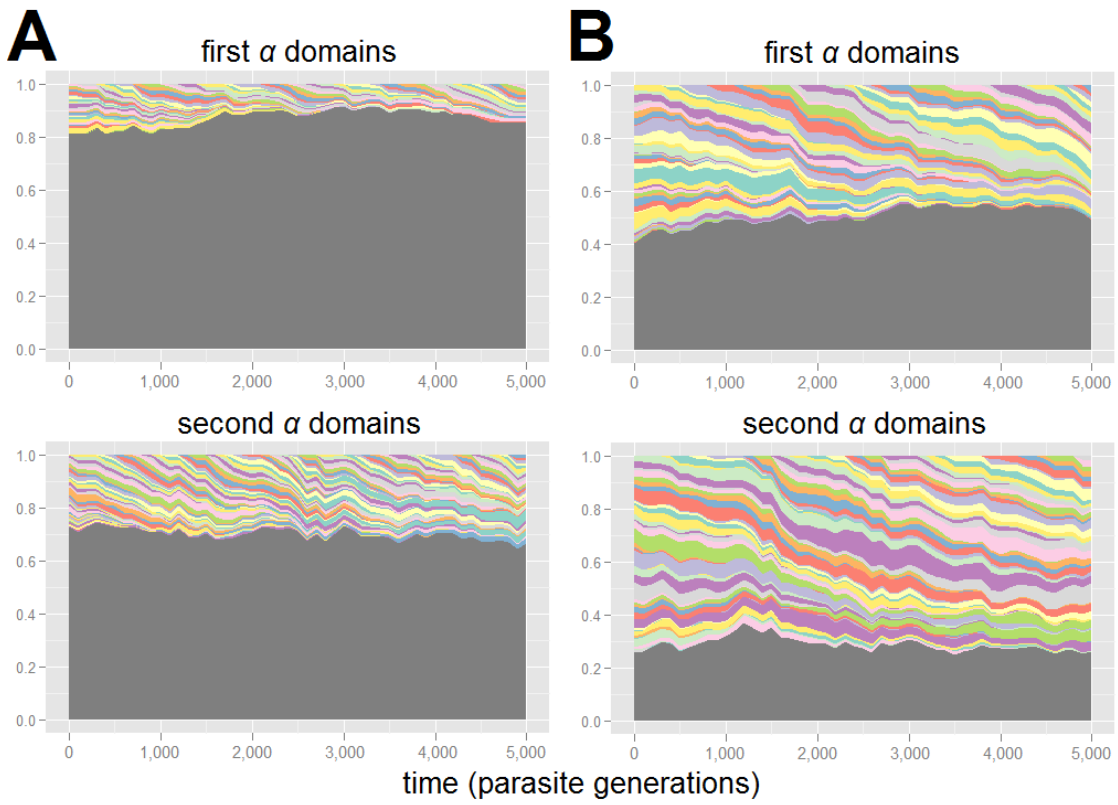


Figure 5.14. Strain structure in the full model with recombination and a functional constraint on one domain. Results are shown for the final 5,000 parasite generations of the simulation. Here there are fewer possible variants for the second domain ($N_{\alpha 1} = N_\beta = 256$ and $N_{\alpha 2} = N_\beta/4$). The recombination rate is 0.05, the contact rate is 0.01, and at most 25% of the repertoire is expressed in each infection. Other parameter values are as listed in Table 5.2. **A.** Each coloured band represents strains with identical domains and dark grey represents strains that never exceeded 1% of the population. **B.** Each coloured band represents strains that differ by 3 or fewer domains. The β sub-repertoire population is so diverse that no strain exceeds 1% frequency, even when strains are grouped together if they differ by 3 or fewer genes.

population diversity.

The results of this model are consistent with the findings of a study that inferred *P. falciparum var* repertoire diversity from serological data [Buckee *et al.*, 2009]. That study found evidence of strong strain structure only among the subset of *var* genes to which hosts first acquire immunity, which generally belong to Group A and encode a large number of binding domains.

5.5 Discussion

Mathematical modelling was used to investigate diversity of *P. falciparum* antigenic repertoires at multiple levels. Consistent with results from an analysis of *P. falciparum* data reported in Chapter 3, it was found that expression order can impose gradients of genetic diversity within *var* repertoires. Additionally, it was shown that differences in gene length can further influence genetic diversity such that model results qualitatively resemble the observations of Chapter 3. This outcome was found to be robust to changes of epidemiological parameters and to different assumptions regarding functional constraints on the encoded binding domains.

Two novel features of the model increased its biological realism. First, a scheme was devised for representing *var* genetic sequences, enabling a distinction to be made between genetic diversity and antigenic diversity. Without attempting to mimic the actual relationships between sequence and antigenic type, the scheme succeeded in reproducing several of its most salient features. Second, the model assumed that novel antigenic types will very rarely emerge unless they are favoured by immune selection. Both factors were found to be crucial in generating the reported results.

In a previous theoretical modelling study, Buckee & Recker [2012], showed that

it can be advantageous for genes to comprise domains with similar levels of antigenic diversity, assuming there is a trade-off between antigenic diversity and within-host fitness. That is, across all domains, genes will evolve either to optimise immune evasion or to optimise transmission probability. In the model used in this chapter, when one domain had low antigenic diversity (due to assumed functional constraints), the second domain in the same gene was unable to confer a benefit from greater antigenic diversity. Therefore the second domain would be under low selective pressure to retain high antigenic diversity, and might instead evolve to increase within-host fitness, as was demonstrated by Buckee & Recker [2012]. The results presented here suggest, however, that even if a domain is unable to affect within-host fitness, its sequence diversity will be reduced by the presence of additional domains in the same gene. Thus, the association between gene length and sequence diversity may be more general and fundamental than previously supposed.

In this chapter, the factors that affected sequence diversity were also found to shape repertoire diversity. As with sequence diversity, gene length was found to have a direct effect on repertoire diversity. Further, a functional constraint affecting one domain of a two-domain gene was found to restrict repertoire diversity with respect to the other, unconstrained domain. With respect to both these domains, the pathogen population was structured by immune selection into a relatively small number of minimally overlapping strains. The remainder of the repertoire, comprising shorter genes, was under weaker immune selection and did not evolve strain structure. These model outcomes are consistent with an analysis of serological data that revealed a similar pattern of multi-level diversity among *var* repertoires [Buckee *et al.*, 2009].

In conclusion, the results of this chapter give reason for caution when inter-

preting measures of *P. falciparum* diversity. There are many reasons why genetic diversity may be an unreliable indicator of antigenic diversity. And while understanding pathogen population structure is necessary for predicting the impact of interventions, it may also be misleading. As explained above, the model considered here indicates that some PfEMP1 domains (or possibly multiple epitopes within domains) may exhibit high sequence conservation and high cross immunity and yet be capable of greater antigenic diversity, without fitness cost, than is seen in a pathogen population. One way to address this issue is to learn more about the functional constraints on individual epitopes.

Chapter 6

Summary and Conclusion

Three features of *Plasmodium falciparum* especially contribute to its burden on human health. First, the parasite is able to infect hosts continually throughout their lives. Second, each infection typically lasts many months, greatly increasing the likelihood of transmission. Lastly, a significant number of infections, especially in young children, lead to severe disease and high risk of mortality. All three features have been linked to variant antigens expressed on the surface of infected red blood cells, and especially to PfEMP1 (Plasmodium falciparum erythrocyte membrane protein 1) [Bull *et al.*, 1998, Chan *et al.*, 2012].

PfEMP1 is enormously variable and it is thought that antibody responses to PfEMP1 are largely variant-specific, enabling parasites to evade immunity at the population level by expressing variants to which the host has had no previous exposure [Giha *et al.*, 1999, Marsh & Howard, 1986, Marsh *et al.*, 1989, Ofori *et al.*, 2002]. During infection, parasites further evade immunity by antigenic variation. This process involves the sequential expression of PfEMP1 variants encoded by a repertoire of approximately 60 *var* genes [Gardner *et al.*, 2002]. Each of these ex-

pressed PfEMP1 variants must perform the essential function of mediating binding to other host cells, to prevent infected erythrocytes being carried to the spleen and destroyed [Boone & Watters, 1995, Buffet *et al.*, 2009]. PfEMP1 variants with particular binding types have been linked to severe malaria (reviewed by Smith *et al.* [2013]). Thus, it appears that interactions between PfEMP1 antigenic diversity, antigenic variation and binding phenotype are important in *P. falciparum* immune evasion and pathology. The aim of this thesis was to shed light on the nature of these interactions and their evolution.

Previous studies have fallen short of fully describing the dynamics of antigenic variation in *P. falciparum*. It is known that each parasite expresses only one *var* gene at a time [Chen *et al.*, 1998, Scherf *et al.*, 1998], and that parasites occasionally change from one active gene to another by transcriptional switching [Kyes *et al.*, 2007]. Non-random switching patterns have been identified [Frank *et al.*, 2007, Horrocks *et al.*, 2004, Recker *et al.*, 2011, Roberts *et al.*, 1992] but an overall network has not been fully elucidated. This is partly due to the difficulty of obtaining sufficient data but is also due to the large dimensionality of the system, which makes parameter estimation challenging. It is here that the method developed in Chapter 2 and applied in Chapter 3 makes an important contribution. The method involves optimisation algorithms designed to estimate parameters of antigenic switching networks from transcriptional measurements of *in vitro* cultures. The dynamical model is the same as that previously used by Recker *et al.* [2011] but the statistical techniques are novel. The method was shown to perform well on artificial data sets subjected to random noise, designed to mimic measurement errors apparent in the results of earlier studies. Previously, conclusions have been based on point estimates of switching parameters, without rigorous statistical analysis [Frank

et al., 2007, Recker *et al.*, 2011]. In contrast, the Markov chain Monte Carlo method described here provides measures of uncertainty for all estimated parameter values and permits statistical testing of competing hypotheses regarding network structure. Given sufficient data, this method enables entire antigenic switching networks to be described in unprecedented detail.

Applying this novel method to experimental data yielded the first complete description of an antigenic switching network in *P. falciparum*. The results confirm previous findings that *var* genes switch from transcriptionally active to inactive at different rates [Roberts *et al.*, 1992] and that these "off-rates" depend on chromosomal location [Frank *et al.*, 2007]. Arguably a more important factor shaping *in vivo* dynamics is the direction of switching – that is, when one gene becomes inactive, what are the probabilities of each other gene being activated? The direction of *var* switching was first considered in detail by Recker *et al.* [2011], who found evidence for a highly structured switching network. In this network, some switches are much more favoured than others and the direction of switching depends on which gene is becoming inactive. As demonstrated in Chapter 2, however, the data set and methods used in that study led to uncertainty in its findings. Instead of a highly structured network, the results of Chapter 3 imply a different pattern in which the direction of switching is largely governed by a hierarchy of activation biases, with relatively little dependence on the deactivating gene. Nevertheless, this outcome is not entirely at odds with previous results. The switching pattern identified here has properties in common with a theoretical network that Recker *et al.* [2011] generated by optimising over evolutionary traits. Genes with high activation biases and low off-rates correspond to the frequently-switched-to "sink" nodes in the theoretical network. Genes with low activation biases and high off-rates correspond to "source"

nodes, which enable a parasite strain to seed diversity in a host that is immune to parts of the strain's antigenic repertoire. There is also evidence of some variation in switching biases dependent on the deactivating gene. Further research is needed to determine the magnitude of this variation and to predict its effects on switching in semi-immune hosts.

The activation hierarchy determined here strongly favours switching to one particular subset of genes within the *var* repertoire. These genes have central chromosomal location, which helps to explain why centrally located *vars* are more highly transcribed *in vitro*, compared to subtelomeric *vars* [Gatton *et al.*, 2003, Peters *et al.*, 2002, 2007, Zhang *et al.*, 2011]. Perhaps more significant is the finding that the most frequently activated genes also encode a particular set of extracellular binding domains. This novel result suggests a link between *var* activation and PfEMP1 binding phenotype. The *vars* found to have highest activation biases encode relatively few extracellular domains, which mediate binding to host receptor CD36 [Baruch *et al.*, 1997, Robinson *et al.*, 2003]. Genes encoding a larger number of domains, associated with binding to other host receptors [Smith *et al.*, 2013], were found to be much less frequently activated by transcriptional switching. A possible explanation for this pattern is the hypothesis that the latter binding types confer higher within-host survival rates [Bull *et al.*, 2000, 1999, Smith *et al.*, 2013], in which case the corresponding genes might not require high activation rates.

Although many authors have cited the hypothesis that PfEMP1 binding types confer different parasite survival rates, the evolutionary advantages of this system have not been investigated. The virulence trade-off hypothesis posits that a pathogen will evolve to balance transmissibility and risk of host mortality [Alizon & van Baalen, 2008, Anderson & May, 1982]. In *P. falciparum*, transmission suc-

cess is only weakly correlated with asexual parasite numbers [Babiker *et al.*, 2008, Bousema & Drakeley, 2011]. Therefore, it can be argued, it is not immediately obvious why some *var* genes would confer high effective multiplication rates in non-immune hosts, leading to very high parasitemia and increased risk of host death. Mathematical modelling in Chapter 4 offers one potential explanation. The model assumed that one *var* binding phenotype is associated with higher transmission fitness for reasons not directly related to parasitemia or asexual parasite survival rate. It then showed that parasites will evolve to express this particular phenotype first during infections. One way to impose such an expression order, perhaps more effectively than by using activation biases, is for PfEMP1 variants to confer different effective multiplication rates. This leads to the speculative hypothesis that variants with higher transmission fitness will evolve to confer higher effective multiplication rates, to ensure that parasites expressing these variants out-compete the rest of the population during infection of a non-immune host. This expression order was favoured even when it was associated with increased risk of host mortality. Therefore it is suggested that transmission fitness, dependent on *in vivo* expression order, might have contributed to the evolution of virulence in *P. falciparum*. Even if this hypothesis is true, however, it is likely that transmission fitness is only one of several contributory factors, possibly including within-host competition between strains. Although multiple *P. falciparum* infections are the norm in high prevalence areas, little is known about how strains interact. Resolving this issue would enable the creation of more sophisticated models, substantially improving understanding of the parasite's evolution and epidemiology.

Chapter 4 also explored how the existence of multiple PfEMP1 binding phenotypes could restrict diversity of *var* repertoire structures. Sequenced *var* repertoires

from around the world are similarly partitioned with respect to major binding phenotypes [Bull *et al.*, 2008, Rask *et al.*, 2010]. Previous mathematical modelling has suggested that consistent repertoire structure might result from a trade-off between within-host fitness (dependent on binding phenotype) and between-host transmission success (dependent on antigenic diversity) [Buckee & Recker, 2012]. That model favoured the evolution of repertoires containing a mix of two antigenic types, one optimised for within-host fitness and the other optimised for antigenic diversity. The evolutionary model in Chapter 4 confirmed that differences in gene expression period and antigenic diversity could affect the mean number of genes of each type per genome. However, these factors by themselves did not lead to the consistency observed in *var* repertoire partitioning. Instead, it was shown that cross immunity, acting more strongly within *var* types than between types, might drive the evolution of consistent *var* repertoire structure in parasite populations. Nevertheless, this result is not inconsistent with the previously invoked idea that PfEMP1 types have diverged to satisfy a trade-off between antigenic diversity and within-host fitness.

Consequences of the latter hypothesis were explored in Chapter 5, which focussed on the evolution of *var* sequence diversity and *var* repertoire diversity. This was partly inspired by a novel result of Chapter 3: the hierarchy of activation biases, which was shown to govern *var* transcription levels *in vitro*, was also found to correlate with *var* sequence diversity. Activation bias is assumed to be correlated with *var* transcription *in vivo* as well as *in vitro*, at least among genes with similar binding phenotypes. Therefore a possible explanation for the observed trend is that *var* transcription has a role in generating *var* sequence diversity. The mathematical model in Chapter 5 was designed to test the hypothesis that immune selection can explain this pattern, even if *var* transcription does not directly facilitate recombi-

nation or mutation. Several studies have measured *var* sequence diversity, which has generally been assumed to correlate with antigenic diversity. A crucial feature of the model used here was the way in which it distinguished between these two forms of diversity, allowing multiple sequence variants to encode the same antigenic type. A second essential feature was the assumptions made about the generation of sequence diversity. It was assumed that although all *var* genes may frequently undergo recombination, host immune selection determines whether novel variants proliferate within the host and are transmitted. These conditions were found to lead to a correlation between expression order and sequence diversity within *var* types, consistent with the observations reported in Chapter 3.

The model of Chapter 5 also sought to explain observed differences in sequence diversity between *var* types. In this case it was found that gene length – meaning the number of encoded extracellular domains – has an important role. Longer genes were found to have lower sequence diversity. This is a result of assuming that novel variants must be favoured by immune selection. Assuming immune responses are domain-specific, a sequence modification will be selected for only if it causes a change in the antigenic type of every domain in a gene. If the recombination rate is the same for each domain then genes with more domains are less likely to achieve antigenic novelty through recombination. Therefore these results suggest that gene length itself – that is, the accumulation of many domains into one gene – has a direct effect on sequence diversity. This is consistent with the observed correlation between *var* gene length and sequence conservation, as recently reported [Buckee & Recker, 2012]. Such modelling results underscore the limitations of *var* diversity studies that are based on sequence comparisons alone.

Further results of the model in Chapter 5 concerned repertoire population di-

versity. As in previous studies, the objective here was to determine what conditions do or do not lead to the evolution of strain structure in *P. falciparum*. The latter can be defined as the structuring of the pathogen population into a relatively small number of strains with high antigenic overlap within strains, and minimal antigenic overlap between strains [Gupta *et al.*, 1998, 1996]. Previously, *P. falciparum* strain structure was shown to arise in a model with only one type of antigenic variant [Artzy-Randrup *et al.*, 2012]. The model in Chapter 5 assumed two types of variant with no recombination between types, so that each repertoire comprised two distinct sub-repertoires. Strain structure was assessed with respect to each of these sub-repertoire populations. As in the analysis of sequence diversity, gene length was found to have a direct effect on limiting sub-repertoire population diversity, corresponding to fewer circulating strains. Adding a second domain to one gene type lowered the diversity of its sub-repertoire population, measured with respect to either domain.

The same model was used to show how functional constraints on PfEMP1 domains might affect strain structure. There is evidence that at least some larger PfEMP1 variants are more antigenically restricted [Cham *et al.*, 2010, 2009, Ghumra *et al.*, 2012, Warimwe *et al.*, 2009]. One hypothesis is that this is due to functional constraints related to binding phenotype [Bull *et al.*, 2000, 1999, Smith *et al.*, 2013]. However, these functional constraints might not apply to other domains in the antigen that are not directly involved in binding. According to the results of Chapter 5, a stronger functional constraint on one domain can lead to strain structure not only with respect to the more constrained domain, but also with respect to less constrained domains within the same antigen. Conversely, in this model, there is no strain structure with respect to antigens containing only less functionally con-

strained domains. The overall pattern is qualitatively similar to one found by inferring *P. falciparum* strain structure from serological data [Buckee *et al.*, 2009]. While these results suggest that it is worth pursuing PfEMP1-based vaccines to combat severe malaria, they also highlight the need to understand the binding phenotypes of PfEMP1 domains, as well as their antigenicities.

In summary, this dissertation reveals evolutionary trade-offs at the heart of *P. falciparum* immune evasion and pathology, and emphasises the central role of host immunity in shaping pathogen population diversity at multiple levels. While the mathematical models considered here are simplifications, they are also strongly informed by biological observations, resulting from novel data analysis as well as from previous research. Hypotheses generated from these models provide stimulus for further laboratory and field studies. In particular, much of what is reported here has been inferred from *in vitro* studies of lab-adapted parasite strains. It will be important to verify these findings using field isolates and experimental conditions more closely resembling the within-host environment. Improved methods of sequencing *var* genes and analysing sequence diversity are also opening up exciting areas for future research. It is becoming possible to sequence very large numbers of *var* genes from field isolates. Analysis of conserved and variable regions within such sequences promises to improve understanding of their short- and long-term evolution. These findings will enable model predictions to be tested and more sophisticated models to be developed. Instead of only two gene types, future models might include multiple domain types with different antigenic and binding characteristics based on empirical studies. Ultimately, this would allow one to predict the impact of vaccines or other interventions aimed at specific PfEMP1 targets. The models developed in this dissertation help provide the groundwork for such an endeavour.

Appendices

OPEN ACCESS Freely available online

PLoS one

A Statistically Rigorous Method for Determining Antigenic Switching Networks

Robert Noble[1], **Mario Recker***

Department of Zoology, University of Oxford, Oxford, United Kingdom

Abstract

Many vector-borne pathogens rely on antigenic variation to prolong infections and increase their likelihood of onward transmission. This immune evasion strategy often involves mutually exclusive switching between members of gene families that encode functionally similar but antigenically different variants during the course of a single infection. Studies of different pathogens have suggested that switching between variant genes is non-random and that genes have intrinsic probabilities of being activated or silenced. These factors could create a hierarchy of gene expression with important implications for both infection dynamics and the acquisition of protective immunity. Inferring complete switching networks from gene transcription data is problematic, however, because of the high dimensionality of the system and uncertainty in the data. Here we present a statistically rigorous method for analysing temporal gene transcription data to reconstruct an underlying switching network. Using artificially generated transcription profiles together with *in vitro var* gene transcript data from two *Plasmodium falciparum* laboratory strains, we show that instead of relying on data from long-term parasite cultures, accuracy can be greatly improved by using transcription time courses of several parasite populations from the same isolate, each starting with different variant distributions. The method further provides explicit indications about the reliability of the resulting networks and can thus be used to test competing hypotheses with regards to the underlying switching pathways. Our results demonstrate that antigenic switch pathways can be determined reliably from short gene transcription profiles assessing multiple time points, even when subject to moderate levels of experimental error. This should yield important new information about switching patterns in antigenically variable organisms and might help to shed light on the molecular basis of antigenic variation.

Citation: Noble R, Recker M (2012) A Statistically Rigorous Method for Determining Antigenic Switching Networks. PLoS ONE 7(6): e39335. doi:10.1371/journal.pone.0039335

Editor: Mike B. Gravenor, University of Swansea, United Kingdom

Received November 8, 2011; **Accepted** May 23, 2012; **Published** June 22, 2012

Copyright: © 2012 Noble, Recker. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: RN is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) studentship. MR is a Royal Society University Research Fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mario.recker@zoo.ox.ac.uk

Introduction

For pathogenic organisms whose opportunities for transmission are episodic, rather than continuous, long infectious periods are crucial for successful transmission. Most notably vector-borne and sexually transmitted pathogens have evolved various strategies to increase their transmission potential by evading their hosts' immune responses. One of the most sophisticated strategies is antigenic variation whereby the pathogen repeatedly changes its antigens over the course of a single infection. Although the underlying mechanisms differ between pathogens, many involve a polymorphic family of genes encoding functionally similar but antigenically diverse variants (reviewed in [1]). Of these genes, only one is actively transcribed while all other genes are transcriptionally silent. During infection, the active gene can 'switch off', leading to the activation, or 'switching on', of a previously silent gene. It is believed that this mono-allelic gene expression pattern helps the pathogen to guard its available antigenic repertoire from the immune system. Furthermore, gene expression has to be coordinated across the whole parasite population, such that the majority of the population expresses the same gene at the same time, as the host would otherwise build up immunity to all variants early on and clear the infection prematurely. As a result, infections with antigenically variable pathogens are often characterised by successive waves of parasitemia that are sequentially dominated by one or only a few antigenic variants.

Some of the best studied organisms employing antigenic variation are African trypanosomes and the causative agent of severe malaria in humans *Plasmodium falciparum*. Various theoretical studies have concentrated on determining the underlying mechanisms responsible for the observed coordination in antigen presentation during infection by these pathogens. In particular, gene activation hierarchies or differences in growth rates have been put forward as potential drivers behind their characteristic infection dynamics [2–7]. Although it was found that parasite intrinsic factors could orchestrate the parasite population in the initial phases of infection, they are insufficient for maintaining sequential dominance of antigenic variants during the later, chronic stages of infections. Instead, immune selection via cross-reactive immune responses has been proposed to offer a more parsimonious solution to this problem, even in the absence of structured differences in switch or growth rates [8]. Nevertheless, *in vitro* studies of malaria parasites have since shown that variant switching is non-random and partly gene specific. For example, Horrocks *et al.* [9] demonstrated that *var* genes, which encode the surface-expressed virulence factor PfEMP1 (*P. falciparum* erythrocyte membrane protein 1), switch on and off at different but constant rates during long term culture. Frank *et al.* [10] also found that these rates differ widely between different genes and

that centrally located genes appear to have lower switch rates than those in subtelomeric loci. This has recently been confirmed by Enderes *et al.* [11] who also showed that these switch preferences appear to be independent of genetic background. Similarly, in *Trypanosoma brucei* it has been shown that switching between *vsg* genes, which encode the pathogens surface coat, is determined by a fixed hierarchy of activation probabilities that appear to depend on features of the genes' loci [7,12,13]. An expression hierarchy dependent on two particular extragenic elements has also been found to underlie antigenic variation in the spirochete *Borrelia hermsii* [14]. Such observations have led to the hypothesis that structured switching or switch hierarchies might be important for structuring the parasite population during the early stages of infection. On the one hand this would make it easier for the adaptive immune response to desynchronise the parasite population and set up a pattern whereby single variants can successively dominate the infection [8]. On the other hand it would also help the parasite to establish an infection in individuals with pre-existing immune responses, as recently shown in [15].

To distinguish between parasite-intrinsic switching and host mediated selection during infection it is important to have a thorough understanding of inherent switch patterns. One approach to investigate these patterns is to analyse longitudinal gene transcription data from *in vitro* cultured parasites in the absence of selection. Given sufficient transcription data, it should theoretically be possible to determine the switching network of a complete gene repertoire. However, a major challenge is the large sample space which makes simple methods unreliable or impractical. For instance, to fully determine the switching network of a pathogen with a repertoire of 10 variant genes requires 90 parameters to describe every gene's rate of activation and silencing (assuming that each gene's activation probability may depend on which gene is being switched off). Using a grid search method testing five values for each parameter would require 8.1×10^{62} simulations and would probably miss the globally optimal solution by a wide margin. Recently, Recker *et al.* [15] addressed this problem using an iterative algorithm to determine the switching pattern of *P. falciparum var* genes, using data from clonal parasite populations followed over 60–80 generations, with *var* transcription levels measured at regular intervals using quantitative 'real-time' PCR (qrt-PCR). Although this method could robustly determine major switch pathways, it was able to explore only a small subset of the many-dimensional parameter space. Furthermore, qrt-PCR data are subject to measurement error, especially when transcription levels are very low, and the algorithm could not provide an indication of the uncertainty of its parameter estimates.

Here we present a statistically rigorous solution to this problem in the form of simulated annealing and Markov Chain Monte Carlo (MCMC) algorithms. Using diverse sets of artificially generated gene transcription time courses we demonstrate that highly structured and complex switch pathways can be resolved reliably from relatively limited data. We further show that although experimental noise can have a major effect on estimates, using transcription measurements from several populations of the same pathogen strain can significantly improve accuracy. Our method can thus be used to resolve complex and high-dimensional switch patterns with high reliability and accuracy from limited and noisy data.

Results

The general aim of our method is to find a set of parameter values relating to antigenic switching, i.e. the switch rates and activation biases, that best fit temporal variant transcript distributions. Under the assumption that a variant's switch rate and biases are constant over time we can describe the proportion of a parasite population transcribing gene variant i at time t, v_i^t, simply as

$$v_i^t = (1-\omega_i)v_i^{t-1} + \sum_j \omega_j \beta_{ji} v_j^{t-1},$$

where ω_i is the rate at which gene i is switched off and β_{ji} is the switch bias, or probability of a switch from variant j to variant i. The switch network or pathway underlying a measured change in transcription levels can thus be described by the combination of a switch matrix $\boldsymbol{\beta} = (\beta_{ij})$ and off-rate vector $\boldsymbol{\omega} = \omega_1, \ldots, \omega_n$. The task then is to use an iterative approach (see Methods) to find $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ such that the deviation between the measured and simulated transcript levels at time points $t_1, \ldots, t_p$ is minimised.

Method Testing

To test our method for accuracy and reliability in resolving antigenic switch patterns from gene transcription time courses we used various test parameter sets representing a wide spectrum of possible switch pathways (see Methods). That is, we constructed a number of artificial switch networks of different degrees of complexity, in terms of the number of genes that each gene switches to, and the number of consecutive switch events, and used these to generate temporal gene transcription data. We then applied our methods in order to reconstruct the most likely switch pathway underlying the data.

In line with previous *var* gene transcription studies we assumed that the parasite population is initially clonal, with every parasite transcribing the same gene, and that relative transcript levels are measured at several time points during *in vitro* culture. A common feature of the previously described transcription profiles is that only a fraction of transcripts reach significant levels during *in vitro* culture [9–11,15]. We have previously argued that less transcribed genes are unlikely to play a dominant role in the switching network in the neighbourhood of the starting gene. We therefore assumed that removing data for genes that are generally transcribed at very low levels would have little effect on the parameter estimates for the interactions between more dominant genes, and that these interactions would be sufficient to determine the most important properties of the switching network. Accordingly, our initial analysis was restricted to a set of 10–16 genes, which could represent the subset of the most dominantly transcribed genes. We later provide a numerical justification for using such a reduced system to determine major transcription pathways.

Resolving switch networks from a single transcription time series. Initially we considered four switching networks of different complexities consisting of 10 genes. The first network describes a situation in which each gene simply switches to only one other gene with very high bias, and is referred to as a one-to-one (1:1) pathway. The second network, referred to as single-many-single (SMS), is similar to that proposed to underlie *var* gene transcription data in a previous study [15]. It describes switching from the initial gene to a group of genes, which all switch with high bias to another single gene. The latter gene in turn switches to a different group of genes, which switch back to the first gene. The third network has a lattice-like structure containing block-diagonal switch biases, and the fourth is a uniform network, where each gene has identical switch biases. The four networks are illustrated in Figure 1 together with the matrix and vector representation of switch biases and off-rates, $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ respectively, and the simulated transcription time courses resulting from these networks of up to 60 generations post cloning.

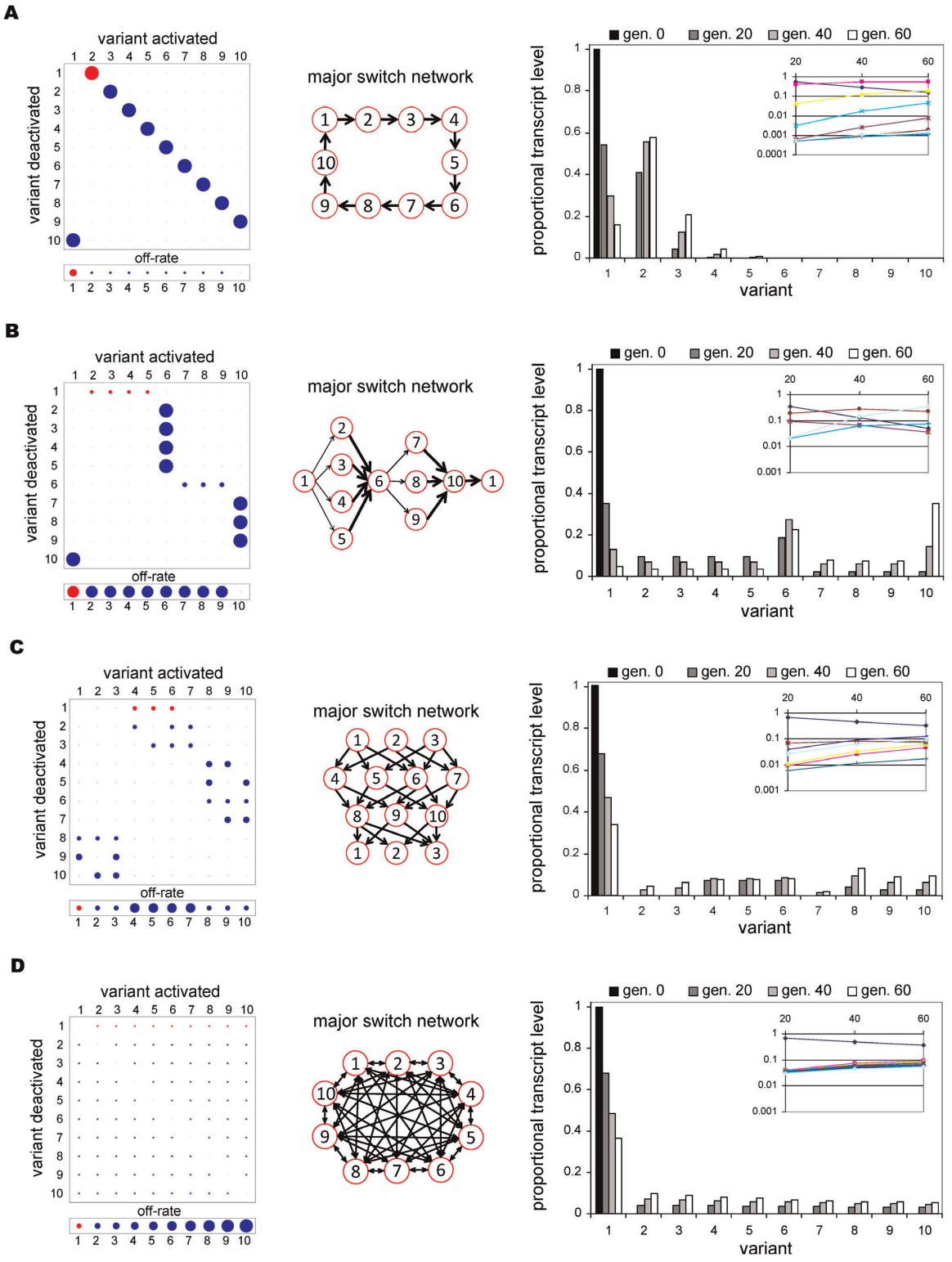

A
variant activated
variant deactivated
off-rate
major switch network
gen. 0
gen. 20
gen. 40
gen. 60
proportional transcript level
variant
B
variant activated
variant deactivated
off-rate
major switch network
gen. 0
gen. 20
gen. 40
gen. 60
proportional transcript level
variant
C
variant activated
variant deactivated
off-rate
major switch network
gen. 0
gen. 20
gen. 40
gen. 60
proportional transcript level
variant
D
variant activated
variant deactivated
off-rate
major switch network
gen. 0
gen. 20
gen. 40
gen. 60
proportional transcript level
variant

Figure 1. Target network and transcript data for initial testing. Four switch networks of different complexities were considered for the initial testing and method calibration: (**A**) one-to-one (1:1), (**B**) single-many-single (SMS), (**C**) lattice, and (**D**) uniform. The sizes of the discs in the switch matrices correspond to the transition biases from variant i to variant j, β_{ij}, and the sizes of the discs in the off-rate vectors are proportional to the per-generation de-activation rates, ω_i. The major switch pathways described by these matrices are highlighted in the middle column and the right column shows the proportional transcription levels of all 10 variants from point of cloning until 60 generations post cloning, taken at 20 generation intervals; the insets depict the proportional transcript levels of the 10 variants on a log scale at generations 20, 40 and 60, with each colour representing a different variant.
doi:10.1371/journal.pone.0039335.g001

Using simulated annealing (see Methods) we aimed to recover the original switching parameters from the transcript levels measured at generations 20, 40 and 60. Figure 2 shows that in each case the method found a good fit between the target and simulated transcript distributions, and it reliably recovered the first four consecutive switch events of the 1:1, SMS and uniform target networks. The switch bias matrix and off-rate vector were less well resolved for the later stages of the 1:1 network (Figure 2A), which comprises ten consecutive switches, and for the lattice network, which has an especially complex structure. Note that parameters associated with the later network stages, i.e. of those genes that are activated only after several switch events, have relatively minor effects on transcript levels and the simulated annealing algorithm is less sensitive to their variation. That is, even when the model output fits the transcript time courses of these variants very tightly (for all networks the average overall error, or deviation between the input data and the model output, was less than 0.1%), the particular pathways between rarely transcribed genes are less accurately described.

