Computational approaches to vaccine development for multi-strain pathogens



Andrew S. Walker

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy, D.Phil

Michaelmas 2017

Contents

Chapter 1 Introduction ---------- 9

1.1 The evolution and fitness of pathogens ---------- 9

1.2 The role of antigenic diversity in evolutionary strategy ---------- 12

1.3 Mathematical modelling of multi-strain pathogens ---------- 14

1.4 Mathematical modelling of infectious diseases and interventions ---------- 18

1.5 Thesis outline ---------- 21

Chapter 2 Reverse Immunodynamics: theory and development of metrics ---------- 25

2.1 Introduction ---------- 26

2.2 Methods ---------- 33

2.2.1 Analysing pairs of allelic variants ---------- 33

2.2.2 Quantifying non-overlap using information theory ---------- 34

2.2.3 Quantifying non-overlap using the f* statistic ---------- 38

2.2.4 Transforming and analysing asymmetrical matrices ---------- 39

2.2.6 Quantifying non-overlap between asymmetrical pairs – W ---------- 42

2.2.7 Mathematical modelling of multi-strain systems ---------- 45

2.2.8 Generating Multiple Sequence Alignments ---------- 45

2.3 Results ---------- 47

2.3.1 Simulating symmetrical dimorphic pairs of sites ---------- 47

2.3.2 Simulating polymorphic symmetrical pairs of sites ---------- 52

2.3.3 An algorithm for accounting for asymmetry when assessing non-overlap 57

2.3.4 Quantifying non-overlap in simulated asymmetrical strain systems ------ 61

2.3.5 Quantifying non-overlap – conclusions ---------- 62

2.4 Discussion ---------- 65

Chapter 3 Reverse Immunodynamics: applications ---------- 70

3.1 Introduction ---------- 71

3.1.1 Genetic sequence data availability ---------- 71

3.1.2 *Neisseria meningitidis* outer membrane proteins ---------- 72

3.1.3 *Plasmodium falciparum* Apical Membrane Antigen 1 ---------- 73

3.1.4 Epstein Barr Virus ---------- 76

3.2 Methods ---------- 80

3.2.1 Multiple Sequence Alignments ---------- 80

3.2.2 Phylogenetic analysis ---------- 82

3.2.3 Biochemistry of amino acid substitution ---------- 83

3.2.4 Quantifying non-overlap ---------- 85

3.3 Results ---------- 87

3.3.1 *Neisseria meningitidis* PorA and FetA ---------- 87

3.3.2 *Plasmodium falciparum* AMA-1 ---------- 95

3.3.3 Analyses of Epstein Barr Virus (EBV) ---------- 101

3.3.3.1 EBV – overall patterns of non-overlap ---------- 101

3.3.3.2 EBV – patterns of non-overlap in latency-associated genes ---------- 104

3.3.3.3 EBV – patterns of non-overlap in lytic phase genes ---------- 107

3.3.3.4 EBV – patterns of non-overlap in genes encoding glycoproteins ---- 111

3.3.3.5 EBV – phylogeny of non-overlapping sites ---------- 114

3.4 Discussion ---------- 116

Chapter 4 Modelling combinations of pre-erythrocytic *Plasmodium falciparum* malaria vaccines ---------- 121

4.1 Introduction ---------- 122

4.2 Methods ---------- 127

4.2.1 The model ---- 127
4.2.2 Model fitting to trial data ---- 130
4.3 Results ---- *133*
4.3.1 Baseline within-host dynamics ---- 133
4.3.2 Within-host dynamics in single vaccine recipients ---- 138
4.3.3 Effects of combining vaccines ---- 140
4.4 Discussion ---- *141*
Chapter 5 Modelling mass malaria vaccine administration ---- 146
5.1 Introduction ---- *147*
5.2 Methods ---- *152*
5.2.1 Data/study design ---- 152
5.2.2 Individual-based survival model (IBSM) ---- 153
5.2.2.1 IBSM – force of infection ---- 156
5.2.2.2 IBSM – modelling vaccination ---- 164
5.2.2.3 IBSM – parameter estimation ---- 166
5.2.3 Epidemiological transmission model ---- 169
5.2.3.1 Transmission model structure ---- 169
5.2.3.2 Transmission model parameter values ---- 172
5.3 Results ---- *175*
5.3.1 Malaria prevalence data ---- 175
5.3.2 Parameterisation of the Individual Based Survival Model ---- 177
5.3.2.1 Model selection by quality of fit ---- 177
5.3.2.2 Estimating force of infection parameter values ---- 181
5.3.2.3 Estimated vaccine efficacy parameter values ---- 186
5.3.3 Simulating malaria dynamics using the transmission model ---- 188

5.3.3.1 Control scenario----------188

5.3.3.2 Disease dynamics under modelled vaccine interventions----------190

5.3.3.3 Efficacy of single interventions---------- 191

5.3.3.4 Efficacy of hypothetical combined interventions ----------198

5.4 Discussion----------*201*

Chapter 6 Summary and conclusions---------- 208

6.1 Immune-driven evolution ---------- *209*

6.2 Rethinking vaccine strategy---------- *211*

6.3 Conclusions ----------*213*

Bibliography----------**216**

Appendices----------**249**

Appendix A: Supplementary Methods for Chapter 3 ---------- *250*

Appendix B: Supplementary Information for Chapter 3----------*255*

Appendix C: Supplementary Information for Chapter 4---------- *266*

Appendix D: Published research paper (within-host modelling)---------- *269*

Computational approaches to vaccine development for multi-strain pathogens

Andrew S. Walker
D.Phil
Hertford College, Michaelmas Term 2017

Abstract

This thesis presents a number of computational approaches that I have used for the study, design, and optimization of vaccination strategy. An effective vaccination regimen against an infectious disease will elicit immune responses that target carefully selected regions of the parasite, and will be implemented in such a way that maximum benefit can be had. Here, I present research into various steps of the vaccine development pipeline, from developing new methods by which targets can be selected, to the best way that a therapy can be implemented in the field. Existing understanding of the ways in which populations of pathogens are structured is used to develop methods for selecting novel vaccination targets, that is, regions of pathogens against which a vaccine might be highly effective. These approaches are combined with mathematical modelling of vaccination at both within-host and population levels to assist in the development and implementation of vaccines against pathogens of global health importance. My within-host model suggests that a highly efficacious sterilely protective malaria vaccine may be possible by combining existing therapies. An epidemiological approach provides evidence that mass vaccination with a malaria vaccine of moderate efficacy, in combination with other interventions, may drive eradication in many transmission contexts. Considered together, the works contained within this thesis explore the ways in which both existing and novel therapies may be combined to gain maximum benefit.

Chapter 1 Introduction

1.1 The evolution and fitness of pathogens

The reproductive fitness of a pathogen, here taken to be its ability to infect a host, produce progeny, and to transmit on to others, is dependent on its successful evasion of destruction by the host immune response. Pathogens that fail to reproduce, fail to pass on their genetic information to the next generation, and so ensuring persistence is the central tenant of a pathogen's evolutionary strategy. Generations of selection and evolution force pathogen populations towards an optimum strategy, the nature of which is dependent on their mode of transmission, and is inextricably linked to host immune responses against the pathogen.

In order to enhance their own fitness, pathogens very often cause pathologies that are conducive to transmission; canonical examples being the coughing and sneezing caused by influenza (reviewed in Taubenberger & Morens 2008), or haemorrhaging caused by ebolavirus (Sanchez, Geisbert and Feldmann, 2006). These pathologies, however, can cause a reduction in fitness of the pathogen if too severe: limiting the life expectancy of the host limits the number of potential opportunities for transmission.

The intrinsic fitness of a pathogen is that which can be measured through *in vitro* assays, where growth rates are measured independently of host immune response (Maree *et al.*, 2000; Arnott *et al.*, 2010) and is a function of factors such as the efficiency of metabolic or cellular processes, protein function, invasion mechanisms, or other such factors. Intrinsic fitness impacts the epidemic potential of a pathogen in a population, as defined by its basic reproductive rate (R_0), that is, the average number

of secondary infections that a single infection in a totally susceptible population is likely to cause (Dublin and Lotka, 1925; Macdonald, 1952). Considerations of optimal pathogen fitness must take account of host immune responses, and a plethora of other environmental factors. This fitness is quantified by the *effective* reproductive number (R_e), which is also reliant on differences in susceptibility to infection of members of the host population, and thereby incorporates interactions of the pathogen with host immunity. Pathogens must maximise their fitness in the face of this selection, and so optimal effective fitness is achieved through an evolutionary process that maintains protein function whilst evading destruction by the host immune response and not causing a pathology so severe that opportunities for transmission are lost. Assessing fitness in the context of immunity has been the subject of much epidemiological modelling and is fundamental to public health. Models such as those discussed in this thesis are regularly used to inform the design of interventions against epidemic infection.

Models seek, with varying degrees of complexity, to simulate the range of factors that summate to a pathogen's evolutionary strategy and resulting fitness. These factors can be considered to comprise a Latin hypercube, with each axis of the hypercube representing discrete values of a variable that impacts fitness, and so a given point within the hypercube corresponds to a particular combination of those factors with a particular ability to pass on its genes. Fitness was first considered in this way by Sewall Wright in 1932, who postulated that, if sets of gene combinations can be graded in terms of their adaptive value, then species will evolve towards a combination which brings the greatest overall gain (Wright, 1932). Many works since Wright have explored this notion of a fitness landscape (Mustonen & La 2009; Romero & Arnold 2009; reviewed in Visser & Krug 2014), although the metaphor of a landscape has been

argued to be of limited utility when considering natural systems (Kaplan, 2008), in part due to the high dimensionality and continuousness of variation. The purpose of a fitness landscape is to represent the range of possible genotypes in terms of the ease with which changes between them may occur. In reality, often even slight genotypic changes can have a profound impact upon protein function and species fitness (Reva, Antipin and Sander, 2011) and, similarly, multiple genotypic changes may occur very rapidly. In keeping with the landscape metaphor, species may appear to instantaneously traverse troughs or tunnel through peaks in order to maintain optimum fitness (Iwasa, Michor and Nowak, 2004).

Classical fitness landscapes may also fail to reflect the transiency of the landscape, which may in fact be better reflected by the metaphor of a transient "seascape" (Mustonen and La, 2009). Time-dependent selection creates shifting peaks of fitness, meaning that a stable equilibrium state (*i.e.*, a constant genotype) may never be reached (Lande 2007; reviewed in Sæther & Engen 2015). In disease systems, these changes can be driven by a plethora of factors, including seasonality in transmission (for example, due to changes in vector dynamics), climate, migration, changes in host behaviour and transmission networks, or co-infection with other pathogens. Transiency in selection seascapes/landscapes can also be considered to be density-dependent – that is, if a reasonable proportion of a species exists in a particular region of the Latin hypercube of variability, then that region will change in fitness, thus driving the species to diversify to new fitness peaks (in the case of negative density-dependent selection) or further drawing the species to that peak (in the case of positive density-dependent selection).

A canonical example of a complex, transient, and high-dimensional landscape is in the context of influenza virus A, where frequency-dependant selection drives diversification in some of the genome, whilst purifying selection maintains conservation in others (Xu *et al.*, 2012). Some sites within the core structural proteins of the virus are hidden from antibodies and so do not face pressure to diversify. Instead, they are often structurally and functionally fragile: changes in them may profoundly impact viral fitness, and so they cannot tolerate mutations that lead to structural changes (Rihn *et al.*, 2013; Wu *et al.*, 2015). In contrast, antigenic regions of the hemagglutinin gene are highly diverse as they must escape the long-lasting host immune response (Fitch *et al.*, 1991, 1997; Plotkin and Dushoff, 2003; Li *et al.*, 2011). As demonstrated later, simplifying assumptions can allow mathematical models to produce meaningful representations of these complex multi-strain pathogen systems (§1.4). Improving our understanding of infectious diseases will better equip us to design novel therapies against them. As discussed later in this thesis, therapies that are sensitive to the pathogen's evolutionary strategy may be more resistant to the emergence of escape.

1.2 The role of antigenic diversity in evolutionary strategy

In the example of influenza, immunodominant epitopes vary over time, however this is not always the case. Pathogens exhibit a plethora of life history strategies, the nature of which is intrinsically linked to the diversity of the regions against which immune responses are effective. Note that this definition of antigenic diversity differs from the more general definition of "genotypic diversity" – the level of variation in the genes which encode all protein products. In some pathogens, for example the measles virus (MeV), immune responses are highly effective against antigenically conserved regions of the pathogen (reviewed in Beaty & Lee 2016; Fulton et al. 2015). This means that a

single infection induces a long-lasting immune response in the host that prevents subsequent infection. The evolutionary strategy of the virus that allows it to be maintained in the population is thus focused on it being highly infectious (having a high R_0) and relying on a steady influx of new immuno-naïve individuals into the host population through birth (Rota *et al.*, 1996, 2011). This pattern of infection is characteristic of many childhood infectious diseases, where strong immunity against conserved antigenic regions protects against re-infection.

This evolutionary strategy contrasts those seen in antigenically diverse pathogens such as influenza A and Human Immunodeficiency Virus (HIV), in which high mutation rates ensure that the effectiveness of immune responses in controlling infection is limited or transient (reviewed in Migueles & Connors 2015), and so repeat or chronic infections are common (Hemelaar, 2012; Redd *et al.*, 2012). In order for the immune system to prevent, clear, or control infection, immune responses must be raised against a large portion of the possible antigenic repertoire. For most individuals infected with HIV, strain-specific immune responses are ineffective against the highly diverse virus, especially given eventual (and almost inevitable without intervention) HIV-induced CD4+ T cell depletion (Okoye and Picker, 2013). This extreme genetic and antigenic diversity is facilitated by high mutation rates (Cuevas *et al.*, 2015), tolerance of mutations in much of the HIV genome (Humphris-narayanan *et al.*, 2012), and redundancies in the cellular and metabolic pathways that mean that a loss of protein function may not necessarily incur a loss of fitness (reviewed in Marx & Alian 2015).

Differences in the diversity of antigenic regions against which immune responses are most effective are also reflected in the relative success of vaccine development efforts.

We have, and commonly utilize, efficacious vaccines against antigenically conserved pathogens such as polio, guinea worm disease, and measles, with these pathogens targeted for eradication by sustained vaccination efforts (Morens *et al.*, 2011; Al-awadi *et al.*, 2014; Groce, Banks and Stein, 2014). In contrast, the development of vaccines against antigenically diverse pathogens is far more limited, despite substantial development efforts and the immeasurable benefits to public health that an effective vaccine against pathogens such as *Plasmodium falciparum* malaria, HIV, or Epstein Barr Virus (EBV) could bring.

MeV and HIV comprise extreme examples of low and high antigenic diversity. Most other pathogens, however, lie somewhere in between: they are largely more constricted in their movements through the fitness landscape by lower mutation rates and less tolerance of changes in essential proteins, whilst still being required to evade or tolerate host immune effectors. Regions of intermediate antigenic diversity within highly diverse pathogens could be of interest to vaccine development: they would be more immunogenic than conserved regions, whilst having more limited avenues of immune escape than highly variable regions. In this thesis, I present analyses of genetic sequence data from a number of pathogens, elucidating regions of intermediate diversity that exhibit certain patterns in congruence with an evolutionary history of immune selection pressure (Chapter 3).

1.3 Mathematical modelling of multi-strain pathogens

Antigenic variation spurns complex aetiologies and ecological interactions between a host and its parasites. By understanding the dynamic systems that emerge as a result this co-evolutionary conflict, we stand a better chance of developing vaccines that are effective in equipping our immune systems against infection.

In considering these systems, it is helpful to consider three archetypal pathogen population structures that may emerge in different conditions. If the host immune response is effective against parts of the genome of the parasite that are conserved between genotypes (variants), then it follows that the variant that will persist will simply be the one with the highest intrinsic fitness. In contrast, immune responses effective against a single *variable* locus, all other regions being conserved and incapable of raising a response, may allow for multiple strains of the pathogen to circulate concurrently. This is because immune responses against the pathogen are raised in a strain-specific manner, and so the effectiveness of responses against each strain in a single host will depend on their history of exposure, and so no single strain is expected to maintain an advantage and overcome the others. Finally, if immune responses are effective against *multiple* loci of a pathogen then complex and dynamic behaviours can emerge, the nature of the which have been shown to depend on immune-mediated competition that occurs between circulating strains, notably the impact that an immune response against a single variant has upon other variants (introduced in full in Chapter 2).

When a given strain induces an immune response that protects against infection with any other strain that shares antigenic components, then the level of cross-protection afforded is said to be high. In this case, individuals may acquire resistance to the entire antigenic repertoire of the pathogen after infection with only a limited number of strains. In contrast, in many systems, immune responses are highly strain-specific, and so responses must be raised against a large proportion of the antigenic repertoire in order to be universally effective (Chattopadhyay *et al.*, 2003; Gatton and Cheng, 2004). Even where there is minimal cross-protection between strains, competition

between those strains inevitably occurs, resulting in a variety of dynamic behaviours (Gupta, Ferguson and Anderson, 1998; Lourenço, Wikramaratna and Gupta, 2015).

Classical, single-strain epidemiological models are of immense use when understanding systems in which there is a single dominant strain, where strains do not compete, or when making simplifying assumptions of more complex systems. That said, in order to robustly consider the roles of immune selection pressure in driving multi-strain pathogen dynamics, more specialized models are required (Wikramaratna *et al.*, 2015; Kucharski, Andreasen and Gog, 2016). The precise topography of strain space (§1.2) has not been defined for any pathogen, however simplifying assumptions allow models of strain competition explore such systems. These models vary in complexity from simple one-dimensional linearizations in which interactions only occur with nearest-neighbours in a circular strain space (Andreasen, Lin and Levin, 1997), to complex models that attempt to incorporate genetic distance as a determinant of the level of cross-immunity between variants (Minayev and Ferguson, 2009).

A seminal multi-locus, multi-strain model of pathogen dynamics (Gupta, Ferguson and Anderson, 1998) predicts that when the levels of cross-protection are very low, strains will circulate independently, i.e., infection with one strain does not preclude infection with others, even those that share antigenic components. In contrast, if infection with one strain induces responses that are largely protective against strains that share antigenic components, only strains that share no antigenic components with the original strain will propagate. In this way, cross-protection leads to the discretisation of strains into non-overlapping variants (Gupta *et al.*, 1996; Gupta, Ferguson and Anderson, 1998; Lourenço, Wikramaratna and Gupta, 2015).

These theoretical predictions are supported by observations in various pathogenic species (Hinnebusch *et al.*, 1998; Areschoug *et al.*, 1999; Johnson *et al.*, 2006; Polley *et al.*, 2007; Buckee *et al.*, 2010), and are apparent at an amino acid level (Chapter 3). If a particular amino acid combination at certain sites in pathogen proteins allows a highly protective allele-specific immune response to be raised, the only other variants that can co-circulate are those which evade that immune response. Thus, one can look at sequence data from multiple isolates of the same pathogen with the expectation that immunodominant sites will have discretized into non-overlapping combinations, showing particular co-variation patterns: changes in those amino acid residues will occur concurrently. Hence, by interrogating large sequence alignments for sites in the genome that exhibit co-variation, we may be able to elucidate regions that are under strong selection pressure. Epitopes identified in this way could be highly immunogenic as part of a vaccine construct, and the combination of multiple variants into a single therapy could induce immune responses that would transcend all of the antigenic variation present in the pathogen population.

This approach, termed Reverse Immunodynamics, can be used to assist in the design of novel, highly efficacious vaccines against variable pathogens. It is sensitive to the evolutionary history and variation in the pathogen, thus can produce vaccines that elicit an immune response that pathogens are less likely to escape. Previous efforts to develop vaccines often fail to take antigenic diversity into account, producing vaccines whose efficacy in the field (where individuals may be infected by one of a range of strains) often fails to match efficacy in the clinic (where their effectiveness is sometimes only tested against single strains) (Ewer *et al.*, 2013; Hodgson *et al.*, 2014; Laurens *et al.*, 2017). In contrast, our approach identifies variable sites that are likely

to be immunogenic and that could comprise a high efficacy, strain-transcending combination vaccine.

In later chapters, I present a comprehensive introduction into the patterns of strain structure that emerge as a result of immune selection pressure, and test the means by which strain structure can be quantified utilizing sequence data simulated by existing multi-locus mathematical models (Chapter 2). I then go on to utilize those methods by applying them to large banks of sequence data derived from pathogens of global health importance, putatively elucidating amino acid residues that may comprise epitopes for use in novel vaccine formulations (Chapter 3).

1.4 Mathematical modelling of infectious diseases and interventions

The most successful disease eradication interventions to date have centred around the widespread use of highly efficacious vaccines (Elwood, 1989; Morens *et al.*, 2011; Groce, Banks and Stein, 2014). Successful vaccination campaigns must use a vaccine that induces a highly protective and broadly neutralizing immune response, and also maximize the impact of the vaccine through careful and deliberate design of administration programs. By reducing the number of individuals infected, vaccines can reduce transmission of an infectious disease thus (through "herd immunity") widespread vaccination reduces rates of disease even in those individuals who have not received immunization. The success of a vaccination campaign is also dependent on its interactivity with demographic, ecological, and evolutionary processes, as well as other concurrently administered interventions. In conjunction with considering pathogen diversity, it is imperative that the epidemiology of a disease, particularly in light of intervention, is well understood.

A better understanding of transmission and the impact of intervention can be gained through the use of simplifying mathematical models. Early mathematical models of disease were developed in 1927 (Kermack and McKendrick, 1927), and consider a host population comprised of three types of individual: those who are yet to be infected (susceptible, S), those who are currently infected (infected, I), and those who are recovered from infection (recovered, R). In the most basic form of the model, some simplifying assumptions are made: individuals can only be infected once (i.e., they acquire totally protective immunity) and are totally susceptible prior to infection, and that the disease does not induce mortality. The Susceptible-Infected-Recovered (SIR) model can be enacted by a system of Ordinary Differential Equations (ODEs, below) with the movement of individuals between compartments for each class (S → I → R) dictated by rates of infection and recovery:

$$\frac{dS}{dt} = \mu(S + I + R) - \beta SI - \mu S$$

$$\frac{dI}{dt} = \beta SI - \sigma I - \mu I$$

$$\frac{dR}{dt} = \sigma I - \mu R$$

where $\frac{dx}{dt}$ is the change in size of compartment x with respect to time, μ is the birth and death rate, β is coefficient of transmission, and σ is the rate of clearance of infection.

As introduced previously, we can quantify the overall reproductive fitness of a pathogen through the basic reproductive number, R_0 (the average number of new infections that are borne of a single infection in a totally susceptible population), which is defined in the SIR model as a function of the rates of infection (β), clearance of

infection (σ), and births/deaths (μ). Births and deaths cause a net reduction in infection of the population: some of the individuals who die are infected, but all individuals are born susceptible. As R_0 is the ratio of the transmission coefficient and the combined rates of birth-death and clearance of infection, it follows that an R_0 of less than 1 can occur when the former is greater than the latter, that is, when new infections happen more slowly (β) than the clearance of infections (σ) and the replacement of individuals with new, susceptible individuals (μ). Therefore, if the R_0 of a particular pathogen is less than 1 then the outbreak will not continue, whereas an outbreak with an R_0 of greater than 1 will, according to the model, continue (Anderson and May, 1991). The equation for R_0 can be defined as:

$$R_0 = \frac{\beta}{\sigma + \mu}$$

R_0 as an intrinsic measure of fitness is useful for theoretical works and some disease systems, however in order to consider interventions against disease we must consider the effective fitness of a pathogen (R_e). As well as being dependent on all of the factors that affect R_0, R_e is also dependent on (often transient) naturally-acquired and vaccine-induced immune pressure, and the presence of other intervention strategies. Interventions that reduce the R_e of an outbreak to < 1 are likely to be successful, although stochasticity, strain-specific efficacy of interventions, heterogeneities in coverage, or intra-host and intra-parasite differences in the effectiveness of treatment can lead to some strains remaining at an R_e of > 1. In some cases, this affords the surviving strains an expanded niche through which they may propagate, further facilitating escape of the pathogen from the effects of the intervention (White *et al.*, 2010).

Accurate estimation of the proportion of a population who are required to be vaccinated in order to drive eradication of an infection is a key topic in infectious disease epidemiology, and is of utmost importance when designing a vaccination campaign, as is conducted in Chapter 5. Campaign design must factor in the efficacy of a vaccine and any loss of efficacy over time, heterogeneities in exposure and vaccine response between host individuals (White *et al.*, 2010), the willingness of a population to be vaccinated, the age at which individuals are to be vaccinated, as well as the potential impact of any loss of naturally acquired immunity that occurs as a result of an infection blocking vaccine. Comprehensive reviews of the challenges faced when modelling these factors and others are provided elsewhere (Childs *et al.*, 2015; Funk *et al.*, 2015; Metcalf *et al.*, 2015; Wikramaratna *et al.*, 2015), and some considerations relevant to the work presented here shall be discussed in full in Chapter 5.

Successful eradication campaigns that target multi-strain infectious diseases are likely to require a careful and considered combination of multiple interventions. In the case of malaria, there are a wide range of interventions that have been successful in reducing incidence and mortality globally, however in order to elimination to be achieved, we should consider the potential synergistic effects that could be borne of multiple prongs of attack. In this thesis, I utilize mathematical models to test the impact of such combinatorial interventions at both within-host and population levels.

1.5 Thesis outline

The ways in which we can better our understanding of the dynamics of infection and use that understanding to optimise vaccination strategy have been discussed in the Chapters that follow. The works presented within this thesis cover multiple key stages

of vaccine development; from understanding the evolutionary history of diverse pathogens and bettering the ways that we can study it, to putatively identifying antigenic regions of parasites that could comprise novel vaccine targets, to the ways in which multiple existing or hypothetical vaccines could be used in combination with each other, and finally in considering the population-level impact of vaccination and its combination with existing interventions in malaria-endemic regions. The structure of this thesis is as follows:

Chapter 2 further introduces multi-locus, multi-strain modelling, and uses existing mathematical models of strain dynamics to develop means by which we can quantify indications of immune selection in a variety of simulated pathogen populations. I evaluate various quantitative methods for analysing sequence data, and develop a novel metric that corrects some of the shortfalls of previously used measures, namely a lack of robustness to increasingly complex antigenic structures and variations in diversity.

Chapter 3 sees the application of this novel metric to real sequence data. Amino acid alignments are interrogated for co-varying sites and patterns of non-overlap (indicative of strain structuring) from three pathogens of global health importance: i) a bacterium, *Neisseria meningitidis*, whose patterns of strain structure are known and well understood; ii) a single protein from the malaria-causing protozoan *Plasmodium falciparum*; and iii) the whole genome of the Epstein Barr Virus. Loci in the pathogens are found to exhibit traits indicative of immune selection pressure, and therefore may comprise vaccine targets that would be immunogenic and, if variants were used in combination, would produce responses effective against the diversity of pathogen variants seen in natural systems.

Chapter 4 presents a mathematical model of malaria on a within-host scale. It is argued that existing low- to moderate-efficacy vaccine candidates may combine synergistically to a high-efficacy regimen. Such a vaccination strategy may be comprised of a combination of components already in the late stages of clinical development, or that have recently been licensed for administration. I provide a mechanistic understanding of this hypothesis, indeed suggesting that synergy is a likely result of a combination of vaccines that act against different pre-erythrocytic stages of infection with *P. falciparum*. A published manuscript corresponding this chapter is included in the appendices (Appendix D, Walker *et al.*, 2015).

Finally, Chapter 5 considers the potential impact of mass vaccine administration in combination with other interventions on *P. falciparum* malaria rates in endemic regions. I use data sourced from field trails of the RTS,S vaccine to fit mathematical models of transmission that are sensitive to seasonality in force of infection and to waning vaccine immunity, providing novel estimates of efficacy robust to those factors and to study design. To date, RTS,S has primarily been considered as an infection blocking agent in young children; here, I consider its potential utility as a part of a mass treatment campaign, one which would aim to block malaria transmission and thus would reduce infection rates even in those individuals not receiving the vaccine. I go on to estimate the potential efficacy of a combination of mass vaccine administration with existing mass drug administration and transmission blocking interventions, again suggesting that a profound impact may be had upon disease rate using tools already at our disposal.

Chapter 2 Reverse Immunodynamics: theory and development of metrics

Andrew S. Walker, Dr José Lourenço, and Professor Sunetra Gupta

2.1 Introduction

As introduced in *Chapter 1*, the unrelenting arms race between parasites and their hosts has shaped the evolution and life history strategies of each. Parasites employ a range of means to maintain a route of transmission, either by attempting to avoid destruction by the immune response, or by evolving ways in which responses may be tolerated. Hosts respond to infection by the production and release of immune effectors aimed at destroying the infectious agent. Host immune cells act against particular components of a parasite, called epitopes, providing selection pressure which drives dynamic properties drastically different to non-immunogenic regions of the genome (Brunham, Plummer and Stephens, 1993; Lipsitch and O'Hagan, 2007; Wisniewski-Dyé and Vial, 2008; Chen and Deng, 2009). By studying patterns of diversity in antigenically variable pathogens, I can elucidate regions of the pathogen that are subject to immune selection pressure.

When studying populations of pathogens, it is important to consider that each antigenic variant does not exist in solitude in the antigenic landscape: immune responses against one variant inevitably have some impact upon the fitness of others, thus the effects of immune selection pressure play out at a population level. The effect that an immune response against one antigenic variant has upon related variants is termed "cross-immunity", the level of which has been shown to play a key role in determining the dynamics of pathogen populations and, by abstraction, their hosts (Gupta *et al.*, 1996; Gupta, Ferguson and Anderson, 1998; Penman *et al.*, 2013). By studying particular properties of populations of pathogens (such as patterns of non-overlap between variable sites in a parasite's proteins, below), I can reveal the impact of selection upon them, as well as elucidate the amino acid mutations at particular

sites that allow for the protein to remain functional and so the parasite to remain viable.

The role of immune selection pressure in driving pathogen dynamics has been well studied and several theoretical frameworks presented (Gupta *et al.*, 1996; Gog and Grenfell, 2002; Gomes, Medley and Nokes, 2002; Kryazhimskiy *et al.*, 2007). Seminal work by Gupta *et al.* presented a model in which a range of multi-strain pathogen population dynamics can be reproduced (Gupta, Ferguson and Anderson, 1998). This includes three discrete dynamic behaviours that the antigenic types of a pathogen population may exhibit (herein termed strain structures) which demonstrate the central role that cross-immunity plays in determining the strain structure that a population takes.

Firstly, when there is no cross-immunity, all antigenic variants circulate as if independent. Immune responses to one strain have no impact upon the others, and so the prevalence of each strain does not differ from that in a classic single-strain SIR model. Systems that exhibit this behaviour are described as having "No Strain Structure" (NSS).