We next investigated the effect of noise, in terms of experimental error (see Methods), on the method's ability to recover underlying switch patterns. As shown in Figure 3 and S1, accuracy dropped markedly when noise was added to the data. However, under moderate levels of noise, major switch pathways could still be determined with reasonable accuracy, dependent on the particular network structure. The mean error increased linearly with the noise level (Figure 3C) but there was only a weak correlation between these two variables. It would therefore be impossible to use this error to estimate precisely the degree of noise in an experimental data set. Furthermore, we found a very weak correlation between the noise level and the error in the switch bias estimates (results not shown), indicating that there would be wide variation in the accuracy of parameter estimates derived from experimental data of this type.

To improve the method's accuracy we increased the number of time points at which transcript levels were measured. For the 1:1 network with noise added, changing from three to seven time points resulted in a 22% decrease in the root mean square deviation between the output switch biases and the target parameters for the first four network stages. More than doubling the number of measurements to 15 yielded only a 16% additional improvement. This non-linear relationship between the number of time points and the deviation of the parameter estimates indicates that the method's accuracy in describing the entire switch pathways of a parasite's gene repertoire is not simply limited by the frequency of measurements.

Resolving switch networks from multiple transcription time series. Since our method resolved the underlying switch networks with limited resolution for those genes whose transcript levels remained low over the whole time course, we next tested whether accuracy could be improved by using several target data sets generated by the same network. That is, using the same global switch pathway we generated transcription time courses from different clonal starter populations and then tried to find the most likely switch pathway by fitting the model to all of the data simultaneously. We increased the size of the parasite's repertoire to consist of 16 genes (requiring 240 parameters) and randomly selected eight of these to be the clonally expressed starter genes for separate data sets, corresponding to eight clones taken from the same parent culture. Transcription levels were recorded for each data set at five time points at generations 20, 30, 40, 60 and 80 post cloning.

As shown in Figure 4 (middle column), despite the high complexity of the underlying networks and size of the parameter space all four switch pathways were consistently recovered with high accuracy and resolution by using multiple time series simultaneously. Even when we considered a high degree of experimental noise, the switch bias and off-rates were accurately predicted for the starter genes (highlighted in red). It should be noted that the level of noise considered here was considerably higher than what we would expect of experimental data and corresponds to transcription level measurements typically differing from the true values by a factor of two. Although the parameter estimates for non-starter genes (blue discs) were less reliable at moderate and high noise levels, they were considerably more accurate than would be expected by chance. Switch biases of non-starter genes were recovered most accurately for the 1:1 network and less accurately for the SMS and lattice network. This is not surprising, however, given the high degree of complexity of these particular switching networks. Furthermore, the noisy data were fitted more closely by the model predictions than by the true, i.e. noise-less, transcription levels (the error was on average 15% smaller, independent of network type or noise level), which confirms that inaccuracies in the parameter estimates were mostly due to the noise and not to a failure of our method to locate good optima. Figure 5 shows two examples of the transcription time courses of the 10 most highly transcribed genes from two different starter genes (2 and 6) of an underlying 1:1 switch pathway, with and without added noise, together with the transcript distributions resulting from the estimated switching networks. This again highlights the robustness of this method to determine genetic switch patterns from limited and noisy data.

Similarly to when we used just one transcript time series, the error increased linearly with the noise level but this time the two variables were strongly correlated (data not shown). The error value obtained from a sufficiently large set of experimental data would therefore be a good estimator of the degree of experimental noise in the data and could be used as an input parameter for determining the uncertainty of the resolved switching network using a Markov Chain Monte Carlo (MCMC) approach described below.

MCMC. One explanation for the inaccuracy of the simulated annealing method in estimating the switch parameters for non-starter genes, and generally genes that remain at low transcript levels during the whole experiment, is that the error is less sensitive to their variation. This would imply the existence of alternative switch networks that fit the target data almost as well as the predicted best-fit solution. To investigate the level of uncertainty and obtain a probability distribution for each parameter, we used a Markov Chain Monte Carlo (MCMC) method (see Methods).

As demonstrated in Figure 6, using the same noisy transcription time series as in Figure 4, we found that the MCMC method produced generally accurate and precise estimates of starter gene parameters (indicated by less fuzzy rings) in all four network types.

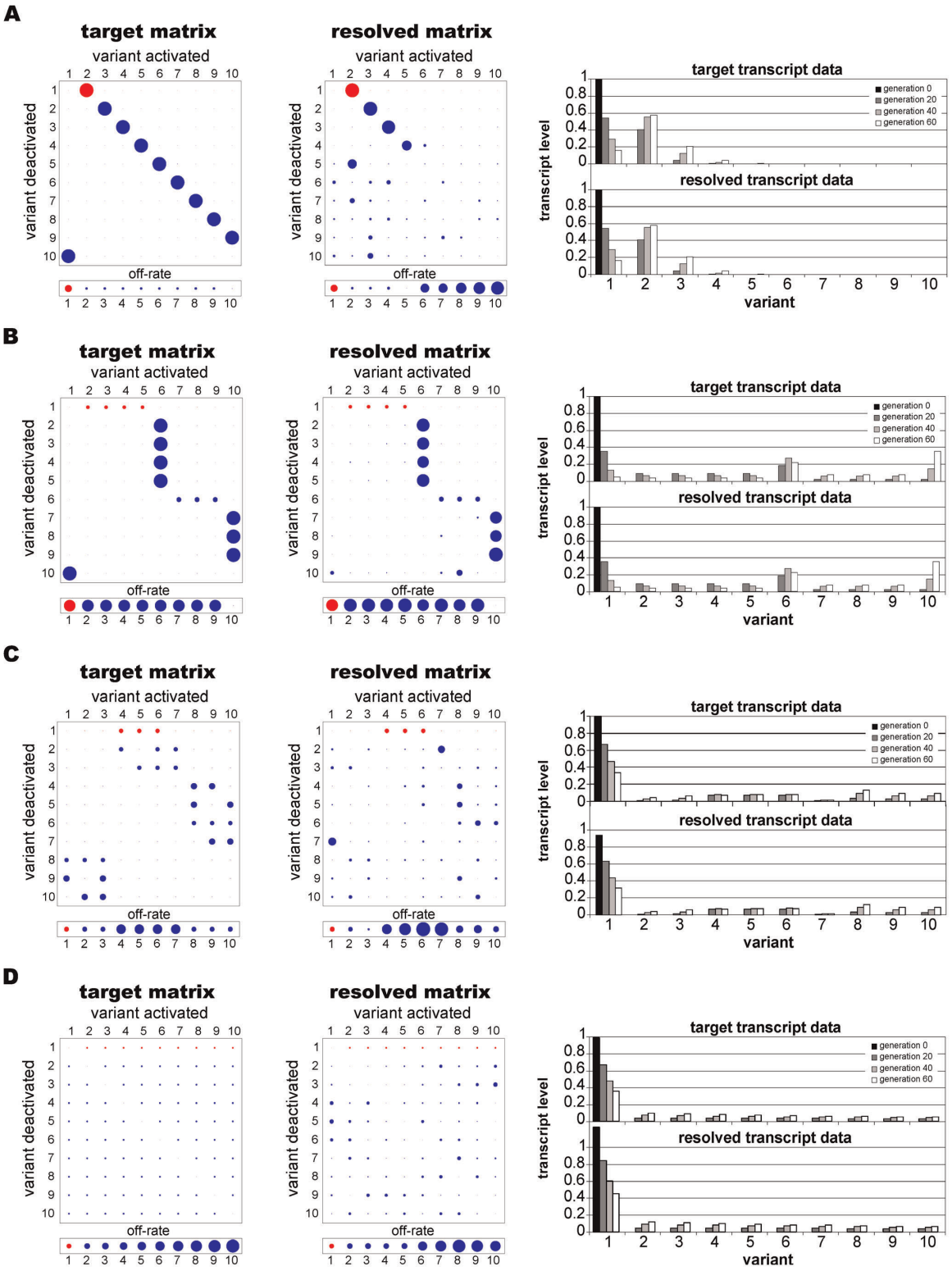

A
target matrix
variant activated
variant deactivated
off-rate
resolved matrix
variant activated
variant deactivated
off-rate
target transcript data
generation 0
generation 20
generation 40
generation 60
resolved transcript data
transcript level
variant
B
target matrix
variant activated
variant deactivated
off-rate
resolved matrix
variant activated
variant deactivated
off-rate
target transcript data
generation 0
generation 20
generation 40
generation 60
resolved transcript data
transcript level
variant
C
target matrix
variant activated
variant deactivated
off-rate
resolved matrix
variant activated
variant deactivated
off-rate
target transcript data
generation 0
generation 20
generation 40
generation 60
resolved transcript data
transcript level
variant
D
target matrix
variant activated
variant deactivated
off-rate
resolved matrix
variant activated
variant deactivated
off-rate
target transcript data
generation 0
generation 20
generation 40
generation 60
resolved transcript data
transcript level
variant

Figure 2. Resolved networks from a single time series. Using simulated annealing the (**A**) 1:1, (**B**) SMS and (**D**) uniform networks are resolved to up to four consecutive switches, although the complex lattice network (**C**) is less well described. The starter genes, which were assumed to be clonally transcribed at generation 0, are highlighted in red. The simulated time courses (right column) show a very good agreement between the target data and the transcription data resulting from the determined network.
doi:10.1371/journal.pone.0039335.g002

Moreover, off-rates for non-starter genes were estimated with a good degree of accuracy and precision, except those for the uniform network which were consistently overestimated. Although estimates of the switch rates of non-starter genes were less precise they were still informative. As is evident from comparing the MCMC output in Figure 6 to the output from simulated annealing (right column of Figure 4), the MCMC method provides a much more reliable indication of the true switching structure. However, the simulated annealing algorithm is still required to estimate the level of noise in the data used to calibrate the MCMC method.

Dimension reduction. We initially argued that using data from only the most dominant genes within the repertoire is justifiable as long as the majority of gene transcripts remain at very low levels throughout the experiment. To verify this assumption we created 60-dimensional versions of our four test networks. Starting genes were selected by first simulating transcription histories for many generations until the transcript levels were close to equilibrium and thus resembled a parent population kept in long-term culture. Eight distinct starting genes were then randomly selected, such that the probability of selecting each gene was proportional to its level of transcription in the parent culture, thus mimicking experimental selection of starting genes by limiting dilution.

For each network we applied the MCMC algorithm using transcription data for only the 16 most transcribed genes. To perform the reduction, genes were ranked by their average transcription levels across all time points and all cultures in the data generated by the 60-dimensional matrix (after adding noise). Data for the 16 most highly ranked genes were then selected, renormalised, and used as input. Note that genes chosen in this way may come from diverse regions of the switching network, so that the reduced matrix, which represents disjoint network elements, may look different from the original, full matrix. For example, the reduced SMS and 1:1 matrices (Figure 7A and 7B) are somewhat similar. Later we show how hypothesis testing may be used to determine the most likely network type, even when the MCMC output appears ambiguous.

For each of our reduced test networks, the MCMC parameter estimates gave a good indication of the network structure (Figure 7 and S2). Where estimates varied from the true parameter values, this was often because genes included in the reduced system took on roles in the network vacated by omitted genes, so that the general network structure was preserved. These results confirm that accurate estimates of the overall switch pathways can be obtained by using only a subset of highly transcribed genes.

Application

As demonstrated in the previous section, our method can reliably determine genetic switch pathways from relatively limited data; however, it also showed how parameter estimates can be affected by experimental noise or sparseness of the available data to be fitted. And although the MCMC approach provides a good indication of the uncertainty in parameter values and thus the underlying switching pathway, the question remains how to proceed if the method results in ambiguous outputs. Here we provide one approach in terms of hypothesis testing, which allows the direct comparison of the likelihoods of alternative switch pathways. Finally we apply our method to real *var* gene transcription data previously described in [15].

Hypothesis testing. A strength of our likelihood-based approach is that it allows testing of hypotheses regarding the network structure where the predicted switch pathway is not immediately obvious or where there is a high degree of uncertainty around the estimated parameter values comprising this pathway. That is, if the outcome of the MCMC method is ambiguous in terms of the most likely switching network underlying the observed change in transcription levels, we can use this approach to specifically test different hypotheses. To demonstrate how this technique might be useful in practice we used likelihood ratio tests to compare the power of two alternative switching models to explain the transcription data generated by the networks considered in the previous section (1:1, SMS, lattice and uniform networks). In the first model, all genes were assumed to have identical, uniform switch biases, corresponding to a situation with completely unbiased switching; the second model allowed all genes to have different sets of switch biases, as in the structured switch pathways considered previously. We then used simulated annealing to find the maximum likelihoods of each model using the noisy, dimension-reduced data sets for each network.

For the 1:1, SMS and lattice networks, the likelihood ratios correctly indicated that, in each case, the switch biases were unlikely to be uniform ($\chi^2_{224} \geq 2{,}350, p < 0.0001$), implying local switch biases and thus rejecting the hypothesis that a change in transcript levels would simply be due to differences in off-rates. Conversely, for the uniform network the test outcome was consistent with uniform switch biases ($\chi^2_{224} = 119, p = 1.00$), thus providing no support for the notion of variant-specific switch biases in this case. It is also possible to test for differences between the parameters of specific genes. For example, likelihood ratio tests indicated that in the SMS matrix, the switch biases of gene 10 were likely to differ from those of gene 12 ($\chi^2_{13} = 282,\ p < 0.0001$) but not from those of gene 9 ($\chi^2_{13} = 10.5$, $p = 0.65$). The latter result implies, correctly, that the network was unlikely to have a 1:1 structure.

Results of F-tests were similar to those of the likelihood ratio tests, and other model selection statistics such as Bayes factors or information criteria are also compatible with our method. Thus, given sufficient experimental data, it should be possible to determine with a high degree of confidence whether a hypothesised switching pattern is likely to be correct.

Experimental data. Finally, we applied our methods to experimental data, comprising a subset of the *P. falciparum var* gene transcription measurements of three different clones (3D7_AS2, IT4_2B2 and IT4_2F6) previously analysed by Recker *et al.* [15]. In agreement with the proposed switch pathways, for each data set the simulated annealing algorithm found that an SMS (single-many-single) network structure gave the best fit to the data. Because of the sparsity of those particular data sets we used the MCMC-based method to assess the uncertainty in the parameter estimates. Although the results confirmed that the switching network was unlikely to have a 1:1 or uniform structure, the output also revealed that other parameter sets representing different network structures were equally likely, as indicated by the blurry matrix entries in Figure 8; unfortunately the data were insufficient to carry out the more detailed hypothesis testing outlined above. Note, we obtained similar results when we used a single, artificial data set generated by a SMS matrix (Figure S3), which strongly

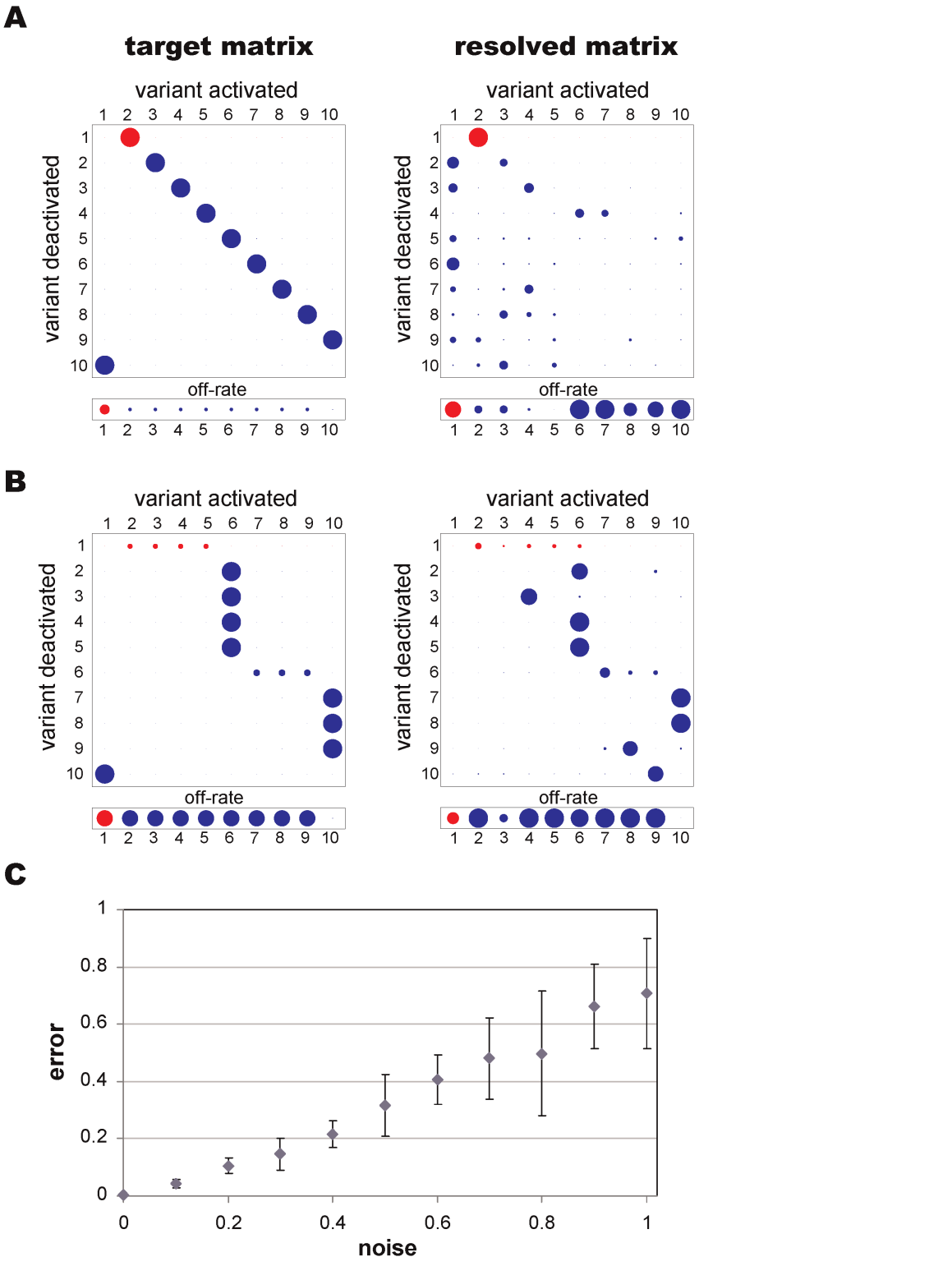

A
target matrix
resolved matrix
variant activated
variant deactivated
off-rate
1 2 3 4 5 6 7 8 9 10
B
variant activated
variant deactivated
off-rate
1 2 3 4 5 6 7 8 9 10
C
error
noise
0 0.2 0.4 0.6 0.8 1

Figure 3. Resolved networks from noisy data. The accuracy in determining switch networks from single transcription time courses is significantly affected by the level of noise in the data (here $\sigma = 0.6$) and can lead to poorly resolved networks, as shown for both the (**A**) 1:1 and (**B**) SMS pathways. (**C**) The error shows a linear correlation with the level of noise.
doi:10.1371/journal.pone.0039335.g003

indicates that more data are required for determining the precise pattern of antigenic switching among *var* genes.

Discussion

Here we described a statistically rigorous method for determining genetic switch pathways from quantitative gene transcription data. The problem in determining switch pathways with direct methods is the wealth of data required to estimate gene intrinsic parameters, such as activation biases and off-rates, especially when the system under consideration consists of more than just a couple of genes. Most antigenically variable organisms have gene repertoires that are orders of magnitude larger, however. We have demonstrated that our iterative method is able to resolve complex switch patterns from minimal data to a high degree of accuracy, even when allowing for a reasonable degree of experimental error or noise. Importantly, we have shown that using multiple sets of data simultaneously increases accuracy such that even highly complex networks can be determined. This should be of particular interest as it suggests that the most reliable information about a gene's or gene repertoire's switch behaviour can be extracted not necessarily from long-term cultures but from analysing transcription profiles which are relatively short but obtained from different clones of the same isolate. In general, our results suggest the following strategy for determining genetic switching networks in three stages. The first step would be to estimate the amount of noise in the data, using simulated annealing. This would be followed by employing our MCMC method – calibrated to the estimated noise level – to return reliable parameter distributions for the parameters comprising the underlying network, i.e. the genes' off-rates and switch biases. Lastly, and strongly dependent on the amount and quality of the available data, one would test hypothesised network structures by comparing the likelihoods of restricted models.

For structured networks we noticed that some variant transcripts remain at very low levels over the whole transcription time courses considered here. This is consistent with experimental studies of longitudinal *var* gene transcription, which also found that some variants are rarely activated, even after long-term *in vitro* culture. Although we could show that data for only the most dominantly transcribed genes are sufficient for determining the overall switch pattern, uncertainty in parameter estimates for genes with low activation probabilities remains a problem for describing the switch behaviour of an entire gene repertoire. However, once identified it should be possible to specifically select these genes for cloning and then apply the method described here. Knowledge about genes with low activation probabilities *in vitro* could yield important information about *in vivo* selection processes and should help to distinguish between parasite-intrinsic switching and immune- or receptor-mediated selection underlying the within-host infection dynamics of antigenically variable organisms.

Investigating intrinsic switch patterns is of major importance for a number of reasons. First of all, the non-random nature at which genes are activated or silenced is likely the result of an evolutionary process shaped by the interaction between the parasite and the immune system. Understanding this pattern would therefore provide valuable insights into the selective pressure acting upon the parasite during infections, which in turn should also shed light on the evolution and structure of the gene repertoire itself. For *P. falciparum* at least, there also is the added complexity of phenotypic as well as antigenic variation. Different PfEMP1 variants mediate cytoadherence of infected red blood cells to different host tissues [16,17] and expression of certain subsets of the *var* repertoire has been associated with disease severity and young host age [18–20]. A full characterisation of the antigenic repertoire in terms of gene activation rates and hierarchies will therefore help to explain age-related pathologies of malaria infections and the observed order of acquisition of protective immunity against certain subsets of genes. It is also hoped that analysing pre-determined switch patterns will contribute to our understanding of the molecular mechanisms behind antigenic variation. For example, differentiation of switch biases between particular genes and "universal" activation preferences should help elucidate how antigenic switching is controlled at the genetic and/or epigenetic level.

Although our method was described with reference to a particular *in vitro* experimental set-up in which *P. falciparum* parasites are followed from an initially clonal population over time, there is no restriction on how the data is generated. In fact, provided that adequate temporal gene transcription data can be obtained, either from *in vitro* culture or *in vivo* infection dynamics, this method can be used to reliably determine the parameters that define the switching network between genes in any antigenically variable organism.

Methods

To analyse antigenic switch pathways from gene transcription data we follow the approach taken by Recker *et al.* [15]. We assume that clonal parasite populations are followed over an extended period of *in vitro* culture with relative transcript levels of all genes measured at various time points. The resulting transcription profiles can then be described by the following time-discrete model:

$$v_{i,c}^{t+1} = (1-\omega_i)v_{i,c}^{t} + \sum_{j \neq i} \omega_j \beta_{ji} v_{j,c}^{t}, \forall i \in \{1, \ldots, n\},$$

where $v_{i,c}^{t}$ is the relative transcript level of variant i in culture c at time t, ω_i is the variant specific off-rate, and β_{ji} is the probability of a switch from variant j to variant i. The task is then to find the parameter values for $\boldsymbol{\omega} = \omega_i$ and $\boldsymbol{\beta} = \beta_{ij}$ that minimise an error function defined by the deviation between the measured – or in our case simulated – transcript level $v_{i,c}^{t}$ and the model outcome $m_{i,c}^{t}$:

$$E = \sqrt{\frac{1}{qpn} \sum_{c=1}^{q} \sum_{t=1}^{p} \sum_{i=1}^{n} \left(\log_2 (v_{i,c}^{t}) - \log_2 (m_{i,c}^{t}) \right)^2},$$

where p is the number of time points where transcript levels were measured and q is the number of cultures.

We assumed that off-rates do not exceed 6% per generation, which is consistent with experimental measurements [9,10,12–14,21].

Simulated Annealing

Because of the large dimension of the system we used iterative approaches to find solutions that minimised the deviation between

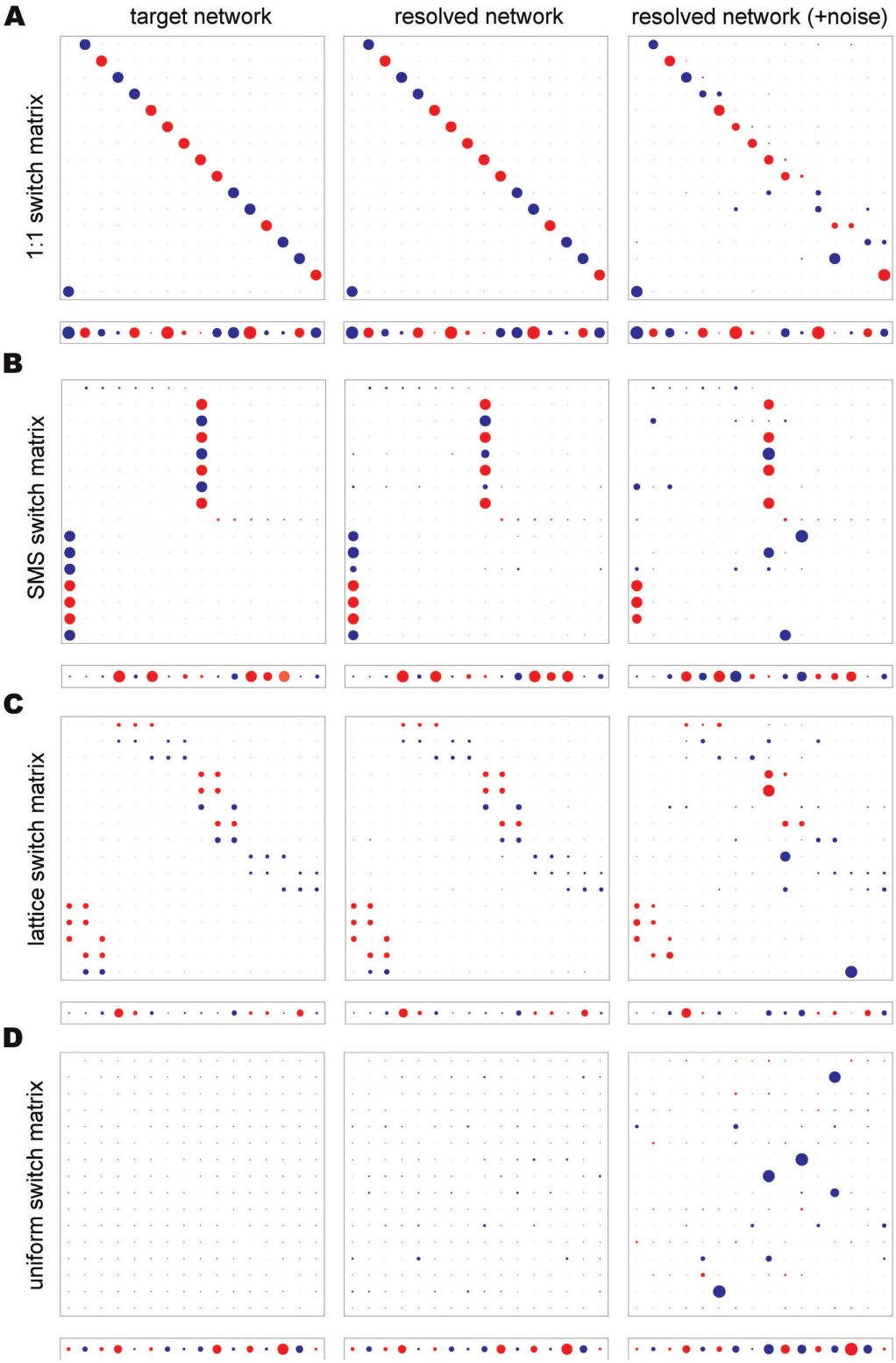

A
target network
resolved network
resolved network (+noise)
1:1 switch matrix
B
SMS switch matrix
C
lattice switch matrix
D
uniform switch matrix

Figure 4. Resolved networks from multiple time series. Transcription histories from eight different clones, each defined by a different starting gene (highlighted in red), were used to resolve four different switch pathways describing 1:1, SMS, lattice and uniform networks (**A–D**, respectively). Without any noise, all matrices and off-rate vectors can be resolved to a high degree of accuracy, even for the non-starter genes (middle column). The use of multiple data sets also yields better estimates when significant levels of noise are added to the data ($\sigma = 1$, right column).
doi:10.1371/journal.pone.0039335.g004

the model output and the data. Simulated annealing is a probabilistic optimisation algorithm which can move between locally optimal parameter sets [22]. The procedure starts with a random parameter set $B=(\beta,\omega)$ containing the switch biases and off-rates, and applies a random perturbation to produce B'. The move from B to B' is accepted if it reduces the error; otherwise the move is accepted with a probability dependent on how much it increases the error and on a "temperature" parameter. The perturbation and acceptance steps are repeated for many iterations while the temperature is gradually lowered so that moves that increase the error become less likely to be accepted. When the temperature becomes sufficiently low the parameter set converges to a local minimum.

Perturbation of a switch bias parameter β_{ji} was achieved by drawing a random variable β'_{ji} from a Cauchy distribution centred at β_{ji}. The distribution was restricted to the permitted range [0,1] by mapping

$$B_i \mapsto \begin{cases} \mathrm{frac}(\beta_{ji}) & \text{if } \mathrm{floor}(\beta_{ji}) \text{is even} \\ 1-\mathrm{frac}(\beta_{ji}) & \text{if } \mathrm{floor}(\beta_{ji}) \text{is odd,} \end{cases}$$

where $\mathrm{floor}(\beta_{ji})$ is the largest integer not greater than β_{ji} and $\mathrm{frac}(\beta_{ji})$ is the fractional part. Off-rates were similarly perturbed after rescaling in proportion to the parameter range [0,0.06]. In each iteration the perturbation procedure was applied to one switch bias per matrix row and the other entries were then rescaled to normalise the row. All off-rates were perturbed in each iteration. We chose a Cauchy distribution because it has a fatter tail than the Gaussian distribution and may therefore be more likely to generate "basin hopping" moves between local optima.

This perturbation procedure is symmetrical, meaning that the probability of moving from B to B' is always equal to the probability of moving from B' to B. However, normalising each row of switch biases caused the algorithm disproportionately to favour small parameter values. Such non-uniformity is permitted in simulated annealing and should not affect the results. The perturbation size was determined by the scale parameter of the Cauchy distribution, which was made constant.

For the acceptance threshold we chose the Fermi distribution; that is, proposed moves were accepted with probability $1/(1+\exp(\frac{E'-E}{T}))$, where E and E' are the errors associated with the current and proposed parameter sets respectively, and T is the temperature. We used a geometric cooling schedule and stopped the algorithm after 10 million iterations.