In the presence of cross-immunity, the model predicts that structure can emerge in two particular forms. At intermediate levels, Cyclic Strain Structure (CSS) arises, in which, at any given time, there are one or a small number of variants that dominate; the set of antigenic variants that dominates quickly switches (Gupta, Ferguson and Anderson, 1998). This pattern is exemplified by influenza infection in which, in each season, there are a number of strains that dominate infection (Andreasen, Lin and Levin, 1997). Within CSS there are a number of structural subclasses that occur across

the range of cross-immunity (Cherif *et al.*, 2017): at low levels of cross-immunity (within CSS) strains switch very rapidly between single or small groups of dominant types, however as cross-immunity increases behaviours become more periodic, that is, sets of strains dominate for an increasing amount of time.

Finally, at very high levels of cross-immunity, competitive exclusion occurs between related antigenic types. Here, the strains segregate into subgroups of variants containing non-overlapping combinations of alleles, and this pattern is termed Discrete Strain Structure (DSS). That is, the only strains that persist are those targeted by immune responses that are not cross-reactive against other persistent strains.

The emergence of DSS can be considered using a hypothetical pathogen population in which there are 2 antigenic loci that are targeted by immune responses. At each of the loci there can be two possible variants – at locus 1 there can be variant A or variant B, and at locus 2 there can be variant X or variant Y. This system therefore gives rise to four possible antigenic types, herein termed "strains" – AX, AY, BX, and BY. Assuming high cross-immunity, an individual who has been infected with AX will be immune to infection with AY (because of the shared A), and also with BX (because of the shared X) – they will, however, be susceptible to infection with BY. Thus, at a population level, solely the strains AX and BY (or strains AY and BX) would be able to co-circulate, while all other variants would be driven to extinction. The alleles that define the persistent strains in this example therefore show a pattern of "non-overlap" – that is, when there is an A at allele 1, there is always an X at allele 2, and when there is a B at allele 1, there can only be a Y at allele 2 (or *vice versa*).

Evidence of this pattern has been found across a number of important human pathogenic species (Hinnebusch *et al.*, 1998; Areschoug *et al.*, 1999; Johnson *et al.*, 2006; Futse *et al.*, 2008; Buckee *et al.*, 2011), and numerous pieces of research have sought to quantify non-overlap between amino acid residues or antigenic regions. Here I build on previous work to present a systematic overview of the theoretical underpinnings of non-overlap, and use sequence data simulated by multi-locus mathematical models of strain structure to develop metrics suitable for robust quantification of natural systems.

A metric that has been previously applied to quantify co-evolution between nucleotide or amino acid sites in multiple sequence alignments (MSAs) is Mutual Information (MI) (Gloor *et al.*, 2005; Little and Chen, 2009). MI is derived from information theory and is a measure of the mutual dependence of two variables (Kullback, 1959). When applied to sequence data, it captures how much information can be gained about one amino acid site from another. For example, when two sites are under total non-overlap (i.e., when site 1 is an A site 2 is always an X, and when site 1 is a B, site 2 is always a Y) then it is possible, with absolute certainly, to ascertain the identity of one site by knowing the other.

Another metric used to study non-overlap is the f^* statistic (Callaghan *et al.*, 2008; Buckee *et al.*, 2010). This was devised to allow investigation of the evolution of antigenic repertoires, with the authors finding that stable combinations of antigenic repertoires were maintained in a meningococcal species over an extended time period, even given high recombination rates. This statistic was related to immunological cross-immunity in both simulated and sampled sequence data, and measures the

extent of non-overlapping structure by considering the dominance of each allelic combination.

A classically utilized measure for measuring non-random associations between genetic loci is linkage disequilibrium (LD) (Lewontin and Kojima, 1960). This measure compares the observed frequency of allelic combinations with those expected to occur by chance, and has been the subject of much study. As with metrics discussed here, LD can be used to identify sites which are highly predictive of each other, however it has some shortcomings which preclude it from further use in this work. Several previous works have noted that LD is frequency dependent, and have sought to develop metrics based on LD which can partially circumvent this intrinsic bias (Lewontin, 1988; discussed in full by Gianola, Qanbari and Simianer, 2013). This bias means that LD-based metrics require careful interpretation: comparisons of LD between populations are notoriously difficult due to variation in allele frequencies.

In addition, whilst LD can be used to predict one locus from another, it does so independently of diversity. In other words, non-overlap in which the primary variant comprises the majority of cases will be predicted by LD metrics to be in linkage (non-overlapping) as one site can be predictive of the other. Of greater interest to this work, however, is pairs in sites in which there is non-overlap *and* the non-overlapping allelic variants are present in reasonable quantities – indicative of immune pressure diversifying the parasite into non-overlapping antigenic groups.

Here, I wish to develop a metric that not frequency dependent and that explicitly measures diverse non-overlap as opposed to linkage or non-overlap with low frequency minor variants: thus, I elect to develop metrics based on information theory.

I should also note that the metrics developed here to analysis non-overlap do so independently of the phylogenetic history of the pathogen. As such, careful *post hoc* analyses should be performed in order to tease apart non-overlap that is driven by immune selection from that which comes as a result of a shared phylogenetic history. Such analyses are discussed in more detail in the Discussion of this chapter, and performed on our pathogens of interest in Chapter 3.

Previous theoretical work has also analysed *in silico* evolved MSAs to examine the role of evolutionary processes in driving the co-evolution of residues, coupled with analyses of real sequence data to find amino acid residues that may be of structural or functional importance either within a protein or by interaction with a substrate (Martin *et al.*, 2005). The *in silico* approach that I present here bears some similarity to this, although interactions that I elucidate between residues are viewed in light of strain theory, and so are attributable to immune selection pressure. In a later chapter, I more thoroughly discuss disentangling the various potential sources of non-overlap.

As mentioned above, several works have sought to quantify patterns of non-overlap: I here rigorously test and improve upon these metrics by analysing sequence data generated from multi-locus models of strain structure. My novel metrics correct some of the shortfalls encountered when studying complex strain structures, paving the way for later work which identifies novel epitopes within real genetic sequence data. These epitopes could be candidates for incorporation into a novel vaccine construct: their patterns of non-overlap indicate that they are naturally immunogenic and, even under sustained evolutionary pressure, have evolved into only a limited set of variants. A multi-allelic vaccine incorporating variants from these epitopes would thus have the

potential to replicate the naturally acquired immunity that would occur under exposure to all possible strains. This formulation would protect against infection and limit the potential for the emergence of vaccine escape variants.

2.2 Methods

2.2.1 Analysing pairs of allelic variants

In order to draw conclusions about strain structure in pathogen populations, as well as to find evidence of immune selection pressure, I here present a number of methods to quantify non-overlap between all possible pairs of amino acid sites within the protein sequence of a pathogen. It is important that methods are robust to all possible topographies of antigenic space: to varying levels entropy (i.e., diversity), morphism (the number of different amino acid residues present at a single site), and symmetry (interactions between highly polymorphic and di- or trimorphic sites).

The combinations of allelic variants at a pair of sites can be visualised using an allelic matrix, as shown in Figure 2-1. Site I and site II are two positions in a multiple sequence alignment: at site I there can be either variant “a” or “b”, and at site II there can be variant “x” or “y”. The allelic matrix shown gives the number of isolates that have each variant pair (of the four possible combinations of a/b and x/y) at each site.

Three examples of different levels of non-overlap are shown in Figure 2-1: the first from a system in which there is total non-overlap between the two sites (Figure 2-1A), the second from a system with intermediate levels of non-overlap (Figure 2-1B), and the third from a system where there is no association between the two sites (Figure 2-1C). These figures provide a useful visualisation of the association between a pair of dimorphic sites (herein denoted by {morphism at I, morphism at II}, e.g., {2,2}).

This method necessitates a means of quantifying non-overlap in a way that is robust to different strain structures. Given the variable levels of polymorphism seen across natural proteins, metrics used must not be biased by high polymorphism (i.e.,

unbiased comparisons of a trimorphic pair of sites {3,3} to a system where each site has 10 possible variants {10,10}) or asymmetrical polymorphism, i.e., comparing a dimorphic with a trimorphic site {2,3}, or tetramorphic site {2,4}, and so forth (up to {2,10}). I later present a metric that is robust to such patterns of polymorphism.

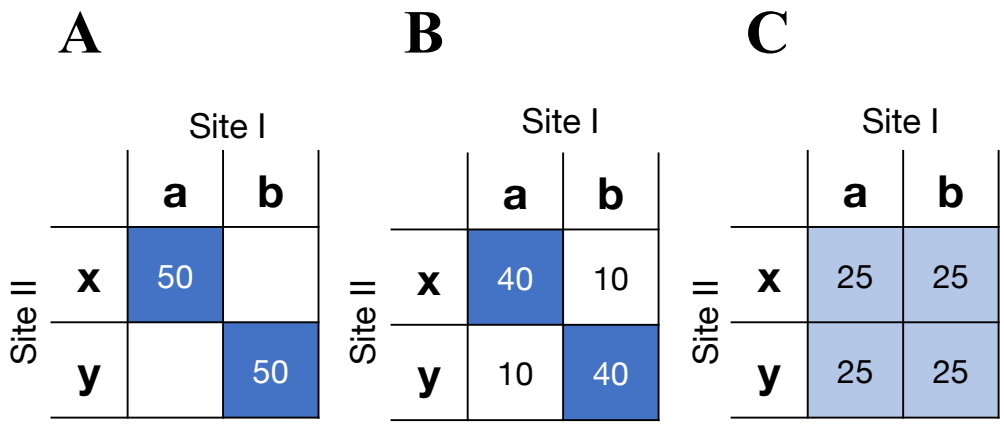


Figure 2-1 Examples of three allelic matrices exhibiting different levels of non-overlap between site I and site II. Both sites are dimorphic: at site I there are variants a and b, and at site II there are variants x and y. In example A, there is total non-overlap between the two sites – when there is an a at site I, there is always an x at site II, and when there is a b at site I there is always a y at site II. In example B, there in some non-overlap, although there are some variant combinations present (i.e., bx and ay) which cause the system to deviate from total non-overlap. In example C, all allelic variants are present in equal quantities – there is total overlap.

2.2.2 Quantifying non-overlap using information theory

Mutual Information is derived from Shannon's entropy, and is a function of the entropy of the two residues in question, as well as their joint entropy, as defined below. Shannon's entropy, H(X), is a measure of the uncertainty of a discrete random variable, X, and k is alphabet length. The possible values of X are defined by an alphabet $K = \{x_1, x_2, \dots, x_k\}$, each letter of which has a probability of occurrence such

that the overall probability distribution is $p(x) = \{p(x_1), p(x_2), \ldots, p(x_k)\}$. The individual entropy of each alphabet member is defined as the product of the probability and log probability, and thus overall entropy is the sum of individual entropies:

$$H(X) = -\sum_{i=1}^{k} p(x_i) \log_b p(x_i) \quad (1)$$

where b is a scalar of entropy. If b=k (i.e., b is equal to the number of possible variants) then the maximum value of entropy is 1. $H(X)$ is maximised when all k alphabet members are present in equal quantity, and minimised when only one alphabet member is present.

Equation (1) defines the entropy for a single variable or, in this case, a single allelic site. Entropy can be expanded to consider the uncertainty of a pair of variables by considering joint entropy, in which the two variables have a joint alphabet of length $k \times l$. In the {2,2} example, the first variable has an alphabet of k = *{a,b}*, and the second an alphabet of l = *{x,y}*, and so the joint alphabet of the two sites is defined as m = *{ax, ay, bx, ay}*. Each joint alphabet member also has an associated probability in $p(x_i, y_j)$. The joint entropy is thus defined as

$$H(X,Y) = -\sum_{i=1}^{k} \sum_{j=1}^{l} p(x_i, y_j) \log_b p(x_i, y_j) \quad (2)$$

A third entropy definition that we are required to consider is the conditional entropies of a pair: that is, the entropy of *X* given that we know *Y*, or the entropy of *Y* given that we know *X*. These can formally be defined as

$$H(X|Y) = H(X,Y) - H(Y) \quad (3)$$

$$H(Y|X) = H(Y,X) - H(X) \quad (4)$$

This can be explained by considering the Venn diagram of entropy space in Figure 2-2A. The joint entropy *H(X,Y)* contains within it the entropies of each variable *H(X)* and *H(Y)*, and so each conditional entropy is the joint entropy minus the entropy of the variable on which the conditional is based. We can thereby deduce that the mutual information (*MI(X,Y)*) is the region of entropy space that is intersected by both individual entropies (Figure 2-2B), a space which can be defined as

$$MI(X,Y) = H(X) - H(X|Y) \quad (5)$$

$$= H(Y) - H(Y|X) \quad (6)$$

$$= H(X) + H(Y) - H(X,Y) \quad (7)$$

Mutual information is thus a measure of the information about X that can be gained from Y. Note that is symmetrical (MI(*Y,X*) is equal to MI(*X,Y*)) and falls in {0,1}.

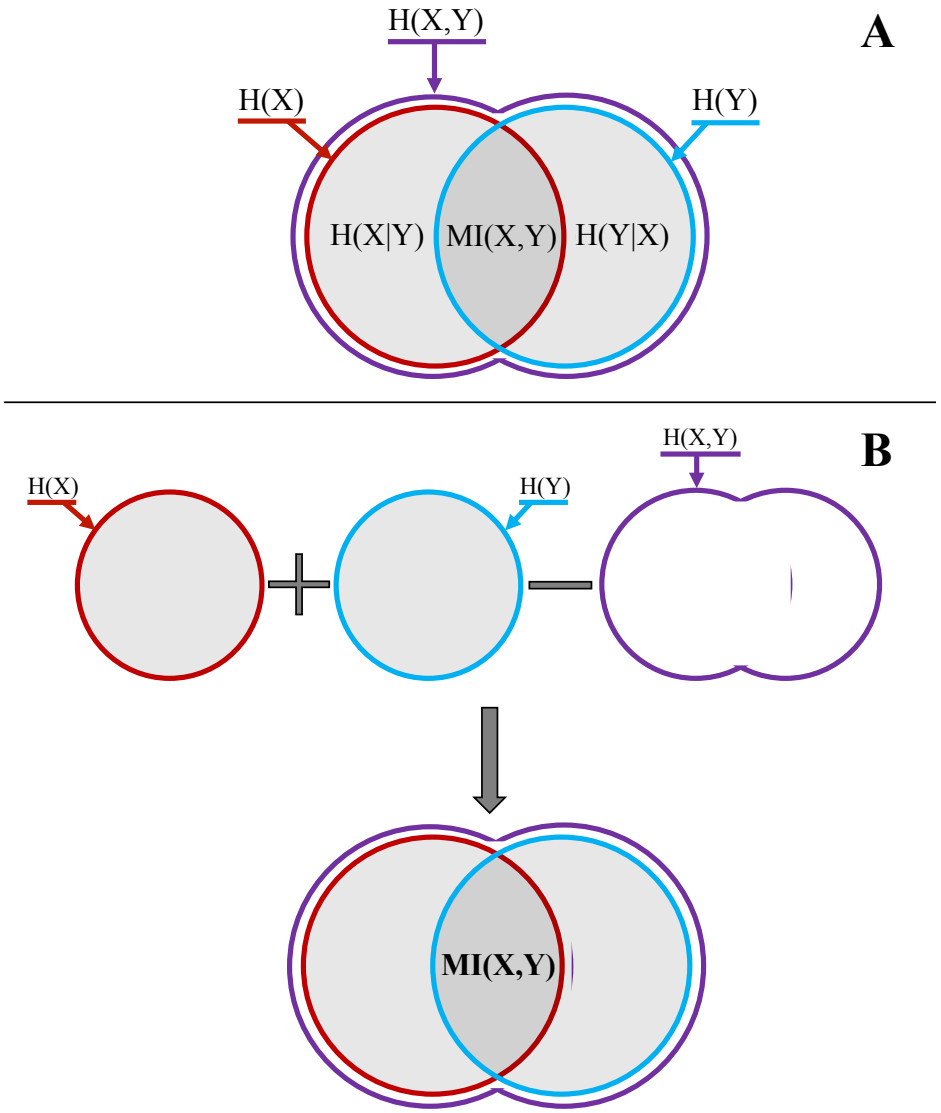


Figure 2-2 **A** Venn diagram of entropy space showing the entropy of two variables (X and Y), their conditional and joint entropies, and their Mutual Information. **B** Transformations of entropy space showing the calculation of Mutual Information from each variable's entropy and their joint entropy.

2.2.3 Quantifying non-overlap using the f statistic*

Buckee *et al.*'s *f** statistic measures the extent of non-overlap by considering the dominance of each allele combination. In a matrix with high non-overlap, the most prevalent allelic combinations will be dominant in both their row and column, and so *f** provides a means with which this dominance can be quantified.

A formal derivation of the *f** metric are given by Buckee et al., 2010 and is as follows. Consider an allelic matrix with two loci, A and B. The per-locus dominance of the highest frequency allelic combination, where p_i is the frequency of the allele i at locus A and q_j is the frequency of the allele j at locus B, is given as:

$$f_A(i) = \frac{x_{ik}^3}{p_i q_k} \text{ for the k that maximises } x_{ik} \quad (8)$$

$$f_B(j) = \frac{x_{kj}^3}{p_k q_j} \text{ for the k that maximises } x_{kj} \quad (9)$$

where the x_{ik} is the frequency of the allelic combination of i at locus A and j at locus B.

Summing all dominance values over both row and column derives a measure of non-overlap for each locus, and the overall degree of non-overlap between two sites is given by the average of each:

$$f* = \frac{1}{2}\left(\sum_i f_A(i) + \sum_j f_B(j)\right) \quad (10)$$

Between a pair of sites with high non-overlap, the most prevalent allelic combinations will be the dominant in both their row and column, and so *f** provides a means with which this dominance can be quantified.

2.2.4 Transforming and analysing asymmetrical matrices

As introduced previously, variable levels of polymorphism between amino acids sites in a protein result in asymmetries between pairs that are compared. Here I present a means with which to "symmetricalize" the allelic matrix, that is, to reduce an $i \times j$ matrix down into an $i \times i$ matrix (where $i < j$) by summing variant frequencies in a way that produces the maximally non-overlapping symmetrical version of the original matrix.

The algorithm devised for symmetricalizing matrices comprises two component steps: first the matrix must be reorganised to be as diagonal as possible, a protocol termed 'trace optimisation'. This rearranges the rows and columns of the matrix so that the most frequent combination of alleles is in position (1,1), and second most frequent combination that does not contain either of the variants that are in (1,1) is placed in (2,2), and so forth. This step is performed using a routine called the Hungarian method (Kuhn, 1955), a combinatorial optimization algorithm typically applied to solve the assignment problem, the classical example of which involves assigning persons to jobs. Say, for example, that workers can each perform one of a number of tasks, however the cost incurred by each individual performing each job differs. The assignment problem thus seeks to assign jobs to individuals in such a way that costs are minimised. The exact steps of the algorithm are formally described in Kuhn, 1955, and here the method is implemented using the "*solve_LSAP*" function of the R package "clue".

In the resulting diagonalised matrix, the rows or columns of the matrix which cause asymmetry (i.e., those that fall outside of an i*i square matrix), are combined with rows or columns in the square matrix such that the sum of the diagonals is maximised (Figure 2-3).

Once matrices have been diagonalised and symmetricalized, I can apply the Mutual Information calculations made previously to generate a new non-overlap score, herein termed "MI_S". As shown later, this measure yields values that are robust to asymmetrical strain systems – note that the MI and MI_S scores for strain systems that are already symmetrical (i.e., {2,2}, {3,3}, etc.) will be identical.

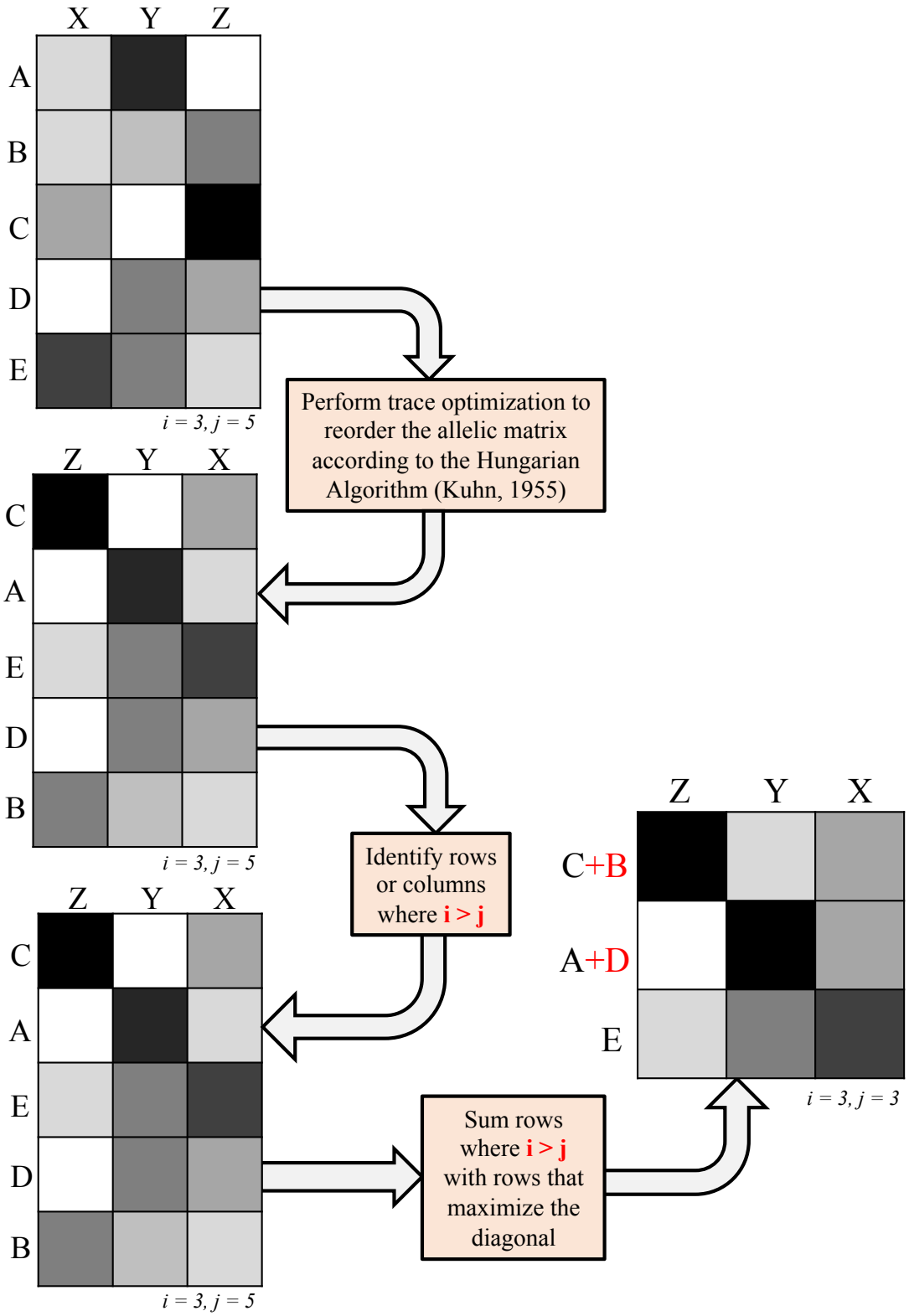


Figure 2-3 Flow diagram showing the steps of the trace optimization and symmetricalisation protocol

2.2.6 Quantifying non-overlap between asymmetrical pairs – W

When considering an allelic matrix, there are two factors that influence the degree of non-overlap that is seen. Firstly, assuming that the matrix is organised in accordance with the above diagonalization protocol, the proportion of all isolates that fall along the trace of the matrix is important. If all isolates fall along the diagonal then there are none than cause the sites to overlap, thus non-overlap should be very high. I may also, however, wish to weight the measure of non-overlap according to the distribution of frequencies along the diagonal. In other words, should a pair of allele combinations distributed with 50% of the isolates belonging to each combination (i.e., 50% AX and 50% BY) be deemed to exhibit equally strong non-overlap as a pair with a weighted distribution (for example, 80% AX, 20% BY)? If I conclude that these two matrices do differ in their non-overlap, then an unbiased metric will be one that is a function of both the total proportion that fall along the diagonal, and their distribution along the diagonal.

Based on this logic, I here present a novel metric, termed W. It is calculated from allelic matrices that have been made symmetrical and had their trace optimized, as described above. I effectively consider each of the variant pairs along the diagonal of the matrix (AX, BY, etc.) to be strains, and so can utilize a standard measure of species diversity to quantify their diversity (Shannon's diversity index, *H*). Note that this measure is analogous to Shannon's entropy, described above, and is given by

$$H = -\sum_{i=1}^{I} p(x_i) \ln p(x_i) \tag{11}$$

where the allelic matrix has dimensions *I x I,* with $i = \{1, 2, \ldots, I\}$.

W is defined as a function of H divided by the natural log of I (thus scaling W to a maximum value of 1) and the proportion of isolates pairs that fall on the diagonal (P):

$$W = \frac{H}{\ln(I)} \times P \tag{12}$$

In symmetrical allelic matrices in which all of the isolates fall along the diagonal, W is equal to the diversity (H) of the isolate frequencies. When the proportion of isolates in each of the strains is equal in a {2,2} system, H is equal to the logarithm of the number of species present. As the proportion increases in either direction, i.e., as one non-overlapping strain dominates, H tends to 0 (Figure 2-4). N.B., in Figure 2-4, a H of 0 is never reached as H cannot be computed on a matrix in which there is only a single value > 0.

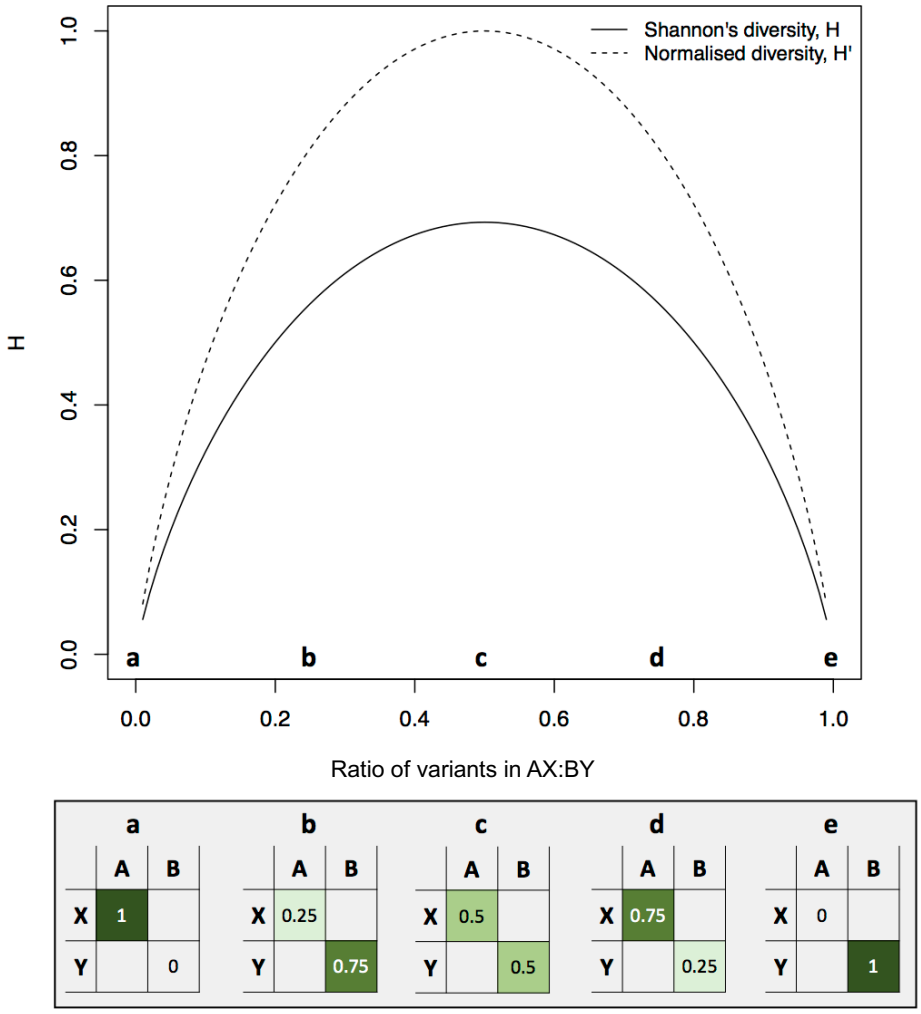


Figure 2-4 Shannon's H in totally non-overlapping strain structures with changing ratios of AX:BY prevalence. In box: example allelic matrices at discrete points along the spectrum of ratios in {0,1}.

2.2.7 Mathematical modelling of multi-strain systems

The MANTIS package simulates a multi-locus model of multi-strain pathogen evolution by wrapping a system of C/C++ Ordinary Differential Equations (ODEs) and a Runge-Kutta solver in a set of R functions (Lourenço, Wikramaratna and Gupta, 2015). The model that underlies MANTIS is from seminal work by Gupta et al. and is described in full elsewhere (Gupta, Ferguson and Anderson, 1998; Recker *et al.*, 2007; Buckee *et al.*, 2011). In short, it is a dynamic epidemiological model of a theoretical disease system in which individuals are susceptible to, infected with, or have been exposed to, and thus developed a protective immune response against, certain strains of a multi-strain pathogen. The model allows for immunity raised against different strains of the pathogen to vary, and for cross-immunity to allow the protection (or partial protection) to develop against multiple strains. The antigenic structure of the pathogen is defined, as are the transmission and recovery rates, the lifespan of the host, and the impact of cross-reactive immune responses each strain induces upon other strains that share antigenic components.

2.2.8 Generating Multiple Sequence Alignments

An output from MANTIS is time-series data for the prevalence of each antigenic variant. By sampling strain prevalences from the model in time, and using them to generate sequence data, I replicate real-world sampling of genetic data from populations of pathogens, allowing me to robustly test, within a well-characterised simulation, methods that can quantitatively assess strain structure in natural populations.

First, I set a frequency with which to sample from the model. Strain relative prevalences (frequencies) from MANTIS at each time point are given as a proportion

of the total population. I set an effective sampling intensity by saying that for a given small proportion of the population, 1 sequence will be "sampled". As a balance between accuracy and computational tractability, I say that for each 10^{-7} of the population, sequence data for one isolate is generated: any strains that exist at below 10^{-7} will not be sampled at that time point.

The ODE system used in MANTIS is deterministic and thus will, if all parameters are kept the same, produce identical outputs over multiple runs. Overall strain dynamics are sensitive to the initial prevalence of each of the antigenic variants. For example, in a four-strain system under DSS the pair of non-overlapping strains that have the initial advantage will, all else being equal, propagate and outcompete the others. In order to generate robust results, I therefore performed 200 runs of the model for each parameter set, each with initial conditions independently sampled from a uniform distribution between ~0 (to a minimum double-precision floating point minimum greater than 0) and 10^{-4}. In all simulations, there is a transient phase of the model before equilibrium is reached, thus results are analysed from time points only once equilibrium conditions have been reached (as assessed by eye) or from when it is clear that cyclic or pseudo-chaotic behaviours will persist.

2.3 Results

Here I present analyses of various measures of non-overlap in systems of varying complexity. Sequence data generated manually, as well as by using well-characterised mathematical models of pathogen dynamics, allow us to robustly test metrics of non-overlap across a full spectrum of strain structures.

2.3.1 Simulating symmetrical dimorphic pairs of sites

First, I investigate how each measure of non-overlap changes over a single simulation of a two-locus, two-allele {2,2} system in MANTIS with the strains AX, AY, BX, and BY. I present outputs from the model at the three values of cross-immunity that have previously been shown to typify the three strain structures: γ = 0.15 for NSS, γ = 0.75 for CSS, and γ = 0.95 for DSS (Figure 2-5, Figure 2-6A-C, Lourenco et al., 2015), with the model run with transmission and recovery parameters fixed and 500 years of simulation.

It is clear that at high and low values of cross-immunity there is a transient phase of the model, as evidenced by highly fluctuating absolute (Figure 2-5) and relative (Figure 2-6A-C) strain prevalences. At all levels of cross-immunity, in the pre-equilibrium phase, there is rapid switching between high and low disease states and therefore the number of sequences sampled is dependent on whether the time point selected for sampling happen to fall on a peak or trough of these transient dynamics. Thus, as described above, values of the metrics are only considered once equilibrium is reached. In all runs of the model, independent of initial conditions and for levels of cross-immunity at which equilibrium is possible (i.e., when the system is not in CSS), equilibrium (as assessed by eye) is reached by t=250. Systems in CSS exhibit cyclic or

pseudo-chaotic dynamics, and so do not reach stable equilibrium. As described previously, for a given proportion (10^{-7}) of the total population infected, sequence data for one isolate is generated (Figure 2-6D-F).

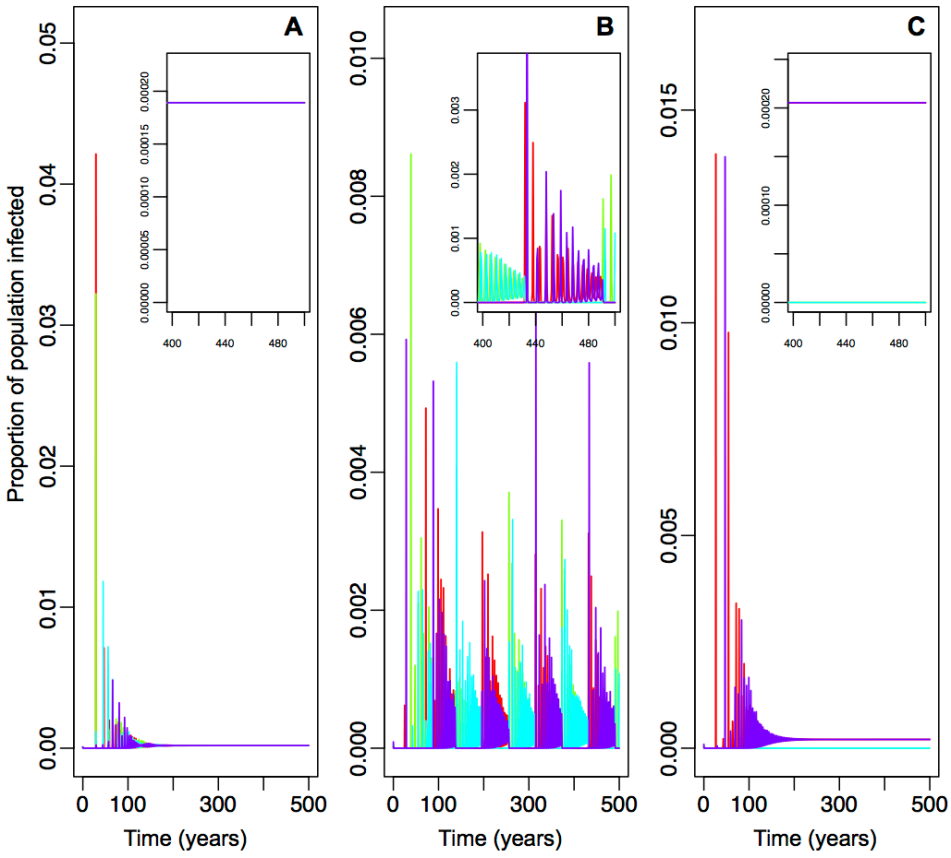


Figure 2-5 Absolute strain frequencies generated by the MANTIS multi-strain multi-locus simulation of a 2-locus, 2-allele (four strain) system at cross immunity levels of 0.05 (A), 0.75 (B), and 0.95 (C). 500 years of the model are shown in main plots, with final 100 years shown inset.

All of the metrics tested are sensitive to changes in strain structure born of differing levels of cross-immunity (Figure 2-7A). The information theory metrics (MI and MI_S) fall on a range from 0 to 1, whereas the minimum value of W is 0.5. When there is strong overlap between a pair of residues (NSS), each variant will exist in approximately equal quantities, and so the proportion of sites along the diagonal will

be 0.5 and the species diversity of those variants (0.5 and 0.5) will be 1, producing a W of 0.5. Similarly, the f^* metric does not scale to 0 – the minimum value of f^* in this system is ~0.125. I later show that the minimum values of W and f^* are dependent on the size of the allelic matrix, that is, the level of polymorphism at each pair.

In symmetrical systems and at all levels of cross-immunity, MI and MI_S scores are identical (Figure 2-7B): the methodologies are equivalent with the exception of the symmetricalisation step applied in MI_S.

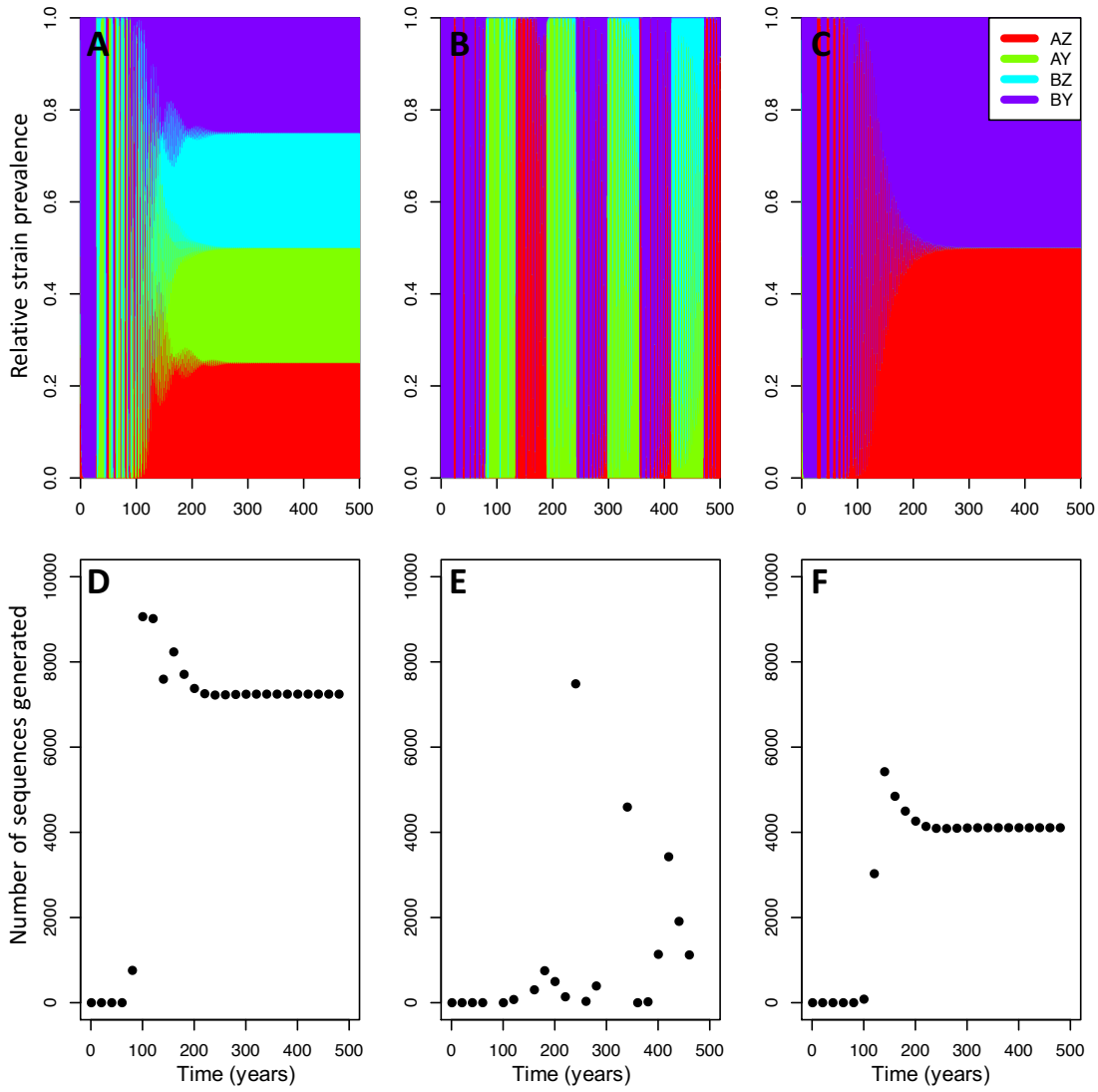


Figure 2-6 A-C: Proportion strain prevalences from the MANTIS simulations of a {2,2} system. D-F: the number of sequences sampled from MANTIS at each sampling point (1 sequence sampled per 10^6 of the proportional population)

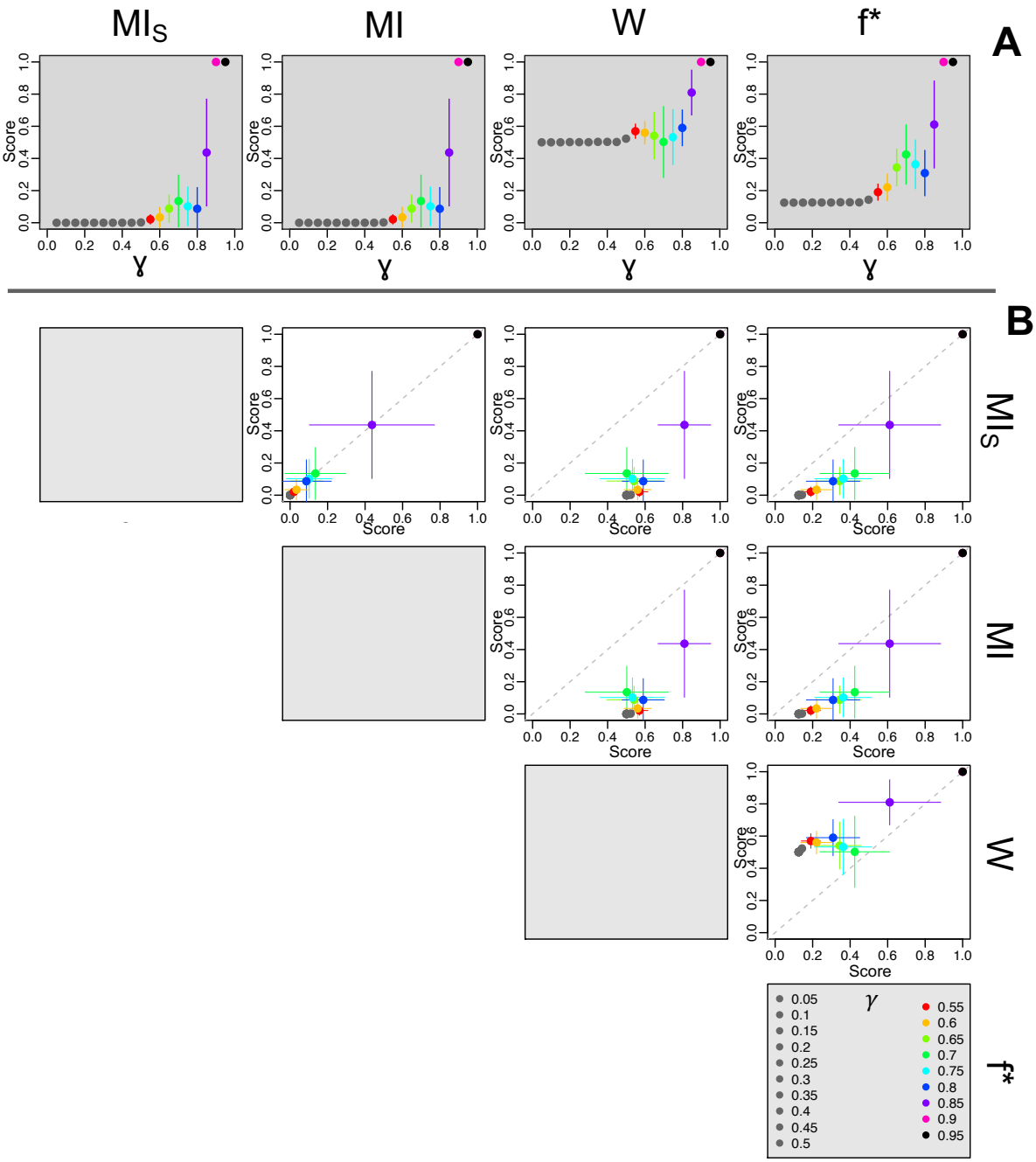


Figure 2-7 A: The mean and SD of the metrics of non-overlap in a {2,2} strain system with varying cross-immunity (indicated by point colours), and B: pair-wise comparisons of the metrics

2.3.2 Simulating polymorphic symmetrical pairs of sites

Symmetrical multi-strain systems exhibit identical patterns of strain structuring to those seen in lower morphism symmetrical systems. That is, at low levels of cross-immunity, the 9 strains present in a {3,3} system (AX, AY, AZ, BX, BY, BZ, CX, CY, CZ) reach an equilibrium in which all are equally prevalent (Figure 2-8A; Figure 2-9A); at intermediate levels of cross-immunity the strains cycle (Figure 2-8B; Figure 2-9B); and at almost total cross-immunity (Figure 2-8C; Figure 2-9C) three non-overlapping strains dominate.

As above, all metrics are sensitive to changes in strain structure (Figure 2-10A). In these symmetrical systems, MI and MI_S again yield identical results, however the minimum values attainable by W and f* differ from the {2,2} system.

By simulating non-overlapping allelic matrices of increasing size, that is, matrices in which all non-overlapping variants are present in equal proportions, I can test the robustness of the metrics to polymorphism (Figure 2-11). This result demonstrates that the MI measures (MI and MI_S) applied to symmetrical matrices will always produce a minimum value of 0: when all variants are present in equal proportions the amount of information about one site that can gained from the other is zero. The same is not true of f* and W: the minimum value of these metrics tends to 0 as the dimensions of the allelic matrix tend to infinity (Figure 2-11).

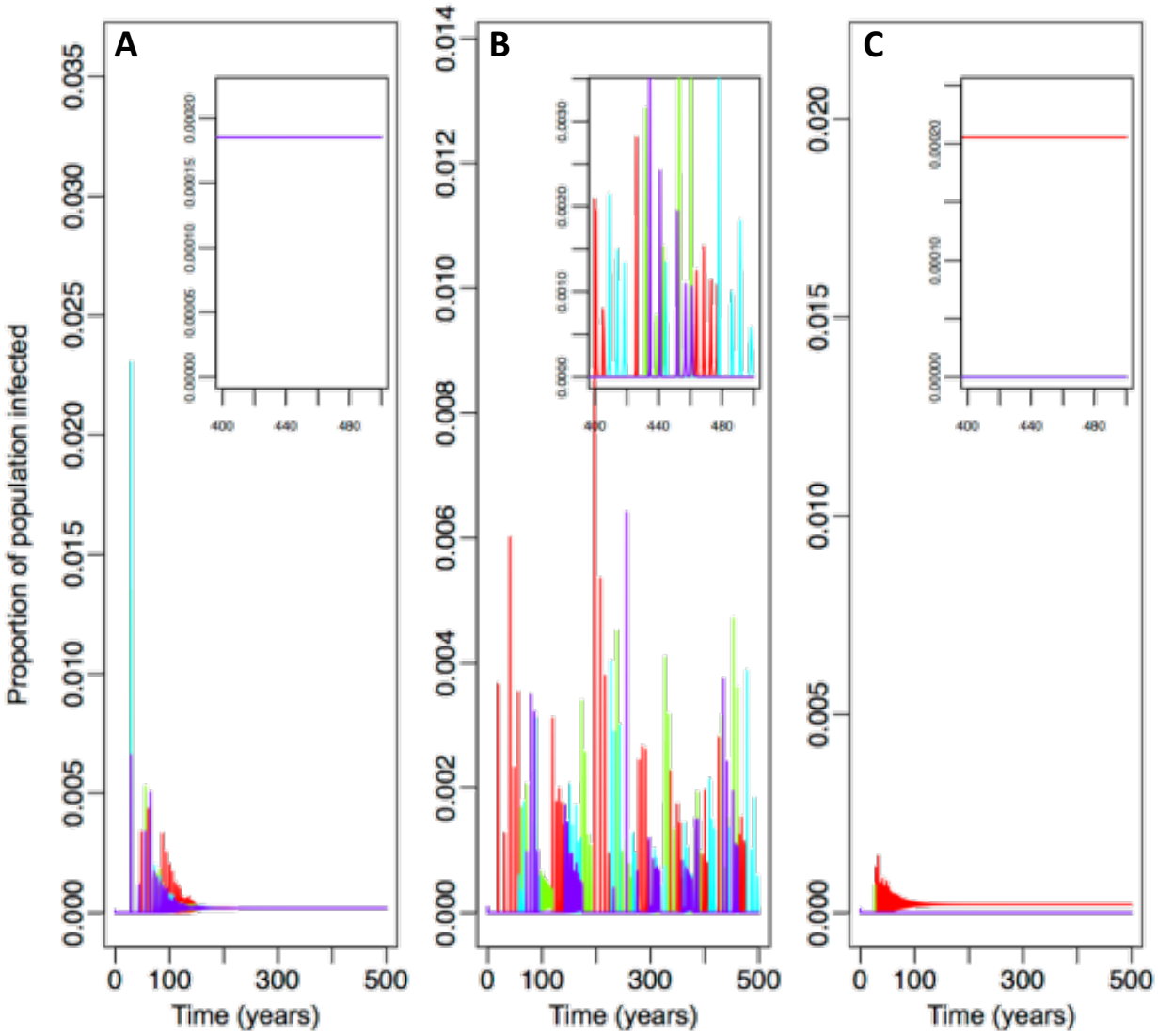


Figure 2-8 Absolute strain frequencies generated from the MANTIS simulation of a 3-locus, 3-allele, 9-strain system at cross immunity levels of 0.05 (**A**), 0.75 (**B**), and 0.95 (**C**). 500 years of the model are shown (main plots), with the final 100 years shown (insert).

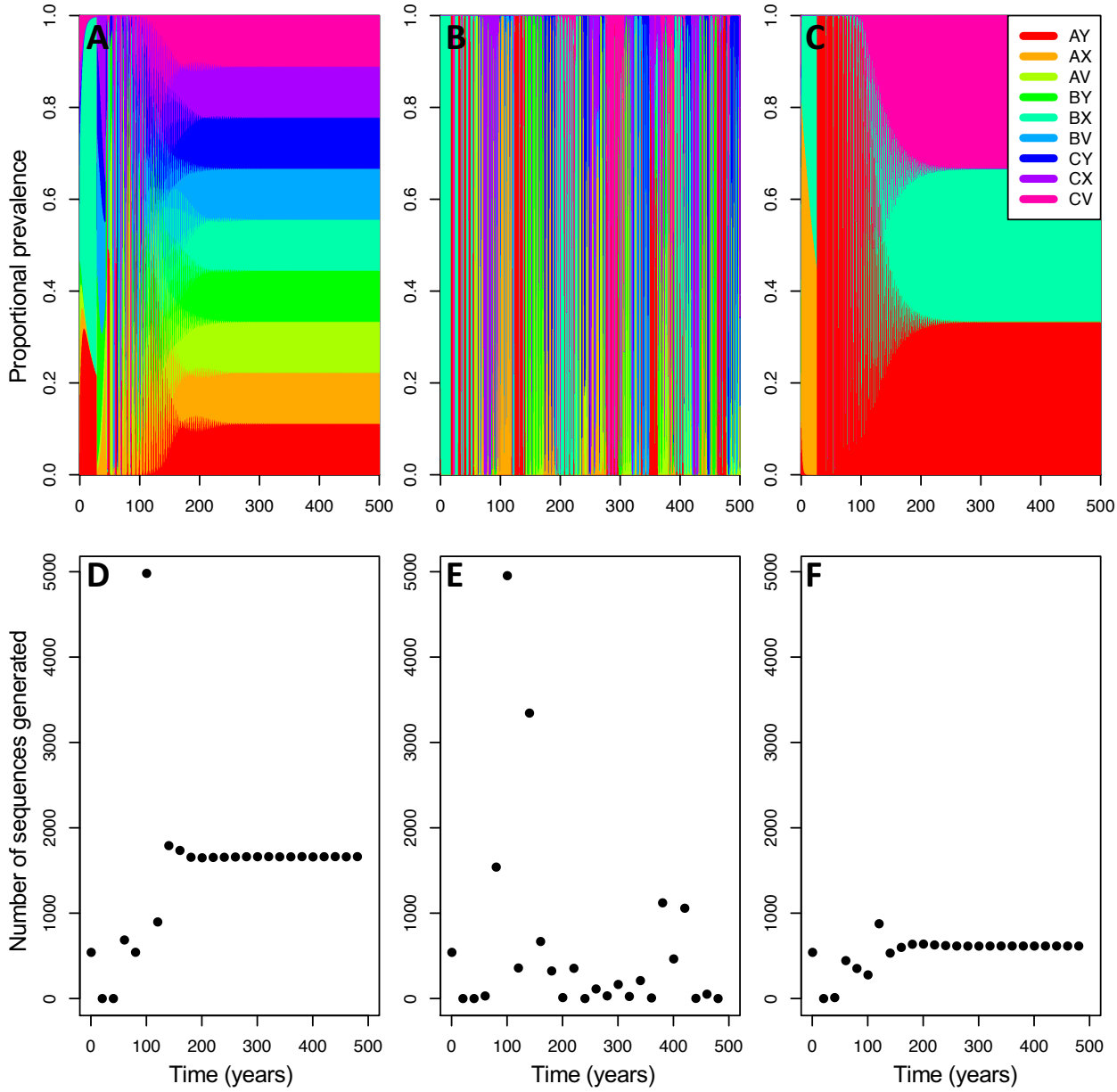


Figure 2-9 **A-C**: Proportion strain prevalences from the MANTIS simulations of a {3,3} system. **D-F**: the number of sequences sampled from MANTIS at each sampling point (1 sequence sampled per 10^6 of the proportional population)

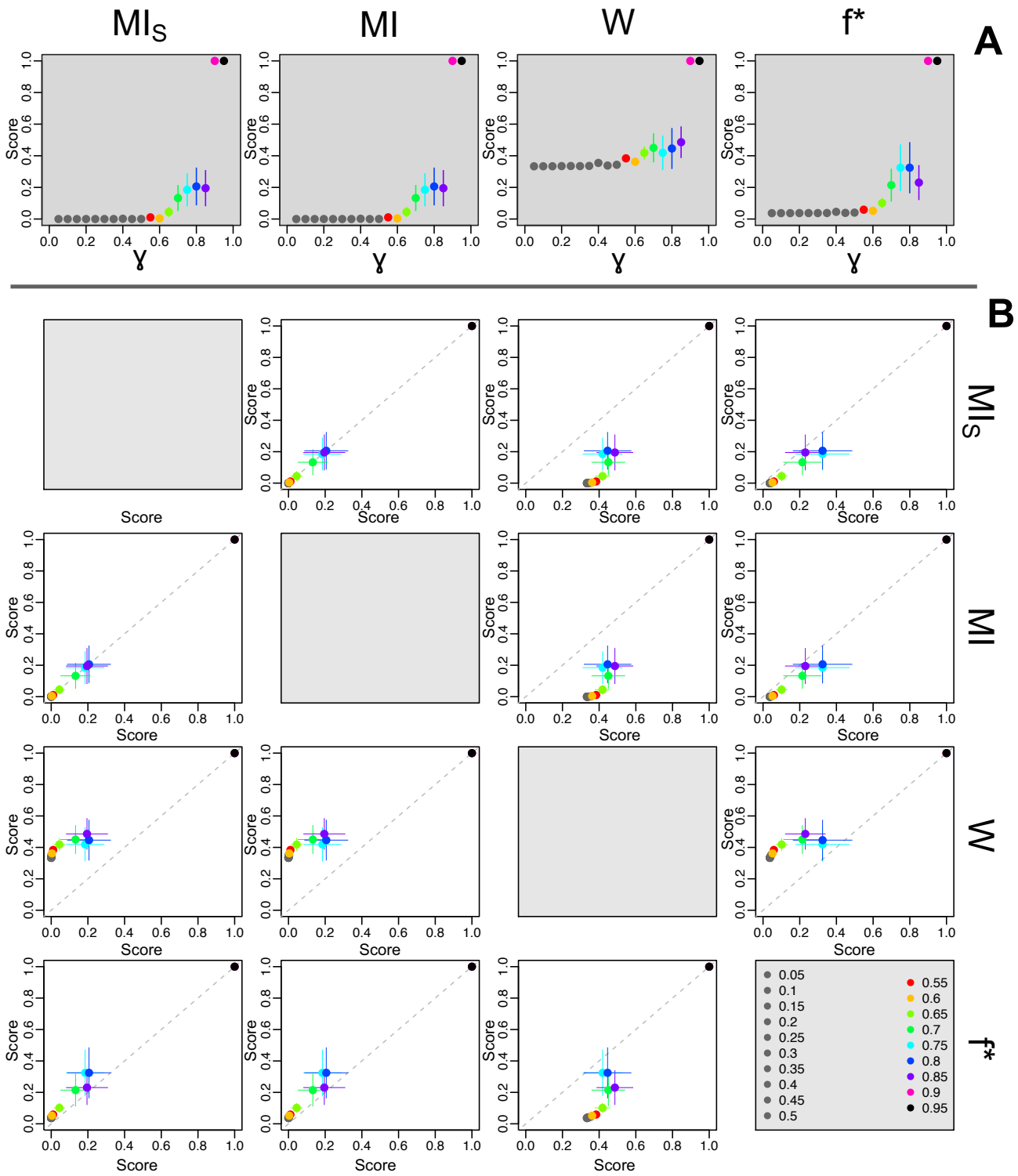


Figure 2-10 **A**: the mean and SD of the metrics of non-overlap in a {3,3} symmetrical strain system with varying cross-immunity, and **B**: pair-wise comparisons of the scores generated by the various metrics used to quantify non-overlap

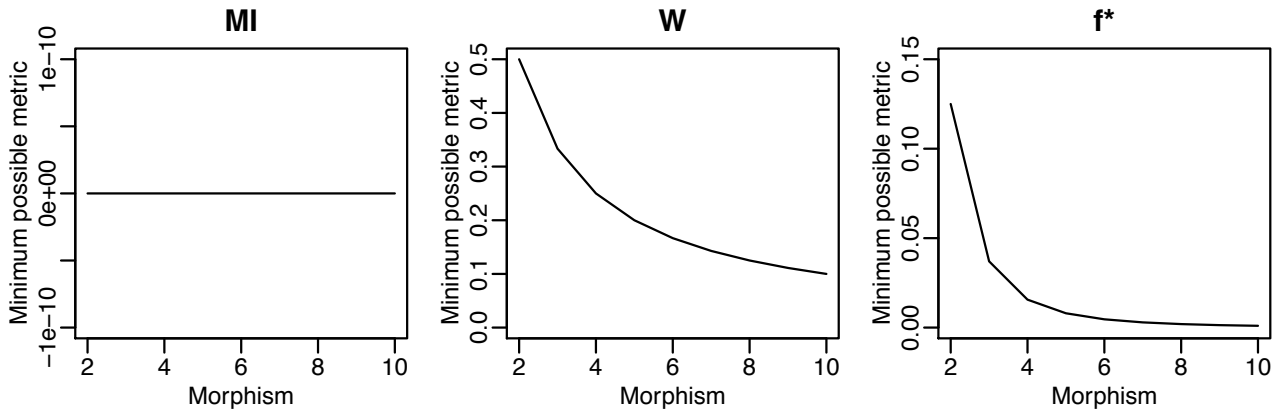


Figure 2-11 A comparison of the minimum possible values of non-overlap scores calculated by the various metrics between increasingly polymorphic maximally-overlapping site pairs. A morphism value (x-axis) of 2 corresponds to a {2,2} allelic matrix, 3 corresponds to {3,3}, and so forth.

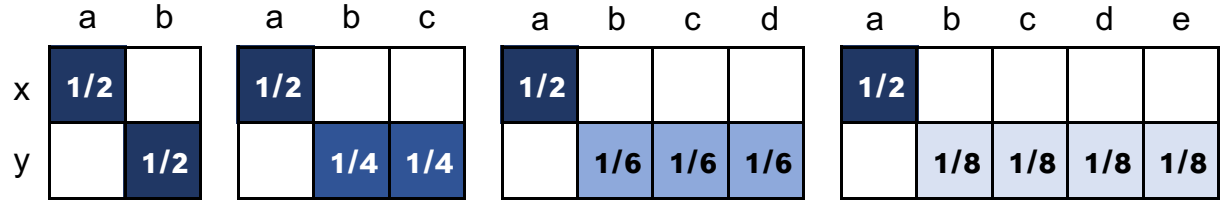


Figure 2-12 Example allelic matrices of increasing asymmetry from a symmetrical pair of sites {2,2} (left), to a highly asymmetrical site pair {5,2} (right). Values are the proportion of isolates that have that particular combination of alleles at the two sites of interest. For example, in the {2,2} system, half of the isolates are AB, the other half are BY. All of these examples show total non-overlap.

2.3.3 An algorithm for accounting for asymmetry when assessing non-overlap

In the multi-strain, multi-locus model of pathogen dynamics, at high levels of cross-immunity, we see that patterns of non-overlap between alleles emerge (DSS). Allelic variants differ phenotypically in terms of the immune responses that recognise them, however it is important for this approach that I can relate these phenotypes to their genotypic underpinnings. One factor to consider is the biochemistry of amino acid substitutions: it is important to recognise that different substitutions differ in their impact on substrate binding or protein function. In other words, an immune response against a particular allelic variant may also impact other variants which have biochemically conservative (similar) changes: although the genotypes of these two variants differ, phenotypically (with respect to immune responses) they are identical.

Here I present an algorithm which reconciles this notion by recognising that multiple amino acid variants may produce a single phenotype. This work seeks to produce the most-complete possible list of candidate of epitopes for further investigation, and so the algorithm seeks to group the amino acid variants at a site such that possible genotypic groupings may be identified and the power with which I can identify non-overlap maximised.

To explore this, consider the toy proportional allelic matrices given in Figure 2-12: in all of these cases there is total association between A and X, however there are a number of residues at the first site that can associate with Y (i.e., BY, CY, DY, and so forth). Despite differing symmetry, it is evident that each of these matrices exhibits total non-overlap: regardless of polymorphism levels, these particular sites have segregated into non-overlapping antigenic variants. While this is a theoretical exercise, it is helpful as such alleles could refer to sets of amino acids in the natural

world, for which biochemical similarities between B, C, D, and E could exist, or in which immune responses to B cross-reactive with C, D, and E, but not A.