Experimental Error

Gene transcript levels are calculated from qrt-PCR output using an exponential formula in which the base is the amplification efficiency of the gene transcript (which should be approximately 2) and the exponent is the number of amplification cycles required for the transcript to reach a threshold abundance [23]. We therefore assumed that error-prone measurements would follow a log-normal distribution and that, at least for the most abundant transcripts, the standard deviation would be proportional to the mean. Accordingly, measurement errors were simulated by applying a noise function $\phi(v_{i,c,t})=2^R v_{i,c,t}$, with $R \sim N(0,\sigma^2)$ to the simulated transcription profiles and then renormalising each

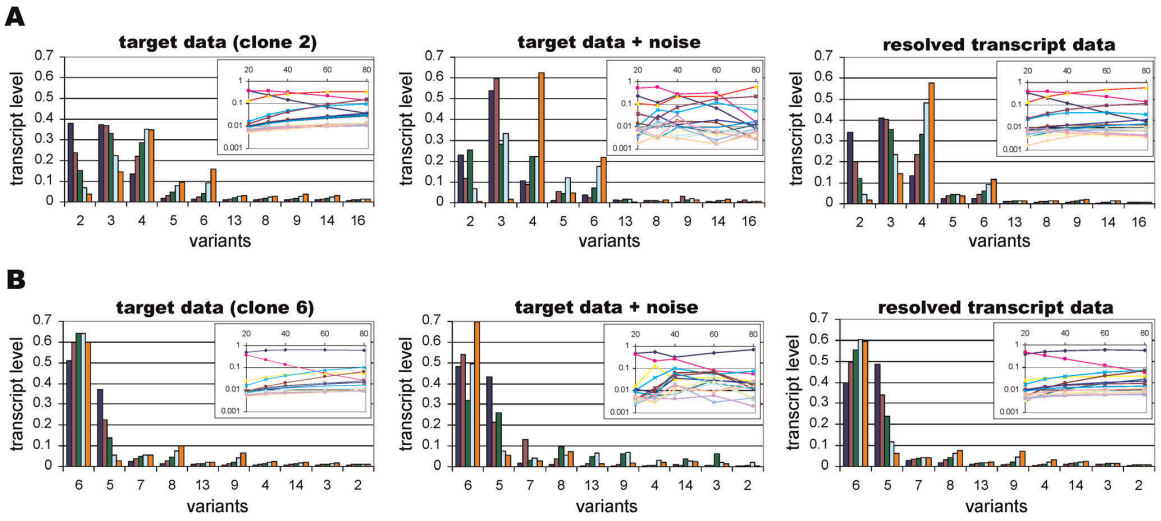


Figure 5. Resolved transcription time courses from noisy data. When the model is fitted to multiple data sets with different starting genes, the transcript levels of the predicted network (right column) are more similar to the noiseless data of the underlying network (left column) than to the noisy data used as input (middle column). Results are shown for two different clones of a 1:1 switch network with transcript levels measured at generation 20 (blue), 30 (purple), 40 (green), 60 (light blue) and 80 (orange). The insets depict the proportional transcript levels of the 16 variants on a log scale at generations 20, 30, 40, 60 and 80, with each colour representing a different variant.
doi:10.1371/journal.pone.0039335.g005

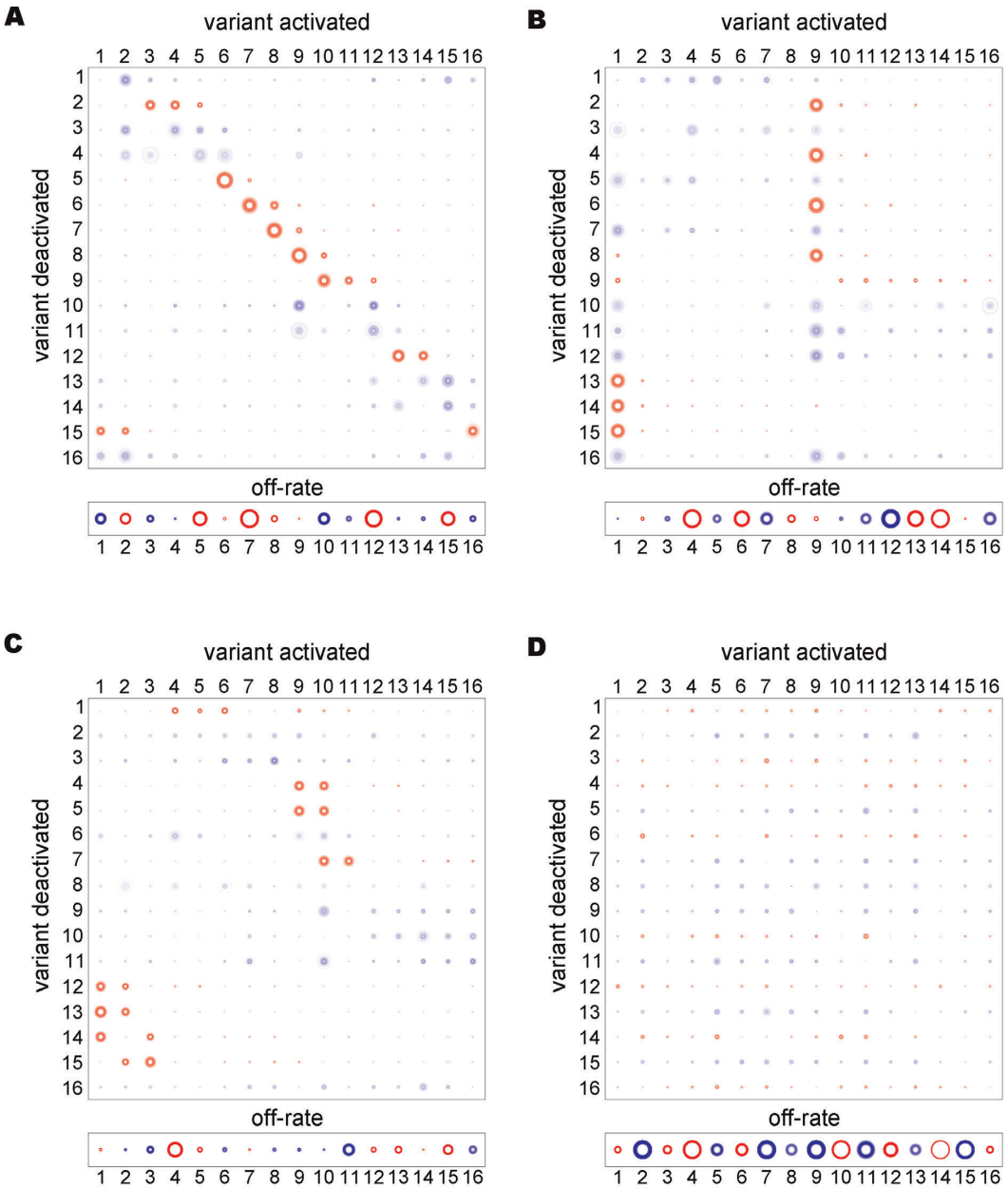


Figure 6. Using MCMC on noisy transcription data. In the MCMC output for the four different switch pathways: 1:1, SMS, lattice and uniform, (**A–D**, respectively), the parameter range for each switch bias and off-rate was divided into bins, represented by rings. If a large proportion of recorded solutions contained similar values for a parameter then the corresponding ring is coloured dark, indicating a high likelihood that the true parameter value lies within that range. The proportions were measured relative to a null distribution H_0, which assumed that all solutions were equally likely to be accepted. The darkest colour corresponds to proportions that differed from the H_0 mean by at least 25 standard deviations.
doi:10.1371/journal.pone.0039335.g006

set $\{v_{1,c,t},\ldots,v_{n,c,t}\}$. In most of our trials $\sigma=1$, which assumes that transcript levels are typically measured to be two times larger or smaller than their true values.

MCMC

To obtain distributions of likely values for the switch biases and off-rates, instead of only the "best-fit" point estimates, we used a Markov chain Monte Carlo (MCMC) method known as the Metropolis-Hastings algorithm [24,25]. This algorithm accepts

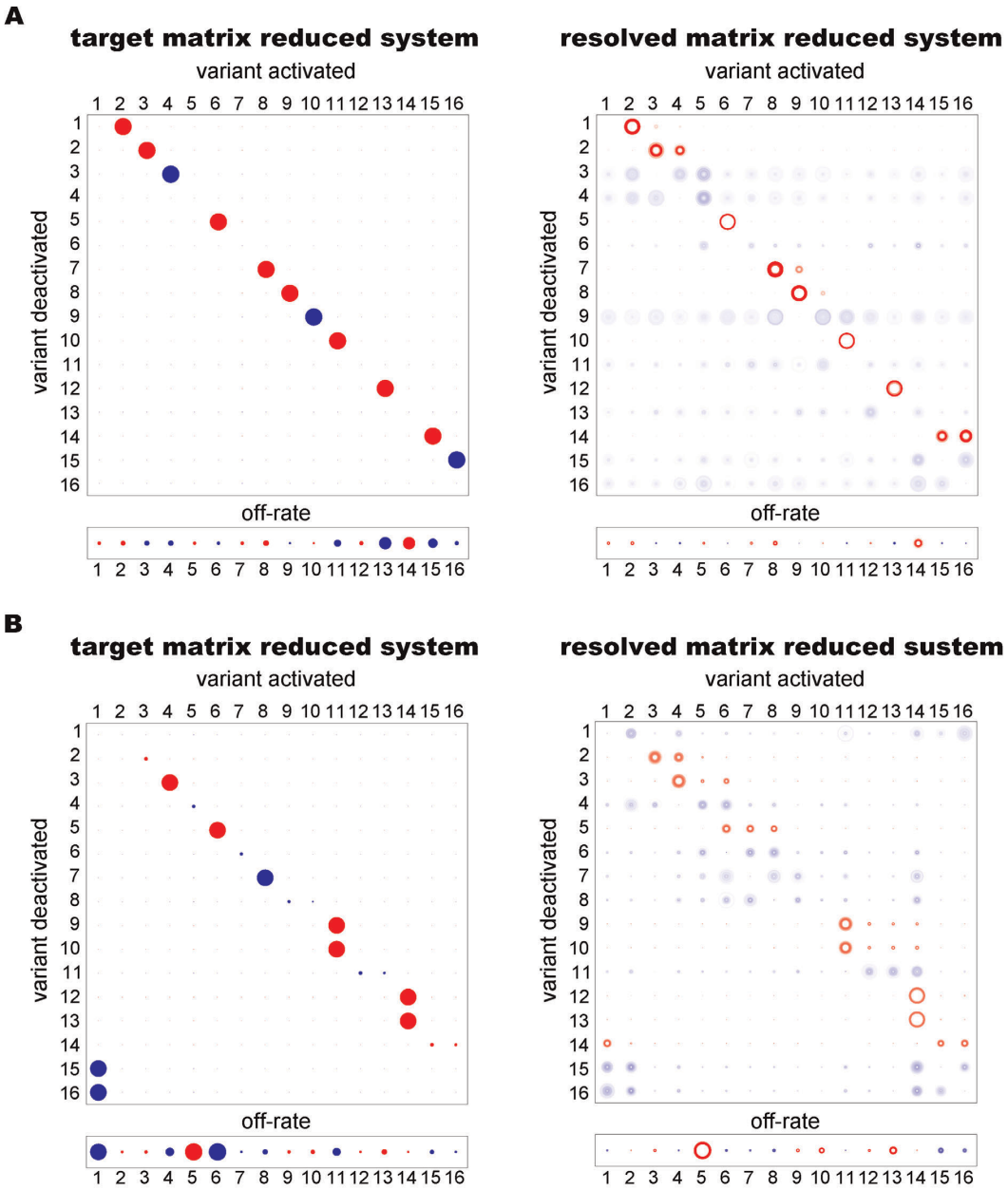


Figure 7. Results following dimension reduction. Left column: target parameters for the (**A**) 1:1 and (**B**) SMS networks reduced from 60 to 16 genes. Right column: MCMC output after adding noise with $\sigma = 1$. To perform the reduction, genes were ranked by their average transcription levels across all time points and all cultures in the data generated by the 60-dimensional matrix (after adding noise). The 16 most highly ranked genes were then selected and their data renormalised. The starter gene parameters are shown in red.
doi:10.1371/journal.pone.0039335.g007

proposed moves with probability $L(B'|v)/L(B|v)$, which is the likelihood ratio of the proposed parameter set B' and the current parameter set B, given the observed data v. To convert the simulated annealing algorithm to instead perform MCMC we replaced the acceptance threshold with the exact likelihood ratio, as derived below, and used the simulated annealing error to estimate the noise parameter σ. The algorithm was run for 10 million iterations and every accepted solution was recorded after an initial "burn-in" period of 1 million iterations.

To simplify the notation, the following derivation considers the case with only one culture and one time point, so $v = \{v_1, \ldots, v_n\}$ is the observed data and $m = \{m_1, \ldots, m_n\}$ is the output of the model

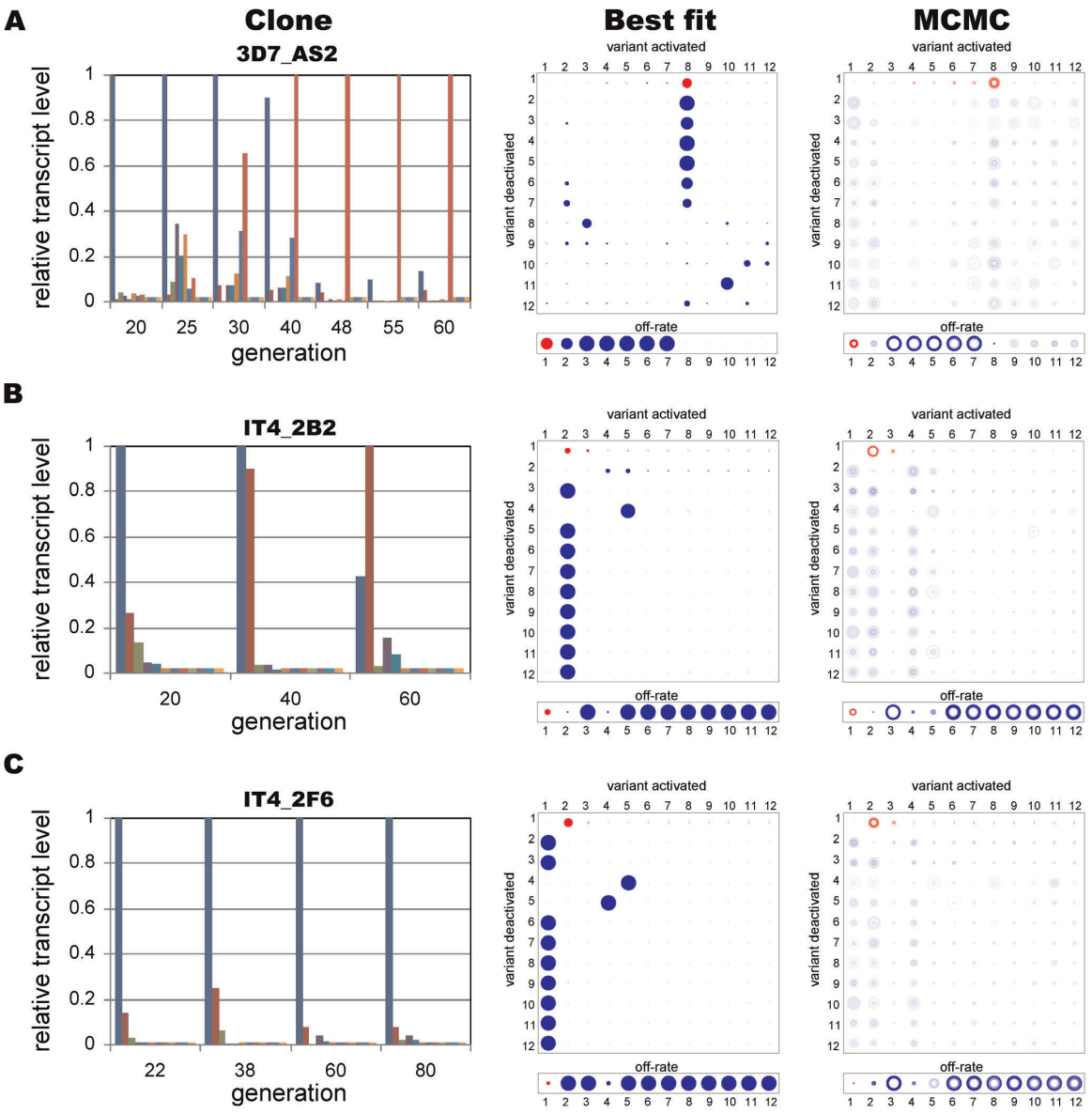


Figure 8. Results from experimental data. Best-fit parameter estimates derived by simulated annealing (top row) and MCMC parameter distributions (bottom row) are shown for three sets of *P. falciparum var* gene transcription data previously analysed by Recker *et al.* [15]. Each data set comprises a single time series of measurements from an initially clonal culture. The results are consistent with an SMS network structure although the MCMC output (right column) also indicates the likelihood of alternative pathways.
doi:10.1371/journal.pone.0039335.g008

with parameter set B. Assuming as before that each $v_i = 2^{r_i} m_i / \sum_{j=1}^{n} 2^{r_j} m_j$, where each r_i is an instance of a random variable $R_i \sim N(0,\sigma^2)$, then

$$r_i - r_1 = \log_2 (v_i m_1/(m_i v_1)),\ \forall i > 1$$

and

$$L(B|v) = P(V = v|B)$$

$$= \int_{-\infty}^{\infty} P(R_1 = x) \prod_{i=1}^{n-1} P(R_i - R_1 = r_i - r_1 | R_1 = x)\, \mathrm{d}x$$

$$= \int_{-\infty}^{\infty} P(R_1 = x) \prod_{i=1}^{n-1} P\left(R_i = \log_2\left(\frac{v_i m_1}{v_1 m_i}\right) + x\right) \mathrm{d}x$$

$$\propto \int_{-\infty}^{\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right) \prod_{i=1}^{n-1} \exp\left(\frac{-(\log_2(v_i m_1/(m_i v_1)) + x)^2}{2\sigma^2}\right) \mathrm{d}x.$$

Changing the variable to $y = 2^x m_1/v_1$ gives

$$L(B|v) = \int_0^{\infty} \frac{1}{y} \prod_{i=1}^{n} \exp\left(\frac{-\log_2^2(y v_i/m_i)}{2\sigma^2}\right) \mathrm{d}y$$

$$\propto \exp\left[\frac{1}{2\sigma^2}\left(-\sum_{i=1}^{n} \log_2^2\left(\frac{v_i}{m_i}\right) + \frac{1}{n}\log_2^2\left(\prod_{i=1}^{n} \frac{v_i}{m_i}\right)\right)\right].$$

It follows that the likelihood ratio for q cultures and p time points is

$$\frac{L(B'|v)}{L(B|v)} = \prod_{c=1}^{q} \prod_{t=1}^{p} \exp\left[\frac{1}{2\sigma^2}\left(-\sum_{i=1}^{n} \log_2^2\left(\frac{v_{i,c,t}}{m'_{i,c,t}}\right) + \sum_{i=1}^{n} \log_2^2\left(\frac{v_{i,c,t}}{m_{i,c,t}}\right)\right.\right.$$

$$\left.\left. + \frac{1}{n}\log_2^2\left(\prod_{i=1}^{n} \frac{v_{i,c,t}}{m'_{i,c,t}}\right) - \frac{1}{n}\log_2^2\left(\prod_{i=1}^{n} \frac{v_{i,c,t}}{m_{i,c,t}}\right)\right)\right]$$

$$= \exp\left[\frac{npq}{2\sigma^2}\left(\underset{\left(\log_2\left(\frac{v}{m}\right)\right)}{\mathrm{Var}} - \underset{\left(\log_2\left(\frac{v}{m'}\right)\right)}{\mathrm{Var}}\right)\right].$$

Supporting Information

Figure S1 Resolved networks from noisy data. The accuracy in determining switch networks from single transcription time courses is significantly affected by the level of noise in the data (here $\sigma = 0.6$) and can lead to poorly resolved networks, as shown for both the (**A**) lattice and (**B**) uniform pathways.
(TIFF)

Figure S2 Results following dimension reduction. Left column: target parameters for the (**A**) lattice and (**B**) uniform networks reduced from 60 to 16 genes. Right column: MCMC output after adding noise with $\sigma = 1$. To perform the reduction, genes were ranked by their average transcription levels across all time points and all cultures in the data generated by the 60-dimensional matrix (after adding noise). The 16 most highly ranked genes were then selected and their data renormalised. The starter gene parameters are shown in red.
(TIFF)

Figure S3 Results from simulating hypothesised *var* gene system. In this case the 60-gene SMS network was reduced to 12 genes by selecting only one starter gene, instead of the eight starter genes used in Figure 7. Simulated transcription levels were recorded at only three time points and a high level of noise was added to the data. These conditions mimic those of the experimental data sets IT4_2B2 (three time points recorded) and IT4_2F6 (four time points) analysed by Recker *et al.* [15] and the results are similar to those found by analysing the original data (Figure 8). The best-fit parameter estimates suggest an SMS network but the MCMC output shows that this result is uncertain because of the relatively small data set.
(TIFF)

Acknowledgments

We would like to thank Richard Hills and Warren Kretzschmar for helpful discussions.

Author Contributions

Conceived and designed the experiments: RN MR. Performed the experiments: RN. Analyzed the data: RN. Wrote the paper: RN MR.

References

1. Deitsch KW, Lukehart SA, Stringer JR (2009) Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. Nature Reviews Microbiology 7: 493–503.
2. Kosinski RJ (1980) Antigenic variation in trypanosomes - a computer-analysis of variant order. Parasitology 80: 343–357.
3. Agur Z, Abiri D, Vanderploeg LHT (1989) Ordered appearance of antigenic variants of African trypanosomes explained in a mathematical-model based on a stochastic switch process and immuneselection against putative switch intermediates. Proceedings of the National Academy of Sciences of the United States of America 86: 9626–9630.
4. Frank SA (1999) A model for the sequential dominance of antigenic variants in african trypanosome infections. Proceedings of the Royal Society of London Series B-Biological Sciences 266: 1397–1401.
5. Molineaux L, Diebner HH, Eichner M, Collins WE, Jeffery GM, et al. (2001) Plasmodium falciparum parasitaemia described by a new mathematical model. Parasitology 122: 379–391.
6. Paget-McNicol S, Gatton M, Hastings I, Saul A (2002) The plasmodium falciparum var gene switching rate, switching mechanism and patterns of parasite recrudescence described by mathematical modelling. Parasitology 124: 225–235.
7. Lythgoe KA, Morrison LJ, Read AF, Barry JD (2007) Parasite-intrinsic factors can explain ordered progression of trypanosome antigenic variation. Proceedings of the National Academy of Sciences of the United States of America 104: 8095–8100.
8. Recker M, Nee S, Bull PC, Kinyanjui S, Marsh K, et al. (2004) Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. Nature 429: 555–558.
9. Horrocks P, Pinches R, Christodoulou Z, Kyes SA, Newbold CI (2004) Variable var transition rates underlie antigenic variation in malaria. Proceedings of the National Academy of Sciences of the United States of America 101: 11129–11134.
10. Frank M, Dzikowski R, Amulic B, Deitsch K (2007) Variable switching rates of malaria virulence genes are associated with chromosomal position. Molecular microbiology 64: 1486–1498.
11. Enderes C, Kombila D, Dal-Bianco M, Dzikowski R, Kremsner P, et al. (2011) Var gene promoter activation in clonal plasmodium falciparum isolates follows a hierarchy and suggests a conserved switching program that is independent of genetic background. Journal of Infectious Diseases 204: 1620–1631.
12. Morrison LJ, Majiwa P, Read AF, Barry JD (2005) Probabilistic order in antigenic variation of trypanosoma brucei. International journal for parasitology 35: 961–972.
13. Aitcheson N, Talbot S, Shapiro J, Hughes K, Adkin C, et al. (2005) Vsg switching in trypanosome brucei: antigenic variation analysed using rnai in the absence of immune selection. Molecular microbiology 57: 1608–1622.
14. Barbour AG, Dai Q, Restrepo BI, Stoenner HG, Frank SA (2006) Pathogen escape from host immunity by a genome program for antigenic variation. Proceedings of the National Academy of Sciences of the United States of America 103: 18290–18295.
15. Recker M, Buckee CO, Serazin A, Kyes S, Pinches R, et al. (2011) Antigenic variation in plasmodium falciparum malaria involves a highly structured switching pattern. Plos Pathogens 7: e1001306.
16. Robinson BA, Welch TL, Smith JD (2003) Widespread functional specialization of plasmodium falciparum erythrocyte membrane protein 1 family members to bind cd36 analysed across a parasite genome. Molecular microbiology 47: 1265–1278.

17. Howell DPG, Levin EA, Springer AL, Kraemer SM, Phippard DJ, et al. (2008) Mapping a common interaction site used by plasmodium falciparum duffy binding-like domains to bind diverse host receptors. Molecular microbiology 67: 78–87.
18. Warimwe GM, Keane TM, Fegan G, Musyoki JN, Newton CRJC, et al. (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: 21801–21806.
19. Cham GKK, Turner L, Lusingu J, Vestergaard L, Mmbando BP, et al. (2009) Sequential, ordered acquisition of antibodies to plasmodium falciparum erythrocyte membrane protein 1 domains. Journal of Immunology 183: 3356–3363.
20. Cham GKK, Turner L, Kurtis JD, Mutabingwa T, Fried M, et al. (2010) Hierarchical, domain type-specific acquisition of antibodies to plasmodium falciparum erythrocyte membrane protein 1 in tanzanian children. Infection and immunity 78: 4653–4659.
21. Roberts DJ, Craig AG, Berendt AR, Pinches R, Nash G, et al. (1992) Rapid switching to multiple antigenic and adhesive phenotypes in malaria. Nature 357: 689–692.
22. Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. Science 220: 671–680.
23. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time rt-pcr. Nucleic acids research 29: e45.
24. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. Journal of Chemical Physics 21: 1087–1092.
25. Hastings WK (1970) Monte-carlo sampling methods using markov chains and their applications. Biometrika 57: 97–109.

eLIFE
elife.elifesciences.org

RESEARCH ARTICLE

The antigenic switching network of *Plasmodium falciparum* and its implications for the immuno-epidemiology of malaria

Robert Noble[1†], Zóe Christodoulou[2†], Sue Kyes[2], Robert Pinches[2], Chris I Newbold[2*], Mario Recker[1*]

[1]Department of Zoology, University of Oxford, Oxford, United Kingdom; [2]Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, United Kingdom

Abstract Antigenic variation in the human malaria parasite *Plasmodium falciparum* involves sequential and mutually exclusive expression of members of the *var* multi-gene family and appears to follow a non-random pattern. In this study, using a detailed in vitro gene transcription analysis of the culture-adapted HB3 strain of *P. falciparum*, we show that antigenic switching is governed by a global activation hierarchy favouring short and highly diverse genes in central chromosomal location. Longer and more conserved genes, which have previously been associated with severe infection in immunologically naive hosts, are rarely activated, however, implying an in vivo fitness advantage possibly through adhesion-dependent survival rates. We further show that a gene's activation rate is positively associated sequence diversity, which could offer important new insights into the evolution and maintenance of antigenic diversity in *P. falciparum* malaria.
DOI: 10.7554/eLife.01074.001

***For correspondence:** chris.newbold@imm.ox.ac.uk (CIN); mario.recker@zoo.ox.ac.uk (MR)

†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 16

Received: 11 June 2013
Accepted: 19 August 2013
Published: 17 September 2013

Reviewing editor: Mercedes Pascual, University of Michigan, United States

Copyright Noble et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Introduction

Acquired antibody-mediated immunity against *Plasmodium falciparum* is directed against the disease-causing, intra-erythrocytic life-stage of the parasite's life-cycle. Continual exposure to *P. falciparum* infection can lead to a form of semi-immunity, whereby protection against life-threatening disease is acquired after a few infections only (***Gupta et al., 1999***) whereas sterile and long-lasting immunity against infection is never achieved through natural exposure (***Greenwood et al., 1987***; ***Marsh, 1992***; ***Tran et al., 2013***). The process of natural acquired immunity therefore coincides with a transition from acute and severe infection in young children to asymptomatic carriage in older individuals (***Marsh and Snow, 1997***). At the core of this variable outcome of infection and poor development of immunity lies the major parasite virulence factor and variant antigen, *P. falciparum* erythrocyte membrane protein 1 (PfEMP1). Members of this family of proteins are expressed on the surface of infected red blood cells and are involved in the binding of parasitised cells to other host cells and tissues (for an overview see ***Kraemer and Smith, (2006)***). The conferred binding phenotypes are believed to enhance parasite survival by avoiding splenic clearance (***Boone and Watters, 1995***; ***Buffet et al., 2009***) and also contribute to malaria pathology through parasite sequestration in the deep vasculature (***Ockenhouse et al., 1991***; ***Pongponratn et al., 1991***). Mutually exclusive transcriptional switching between PfEMP1 variants during the course of an infection thus affords the parasite an immune evasion strategy and a means to exploit different host tissues (***Roberts et al., 1992***; ***Smith et al., 1995***).

Each haploid parasite genome contains approximately 60 members of the *var* gene family (***Gardner et al., 2002***) encoding distinct PfEMP1 variants, with a high degree of divergence between the antigenic repertoires of any two parasites. The high diversity of *var* genes and *var* gene repertoires is predominantly

eLife digest Our ability to acquire immunity to a disease depends on our immune system learning to recognise foreign molecules—called antigens—that are specific to the disease-causing virus, bacterium or parasite. However, some pathogens, such as the malaria-causing parasite *Plasmodium falciparum*, get around this defence through a process called antigenic variation. This involves the parasite switching between different antigens over the course of an infection, preventing the host immune system from learning to recognise them and leading to infections that last many weeks or even months.

The main antigen in *P. falciparum* is a protein called PfEMP1, which is encoded by a family of genes called *var* ('variable'). *Var* genes have evolved to be highly diverse, and different parasites have different repertoires of around 50–60 *var* genes. This ensures that there are a huge number of distinct variants of the PfEMP1 antigen available within the population, allowing the malaria parasite to maintain long-lasting infections and also to infect the same individuals again and again.

Previous work has shown that the expression of *var* genes is not random, but it is not clear what determines which genes are expressed at any given time. Now, Noble et al. have performed a detailed investigation of antigenic switching in *P. falciparum*. Using clonal parasites, they closely monitored the expression of the entire *var* gene repertoire during many generations of parasite culture. They observed that although different cultures initially expressed distinct *var* genes, most of them ended up expressing two particular genes—*var27* and *var29*—at high levels, indicating a hard-wired gene 'activation hierarchy'.

Noble et al. found that whenever the parasites switched antigens, *var* genes that were centrally located on chromosomes—such as *var27* and *var29*—were more likely to be activated than those at the ends of chromosomes. Moreover, *var* genes that were highly diverse were more likely to be activated than more conserved genes: this is the first evidence linking *var* gene evolution with gene activation probabilities. Together, these factors gave rise to the proposed activation hierarchy, which favours genes optimised for immune evasion and aids their continued evolution and diversification. Further work is now needed to identify the molecular mechanisms that control antigenic switching and to determine whether these could represent new therapeutic targets for malaria.

DOI: 10.7554/eLife.01074.002

generated and maintained by frequent recombination and gene conversion events (**Deitsch et al., 1999**; **Ward et al., 1999**; **Freitas-Junior et al., 2000**; **Taylor et al., 2000**; **Kraemer et al., 2007**; **Frank et al., 2008**), which has also resulted in a high degree of mosaicism within these genes (**Ward et al., 1999**; **Taylor et al., 2000**; **Bull et al., 2008**). The modular and highly polymorphic, extracellular portion of PfEMP1 is composed of a variable number of Duffy-binding-like domains (DBL; at least two and up to seven per gene) and cysteine-rich interdomain regions (CIDR; up to two per gene) (**Kraemer et al., 2007**). Despite the overall diversity of *var* genes at the sequence level, distinct associations with severe malaria have been identified for several such genes (**Salanti et al., 2003**; **Jensen et al., 2004**), with the involvement of *var2csa* in pregnancy-associated malaria being the clearest example to date (**Salanti et al., 2004**), as well as associations with particular DBL domains (**Warimwe et al., 2009**) and combinations of DBL and CIDR (**Avril et al., 2012**; **Claessens et al., 2012**; **Lavstsen et al., 2012**).

Var genes can also be classified according to a conserved upstream promoter sequence into four types termed UpsA, UpsB, UpsC, and UpsE (**Lavstsen et al., 2003**). Of particular interest are the UpsA-type genes, which appear to have diverged from the other groups (**Kraemer et al., 2007**; **Bull et al., 2008**) and have been shown to be upregulated during severe infections in young hosts (**Jensen et al., 2004**; **Bull et al., 2005**; **Kaestli et al., 2006**; **Kyriacou et al., 2006**; **Rottmann et al., 2006**; **Warimwe et al., 2009**). As a possible consequence, acquisition of anti-PfEMP1 immunity appears to follow a particular order, with UpsA variants generally being the first to be broadly recognised in older children (**Warimwe et al., 2009**; **Cham et al., 2010**). Nevertheless, clear associations between the expression of particular genes or Ups-groups, host age and severity of disease are still missing and could in fact be strain- or location-specific. Understanding the underlying pattern of antigenic switching between *var* genes is therefore important to explain not only the mechanisms and dynamics of persistent

infections but also the relationship between immune-mediated expression of particular PfEMP1 sub-types and infection outcomes.

Various studies have investigated *var* gene expression during natural or experimental infections (**Peters et al., 2002**; **Lavstsen et al., 2005**; **Wunderlich et al., 2005**; **Blomqvist et al., 2010**). However, as the diversity and order of PfEMP1 variants in the human host is influenced by immune responses and other host factors (**Kyriacou et al., 2006**; **Warimwe et al., 2009**), attempts to elucidate the underlying patterns of antigenic change and estimate related switching parameters are commonly based on in vitro cultured parasites in the absence of selection pressure. In this setting, Horrocks et al. (**Horrocks et al., 2004**) found that the rates at which *var* genes activate or deactivate are non-random, gene-specific and highly dissimilar. It has furthermore been suggested that *var* genes occupying subtelomeric loci tend to switch off faster than those positioned in central chromosomal regions (**Frank et al., 2007**), offering a possible explanation for the rapid decline in the transcription of subtelomeric genes in parasites during culture adaptation (**Peters et al., 2007**; **Zhang et al., 2011**). More recently, **Recker et al. (2011)** proposed that *var* gene transcriptional change involves a highly structured pathway that has evolved in response to a trade-off between the within-host and the between-host level fitness. However, most of these previous studies were limited in analytical depth due to restrictions in the number of variants considered and/or time points at which gene transcription levels were measured, and could therefore describe only small fragments of the whole switching network.

In this study, we used a novel approach to provide the first detailed characterisation of antigenic switching in the HB3 isolate of *P. falciparum*. We analysed pooled *var* gene transcription data from verified quantitative real time PCR measurements of a diverse set of clonal parasite cultures using a statistically rigorous method of parameter estimation. This revealed a global hierarchy in *var* gene activation favouring a highly diverse set of genes in central chromosomal locations. Our results further suggest a role of active gene transcription in the generation of antigenic diversity and will have important implications for understanding the age- and exposure-dependent pathology of *P. falciparum* malaria.

Results

11 subclones of the lab-adapted HB3 isolate of *P. falciparum* were obtained by serial dilution of the parent culture. The clones were cultivated and transcript levels of 38 *var* genes measured by quantitative PCR, first at 29–33 parasite generations (8–9 weeks) after cloning and then at several time points over the following 60–79 generations in each culture (**Table 1**); a detailed description can be found in the 'Materials and methods' section. These 38 genes comprised almost all known functional HB3 *var* genes (**Kraemer et al., 2007**; **Rowe et al., 2009**) and one partial gene. The HB3 parent culture was also observed at seven time points. At the first observation, each of the 11 clones predominantly transcribed one of the nine different *var* genes, which were most likely the genes transcribed at the time of cloning. Importantly, these nine genes—hereafter referred to as 'starter genes'—were representative of the *var* repertoire diversity in terms of chromosomal location and Ups type.

Var gene transcription dynamics

The proportional transcript levels of the different starter genes declined at notably different rates over the course of the experiment (**Figure 1**, **Figure 1—figure supplements 1 and 2**). For example, the centrally located, UpsC starter gene *var29* in culture #8 accounted for more than 90% of the total transcript level at the first measurement and remained the dominant variant over the following 60 generations. In contrast, the transcript level of the *var35* starter gene of culture #11 (subtelomeric, UpsA promoter type) was already below 50% of the total transcript level at the point of the first observation (approximately 30 generations post cloning) and continued to decline during the time course of culture, to be eventually replaced by other genes as dominant transcripts. Similar patterns were seen in cultures #3 and #7, which both had centrally located, UpsC starter genes, whereas the remaining five starter genes showed intermediate rates of decline.

Of particular interest was the apparent convergence of most cultures towards high transcription of *var27* and/or *var29* (**Figure 1**, **Figure 1—figure supplements 1 and 2**), which are both situated in central chromosomal location but have different promoter types (UpsB and UpsC, respectively). In light of the very different initial conditions, in terms of *var* gene transcription levels at which these cultures were established, this indicates that antigenic switching in *P. falciparum* might be governed by an activation hierarchy.

Table 1. Parasite culture and RNA sampling information

Culture	Starter gene	Observed generations
1	*var11*	29, 48, 59, 69, 89
2	*var13*	29, 48, 69, 89, 99, 108
3	*var13*	30, 49, 70, 90, 100, 109
4	*var5*	30, 39, 49, 60, 70, 90
5	*var27*	30, 49, 70, 90
6	*var27*	29, 38, 48, 69, 89
7	*var28*	33, 52, 63, 73, 93
8	*var29*	30, 49, 70, 80
9	*var30*	29, 48, 59, 69, 89
10	*var31*	29, 48, 59, 69, 89, 99, 108
11	*var35*	32, 51, 65, 72, 92
Parent	–	0, 9, 19, 40, 60, 70, 79

DOI: 10.7554/eLife.01074.003

To fully understand the patterns of antigenic switching underlying the observed transcriptional changes in our parasite cultures, we estimated relevant switch parameters using a previously described Markov chain Monte Carlo (MCMC) method (***Noble and Recker, 2012***) ('Materials and methods'). We have previously proposed that *var* gene switching can be fully described dynamically by two sets of parameters ('Materials and methods' and [***Recker et al., 2011***]). First, each gene *i* has an off-rate, ω_i, which is the per generation probability of switching from active to silenced. Second, each gene has a switch bias, β_{ij}, which is the probability that when gene *i* switches off, gene *j* becomes activated. We applied our algorithm to the full 38-gene data set and separately to a reduced set of the 16 most transcribed genes, and both models generally gave good and similar fits between the data and predicted transcript levels (***Figure 1—figure supplements 1 and 2***), with deviations resembling stochastic measurement errors.

Our method provided estimated switch parameters for the entire *var* repertoire, which are illustrated in ***Figure 2*** for the 16 most highly transcribed genes. Due to higher signal to noise ratios, parameter estimates of the starter genes (indicated in red) were more accurate and resulted in much narrower credible intervals, as indicated by less fuzzy rings that represent respective switch probabilities. As clearly demonstrated in ***Figure 2***, there was wide distribution in switch biases and off-rates, and we next analysed for specific associations between these parameters and other gene characteristics.

Off-rate estimates suggest a role of chromosomal location, not promoter type

Estimated off-rates ranged from 0.3% to more than 5% per generation, with a mean of approximately 2.9% (***Table 2***). Notably, the two most transcribed genes, *var29* and *var27* (both centrally located on chromosome 4), had exceptionally low off-rates of approximately 0.3% and 0.8%, respectively, whereas the other starter genes switched off at rates between 1.4% and 3.5% per generation. In general we found off-rates of centrally located genes to be significantly lower than of genes in subtelomeric regions (Welch's t = 3.3, p=0.002), as shown in ***Figure 3***. This result remained significant (p = 0.035) when we accounted for uncertainty in our estimates ('Materials and methods').

Previous studies reported that UpsA *var* genes from patient isolates were down-regulated faster than other *var* genes during adaptation to culture (***Peters et al., 2007***; ***Zhang et al., 2011***). The HB3 repertoire contains seven UpsA genes, all but one of which are subtelomeric. We compared the off-rates of the subtelomeric UpsA genes with those of other subtelomeric genes and found no significant difference (Welch's t = 0.65, p=0.54; accounting for uncertainty p=0.55), as shown in ***Figure 3***. Among centrally located genes we compared UpsB with UpsC and again found no significant difference (***Figure 3***; Welch's t = 0.44, p=0.44; accounting for uncertainty p=0.57). This lends support to a previous study suggesting that the rate at which transcriptionally active *var* genes become silenced is less dependent on promoter type than on chromosomal location (***Frank et al., 2007***).

Var gene switching is governed by an activation hierarchy

In each of our cultures, gene transcription levels spanned several orders of magnitude, and only a minority of genes were ever highly transcribed during the experiment. As mentioned earlier, despite the diversity of the starter genes, the relative transcript levels of the activated genes appeared to converge towards similar levels across all cultures (***Figure 1***, ***Figure 1—figure supplements 1 and 2***). To test the degree of similarity in gene activation, we excluded the nine starter genes and ranked the rest of the repertoire according to average transcript levels at later time points (more than 65 generations) for each culture. Pair-wise comparison showed that the ranking was highly correlated between

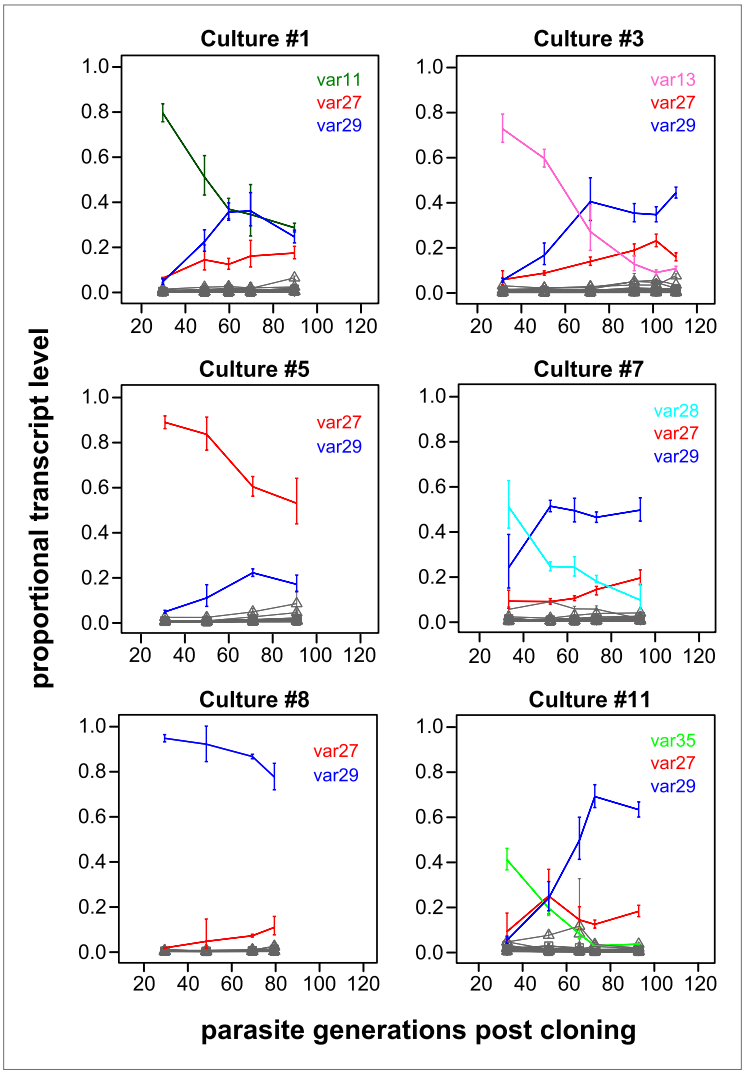


Figure 1. Proportional var transcript levels for six in vitro cultures. The parasite cultures initially expressed a variety of dominant 'starter genes', which switched off at notably different rates. Nevertheless, most cultures converged towards high level transcription of two centrally located genes *var27* and *var29* (red and blue lines, respectively), whereas most other gene transcripts (grey lines) remained at relatively low levels.

DOI: 10.7554/eLife.01074.004

The following figure supplements are available for figure 1:

Figure supplement 1. Proportional *var* transcript levels and model output for cultures 1–6.

DOI: 10.7554/eLife.01074.005

Figure supplement 2. Proportional *var* transcript levels and model output for cultures 7–11.

DOI: 10.7554/eLife.01074.006

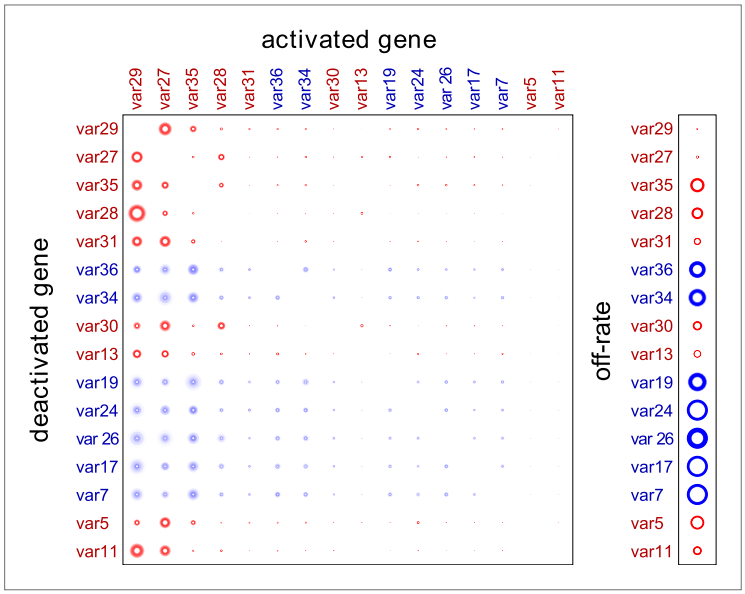


Figure 2. Estimation of switch parameters of highly transcribed genes. Parameter estimations for the 16 most transcribed *var* genes are represented as a switch matrix and an off-rate vector. The diameter of a circle in the *i*th row and *j*th column of the matrix is proportional to the switch bias β_{ij} from gene *i* to gene *j*, and the diameter of a circle in the off-rate vector is proportional to the rate ω_i at which gene *i* becomes silenced. The fuzziness indicates uncertainty in the estimate, such that the darkness of each concentric ring is proportional to the likelihood that the parameter is within the corresponding range (***Noble and Recker, 2012***). Starter gene parameters, which are more precisely estimated, are in red, and genes are arranged from left to right, and top to bottom, in order of average transcript level.

DOI: 10.7554/eLife.01074.007

The following source data and figure supplements are available for figure 2:

Source data 1. Estimated switch biases and 95% credible intervals (CI).

DOI: 10.7554/eLife.01074.008

Figure supplement 1. Validation of estimated switching parameters.

DOI: 10.7554/eLife.01074.009

the cultures (Spearman's $\rho > 0.8$, $p=10^{-5}$ for all pairs), suggesting that the outcome of antigenic switching in *P. falciparum*, or at least within the HB3 isolate, is largely governed by an activation hierarchy that is independent of the currently activated gene.

Consistent with this hypothesis, the estimated switch biases between genes, β_{ij}, were remarkably similar, so that $\beta_{ij} \approx \beta_{kj}$ for each *i*, *j* and *k*. (***Figure 2***, ***Figure 2—source data 1***, ***Figure 2—figure supplement 1***). All genes were found to have high switch biases towards the same few genes, which explain the observed convergence in *var* transcription in our cultures. As expected, the highest biases across the repertoire were found towards *var29* and *var27*, with some genes also frequently switching towards *var35* and/or *var28* (both UpsC-types in central chromosomal location).

To better describe the observed activation hierarchy, we used our MCMC method to estimate the average switch bias towards each gene, which we hereafter refer to as the gene's activation bias. That is, the activation bias of gene *j* is defined as the average of the switch biases β_{ij} for all deactivating genes *i*. It can be understood as a gene specific, *per switch* activation probability and is therefore distinct from the previously defined *on-rate* (***Horrocks et al., 2004***), in that it is invariant to transcript levels. As shown in ***Table 2***, there was a clear hierarchy in gene activation biases spanning over three orders of magnitude. The ranking also appeared to be non-random so that, for example, most genes within central chromosomal location were situated in the top half of the table. Analysing for possible associations with genetic attributes

Table 2. Parameter estimates for the HB3 *var* repertoire with 95% credible intervals

Gene	Location	Domains*	Ups	Off-rate (95% CI)	Activation bias (95% CI)
var29	Central	4	C	0.3% (0.1, 0.5)	27% (22, 32)
var27	Central	4	B	0.8% (0.5, 1.2)	23% (19, 28)
var35	Central	4	C	3.3% (2.2, 4.7)	8.0% (5.8, 11)
var28	Central	4	C	2.2% (1.4, 3.1)	6.9% (5.2, 9.1)
var32	Central	4	C	4.7% (2.5, 5.9)	2.5% (1.5, 3.3)
var36	Central	4	C	1.8% (0.2, 4.2)	2.5% (1.5, 4.1)
var31	Central	4	C	1.4% (0.9, 2.0)	2.2% (1.7, 2.9)
var34	Central	6	C	1.3% (0.1, 3.4)	2.1% (1.4, 3.6)
var13	Subtelomeric	4	B	1.9% (1.5, 2.4)	2.1% (1.6, 2.8)
var30	Central	4	C	1.9% (1.3, 2.7)	1.9% (1.4, 2.5)
var10	Subtelomeric	5	B	3.4% (1.1, 5.7)	1.7% (0.94, 2.7)
var17	Central	6	B	3.0% (0.8, 5.6)	1.7% (0.90, 2.8)
var24	Central	6	B	1.9% (0.1, 5.2)	1.6% (0.85, 3.1)
var7	Subtelomeric	7	B	2.2% (0.3, 5.1)	1.5% (0.83, 2.7)
var19	Central	4	B	1.3% (0.1, 3.2)	1.4% (0.93, 2.5)
var26	Central	4	B	1.5% (0.2, 3.4)	1.4% (0.90, 2.3)
var25	Central	6	B	2.7% (0.4, 5.6)	1.2% (0.60, 2.1)
var1csa	Subtelomeric	8	A	5.5% (4.4, 6.0)	1.2% (0.89, 1.5)
var33	Central	4	C	2.7% (0.5, 5.8)	1.1% (0.57, 1.9)
var16	Subtelomeric	4	B	2.8% (1.0, 5.3)	0.91% (0.52, 1.5)
var4	Subtelomeric	8	A	4.3% (2.1, 5.9)	0.86% (0.52, 1.3)
var14	Subtelomeric	4	B	4.0% (2.2, 5.9)	0.73% (0.45, 1.1)
var22	Central	7	B	1.9% (0.2, 5.1)	0.72% (0.42, 1.3)
var21	Central	7	B	4.7% (2.7, 5.9)	0.70% (0.46, 0.94)
var8	Subtelomeric	7	B	3.7% (1.8, 5.8)	0.66% (0.39, 1.0)
var11	Subtelomeric	5	B	1.5% (0.9, 2.2)	0.60% (0.45, 0.80)
var18	Subtelomeric	4	B	5.2% (3.8, 6.0)	0.60% (0.42, 0.79)
var5	Subtelomeric	6	A	3.5% (2.7, 4.3)	0.53% (0.39, 0.69)
var39p†	Subtelomeric	2	B	4.7% (2.7, 6.0)	0.48% (0.30, 0.66)
*var50*Ψ‡	Central	6	C	4.4% (2.0, 5.9)	0.46% (0.26, 0.67)
var23	Central	6	B	0.6% (0.0, 2.0)	0.38% (0.26, 0.55)
var9	Subtelomeric	6	B	2.8% (0.7, 5.3)	0.36% (0.20, 0.58)
var2csa	Subtelomeric	6	E	4.7% (2.3, 5.9)	0.28% (0.16, 0.40)
var2	Subtelomeric	6	A	1.4% (0.2, 4.2)	0.27% (0.17, 0.50)
var12	Subtelomeric	4	B	3.8% (1.5, 5.8)	0.26% (0.15, 0.40)
var20	Subtelomeric	4	B	4.1% (1.8, 5.9)	0.25% (0.14, 0.37)
var1	Subtelomeric	7	A	5.5% (4.2, 6.0)	0.024% (0.018, 0.031)
var6	Central	8	A	2.9% (0.4, 5.9)	0.018% (0.0086, 0.032)

*Number of encoded DBL (Duffy binding-like) and CIDR (Cys rich inter-domain region).
†Partial gene.
‡Psuedogene.
DOI: 10.7554/eLife.01074.010

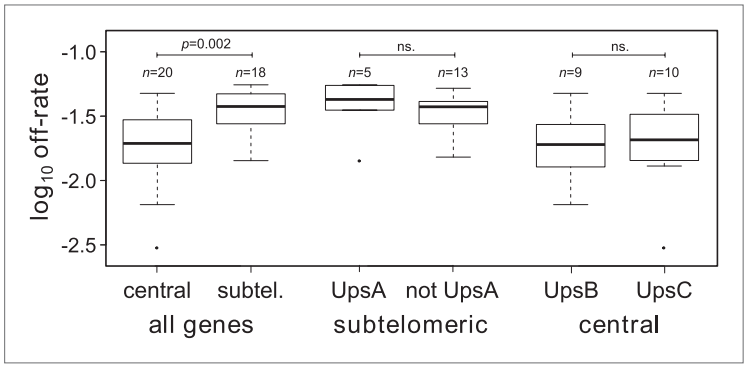


Figure 3. Off-rate estimates show strong dependency on chromosomal location. The estimated rates at which active var genes become silenced are significantly lower for centrally located var genes than for those in subtelomeric locations. There was no significant effect of upstream promoter type when differentially testing off-rates of genes in subtelomeric location (UpsA vs non-UpsA) or in central location (UpsB vs UpsC).

DOI: 10.7554/eLife.01074.011

confirmed that activation biases of centrally located *var* genes are significantly higher than those of subtelomeric genes (***Figure 4***, Welch's t = 2.8, p=0.009; accounting for uncertainty p=0.011). In fact, of the highest four activation biases, three belonged to neighbouring genes on chromosome 4: *var29*, *var27* and *var28*. Importantly, no significant and independent effect of promoter type was found, which again points towards a strong influence of chromosomal location on *var* gene activation.

Additionally, we considered whether gene architecture had an effect on the observed activation biases, as recent research suggests that gene length, in terms of the number of binding domains, is another important *var* gene characteristic besides chromosomal location and promoter type (***Buckee and Recker, 2012***). Grouping genes depending on whether they encoded only four domains (short

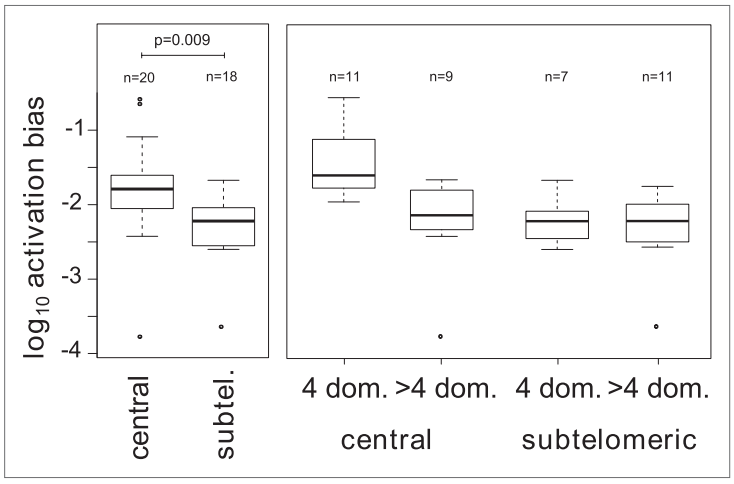


Figure 4. Associations between activation biases and genetic attributes. The mean activation bias is higher for centrally located *var* genes. Within the set of centrally located genes, those encoding only four binding domains have higher activation biases than longer genes. Together, short, central genes have activation biases ≈ 10 times higher than the rest of the repertoire.

DOI: 10.7554/eLife.01074.012

genes) or more than four domains (long genes) showed that activation biases differed significantly depending on gene length (***Figure 4***, $F = 9.7$, p=0.004) as well as chromosomal location ($F = 8.1$, p=0.007) and the interaction of these two factors ($F = 5.0$, p=0.032). As such, central *var* genes encoding only four binding domains (also known as Type 1 *var* genes [***Gardner et al., 2002***]) had on average a 10 times higher activation bias than the rest of the repertoire. Results were similar when we accounted for uncertainty in our estimates (p=0.005, p=0.0096, p=0.037, respectively).

Var gene switching is influenced by the deactivating gene

We next tested explicitly whether the currently active gene has an influence on the direction of switching, that is if in general $\beta_{ij} \neq \beta_{kj}$. We used an optimisation algorithm ('Materials and methods') to fit two alternative models to the data for the 16 most highly transcribed *var* genes: (i) a general model including off-rates and individual switch biases (240 independent parameters), and (ii) a simpler model, in which all sets of switch biases were required to be identical, that is where $\beta_{ij} = \beta_j$ for all i and j (31 independent parameters). A likelihood ratio test indicated that the model with individual switch biases was significantly more likely than the second model ($F = 1.84$, p<0.0001), confirming that antigenic switching in *P. falciparum* is not only controlled by an activation hierarchy but also influenced by the currently active gene. We again tested for associations with chromosomal location and promoter type ('Materials and methods'), which indicated that switch biases were higher between *var* genes with matching chromosomal locations than between differently located genes (***Figure 5***; Welch's $t = 2.2$, p=0.029), whereas no significant association was found between switch biases and matching Ups types or gene lengths. This result needs to be confirmed with larger data sets, however, as statistical significance was lost when we accounted for uncertainty in our parameter estimates.

Var gene sequences suggest a role of activation bias in antigenic evolution

Var gene sequences can be described using a set of 628 conserved homology blocks, with an average length of 19 amino acids (***Rask et al., 2010***). Using homology blocks found only within the DBL1 and CIDR1 domains, which are present in all HB3 *var* genes except *var1csa*, *var2csaA*, *var2csaB* and the partial gene *var39*, we constructed a *var* gene homology network, in which edges between genes are defined and weighted by the number of shared homology blocks (***Figure 6—figure supplement 1***). We then calculated the centrality of each gene within this network. This is a distance measure between all pairs of nodes (i.e., genes) and can therefore be used as a proxy for the relatedness of a gene to the rest of the repertoire. The inclusion of network centrality as an explanatory variable in the previous ANOVA model confirmed that the degree by which *var* genes share sequence blocks is positively correlated with activation bias, independently of chromosomal location and gene length (***Figure 6A***; $F = 8.8$, p=0.006 for the model comparison; accounting for uncertainty p=0.006).

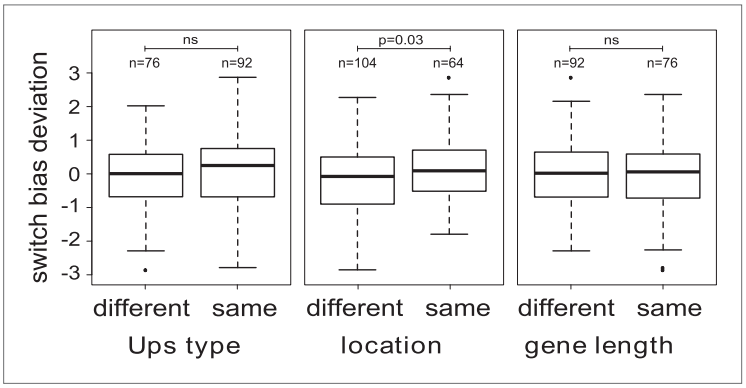


Figure 5. Variation in switch biases (deviation from the mean). Antigenic switching between *var* genes is more frequent between genes with matching chromosomal locations, for example from central to central or from subtelomeric to subtelomeric. No significant associations between switch bias and matching Ups type or gene length are found.
DOI: 10.7554/eLife.01074.013

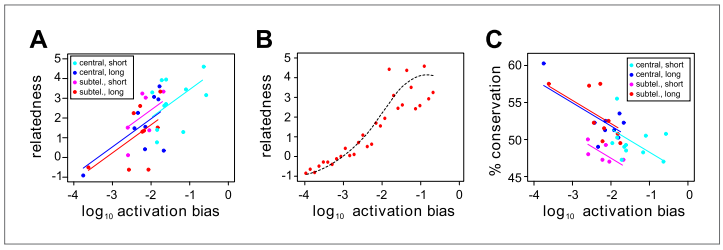


Figure 6. Effect of *var* gene activation on sequence evolution. (**A**) *Var* gene activation biases are significantly correlated with each gene's relatedness to the rest of the HB3 repertoire, as measured by the gene's centrality within a shared homology block network, independently of chromosomal location or gene length. Trend lines show the fit of a linear regression model. (**B**) Simulation of *var* gene evolution by gene conversion, whereby homology blocks are swapped among pairs of genes chosen at random according to their activation biases, shows a similar correlation between the genes' activation biases and their relatedness to the rest of the repertoire. The red points show the outcome of one simulation, and the dashed line is the smoothed average of 50 runs. (**C**) Activation biases are negatively correlated with the gene's average domain sequence conservation, as a measure of population-level diversity, independently of chromosomal location and gene length. Trend lines show the fit of a linear regression model.

DOI: 10.7554/eLife.01074.014

The following figure supplements are available for figure 6:

Figure supplement 1. Relatedness network of the HB3 repertoire.

DOI: 10.7554/eLife.01074.015

We hypothesised that this could indicate a possible role of activation bias, or gene transcription, in *var* gene sequence evolution. Under the assumption that gene transcription facilitates recombination we would expect more frequently activated genes to share more sequence blocks with other genes, simply because they more often act as recombination sites. To test this we explicitly simulated gene conversion, in which pairs of genes were chosen at random, weighted by their activation bias and homology blocks copied between them ('Materials and methods). As expected, we found a strong positive correlation between activation bias and our measure of relatedness in terms of network centrality (***Figure 6B***).

Under the same assumption we would also expect genes with higher activation biases to be overall more diverse. Indeed, using the average sequence conservation of all domains as a measure of population level diversity (***Buckee and Recker, 2012***), we again found this to be significantly and positively correlated, independently of chromosomal location and gene length (***Figure 6C***; $F = 13.6$, p=0.0009 for the model comparison; accounting for uncertainty p=0.0006), suggesting a role of active gene transcription in generating antigenic diversity in *P. falciparum* malaria.

Discussion

Several studies have noted that *var* genes located near the centres of chromosomes tend to be more highly transcribed in vitro than those in subtelomeric location (***Frank et al., 2007***; ***Peters et al., 2007***; ***Enderes et al., 2011***; ***Zhang et al., 2011***; ***Fastman et al., 2012***), and it has been suggested that this might be due to low deactivation rates (***Frank et al., 2007***). In this study, we found that high transcription levels are confined to a subset of central genes that encode only four PfEMP1 binding domains (termed Type 1 *var* genes [***Gardner et al., 2002***]) and, importantly, are the result of high activation biases rather than low off-rates. That is, even though some of the most highly transcribed genes had very low off-rates, the total sum of other genes switching off with high bias towards these genes significantly outweighs the effect of off-rates. This is consistent with our previous study using artificial transcription data, which also showed that off-rates, at least within a biologically plausible range, can have only a modest effect on the total transcript levels within parasite populations (***Noble and Recker, 2012***).

Reanalysis of data from a recent study on the NF54 parasite strain by ***Fastman et al. (2012)*** supports these findings. Consistent with our results for HB3, centrally located *var* genes encoding only four binding domains were more highly transcribed than the rest of the repertoire (Welch's $t = 2.3$, p=0.02), with PF3D7_0809100, PF3D7_0421300 and PF3D7_0800100 being the most highly transcribed genes. According to our model, these are predicted to have high activation bias, and indeed, PF3D7_0809100 has previously been reported as the most transcribed *var* gene in NF54 bulk culture, whereas PF3D7_0421300

and PF3D7_0800100 were found to be the two genes most commonly activated by *var* gene switching in previous studies (**Dzikowski et al., 2006**; **Frank et al., 2007**; **Enderes et al., 2011**). Also in agreement with our results, NF54 gene transcription levels were positively correlated with relatedness to the rest of the repertoire in terms of shared DBL1 and CIDR1 homology blocks (Spearman's $\rho = 0.3$, p=0.044).

Our results further indicate that the direction of transcriptional switching between *var* genes depends not only on intrinsic activation biases but also on the particular preferences of each deactivating gene. Even though these variations were found to have only a modest effect on in vitro switching, they are expected to play a much greater role during infection, where factors such as immune responses and binding affinities can significantly alter the *var* gene expression profiles. Furthermore, the combination of an activation hierarchy and these specific variations could explain our previously proposed source-sink model underlying antigenic switching in *P. falciparum* (**Recker et al., 2011**), in which genes with high in-degrees, i.e., those most commonly switched to, correspond to the genes here identified as having high activation biases. Intriguingly, also, our results suggest that switch biases may be stronger between genes in similar chromosomal locations: central genes prefer to switch to other central genes, and subtelomeric genes prefer to switch to other subtelomeric genes. More data are needed to fully establish the factors influencing variation in switch biases, however, especially for less activated genes, where parameter estimations are more prone to uncertainties.

Despite recent advances in elucidating the genetics and epigenetics of *var* gene activation and silencing, not much is known about how mutually exclusive switching between the 60 *var* genes is controlled at the molecular level. The discovery of intrinsic *var* gene activation biases and their associations with chromosomal locations and other genetic features could therefore have important implications for future studies. These factors are clearly correlated, however, and it is uncertain how they are linked by cause and effect. For example, in a recent study it was found that gene length is positively correlated with sequence conservation (**Buckee and Recker, 2012**), and long *var* genes are more often positioned in subtelomeric regions. Nevertheless, the correlation between activation bias and relatedness, which was found to be independent of chromosomal location, gene length and upstream promoter, suggests a role of gene transcription in facilitating recombination and therefore in generating and maintaining antigenic diversity in *P. falciparum*. Transcription-associated recombination has been observed in other eukaryotes and may result from increased accessibility to recombination proteins or from stalled replication forks (**Prado and Aguilera, 2005**; **Gottipati and Helleday, 2009**). Importantly, for *P. falciparum* this phenomenon could help to explain why different groups of *var* genes within the repertoire display very different levels of diversity within the population (**Buckee et al., 2009**; **Buckee and Recker, 2012**).

It is further possible that some *var* genes are under balancing selection, whereby the relative sequence conservation of longer genes could be due to optimised functionality, for example in terms of binding avidity and/or affinity. This assumption would also help to reconcile apparent contradictions between our observations and *var* gene transcription in vivo. That is, various in vivo studies have reported UpsA *var* genes to be over-expressed during infections in individuals with little previous exposure (**Jensen et al., 2004**; **Bull et al., 2005**; **Lavstsen et al., 2005**; **Kyriacou et al., 2006**; **Warimwe et al., 2009**), despite their low activation rates in vitro. Each *var* repertoire consists of a set of relatively long and conserved genes, such as those predominantly found with in the UpsA group, and another, short and diverse set of genes, such as those predominantly found within the UpsB and UpsC groups (**Buckee and Recker, 2012**). Under the assumption that more conserved genes have an in vivo survival advantage, possibly due to higher binding affinities (**Avril et al., 2012**; **Claessens et al., 2012**; **Lavstsen et al., 2012**), these variants will dominate early infections in naive hosts, regardless of activation rates. Their lower sequence diversity, however, will then lead to a more rapid acquisition of protective immunity, causing a transition towards the expression of more diverse variants, facilitated by their inherent activation biases. In light of this we would also speculate that the proposed order in which individuals acquire immunity (anti-UpsA followed by anti-UpsB/C), as well as the proposed associations between the expression of UpsA *var* genes and severe disease outcome, especially in young children, is less determined by the Ups group itself and is more a consequence of within-host selection of long, conserved genes, which are often but not exclusively found within the UpsA group.

Together, these results demonstrate how selection pressures operating on multiple ecological scales have shaped the phenotypic plasticity embodied within the antigenic repertoire of *P. falciparum*. Within this setup, *var* genes and the molecular mechanisms that underlie their sequential activation and silencing have co-evolved to allow the parasite to express the most advantageous phenotype in response to its current environment.