Of the metrics presented previously, two (MI and f^*) can generate differing scores for systems that I *a priori* deem to not differ in their level of non-overlap. By means of example, Figure 2-13 presents 4 allelic matrices that all exhibit maximum non-overlap. Although there are multiple allelic variants at one site that associate with a single allele at the other, all can be reduced to a symmetrical 2x2 matrix identical to Figure 2-13A. The f^* and MI measures produce different values for the example matrices, however W and MI_S are calculated from the symmetricalized version of an allelic matrix, and thus successfully correct for differing morphism.

It should be noted that a given asymmetrical allelic matrix may "symmetricalize" to one of a number of possible symmetrical matrices, each with an equal trace (sum of the diagonal). For example, as seen in Figure 2-14A, there are multiple re-arrangements of Figure 2-13's matrix C that could produce allelic matrices with equal traces (Figure 2-14B/C). Both of the rearrangements have, along the diagonal, the same total number of variants (n=70), as well an equal ratio (40:30, or 30:40), thus the W scores would not differ. In the other measures, however, the scores would vary. Examples such as this justify an additional step in the algorithm: when multiple rearrangements are possible, the re-arrangement with the highest non-overlap score (as quantified by MI_S) is selected. Note that the probability of cases such occurring in natural systems is low (Chapter 3).

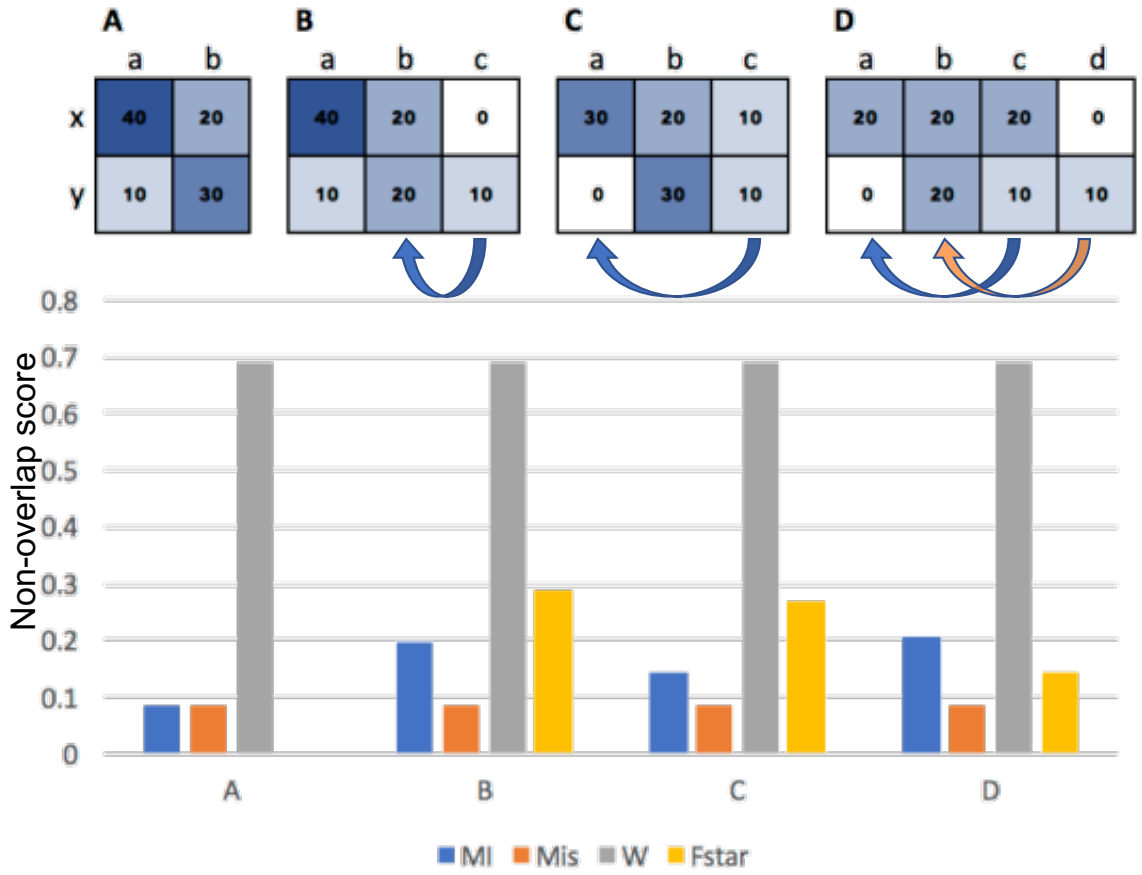


Figure 2-11 Comparing the measures of non-overlap between three allelic matrices (B-D) that can all be rearranged to the same maximally non-overlapping matrix in A. Arrows indicates rearrangement steps. Bar plots show the non-overlap scores that each metric applied to the corresponding allelic matrix produces.

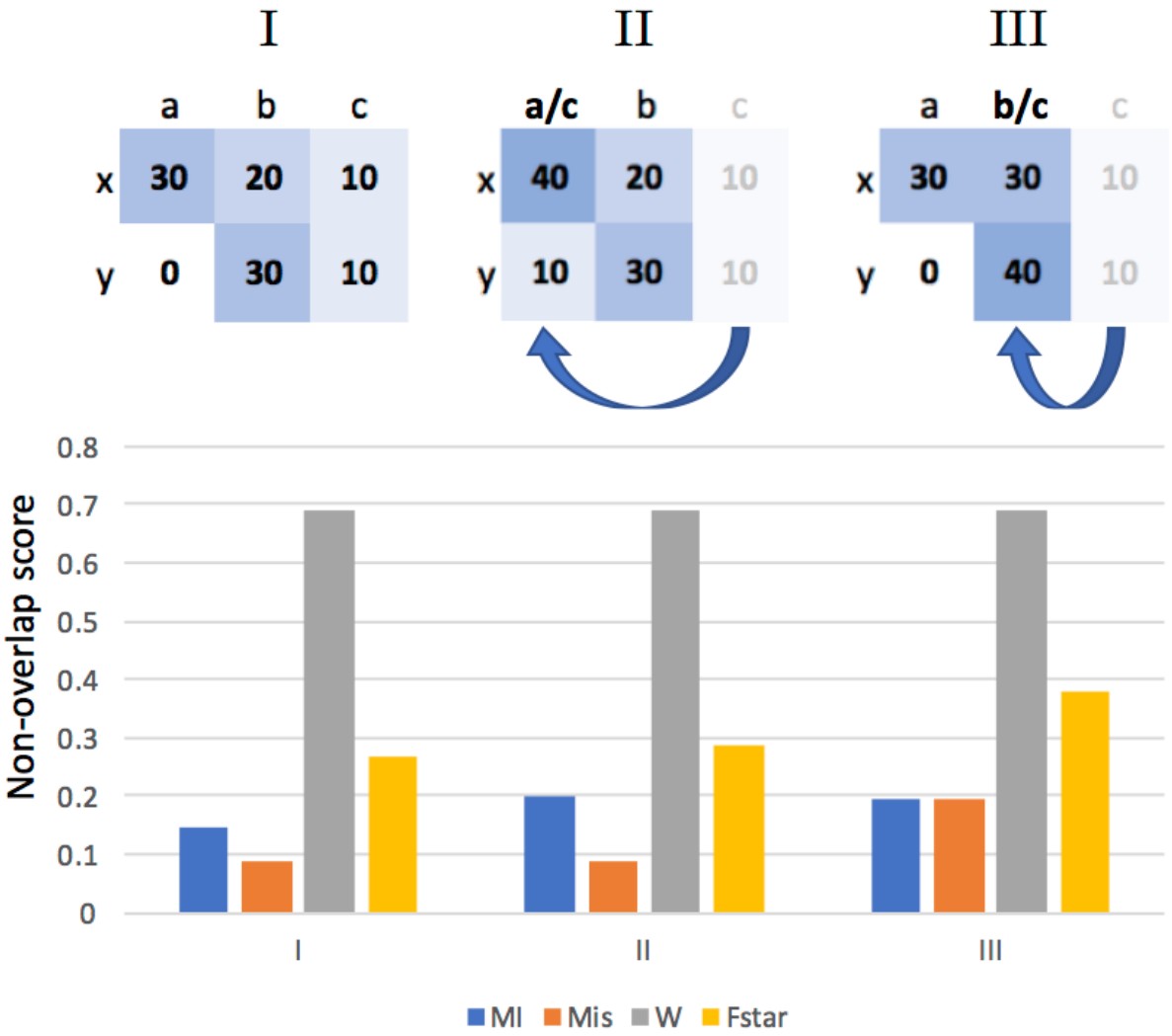


Figure 2-12 The multiple possible allelic matrix rearrangements (II and III) of the matrix in I (and Figure 2-13C), and their non-overlap scores according to each of the 4 metrics.

2.3.4 Quantifying non-overlap in simulated asymmetrical strain systems

Here I simulate systems of increasing complexity, beginning with a symmetrical {2,2} system, and progressing up to a highly asymmetrical {2,8} system ({2,3},{2,4},...,{2,8}). For each level of cross-immunity (with each corresponding strain structure: NSS, CSS, and DSS), 200 repeats are performed and the mean and standard deviation of each metric presented (Figure 2-15A). As was conducted previously, values for non-overlap here are calculated from sequence data sampled from the equilibrium phase of the model (if such a phase exists, otherwise from the times at which equilibrium would have been reached in a DSS system). Initial strain prevalences for the MANTIS simulations are selected from a random uniform distribution between ~0 and 10^{-5}, and the same 200 values are used for replicates at every level of cross-immunity and each strain structure.

Pairwise comparisons between the mean values of the metrics calculated for each strain structure and value of cross-immunity show differences in the ability of the metrics to deal with asymmetry (Figure 2-15B). Of particular interest is the comparison between MI and MI_S – these metrics are calculated using the same basic methodology and so allow us to better disentangle the effect of increasing asymmetry (Figure 2-15 subplots [row 1, column 2] and [row 2, column 1]). As summarized in Figure 2-16, there is no difference between these metrics in symmetrical systems however as asymmetry increases the values diverge. This is due to the fact that in highly asymmetrical systems the increase in entropy of the more polymorphic site is not fully offset by a decrease in the pair's joint entropy, thus the MI score decreases. The MI_S metric, however, is calculated from symmetricalized versions of the matrices, and so is not impacted by differences in polymorphism between amino acid sites.

My analyses of symmetrical polymorphic systems noted that the minimum possible values calculated by W and f^* are a function of the level of polymorphism at each site. The same observation holds here for the f^* metric (Figure 2-15A): the minimum value of f^* decreases with increasing asymmetry (i.e., increasing polymorphism at one of the two sites) at all levels of cross-immunity, and thus this metric is not robust to comparison of sites that differ in their polymorphism. As shown in Chapter 3, levels of polymorphism vary wildly, even within a single protein; methods previously described by the literature fail to account for this variation.

2.3.5 Quantifying non-overlap – conclusions

Of the various metrics compared here, I postulate that MI_S is the most suitable for performing analyses of complex strain structure in real sequence data. It is robust to increasing asymmetry between sites (unlike MI or f*), it is not a function of polymorphism levels (unlike W or f*), and it, through selection of maximally non-overlapping matrices, provides the best means with which a most complete candidate list of non-overlapping sites may be predicted. In Chapter 3, I utilize this metric to provide robust analyses of non-overlap in sequence data from multiple pathogenic species.

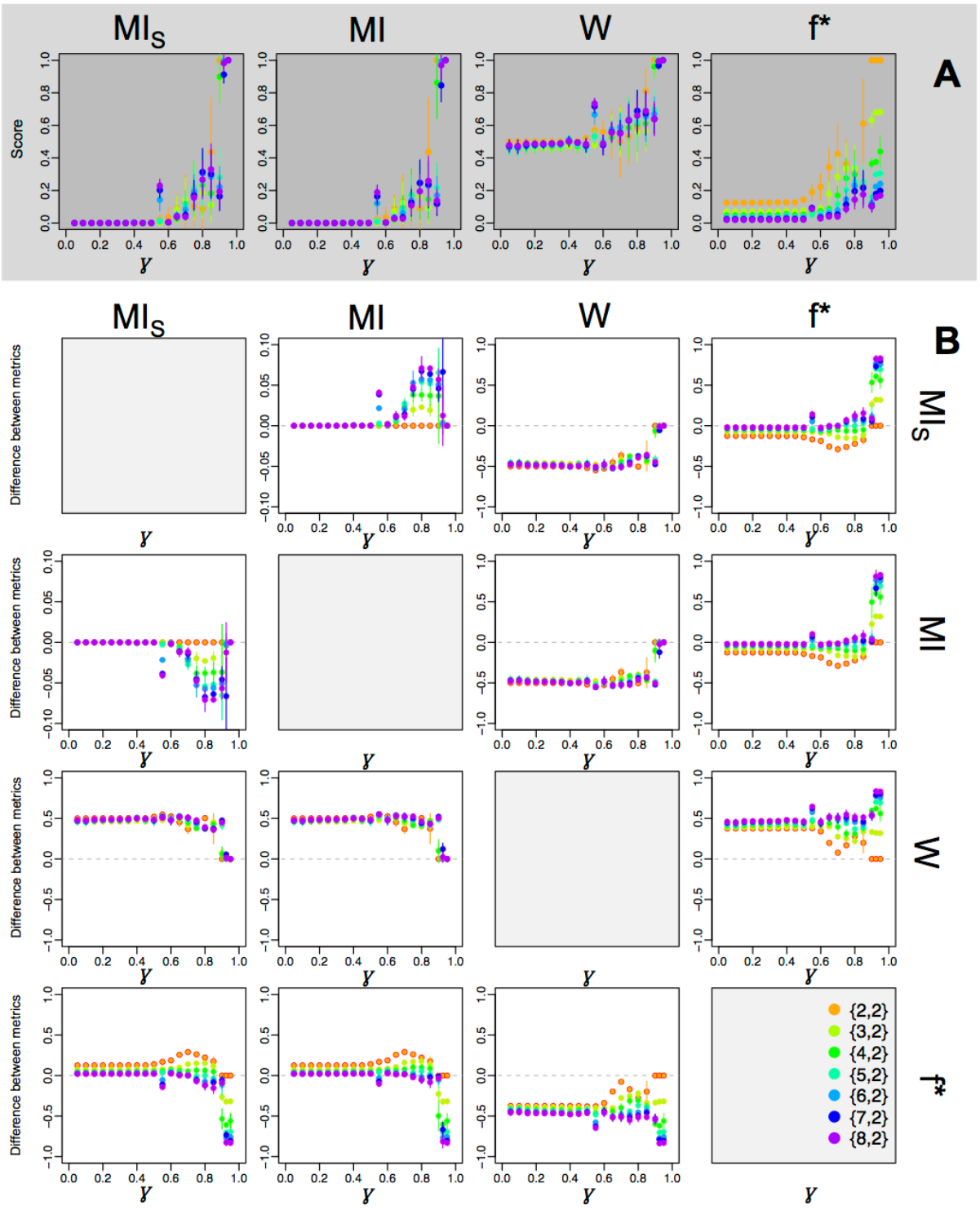


Figure 2-13 **A**: Mean and standard deviation of values of each of the metrics at each level of cross-immunity (x-axis) for increasing levels of asymmetry from a symmetrical {2,2} to highly asymmetrical {8,2} (symmetry denoted by point colour) and **B**: pair-wise differences (i.e., MIs vs MI; MIs vs W; etc.) between the metrics. The difference value corresponds to the value of the row metric minus the value of the column metric. Non-overlap scores are generated from analysis of the MATNIS model outputs across different levels of cross-immunity (x-axes) and

increasingly asymmetrical strain systems (point colours). Note that the y-axis scale differs in comparisons between MI and MIs (positions [1,2] and [2,1]), -0.1 to +0.1) from the other comparisons (-1 and 1).

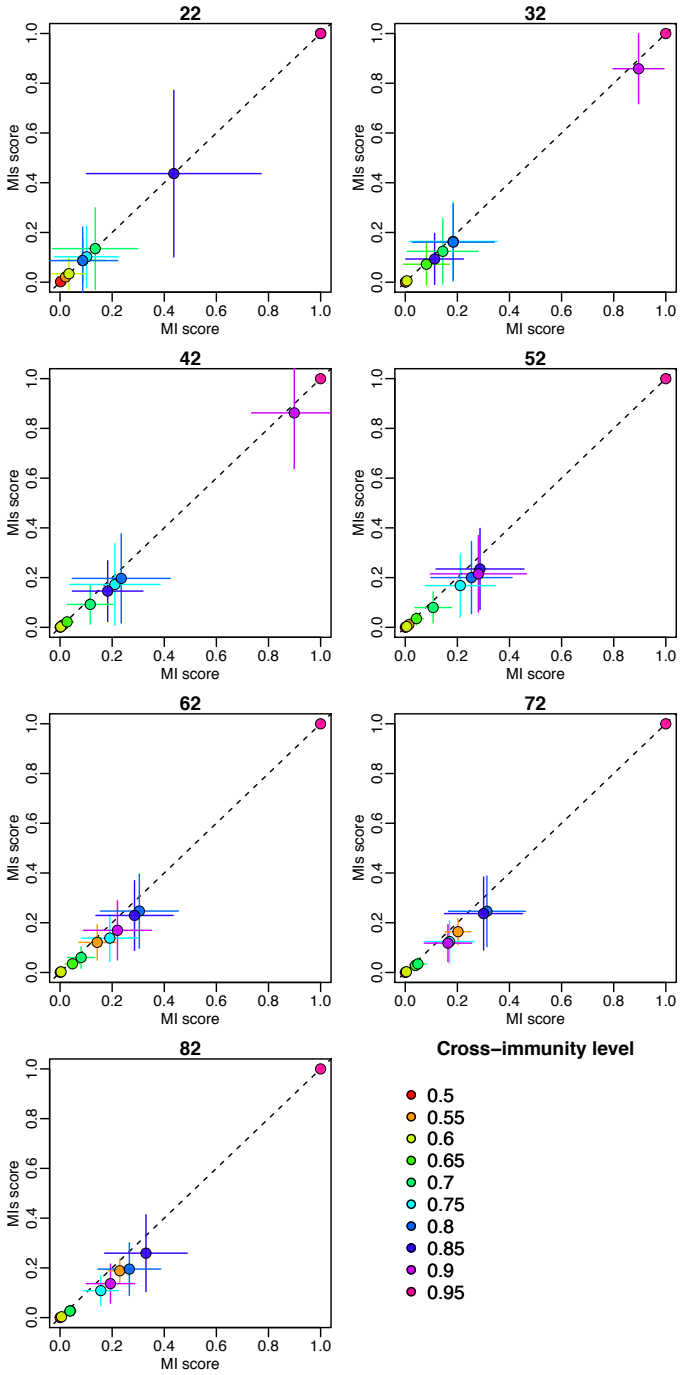


Figure 2-14 A comparison of mean (and SD) MI and MIS scores generated at each level of cross-immunity from 0.5 (NSS) to 0.95 (DSS) in systems of increasing asymmetry. "22" corresponds to a totally symmetrical {2,2} pair of sites. , and "28" to a highly asymmetrical {2,8} pair. MI and MIS scores of levels of cross immunity of <05 have mean and SDs of ~0 and are not shown. Points correspond to the mean of 100 repetitions of the model, with error bars showing 1 SD either side of the mean.

2.4 Discussion

In accordance with the principles of strain theory, pathogen populations under immune selection pressure at high levels of cross-immunity will segregate into non-overlapping sets of strains (DSS). By searching for signatures of DSS in sequence data, I can elucidate sites that are antigenic and under immune selection pressure. For this to be done efficiently and without bias, a robust metric for defining non-overlap is required. Here, I present a systematic analysis of various published and novel methods that can be used to quantify non-overlap in sequence data – this work postulates that MI_S is such a metric, and paves the way for analysis of amino acid sequence information from key human pathogen species (Chapter 3).

Patterns of DSS are of particular interest when it comes to considering vaccines against disease. As described above, vaccine development efforts have so far been most successful against pathogens where the most effective natural immune responses target conserved epitopes. There is a dire need, however, for vaccines against pathogens where this is not the case, and so development efforts should consider multi-allelic combination vaccines that target sites of intermediate variability that are in DSS and are therefore under strong immune pressure.

In systems that are in CSS, within the dominant set of variants at a given time, there is likely to be non-overlap, although this pattern is transient and the duration of dominance often difficult to predict. The canonical example of this is Influenza A, for which seasonal vaccines must be designed based on predictions of the strains that are likely to dominate. In order to explore non-overlap in pathogens that show CSS, I would be required to have a reasonable sample of sequence data at each step of a fine temporal scale. For this reason, and those given previously, I here elect to use identify

epitopes that exhibit stable patterns of non-overlap indicative of DSS; here studied in theoretical models, and later applied to real sequence data (Chapter Three).

Non-overlap has been explored by previous research, and a number of metrics used to quantify it. These metrics, whilst suitable for the purpose for which they were developed, often do not perform well when applied to systems of increasing complexity (due to natural variation in diversity levels). Here I demonstrate this, and present the development of two novel metrics for quantifying non-overlap, W and MI_S. The former is entirely novel and stems from the observation that the level of non-overlap between a pair of sites may be considered a function of two entities: the proportion of isolates that exhibit non-overlap at those sites, and the division of isolates in the population between the non-overlapping variants. The latter is an adaptation of the previously used MI score, derived from information theory, with a correction that allows for sites of differing levels of polymorphism to be systematically assessed. These asymmetries can also inevitably arise as a result of differing functional and structural constraints across a protein – at many amino acid sites, if any but two or three of the 21 possible amino acids are mutated to, the protein may lose its function, the pathogen loses fitness, and so mutations may not be passed on (they will not fixate and be present in a meaningful proportion of the pathogen population).

There have been discussions previously made of the role of polymorphic asymmetry, most notably by Buckee *et al.* who demonstrated the importance of including stochastic processes when modelling multi-strain pathogen populations (Buckee *et al.*, 2011). Earlier theoretical works demonstrated that the signals of co-evolution in a pathogen population can be lost to noise when the sites compared differ in diversity levels (Martin *et al.*, 2005). In spite of this, highly non-overlapping population

structures have been found in several pathogen populations, including the highly variable regions of *Neisseria meningitidis* (Gupta *et al.*, 1996; Buckee *et al.*, 2008, 2010), *Anaplasma marginale* (Futse *et al.*, 2008), *Streptococcus pneumoniae* (Lourenço *et al.*, 2017), and a number of *Plasmodium falciparum* antigens (Bull *et al.*, 2005; Polley *et al.*, 2007).

The MI_S metric developed here facilitates robust quantification of non-overlap in pathogens with varying genotypic diversity. It does this, however, without considering the phylogenetic history of the pathogen in question. It is possible that sites that exhibit non-overlap could do so as a result of their phylogenetic history: it could be the case that a non-overlapping pair of variants emerged once in the distant evolutionary history of a pathogen and have been maintained without positive evolutionary pressure. Further work could seek to incorporate shared a shared history of variants into a metric for non-overlap, however this falls outside of the remit of this Chapter. In Chapter 3, I perform such an analysis *post hoc*.

Another arm of analysis that could comprise further work is introduce significance testing to the metric; *i.e.*, to be able to state whether or not a pair of sites is statistically more non-overlapping that may be expected by chance. This approach could be conducted by, for example, comparing real data to phylogeny-aware simulated sequences such as that generated by Seq-Gen (Rambaut and Grassly, 1997). The simulated data would comprise a null distribution of MI_S scores against which real data could be compared. This approach would account for population structure, the number of isolates present, and differences in entropy across the genome. It would allow us to i) assign statistical significance to the findings outlined here and would ii) do so in light of the phylogenetic history of the pathogen.

As is discussed in further detail later (Chapter 3), this work aims to produce the most complete possible list of amino acid sites in pathogens that show evidence of immune selection pressure. By presenting this list in its entirety, I maximise the amount of information that can be gained from the sequence data, albeit at the cost of introducing potential false positives. Further phylogenetic and bioinformatic analyses such as those discussed in this thesis, could seek to reduce this list, allowing for triage of candidates for vaccine development.

Of the various metrics analysed here, MI_S provides the best measure of non-overlap in a variety of strain structure systems. I demonstrate it to be robust to varying levels of polymorphism between sites, even at very high levels of asymmetry. This metric is adapted from an existing well-characterised method of quantifying protein co-evolution (Mutual Information), and will be used to interrogate alignments of real genetic sequence data for indications of strain structuring (Chapter 3).

Chapter 3 Reverse Immunodynamics: applications

Andrew S. Walker, Katrina Spensley, Dr Paul Wikramatna, Fanny Wegner, Dr José Lourenço, Professor Sunetra Gupta

Author contributions:

FW curated whole genome EBV alignments. Analyses of non-overlap and phylogeny in AMA-1 were first performed by KS and WM (utilizing the traditional Mutual Information metric). ASW repeated the AMA-1 analyses (with a novel metric) using sequence data more recently curated from online databases and did all other work.

3.1 Introduction

Across the myriad of infectious agents that humans are infected with, pathogens employ an incredibly diverse range of strategies in order to evade or tolerate host immune responses. By bettering our understanding of these strategies and their impact on the co-evolution of host and pathogen, we stand a greater chance of developing new vaccines or treatment regimens against them. The impact of the arms race between parasite and host is evident in both: in this Chapter, I utilize genetic sequence information to investigate a number of important infectious agents.

3.1.1 Genetic sequence data availability

As discussed in Chapter Two, a population of parasites under immune selection, assuming high levels of cross-immunity, will segregate into non-overlapping groups of strains. In 1998 this strain theory was first proposed and the potential for reversal of the theory, that is, identifying sites under immune selection by studying non-overlap, first postulated (Gupta, Ferguson and Anderson, 1998). A review of the literature later concluded that applications of information theory to bioinformatics are in part limited by "the dearth of multiply sequenced genomes" (Adami, 2004). Here, I postulate that the "dearth" has ended, and that the recent advent of relatively cheap and accessible next-generation sequencing methods now stands to bring previously impossible bioinformatic approaches back into the fore. In 2014 alone, in excess of 14,000 bacterial and archaeal genomes were sequenced and submitted to the NCBI database (Land *et al.*, 2015) and, in this chapter, I use some of this genetic sequence data to better our understanding of the strain dynamics of a number of pathogens, allowing us to predict the specific location of epitopes in pathogens of global importance. These epitopes could comprise candidates for novel vaccines – I provide evidence of these sites being under immune selection pressure, and present an analysis of the genotypic

diversity seen at each site. A multi-component vaccine incorporating these immuno-dominant targets has the potential to be highly immunogenic and to induce responses that are cross-reactive against diverse natural pathogen populations.

3.1.2 Neisseria meningitidis outer membrane proteins

In order to robustly test the Reverse Immunodynamic approach, I first apply it to a pathogen in which patterns of strain structuring and non-overlap between regions under selective pressure are well known and understood – *Neisseria meningitidis*. This Gram-negative bacterium is capable of causing large outbreaks of bacterial meningitis, with approximately 500,000 cases occurring per year resulting in over 50,000 deaths (Pollard, 2004; Girard *et al.*, 2006). It circulates as a number of discrete groups, types, and subtypes, the distribution of which vary globally and are determined by the antigenic variants present on the polysaccharide capsule that encases the bacterium. Of the thirteen serogroups (that is, groups that induce similar antigenic responses) of *N. meningitidis*, six are responsible for virtually all cases of human disease (known as A, B, C, W, X, and Y). Serogroup-specific vaccine components have been developed against the most prevalence serotypes in developed countries (A, C, W, and Y), however development of a vaccine against serotype B has been limited by its similarity to the host antigen NCAM (Jodar *et al.*, 2002). There is also no vaccine against serogroup X which has recently been responsible for a number of outbreaks of meningococcal disease in Africa (Boisier *et al.*, 2007; Materu *et al.*, 2007). A comprehensive serotype-transcending vaccine would therefore be of huge public health benefit.

Naturally acquired immune responses predominantly target the serotype-defining polysaccharide capsule and sub-capsular outer membrane proteins (OMPs) of the

bacteria (Pollard *et al.*, 1999). OMPs comprise an alternative target for vaccination to classical capsule-based approaches and, by improving our understanding of their epidemiology and strain (serotype) structuring, particularly the stability of associations between different antigenic types, I will be better equipped to develop a vaccine that induces responses that act against all serogroups.

The best studied OMP is the PorA protein which is used in tandem with the polysaccharide capsule to define meningococcal serotypes and has been included in strain-specific capsule-targeting vaccines (Tappero *et al.*, 1999; Luijkx *et al.*, 2003). PorA, along with another OMP called FetA, has previously been shown to elicit bactericidal antibodies in humans (Russell *et al.*, 2004; Marsay *et al.*, 2015). Variable regions (VRs) of the genes that encode the OMPs have been demonstrated to, in accordance with theoretical predictions described previously, segregate into non-overlapping combinations and to maintain this pattern in spite of high levels of genetic recombination (Gupta *et al.*, 1996; Buckee *et al.*, 2008, 2010). These previous works analysed non-overlap in a non-site-specific manner. Rather than elucidating particular amino acid residues that co-varied, they studied the co-occurrence of subclasses of the PorA genes, demonstrating that the pathogen segregates into non-overlapping strains as a result of selection pressure. Here, I extend this work to look at site-specific non-overlap, with the aim of finding particular amino acid residues which may drive the broader patterns observed previously.

3.1.3 Plasmodium falciparum Apical Membrane Antigen 1

It is also widely acknowledged that there is a pressing need for an efficacious vaccine against *Plasmodium falciparum* malaria, and decades of research into various candidates have yet to produce a vaccine that has been widely implemented (Hill,

2011; Riley and Stewart, 2013; Arama and Troye-Blomberg, 2014). The only licensed malaria vaccine to date, RTS,S, is a subunit vaccine containing part of the circumsporozoite (CSP) protein. CSP is a sporozoite stage-specific protein that plays a role in migration of the sporozoite to the salivary glands of the Anophelene vector, and in hepatocytic invasion in the human host.

A formulation of RTS,S with adjuvant has shown moderate efficacy in large scale phase 3 clinical trials (RTSS Clinical Trials Partnership, 2015), has been approved by the European Medicines Agency for vaccination of young children (European Medicines Agency, 2015), and is also under investigation for inclusion as part of an intensive malaria elimination strategy in low-endemicity transmission settings (Gosling and von Seidlein, 2016). If targeted at the pre-erythrocytic stage of infection (immediately post-inoculation of parasites by the mosquito, but prior to hepatocytic invasion), vaccines must seek to eradicate all of the inoculated parasites. Studies suggest that the survival of tiny numbers of (or even single) parasites is sufficient for full infection to progress (Lyke et al. 2010; Bejon et al. 2005; Chapter 4). That said (as discussed and studied in Chapter 5), as a result of their population-level effects (herd immunity), the ability of vaccines to limit spread disease at a population level can outstrip their individual-level efficacy.