Materials and methods

Parasite culture and cDNA preparation

Parasites were cultured and sorbitol-synchronised using standard techniques (**Trager and Jensen, 1976**; **Lambros and Vanderberg, 1979**), and RNA was extracted from saponin lysed, mid-to-late ring stage parasites with Trizol as previously described (**Kyes et al., 2000**). For cDNA preparation, 5 µg RNA was treated in 20 µl reactions with TurboFree DNase (Ambion®; following manufacturer's recommendations except incubation was at 16°C for 40 min). All DNase treated RNA samples were tested by PCR with primers to fructose-bisphosphate aldolase (PF14_0425) (**Salanti et al., 2003**) to ensure complete removal of gDNA. Duplicate 20 µl reactions consisting of 6 µl RNA were reversed transcribed with 100 ng random hexamers (Invitrogen) and Bioscript reverse transcriptase (Bioline; following manufacturer's instructions) at 40°C for 40 min. The resulting cDNA was stored at −20°C in single-use aliquots and diluted 1:6 for real time PCR analysis. cDNA was synthesised in triplicate (84% of all data points) or duplicate, and real time PCR was done in duplicate for each primer pair on each cDNA.

Primer design and initial testing

Primers specific for the HB3 *var* exon 1 repertoire were optimised to meet several basic criteria: that they distinguished between the known *var* gene sequences within the HB3 genome; that they were specific for HB3 *var* genes with no cross-reactivity to 3D7 *var* genes (for future application to studies on 3D7 × HB3 cross progeny); and that they reflected measurement of full-length *var* transcript. Primer pairs were designed using Primer3plus (**Kraemer et al., 2007**; **Untergasser et al., 2007**), with parameters manually set for product size range (100–150 nt), primer length (20 nt), Tm (60°C), GC clamp (1 nt), salt correction formula and thermodynamic parameters (SantaLucia 1998); all other parameters were default settings (see **Table 3** for primer sequences). To avoid general cross-detection of all *var* types, the first kilobase of exon 1 containing the conserved DBLα domain was avoided. The final kilobase was also excluded to avoid sterile transcripts (**Su et al., 1995**; **Kyes et al., 2007**). Each primer pair was specific for a single gene except, unavoidably, the *var3* primer pair and the two pairs designed to the highly similar copies of *var2csa*. Each HB3 primer pair was tested five times on gDNA; pairs that varied more than ±1.5 threshold cycles (Ct) values from the median or had efficiency value (Amplification) less than 1.85 were redesigned. Primer pairs that amplified 3D7 genomic DNA were discarded and redesigned so that they were specific for HB3 *var* genes only.

qRT-PCR

Quantitative real time PCR (qRT-PCR) reactions were prepared with a Corbett CAS1200 liquid handling system and run on a Corbett Rotor-Gene 6000. Each 10 µl reaction contained 1 µl template, 5 µl SensiMix SYBR (Bioline), 3 µl water and 1 µl primer mix (1.6 µM of each primer). The cycling conditions were as follows: hold, 95°C/10 min; 45 cycles (95°C/25 s, 58°C/25 s, 68°C/30 s); melt, 55°C−99°C. Amplicons were checked for a single product of the expected size by agarose gel electrophoresis. Melt analysis confirmed single products.

Data collection, processing and verification

Transcription levels for each replicate were calculated by the comparative quantification method of **Pfaffl (2001)**. Comparative quantification results are given as expression levels relative to a control (housekeeping) gene, seryl-tRNA synthetase (Seryl-tRNAF agctacctcagaacaaccattatgtgc, Seryl-tRNAR atcctttc-catgtgcccctgc), which is arbitrarily set at 1.0. Minor differences in Take Off values between primer pairs were accounted for using a correction factor derived from five replicates on gDNA. Expression was determined using $A^{T.O.-CF}$, where *A* is the average amplification, *T.O.* is the average Ct and *CF* is the correction factor. To express results relative to the control gene, the following formula was used: $A_{ref}^{TO-CF}/A_{x}^{TO-CF}$. Expression levels were converted to proportions before calculating geometric means of the replicates. Replicates were excluded as outliers (1.1%) if they deviated from the mean by a factor of more than $10^{1.5}$ or had an amplification efficiency below 1.5 (most excluded measurements met both criteria).

Dominant *var* gene expression detected by qRT-PCR was verified as both full-length and dominant by northern blot analysis, comparing specific probes for each dominant gene vs a generic exon 2 probe (data not shown). The primer pairs for *var* genes 5, 11, 13, 27, 28, 29, 31 and 35 were thus verified as specific for detecting full-length transcripts. The *var3* primer pair, which was known to also detect *var5* and pseudogene *var46*, indicated dominant *var3* alongside *var5* in culture #4. However, we were unable to

Table 3. HB3 qPCR primers and cross-referenced *var* gene identifiers

Gene name	Contig name	Broad locus name	F oligonucleotide	R oligonucleotide
HB3 var16	HB3-1000-1	PFHG_03232.1	ccctgtccacaaccatcagc	cgtcgtcgtcatcagtgtcc
HB3 var1	HB3-1000-2	PFHG_03234.1	ccaaaggagaaggcaccacc	acctatggcacccctctcac
HB3 var12	HB3-1040	PFHG_03416.1	gatgctacaaccaccccacc	gtgttaccactcgcccactc
HB3 var27	HB3-1704_1	PFHG_03476.1	gctcccaaccaccacgttcc	gcttcctgctggtggctgtc
HB3 var28	HB3-1074-2	PFHG_03478.1	gatggcacaaaagttggcgg	tgttctgggtcgacctcctc
HB3 var29	HB3-1074-3	PFHG_03480.1	aagaagatggcgacgaaggc	tccggtgatccctcttctgg
HB3 var13	HB3-1107	PFHG_03516.1	tggtaaatgcaagggtgatacagg	tgcatcgttatcactcaccagc
HB3 var1csa	HB3-1108	PFHG_03521.1	cgcaatatgcaactaatgac	acttggcaatattctgaacg
HB3 var2	HB3-1210	PFHG_03840.1	cgaggacaccacggaggagg	ttggtgctgctggttgtggc
HB3 var5	HB3-1235	PFHG_03671.1	aggtctgctccttcagatgcgtg	tgttttccctaccatgacaaggatgcc
HB3 var36	HB3-1296	PFHG_04012.1	atggacaaatgatggtaagg	taggagtaggtgttgcgttc
HB3 var17	HB3-1296-2	PFHG_04014.1	agatggcgacaaaggccaag	ttgggtttggcaccactagc
HB3 var35	HB3-1296-3	PFHG_04015.1	aaacggaaaacctggcctcc	tcgtcttggcctttggcttc
HB3 var14	HB3-1308	PFHG_04035.1	ggtggtggtgccgatcccgcc	tgtgacgcctccgtcttagtggccc
HB3 var8	HB3-1334	PFHG_04081.1	ggcggtgtctgtattccacc	gctgcctcaccacctgttag
HB3 var9	HB3-1408	PFHG_04057.1	tgctatgacgtgtaatgcccc	acttacatgagtcccatctggtg
HB3 var32	HB3-1459	Fbadf	ggaaaccgcggtggactcac	acttgtgggtgctttggggc
HB3 var11	HB3-1499	PFHG_04491.1	attggatgatgcctgtcgcc	ggcacctggtttagtggtgg
HB3 var19	HB3-1514	PFHG_04620.1	aaactgacaatggccccgac	gttgttgaggggggtcttcgg
HB3 var18	HB3-1523	PFHG_04593.1	gcggctcacccgacatcttc	gccgcctcgtcttcttcgtc
HB3 var10	HB3-1587	PFHG_04749.1	accactcgtgccaccacctc	gagtttgtacctggcaccccc
HB3 var7	HB3-1604-1	PFHG_04769.1	agcgagtggtactcaggagg	gatggaccacgagatgtgcc
HB3 var20	HB3-1604-2	PFHG_04770.1	acgaagaagacgatgccacc	gaagtcttcggagcgaccac
HB3 var4	HB3-1703	PFHG_04861.1	tggtgccaaagacccctccc	ggccactcgctgtgtctgtg
HB3 var2csaA	HB3-1727	PFHG_05046.1	gggggaaatgtggggtgccg	gggggataccccacactcattaccag
HB3 var3	HB3-1737	PFHG_05052.1	aaagtgcgaagcacctcccc	cgccactgcagggattagctg
HB3 var2csaB	HB3-1817	PFHG_05155.1	tggtacagctgatggtggtacttccg	tgtgcccgctttacggtttcg
*HB3 var39p**	HB3-2007	efbe	agccattacgtgcgaagctgg	agcggcacatcggcatttttg
HB3 var34	HB3-209	PFHG_00592.1	agtggtgctgtagagccaaaagac	cctgcggcggtgctgtaagg
HB3 var23	HB3-699-1	PFHG_02272.1	gaacccttgacgacgacac	ctcaacacacgtcaaaggcg
HB3 var30	HB3-699-2	PFHG_02273.1	aagacgacaaacctggcacc	gtcgttgcttttggcttcgg
HB3 var6	HB3-699-3	PFHG_02274.1	attcacagcactgaaagtcc	tcacaatcattaaaagcatcc
HB3 var26	HB3-699-4	PFHG_02276.1	aagcagctgatggaacggac	tggttgttgtgggtcttggc
HB3 var31	HB3-699-5	PFHG_02277.1	cgcgaagacgaaaacgtcac	gtttcatccggaccgtcctc
*HB3 var50*Ψ†	HB3-752-1	PFHG_02419.1	tggtaatgatgaagatgacg	gaattggcttcactttgttc
HB3 var24	HB3-752-2	PFHG_02421.1	gctcgctctttaccacccgc	ttccgtctcctccttcgccg
HB3 var25	HB3-752-3	PFHG_02423.1	agtggtgccaaaactgtcgg	accacaaaagtcgcttcccc
HB3 var22	HB3-752-4	PFHG_02425.1	ccaccacaaaaccccctccag	tccgcttgtggttcgtcttc
HB3 var33	HB3-752-5	PFHG_02429.1	acagaaagttggacaggatg	atggttgttttgagaattgc

*Partial gene.
†Pseudogene.
DOI: 10.7554/eLife.01074.016

detect *var3* expression by northern blot. As this sample was positive by northern blot probed with *var5*, and it was the only one with 'dominant' *var3* expression by qRT-PCR, we discarded all *var3* qRT-PCR data as ambiguous. We also discarded all the data from one of the two *var2csaA*/*var2csaB* primer pairs.

Parameter estimation

Following **Recker et al. (2011)**, we assumed that *var* transcription profiles can be described by a time-discrete model:

$$v_{i,c,t+1} = (1-\omega_i)v_{i,c,t} + \sum_{j\neq i}\omega_j\beta_{ji}v_{j,c,t},$$

where $v_{i,c,t}$ is the relative transcript level of variant i in culture c at time t, ω_i is the variant specific off-rate and β_{ji} is the switch bias (probability of a switch) from variant j to variant i. We assumed a maximum off-rate of 6% per generation, which is consistent with previous studies and theory (**Roberts et al., 1992**; **Horrocks et al., 2004**; **Frank et al., 2007**; **Recker et al., 2011**).

We used a Markov chain Monte Carlo (MCMC) method to obtain posterior distributions of likely values for the switch biases and off-rates, using the pooled data from all observed cultures as input. For model comparisons we used a simulated annealing algorithm to find maximum likelihood parameter sets. These methods have previously been described in detail and tested using diverse artificial data sets (**Noble and Recker, 2012**). For some of our analyses it was desirable to use a model of reduced dimension. Previously we have shown that reliable parameter estimates can be obtained from data for the 16 most highly transcribed *var* genes (**Noble and Recker, 2012**), which in this case comprised the 9 starter genes and the 7 others with highest average transcript levels. These 16 genes together accounted, on average, for more than 92% of the sum of transcript levels in each culture (excluding the culture's starter gene), whereas none of the excluded genes accounted for more than 1%. To estimate activation biases, we modified the model so that the direction of switching was independent of the previously active gene (i.e., $\beta_{ji} = \beta_{ki}$ for all i, j and k).

Before conducting statistical tests, we transformed switch biases using the logit function $\mathrm{logit}(x) = \log(x/(1-x))$ to map from the range (0,1) to the range $(-\infty, \infty)$. Additionally, to account for uncertainty in our parameter estimates, we performed tests on the collected output of our MCMC method. This method samples from the posterior distribution, which means it selects numerous parameter sets such that the probability of each being chosen depends on how likely it is to explain the data. We performed our statistical tests on all the selected parameter sets and calculated mean p-values. We used the same parameter sets to obtain credible intervals for our estimates, as shown in **Table 2**.

Variation in switch biases

We found that all *var* genes had similar sets of switch biases, such that $\beta_{ij} \approx \beta_{kj}$ for each i, j and k. A model comparison indicated that these sets were not identical but varied depending on the deactivating gene. To estimate these variations we took a subset of the MCMC switch bias estimates and standardised each set $\{\beta_{1j}, \beta_{2j}, \ldots, \beta_{nj}\}$ by subtracting the mean and dividing by the standard deviation (after rescaling each set to compensate for $\beta_{jj} = 0$). Negative values then corresponded to biases that were smaller than average and positive values to those that were larger. We used the subset of switch biases from the 9 starter genes to the 25 most transcribed genes, omitting the diagonal terms.

Relatedness network

We constructed a network in which each node represented a gene and each edge was weighted by the number of shared variable DBL1 and CIDR1 homology blocks. Each gene's centrality within the network was calculated using a weighted closeness measure (**Newman, 2001**) implemented by the tnet package for R (**Opsahl, 2009**). We applied variance stabilising transformations to the data prior to analysis. Specifically, the centrality measure was rescaled and squared, and to each switch bias we applied the logit transformation, added five and then square-rooted the magnitude. These transformations made the data conform better to the assumptions of our statistical model; test results would have been more significant had we instead used the untransformed data.

Gene conversion model

We simulated gene conversion in a repertoire of 30 genes, each comprising 40 homology blocks, with 30 possible variants for each homology block. Genes were initially constructed of randomly chosen homology blocks and were assigned activation biases according to a geometric sequence. For 30,000

iterations, pairs of genes were selected at random, with the probability of selection being proportional to activation bias. A randomly chosen homology block was copied from the first of the chosen genes to replace the corresponding block in the second gene. Homology blocks in all genes also underwent random mutation at a rate 10 times lower than the gene conversion rate. We constructed a relatedness network and calculated centrality in the same way as for the HB3 data.

Initial conditions of the model

Our switching model required a set of initial conditions. Assuming that each culture, excluding the parent, began with a single parasite and that populations expanded approximately fivefold per generation, we deduced that the first transcriptional switch most likely occurred, on average, two or three generations after cloning. Therefore we assumed that each of these cultures remained clonal for the first three generations. For the parent culture, we used the first observation as the initial condition. If a switch had occurred in one or more of the initially clonal cultures within the first few generations, when the population was very small, then our assumed initial conditions would be inaccurate. Therefore we also ran our algorithms using the first observations as the initial conditions for all cultures. The parameter estimates were highly consistent with those obtained using clonal initial conditions.

Cross-validation of parameter estimates

The simplest way to estimate switch biases and off-rates for a single *var* gene is to obtain a clone transcribing the gene, grow it in culture and analyse changes in transcript levels over time. A particular strength of our method was that it also estimated parameters for genes we did not clone. To assess the accuracy of results obtained in this way, we performed cross-validation by sequentially excluding data sets. For example, to derive alternative estimates for *var29* we ran the MCMC algorithm without the data for culture #8 (which initially expressed *var29*), and to derive alternative estimates for *var27* we excluded data for cultures #5 and #6. The results, shown in ***Figure 2—figure supplement 1***, confirmed our method's ability to estimate switch biases of non-starter genes with sufficient accuracy to determine the general switching pattern. Off-rate estimates were however not always accurate, especially for genes with very low transcript levels.

We used a similar method to test the accuracy of our method for estimating variations in switch biases. Cultures #2 and #3 initially expressed the same starter gene *var13*, and both cultures #5 and #6 initially expressed *var27*. By excluding the data from culture #2 we derived estimates for the *var13* switch biases mostly dependent on the culture #3 data, and vice versa. Similarly we excluded data from each of cultures #5 and #6 to obtain two sets of estimates for the *var27* switch biases. For each of the two different data sets we calculated the deviations from the mean switch biases and found these to be positively correlated (Pearson's $\rho = 0.43$, p=0.015), confirming the validity of our method.

Validation of model assumptions

Our likelihood function, used for parameter estimation and model comparisons, assumed that input data were subject to measurement errors following a log-normal distribution. We have previously shown that our MCMC algorithm reliably determines network structures when the measurement error distribution is log-normal (base 2) with standard deviation $\sigma \leq 1$, and that σ can be reliably estimated using our simulated annealing algorithm (***Noble and Recker, 2012***). In the 16-gene model, $\sigma \approx 0.73$, which was well within the required range.

To test whether the measurement errors obeyed a log-normal distribution, we analysed the distribution of the deviations between the data and the simulated annealing output, after applying log transformations. The distribution for the 16-gene model was unimodal and only slightly asymmetrical (skewness 0.4) but was more peaked than a normal distribution (excess kurtosis 2.4), so that it somewhat resembled a Laplace distribution. We tried running the simulated annealing algorithm with an error function that assumed log-Laplace-distributed noise, and this made very little difference to the results. The deviations for the constrained 38-gene model were similarly distributed, excluding a very small minority of outliers.

Unusually large deviations occurred at 13 of the 64 observed time points, 8 of which were between generations 48 and 52 and therefore belonged to the same round of measurements. Transcript levels deviated more for longer genes than for shorter genes, and northern blotting of the samples taken at the affected time points showed unusually low levels of long, for example *var* length, RNA, confirming that the cause was not PCR error but most likely an as yet unidentified artefact of RNA preparation. Excluding the 13 data points affected by this phenomenon made very little difference to our parameter estimates and statistical test results.

Analysis of the qRT-PCR replicates revealed that measurements of lower transcript levels had higher standard errors, but the average standard error increased only threefold for each 1000-fold difference in transcript level.

Growth rates

It has been suggested that *var* gene transcript levels in culture may be affected by small differences in parasite growth rates (***Horrocks et al., 2004***), although no evidence has been found to support this hypothesis. To test whether such differences might affect our estimates we added to our model an excess growth rate parameter ϕ_i for each gene i:

$$v_{i,c,t+1} = (1-\omega_i)(1+\phi_i)v_{i,c,t} + \sum_{j\neq i}\omega_j\beta_{ji}(1+\phi_j)v_{j,c,t}.$$

When we allowed growth rates to vary by up to 0.5% per generation (equivalent to approximately 50% over the duration of the experiment), the extended model was not significantly more likely than the model with only switch biases and off-rates. When we increased the upper bound on the excess growth rates to 1% per generation, the extended model became significantly more likely. However, such large differences in growth rates in vitro are biologically implausible and have never been observed. Importantly, the inclusion of growth rate differences had very little effect on our switch bias and off-rate estimates.

Acknowledgements

We would like to thank Caroline O Buckee for providing estimates of *var* gene domain sequence conservation. The work was funded by the Wellcome Trust (Grant No. 082130/Z/07/Z to CN), the Biotechnology and Biological Sciences Research Council (studentship to RN) and the Royal Society (University Research Fellowship to MR).

Additional information

Funding

Funder	Grant reference number	Author
Wellcome Trust	082130/Z/07/Z	Chris I Newbold
BBSRC		Robert Noble
Royal Society		Mario Recker

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

RN, Analysis and interpretation of data, Drafting or revising the article; ZC, Acquisition of data, Analysis and interpretation of data; SK, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; RP, Acquisition of data, Analysis and interpretation of data; CIN, MR, Conception and design, Analysis and interpretation of data, Drafting or revising the article

References

Avril M, Tripathi AK, Brazier AJ, Andisi C, Janes JH, Soma VL, et al. 2012. A restricted subset of *var* genes mediates adherence of *Plasmodium falciparum*-infected erythrocytes to brain endothelial cells. *Proc Natl Acad Sci USA* **109**:E1782–90. doi: 10.1073/pnas.1120534109.

Blomqvist K, Normark J, Nilsson D, Ribacke U, Orikiriza J, Trillkott P, et al. 2010. *var* gene transcription dynamics in *Plasmodium falciparum* patient isolates. *Mol Biochem Parasitol* **170**:74–83. doi: 10.1016/j.molbiopara.2009.12.002.

Boone KE, Watters DA. 1995. The incidence of malaria after splenectomy in Papua New Guinea. *BMJ* **311**:1273. doi: 10.1136/bmj.311.7015.1273.

Buckee CO, Recker M. 2012. Evolution of the multi-domain structures of virulence genes in the human malaria parasite, *Plasmodium falciparum*. *PLOS Comput Biol* **8**:e1002451. doi: 10.1371/journal.pcbi.1002451.

Buckee CO, Bull PC, Gupta S. 2009. Inferring malaria parasite population structure from serological networks. *Proc Biol Sci* **276**:477–85. doi: 10.1098/rspb.2008.1122.

Buffet PA, Safeukui I, Milon G, Mercereau-Puijalon O, David PH. 2009. Retention of erythrocytes in the spleen: a double-edged process in human malaria. *Curr Opin Hematol* **16**:157–64. doi: 10.1097/MOH.0b013e32832a1d4b.
Bull PC, Berriman M, Kyes S, Quail MA, Hall N, Kortok MM, et al. 2005. *Plasmodium falciparum* variant surface antigen expression patterns during malaria. *PLOS Pathog* **1**:202–13. doi: 10.1371/journal.ppat.0010026.
Bull PC, Buckee CO, Kyes S, Kortok MM, Thathy V, Guyah B, et al. 2008. *Plasmodium falciparum* antigenic variation. Mapping mosaic *var* gene sequences onto a network of shared, highly polymorphic sequence blocks. *Mol Microbiol* **68**:1519–34. doi: 10.1111/j.1365-2958.2008.06248.x.
Cham GK, Turner L, Kurtis JD, Mutabingwa T, Fried M, et al. 2010. Hierarchical, domain type-specific acquisition of antibodies to *Plasmodium falciparum* erythrocyte membrane protein 1 in Tanzanian children. *Infect Immun* **78**:4653–9. doi: 10.1128/IAI.00593-10.
Claessens A, Adams Y, Ghumra A, Lindergard G, Buchan CC, Andisi C, et al. 2012. A subset of group a-like *var* genes encodes the malaria parasite ligands for binding to human brain endothelial cells. *Proc Natl Acad Sci USA* **109**:E1772–81. doi: 10.1073/pnas.1120461109.
Deitsch KW, del Pinal A, Wellems TE. 1999. Intra-cluster recombination and *var* transcription switches in the antigenic variation of *Plasmodium falciparum*. *Mol Biochem Parasitol* **101**:107–16. doi: 10.1016/S0166-6851(99)00062-6.
Dzikowski R, Frank M, Deitsch K. 2006. Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. *PLOS Pathog* **2**:e22. doi: 10.1371/journal.ppat.0020022.
Enderes C, Kombila D, Dal-Bianco M, Dzikowski R, Kremsner P, Frank M. 2011. *Var* gene promoter activation in clonal *Plasmodium falciparum* isolates follows a hierarchy and suggests a conserved switching program that is independent of genetic background. *J Infect Dis* **204**:1620–31. doi: 10.1093/infdis/jir594.
Fastman Y, Noble R, Recker M, Dzikowski R. 2012. Erasing the epigenetic memory and beginning to switch–the onset of antigenic switching of *var* genes in *Plasmodium falciparum*. *PLOS ONE* **7**:e34168. doi: 10.1371/journal.pone.0034168.
Frank M, Dzikowski R, Amulic B, Deitsch K. 2007. Variable switching rates of malaria virulence genes are associated with chromosomal position. *Mol Microbiol* **64**:1486–98. doi: 10.1111/j.1365-2958.2007.05736.x.
Frank M, Kirkman L, Costantini D, Sanyal S, Lavazec C, Templeton TJ, et al. 2008. Frequent recombination events generate diversity within the multi-copy variant antigen gene families of *Plasmodium falciparum*. *Int J Parasitol* **38**:1099–109. doi: 10.1016/j.ijpara.2008.01.010.
Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, et al. 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* **407**:1018–22. doi: 10.1038/35039531.
Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**:498–511. doi: 10.1038/nature01097.
Gottipati P, Helleday T. 2009. Transcription-associated recombination in eukaryotes: link between transcription, replication and recombination. *Mutagenesis* **24**:203–10. doi: 10.1093/mutage/gen072.
Greenwood BM, Bradley AK, Greenwood AM, Byass P, Jammeh K, Marsh K, et al. 1987. Mortality and morbidity from malaria among children in a rural area of the Gambia, West Africa. *Trans R Soc Trop Med Hyg* **81**:478–86. doi: 10.1016/0035-9203(87)90170-2.
Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C. 1999. Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nat Med* **5**:340–3. doi: 10.1038/6560.
Horrocks P, Pinches R, Christodoulou Z, Kyes SA, Newbold CI. 2004. Variable *var* transition rates underlie antigenic variation in malaria. *Proc Natl Acad Sci USA* **101**:11129–34. doi: 10.1073/pnas.0402347101.
Jensen AT, Magistrado P, Sharp S, Joergensen L, Lavstsen T, et al. 2004. *Plasmodium falciparum* associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group a *var* genes. *J Exp Med* **199**:1179–90. doi: 10.1084/jem.20040274.
Kaestli M, Cockburn IA, Cortes A, Baea K, Rowe JA, Beck HP. 2006. Virulence of malaria is associated with differential expression of *Plasmodium falciparum var* gene subgroups in a case-control study. *J Infect Dis* **193**:1567–74. doi: 10.1086/503776.
Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, Christodoulou Z, et al. 2007. Patterns of gene recombination shape *var* gene repertoires in *Plasmodium falciparum*: comparisons of geographically diverse isolates. *BMC Genomics* **8**:45. doi: 10.1186/1471-2164-8-45.
Kraemer SM, Smith JD. 2006. A family affair: *var* genes, PfEMP1 binding, and malaria disease. *Curr Opin Microbiol* **9**:374–80. doi: 10.1016/j.mib.2006.06.006.
Kyes S, Christodoulou Z, Pinches R, Kriek N, Horrocks P, Newbold C. 2007. *Plasmodium falciparum* var gene expression is developmentally controlled at the level of RNA polymerase II-mediated transcription initiation. *Mol Microbiol* **63**:1237–47. doi: 10.1111/j.1365-2958.2007.05587.x.
Kyes S, Pinches R, Newbold C. 2000. A simple RNA analysis method shows *var* and *rif* multigene family expression patterns in *Plasmodium falciparum*. *Mol Biochem parasitol* **105**:311–5. doi: 10.1016/S0166-6851(99)00193-0.
Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, Thera MA, et al. 2006. Differential *var* gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria compared to hyperparasitaemia. *Mol Biochem Parasitol* **150**:211–8. doi: 10.1016/j.molbiopara.2006.08.005.
Lambros C, Vanderberg JP. 1979. Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. *J Parasitol* **65**:418–20. doi: 10.2307/3280287.
Lavstsen T, Magistrado P, Hermsen CC, Salanti A, Jensen AT, Sauerwein R, et al. 2005. Expression of *Plasmodium falciparum* erythrocyte membrane protein 1 in experimentally infected humans. *Malar J* **4**:21. doi: 10.1186/1475-2875-4-21.

Lavstsen T, Salanti A, Jensen ATR, Arnot DE, Theander TG. 2003. Sub-grouping of *Plasmodium falciparum* 3D7 *var* genes based on sequence analysis of coding and non-coding regions. *Malar J* **2**:27. doi: 10.1186/1475-2875-2-27.

Lavstsen T, Turner L, Saguti F, Magistrado P, Rask TS, Jespersen JS, et al. 2012. *Plasmodium falciparum* erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children. *Proc Natl Acad Sci USA* **109**:E1791–800. doi: 10.1073/pnas.1120455109.

Marsh K. 1992. Malaria–a neglected disease? *Parasitology* **104**(Suppl):S53–69. doi: 10.1017/S0031182000075247.

Marsh K, Snow RW. 1997. Host-parasite interaction and morbidity in malaria endemic areas. *Philos Trans R Soc Lond B Biol Sci* **352**:1385–94. doi: 10.1098/rstb.1997.0124.

Newman MEJ. 2001. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Phys Rev E Stat Nonlin Soft Matter Phys* **64**:016132. doi: 10.1103/PhysRevE.64.016132.

Noble R, Recker M. 2012. A statistically rigorous method for determining antigenic switching networks. *PLOS ONE* **7**:e39335. doi: 10.1371/journal.pone.0039335.

Ockenhouse CF, Ho M, Tandon NN, Vanseventer GA, Shaw S, White NJ, et al. 1991. Molecular-basis of sequestration in severe and uncomplicated *Plasmodium falciparum* malaria - differential adhesion of infected erythrocytes to CD36 and ICAM-1. *J Infect Dis* **164**:163–169. doi: 10.1093/infdis/164.1.163.

Opsahl T. 2009. *Structure and evolution of weighted networks*. p. 104–122 [Online]. Available: http://toreopsahl.com/publications/thesis/.

Peters J, Fowler E, Gatton M, Chen N, Saul A, Cheng Q. 2002. High diversity and rapid changeover of expressed *var* genes during the acute phase of *Plasmodium falciparum* infections in human volunteers. *Proc Natl Acad Sci USA* **99**:10 689–94. doi: 10.1073/pnas.162349899.

Peters JM, Fowler EV, Krause DR, Cheng Q, Gatton ML. 2007. Differential changes in *Plasmodium falciparum var* transcription during adaptation to culture. *J Infect Dis* **195**:748–55. doi: 10.1086/511436.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29**:e45. doi: 10.1093/nar/29.9.e45.

Pongponratn E, Riganti M, Punpoowong B, Aikawa M. 1991. Microvascular sequestration of parasitized erythrocytes in human falciparum-malaria - a pathological study. *Am J Trop Med Hyg* **44**:168–75.

Prado F, Aguilera A. 2005. Impairment of replication fork progression mediates RNA polII transcription-associated recombination. *EMBO J* **24**:1267–76. doi: 10.1038/sj.emboj.7600602.

Rask TS, Hansen DA, Theander TG, Pedersen AG, Lavstsen T. 2010. *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes - divide and conquer. *Plos Comput Biology* **6**:e1000933. doi: 10.1371/journal.pcbi.1000933.

Recker M, Buckee CO, Serazin A, Kyes S, Pinches R, Christodoulou Z, et al. 2011. Antigenic variation in *Plasmodium falciparum* malaria involves a highly structured switching pattern. *Plos Pathog* **7**:e1001306. doi: 10.1371/journal.ppat.1001306.

Roberts DJ, Craig AG, Berendt AR, Pinches R, Nash G, Marsh K, et al. 1992. Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature* **357**:689–92. doi: 10.1038/357689a0.

Rottmann M, Lavstsen T, Mugasa JP, Kaestli M, Jensen AT, Muller D, et al. 2006. Differential expression of *var* gene groups is associated with morbidity caused by *Plasmodium falciparum* infection in Tanzanian children. *Infect Immun* **74**:3904–11. doi: 10.1128/IAI.02073-05.

Rowe JA, Claessens A, Corrigan RA, Arman M. 2009. Adhesion of *Plasmodium falciparum*-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications. *Expert Rev Mol Med* **11**:e16. doi: 10.1017/S1462399409001082.

Salanti A, Dahlback M, Turner L, Nielsen MA, Barfod L, Magistrado P, et al. 2004. Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. *J Exp Med* **200**:1197–203. doi: 10.1084/jem.20041579.

Salanti A, Staalsoe T, Lavstsen T, Jensen ATR, Sowa MPK, Arnot DE, et al. 2003. Selective upregulation of a single distinctly structured *var* gene in chondroitin sulphate A-adhering *Plasmodium falciparum* involved in pregnancy-associated malaria. *Mol Microbiol* **49**:179–191. doi: 10.1046/j.1365-2958.2003.03570.x.

Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, et al. 1995. Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**:101–10. doi: 10.1016/0092-8674(95)90056-X.

Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. 1995. The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82**:89–100. doi: 10.1016/0092-8674(95)90055-1.

Taylor HM, Kyes SA, Newbold CI. 2000. *Var* gene diversity in *Plasmodium falciparum* is generated by frequent recombination events. *Mol Biochem Parasitol* **110**:391–7. doi: 10.1016/S0166-6851(00)00286-3.

Trager W, Jensen JB. 1976. Human malaria parasites in continuous culture. *Science* **193**:673–675. doi: 10.1126/science.781840.

Tran TM, Li S, Doumbo S, Doumtabe D, Huang CY, Dia S, et al. 2013. An intensive longitudinal cohort study of Malian children and adults reveals no evidence of acquired immunity to *Plasmodium falciparum* infection. *Clin Infect Dis* **57**:40–7. doi: 10.1093/cid/cit174.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* **35**:W71–4. doi: 10.1093/nar/gkm306.

Ward CP, Clottey GT, Dorris M, Ji DD, Arnot DE. 1999. Analysis of *Plasmodium falciparum* PfEMP-1/*var* genes suggests that recombination rearranges constrained sequences. *Mol Biochem Parasitol* **102**:167–77. doi: 10.1016/S0166-6851(99)00106-1.

Warimwe GM, Keane TM, Fegan G, Musyoki JN, Newton CR, Pain A, et al. 2009. *Plasmodium falciparum var* gene expression is modified by host immunity. *Proc Natl Acad Sci USA* **106**:21 801–6. doi: 10.1073/pnas.0907590106.

Wunderlich G, Alves FP, Golnitz U, Tada MS, Camargo EF, Pereira-da Silva LH. 2005. Rapid turnover of *Plasmodium falciparum var* gene transcripts and genotypes during natural non-symptomatic infections. *Rev Inst Med Trop Sao Paulo* **47**:195–201. doi: 10.1590/S0036-46652005000400004.

Zhang Q, Zhang Y, Huang Y, Xue X, Yan H, Sun X, et al. 2011. From in vivo to in vitro: dynamic analysis of *Plasmodium falciparum var* gene expression patterns of patient isolates during adaptation to culture. *PLOS ONE* **6**:e20591. doi: 10.1371/journal.pone.0020591.