There have also been considerable efforts towards developing vaccines that elicit an immune response against erythrocytic infection. Such vaccines would target blood stage parasites (merozoites) that egress from the liver, as opposed to seeking to block liver infection altogether (reviewed in Miura 2016; Richards & Beeson 2009). For a vaccine that induces an immune response against the blood stages of infection to be effective it must reduce the blood stage parasite multiplication rate (PMR), that is,

fold-increase of parasitemia per 48-h cycle, to an average of <1. A number of blood-stage antigens have been identified as potential targets for vaccination (Osier *et al.*, 2014), and here I investigate a leading target, the protein Apical Membrane Antigen 1 (AMA-1).

Several phase II trials on AMA-1 have been completed to date. In a recent trial, AMA-1 protein from the 3D7 falciparum strain was adjuvanted with $AS02_A$ (FMP2.1/$AS02_A$) and administered as part of a trial of 383 Malian children. This trial saw no significant impact on clinical malaria across all participants, however an analysis of just those individuals who saw infection with AMA1 3D7-type parasites found the vaccine formulation to have significant protective effects (Thera *et al.*, 2011). Earlier investigations also suggested that diversity of the protein may be hampering development efforts: a combination of alleles used in a phase I trial (AMA1-3D7 and AMA1-FVO) was found to be unlikely to cover the natural variation seen in the field (Mullen *et al.*, 2008). An additional phase II trial tested a mixture of 3D7 and FVO-type AMA-1 and found no impact on clinical malaria (strain-specific or otherwise) in a cohort of 279 Malian children (Sagara *et al.*, 2009; Ouattara *et al.*, 2010). Collectively, these studies, in tandem with biochemical structural analyses (Lim *et al.*, 2014), conclude that a successful malaria vaccine based on AMA1 will be required to elicit strain-transcending antibodies.

Here, I utilize the Reverse Immunodynamic approach to elucidate strain structuring of and immune selection on the AMA-1 protein. This work builds on previous research (Spensley et al. 2017; Chapter 2) to apply new methodologies devised in Chapter Two to this important blood stage antigen. By identifying key epitopes that are under immune selection pressure and evolve within a fixed antigenic space, we may be able

to inform the design of novel vaccine constructs that would be highly immunogenic and cover the considerable allelic variability seen in malaria infection.

3.1.4 Epstein Barr Virus

The third pathogen that will be studied here is human herpesvirus 4, the Epstein Barr virus (EBV). EBV infects almost ubiquitously worldwide, with 90% of adults showing some serological response to infection (reviewed in Cohen 2000). In developing countries, the majority of children are infected by age 5, whereas in developed countries first infection is often delayed until adolescence or early adulthood. A later age of first infection is associated with infectious mononucleosis (glandular fever) and the virus is also strongly implicated in a number of lymphatic, gastric, and nasopharyngeal cancers (first noted by Old et al. 1966; reviewed in Hjalgrim et al. 2007). Infection is also associated with several autoimmune conditions and immune-deficiencies such as those that cause multiple sclerosis (Wekerle and Hohlfeld, 2003), lymphoid interstitial pneumonitis (reviewed in Swigris et al. 2002), and oral hairy leucoplakia (Young *et al.*, 1991).

EBV is most commonly transmitted through oropharyngeal secretions, although there is also evidence of vertical (mother to child *in utero*) transmission (Meyohas *et al.*, 1996). Upon transmission, EBV infects B cells by interaction of its glycoproteins with host cellular receptors: viral gp350 binds to the cellular receptor CD21, and then viral gp42 interacts with host MHC class II molecules to trigger fusion of the viral envelope and cell membrane, facilitating entry of the virus into the B cell (Speck, Haan and Longnecker, 2000). Whilst this is the most common route of entry, the virus can also invade via interaction with human CD35 on B cells (Ogembo *et al.*, 2013) and cellular

β-1 integrin on host epithelial cells (which lack the CD21 receptor) (Tugizov, Berline and Palefsky, 2003; Xiao *et al.*, 2007, 2009).

Once successfully transported to nucleus of the B or epithelial cell, the lytic cycle may proceed. The circular genome of EBV is linearized, and then replicated by viral DNA polymerases. Viral lytic gene products are characterised in accordance with the temporality of their expression; either early (E), intermediate-early (IE), or late (L) in the viral lytic program (reviewed in Tsurumi et al. 2005). The early gene products perform a wide variety of functions, whereas later products tend to code products that play some role in immune evasion or that are viral capsid antigens (VCAs) (reviewed in Tsurumi et al. 2005).

As with all herpesviruses, EBV also has a latent stage of infection. In this state, there is no active viral replication, and the virus persists primarily in memory B-cells. Replication occurs at this stage only by manipulation of the host's own cellular machinery. In contrast to the lytic phase, host DNA polymerases allow for replication of extra-nuclear EBV genomes called episomes (Ambinder and Lin, 2005). There are a number of programs of latency, each differing in the set of genes that are expressed and resultant malignancies (Blake *et al.*, 1997; Bornkamm and Hammerschmidt, 2001; Thorley-Lawson and Gross, 2004; Speck, 2005). Infection can re-enter the lytic cycle, governed by a suite of molecular pathways (Israel and Kenney, 2005), allowing active production of virus that can infect further B cells and epithelial cells within the host and contribute to transmission.

Naturally acquired immune responses target at least 90 known epitopes located with 15 of the approximately 80 proteins encoded by EBV. EBV induces immune responses

at both the lytic and latent stages, and responses are both cellular and humoral, reviewed elsewhere (Landais, Saulquin and Houssaint, 2005; Rowe and Zuo, 2010; Chen, 2011). These immune responses are highly effective in controlling viral load and thus clinical prognoses, however infection with EBV is persistent and lifelong. Persistent viruses such as EBV utilize a range of methods to modulate host immunity; maintenance of constant low-titre infection requires a careful balance between transmissibility, pathogenicity, and survival.

It is widely accepted that an efficacious vaccine against EBV is needed, to prevent both IM and the 200,000 or so cases of associated disease that occur per year (Cohen *et al.*, 2011). A highly efficacious vaccine administered to young infants could reduce the incidence of Burkett lymphoma in Africa by as much as 85%, anaplastic nasopharyngeal carcinomas almost entirely, and gastric lymphoepitheliomas by up to 90%. Thus far no vaccine against EBV has been licensed, although a total 9 clinical trials on vaccine candidates in humans have been conducted; 5 tested as possible prophylaxis (Gu *et al.*, 1995; Moutschen *et al.*, 2007; Sokal *et al.*, 2007; Elliott *et al.*, 2008; Rees *et al.*, 2009), and 4 as possible therapeutic administrations (Lin *et al.*, 2002; Chia *et al.*, 2012; Edwin P. Hui *et al.*, 2013; Taylor *et al.*, 2014).

All but one of these prophylactic vaccine regimens aimed to induce responses against the envelope glycoprotein gp350, and have been tested with or without various adjuvants. The trials varied in their outcomes: all formulations induced neutralising antibodies against gp350, however not all successfully reduced infection rates or clinical prognoses (where tested). A single human clinical trial has also been conducted on a peptide from the EBNA3A; this found a statistically insignificant trend

towards a reduction in IM, but no impact upon overall infection incidence (Elliott *et al.*, 2008).

In contrast to a prophylactic vaccine, a successful therapeutic vaccine would not block infection, rather it would seek to reduce the associated malignancies. The primary goal of such a vaccine would be to induce cellular immunity to particular proteins that are associated with disease and expressed in the latent stage of infection. For example, many EBV-associated lymphomas are seen when EBV assumes a type 2 latency pattern in which EBNA-1, LMP1, and LMP2 are expressed thus targeting T cell responses against these proteins may reduce the incidence of disease. Of the 4 therapeutic human vaccine trials conducted to date, two tested LMP2 in combination with a peptide from EBNA-1 and induced T cell responses in more than half of patients (Edwin P Hui *et al.*, 2013; Taylor *et al.*, 2014), one tested autologous dendritic cells containing LMP2 peptides and found a boosting of specific T cells and tumour regression in some patients (Lin *et al.*, 2002), and one found a combination of autologous dendritic cells containing both LMP1 peptides and LMP2 to induce responses in a small proportion of patients (Chia *et al.*, 2012).

A recent increase in the availability of high quality whole genome sequences of EBV means that Reverse Immunodynamics can be used in an attempt to better our understanding of the molecular epidemiology of the virus. Finding epitopes which exhibit patterns of non-overlap congruent with immune selection pressure under high cross-immunity may allow for the development of novel highly immunogenic and broadly cross-reactive vaccine strategies. Here I elucidate interactions between epitopes that occur between different proteins, expanding on the single protein and dual protein approaches described above.

3.2 Methods

3.2.1 Multiple Sequence Alignments

Amino acid sequences for the *P. falciparum* antigen AMA-1 were sourced from GenBank (National Centre for Biotechnology Information). The Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) was used to search the GenBank *nr* (non-redundant) database for unique sequences which matched an AMA-1 reference sequence (UniProtKB ID: Q7KQK5) with greater than 90% identity.

A total of 1198 unique AMA-1 sequences (downloaded from GenBank, 27/01/2017) were aligned using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) method, implemented in MEGA6 (Edgar, Drive and Valley, 2004; Tamura *et al.*, 2013). The quality of the alignment was checked by eye in a visualisation tool, and by assessing the pairwise distances between all sequences to check for any misalignments or foreign sequences.

Whole genome *EBV* nucleotide sequences were curated and assembled from sequenced contigs by a collaborator at University College London (Fanny Wegner, see author contributions). In order to derive amino acid sequence alignments for particular proteins from those whole genome sequences, I devised a pipeline, comprising the following steps:

1. A reference amino acid sequence for each protein was downloaded from UniProt (UniProt Proteome ID: UP000153037, downloaded 29/02/2016)
2. Each reference amino acid sequence was reverse translated to its most likely nucleotide sequence using the EMBOSS Backtranseq tool (Rice, Longden and Bleasby, 2000) on the EMBL-EBI website (McWilliam et al., 2013)

3. BLAST was then used to search each EBV whole genome sequence for each reference nucleotide sequence. In some cases, a gene product is encoded by a number of non-consecutive reading frames, and so care had to be taken to reassemble those frames correctly
4. The curated nucleotide sequences for each gene were then translated back to amino acid sequences. Correct reading frames and sense (3' to 5' or 5' to 3' direction) were found by testing each reading frame and sense in turn, and selecting the translation most similar to the original reference sequence
5. Translated amino acid sequences were then aligned using the MUSCLE tool, implemented in MEGA6.

N. meningitidis PorA and *FetA* sequences were downloaded from the PubMLST database (Jolley and Maiden, 2010). Given the extremely high levels of diversity seen in Variable Region 1 and 2 (VR1 and VR2) of *PorA* and VR1 of *FetA*, a bespoke alignment routine was required. For each of the variable regions, there are a number of known families and subfamilies of types. For each isolate, the sequence for each Variable Region was compared to a reference sequence (from PubMLST) for each VR family (Jolley and Maiden, 2010). The VR sequence was assigned to the best fitting family, and per-family alignments created.

The wealth of sequence data available for *N. meningitidis* (>6000 sequences), as well as the high quality of the curation, also allowed us to stratify our analysis by a number of factors, including serotype, location of sample, and year of sampling.

As part of this work, I also curated a database of amino acid sequence alignments for other malaria proteins of interest (Appendix A). Analysis of this dataset, however, falls outside of the remit of this thesis.

More information about numbers of sequences for each pathogen (i.e., year and country and collection) are given in Appendix B (Figures S1, S2, and S3).

3.2.2 Phylogenetic analysis

In order to strengthen claims of the role of immune selection in driving patterns of strain structure, we must seek to identify any other factors that may play a role in shaping the population dynamics of pathogens. One such factor that should be considered is the evolutionary history of the parasite population – if the population exists as two distinct clades, and these clades differ from each other at a few keys sites, then ancestry may present itself as non-overlap. It is, therefore, necessary for us to assess ancestry, paying particular attention to the variants that appear at sites of interest.

Reconstruction of evolutionary history (ancestry) of pathogens can be performed by inferring phylogenetic trees. Here we assess the impact of phylogeny on the patterns of non-overlap seen in *P. falciparum* AMA-1 by creating neighbour joining trees in the MEGA software package (Tamura *et al.*, 2013). We then paint onto the tips of each tree the variants that exist at particular sites of interest, and assess how each variant presents in relation to the natural ancestry of the isolates.

This relationship is quantified for each amino acid site using a measure called the Parsimony Score (PS). The PS describes the minimum number of changes required for

the site of interest to map onto a phylogenetic tree derived from all other sites. Pairs of sites are assessed in this way by computing the geometric mean of the two parsimony scores (GMPS).

This analysis is performed for AMA-1 and EBV, although recombination across the genomes of each (Kwok *et al.*, 2014; Tibayrenc and Ayala, 2015) means that the results of such phylogenetic analyses should be carefully interpreted. The analysis does, however, allow us to determine whether sites that exhibit non-overlap are likely to do so due to some shared evolutionary history.

For *N. meningitis*, population genetic forces (aside from those driven by immune selection pressure) are known to drive clonal expansion (Tibayrenc and Ayala, 2015), which means that non-overlapping variants may appear to be evolved due to a shared phylogenetic history rather than immune selection pressure. The species was included in this analysis, however, due to it previously being shown to exhibit non-overlap as a result of immune pressure (it is effectively a positive control): further investigations into patterns of non-overlap in the context of phylogeny are not required.

3.2.3 Biochemistry of amino acid substitution

When performing comparative analysis of amino acid substitutions within protein sequences it is important to consider the biochemical properties of the amino acid residues present. 17 of the 21 amino acids can be divided into four groups according to the charge of their side chains, with the remainder representing “special cases” that can form covalent disulphide bonds (selenocysteine and cysteine), form cyclic compounds (proline), or show enhanced flexibility relative to other amino acids (glycine). The four groups are as follows: those with positively charged side chains,

those with negatively charged side chains, those with polar uncharged side chains, and those with hydrophobic side chains.

Within each of the groups the amino acids exhibit limited structural diversity, and so amino acid substitutions occur much more readily within than between groups, with within-group substitutions often do not lead to functional or structural changes. Such substitutions are deemed "conservative", as opposed to substitutions between biochemically diverse residues which are termed "radical". There has been a wealth of work aiming to quantify the interchangeability of amino acids in different evolutionary contexts, each producing a matrix of the scores to changes between each residue pair. Two of the most widely used series of substitution matrices are BLOSUM (BLOcks Substitution Matrix) and PAM (Percent Accepted Mutation) (Dayhoff and Schwartz, 1978; Henikoff and Henikoff, 1992).

BLOSUM was first introduced in 1992 by Henikoff and Henikoff and was computed by analysing the frequency of substitutions between each residue in conserved regions of proteins families in a large database. There are a range of BLOSUM matrices, each defined according to the sequence similarity from which the scores were computed – i.e., the BLOSUM62 matrix was computed from sequence with less than 62% similarity. Higher rank BLOSUM matrices are typically used to assess less divergent sequence sets.

PAM matrices were introduced in 1972 by Dayhoff *et al.* and were calculated from observations of phylogenetic trees of a set of highly related proteins (>85% similarity). As with BLOSUM, there are a range of PAM matrices that differ in the divergence of sequences that can assessed. In these matrices, each value indicates the likelihood of

the row's amino acid being replaced by the column's amino acid through point mutations over a given evolutionary time. A commonly used PAM matrix is PAM70: in this matrix, the values estimate the rate of substitution given that 70% of the amino acids have changed over some evolutionary time. Thus, in the matrix, high values correspond to more likely changes between residues (i.e., conservative changes), and low values correspond to changes that are biochemically radical and so less likely to occur.

Here, I use a PAM70 substitution matrix to quantify the nature of the amino acid substitutions that is seen in the protein sequence data from each pathogen species. Matrix-based approaches are typically utilized to assess the homology between and phylogeny of aligned sequences. I apply the approach to analyse the likelihood of given single accepted point substitutions at our amino acid sites of interest. For each site of interest, we calculate a mean PAM70 log likelihood score by finding and then averaging over the pairwise likelihoods of changes between each of the variants present. For example, in the *N. meningitidis* PorA antigen, at site 262, there are 3 amino variants: Glycine, Threonine, and Aspartic acid. The PAM70 log likelihood of changes between each residue are as follows: Glycine to Threonine is -3, Threonine to Aspartic Acid is -2, Glycine to Aspartic acid is -1, and so the mean PAM70 likelihood for PorA site 262^{GTD} is -2. This score would agree with our expectation of a low score given radical changes: the three amino acid residues at that site belong to different biochemical groups.

3.2.4 Quantifying non-overlap

The specific methodologies used here to quantify the degree of non-overlap between amino acid sites in the pathogens of interest are described in detail in Chapter Two.

The theoretical investigations conducted previously found that the MI_S measure developed in this thesis is robust to asymmetrical levels of polymorphism between pairs, as well as overall entropy, and so here I present non-overlap using that new measure.

3.3 Results

3.3.1 Neisseria meningitidis PorA and FetA

Analysis of amino acid sequence data translated from nucleotide data downloaded from the PubMLST database allows us to identify sites within the *PorA* and *FetA* proteins of *N. meningitidis* that exhibit patterns of strain structure congruent with immune selection pressure. The numbers of sequences used in the analysis are given in Table 3-1. Previous works have found the variable regions (VRs) within these key antigenic proteins to exhibit non-overlap, and so our analysis here serves to test whether our methods successfully identify amino acid sites already known to be antigenic, under selection pressure, and displaying DSS.

The variable regions of PorA in the sequence alignment are as follows: VR1 is from sites 47 to 62, VR2 is from sites 195 to 213, VR3 is from sites 250 to 274. In FetA the Variable Region is from site 166 to 208. As expected, these regions all exhibit higher levels of polymorphism and entropy when compared to surrounding regions (Figure 3-1). Our findings also agree with previous observations of VR3 being less diverse than VR1 and VR2: it has been hypothesised that this could be due to the shorter length of its surface loop meaning that it is less exposed to immune effectors and so faces a lesser selection pressure (De Filippis *et al.*, 2007).

The VR in FetA also demonstrates greater entropy (Figure 3-1B,C) and polymorphism levels (Figure 3-1E,F) than surrounding regions. This data also indicates that there are other non-“VR” regions of *FetA* that demonstrate intermediate levels of variability, namely in regions between sites 50 and 80, and between 440 and 470.

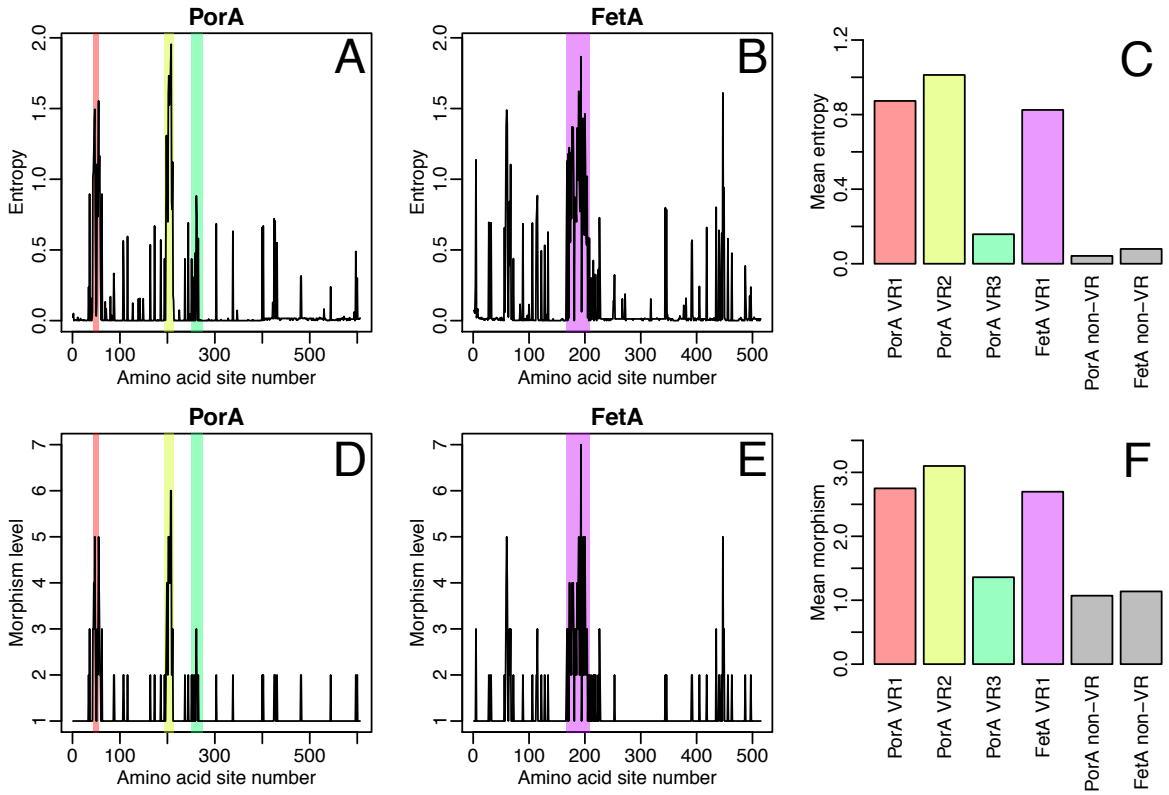


Figure 3-1 Levels of entropy and polymorphism across the alignments of *Neisseria meningitidis* PorA and FetA. Note that all variants present in less than 5% of isolates are excluded from this analysis. **A** Entropy levels across the PorA protein, with variable regions highlighted in red (VR1), yellow (VR2), and green (VR3). **B** Entropy levels across the FetA protein, with the variable region (VR) highlighted in purple. **C** The mean entropies of sites falling within and outside of VRs in PorA and FetA. **D** Morphism levels across the PorA protein, with VRs highlighted as in A. **E** Morphism levels across the FetA protein, with VR highlighted as in B. **F** The mean polymorphism levels of sites falling within and outside of VRs in PorA and FetA.

Our analysis finds sites within both *PorA* and *FetA* that exhibit patterns indicative of immune selection pressure (Figure 3-2). The vast majority of amino acids sites that show non-overlap in either protein fall within the known VRs. Of the 34 sites that exhibit non-overlap with other sites in PorA at an MI_S of greater than 0.6, 23 fall directly within the VRs, and only four fall more than 10 amino acids outside of a VR. Similarly, in FetA, of the 17 amino acid residues exhibiting non-overlap with some other, 11 fall within the VR; the others are mostly located between amino acid residues 440^{AV} and 456^{KQ}, a region that shows higher entropy than most of the protein. In both proteins, the mean PAM70 scores indicate that many of the amino acid changes present are likely to cause significant structural/functional changes to the proteins (Table 3-2).

There are four dimorphic sites outside of the VRs of PorA which show non-overlap with sites inside the VRs in a significant portion of the subsets of isolates: 116^{IV} falls some distance away from a VR and shows non-overlap with sites in VR1 and VR2; 173^{ED} shows strong non-overlap with polymorphic residues in VR2, 237^{NS} shows strong non-overlap with a single highly polymorphic residue in VR2 ($206^{KNTAGDVMQI}$); and 338^{FL} shows non-overlap with an adjacent pair of residues in VR3 (260^{RS} and 261^{IATG}). All four of these residues are located within transmembrane regions of PorA and are dimorphic. They are located on exposed loops of the protein (loops I, IV, and V respectively), all of which are capable of inducing production of bactericidal antibodies.

As discussed previously, an important factor when considering amino acid substitutions is the biochemical conservativism of changes – that is, the degree of biochemical similarity that exists between the amino acid variants present at a

particular site. Amino acid substitutions between biochemically similar pairs typically have less of an impact upon structure and function of the protein. The four transmembrane sites above all show conservative changes – both 116 and 338 switch between amino acids with hydrophobic side chains (valine to isoleucine and phenylalanine to leucine respectively), 173 between amino acids with negatively charged side chains (aspartic acid to glutamic acid), and 237 between amino acids with polar uncharged side chains (serine to asparagine). This is reflected by their relatively high mean PAM70 scores (Table 3-2) and, combined with their location in transmembrane domains of PorA, suggests that these sites vary in response to being targeted by immune responses.

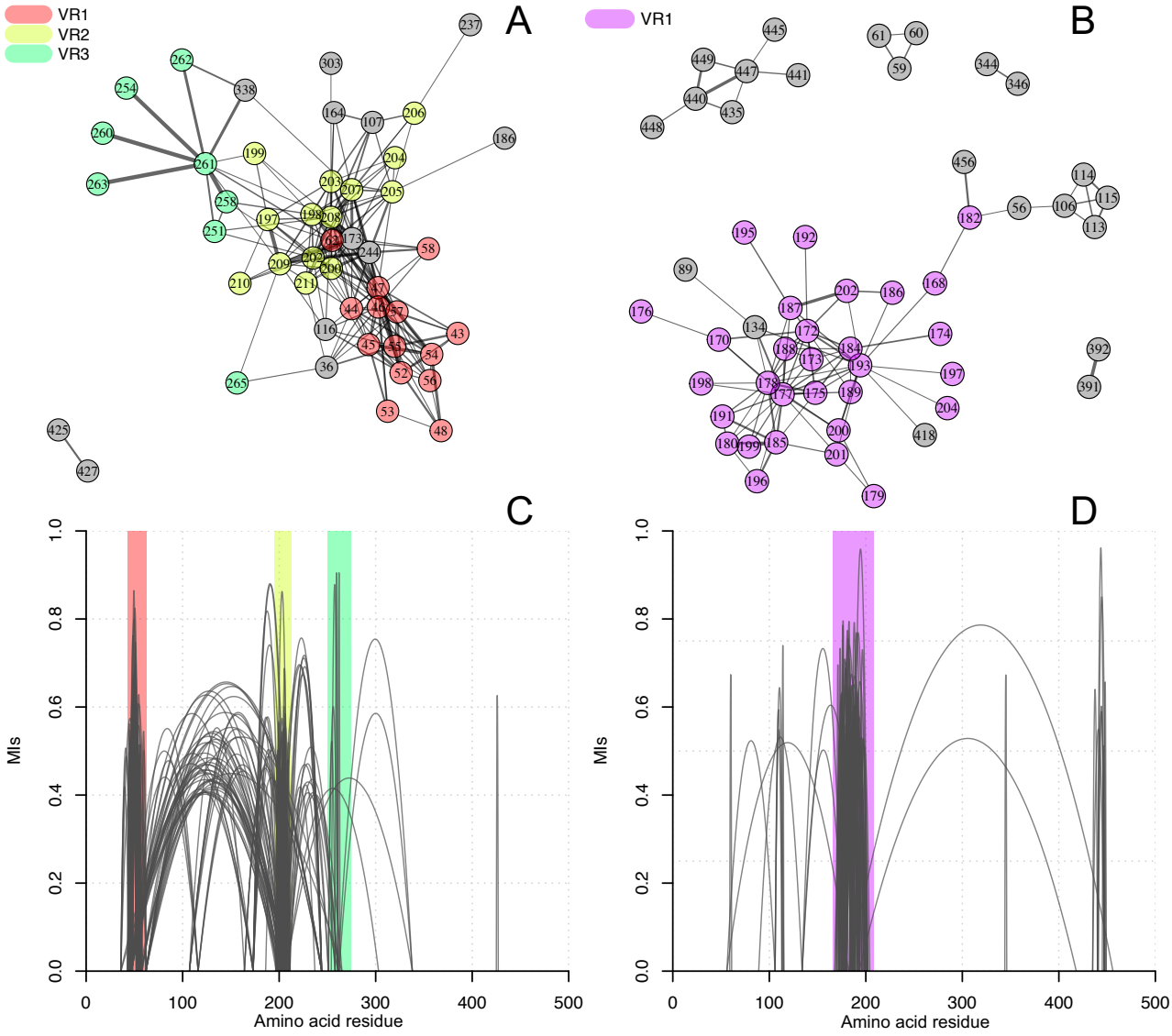


Figure 3-2 **A,B** Network maps showing pairs of amino acid residues in PorA (**A**) and FetA (**B**) that show some non-overlap (MI_S>0.4). Nodes are coloured by presence in a variable region (see legend), and edge thicknesses correspond to the strength of interaction between a pair. **C, D** Mapping interactive sites onto a linear sequence of the protein (x-axis). A line connecting points *X* and *Y* on the x-axis corresponds to non-overlap between amino acid residues *X* and *Y*. The strength of non-overlap is indicated by the height of the curved line, and the coloured regions indicate the VR1 (red), VR2 (yellow), and VR3 (green) regions of *PorA*, and VR (purple) region of *FetA*.

All of the previous analyses were conducted upon all available sequence data from the PubMLST database, with the following serogroups represented in meaningful numbers (n>100): serogroup B (n=2694), serogroup W (n=889), serogroup C (n=812), serogroup Y (n=774), and serogroup A (n=221). The large number of samples present for serotype B facilitates subsetting this data to test whether patterns of non-overlap that we find across all samples are maintained at different spatial and temporal scales. The number of serogroup B sequence isolates from each year of sampling distinguished by global samples and UK samples are shown in Table 3-1.

	Global	UK
2010	221	192
2011	413	373
2012	352	293
2013	467	293
2014	364	210
All years	2696	1825

Table 3-1 Numbers of sequence isolates available for *Neisseria meningitidis* serogroup B PorA from the PubMLST database.

If a particular pair of residues are co-dependant in just one spatio-temporal natural setting of the pathogen's population then the signal of that interaction will be lost to noise when the dataset is analysed as a whole. By analysing the data in this way, we increase the power with which we can detect amino acid residues that may be interest in vaccine development or in understanding the role that immune selection pressure plays in driving evolution of meningococcal disease. Conducting RID on PorA

sequence data from each year of collection (2010-2014) reveals some amino acid sites under strong non-overlap that were not found in the analysis of the whole data set (Table 3-2; 62, 186, 198, 204, 205, 210, 303).

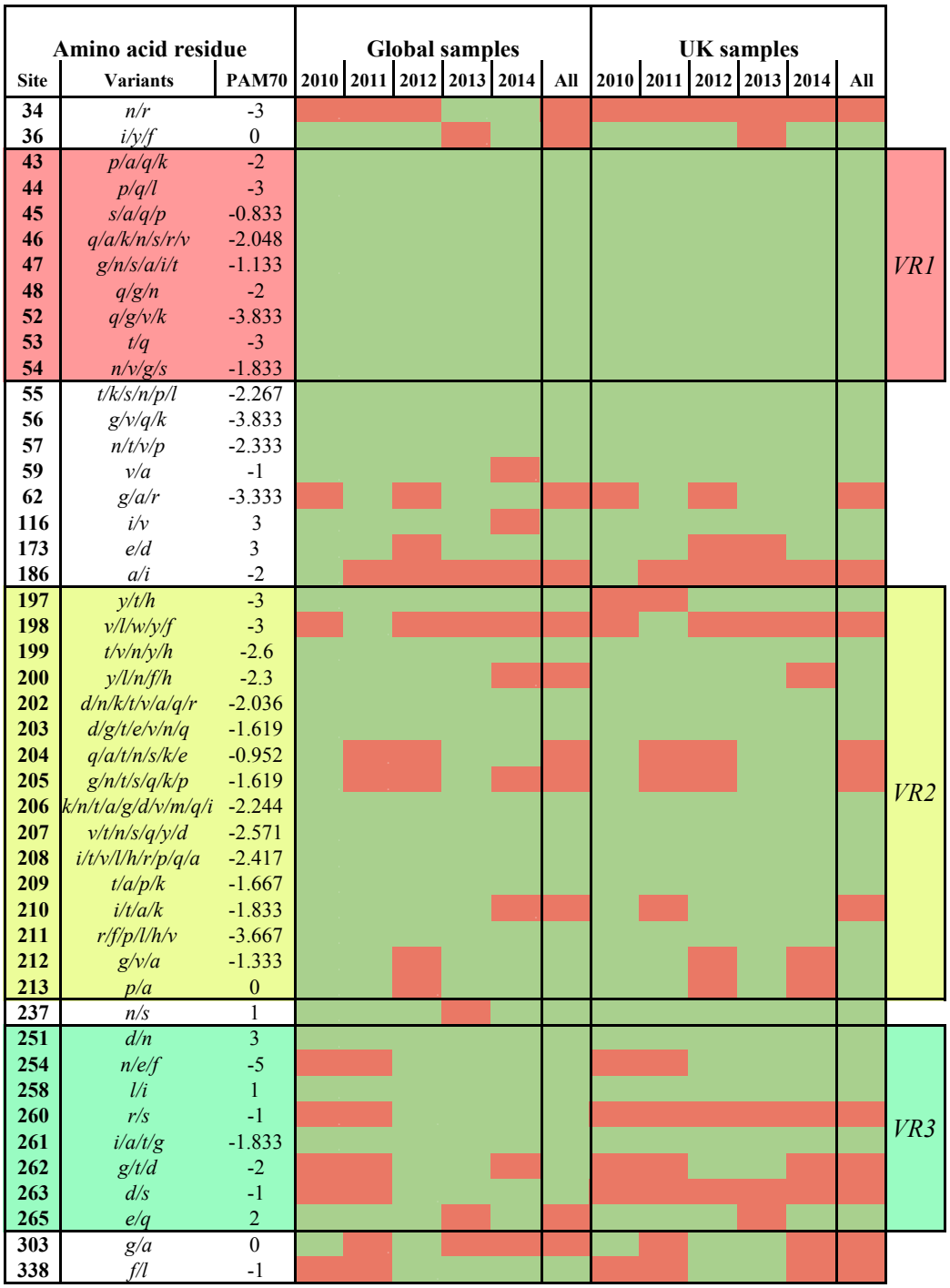


Table 3-2 Amino acid residues that show non-overlap with others within the PorA sequence of each of the spatiotemporal subsets *N. meningitidis*. If a particular "site of interest" shows non-overlap with at least one other residue in a particular subset of isolates it is marked in green, otherwise it is marked in red. Residue numbers in the alignment, as well as the variants present (in > 1% of serogroup B samples), are shown.

3.3.2 Plasmodium falciparum AMA-1

Of the 386 amino acid residues in our alignment of 1198 *P. falciparum* AMA1 sequences, 168 showed some polymorphism. The distribution of numbers of sites at each level of polymorphism is shown in Figure 3-3. A total of 12 sites in AMA-1 showed strong non-overlap with other sites ($MI_S > 0.4$; Figure 3-3), and can be split into four discrete networks of non-overlapping sites (Figure 3-4).

The first (in order of 3' to 5' location) comprises the sites 197, 200, 206, and 225, all located within the first ectodomain of AMA-1. There is also non-overlap between sites 283 and 285, although the close proximity of these sites means that their interaction could be due to compensatory mutations that act to maintain the local tertiary structure of the protein. There are then two trios of interactive sites located in (or immediately adjacent to) domain III of the protein: i) 439, 451, and 485, and ii) 496, 503, and 512.

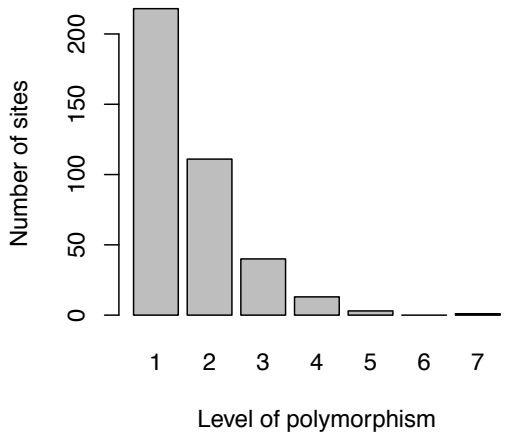


Figure 3-3 Counts of numbers of sites at each levels of polymorphism

Domain	Site	Residues	PAM70 likelihood	Parsimony score
I	197	*q/e/d/g/h/r*	-1.7333333	168
I	200	*h/d/r/l/v*	-3.9	117
I	206	*k/e*	-2	37
I	225	*i/n*	-3	40
I	283	*l/s*	-6	93
I	285	*e/q*	2	80
II	439	*n/h/d*	1	130
III	451	*k/m*	0	137
III	485	*k/i*	-4	139
III	496	*i/m*	1	156
III	503	*n/r/h*	-0.6666667	163
III	512	*k/r*	2	170

Table 3-3 Location, domain, amino acids present, conservativism, and phylogeny of the non-overlapping sites of interest in AMA1

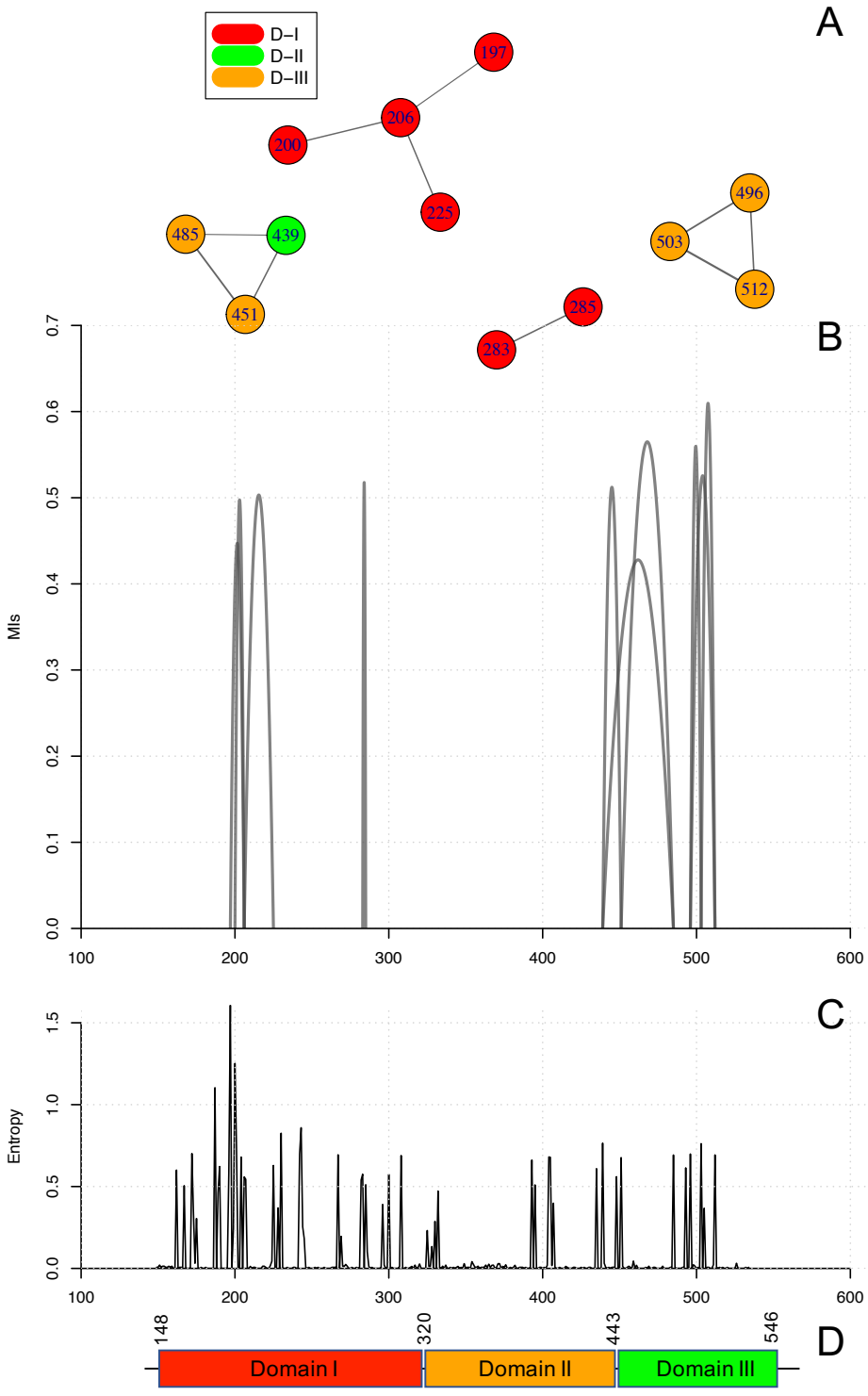
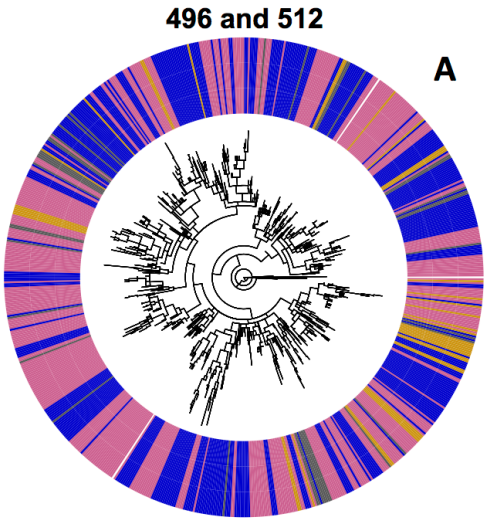
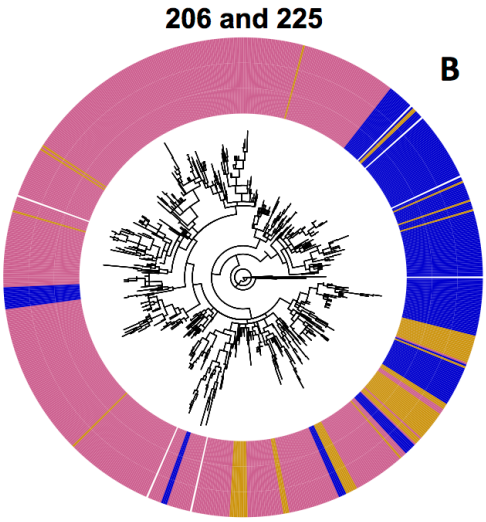


Figure 3-4 **A** The four sets of amino acid sites within AMA-1 that show non-overlap coloured by ectodomain (see D). **B** The interactive sites within AMA1; curved lines link sites that show non-overlap (MIS>0.4), height of the curve equates to MIS non-overlap score. **C** Entropies of polymorphic sites along the AMA1 protein. **D** Locations of the ectodomains on the linear protein sequence.

As discussed previously, it is possible that patterns of non-overlap could be generated by the phylogenetic history of variants, rather than by immune selection pressure alone. If, at a particular pair of sites of interest, the variant pairs are interdigitated across the tree, then we can say that the pattern that we see has emerged independently of the isolates phylogenetic history. This pattern can be quantified using a Parsimony Score (PS), as described in *Methods* and given for each site of interest in Table 3-3. These patterns can also be assessed visually by mapping variants at pairs of sites onto a phylogenetic history of the protein (Figure 3-5). In the centre of each of these figures is a phylogenetic tree drawn from the entire AMA1 sequence, and at the tips are coloured the amino acid variant pairs from the corresponding pair of interactive sites identified as being in non-overlap. Note that these phylogenetic trees are stylised into a circular form; the two tips of the tree either side of the 3 o'clock position are only distantly related, despite appearing adjacently to each other in the circularised tree.

From Figure 3-5 it is evident that the first pair (496 and 512; Figure 3-5A) have a high geometric mean parsimony score (GMPS) and variants appear dispersed in the tree in a pattern not suggestive of being easily explained by ancestry alone. In the second pair of sites (206 and 225; Figure 3-5B), there is a relatively low GMPS and so a pattern of amino acid variants in congruence with a pattern likely to emerge from ancestry alone.





GMPS = 162.8496

Colour of tree tip	Count of pair	Amino acid variants	
		Site 496	Site 512
	552	i	k
	523	m	r
	68	i	r
	54	m	k
	1	k	k
	0	k	r

GMPS = 38.4708

Colour of tree tip	Count of pair	Amino acid variants	
		Site 206	Site 225
	806	e	n
	209	k	i
	96	e	i
	6	k	n

Figure 3-5 Amino acid variants at two pairs of non-overlapping sites painting onto the tips of phylogenetic trees drawn from the *P. falciparum* AMA1 protein. Tables below give the amino acid variants that exist at each of the sites, a count of each variant pair, and the colour in the figure that the pair corresponds to for **A** 496 and 512, and **B** 206 and 225.

		MIs		
		> 0.5	> 0.7	> 0.9
Between and within genes	Genes with SoI	58	45	21
	Total SoI	443	189	57
	SoI in EBNAs	132	53	20
	SoI in LMPs	37	30	6
	SoI in glycoproteins	38	24	3
Between genes	Genes with SoI	58	36	12
	Total SoI	319	94	14
	SoI in EBNAs	77	17	0
	SoI in LMPs	28	15	0
	SoI in glycoproteins	33	13	1

Table 3-4 The numbers of sites of interest (SoI) in across the whole of EBV, and within certain groups of genes: those encoding Epstein Barr Nuclear Antigens (EBNAs), the Latent Membrane Proteins (LMPs), and glycoproteins.

3.3.3 Analyses of Epstein Barr Virus (EBV)

3.3.3.1 EBV – overall patterns of non-overlap

Analyses of amino acid sequences translated from protein coding regions of the whole genome of EBV found many sites that exhibit strong patterns of non-overlap consistent with strain structuring borne of immune selection pressure. There is evidence of selection driving co-evolution of amino acid pairs both within and between genes. Above the most stringent cut-off of MI_S score (0.9, "very strong non-overlap"), there are 57 amino acid sites that interact with others both within the given gene and with other genes, 14 of which interact with amino acid sites in other genes. Above the less stringent threshold of MI_S of 0.7 ("strong non-overlap") there are 189 sites that show non-overlap with other sites, 96 of which interact with other genes. The numbers of sites of interest (SoI) in each gene class is shown in Table 3-4, and complete lists of numbers of sites of interest per gene at each threshold level of non-overlap are in Appendix B; Table S2. The numbers of interactions that occur between each gene at each threshold level of non-overlap are shown in Figure 3-6.

The vast majority of amino acid residues that we find to be under strong non-overlap do not fall within known T cell epitopes or antibody binding sites (Appendix A; Table S1). This can be explained by the fact that most of the known epitopes show almost total conservation, a stark contrast to our sites of interest which, by their very definition, show at least intermediate levels of entropy. Indeed, the minimum entropy of any residue that shows strong non-overlap is 0.472, whereas many of the known T cell epitopes show total or near-total conservation (Appendix B; Table S1).

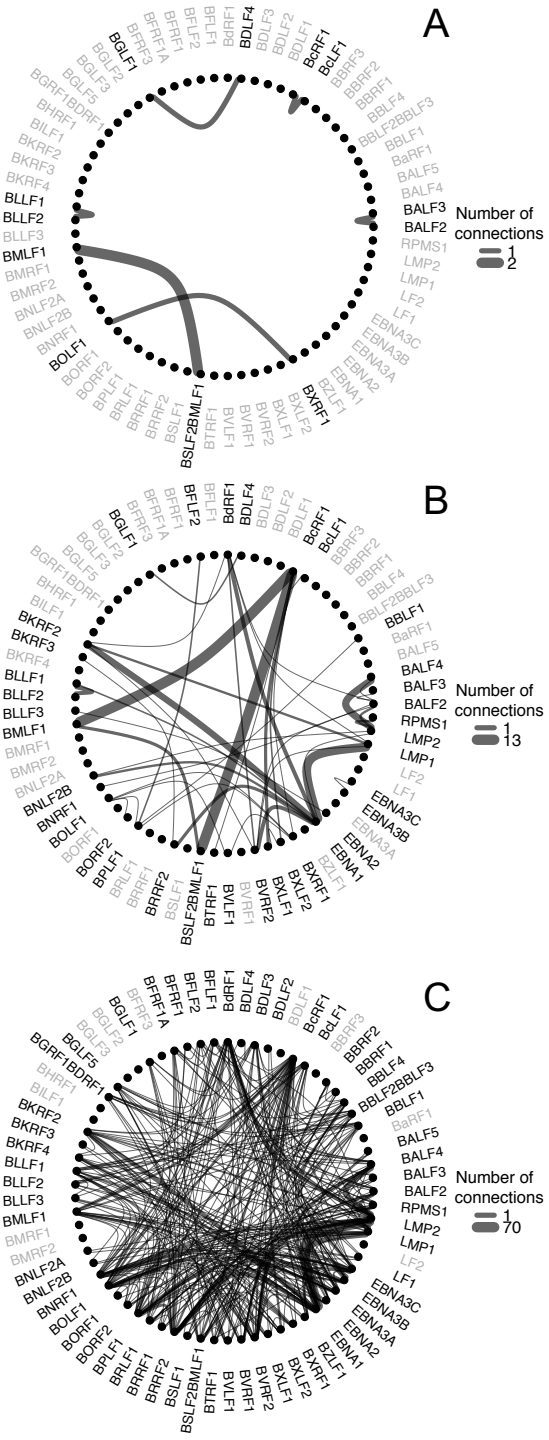


Figure 3-6 The numbers of interactions between genes of EBV. Note that genes are arranged alphabetically, and not according to genomic location – for the precise genomic location of interactive sites see Figure 3-7. The thickness of links between genes indicate the number of non-overlapping site pairs between those genes at levels of non-overlap: **A** MI_S>0.9, **B** MI_S>0.7, and **C** MI_S>0.5.

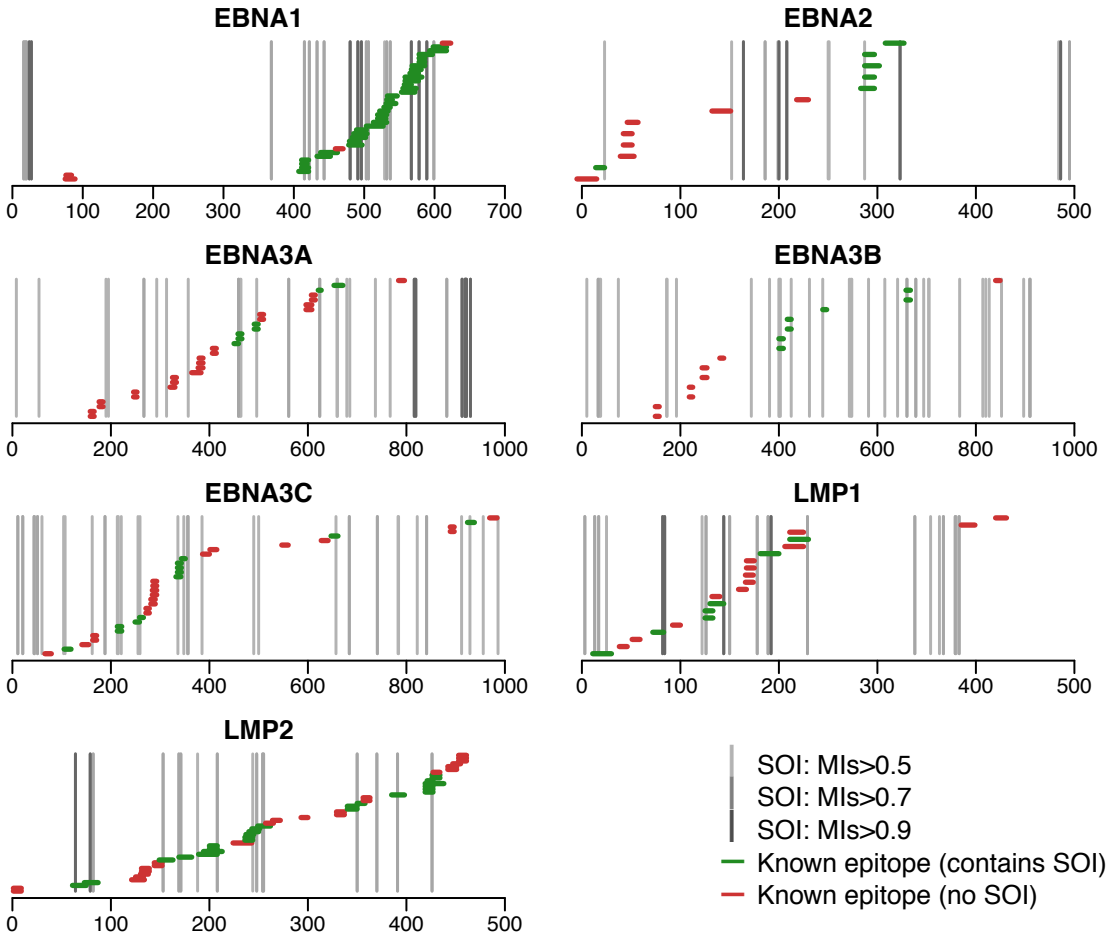


Figure 3-7 Maps of sites of interest found by our analysis (vertical lines) and known T cell epitopes (horizontal lines) in each of the latency-associated genes in which sites of interest (SOI) were found. Values on x-axis correspond to amino acid site in the protein.

3.3.3.2 EBV – patterns of non-overlap in latency-associated genes

The latent infection of B cells is crucial to viral persistence and so genes expressed at this stage of infection have received substantial scientific interest. Numerous works have found immune responses that act against latently expressed EBV proteins, with several comprising potential targets for vaccination (Elliott et al., 2008). Recent work describes the production of virus-like particles containing functionally inactive latent proteins (EBNA-3A, EBNA-3B, EBNA-3C, LMP1), along with a lytic protein (BZLF1) which showed high immunogenicity through both humoral and T cell responses *in vitro* and in a murine model (Ruiss *et al.*, 2011).

Therapeutic vaccine development against latent proteins has focussed on those expressed in the type 1 and type 2 latency patterns – those most associated with many of the clinical implications of EBV infection. Expression of EBNA-1 is associated with the type 1 latency pattern, which has been implicated in gastric carcinoma. A number of other carcinomas and lymphomas are associated with the type 2 latency pattern, which sees expression of both the LMPs and EBNA-1. Experimental infusion of EBV-specific cytotoxic T lymphocytes (CTLs), or LMP1 and LMP2-specific CTLs has been found to reduce EBV DNA loads and/or improve clinical prognosis in individuals with Hodgkin or non-Hodgkin lymphoma or nasopharyngeal carcinoma (Roskrow *et al.*, 1998; Straathof *et al.*, 2005; Smith *et al.*, 2012; Bollard *et al.*, 2014; Chia *et al.*, 2014). This suggests that a vaccine that induces such T cells may be effective, although development of such a vaccine has, so far, been limited.

Here, I find 169 sites within genes expressed at the latent stage of infection that exhibit non-overlap with other sites (MIs>0.5), 83 that exhibit strong non-overlap (MIs>0.7), and 26 that exhibit very strong non-overlap (MIs>0.9). These are sites that are located

in all of the EBNA and LMP families of genes (Appendix B; Table S2). In LMP2 and EBNA1 the majority sites of interest fall within known CD4 or CD8 T cell epitopes (Figure 3-7). This is in contrast to other latency-associated genes in which the majority of our sites of interest fall outside of known epitopes. Note, sequence data were not available for the latently expressed gene EBNA-LP or for the EBERs, a pair of short non-coding RNAs that, unlike EBNA-2, EBNA-3A, EBNA-3C, and LMP1 (Wang *et al.*, 1991; Kaye, Izumi and Kieff, 1993; Tomkinson, Robertson and Kieff, 1993), are not essential for transformation of B-lymphocytes (Yates, Warren and Sugden, 1985).

Most of the non-overlapping sites found within EBNA-1 fall within a 200-amino acid C-terminal region of the gene known to be epitope-rich and highly immunogenic (Chen et al 1999). EBNA-1 is amongst the most clinically important of the latent stage genes. (Wilson, Bell and Levine, 1996; Kennedy, Komano and Sugden, 2003). It is known to play a key role in actively binding viral DNA plasmids to host metaphase chromosomes to ensure that a lifelong pool of latently infected cells may be maintained. This allows the virus to persist without requiring lytic replication, the presence of which can alert the host's immune system to the presence of viral infection (Middleton and Sugden, 1994). It is also known to play a non-essential role in lytic activation: *in vitro* downregulation of EBNA1 slows the outgrowth of infected B cells (Wilson, Bell and Levine, 1996) and has been directly implicated in a number of carcinomas by inhibiting apoptosis of infected B cells and modification of other normal host regulatory function (Sivachandran *et al.*, 2012; Wang *et al.*, 2014). Potent vaccine induced immune responses against this protein could therefore act to drastically reduce the clinical prognoses of EBV, as well as having the potential to clear latent infection in infected individuals.

The most advanced EBV vaccine candidate to date, MVA-EL, acts by induction of immune responses against LMP2 and EBNA1. A phase 1 trial of the vaccine found immunity against both antigens to be present in at least half of patients, and to be raised against epitopes that vary between strains of EBV that vary geographically (Taylor *et al.*, 2014).

3.3.3.3 EBV – patterns of non-overlap in lytic phase genes

The EBV lytic cycle is the process by new virions are made, with expression of a specific set of genes required for each step in the production process. The lytic phase can be divided into three stages, as identified by the genes expressed in each: i) immediate-early genes, the products of which primarily serve to induce expression of genes in the later stages; ii) early genes, the products of which play a variety of roles in replication, metabolism, and immune evasion; and iii) late genes that encode proteins that are structural and/or play a role in immune evasion, for example in forming the viral capsid antigen.

The majority of known T cell epitopes in genes expressed in the lytic phase fall in the genes BHRF1, BMLF1, BMRF1, BRLF1, and BZLF1. Our approach, however, finds only one amino acid site within a known epitope, site 195^{QH} in BZLF1, to show non-overlap. This can be explained by the fact that, unlike the latent gene epitopes discussed previously, the majority of the known T cell epitopes in lytic phase genes show total or near total conservation (Appendix B; Table S2). T cell responses to lytic cycle genes are heavily skewed towards immediate-early and early genes (Pudney *et al.*, 2005). Our analysis finds a large number of sites that exhibit non-overlap with some other sites in genes from all stages of the lytic cycle (Figure 3-6).

Two genes are simultaneously expressed at the immediate early stage: BZLF1 and BRLF1. Whilst we find no interactions to occur between these genes, there are sites within each gene that show intragenic non-overlap: in BRLF1 two sites (273^{MR} and 316^{EK}) show very strong non-overlap with each other (MI_S of >0.9), and in BZLF1 there are a network of 5 sites that show strong non-overlap (MI_S of >0.7) with each other (68^{TA}, 76^{SP}, 124^{TP}, 152^{VA}, and 163^{QL}).

A previous study identified a number of variant types in the immediate early genes, and sought to associate those types with lymphatic tumorigenesis (Yang *et al.*, 2014). The study found amino acid changes in BZLF1 and BRLF1 that correspond to those described here in EBV samples taken from lymphomas. That work presents mutations that occur at two sites in the DNA binding domain and four sites in the transactivation domain of BRLF1: the two DNA binding domain sites are those identified here as being under strong non-overlap (273^{MR} and 316^{EK}). This contrasts the sites found here in BZLF1; sites 68^{TA} and 76^{SP} lie within the transactivation domain, and the others lie within the "less important" sub-regions of the transactivation domain, and thus may play a lesser role in activation of subsequent genes in the lytic phase of infection.

A further amino acid change in the functional domains of BZLF1 in Yang et al was detected as being under non-overlap (MI_S>0.5) by our analysis. Changes in site 195^{QH}, which is located within the DNA binding domain, were strongly associated with samples from lymphomas and not with samples from healthy individuals. In our isolates, 82% of individuals had a H at site 195 rather than a Q, in agreement with the 83.3% of healthy samples which had a 195^{Q} in Yang et al (compared to 53.8%) of lymphoma samples. The rarity of the alternative sequence variant (Q) at this site could explain its ranking as being under "non-overlap", as opposed to "strong non-overlap" – this pattern could be reassessed in the future as more sequences become available.

The development of vaccines that induce immune responses against Immediate Early (IE) stage genes has sought to capitalise on the high natural immunogenicity of regions of BZLF1 and BRLF1 (Ruiss *et al.*, 2011; Hartlage *et al.*, 2015). The early immune responses against infection with EBV are dominated by T cell responses against BZLF1

(Bogedain *et al.*, 1995; Tan *et al.*, 1999; Precopio *et al.*, 2003), and a vaccine construct was shown to be capable of eliciting robust, memory T-cell responses in healthy individuals, cementing its place as a candidate component of an EBV vaccine regimen (Hartlage *et al.*, 2015).

In addition to the sites in the IE genes, we also find a number of amino acid sites to exhibit non-overlap with others in genes expressed in the early and late lytic phase. At the early lytic phase the following genes are known to contain T cell epitopes: BaRF1, BHRF1, BLLF3, and BMLF1. In BaRF1 and BHRF1 we do not find any sites to exhibit non-overlap, whereas in BLLF3 we find one site, 258^{SG}, which shows strong non-overlap with site 495^{SP} in the latent phase gene, EBNA2. We find there to be three sites that exhibit non-overlap in BMLF1: sites 2^{ED}, 32^{RQ}, 427^{GA}. Sites 32^{RQ} and 427^{GA} in particular show strong non-overlap with a number of sites in the lytic early protein BcRF1 (72^{MI}, 134^{NS}, 172^{RQ}, 182^{NK}, 309^{SA}, 318^{SA}, 401^{SF}), as well as slightly weaker interactions with other sites including BXLF1 sites 268^{MT} and 500^{TM}, BRRF2 site 374^{TA} and 385^{MT}, and EBNA3C site 348^{ILM}. BMLF1 site 427^{GA} falls just outside of a known CD8+ T cell epitope, whereas 32^{RQ} falls within a region that is previously not known to be epitope-rich.

We find amino acid sites that exhibit patterns of non-overlap in a number of lytic stage genes in which potential epitopes have not previously been reported (Appendix B; Table S1). These are too numerous to explore in detail here, however may comprise interesting targets for further bioinformatic or experimental analysis as potential vaccine candidates.