OPEN ACCESS Freely available online

PLoS one

Erasing the Epigenetic Memory and Beginning to Switch—The Onset of Antigenic Switching of *var* Genes in *Plasmodium falciparum*

Yair Fastman[1], Robert Noble[2], Mario Recker[2], Ron Dzikowski[1]*

1 Department of Microbiology & Molecular Genetics, The Kuvin Center for the Study of Infectious and Tropical Diseases, IMRIC, The Hebrew University-Hadassah Medical School, Jerusalem, Israel, **2** Department of Zoology, University of Oxford, Oxford, United Kingdom

Abstract

Antigenic variation in *Plasmodium falciparum* is regulated by transcriptional switches among members of the *var* gene family, each expressed in a mutually exclusive manner and encoding a different variant of the surface antigens collectively named PfEMP1. Antigenic switching starts when the first merozoites egress from the liver and begin their asexual proliferation within red blood cells. By erasing the epigenetic memory we created parasites with no *var* background, similar to merozoites that egress from the liver where no *var* gene is expressed. Creating a null-*var* background enabled us to investigate the onset of antigenic switches at the early phase of infection. At the onset of switching, *var* transcription pattern is heterogeneous with numerous genes transcribed at low levels including upsA *vars*, a subtype that was implicated in severe malaria, which are rarely activated in growing cultures. Analysis of subsequent *in vitro* switches shows that the probability of a gene to turn on or off is not associated with its chromosomal position or promoter type *per se* but on intrinsic properties of each gene. We concluded that *var* switching is determined by gene specific associated switch rates rather than general promoter type or locus associated switch rates. In addition, we show that fine tuned reduction in *var* transcription increases their switch rate, indicating that transcriptional perturbation can alter antigenic switching.

Citation: Fastman Y, Noble R, Recker M, Dzikowski R (2012) Erasing the Epigenetic Memory and Beginning to Switch—The Onset of Antigenic Switching of *var* Genes in *Plasmodium falciparum*. PLoS ONE 7(3): e34168. doi:10.1371/journal.pone.0034168

Editor: Thomas J. Templeton, Weill Cornell Medical College, United States of America

Received January 11, 2012; **Accepted** February 23, 2012; **Published** March 26, 2012

Copyright: © 2012 Fastman et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a grant from the German Israeli Foundation [997/2008]. RD is supported by the Marie Curie International Reintegration Grant (IRG) [203675], the United States-Israel Binational Science Foundation [2007350], and the Israeli Academy for Science [660/09]. RD is also supported by the Jacob and Lena Joels Memorial Foundation Senior Lectureship for Excellence in the Life and Medical Sciences. RN is supported by a BBSRC studentship. MR is funded by a Royal Society University Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rond@ekmd.huji.ac.il

Introduction

Plasmodium falciparum causes the deadliest form of human malaria and its virulence is attributed in part to the ability of the parasites to undergo antigenic variation of immunodominant cell surface proteins to avoid immune recognition [1,2,3].

The major antigenic ligands responsible for cytoadhesion are members of the highly polymorphic PfEMP1 [4] proteins expressed on the surface of infected red blood cells (iRBC). The placement of parasite-encoded proteins on the cell surface stimulates an antibody response that often clears the majority of infected cells from the circulation. However, small sub-populations of parasites switch expression to an alternative PfEMP1 variant [5] and thus avoid the antibody response and re-establish the infection. This process of antigenic variation is responsible for the persistent nature of the disease as well as the waves of parasitemia frequently observed in *P. falciparum* infections [6]. PfEMP1 are encoded by a multi-copy gene family named *var* [7] and each parasite genome has a repertoire of ~60 *var* genes [8]. Different parasite isolates typically have completely different *var* complements resulting in a virtually limitless repertoire of antigenic determinants [9]. Despite the high variability of the *var* gene family its genomic organization is well conserved across different parasite isolates. Approximately two thirds of the *var* genes are located in subtelomeric regions and the rest are found in the internal regions of the chromosomes. *Var* genes have been classified into several subgroups based on the sequence similarity of their 5′ UTRs [10,11] with subtelomeric *var* genes having either upsA or upsB promoters, whereas centrally located *var* genes have upsC promoters; *var2csa*, the *var* gene that was implicated in pregnancy associated malaria, has a unique upsE promoter. The proportions of these groups seem to be highly conserved between different parasite isolates [12,13].

Immune evasion of *P. falciparum* depends on its ability to express only a single *var* gene at a time in order to limit antigenic exposure [14,15], and to undergo intrinsic transcriptional switches between different *var* genes [16]. These switches need to occur often enough to generate parasite subpopulations that escape the human immune response, yet also need to be tightly controlled in order to avoid premature expenditure of the 60 different proteins. Horrocks *et al.* [17] were the first to suggest the existence of a hierarchy in *var* gene switching and showed that different *var* genes have different intrinsic switch rates [17]. These rates were later proposed to be associated with their chromosomal location [18]. More recently, models have demonstrated that this highly structured switching pattern, based on different switch rates and

activation probabilities, is crucial to prevent exhaustion of the antigenic repertoire in the early stages of infection [19].

Passage of *P. falciparum* through the liver of the mammalian host appears to reset the *var* gene expression pattern. A parasite population recovered from a patient only 11 days post infection indicated that most *var* genes were expressed at this stage [20]. Similar results were obtained when parasite populations recovered from human volunteers were put in cultures for a month prior to transcriptional analysis [21]. However, obtaining conclusive data on the entire *var* expression profile in early merozoites that emerge from the liver is challenging since experiments done in human volunteers requires treatment when low parasitemia is observed and such low parasitemia does allow recovering enough RNA for transcriptional profiling. Thus far no experimental system exists that enables studying the onset of antigenic switching in the very early merozoites that egress from the liver and enter the blood stream at the early phase of infection. This represents a fundamental deficiency in our ability to understand parasite biology since the ability to establish an initial blood stage infection is of paramount importance for disease transmission.

Previously, we have demonstrated a new procedure to silence all *var* genes in the genome by promoter titration [22]. However, transcription patterns were documented only 10 weeks after switching initiation and therefore *var* gene expression at the onset of switching remained unknown. The ability to silence all *var* genes in a reversible way and erase their epigenetic memory enables us for the first time to study *in vitro var* expression at the onset of antigenic switching and thus, mimicking the moment when parasites exit the liver and first invade RBCs. In addition, this procedure allowed us to fine tune interference with *var* transcription and provided new insights for the possible mechanisms that influence switch rates that regulate antigenic variation.

Results

Erasing epigenetic memory and beginning to switch: At the onset of antigenic switching var transcription pattern is heterogeneous

In order to investigate the onset of *var* gene switching activation we used promoter titration, briefly described below, to silence and erase the epigenetic memory of different parasite populations and thus created parasite lines with a null-*var* expression background similar to the initial merozoite population at the beginning of blood stage infection. The first clone that we erased expressed a subtelomeric *var* gene with an upsB promoter (E-1, expressing predominantly PFL0005w); the second line expressed a central *var* gene with upsC promoter (G-6, expressing predominantly PFD1005c). Both clones, each expressing a different PfEMP1 have similar growth rate which is also similar to the growth rate of transgenic parasite lines expressing no PfEMP1 (unpublished data). This fact argues that dynamics in transcription patterns of each population reflects switch rates rather than fitness cost of particular PfEMP1 expression. Each of these parasite lines was transfected with the pVBH plasmid by electroporation as described [23] and selected on 10 μg/ml blasticidin for populations that carry high copy numbers of active episomal *var* promoters. The competing pVBH episomes effectively silenced the predominant *var* genes, PFL0005w in the E-1 clone (Fig. 1A, second panel) and PFD1005c in the G-6 clone (Fig. 2A, second panel), by promoter titration. Both parasite populations transcribed high levels of the *bsd* gene from ~8–10 episomal copies of pVBH (Fig. 1B & 2B).

In order to allow *in vitro* switching to begin we removed drug pressure from the transfected populations, thus enabling the parasites to shed the pVBH episomes. qPCR of parasite populations growing for one month under no drug pressure indicated that these parasites no longer carried the pVBH plasmid and did not express *bsd* (Fig. 1B & 2B). We then followed the switching patterns of the entire *var* gene family of these parasite populations over a period of half a year (92 generations) using RT-qPCR as described [22]. RNA was extracted every two weeks for the first 3 months and then once a month until the end of the experiment. Two weeks (7 generations) after drug pressure was removed low levels of endogenous *var* transcripts could be detected, indicating that the parasites had started to actively express endogenous *var* genes. At this early time point, low transcription levels of many genes could be detected both in the E-1 and the G-6 population (Fig. 1A & 2A, third and fourth panels respectively), indicating a high activation probability of these genes from a null-*var* background. In both the E-1 and the G-6 populations transcription of upsA *var* genes was also detected in the early generations (up to one month) following switching initiation indicating that these genes can be activated from a null-*var* background at the early phase of infection. Moreover, at the onset of switching (up to one month after switching initiation) the relative proportion of transcript levels per *var* gene was similar among the different *var* subtypes, including upsA, both in the E-1 and G-6 populations (Fig. 1C & Fig. 2C).

Switching dynamics of different var subsets are independent of chromosomal location and promoter type per se

The dynamics of switching for each of the *var* subsets showed a bias in the relative transcription over-time towards particular genes in each population due to hierarchical switch order (Fig. 3). These results indicate that the switching patterns of parasites with no previous memory are similar to switches of clonal populations with previous epigenetic memory of *var* expression [17,18,19]. However, contrary to what was previously concluded, our gene specific analyses of the transcription time courses imply that promoter type or chromosomal location *per se* could not explain these dynamics. Instead, intrinsic properties within each particular gene seem to underlie the observed dynamics. Over time, there was an increase in the ratio of the upsC type in the G-6 populations and an increase in the upsB/C in the E-1 population. At the same time we observed a significant decrease in the relative levels of transcripts from the upsA promoter type in both populations (Fig. 1 & 2 and Fig. S1 & S2). These results indicate that while some particular *var* genes have a high probability to be activated and dominate the culture over time, other *var* genes are rarely activated. In order to test whether the genes that are rarely activated share common characteristics, we sorted all *var* genes by their transcription levels at each time point after the first month of switching initiation in both populations. We scored the genes in which their cumulative transcripts represent the lowest fifth percentile of the total *var* transcripts in the population. This was done for both populations at each time point. We found that 19 of the genes that are rarely activated are common in both populations (Fig. 4A). In addition, it appears that *var* genes with upsA promoters are highly represented in this group (7/8 upsA *var* genes) (Fig. 4B). Moreover, we noted that some of the upsC *var* genes (MAL7P1.50, MAL7P1.56, PFD0630c/0635c and PF08_0106) that are found in gene's clusters in internal chromosomal loci are also rarely activated even though the adjacent genes in the cluster are amongst the most highly activated genes in the genome. Similarly, we scored the highly activated *var* genes as the ten most active genes in both clones at each time point from one month throughout the entire

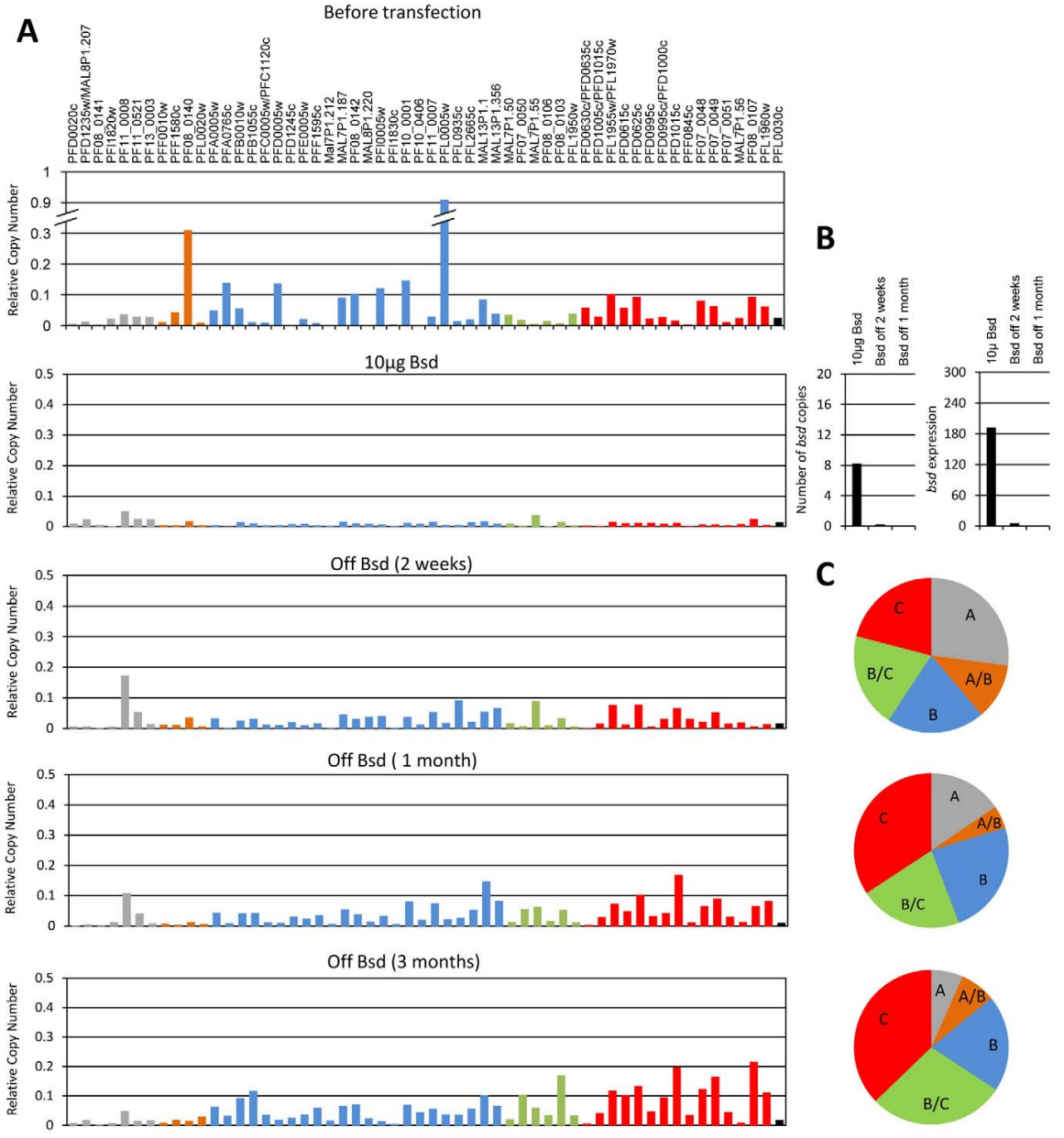


Figure 1. Erasing the epigenetic memory of the entire *var* gene family by promoter titration and the onset of switching of the E-1 population. (**A**), Steady state mRNA levels of each individual *var* gene are presented as relative copy number. All values are presented as relative copy number to the housekeeping genes arginyl-tRNA synthetase (PFL0900c), fructose biphosphate aldolase (PF14_0425), and actin (PFL2215). Each *var* gene is colored by its promoter type: Grey, upsA; Orange, upsA/B; Blue, upsB; Green, upsB/C; Red, upsC; Black, upsE. Transcription profiles are presented before transfection (upper panel), under titration using selection on 10 µg/ml blasticidin (second panel), two weeks, one month and three months (third to lower panel respectively) after drug removal and switching initiation. (**B**), pVBH copy numbers (left panel) and *bsd* transcription levels (right panel) of the E-1 populations under selection on 10 µg/ml blasticidin, as well as at two weeks and one month after drug removal. The number of pVBH copies was measured by qPCR on gDNA and transcription levels were measured by RT-qPCR on cDNA. (**C**), Dynamics in the relative transcription of each *var* gene subsets two weeks, one month and three months after drug removal (upper to lower pies respectively). The relative proportion of each *var* gene subsets, A, A/B, B (telomeric) and B/C, C (central) from the total transcripts is displayed for each population at each time points. The relative proportion of transcript was normalized to the number of genes belonging to each *var* subset.
doi:10.1371/journal.pone.0034168.g001

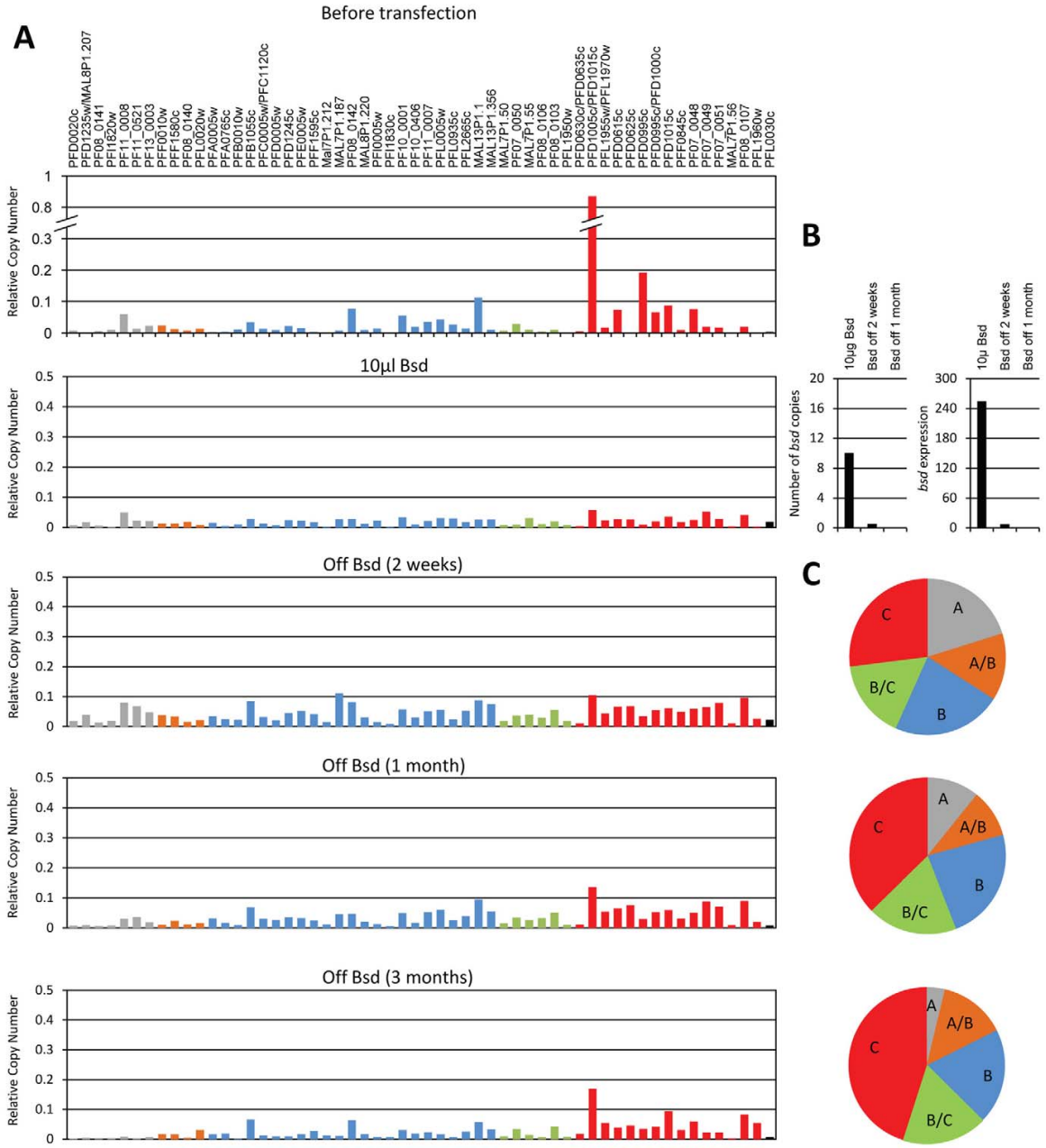


Figure 2. Erasing the epigenetic memory of the entire *var* gene family by promoter titration and the onset of switching of the G-6 population. (**A**), Steady state mRNA levels of each individual *var* gene are presented as relative copy number. All values are presented as relative copy number to the housekeeping genes arginyl-tRNA synthetase (PFL0900c), fructose biphosphate aldolase (PF14_0425), and actin (PFL2215). Each *var* gene is colored by its promoter type: Grey, upsA; Orange, upsA/B; Blue, upsB; Green, upsB/C; Red, upsC; Black, upsE. Transcription profiles are presented before transfection (upper panel), under titration using selection on 10 µg/ml blasticidin (second panel), two weeks, one month and three months (third to lower panel respectively) after drug removal and switching initiation. (**B**), pVBH copy numbers (left panel) and *bsd* transcription levels (right panel) of the G-6 populations under selection on 10 µg/ml blasticidin, as well as at two weeks and one month after drug removal. The number of pVBH copies was measured by qPCR on gDNA and transcription levels were measured by RT-qPCR on cDNA. (**C**), Dynamics in the relative transcription of each *var* gene subsets two weeks, one month and three months after drug removal (upper to lower pies respectively). The relative proportion of each *var* gene subsets, A, A/B, B (telomeric) and B/C, C (central) from the total transcripts is displayed for each population at each time points. The relative proportion of transcript was normalized to the number of genes belonging to each *var* subset.
doi:10.1371/journal.pone.0034168.g002

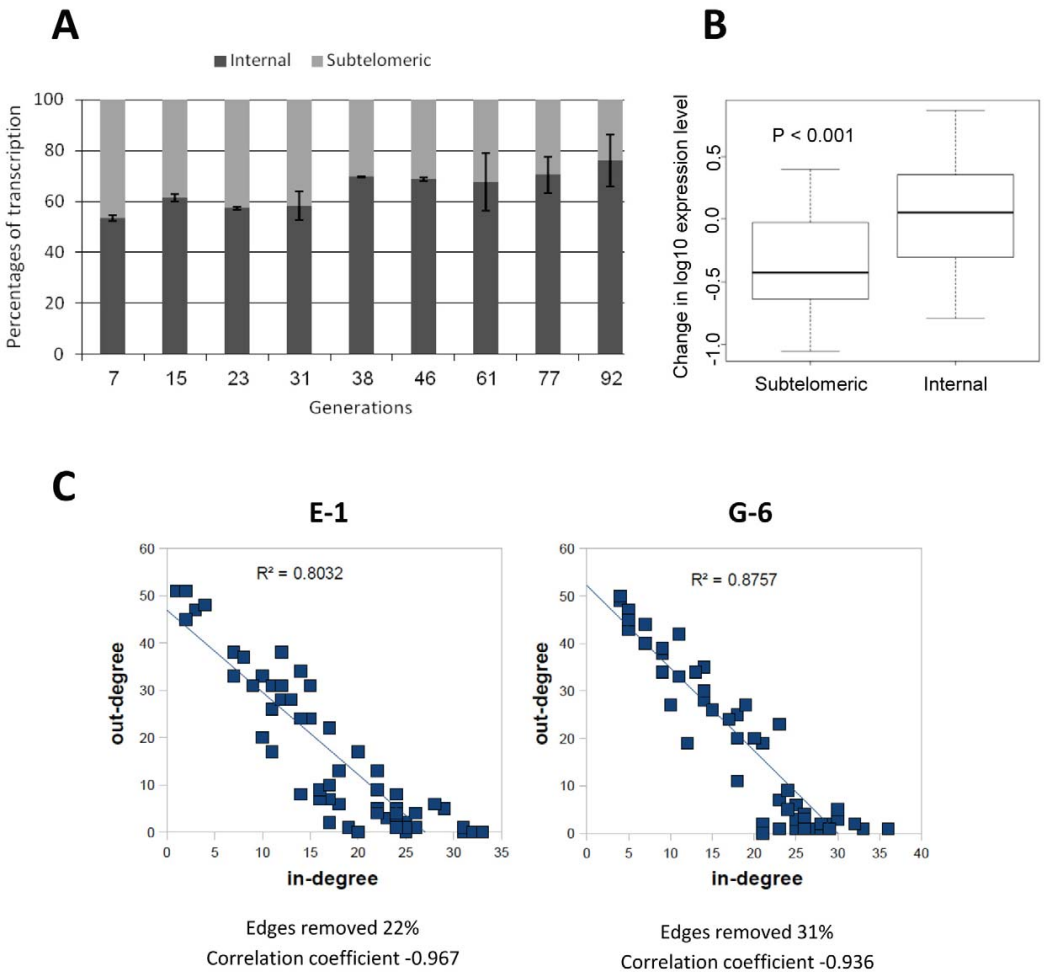


Figure 3. Switching dynamics of *var* genes in the following generations after erasing the epigenetic memory are biased. (**A**) Over time *var* transcription is biased toward internal *var* genes. The ratio of the mean total transcripts from internal and subtelomeric *var* genes (normalized to the number of genes in each sub-group) of the E-1 and G-6 cultures is presented for each time point. Error bars represents standard errors. Differences between the first and the last time points were tested by chi-square distribution and had values of $P<0.05$ (**B**), Comparison of overall transcript levels between the first 4 time points and the last 5 time points shows significant decrease in transcription levels of subtelomeric genes over time ($P = 0.005$; Mann-Whitney U test) (**C**), Analyzing the transcription data for underlying switch dynamics of each individual *var* gene over half a year after switching initiation demonstrates biased activation and silencing patterns. The *var* gene repertoire can be envisaged as a switching network whose nodes represent individual *var* genes and edges the probability of a switch between them [19]. The strong negative correlation between a variant's in - and out - degree persists even after removing a large proportion of the network, thus supporting the non-random pattern of *var* gene switching in our data.
doi:10.1371/journal.pone.0034168.g003

experiment. We found that 12 of the most active genes are common for both clones and that most of them are upsC (Fig. 4C). This suggests that a large proportion of the *var* repertoire has low probability to be expressed *in vitro*. In addition, while most upsA *var* genes are rarely activated, in other promoter types some genes are highly activated and some are rarely activated. This implies that switch rates are determined by gene specific associated mechanisms rather than a general promoter type or chromosomal locus associated mechanism. Therefore, we concluded that the switching properties of each *var* gene are independent of its chromosomal location or promoter type *per se* (Fig. 5).

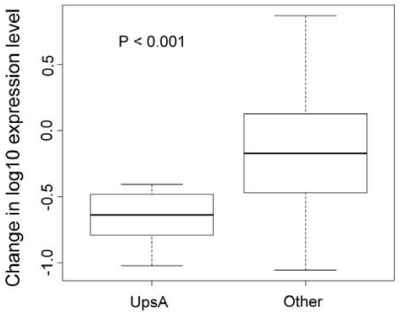
Partial var repression promotes switching but does not erase the epigenetic memory

Passing a certain threshold of transcriptional down-regulation of *var* genes by promoter titration has been shown to result in the loss of their epigenetic memory [22]. However, the effect of only a partial repression of the endogenous gene by low copy number of competing episomes is unknown. We reasoned that if the maintenance of epigenetic memory is associated with the act of transcription *per se* then it is possible that transcriptional interference by fine tuned promoter titration will trigger faster switching to other genes. We therefore analyzed the entire *var* transcription profile of two parasite lines with low switch rate: the

A

Rare *var* genes in both cultures	
Gene	Promoter type
PFD0020c	A
PFD1235w/MAL8P1.207	A
PFI1820w	A
PF08_0141	A
PF13_0003	A
PF11_0521	A
PF08_0140	B/A
Mal7P1.212	B
MAL8P1.220	B
PFI0005w	B
PFI1830c	B
PFL0935c	B
PFA0765c	B
MAL7P1.50	B/C
PF08_0106	B/C
PFD0630c/PFD0635c	C B/C
MAL7P1.56	C

B



C

Common *var* genes in both cultures	
Gene	Promoter type
PFD0615c	C
PFD0625c	C
PFD1015c	C
PF07_0048	C
PF08_0107	C
PFL1955w/PFL1970w	B/C C
PF07_0050	B/C
PF08_0103	B/C
PFB1055c	B
PF08_0142	B
PF13_0001	B

Figure 4. Highly and rarely activated *var* genes. *var* gene were sorted by their relative mRNA levels to the total *var* transcript of the entire family, at each time point for all three populations. The *var* genes that were common among both populations at least twice are displayed. (**A**), *var* genes that are rarely activated. Note that among the 19 *var* genes that are rarely activated, 14 are subtelomeric. In addition, *var* genes with upsA promoters are highly represented in this group. (**B**), UpsA *var* genes become less transcribed over time. Comparison of overall upsA transcript levels between the first 4 time points and the last 5 time points shows significant decrease in transcription levels of upsA *var* genes over time ($P = 0.007$; Mann-Whitney U test (**C**), *var* genes that have high probability to be activated. Among the 12 *var* genes that are commonly activated internal *vars* with upsC promoters are highly represented.
doi:10.1371/journal.pone.0034168.g004

G-6 and CSA-selected populations after transfection with pVBH and subsequent selection on low dose of only 2 μg/ml blasticidin. Under this low drug pressure these parasite populations carried only two copies of the competing pVBH episome (Fig. 6) and transcription levels of PFD1005c and PFL0030c were down-regulated by 2–3 fold; however, both remained the dominant *var* genes in the G-6 and the CSA-selected populations respectively.

In both clones the reduced transcription levels of the dominant *var* gene was associated with the appearance of detectable transcripts from other endogenous genes (Fig. 6). In both the G-6 and the CSA-selected populations that were not transfected with pVBH and kept growing in culture during the course of the experiment, no apparent switches were observed and their transcription profiles remained similar to the original clonal populations (Fig. S3). These results indicate that the observed switches to other *var* genes were specific to the populations in which transcription levels of the dominant *var* genes were reduced by titration. Moreover, the fact that under such low levels of promoter titration, PFD1005c and PFL0030c remained the dominant *var* genes suggests that the epigenetic memory that marks these genes as the active genes is still maintained. To test this hypothesis we simply removed drug pressure and let these populations remove the competing episomes. We found that contrary to the promoter titration with multi-copy pVBH that completely erased the epigenetic memory and resulted in heterogenous *var* transcription patterns, both the G-6 and the CSA-selected populations reverted back to clonal transcription patterns in which each of the populations was expressing exclusively either PFD1005c or PFL0030c respectively (Fig. 6). These results suggest that the epigenetic memory is weakened when transcription levels are repressed below the threshold that leads to complete silencing and that this partial repression can induce higher switch rates.

Discussion

Antigenic variation has evolved in many pathogens as a successful mechanism to evade host immunity and to sustain

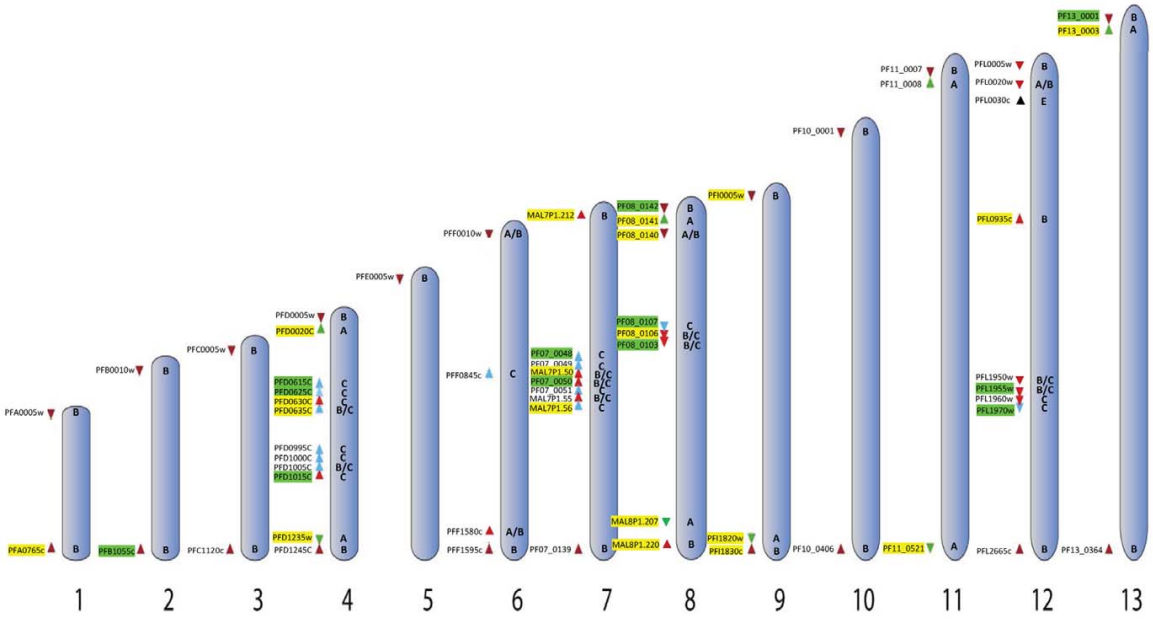


Figure 5. Schematics of the *var* gene family in the NF54 genome presenting the chromosomal location of *var* gene that are highly or rarely activated. Chromosomes containing *var* genes are numbered, and each arrowhead represents a single *var* gene (modified from: [46]). Accession numbers and promoter type are marked for each individual gene. *var* genes that have high probability to be activated are displayed in green and *var* genes that are rarely activated are in yellow.
doi:10.1371/journal.pone.0034168.g005

long-term chronic infections [24]. This is especially important for vector-born pathogens such as *P. falciparum* in order to ensure the continuity of transmission in regions where mosquito abundance is seasonal. Antigenic variation created by *P. falciparum* is tightly regulated by continuous switches between expressions of different *var* genes. It has been proposed that different *var* genes turn on and off at different rates [18,19,25]. These studies were able to demonstrate that the history of expression in a clonal population expressing a particular *var* gene would affect the future switching pattern in the population and proposed a model of hierarchical switching order which would help the parasite to avoid fast exposure of its antigenic repertoire. However, since *var* genes are expressed exclusively during the intra-erythrocytic development cycle (IDC) and not in liver stages, the switching patterns that occur at the onset of malaria when the first merozoites burst out of hepatocytes and start replicating within iRBC begin from a null-*var* background rather than from parasites expressing any particular *var* gene. Here we used a controlled *in vitro* system to erase the epigenetic memory and create a null-*var* background in blood stages parasites. This allowed us to study the switching patterns from the onset of switching and over a period of six months. Our results showing a heterogeneous *var* expression at the onset of switching is similar to the parasite population recovered from a patient only 11 days post infection [20] and may indicate that merozoites leaving the liver with a null-*var* background employ similar epigenetic marks to intra-erythrocytic parasites.