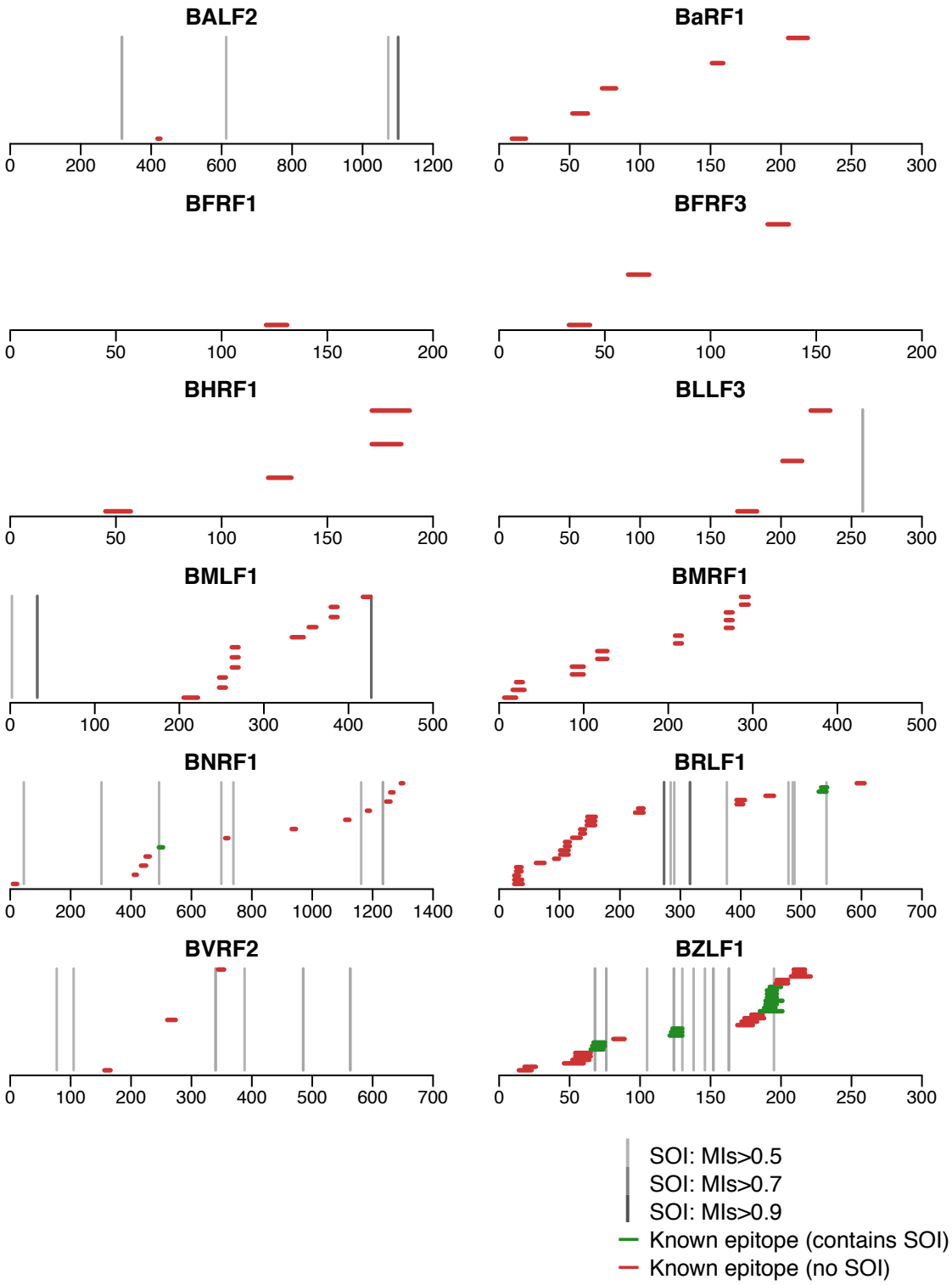


Figure 3-8 Maps of sites of interest (SOI) found by our analysis (vertical lines) and known T cell epitopes (horizontal lines) in each of the lytic stage-associated genes. Values on x-axis correspond to amino acid site in the protein.

3.3.3.4 EBV – patterns of non-overlap in genes encoding glycoproteins

The genome of EBV encodes 13 glycoproteins that perform a variety of functions including viral entry and spread, viral assembly, and host cell manipulation. They are often highly immunogenic, with 11 of them comprising components of the virion envelope. Several glycoproteins play a key role in the initial invasion of host B cells. Upon first contact of virion and host cell, gp350 (encoded by the BLLF1 gene) mediates attachment of the virus to the host CD21 receptor on the host B cell surface. A complex of glycoproteins then binds another host receptor, followed by envelope fusion mediated by gH (formerly known as gp85), gL (formerly gp25), and gB (formerly gp110), encoded by BXLF2, BKRF2, and BALF4, respectively.

As discussed in the Introduction, the most advanced prophylactic vaccine trials to date have focussed on the glycoprotein gp350, encoded by the BLLF1 gene. Until recently only two known epitopes in BLLF1 were described by the literature, however recent computer modelling of the interaction between gp350 and the monoclonal antibody 72A1 revealed a number of regions of the glycoprotein that bind to the antibody through H bonds or salt bridges (Tanner et al., 2015). Here we find 17 amino acid sites that exhibit non-overlap, 11 of which show strong non-overlap, and one of which shows very strong non-overlap. The site under very strong non-overlap, 819^{GD}, interacts with BLLF2 site 92^{VM}. None of the sites of interest found here fall within known T cell epitopes, however site 441^{AD} (which interacts with BLLF1 site 724^{TI}) does fall within a potential antibody-binding site.

Amongst the other glycoproteins, none of our sites of interest fall within known epitopes (Figure 3-9), although we do find sites exhibiting non-overlap congruent with immune selection pressure in gB (encoded by BALF4, interacts with BALF2 and

LMP2) and BDLF2 (interacts with EBNA3A and BBLF2/3), as well as BDLF3, gL (encoded by BKRF2), and gH (encoded by BXLF2), all of which show interactions with numerous sites in latent and lytic stage genes (Figure 3-6).

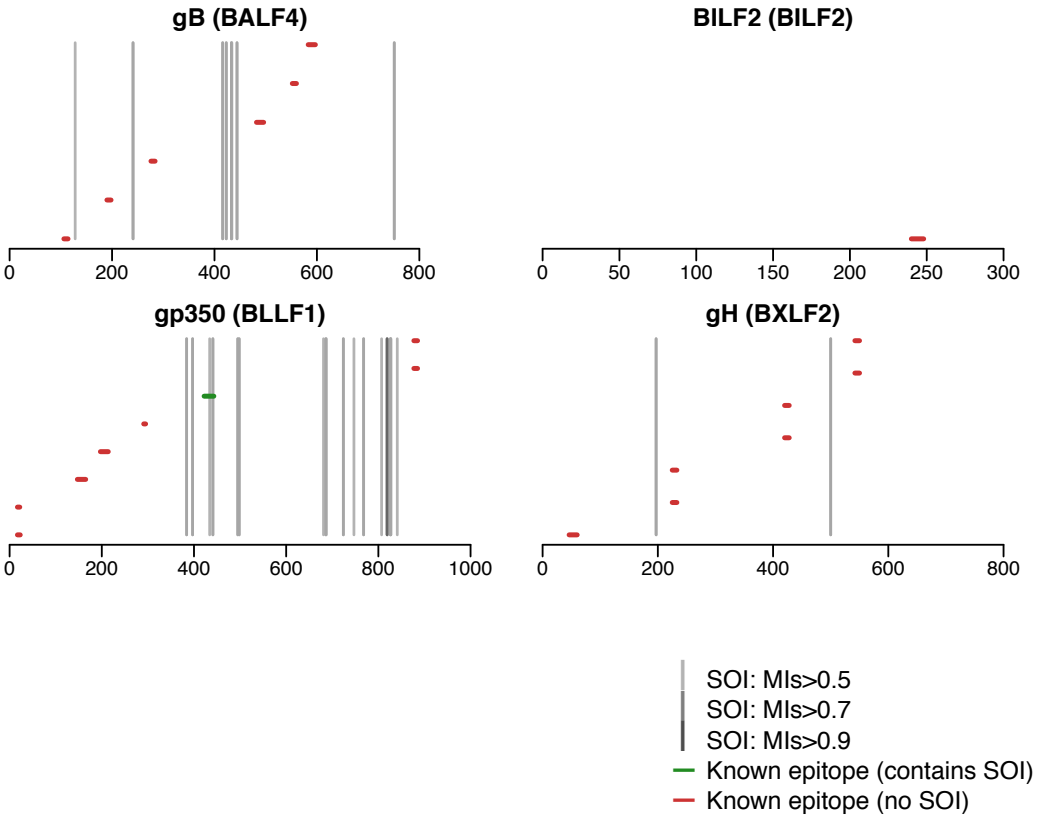


Figure 3-9 Maps of sites of interest (SOI) found by our analysis (vertical lines) and known T cell epitopes (horizontal lines) in each of the genes encoding glycoproteins. Values on x-axis correspond to amino acid site in the protein.

3.3.3.5 EBV – phylogeny of non-overlapping sites

By computing the Parsimony Scores of each amino acid site in the genome of EBV, I am able to analyse whether or not our non-overlapping sites of interest deviate from normal evolutionary patterns across the virus. As mentioned previously, the phylogenetic information presented herein does not account for frequent recombination in the EBV genome, however will still provide an extra insight into the role of immune selection pressure in shaping the pathogen's dynamics.

Note also that the maximum possible values of Parsimony Score is dependent on the number of isolates that are present in the tree. Thus, the scores found for EBV (with 83 sequences) are much lower than those presented for AMA-1 (with >1000 sequences).

The distribution of non-zero Parsimony Scores for sites in the EBV alignment are given in Figure 3-10. A parsimony scores of 0 indicates that a site fits perfectly with the phylogenetic history of the virus or, as is more commonly seen, is invariant. Ordered by MI_S score, the 17 of the top 20 non-overlapping sites in EBV have higher Parsimony Scores than 85% of all non-zero PS sites. The remaining 3 falls in the top 71%, 79%, and 79%. This result indicates that these sites have been driven to evolve to evade immune responses at multiple points along the virus's evolutionary history, as opposed to evolving at one point and then being maintained by neutral processes.

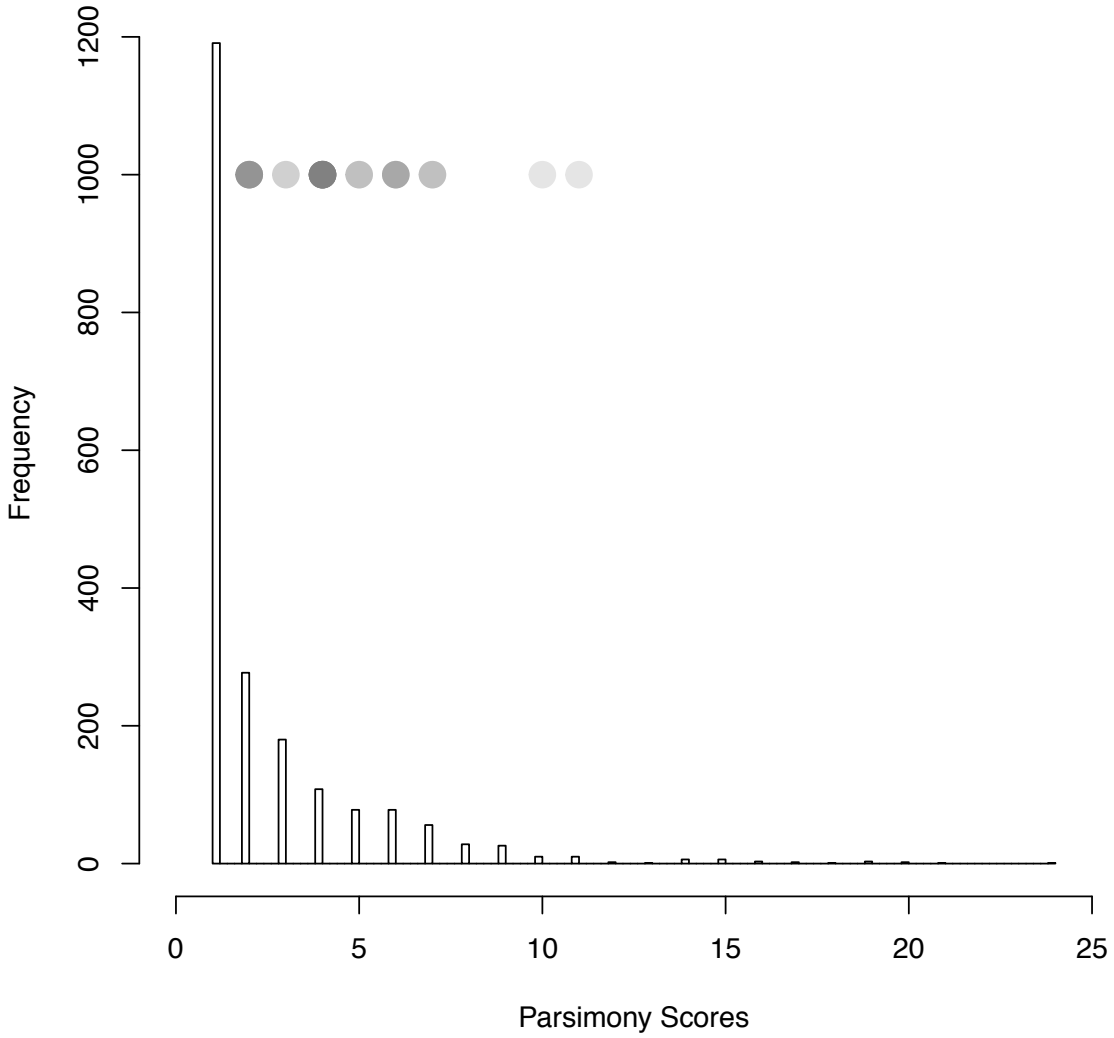


Figure 3-10 The distribution of Parsimony Scores of all sites in the EBV genome (sites with a Parsimony Score of 0 are not shown). The 20 amino acid sites in the top 10 non-overlapping pairs (as determined by MI_S) are indicated by translucent grey circles (multiple light-grey circles overlap to a dark grey circle).

3.4 Discussion

Here, I present an in-depth analysis of patterns of non-overlap in three pathogens from different orders: one viral, one bacterial, and one protozoan.

By exploring co-evolutionary dynamics in the OMPs of *N. meningitidis*, a set of proteins known to exhibit immune-mediated non-overlap (Buckee *et al.*, 2010; Watkins and Maiden, 2012), I build on the previous validation of methods conducted in Chapter Two. As well as corroborating previously identified interactions between variable regions (Gupta *et al.*, 1996), I identify a number of amino acid sites outside of the classically considered variable regions which may be of further interest to the design of the next generation of vaccines against bacterial meningitis. There are currently capsule polysaccharide (CSP)-conjugate vaccines against available against meningococcus types A, C, W-135, and Y, and an outer membrane vesicle (OMV) vaccine successfully used to control meningococcus B infection. The efficacy of these vaccines is estimated at 85-100% and lasts for at least 2 years (World Health Organization, 2011). Ongoing efforts seek to improve the longevity and efficacy of meningococcal vaccines through exploring new approaches to development, including developing synthetic constructs that present particular epitopes, DNA vaccines, recombinant proteins, live-attenuated inoculation, and refinement of the previously successful OMV and CPS approaches.

I then go on to investigate a key *P. falciparum* malaria antigen, AMA-1. This antigen has been the focus of much study, with several vaccine trials showing low efficacy and finding pathogenic diversity to be the likely cause. Here I postulate that these failures may be surmountable by using combinations of non-overlapping regions of the protein, specifically regions in Domain I and Domain III of the protein. My work

suggests that a vaccine construct comprising a combination of antigenic types containing natural combinations of amino acids at the identified sites may be highly immunogenic, and may be capable of enacting protective responses that are able to neutralise the multitude of parasite variants seen in nature.

The failures of numerous trials of AMA-1-based vaccines have been attributed to variability in this blood stage antigen. Strain-specific protection has been achieved, however challenges with alternate variants or naturally occurring mixtures of variants reduce efficacy to near 0 (Sagara *et al.*, 2009). Growth inhibition assays after immunization with a mixture of AMA-1 proteins derived from two reference strains of *P. falciparum*, 3D7 and FVO, found the mixture to be unable to cover the variant seen in natural settings (Mullen *et al.*, 2008). Increasing the number of different AMA-1 variants in the mixture showed promising results in animal trials (Drew *et al.*, 2012; Dutta *et al.*, 2013), with cross-reactivity shown to be improved by sequential, rather than simultaneous, administration of vaccine variants (Kusi *et al.*, 2011). The data presented here could assist in developing novel AMA-1 vaccine constructs which cover the diversity seen at sites showing patterns congruent with immune selection pressure.

Finally, I present an analysis of non-overlap across the whole genome of EBV. Immunodominant responses against viral infection are known to be T-cell focussed. T cell responses act against expressed viral proteins which have been cleaved intracellularly into short peptide sequences by the infected cell, transported to the cells surface, formed into complex with the host MHC molecules, and presented on the cell surface. By this mechanism the cellular immune response has access to many, if not all, of the proteins that are expressed intracellularly, and so we might expect that patterns of non-overlap may be likely to occur between, as well as within, viral

proteins. Indeed, we here find that patterns of non-overlap between proteins occur almost as readily as within proteins (Table 3-4). This mechanism of cellular immunity contrasts antibody-mediated responses which act against larger regions of the outside of a parasite and depend on the tertiary structure of the target. Thus, in pathogens where cellular immunity dominates, we expect to see non-overlap occurring inter-protein, whereas where antibody-mediated responses dominate, such patterns may be limited to occurring on an intra-protein level.

Previously identified epitopes in EBV are largely conserved. I here present a list of potential epitopic sites that are variable, suggest that immune responses may be effective against these sites, and postulate that they could comprise a novel vaccine candidate. The comprehensive list produced here may be refined by further bioinformatic analysis. For example, as described above, for an immune response to be raised against a particular region of a viral protein, it must be expressed, processed in a short peptide sequence by the 9mer, transported by the TAP (Transporter associated with antigen processing) protein, and finally bound to the host's HLA molecular. Each of these steps could be subject to extensive further study. First, gene expression analyses could be conducted on transfected cells *in vitro* using RNASeq, which can be used to quantify differential gene expression (Finotello and Di Camillo, 2015). If a particular gene is found to be expressed, then existing models can be used to model the MHC class I pathway, including quantifying proteasomal cleavage and TAP transport (Tenzer *et al.*, 2005).

Finally, after transport, the short peptide sequences (8mer – 13mer) that this pathway produces form a protein-MHC (pMHC) complex with the host MHC molecule. A number of tools are available (IEDB - National Institute of Allergy and Infectious

Diseases, 2017) to predict the binding affinity and stability of these complexes and so an estimate can be produced of the likelihood that a particular region of a protein of interest, i.e., those identified here for EBV, may be presented to the cellular immune system. Work has also sought to model the natural immunogenicity of a particular peptide sequence by analysing features and motifs in peptides of known immunogenicity (Calis *et al.*, 2013).

Data presented here could assist in the development of vaccines against *P. falciparum* malaria, *N. meningitidis*, and Epstein Barr Virus. I use large banks of sequence data curated from online repositories to identify and analyse epitopic regions of proteins from these three key pathogens, providing evidence in each of patterns of non-overlap that are in congruence with immune selection pressure. As such, inclusion of these regions in a vaccine construct could facilitate creation of novel therapies that are i) highly immunogenic, and ii) cover key strain-defining variation that occur in natural systems. Further analyses should explore potential structural-functional limitations that occur in the pathogens, and in the case of T cell-mediated immunity, the propensity for a given region of interest to be presented to the cellular immune system. It would also be possible to replicate the work presented here to investigate antigens within other pathogenic species for which novel vaccination strategies are required, and potentially expand analyses to investigate pathogens known to exhibit Cyclic, rather than Discrete, Strain Structuring.

Chapter 4 Modelling combinations of pre-erythrocytic *Plasmodium falciparum* malaria vaccines

Andrew S. Walker, Dr José Lourenço, Professor Adrian V. S. Hill, and Professor Sunetra Gupta

4.1 Introduction

In 2016, the WHO estimated there to be around 438,000 deaths from malaria per year, although others have predicted this value to be as high as 1,280,000 (Murray *et al.*, 2012). Despite this discrepancy, it is clear that malaria incidence has declined substantially, according to WHO estimates by 37% between 2010 and 2015, largely due to the increased distribution of insecticide-treated bed nets and artemisinin-based combination therapies (World Health Organization, 2015).

It is widely accepted, however, that further concerted efforts are required if the WHO's strategic targets for a 90% reduction in incidence and mortality by 2030 (relative to 2015) are to be met (World Health Assembly, 2016). An effective vaccine is likely to form a crucial component of control and eradication efforts, especially in light of recently emerged artemisinin resistance (Dondorp, Nosten and Yi, 2009) and multiple insecticide resistance (Yewhalaw *et al.*, 2011). The malaria parasite has a complex lifecycle, and vaccine research targets each stage (Hill, 2011); beginning with the initial sporozoite inoculum and subsequent liver stages, and onto blood stage infection, as well as aiming to prevent transmission by blocking sexual stage development of the parasite.

Much of the focus has been on development of a vaccine that induces an immune response against pre-erythrocytic (PE) parasite stages using either attenuated whole parasites (Epstein and Richie, 2013) or recombinant formulations that induce immune responses against particular epitopes. The former has shown greater efficacy against homologous challenge, with total short term sterile protection seen in individuals inoculated with very large doses of irradiated sporozoites (Seder *et al.*, 2013). This approach, however, faces substantial challenges in breeding large numbers of

modified *Plasmodium falciparum* parasites on an industrial scale, delivering cryogenically preserved live parasites, and administering them intravenously in the field (Seder *et al.*, 2013; Sheehy *et al.*, 2013). Small-scale studies into the durability of protection and efficacy against heterologous strains have shown promising preliminary results (Ishizuka *et al.*, 2016; Lyke *et al.*, 2017), and further field trials are ongoing.

There continues to be a wealth of research focusing upon recombinant vaccines that could provide readily deployable vaccine options. RTS,S/AS01 (trade name Mosquirix®) is a subunit vaccine based on the circumsporozoite protein (CSP) that was first developed in a collaboration between GSK and the Walter Reid Army Institute of Research in the early 90s (Vreden *et al.*, 1991). It is comprised of genes from the repeat and T-cell epitope of CSP with a viral envelope protein from hepatitis B (HbsAg) and a chemical adjuvant (AS01). RTS,S/AS01 is the first vaccine to be licensed against malaria, indeed the first to be licensed against any human parasitic disease. A large-scale phase III clinical trial completed in 2014 demonstrating modest efficacy against severe malaria, hospitalisation, and clinical malaria in children (Agnandji *et al.*, 2012; RTSS Clinical Trials Partnership, 2014; Olotu *et al.*, 2016). Protection is conferred mainly by anti-circumsporozoite protein (CSP) antibodies, although a minor contribution from CSP-specific $CD4^+$ T-cells that help antibody production cannot be ruled out (White *et al.*, 2013). In 2009, a phase IIa Controlled Human Malaria Infection (CHMI) trial of RTS,S/AS01B on immune-naive adults showed a sterile efficacy of around 50% (95% CI, 32.9% - 67.1%) (Kester *et al.*, 2009). This level of protection has been computed to result from a 96.1% (95% CI, 93.4% - 97.8%) reduction in sporozoite numbers (White *et al.*, 2013).

ChAd63-MVA ME-TRAP is a PE subunit vaccine candidate in development that has been shown to induce modest sterile efficacy in clinical trials. It enacts a protective $CD8^+$ T-cell response against infected hepatocytes, and was demonstrated to provide 21% sterile protective efficacy in a phase IIa CHMI trial, estimated to result from a 96% reduction in liver parasite burden (Ewer *et al.*, 2013). This trial found a weak correlation between the time to patency and concentration of IFN-γ^+ mono-function $CD8^+$ T-cells. Optimal results were seen when the vaccine was administered by a prime-boost regimen in which multiple epitopes of the thrombospondin-related adhesion protein are inserted into chimpanzee-derived simian adenovirus 63 (ChAd63) and a modified virus Ankara (MVA) (Ewer *et al.*, 2013).

In conjunction to measuring sterile efficacy (that is, the proportion of patients who do not get any infection on challenge), vaccine trials also aim to quantify partial efficacy (that is, reductions in numbers of parasites as a result of vaccine-induced immune clearance). The latter can be measured by a delay in the time to blood-stage infection, which is indicative of a substantial reduction in parasite load. Previous models have suggested that more than 95% of pre-erythrocytic parasites must be cleared for there to be a 2 day delay in time to patency, as measured by microscopic blood slide analysis (Bejon *et al.*, 2005). These results suggest that the reported sterile efficacies may not truly reflect the high potencies of the vaccines tested, as evasion of a vaccine-induced immune response by only a very small number of pre-erythrocytic parasites is sufficient for blood-stage infection (Bejon *et al.*, 2005). In other words, PE vaccine efficacy depends upon sterile protection. In contrast, a protection blood stage vaccine would instead require reduction of the erythrocytic Parasite Multiplication Rate (PMR) below a threshold at which infection cannot persist (D. Douglas et al., 2011).

Mathematical models suggest that incremental potency increases at high levels of sporozoite destruction (or impediment) can have profound effects on sterile efficacy (White and Smith, 2013). This supports the previously proposed notion that combinations of vaccines acting via distinct biological mechanisms could act synergistically (Duffy and Kaslow, 1997; Hill, 2006; Hutchings *et al.*, 2007). We might not expect there to be a linear relationship between degree of parasite killing by a vaccine induced response and the proportion of individuals who are protected as, even at moderate levels of killing, small numbers of parasites may evade destruction and allow infection to persist. Combinations of two moderately efficacious (but highly "potent") vaccines, may circumvent this problem: one will remove the small numbers of parasites left behind by the other.

Thus, the aforementioned PE vaccines, each with well in excess of a 90% reduction in parasite numbers (Bejon *et al.*, 2005; Ewer *et al.*, 2013) and discrete methods of action, could prove a potent combination. Such an effect has previously been noted in a murine model, in which a pair of analogous T-cell and antibody inducing vaccines, each with around 30-35% sterile efficacy when administered alone, elicited 90% sterile efficacy upon their combination (Hutchings *et al.*, 2007). Similarly, it has been postulated that the efficacy of whole parasite approaches may be due to induction of numerous, independent arms of the immune system (White and Smith, 2013).

Here I design a within-host mathematical model that allows us to investigate the effects of combining anti-malarial vaccines acting at different stages of the life cycle. I fit the model to phase IIa trial data from a moderate efficacy PE subunit vaccine (RTS,S/AS01) and a candidate currently under trial (ME-TRAP), to parameterise the within-host dynamics of *Plasmodium falciparum* with and without treatment (Kester

et al., 2009; Ewer *et al.*, 2013). I create a baseline model of within-host dynamics, test the implications of vaccine-induced immune responses for efficacy and partial efficacy, and then simulate the concurrent administration of multiple vaccines. I show that vaccines targeting parasites at different life history stages can act synergistically: near total sterile protection could be obtained by combining two vaccines which, administered alone, show modest efficacy.

4.2 Methods

4.2.1 The model

Within-host parasite dynamics are modelled by the following system of equations describing the rates of change in numbers of infected hepatocytes (H) and merozoites (M), with parameter described in Table 4-1.

$$\frac{dH}{dt} = -\alpha_2 H \quad (1)$$

$$\frac{dM}{dt} = \gamma M \quad (2)$$

$$H_0 = (1-\alpha_1) f S_0 \quad (3)$$

$$M_T = r H_T \quad (4)$$

Here, H_0 and M_T specify starting conditions for the respective populations, with M=0 when $t < T$, where T is the length of the incubation period within the liver (in days). There are several stochastic events in the development of malaria infection, made apparent in CHMI trials by differences between individuals in the time taken to reach defined thresholds of parasitemia. The model captures this behaviour by incorporating inter-individual variation in the sporozoite inoculum (S_o), probability of liver invasion (f), incubation period (T), and erythrocytic growth rate (γ). It is assumed that an individual receives 5 bites (as per protocol), where the size of each sporozoite inoculum, S_i (such that the total initial inoculum $S_o = \Sigma S_i$), is sampled from a negative binomial distribution obtained by fitting a negative binomial model to counts of parasites inoculated per-bite by anophelene mosquitoes (Medica and Sinnis, 2005). The proportion of sporozoites invading liver cells (f) is drawn from a Poisson distribution with shape parameter $\varepsilon = \frac{S_0}{\lambda}$, where λ is a modifier of sporozoite success.

Vaccine-induced anti-sporozoite immune responses reduce the probability of invasion by $(1 - \alpha_1)$, while infected hepatocytes are removed by a liver-stage vaccine at a rate α_2. Intra-hepatocytic development takes approximately 7 days, and at the end of this period (T), the merozoite population size is set to rHT, where r is the number of merozoites released per hepatocyte. Erythrocytic growth occurs at a rate γ, sampled from a Gaussian distribution with a given mean and standard deviation (SD; $\gamma_{mean}$and γ_{sd}, respectively). The model was implemented in MATLAB R2013b (The MathWorks, Inc., Natick, MA), and the ODE system solved using a non-stiff Runge-Kutta solver, ode45.

Parameter	Explanation	Value
α_1	Vaccine-induced modifier of sporozoite invasion probability: $\alpha_{1,min} = 0$, $\alpha_{1,max} = 1$	MCMC fitted
α_2	Vaccine-induced rate of removal of infected hepatocytes: $\alpha_{2,min} = 0$, $\alpha_{2,max} = 1$	MCMC fitted
μ	Rate of sporozoite loss, set such that sporozoites are removed from system at a realistic rate	20
T	Liver incubation time: time from sporozoite inoculum to merozoite release	Selected from Gaussian distribution, mean = 7 days, SD = 0.5 days
r	Successful merozoites per hepatocyte	10,000
f	Proportion of sporozoites that successfully invade hepatocytes: if, $f > 1$, $f = 1$	Stochastically selected from Poisson distribution (see Methods)
λ	Shape determinant for Poisson cumulative density function from which sporozoite success rate is stochastically sampled: $\lambda_{max} = 10$; $\lambda_{min} = 1$	MCMC fitted
S_0	Initial inoculum size, based on sum of five samples (five bites) from negative binomial distribution	–
P	Success parameter for sampling from negative binomial to give per bite inoculum size	MCMC fitted
R	Shape parameter of inoculum size negative binomial[19]	0.246
γ_{mean}	Mean of EGR: $\gamma_{mean,max} = 5$; $\gamma_{mean,max} = 0$	MCMC fitted
γ_{sd}	SD of EGR: $\gamma_{sd,max} = 1.25$; $\gamma_{sd,max} = 0$	MCMC fitted
γ	EGR: if $\gamma < 0$ then $\gamma = 0$	Sampled from Gaussian of mean γ_{mean} and SD γ_{sd}

EGR = erythrocytic growth rate; MCMC = Markov chain Monte Carlo; SD = standard deviation.