We have previously proposed that in clonal populations, switches in expression of subtelomeric *var* genes are more frequent than the switches of *var* genes that are located in internal regions of the chromosome, which led to the conclusion that the intrinsic switch rate of a gene is correlated with its chromosomal position [18]. The results of our current study imply that many of the genes that are located in internal loci, mainly those with upsC promoters, have a high probability to be activated. However, some of the most highly activated genes in both clones are subtelomeric *var* genes. In addition, we found that genes with low activation probability are located mostly in subtelomeric regions, but a few are found in the internal gene clusters and have either upsC or upsB/C promoters. This clearly indicates that a simple model in which the chromosomal location or the type of promoter *per se* determines the fate of a *var* gene does not explain the differences in the probability of a particular *var* gene to get activated. Rather, this indicates for a gene specific associated mechanism that determines its switch rate. It is still possible that the odds of each particular gene to get activated depends on its position in the overall switching network as previously suggested [19]. These authors also demonstrated that the hierarchies of switching often show patterns that favor switching to genes that are located at different chromosomes. Such switching hierarchies could also explain why the *var* subtypes of genes that were predominantly transcribed in cultures after half a year in both the E-1 and G-6 clones were similar to those of the original clonal populations before erasing their memory. We show that some of the genes that are highly activated and those that are rarely transcribed are located adjacent to each other at the same chromosomal locus. Remodeling the chromatin structure was implicated in *var* gene regulation; however, having the two extremes, highly and rarely activated genes, at the same locus suggests that switching does not necessarily occur through the spreading of euchromatin from the active gene to the adjacent gene. In addition, our results support a model by which each *var* gene is an autonomous unit that contains all DNA elements that

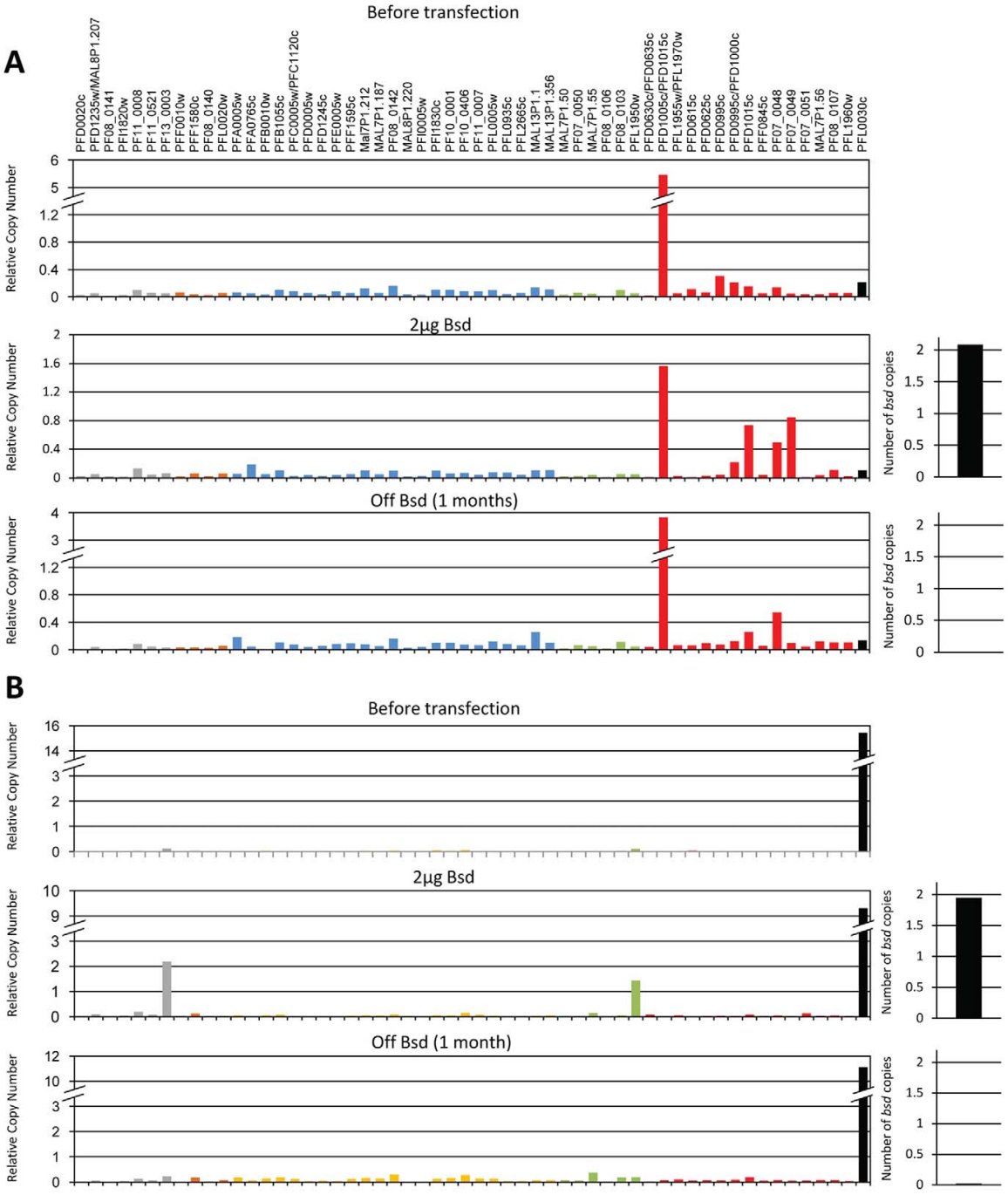


Figure 6. Low levels of promoter titration that reduce transcription promote switching but do not erase the epigenetic memory. Two clonal populations G-6 (**A**) and CSA (**B**) (upper panels) were transfected with the plasmid pVBH and selected on low dose of 2 µg/ml blasticidin (middle panels). These populations carried two copies of the competing episomes (right diagram). Under drug selection the transcription levels of the genes that were exclusively expressed prior to titration: PFD1005c and PFL0030c were reduced in both the G-6 and the E-1 respectively, while transcription of other *var* genes was detected. A month after the removal of drug pressure, both populations returned to their previous clonal expression patterns (lower panels). All values are presented as relative copy number to the housekeeping gene arginyl-tRNA synthetase (PFL0900c). *var* genes are colored by promoter type: Grey, upsA; Orange, upsA/B; Blue, upsB; Green, upsB/C; Red, upsC; Black, upsE.
doi:10.1371/journal.pone.0034168.g006

are required to regulate antigenic switching and mutually exclusive expression. Indeed, recombination in the *var2csa* regulatory intron had altered its switch rate [26].

In order to silence *var* genes and to erase the epigenetic memory it has been proposed that the level of transcriptional repression needs to be below a certain threshold. This implied that active transcription is required for the maintenance of the epigenetic memory that regulates *var* gene expression [22] possibly through the involvement of PolII [27]. However, the exact mechanism by which the act of transcription is involved in maintaining the epigenetic memory is unknown. It is possible that by reducing the level of transcription of a particular gene as demonstrated here, the maintenance of the epigenetic marks required for epigenetic memory were also weakened which in turn increased the probability of expression to switch to another gene. This could explain why in the populations subjected to fine tuned transcriptional repression additional *var* transcripts were detected, since there is no evidence that the mechanisms of mutually exclusive expression were interrupted in those parasites. How these parasite populations revert back to the original dominant gene is yet to be discovered, however, similar phenomenon was recently reported in parasite populations switching *in vitro* without prior manipulation of the epigenetic memory [25]. Further investigation is needed to explore the possibility that some *in vivo* situations may perturb *var* transcription and alter antigenic switching.

Various attempts have been made to link *var* gene expression with the clinical manifestations of malaria and to investigate the role of specific PfEMP1 variants with pathogenesis. Profiling *var* expression in field isolates taken from Kenyan children showed an association between the abundance of specific *var* transcripts and the "rosetting" properties of the iRBC [28]. Cerebral malaria was also associated with specific PfEMP1 expression. Parasites that were isolated from organs of children dying from malaria showed clear dominance of a small number of certain *var* transcripts in parasites that were sequestered in the brain. This suggests that PfEMP1 proteins encoded by only a small number of *var* genes are responsible for sequestration in the brain microvasculature [29], possibly through specific binding to ICAM1 [30,31]. Furthermore, using *in vitro* selection on sera taken from children with severe malaria it was shown that parasites that cause severe childhood malaria preferentially express PfEMP1 variants that are encoded by upsA *var* genes [32]. UpsA *var* genes were found to be predominantly transcribed in children with cerebral malaria [33] and a particular upsA *var* gene (PF13_0003) was recently linked to the "rosetting" phenotype [34]. We found that upsA *var* genes are activated at the onset of switching similar to other *var* subtypes but quickly switch away in growing cultures. Moreover, transcripts from these genes did not appear in any of the clones even after half a year in culture suggesting a very low on rate outside the human host. However, we did not obtain any upsA expressing clone and therefore could not investigate the effect of promoter titration on such parasite line. UpsA *var* genes have been shown to switch away during adaptation of wild-type parasites to culture [35] and seem to be rarely activated in clones of the IT line [36]. Recent computational analysis demonstrated that upsA *var* genes maintain a high level of conservation of specific PfEMP1 blocks across several genetic lines [37]. The low probability of upsA *var* genes to be activated and their association with severe malaria might offer a possible explanation for this relatively high degree of conservation. In chronic infections upsA *vars* are rarely activated and therefore are not exposed to immune pressure. However, when upsA *var* genes are activated in non-immune individuals they would lead to severe malaria that would either have fatal consequences or immunity acquisition. This explanation is supported by a clinical study that showed that PfEMP1 variants encoded by upsA *var* genes with cytoadhesive properties that confer a growth advantage dominate infections in naïve children in the absence of a protective immune response [38], while the cytoadhesive properties of these variants lead to severe malaria. The authors suggested that rapid immunity acquisition to this relatively conserved group may explain why children are less susceptible to severe malaria after few infections [39]. While it is possible that at the onset of switching activation of upsA *var* genes could lead to severe malaria in the absence of immunity by cytoadhesive selection. Our results may suggest that evolutionary evolvement of the high "off" rates of these *var* genes enables the parasites to evade immune attack in pre-immune patient by pre-programmed rapid switching to less conserved PfEMP1. Understanding antigenic gene expression at the onset of malaria as well as host-parasite interactions that influence parasites' establishment and pathogenicity at the early phase of infection will hopefully lead to the development of new intervention strategies.

Materials and Methods

Parasite Culture

All parasites used were derivatives of the NF54 parasite line and were cultivated at 5% hematocrit in RPMI 1640 medium, 0.5% Albumax II (Invitrogen), 0.25% sodium bicarbonate, and 0.1 mg/ml gentamicin. Parasites were incubated at 37°C in an atmosphere of 5% oxygen, 5% carbon dioxide and 90% nitrogen. Clonal populations were obtained by limiting dilutions as previously described [18]. Each population expressed a different *var* gene: the G6 clone expressed PFD1005c; the E1, predominantly expressed PFL0005w and CSA exclusively express PFL0030c.

Parasite cultures were synchronized using percoll/sorbitol gradient centrifugation as previously described [40,41]. Briefly, infected RBCs were layered on a step gradient of 40%/70% percoll containing 6% sorbitol. The gradients were then centrifuged at 12,000 g for 20 minutes at room temperature. Highly synchronized, late stage parasites were recovered from the 40%/70% interphase, washed twice with complete culture media and placed back in culture.

Plasmid Construction and Parasite Transfection

The plasmids pVBH, is described [22]. Parasites were transfected as described [42]. Briefly, 0.2 cm electroporation cuvettes were loaded with 0.175 ml of erythrocytes and 100 µg of plasmid DNA in incomplete cytomix solution. Stable transfectants were initially selected on 2 µg/ml blasticidin (Invitrogen, USA). In order to obtain parasites carrying large plasmid copy numbers, these cultures were then subjected to 10–15 µg/ml blasticidin for two weeks or until normal growth rates were observed. At this stage parasite were grown to 5% parasitemia in 20 ml culture for RNA extraction.

Genomic DNA extraction, RNA extraction and cDNA synthesis

Genomic DNA was extracted as described [22]. Briefly, iRBCs were pelleted and lysed with Saponin. Parasites from 20 ml cultures were pelleted, washed with PBS, and taken up in 200 µl TSE buffer (100 mM NaCl, 50 mM EDTA, 20 mM Tris, pH 8), 40 µl of 10% SDS and 20 µl 6 M $NaClO_4$. This suspension was rocked for twelve hours and genomic DNA was extracted with phenol/chloroform. The resulting DNA was taken up in 100 µl dH_2O.

RNA extraction and cDNA synthesis was performed as described [16]. Briefly, RNA was extracted from synchronized

parasite cultures at 18 hours post invasion, corresponding with the peak RNA expression levels for determination of *var* transcript levels. RNA was extracted with the TRIZOL LS Reagent® as described [43] and purified on PureLink column (Invitrogen) according to manufacturer's protocol. Isolated RNA was then treated with Deoxyribonuclease I® (Invitrogen) to degrade contaminating gDNA. cDNA synthesis was performed from 800 ng total RNA with Superscript II Rnase H reverse transcriptase ® (Invitrogen) with random primers ® (Invitrogen) as described by the manufacturer.

Real-time RT-qPCR

For RT-qPCR reactions to detect transcription from all *var* genes present in the 3D7 genome, we used the primer sets published by Salanti et. al. [44] with few modifications published in [45]. Transcript copy numbers were determined using the formula $2^{-\Delta\Delta CT}$ as described in the Applied Biosystems User Bulletin 2 using NF54 gDNA as the calibrator. Specifically, relative copy number was calculated as 2 exponential negative ((Ct target gene in cDNA – Ct reference gene in cDNA)-(Ct target gene in gDNA-Ct target gene in gDNA)). Copy numbers of episomal promoters were calculated using RT-qPCR of gDNA and comparing the ΔCT of *bsd* to that of the single copy housekeeping gene arginyl-tRNA synthetase (PFL0900c). This gene as well as P60-seryl-tRNA synthetase (PF07_0073), P61-fructose bisphosphate aldolase (PF14_0425), P100-actin (PFL2215), and glutaminyl-tRNA synthetase (PF13_0170) were used as control genes in all RT-qPCR assays as described [45]. The reference gene used for result presentations in all of the presented graphs presented in the manuscript is arginyl-tRNA synthetase. All RT-qPCR assays were performed at least in duplicate for each template with no apparent differences, and each experiment was completed at least twice in its entirety, again with no significant differences.

Var gene transcription analysis

We used the same computational method previously employed by Recker et al., [19] to analyze the long-term transcription profiles. In brief, *var* gene switching can be described simply by the rate an active gene becomes silent, r_i, and the probability of a switch from gene i to gene j, b_{ij}.The method then determines the switch rates and switch biases that best fit the experimental data and which can then be expressed as an overall, directed switching network where the nodes represent the individual genes and the edges the net transition between them, i.e. the difference $r_i{*}b_{ij} - r_j{*}b_{ji}$. Because of the deterministic nature of this method the resulting network is fully connected and will naturally exhibit a negative correlation between a node's out-degree, i.e. the number of genes it switches to, and its in-degree, i.e. the number of genes that switch towards this gene. To investigate the robustness of this correlation we therefore progressively removed edges with the weakest transition rates and compared the resulting correlation to a randomly connected network with a comparable number of connections removed.

Supporting Information

Figure S1 Switching dynamics of parasite population expressing a subtelomeric *var* gene (E-1) in the following generations after erasing its epigenetic memory. (**A**), Steady state mRNA levels of each individual *var* gene are presented as relative copy number. Each *var* gene is colored by its promoter type: Grey, upsA; Orange, upsA/B; Blue, upsB; Green, upsB/C; Red, upsC; Black, upsE. Transcription patterns are presented at different time points after drug removal and switching initiation: two months (upper panel), four months (second panel), and six months (lower panel). Transcription levels were measured by RT-qPCR. All values are presented as relative copy number the to the housekeeping gene arginyl-tRNA synthetase (PFL0900c). (**B**), The relative proportion of transcripts of each *var* gene subsets, A, A/B, B, (telomeric) and B/C, C, (central) from the total *var* transcripts is displayed for each time points as a pie chart on the right. The relative proportion of transcript was normalized to the number of genes belonging to each *var* subset.
(TIF)

Figure S2 Switching dynamics of parasite population expressing an internal *var* gene (G-6) in the following generations after erasing its epigenetic memory. (**A**), Steady state mRNA levels of each individual *var* gene are presented as relative copy number. Each *var* gene is colored by its promoter type: Grey, upsA; Orange, upsA/B; Blue, upsB; Green, upsB/C; Red, upsC; Black, upsE. Transcription patterns are presented at different time points after drug removal and switching initiation: two months (upper panel), four months (second panel), and six months (lower panel). Transcription levels were measured by RT-qPCR. All values are presented as relative copy number to the housekeeping gene arginyl-tRNA synthetase (PFL0900c). (**B**), The relative proportion of transcripts of each *var* gene subsets, A, A/B, B, (telomeric) and B/C, C, (central) from the total *var* transcripts is displayed for each time points as a pie chart on the right. The relative proportion of transcript was normalized to the number of genes belonging to each *var* subset.
(TIF)

Figure S3 *Var* transcription patterns of the un-transfected clonal populations after one month of *in vitro* switching. Both the untransfected G-6 (**A**) and the CSA-selected (**B**) populations used for the fine tune down regulation experiment on 2 μg/ml blasticidin presented in figure 6 were kept in culture during the course of the experiment. Steady state mRNA levels measured by RT-qPCR on cDNA of each individual *var* gene are presented as relative copy number. Each *var* gene is colored by its promoter type: Grey, upsA; Orange, upsA/B; Blue, upsB; Green, upsB/C; Red, upsC; Black, upsE. All values are presented as relative copy number to the housekeeping genes arginyl-tRNA synthetase (PFL0900c).
(TIF)

Author Contributions

Conceived and designed the experiments: RD YF. Performed the experiments: RD YF. Analyzed the data: RD YF MR RN. Contributed reagents/materials/analysis tools: MR RN. Wrote the paper: RD.

References

1. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI (2005) The global distribution of clinical episodes of Plasmodium falciparum malaria. Nature 434: 214–217.
2. Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415: 673–679.
3. Dzikowski R, Templeton TJ, Deitsch K (2006) Variant antigen gene expression in malaria. Cell Microbiol 8: 1371–1381.
4. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, et al. (1995) Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82: 77–87.

5. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, et al. (1995) Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82: 101–110.
6. Miller LH, Good MF, Milon G (1994) Malaria Pathogenesis. Science 264: 1878–1883.
7. Su X, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JV, et al. (1995) A large and diverse gene family (var) encodes 200–350 kD proteins implicated in the antigenic variation and cytoadherence of Plasmodium falciparum -infected erythrocytes. Cell 82: 89–100.
8. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498–511.
9. Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, et al. (2007) Population genomics of the immune evasion (var) genes of Plasmodium falciparum. PLoS Pathog 3: e34.
10. Lavstsen T, Salanti A, Jensen ATR, Arnot DE, Theander TG (2003) Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions. Malaria Journal 2: 27.
11. Kraemer SM, Smith JD (2003) Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family. Molecular Microbiology 50: 1527–1538.
12. Kyes SA, Kraemer SM, Smith JD (2007) Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family. Eukaryot Cell 6: 1511–1520.
13. Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, et al. (2007) Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates. BMC Genomics 8: 45.
14. Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, et al. (1998) Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. EMBO J 17: 5418–5426.
15. Chen Q, Fernandez V, Sundstrom A, Schlichtherle M, Datta S, et al. (1998) Developmental selection of var gene expression in Plasmodium falciparum. Nature 394: 392–395.
16. Dzikowski R, Frank M, Deitsch K (2006) Mutually Exclusive Expression of Virulence Genes by Malaria Parasites Is Regulated Independently of Antigen Production. PLoSPathog 2: e22.
17. Horrocks P, Pinches R, Christodoulou Z, Kyes S, Newbold C (2004) Variable var transition rates underlie antigenic variation in malaria. ProcNatlAcadSciUSA.
18. Frank M, Dzikowski R, Amulic B, Deitsch K (2007) Variable switching rates of malaria virulence genes are associated with chromosomal position. Mol Microbiol 64: 1486–1498.
19. Recker M, Buckee CO, Serazin A, Kyes S, Pinches R, et al. (2011) Antigenic Variation in Plasmodium falciparum Malaria Involves a Highly Structured Switching Pattern. PLoS Pathog 7: e1001306.
20. Wang CW, Hermsen CC, Sauerwein RW, Arnot DE, Theander TG, et al. (2009) The Plasmodium falciparum var gene transcription strategy at the onset of blood stage infection in a human volunteer. Parasitol Int 58: 478–480.
21. Lavstsen T, Magistrado P, Hermsen CC, Salanti A, Jensen AT, et al. (2005) Expression of Plasmodium falciparum erythrocyte membrane protein 1 in experimentally infected humans. MalarJ 4: 21.
22. Dzikowski R, Deitsch KW (2008) Active Transcription is Required for Maintenance of Epigenetic Memory in the Malaria Parasite Plasmodium falciparum. J Mol Biol.
23. Wu Y, Sifri CD, Lei HH, Su X, Wellems TE (1995) Transfection of Plasmodium falciparum within human red blood cells. Proceedings of the National Academy of Sciences USA 92: 973–977.
24. Deitsch KW, Lukehart SA, Stringer JR (2009) Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. Nat Rev Microbiol 7: 493–503.
25. Enderes C, Kombila D, Dal-Bianco M, Dzikowski R, Kremsner P, et al. (2011) Var Gene promoter activation in clonal Plasmodium falciparum isolates follows a hierarchy and suggests a conserved switching program that is independent of genetic background. J Infect Dis 204: 1620–1631.
26. Duffy MF, Byrne TJ, Carret C, Ivens A, Brown GV (2009) Ectopic recombination of a malaria var gene during mitosis associated with an altered var switch rate. J Mol Biol 389: 453–469.
27. Kishore SP, Perkins SL, Templeton TJ, Deitsch KW (2009) An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases. J Mol Evol 68: 706–714.
28. Bull PC, Berriman M, Kyes S, Quail MA, Hall N, et al. (2005) Plasmodium falciparum variant surface antigen expression patterns during malaria. PLoS Pathog 1: e26.
29. Montgomery J, Mphande FA, Berriman M, Pain A, Rogerson SJ, et al. (2007) Differential var gene expression in the organs of patients dying of falciparum malaria. MolMicrobiol 65: 959–967.
30. Kraemer SM, Smith JD (2006) A family affair: var genes, PfEMP1 binding, and malaria disease. Curr Opin Microbiol 9: 374–380.
31. Ochola LB, Siddondo BR, Ocholla H, Nkya S, Kimani EN, et al. (2011) Specific Receptor Usage in *Plasmodium falciparum* Cytoadherence Is Associated with Disease Outcome. PLoS ONE 6: e14741.
32. Jensen ATR, Magistrado P, Sharp S, Joergensen L, Lavstsen T, et al. (2004) Plasmodium falciparum associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group A var genes. Journal of Experimental Medicine 199: 1179–1190.
33. Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, et al. (2006) Differential var gene transcription in Plasmodium falciparum isolates from patients with cerebral malaria compared to hyperparasitaemia. Mol Biochem Parasitol 150: 211–218.
34. Vigan-Womas I, Guillotte M, Juillerat A, Vallieres C, Lewit-Bentley A, et al. (2011) Allelic diversity of the Plasmodium falciparum erythrocyte membrane protein 1 entails variant-specific red cell surface epitopes. PLoS One 6: e16544.
35. Peters JM, Fowler EV, Krause DR, Cheng Q, Gatton ML (2007) Differential changes in Plasmodium falciparum var transcription during adaptation to culture. J Infect Dis 195: 748–755.
36. Janes JH, Wang CP, Levin-Edens E, Vigan-Womas I, Guillotte M, et al. (2011) Investigating the Host Binding Signature on the Plasmodium falciparum PfEMP1 Protein Family. PLoS Pathog 7: e1002032.
37. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T (2010) Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes–divide and conquer. PLoS Comput Biol 6.
38. Warimwe GM, Keane TM, Fegan G, Musyoki JN, Newton CR, et al. (2009) Plasmodium falciparum var gene expression is modified by host immunity. Proc Natl Acad Sci U S A 106: 21801–21806.
39. Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C (1999) Immunity to non-cerebral severe malaria is acquired after one or two infections. Nat Med 5: 340–343.
40. Aley SB, Sherwood JA, Howard RJ (1984) Knob-positive and knob-negative Plasmodium falciparum differ in expression of a strain-specific malarial antigen on the surface of infected erythrocytes. JExpMed 160: 1585–1590.
41. Calderwood MS, Gannoun-Zaki L, Wellems TE, Deitsch KW (2003) Plasmodium falciparum var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron. Journal of Biological Chemistry 278: 34125–34132.
42. Deitsch KW, Driskill CL, Wellems TE (2001) Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. Nucleic Acids Research 29: 850–853.
43. Kyes S, Pinches R, Newbold C (2000) A simple RNA analysis method shows var and rif multigene family expression patterns in Plasmodium falciparum. MolBiochemParasitol 105: 311–315.
44. Salanti A, Staalsoe T, Lavstsen T, Jensen AT, Sowa MP, et al. (2003) Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering Plasmodium falciparum involved in pregnancy-associated malaria. Mol Microbiol 49: 179–191.
45. Frank M, Dzikowski R, Costantini D, Amulic B, Berdougo E, et al. (2006) Strict pairing of var promoters and introns is required for var gene silencing in the malaria parasite Plasmodium falciparum. J Biol Chem 281: 9942–9952.
46. Deitsch KW, Hviid L (2004) Variant surface antigens, virulence genes and the pathogenesis of malaria. Trends Parasitol 20: 562–566.

Bibliography

Agur, Z., Abiri, D. & Vanderploeg, L.H.T. (1989) Ordered appearance of antigenic variants of african trypanosomes explained in a mathematical-model based on a stochastic switch process and immune-selection against putative switch intermediates. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 9626–9630.

Aitcheson, N., Talbot, S., Shapiro, J., Hughes, K., Adkin, C., Butt, T., Sheader, K. & Rudenko, G. (2005) Vsg switching in trypanosoma brucei: antigenic variation analysed using rnai in the absence of immune selection. *Molecular microbiology*, **57**, 1608–1622.

Alizon, S. & van Baalen, M. (2008) Multiple infections, immune dynamics, and the evolution of virulence. *The American Naturalist*, **172**, E150–68.

Anderson, R.M. & May, R.M. (1982) Coevolution of hosts and parasites. *Parasitology*, **85 (Pt 2)**, 411–426.

Artzy-Randrup, Y., Rorick, M.M., Day, K., Chen, D., Dobson, A.P. & Pascual, M. (2012) Population structuring of multi-copy, antigen-encoding genes in plasmodium falciparum. *eLife*, **1**, e00093.

Avril, M., Tripathi, A.K., Brazier, A.J., Andisi, C., Janes, J.H., Soma, V.L., Jr, D.J.S., Bull, P.C., Stins, M.F. & Smith, J.D. (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**, E1782–90.

Avril, M., Brazier, A.J., Melcher, M., Sampath, S. & Smith, J.D. (2013) Dc8 and dc13 var genes associated with severe malaria bind avidly to diverse endothelial cells. *PLoS pathogens*, **9**, e1003430–e1003430.

Babiker, H.A., Schneider, P. & Reece, S.E. (2008) Gametocytes: insights gained during a decade of molecular monitoring. *Trends in parasitology*, **24**, 525–530.

Barbour, A.G., Dai, Q., Restrepo, B.I., Stoenner, H.G. & Frank, S.A. (2006) Pathogen escape from host immunity by a genome program for antigenic variation. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 18290–18295.

Barry, A.E., Leliwa-Sytek, A., Tavul, L., Imrie, H., Migot-Nabias, F., Brown, S.M., McVean, G.A.V. & Day, K.P. (2007) Population genomics of the immune evasion (var) genes of plasmodium falciparum. *Plos Pathogens*, **3**, e34.

Baruch, D.I., Ma, X.C., Singh, H.B., Bi, X., Pasloske, B.L. & Howard, R.J. (1997) Identification of a region of pfemp1 that mediates adherence of plasmodium falciparum infected erythrocytes to cd36: conserved function with variant sequence. *Blood*, **90**, 3766–3775.

Baruch, D.I., Pasloske, B.L., Singh, H.B., Bi, X., Ma, X.C., Feldman, M., Taraschi, T.F. & Howard, R.J. (1995) Cloning the p. falciparum gene encoding pfemp1,

a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, **82**, 77–87.

Beeson, J.G., Rogerson, S.J., Cooke, B.M., Reeder, J.C., Chai, W., Lawson, A.M., Molyneux, M.E. & Brown, G.V. (2000) Adhesion of plasmodium falciparum-infected erythrocytes to hyaluronic acid in placental malaria. *Nature medicine*, **6**, 86–90.

Blomqvist, K., Normark, J., Nilsson, D., Ribacke, U., Orikiriza, J., Trillkott, P., Byarugaba, J., Egwang, T.G., Kironde, F., Andersson, B. & Wahlgren, M. (2010) var gene transcription dynamics in plasmodium falciparum patient isolates. *Molecular and Biochemical Parasitology*, **170**, 74–83.

Boone, K.E. & Watters, D.A. (1995) The incidence of malaria after splenectomy in papua new guinea. *BMJ (Clinical research ed)*, **311**, 1273.

Bopp, S.E., Manary, M.J., Bright, A.T., Johnston, G.L., Dharia, N.V., Luna, F.L., McCormack, S., Plouffe, D., McNamara, C.W., Walker, J.R., Fidock, D.A., Denchi, E.L. & Winzeler, E.A. (2013) Mitotic evolution of plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS genetics*, **9**, e1003293.

Bousema, T. & Drakeley, C. (2011) Epidemiology and infectivity of plasmodium falciparum and plasmodium vivax gametocytes in relation to malaria control and elimination. *Clinical microbiology reviews*, **24**, 377–+.

Bretscher, M.T., Maire, N., Chitnis, N., Felger, I., Owusu-Agyei, S. & Smith, T. (2011) The distribution of plasmodium falciparum infection durations. *Epidemics*, **3**, 109–118.

Buckee, C.O. & Recker, M. (2012) Evolution of the multi-domain structures of virulence genes in the human malaria parasite, plasmodium falciparum. *PLoS computational biology*, **8**, e1002451.

Buckee, C.O., Bull, P.C. & Gupta, S. (2009) Inferring malaria parasite population structure from serological networks. *Proceedings of the Royal Society B-Biological Sciences*, **276**, 477–485.

Buffet, P.A., Safeukui, I., Milon, G., Mercereau-Puijalon, O. & David, P.H. (2009) Retention of erythrocytes in the spleen: a double-edged process in human malaria. *Current opinion in hematology*, **16**, 157–164.

Bull, P.C., Kortok, M., Kai, O., Ndungu, F., Ross, A., Lowe, B.S., Newbold, C.I. & Marsh, K. (2000) Plasmodium falciparum-infected erythrocytes: Agglutination by diverse kenyan plasma is associated with severe disease and young host age. *Journal of Infectious Diseases*, **182**, 252–259.

Bull, P.C., Lowe, B.S., Kortok, M. & Marsh, K. (1999) Antibody recognition of plasmodium falciparum erythrocyte surface antigens in kenya: Evidence for rare and prevalent variants. *Infection and immunity*, **67**.

Bull, P.C., Lowe, B.S., Kortok, M., Molyneux, C.S., Newbold, C.I. & Marsh, K. (1998) Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature medicine*, **4**, 358–360.

Bull, P.C., Berriman, M., Kyes, S., Quail, M.A., Hall, N., Kortok, M.M., Marsh, K. & Newbold, C.I. (2005) Plasmodium falciparum variant surface antigen expression patterns during malaria. *Plos Pathogens*, **1**, 202–213.

Bull, P.C., Buckee, C.O., Kyes, S., Kortok, M.M., Thathy, V., Guyah, B., Stoute, J.A., Newbold, C.I. & Marsh, K. (2008) Plasmodium falciparum antigenic variation. mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Molecular microbiology*, **68**, 1519–1534.

Calderwood, M.S., Gannoun-Zaki, L., Wellems, T.E. & Deitsch, K.W. (2003) Plasmodium falciparum var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron. *Journal of Biological Chemistry*, **278**, 34125–34132.

Cham, G.K.K., Turner, L., Kurtis, J.D., Mutabingwa, T., Fried, M., Jensen, A.T.R., Lavstsen, T., Hviid, L., Duffy, P.E. & Theander, T.G. (2010) Hierarchical, domain type-specific acquisition of antibodies to plasmodium falciparum erythrocyte membrane protein 1 in tanzanian children. *Infection and immunity*, **78**, 4653–4659.

Cham, G.K.K., Turner, L., Lusingu, J., Vestergaard, L., Mmbando, B.P., Kurtis, J.D., Jensen, A.T.R., Salanti, A., Lavstsen, T. & Theander, T.G. (2009) Sequential, ordered acquisition of antibodies to plasmodium falciparum erythrocyte membrane protein 1 domains. *Journal of Immunology*, **183**, 3356–3363.

Chan, J.A., Howell, K.B., Reiling, L., Ataide, R., Mackintosh, C.L., Fowkes, F.J., Petter, M., Chesson, J.M., Langer, C., Warimwe, G.M., Duffy, M.F., Rogerson, S.J., Bull, P.C., Cowman, A.F., Marsh, K. & Beeson, J.G. (2012) Targets of antibodies against plasmodium falciparum-infected erythrocytes in malaria immunity. *The Journal of clinical investigation*.

Chen, D.S., Barry, A.E., Leliwa-Sytek, A., Smith, T.A., Peterson, I., Brown, S.M.,

Migot-Nabias, F., Deloron, P., Kortok, M.M., Marsh, K., Daily, J.P., Ndiaye, D., Sarr, O., Mboup, S. & Day, K.P. (2011) A molecular epidemiological study of var gene diversity to characterize the reservoir of plasmodium falciparum in humans in africa. *PloS one*, **6**, e16629.

Chen, Q.J., Fernandez, V., Sundstrom, A., Schlichtherle, M., Datta, S., Hagblom, P. & Wahlgren, M. (1998) Developmental selection of var gene expression in plasmodium falciparum. *Nature*, **394**, 392–395.

Chookajorn, T., Dzikowski, R., Frank, M., Li, F., Jiwani, A.Z., Hartl, D.L. & Deitsch, K.W. (2007) Epigenetic memory at malaria virulence genes. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 899–902.

Claessens, A., Adams, Y., Ghumra, A., Lindergard, G., Buchan, C.C., Andisi, C., Bull, P.C., Mok, S., Gupta, A.P., Wang, C.W., Turner, L., Arman, M., Raza, A., Bozdech, Z. & Rowe, J.A. (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**, E1772–81.

Collins, W.E. & Jeffery, G.M. (1999a) A retrospective examination of secondary sporozoite- and trophozoite-induced infections with plasmodium falciparum: development of parasitologic and clinical immunity following secondary infection. *The American Journal of Tropical Medicine and Hygiene*, **61**, 20–35.

Collins, W.E. & Jeffery, G.M. (1999b) A retrospective examination of sporozoite- and trophozoite-induced infections with plasmodium falciparum: development

of parasitologic and clinical immunity during primary infection. *The American Journal of Tropical Medicine and Hygiene*, **61**, 4–19.

de Roode, J.C., Pansini, R., Cheesman, S.J., Helinski, M.E., Huijben, S., Wargo, A.R., Bell, A.S., Chan, B.H., Walliker, D. & Read, A.F. (2005) Virulence and competitive ability in genetically diverse malaria infections. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 7624–7628.

DeAngelis, D.L. & Mooij, W.M. (2005) Individual-based modeling of ecological and evolutionary processes. *Annual Review of Ecology Evolution and Systematics*, **36**, 147–168.

Deitsch, K.W., Calderwood, M.S. & Wellems, T.E. (2001) Malaria - cooperative silencing elements in var genes. *Nature*, **412**, 875–876.

Deitsch, K.W., del Pinal, A. & Wellems, T.E. (1999) Intra-cluster recombination and var transcription switches in the antigenic variation of plasmodium falciparum. *Molecular and biochemical parasitology*, **101**, 107–116.

Deitsch, K.W., Lukehart, S.A. & Stringer, J.R. (2009) Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nature Reviews Microbiology*, **7**, 493–503.

Doolan, D.L., Dobano, C. & Baird, J.K. (2009) Acquired immunity to malaria. *Clinical microbiology reviews*, **22**, 13–36.

Duffy, M.F., Maier, A.G., Byrne, T.J., Marty, A.J., Elliott, S.R., O'Neill, M.T., Payne, P.D., Rogerson, S.J., Cowman, A.F., Crabb, B.S. & Brown, G.V. (2006) Var2csa is the principal ligand for chondroitin sulfate a in two allogeneic isolates of plasmodium falciparum. *Molecular and biochemical parasitology*, **148**, 117–124.

Duffy, M.F., Byrne, T.J., Carret, C., Ivens, A. & Brown, G.V. (2009) Ectopic recombination of a malaria var gene during mitosis associated with an altered var switch rate. *Journal of Molecular Biology*, **389**, 453–469.

Duraisingh, M.T., Voss, T.S., Marty, A.J., Duffy, M.F., Good, R.T., Thompson, J.K., Freitas-Junior, L.H., Scherf, A., Crabb, B.S. & Cowman, A.F. (2005) Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in plasmodium faiciparum. *Cell*, **121**, 13–24.

Dzikowski, R. & Deitsch, K.W. (2008) Active transcription is required for maintenance of epigenetic memory in the malaria parasite plasmodium faiciparum. *Journal of Molecular Biology*, **382**, 288–297.

Dzikowski, R., Li, F., Amulic, B., Eisberg, A., Frank, M., Patel, S., Wellems, T.E. & Deitsch, K.W. (2007) Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites. *EMBO reports*, **8**, 959–965.

Dzikowski, R., Frank, M. & Deitsch, K. (2006) Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. *Plos Pathogens*, **2**, e22.

Enderes, C., Kombila, D., Dal-Bianco, M., Dzikowski, R., Kremsner, P. & Frank, M. (2011) Var gene promoter activation in clonal plasmodium falciparum isolates follows a hierarchy and suggests a conserved switching program that is independent of genetic background. *Journal of Infectious Diseases*, **204**, 1620–1631.

Falk, N., Kaestli, M., Qi, W., Ott, M., Baea, K., Cortes, A. & Beck, H.P. (2009) Analysis of plasmodium falciparum var genes expressed in children from papua new guinea. *Journal of Infectious Diseases*, **200**, 347–356.

Fastman, Y., Noble, R., Recker, M. & Dzikowski, R. (2012) Erasing the epigenetic memory and beginning to switch–the onset of antigenic switching of var genes in plasmodium falciparum. *PloS one*, **7**, e34168.

Flueck, C., Bartfai, R., Volz, J., Niederwieser, I., Salcedo-Amaya, A.M., Alako, B.T.F., Ehlgen, F., Ralph, S.A., Cowman, A.F., Bozdech, Z., Stunnenberg, H.G. & Voss, T.S. (2009) Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *Plos Pathogens*, **5**.

Frank, M., Dzikowski, R., Costantini, D., Amulic, B., Berdougo, E. & Deitsch, K. (2006) Strict pairing of var promoters and introns is required for var gene silencing in the malaria parasite plasmodium falciparum. *Journal of Biological Chemistry*, **281**, 9942–9952.

Frank, M., Kirkman, L., Costantini, D., Sanyal, S., Lavazec, C., Templeton, T.J. & Deitsch, K.W. (2008) Frequent recombination events generate diversity within the multi-copy variant antigen gene families of plasmodium falciparum. *International journal for parasitology*, **38**, 1099–1109.

Frank, M., Dzikowski, R., Amulic, B. & Deitsch, K. (2007) Variable switching rates of malaria virulence genes are associated with chromosomal position. *Molecular microbiology*, **64**, 1486–1498.

Frank, S.A. (1999) A model for the sequential dominance of antigenic variants in african trypanosome infections. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **266**, 1397–1401.

Freitas-Junior, L.H., Bottius, E., Pirrit, L.A., Deitsch, K.W., Scheidig, C., Guinet, F., Nehrbass, U., Wellems, T.E. & Scherf, A. (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of p. falciparum. *Nature*, **407**, 1018–1022.

Fried, M. & Duffy, P.E. (1996) Adherence of plasmodium falciparum to chondroitin sulfate a in the human placenta. *Science*, **272**.

Fried, M., Nosten, F., Brockman, A., Brabin, B.J. & Duffy, P.E. (1998) Maternal antibodies block malaria. *Nature*, **395**, 851–852.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M. & Barrell, B. (2002) Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, **419**, 498–511.

Gatton, M.L., Peters, J.M., Fowler, E.V. & Cheng, Q. (2003) Switching rates of plasmodium falciparum var genes: faster than we thought? *Trends in parasitology*, **19**, 202–208.

Ghumra, A., Semblat, J.P., Ataide, R., Kifude, C., Adams, Y., Claessens, A., Anong, D.N., Bull, P.C., Fennell, C., Arman, M., Amambua-Ngwa, A., Walther, M., Conway, D.J., Kassambara, L., Doumbo, O.K., Raza, A. & Rowe, J.A. (2012) Induc-

tion of strain-transcending antibodies against group a pfemp1 surface antigens from virulent malaria parasites. *PLoS pathogens*, **8**, e1002665.

Giha, H.A., Staalsoe, T., Dodoo, D., Elhassan, I.M., Roper, C., Satti, G.M., Arnot, D.E., Theander, T.G. & Hviid, L. (1999) Nine-year longitudinal study of antibodies to variant antigens on the surface of plasmodium falciparum-infected erythrocytes. *Infection and immunity*, **67**, 4092–4098.

Gjini, E., Haydon, D.T., Barry, J.D. & Cobbold, C.A. (2010) Critical interplay between parasite differentiation, host immunity, and antigenic variation in trypanosome infections. *American Naturalist*, **176**.

Gottipati, P. & Helleday, T. (2009) Transcription-associated recombination in eukaryotes: link between transcription, replication and recombination. *Mutagenesis*, **24**, 203–210.

Grimm, V. (1999) Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological Modelling*, **115**, 129–148.

Guizetti, J. & Scherf, A. (2013) Silence, activate, poise, and switch! mechanisms of antigenic variation in plasmodium falciparum. *Cellular microbiology*.

Gupta, S., Ferguson, N. & Anderson, R. (1998) Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science*, **280**, 912–915.

Gupta, S., Maiden, M.C.J., Feavers, I.M., Nee, S., May, R.M. & Anderson, R.M. (1996) The maintenance of strain structure in populations of recombining infectious agents. *Nature medicine*, **2**, 437–442.

Gupta, S., Snow, R.W., Donnelly, C.A., Marsh, K. & Newbold, C. (1999) Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nature medicine*, **5**, 340–343.

Hall, J.P.J., Wang, H. & Barry, J.D. (2013) Mosaic vsgs and the scale of trypanosoma brucei antigenic variation. *PLoS pathogens*, **9**, e1003502–e1003502.

Hastings, W.K. (1970) Monte-carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.

Hay, S.I., Okiro, E.A., Gething, P.W., Patil, A.P., Tatem, A.J., Guerra, C.A. & Snow, R.W. (2010) Estimating the global clinical burden of plasmodium falciparum malaria in 2007. *PLoS medicine*, **7**, e1000290.

Horrocks, P., Pinches, R., Christodoulou, Z., Kyes, S.A. & Newbold, C.I. (2004) Variable var transition rates underlie antigenic variation in malaria. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11129–11134.

Howell, D.P.G., Levin, E.A., Springer, A.L., Kraemer, S.M., Phippard, D.J., Schief, W.R. & Smith, J.D. (2008) Mapping a common interaction site used by plasmodium falciparum duffy binding-like domains to bind diverse host receptors. *Molecular microbiology*, **67**, 78–87.

Jensen, A.T.R., Magistrado, P., Sharp, S., Joergensen, L., Lavstsen, T., Chiucciuini, A., Salanti, A., Vestergaard, L.S., Lusingu, J.P., Hermsen, R., Sauerwein, R., Christensen, J., Nielsen, M.A., Hviid, L., Sutherland, C., Staalsoe, T. & Theander, T.G. (2004) Plasmodium falciparum associated with severe childhood

malaria preferentially expresses pfemp1 encoded by group a var genes. *Journal of Experimental Medicine*, **199**, 1179–1190.

Jiang, L., Mu, J., Zhang, Q., Ni, T., Srinivasan, P., Rayavara, K., Yang, W., Turner, L., Lavstsen, T., Theander, T.G., Peng, W., Wei, G., Jing, Q., Wakabayashi, Y., Bansal, A., Luo, Y., Ribeiro, J.M.C., Scherf, A., Aravind, L., Zhu, J., Zhao, K. & Miller, L.H. (2013) Pfsetvs methylation of histone h3k36 represses virulence genes in plasmodium falciparum. *Nature*, **499**, 223–7.

Joergensen, L., Turner, L., Magistrado, P., Dahlback, M.A., Vestergaard, L.S., Lusingu, J.P., Lemnge, M., Salanti, A., Theander, T.G. & Jensen, A.T.R. (2006) Limited cross-reactivity among domains of the plasmodium falciparum clone 3d7 erythrocyte membrane protein 1 family. *Infection and immunity*, **74**, 6778–6784.

Kaestli, M., Cockburn, I.A., Cortes, A., Baea, K., Rowe, J.A. & Beck, H.P. (2006) Virulence of malaria is associated with differential expression of plasmodium falciparum var gene subgroups in a case-control study. *Journal of Infectious Diseases*, **193**, 1567–1574.

Kalmbach, Y., Rottmann, M., Kombila, M., Kremsner, P.G., Beck, H.P. & Kun, J.F.J. (2010) Differential var gene expression in children with malaria and antidromic effects on host gene expression. *Journal of Infectious Diseases*, **202**, 313–317.

Kerr, P.J., Ranford-Cartwright, L.C. & Walliker, D. (1994) Proof of intragenic recombination in plasmodium falciparum. *Molecular and biochemical parasitology*, **66**, 241–248.

Kirchgatter, K. & del Portillo, H.A. (2002) Association of severe noncerebral plasmodium falciparum malaria in brazil with expressed pfemp1 dbl1 alpha sequences lacking cysteine residues. *Molecular Medicine*, **8**, 16–23.

Kirkpatrick, S., Gelatt, C. & Vecchi, M. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Koepfli, C., Schoepflin, S., Bretscher, M., Lin, E., Kiniboro, B., Zimmerman, P.A., Siba, P., Smith, T.A., Mueller, I. & Felger, I. (2011) How much remains undetected? probability of molecular detection of human plasmodia in the field. *Plos One*, **6**, e19010.

Kosinski, R.J. (1980) Antigenic variation in trypanosomes - a computer-analysis of variant order. *Parasitology*, **80**, 343–357.

Kraemer, S.M. & Smith, J.D. (2003) Evidence for the importance of genetic structuring to the structural and functional specialization of the plasmodium falciparum var gene family. *Molecular microbiology*, **50**, 1527–1538.

Kraemer, S.M., Kyes, S.A., Aggarwal, G., Springer, A.L., Nelson, S.O., Christodoulou, Z., Smith, L.M., Wang, W., Levin, E., Newbold, C.I., Myler, P.J. & Smith, J.D. (2007) Patterns of gene recombination shape var gene repertoires in plasmodium falciparum: comparisons of geographically diverse isolates. *Bmc Genomics*, **8**, 45.

Kraemer, S.M. & Smith, J.D. (2006) A family affair: var genes, pfemp1 binding, and malaria disease. *Current opinion in microbiology*, **9**, 374–380.

Krause, D.R., Gatton, M.L., Frankland, S., Eisen, D.P., Good, M.F., Tilley, L. & Cheng, Q. (2007) Characterization of the antibody response against plasmodium

falciparum erythrocyte membrane protein 1 in human volunteers. *Infection and immunity*, **75**, 5967–5973.

Kyes, S., Christodoulou, Z., Pinches, R., Kriek, N., Horrocks, P. & Newbold, C. (2007) Plasmodium falciparum var gene expression is developmentally controlled at the level of rna polymerase ii-mediated transcription initiation. *Molecular microbiology*, **63**, 1237–1247.

Kyes, S., Pinches, R. & Newbold, C. (2000) A simple rna analysis method shows var and rif multigene family expression patterns in plasmodium falciparum. *Molecular and biochemical parasitology*, **105**, 311–315.

Kyriacou, H.M., Stone, G.N., Challis, R.J., Raza, A., Lyke, K.E., Thera, M.A., Kone, A.K., Doumbo, O.K., Plowe, C.V. & Rowe, J.A. (2006) Differential var gene transcription in plasmodium falciparum isolates from patients with cerebral malaria compared to hyperparasitaemia. *Molecular and biochemical parasitology*, **150**, 211–218.

Lambros, C. & Vanderberg, J.P. (1979) Synchronization of plasmodium-falciparum erythrocytic stages in culture. *Journal of Parasitology*, **65**, 418–420.

Langhorne, J., Ndungu, F.M., Sponaas, A.M. & Marsh, K. (2008) Immunity to malaria: more questions than answers. *Nature immunology*, **9**, 725–732.

Lavstsen, T., Magistrado, P., Hermsen, C.C., Salanti, A., Jensen, A.T.R., Sauerwein, R., Hviid, L., Theander, T.G. & Staalsoe, T. (2005) Expression of plasmodium falciparum erythrocyte membrane protein 1 in experimentally infected humans. *Malaria Journal*, **4**, 21.

Lavstsen, T., Salanti, A., Jensen, A.T.R., Arnot, D.E. & Theander, T.G. (2003) Subgrouping of plasmodium falciparum 3d7 var genes based on sequence analysis of coding and non-coding regions. *Malaria Journal*, **2**, 27.

Lavstsen, T., Turner, L., Saguti, F., Magistrado, P., Rask, T.S., Jespersen, J.S., Wang, C.W., Berger, S.S., Baraka, V., Marquard, A.M., Seguin-Orlando, A., Willerslev, E., Gilbert, M.T., Lusingu, J. & Theander, T.G. (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**, E1791–800.

Lemieux, J.E., Kyes, S.A., Otto, T.D., Feller, A.I., Eastman, R.T., Pinches, R.A., Berriman, M., Su, X.Z. & Newbold, C.I. (2013) Genome-wide profiling of chromosome interactions in plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic variation. *Molecular microbiology*.

Lindenthal, C., Kremsner, P.G. & Klinkert, M.Q. (2003) Commonly recognised plasmodium falciparum parasites cause cerebral malaria. *Parasitology research*, **91**, 363–368.

Lopez-Rubio, J.J., Gontijo, A.M., Nunes, M.C., Issar, N., Rivas, R.H. & Scherf, A. (2007) 5 ' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Molecular microbiology*, **66**, 1296–1305.

Lusingu, J.P.A., Jensen, A.T.R., Vestergaard, L.S., Minja, D.T., Dalgaard, M.B., Gesase, S., Mmbando, B.P., Kitua, A.Y., Lemnge, M.M., Cavanagh, D., Hviid, L. & Theander, T.G. (2006) Levels of plasma immunoglobulin g with specificity

against the cysteine-rich interdomain regions of a semiconserved plasmodium falciparum erythrocyte membrane protein 1, var4, predict protection against malarial anemia and febrile episodes. *Infection and immunity*, **74**, 2867–2875.

Lythgoe, K.A., Morrison, L.J., Read, A.F. & Barry, J.D. (2007) Parasite-intrinsic factors can explain ordered progression of trypanosome antigenic variation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8095–8100.

Magistrado, P.A., Lusingu, J., Vestergaard, L.S., Lemnge, M., Lavstsen, T., Turner, L., Hviid, L., Jensen, A.T.R. & Theander, T.G. (2007) Immunoglobulin g antibody reactivity to a group a plasmodium falciparum erythrocyte membrane protein 1 and protection from p-falciparum malaria. *Infection and immunity*, **75**, 2415–2420.

Marcello, L. & Barry, J.D. (2007) Analysis of the vsg gene silent archive in trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome research*, **17**, 1344–1352.

Marsh, K., Forster, D., Waruiru, C., Mwangi, I., Winstanley, M., Marsh, V., Newton, C., Winstanley, P., Warn, P. & Peshu, N. (1995) Indicators of life-threatening malaria in african children. *The New England journal of medicine*, **332**, 1399–1404.

Marsh, K. & Howard, R.J. (1986) Antigens induced on erythrocytes by p. falciparum: expression of diverse and conserved determinants. *Science (New York, NY)*, **231**, 150–153.

Marsh, K., Otoo, L., Hayes, R.J., Carson, D.C. & Greenwood, B.M. (1989) Antibodies to blood stage antigens of plasmodium falciparum in rural gambians and their relation to protection against infection. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **83**, 293–303.

Marsh, K. & Snow, R.W. (1997) Host-parasite interaction and morbidity in malaria endemic areas. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **352**, 1385–1394.

Marty, A.J., Thompson, J.K., Duffy, M.F., Voss, T.S., Cowman, A.F. & Crabb, B.S. (2006) Evidence that plasmodium falciparum chromosome end clusters are cross-linked by protein and are the sites of both virulence gene silencing and activation. *Molecular microbiology*, **62**, 72–83.

Maubert, B., Fievet, N., Tami, G., Cot, M., Boudin, C. & Deloron, P. (1999) Development of antibodies against chondroitin sulfate a-adherent plasmodium falciparum in pregnant women. *Infection and immunity*, **67**, 5367–5371.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Molineaux, L., Diebner, H.H., Eichner, M., Collins, W.E., Jeffery, G.M. & Dietz, K. (2001) Plasmodium falciparum parasitaemia described by a new mathematical model. *Parasitology*, **122**, 379–391.

Morrison, L.J., Majiwa, P., Read, A.F. & Barry, J.D. (2005) Probabilistic order in antigenic variation of trypanosoma brucei. *International journal for parasitology*, **35**, 961–972.

Mueller, I., Schoepflin, S., Smith, T.A., Benton, K.L., Bretscher, M.T., Lin, E., Kiniboro, B., Zimmerman, P.A., Speed, T.P., Siba, P. & Felger, I. (2012) Force of infection is key to understanding the epidemiology of plasmodium falciparum malaria in papua new guinean children. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 10030–10035.

Mugasa, J., Qi, W., Rusch, S., Rottman, M. & Beck, H.P. (2012) Genetic diversity of expressed plasmodium falciparum var genes from tanzanian children with severe malaria. *Malaria journal*, **11**, 230.

Newman, M.E.J. (2001) Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, **64**, 016132.

Nielsen, M.A., Staalsoe, T., Kurtzhals, J.A.L., Goka, B.Q., Dodoo, D., Alifrangis, M., Theander, T.G., Akanmori, B.D. & Hviid, L. (2002) Plasmodium falciparum variant surface antigen expression varies between isolates causing severe and non-severe malaria and is modified by acquired immunity. *Journal of Immunology*, **168**, 3444–3450.

Ockenhouse, C.F., Ho, M., Tandon, N.N., Vanseventer, G.A., Shaw, S., White, N.J., Jamieson, G.A., Chulay, J.D. & Webster, H.K. (1991) Molecular-basis of sequestration in severe and uncomplicated plasmodium-falciparum malaria - differential adhesion of infected erythrocytes to cd36 and icam-1. *Journal of Infectious Diseases*, **164**, 163–169.

Ofori, M.F., Dodoo, D., Staalsoe, T., Kurtzhals, J.A., Koram, K., Theander, T.G., Akanmori, B.D. & Hviid, L. (2002) Malaria-induced acquisition of antibodies to

plasmodium falciparum variant surface antigens. *Infection and immunity*, **70**, 2982–2988.

Opsahl, T. (2009) Structure and evolution of weighted networks.

Paget-McNicol, S., Gatton, M., Hastings, I. & Saul, A. (2002) The plasmodium falciparum var gene switching rate, switching mechanism and patterns of parasite recrudescence described by mathematical modelling. *Parasitology*, **124**, 225–235.

Perez-Toledo, K., Rojas-Meza, A.P., Mancio-Silva, L., Hernandez-Cuevas, N., Delgadillo, D.M., Vargas, M., Martinez-Calvillo, S., Scherf, A. & Hernandez-Rivas, R. (2009) Plasmodium falciparum heterochromatin protein 1 binds to tri-methylated histone 3 lysine 9 and is linked to mutually exclusive expression of var genes. *Nucleic acids research*, **37**, 2596–2606.

Peters, J., Fowler, E., Gatton, M., Chen, N.H., Saul, A. & Cheng, Q. (2002) High diversity and rapid changeover of expressed var genes during the acute phase of plasmodium falciparum infections in human volunteers. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10689–10694.

Peters, J.M., Fowler, E.V., Krause, D.R., Cheng, Q. & Gatton, M.L. (2007) Differential changes in plasmodium falciparum var transcription during adaptation to culture. *Journal of Infectious Diseases*, **195**, 748–755.

Pfaffl, M.W. (2001) A new mathematical model for relative quantification in real-time rt-pcr. *Nucleic acids research*, **29**, e45.

Pollitt, L.C., MacGregor, P., Matthews, K. & Reece, S.E. (2011) Malaria and trypanosome transmission: different parasites, same rules? *Trends in parasitology*, **27**, 197–203.

Pongponratn, E., Riganti, M., Punpoowong, B. & Aikawa, M. (1991) Microvascular sequestration of parasitized erythrocytes in human falciparum-malaria - a pathological-study. *American Journal of Tropical Medicine and Hygiene*, **44**, 168–175.

Prado, F. & Aguilera, A. (2005) Impairment of replication fork progression mediates rna polii transcription-associated recombination. *Embo Journal*, **24**, 1267–1276.

Ralph, S.A., Scheidig-Benatar, C. & Scherf, A. (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**, 5414–5419.

Rask, T.S., Hansen, D.A., Theander, T.G., Pedersen, A.G. & Lavstsen, T. (2010) Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes - divide and conquer. *Plos Computational Biology*, **6**, e1000933.

Recker, M., Nee, S., Bull, P.C., Kinyanjui, S., Marsh, K., Newbold, C. & Gupta, S. (2004) Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. *Nature*, **429**, 555–558.

Recker, M., Arinaminpathy, N. & Buckee, C.O. (2008) The effects of a partitioned var gene repertoire of plasmodium falciparum on antigenic diversity and the acquisition of clinical immunity. *Malaria Journal*, **7**, 18.

Recker, M., Buckee, C.O., Serazin, A., Kyes, S., Pinches, R., Christodoulou, Z., Springer, A.L., Gupta, S. & Newbold, C.I. (2011) Antigenic variation in plasmodium falciparum malaria involves a highly structured switching pattern. *Plos Pathogens*, **7**, e1001306.

Ricke, C.H., Staalsoe, T., Koram, K., Akanmori, B.D., Riley, E.M., Theander, T.G. & Hviid, L. (2000) Plasma antibodies from malaria-exposed pregnant women recognize variant surface antigens on plasmodium falciparum-infected erythrocytes in a parity-dependent manner and block parasite adhesion to chondroitin sulfate a. *Journal of immunology (Baltimore, Md: 1950)*, **165**, 3309–3316.

Roberts, D.J., Craig, A.G., Berendt, A.R., Pinches, R., Nash, G., Marsh, K. & Newbold, C.I. (1992) Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature*, **357**, 689–692.

Robinson, B.A., Welch, T.L. & Smith, J.D. (2003) Widespread functional specialization of plasmodium falciparum erythrocyte membrane protein 1 family members to bind cd36 analysed across a parasite genome. *Molecular microbiology*, **47**, 1265–1278.

Rottmann, M., Lavstsen, T., Mugasa, J.P., Kaestli, M., Jensen, A.T.R., Mueller, D., Theander, T. & Beck, H.P. (2006) Differential expression of var gene groups is associated with morbidity caused by plasmodium falciparum infection in tanzanian children. *Infection and immunity*, **74**, 3904–3911.

Rowe, J.A., Claessens, A., Corrigan, R.A. & Arman, M. (2009) Adhesion of plasmodium falciparum-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications. *Expert Reviews in Molecular Medicine*, **11**, e16.

Salanti, A., Dahlback, M., Turner, L., Nielsen, M.A., Barfod, L., Magistrado, P., Jensen, A.T., Lavstsen, T., Ofori, M.F., Marsh, K., Hviid, L. & Theander, T.G. (2004) Evidence for the involvement of var2csa in pregnancy-associated malaria. *The Journal of experimental medicine*, **200**, 1197–1203.

Salanti, A., Staalsoe, T., Lavstsen, T., Jensen, A.T.R., Sowa, M.P.K., Arnot, D.E., Hviid, L. & Theander, T.G. (2003) Selective upregulation of a single distinctly structured var gene in chondroitin sulphate a-adhering plasmodium falciparum involved in pregnancy-associated malaria. *Molecular microbiology*, **49**, 179–191.

Sama, W., Killeen, G. & Smith, T. (2004) Estimating the duration of plasmodium falciparum infection from trials of indoor residual spraying. *American Journal of Tropical Medicine and Hygiene*, **70**, 625–634.

Scherf, A., Hernandez-Rivas, R., Buffet, P., Bottius, E., Benatar, C., Pouvelle, B., Gysin, J. & Lanzer, M. (1998) Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in plasmodium falciparum. *Embo Journal*, **17**, 5418–5426.

Severins, M., Klinkenberg, D. & Heesterbeek, H. (2012) How selection forces dictate the variant surface antigens used by malaria parasites. *Journal of the Royal Society, Interface / the Royal Society*, **9**, 246–260.

Smalley, M.E., Abdalla, S. & Brown, J. (1981) The distribution of plasmodium falciparum in the peripheral blood and bone marrow of gambian children. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **75**, 103–105.

Smith, D.L., Battle, K.E., Hay, S.I., Barker, C.M., Scott, T.W. & McKenzie, F.E. (2012) Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS pathogens*, **8**, e1002588.

Smith, J.D., Chitnis, C.E., Craig, A.G., Roberts, D.J., Hudson-Taylor, D.E., Peterson, D.S., Pinches, R., Newbold, C.I. & Miller, L.H. (1995) Switches in expression

of plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell*, **82**, 101–110.

Smith, J.D., Rowe, J.A., Higgins, M.K. & Lavstsen, T. (2013) Malaria's deadly grip: Cytoadhesion of plasmodium falciparum infected erythrocytes. *Cellular microbiology*.

Smith, J.D., Subramanian, G., Gamain, B., Baruch, D.I. & Miller, L.H. (2000) Classification of adhesive domains in the plasmodium falciparum erythrocyte membrane protein 1 family. *Molecular and biochemical parasitology*, **110**, 293–310.

Su, X.Z., Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch, J.A. & Wellems, T.E. (1995a) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. *Cell*, **82**, 89–100.

Su, X.Z., Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch, J.A. & Wellems, T.E. (1995b) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium-falciparum-infected erythrocytes. *Cell*, **82**, 89–100.

Swamy, L., Amulic, B. & Deitsch, K.W. (2011) Plasmodium falciparum var gene silencing is determined by cis dna elements that form stable and heritable interactions. *Eukaryotic Cell*, **10**, 530–539.

Taylor, H.M., Kyes, S.A., Harris, D., Kriek, N. & Newbold, C.I. (2000) A study of var gene transcription in vitro using universal var gene primers. *Molecular and biochemical parasitology*, **105**, 13–23.

Trager, W. & Jensen, J.B. (1976) Human malaria parasites in continuous culture. *Science*, **193**, 673–675.

Tran, T.M., Li, S., Doumbo, S., Doumtabe, D., Huang, C.Y., Dia, S., Bathily, A., Sangala, J., Kone, Y., Traore, A., Niangaly, M., Dara, C., Kayentao, K., Ongoiba, A., Doumbo, O.K., Traore, B. & Crompton, P.D. (2013) An intensive longitudinal cohort study of malian children and adults reveals no evidence of acquired immunity to plasmodium falciparum infection. *Clinical Infectious Diseases*, **57**, 40–47.

Trimnell, A.R., Kraemer, S.M., Mukherjee, S., Phippard, D.J., Janes, J.H., Flamoe, E., Su, X.Z., Awadalla, P. & Smith, J.D. (2006) Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Molecular and biochemical parasitology*, **148**, 169–180.

Turner, L., Lavstsen, T., Berger, S.S., Wang, C.W., Petersen, J.E.V., Avril, M., Brazier, A.J., Freeth, J., Jespersen, J.S., Nielsen, M.A., Magistrado, P., Lusingu, J., Smith, J.D., Higgins, M.K. & Theander, T.G. (2013) Severe malaria is associated with parasite binding to endothelial protein c receptor. *Nature*, **498**, 502–+.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. & Leunissen, J.A.M. (2007) Primer3plus, an enhanced web interface to primer3. *Nucleic acids research*, **35**, W71–W74.

van Baalen, M. & Sabelis, M.W. (1995) The dynamics of multiple infection and the evolution of virulence. *American Naturalist*, **146**, 881–910.

van Noort, S.P., Nunes, M.C., Weedall, G.D., Hviid, L. & Gomes, M.G.M. (2010) Immune selection and within-host competition can structure the repertoire of variant surface antigens in plasmodium falciparum - a mathematical model. *Plos One*, **5**, e9778.

Vestergaard, L.S., Lusingu, J.P., Nielsen, M.A., Mmbando, B.P., Dodoo, D., Akanmori, B.D., Alifrangis, M., Bygbjerg, I.C., Lemnge, M.M., Staalsoe, T., Hviid, L. & Theander, T.G. (2008) Differences in human antibody reactivity to plasmodium falciparum variant surface antigens are dependent on age and malaria transmission intensity in northeastern tanzania. *Infection and immunity*, **76**, 2706–2714.

Viebig, N.K., Gamain, B., Scheidig, C., Lepolard, C., Przyborski, J., Lanzer, M., Gysin, J. & Scherf, A. (2005) A single member of the plasmodium falciparum var multigene family determines cytoadhesion to the placental receptor chondroitin sulphate a. *EMBO reports*, **6**, 775–781.

Voss, T.S., Healer, J., Marty, A.J., Duffy, M.F., Thompson, J.K., Beeson, J.G., Reeder, J.C., Crabb, B.S. & Cowman, A.F. (2006) A var gene promoter controls allelic exclusion of virulence genes in plasmodium falciparum malaria. *Nature*, **439**, 1004–1008.

Voss, T.S., Tonkin, C.J., Marty, A.J., Thompson, J.K., Healer, J., Crabb, B.S. & Cowman, A.F. (2007) Alterations in local chromatin environment are involved in silencing and activation of subtelomeric var genes in plasmodium falciparum. *Molecular microbiology*, **66**, 139–150.

Ward, C.P., Clottey, G.T., Dorris, M., Ji, D.D. & Arnot, D.E. (1999) Analysis

of plasmodium falciparum pfemp-1/var genes suggests that recombination rearranges constrained sequences. *Molecular and biochemical parasitology*, **102**, 167–177.

Warimwe, G.M., Keane, T.M., Fegan, G., Musyoki, J.N., Newton, C.R.J.C., Pain, A., Berriman, M., Marsh, K. & Bull, P.C. (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**, 21801–21806.

WHO (2012) World malaria report 2012. Technical report.

Wunderlich, G., Alves, F.P., Golnitz, U., Tada, M.S., Camargo, E.F. & da Silva, L.H.P. (2005) Rapid turnover of plasmodium falciparum var gene transcripts and genotypes during natural non-symptomatic infections. *Revista do Instituto de Medicina Tropical de Sao Paulo*, **47**, 195–201.

Zhang, Q., Zhang, Y., Huang, Y., Xue, X., Yan, H., Sun, X., Wang, J., McCutchan, T.F. & Pan, W. (2011) From in vivo to in vitro: Dynamic analysis of plasmodium falciparum var gene expression patterns of patient isolates during adaptation to culture. *Plos One*, **6**, e20591.