Table 4-1 Parameter values and initial conditions for the model

4.2.2 Model fitting to trial data

Malaria challenge trials are terminated in a given individual when parasites are detected by microscopy (thick smear analysis), although qPCR detection may occur several days prior to this. Net parasite numbers at each of these clinical endpoints are computed on the following basis:

i. qPCR detection methods can reliably detect 20 parasites/mL of blood (Andrews *et al.*, 2005) and, given ~4.7L of blood in an adult human, this equates to 94,000 parasites in a single host
ii. Blood slide methods generally detect parasitemia at when between 20 and 50 parasites per microliter are present (Guerin *et al.*, 2002): an estimated mean of 35 was therefore used, equating to 1.645×10^8 parasites across the body.

These estimates assume homogeneity in parasite density across the host.

Survival curves of the proportion of individuals who have reached these two levels of parasitemia are compared to the model outputs. A Markov-chain Monte Carlo (MCMC) fitting regime is used to find combinations of parameters that provide the best fit of the model to the data, with each step of the MCMC routine effectively representing one distinct, *in silico* clinical trial. To establish a baseline parameter set, the model was first fit to data for individuals who did not receive a vaccine (parameters marked "MCMC fitted" in Table 4-1) – α_1 and α_2 are set to 0 in this exercise. These parameters were then fixed and the model refit using the same MCMC protocol, except this time to data from vaccine recipients in the two CHMI trials: α_1 was fit to data from the RTS,S recipients, and α_2 to data from ME-TRAP recipients.

The fitting process follows a protocol described previously (Lourenço and Recker, 2014), which is effectively a random walk Metropolis-Hasting algorithm. First, a set of initial guesses for each of the parameters that is to be fit is generated. The model is run 500 times with these parameter values – this repetition accounts for stochasticity in the model (and seen in the biological system), and allows for the proportion of individuals that are at threshold levels of parasitemia at each time to be calculated. The least squared distance between this model output and the trial data is then computed.

New parameter values are then selected from a Gaussian distribution with a mean of the previous accepted parameter value, the model rerun, and a new least squared distance calculated. The following equation is then used to generate a probability of acceptance, ρ:

$$\rho = min\left\{1, \mathrm{e}^{-\chi^2(O(Y))-\chi^2\left(O(M^t)\right)}\right\}$$

where the quality of fit (χ^2) of model outputs (O) to the data are compared between the previous accepted parameter set ($O(M^t)$) and the proposed parameter set ($O(Y)$).

A random uniform number (between 0 and 1) is then generated, and if the random number is less than ρ, the new guesses are accepted, and the whole process repeated. If the new parameter set produces a better fit to the data than the old parameter set then ρ will be high (>0.5), and so there is a higher probability that the uniform random number will be less than ρ. In this case the parameter set is accepted and used to make the next parameter guesses.

Each MCMC chain was run for 500,000 steps, with parameters values selected by computing the median of the best fitting 100,000 steps (by Least Squares Distance between model and data).

A sensitivity analysis was performed by fixing the parameters at the values found by the MCMC procedure, and then sampling each parameter independently from a Gaussian. For each parameter value, the least-squares distance between the model output to the trial data was calculated and plotted. This fitting was conducted for inoculum size (S_0), lambda factor (λ), and erythrocytic growth rate mean (γ_{mean}) and standard deviation (γ_{sd}), and vaccine potencies (α_1 and α_2).

4.3 Results

4.3.1 Baseline within-host dynamics

Control subjects from CHMI trials provide an opportunity to model parasite dynamics within immune-naive individuals. The two studies used to parameterise the model (Kester *et al.*, 2009; Ewer *et al.*, 2013) utilized similar challenge protocols and the data presented no significant differences in times to patency ($t_{df} = 1.5409_{34}$, p=0.1326), nor was there a difference between either study and an additional study investigating times to patency following infected *Anopheles* bites (Lyke *et al.*, 2010). Similarly, there was no significant difference between the times to qPCR detection in the Ewer and Lyke studies ($t_{df} = 1.320_{22.755}$, p=0.2002) – qPCR data were not available for the RTS,S CHMI study (Kester *et al.*, 2009).

When the model was fit to the control subjects, the mean time to patency by blood slide analysis was 11.290 days (SD: 1.145), this compares to 11.038 (95% CI 8.074 to 14.002, n = 53) in the CHMI trials (Figure 4-1a). The mean time to the qPCR detection threshold of parasitemia in the model was 7.319 days (SD: 0.264), compared to 7.759 days (95% CI 5.573 to 9.944, n = 29) in the CHMI trials (Figure 4-1b). These values were obtained by running 500 stochastic simulations of the model with fixed parameters as estimated by the MCMC approach ("MCMC fitted" in Table 4-1). Approximately Gaussian parameter distributions indicate that the model converged on reasonable parameter estimates, and I found there not to be identifiability issues in the model (Appendix C; Figure S1/S2).

For completeness, I performed a sensitivity analysis (Figure 4-2) by resampling each parameter independently from a Gaussian distribution about a mean value found by fitting. We found the fit of the model to be poor at low inoculum sizes as, even without

vaccination, some individuals may not experience blood-stage infection due to the high rates of stochastic sporozoite destruction. A similar effect is seen at high λ values (Figure 4-2), where the success rate of sporozoites is low.

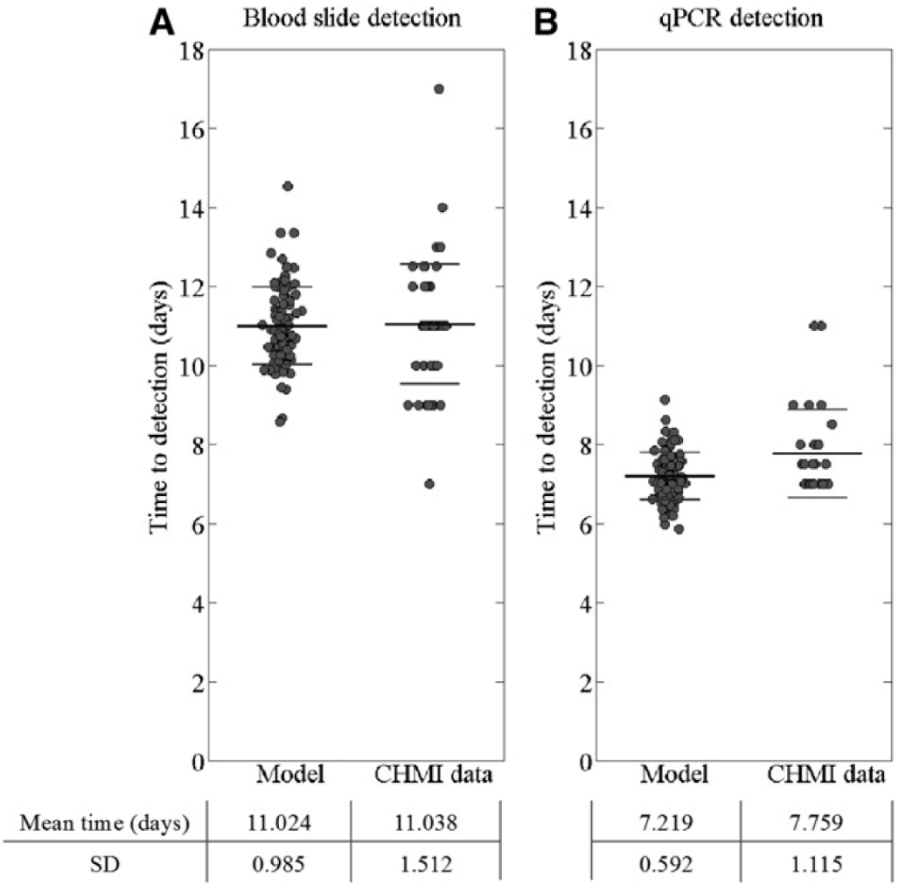


Figure 4-1 Times to (**A**) blood slide patency and (**B**) detection by qPCR in CHMI trials analysis (Kester *et al.*, 2009; Lyke *et al.*, 2010; Ewer *et al.*, 2013) and the model for control (non-vaccine receiving) individuals. Model data is based upon a random sample (n = number of individuals in corresponding CHMI data) of 1000 runs with parameters fixed at the values found by MCMC fitting. Median lines are shown.

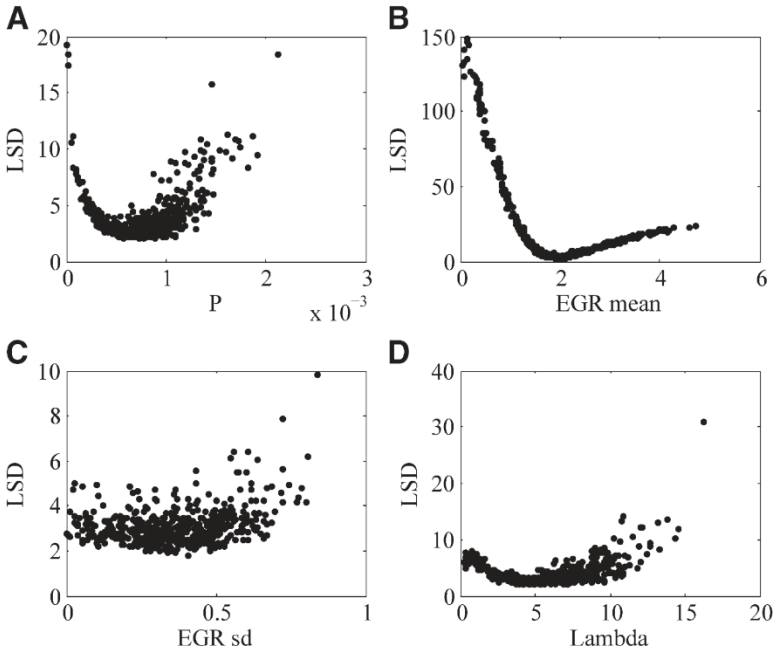


Figure 4-2 Results of a sensitivity analysis conducted on the model, demonstrating change in least squared distance (LSD) between the model and the CHMI data as a result of varying single parameters about a mean set by the MCMC procedure. Inoculum mean (**A**), λ factor (**B**), γ_{mean} (**C**) and γ_{sd} (**D**) are varied in a system with no vaccine.

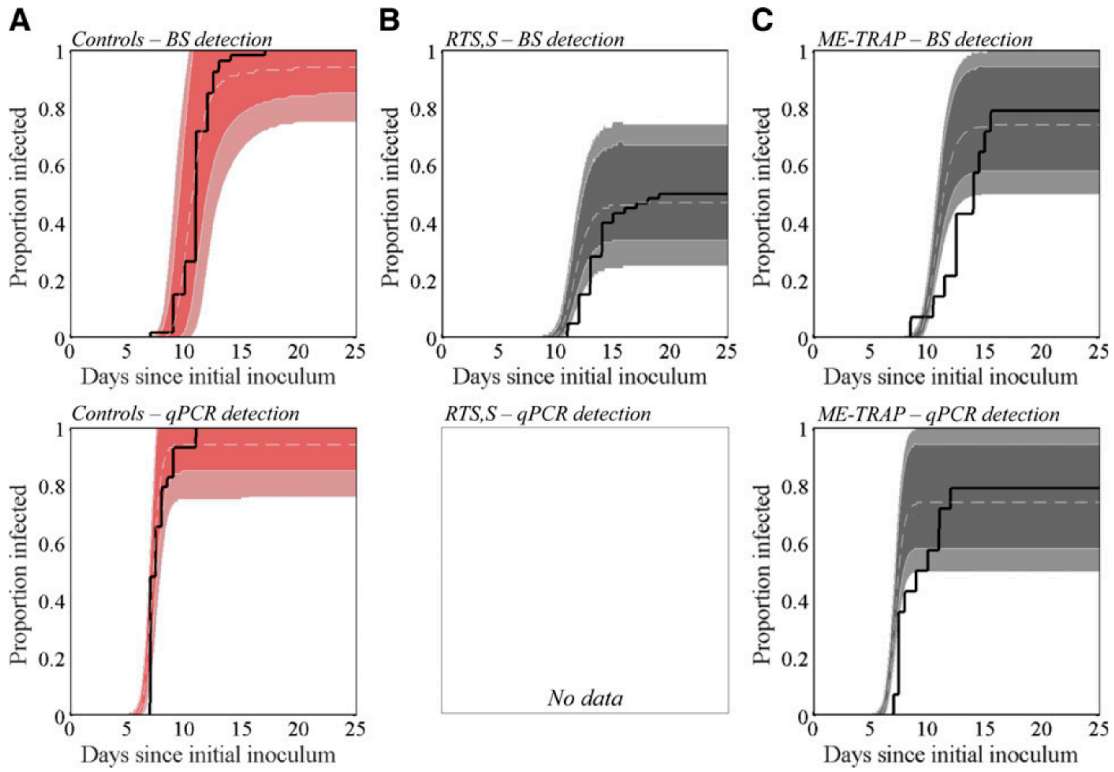


Figure 4-3 Survival curves showing the proportion of individuals who have not reached the thresholds of blood stage parasitemia required for detection by either blood slide (BS, top) or qPCR (PCR, bottom) analysis – no qPCR data were available for the RTS,S/AS01B study analysed. Solid lines represent data from CHMI trials on (**B**) ChAd63-MVA ME-TRAP (Ewer *et al.*, 2013) and (**C**) RTS,S/AS01B (Kester *et al.*, 2009), as well as on (**A**) controls collated from these studies and another (Lyke *et al.*, 2010). The dashed line represents the median of all post burn-in chain steps of the MCMC parameter inference protocol, and dark and light blue areas represent 99 and 90% credible intervals respectively.

4.3.2 Within-host dynamics in single vaccine recipients

Further MCMC fitting was conducted on both parameters that determine vaccine potency, with the model fit to CHMI data for vaccine recipients in trials of RTS,S and ME-TRAP (Figure 4-3; α_1 and α_2 respectively). Gaussian parameter distributions indicate convergence of the chain on a good parameter estimate (Appendix C; Figure S1/S2).

According to the model the sterile protective efficacy of RTS,S is 55.6% (SD: 8.21%), compared to 50% (95% CI 32.9% to 67.1%) in the CHMI trial (Kester *et al.*, 2009). In those individuals who were not protected the mean time to patency by blood slide analysis in the model was 12.881 days (SD: 1.53), compared to the CHMI result of 13.57 days (95% CI 9.78 to 17.37, Figure 4-3). Time to qPCR detection in the model was 8.825 days (SD: 0.79) (Figure 4-3) – this cannot be compared to trial output as qPCR data were not available in this trial.

ChAd-63 MVA ME-TRAP elicited 24.35% efficacy (SD: 8.37%) in the model, compared to 21.4% (95% CI 3.2% to 46.0%) in the CHMI study (Ewer *et al.*, 2013). The mean times to blood slide and qPCR detection in non-protected individuals in the model were 12.51 days (SD: 1.60) and 8.442 days (SD: 0.76) respectively, compared to 12.82 (95% CI 8.72 to 16.91) and 8.91 (95% CI 5.41 to 12.41) days in the vaccine trial (Figure 4-3).

We further demonstrate that minor variations in anti-infected hepatocyte vaccines (Figure 4-4, α_2) have a much more profound effect upon the fit of the model than variations in anti-sporozoite potency (Figure 4-4, α_1). This is due to there being a far

greater number of sporozoites than infected liver cells, thus small changes in α_2 are more likely to affect time to patency and sterile protection rates. Overall, the fit of the model to the RTS,S (α_1) data is also better than to ME-TRAP (α_2) data, in part due to only blood slide patency data being available, whereas both blood slide and qPCR data were available for ME-TRAP.

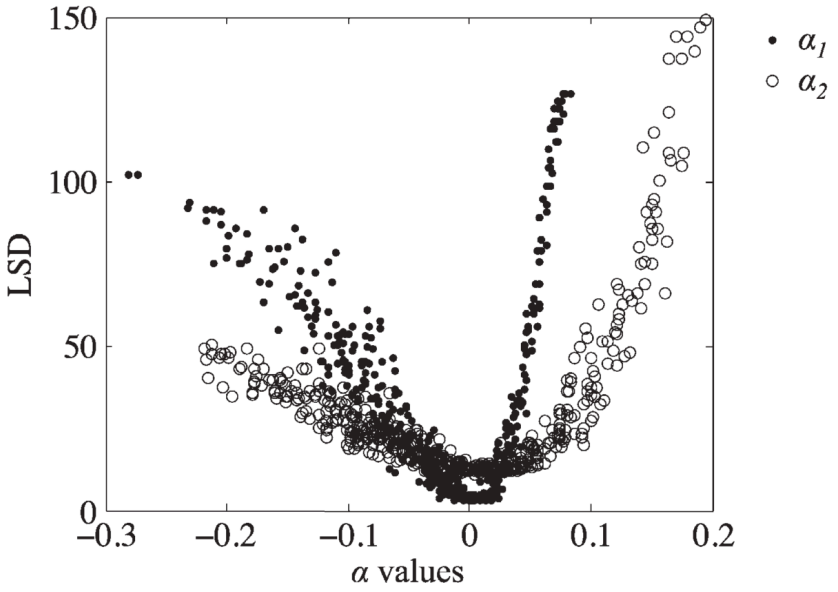


Figure 4-4 A sensitivity analysis of the vaccine parameters. The effect of independently varying α_1 and α_2 upon the least-squared distance (LSD) between the model and the data, with all other parameters fixed. Each point represents the mean of 100 runs of the ODE model, n = 1000.

4.3.3 Effects of combining vaccines

Sterile protective efficacies of every combination of possible vaccine potencies (α_1 and α_2) were evaluated (Figure 4-5). The model suggests that a combination of the potencies identified previously for RTS,S and ME-TRAP can confer almost total sterile protection. In fact, in 1000 independent stochastic runs of the model, less than 1% of individuals experienced any blood stage infection – in excess of 99% of infections were eliminated (Figure 4-5, square region contained within intersection of the pairs of vertical and horizontal lines). Interestingly, I also noted that combinations of vaccines that have <1% protective sterile efficacy when used alone are shown to have up to 27% efficacy (Figure 4-5) when combined.

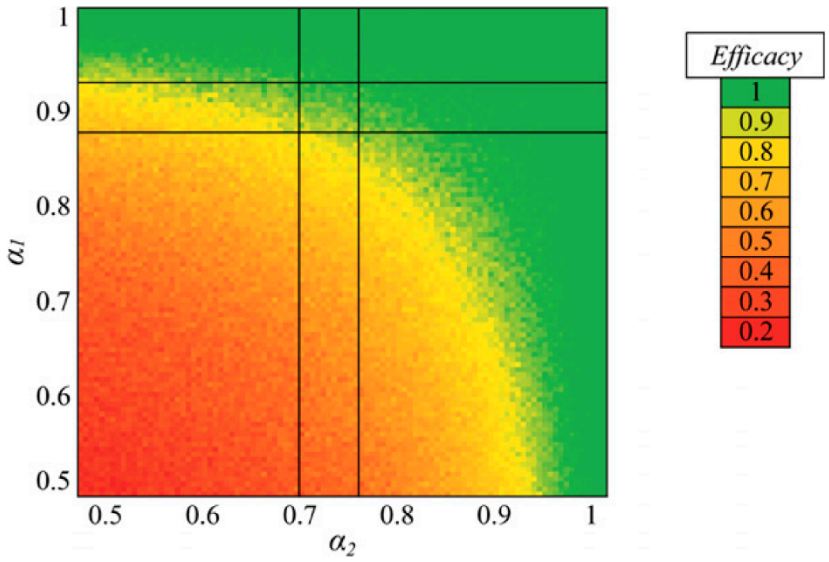


Figure 4-5 Proportional sterile protective efficacy of combinations of vaccines, with colour scale indicating proportional sterile efficacy from green (total protection, 100% sterile efficacy) to red (20% efficacy protection). Double lines indicate the potencies of RTS,S/AS01B (horizontal line, Kester, 2009) and ChAd63-MVA ME-TRAP (vertical line, Ewer, 2013) vaccines in CHMI trials: modelling their combination gives an efficacy of ~100%.

4.4 Discussion

The development of an effective vaccine against *P. falciparum* malaria is regarded to be an essential step towards eradication of the disease. In this chapter, I provide support for synergistic efficacy in combinations of vaccines currently in late stage development. Underpinning this is the requirement of vaccine-induced immune responses to elicit sterile protection: in individuals in whom all but a few parasites are eradicated, disease is still likely to emerge, albeit with a slight delay. Previous studies have estimated that two PE subunit vaccines currently in late-stage development or recently licensed can independently eradicate in excess of 90% of parasites, and yet sterile efficacies remain no greater than 50% (Kester *et al.*, 2009; Ewer *et al.*, 2013). I provide data-driven support that combinations of vaccines with such strong parasite-killing effects will be drastically more efficacious than the aforementioned, although highly promising, single-vaccine approaches.

CHMI studies remain a crucial step in the development of novel malaria vaccine candidates, providing a stringent test of efficacy and generating invaluable data. They allow for identification and presentation of candidates suitable for field studies on larger cohorts, as well as consideration of the longevity of induced protection. This work presents a within-host model of the dynamics of *P. falciparum* infection, and uses it to predict that the efficacy of a vaccine comprising a combination of the best two PE subunit vaccines to date (RTS,S/AS01 and ChAd63-MVA ME-TRAP) would demonstrate synergy, far outstripping that of its components. It also provides a framework for estimating the efficacy of combinations of known and hypothetical vaccine candidates.

The five bite challenge faced by CHMI participants is a stringent test of vaccine efficacy, with each mosquito conservatively estimated to inoculate ~11 sporozoites into the skin (Beier *et al.*, 1992). If each of these sporozoites went on to infect a liver cell and complete development then in excess of 10^6 merozoites would be released, with a single schizont shown to release around 20000 merozoites (Meis *et al.*, 1984, 1986). This estimate is well above the threshold for qPCR detection, which is on the order of 20 parasites per ml (Andrews *et al.*, 2005), extrapolated to ~10^4 parasites across the entire host. My results endorse the previously held notion that only a fraction of sporozoites successfully produce progeny, and that the full cohort of merozoites released from each liver cell may not be successful (Bejon *et al.*, 2005). Indeed, for the time to qPCR detection to be delayed beyond the period observed in non-vaccinees, even with only one infected liver cell, we would require only a fraction of its merozoite population to successfully undergo erythrocytic replication.

There are several potential explanations for this discrepancy. It has been found that merozoite egress from hepatocytes does not happen instantaneously or through cellular rupture, rather vesicles called merosomes, each containing 100-200 merozoites, bud from the infected cell (Sturm *et al.*, 2006). It is possible that some merosomes fail to leave the liver, and this may be exacerbated by known inflammatory immune responses restricting blood flow in the infected liver (Viriyavejakul, Khachonsaksumet and Punsawad, 2014). Furthermore, only a small number of the merozoites released may successfully invade red blood cells (Bejon *et al.*, 2005): this natural stochasticity may be partially responsible for the range in times to qPCR detection and patency by blood slide analysis seen in CHMI studies (Roestenberg *et al.*, 2012). In addition, it is likely that there is heterogeneity in the spread of parasites across the body. For example, release of the merosome cargo has been shown in a

murine model to occur mostly in the lungs (Baer *et al.*, 2007): parasites are released into deep vasculature, whereas blood for qPCR detection and thick smear analysis is taken from more peripheral sources.

Recent experiments have sought to more accurately measure the erythrocytic parasite replication rate by conducting malaria challenge trials using a merozoite inoculum rather than the classical mosquito-bite sporozoite challenge (Sanderson *et al.*, 2008; Duncan and Draper, 2012). Such methods allow merozoite dosage to be standardized between clinical trials. As suggested by our model, the number of sporozoites inoculated by mosquitoes, the proportion of those that successfully undergo hepatic replication, and the proportion of those released from the liver that successful enter the erythrocytic cycle are all factors that influence size of the final *effective* liver-blood inoculum. Blood stage inoculation with a fixed dose will negate the effects of all but the latter of these factors, allowing for more accurate estimation of *in vitro* merozoite dosage. The hepatocytic stages of development represent something of a "black box" in terms of our understanding of the life cycle, however blood stage inoculation experiments, combined with microscopy (Baer *et al.*, 2007), may enhance our understanding of this key stage. New estimates may be sought for the numbers of infected hepatocytes, per-hepatocyte merozoite dose, and incubation time, which would in turn allow our model to be refined.

An effective vaccine against malaria will provide durable, sterile protection against infection. This work predicts that modestly efficacious vaccines may combine synergistically to produce a higher efficacy, thus suggesting that it may be beneficial to invest in combinations of lower efficacy, perhaps higher durability single vaccines rather than their high efficacy but low durability counterparts. 13 pre-erythrocytic and

9 erythrocytic vaccines have entered clinical trials (WHO, 2014), but very few have been found to demonstrate lasting protective efficacy in humans, despite high antibody titres or T cell levels (reviewed in Schwartz, Brown, Genton, & Moorthy, 2012). The work presented here supports the notion that a highly efficacious pre-erythrocytic subunit vaccine may well be within reach; perhaps utilizing some of the tools presently available.

Chapter 5 Modelling mass malaria vaccine administration

Andrew S. Walker and Professor Lisa J. White

5.1 Introduction

There have been substantial reductions in infection prevalence, clinical disease, and mortality caused by *Plasmodium falciparum* malaria in recent decades. The Millennium Development Goals (MDG) were adopted in 2000 and provided an aim to, by 2015, reduce the incidence of malaria by increased access to antimalarial drugs and Insecticide Treatment Nets (ITNs). These efforts, teamed with interventions such as indoor residual spraying (IRS), mass drug administration campaigns, and increased surveillance, have driven an estimated 37% decline in case incidence and 60% decline in malaria mortality (Bhatt *et al.*, 2015; World Health Organization, 2015).

Despite this positive progress, sustained efforts are required in order to continue reductions in caseloads. Resistance of malarial parasites to anti-malarial drugs is commonplace in parts of South East Asia (Ashley *et al.*, 2014; Hanboonkunupakarn and White, 2016), and some reports paint a worrying picture of the potential for spread of resistance in Africa (Takala-Harrison and Laufer, 2015; Slater *et al.*, 2016; Lu *et al.*, 2017), where the *P. falciparum* parasite is highly endemic, comprising 90% of global annual malaria incidence (WHO, 2016). Between 2005 to 2015, financial investment in malaria research and treatment more than doubled, however it is estimated that annual investments will need to greatly exceed current levels to achieve the WHO's Global Technical Strategy for malaria 2016-2030 goals (Cibulskis *et al.*, 2016). There is a real risk that, as we set our sights on eradication of malaria, intervention funding may wane and render us unable to prevent a resurgence of disease. This means that a novel treatment must aim to be cost-effective as well as efficacious. That said, a single "silver bullet" against malaria may not be possible, however it may instead be likely that malaria eradication will be reached upon the careful combination of existing and

novel therapies, potentially including a prophylactic vaccine that is at least partially protective against infection.

As discussed in Chapter 4, the most advanced (in terms of clinical trial progression) candidate vaccine for widespread administration in the field is the recombinant protein construct RTS,S/AS01. This construct contains part of, and induces an immune response against, the pre-erythrocytic circumsporozoite protein (CSP). A CSP peptide is co-expressed with the hepatitis B surface antigen, and protection is afforded by a combination of cellular and T cell-mediated immune responses (White *et al.*, 2013). A wealth of clinical trials have been conducted on vaccine candidate formulations aiming to elicit responses to the CSP protein since its cloning and sequencing in the 1980s (Dame *et al.*, 1984). Most recently, a large-scale phase 3 field trial involving in excess of 6000 children and infants (RTSS Clinical Trials Partnership, 2014) found responses to induce total sterile protection in some individuals, however this protection waned to undetectable levels within a number of years (Olotu *et al.*, 2016).

A number of follow-ups of the large-scale phase 3 trial have been published, tracking the efficacy of the vaccine over time, and searching for potential "rebound" effects from vaccination (The RTSS Clinical Trials Partnership, 2011; Olotu *et al.*, 2013, 2016; RTSS Clinical Trials Partnership, 2015). A recent follow-up of the phase 3 RTS,S trial found that a three-dose vaccination regimen of RTS,S/AS01 was initially protective against infection and thus reduced clinical malaria rates, however the number of cases that were prevented was largely offset by an higher case load in later years in medium to high transmission areas (Olotu *et al.*, 2016). That said, other recent work has found

that delay of the final dose of the three-dose regimen may substantially boost immunogenicity and protection (Regules *et al.*, 2016).

Shortcomings aside, RTS,S is the leading malaria vaccine candidate, and does induce potent and protective (albeit transient) responses against malaria infection. In light of this, it has been suggested that alternative administration regimens for RTS,S should be considered (Gosling and von Seidlein, 2016). Non-cerebral clinical manifestations of malaria occur primarily in children and infants, and most trials of the RTS,S vaccine have only been conducted on those individuals. Although large-scale, the trials of RTS,S that have been conducted thus far will not have been sufficient to disrupt local transmission dynamics and have any appreciable impact on overall incidence rate. This contrasts studies of those interventions which have seen most success in reducing incidence; those that act to directly block transmission of parasites.

Systematic analysis of a number of models concluded that the cost-effectiveness of childhood vaccination with RTS,S outstrips that of current interventions (Penny *et al.*, 2016). We might therefore expect that mass vaccination, which would afford protection to vaccine recipients as well as reduce incidence and thus transmission, may be highly cost-effective and drive elimination, particularly in combination with existing interventions.

Campaigns of Mass Drug Administration (MDAs) formed a key component of early malaria eradication programs, however interest in them as part of campaigns to reduce transmission and drive eradication has waned in light of drug resistance and concerns about efficacy (Newby *et al.*, 2015). The short-term action of the drugs means that, for a campaign to be successful, it must be sustained for a long period. Shorter

interventions in endemic regions are unlikely to be successful alone, with a recent study suggesting that malaria incidence rates return to pre-intervention levels within 4 weeks of an intensive administration campaign (Aregawi *et al.*, 2016). If primaquine is administered as part of the regimen then it may have some impact on transmission, but this is due to its gametocytocidal properties rather than population-level effects of a prevalence reduction (Dicko *et al.*, 2016).

Substantial reductions in malaria prevalence have also been borne of widespread use of Insecticide Treated Nets (ITNs). The proportion of the at-risk population in sub-Saharan Africa with household access to an ITN increased from 7% in 2005 to 67% in 2015, with an estimated 82% of those with access to nets frequently sleeping under them (Cibulskis *et al.*, 2016).

Mathematical models are routinely utilized to unpick the mechanistic determinants of infectious disease dynamics (Grassly and Fraser, 2008). They allow for intervention efforts to be trialled *in silico*, and comprise a key step in healthcare policy decision making (Metcalf, Edmunds and Lessler, 2015). With reference to malaria, almost every stage of the infection cycle has been subject to modelling (Mandal, Sarkar and Sinha, 2011): from complex models of vector dynamics to understand transmission (Traoré, Sangaré and Traoré, 2017), to modelling of immune responses and parasite dynamics within a single host (Recker *et al.*, 2004; De Leenheer and Pilyugin, 2008; Walker *et al.*, 2015), to epidemiological modelling of transmission chains (Liu *et al.*, 2012). Our approach here uses data from vaccine field trials to predict outcomes of interventions on an epidemiological scale. Our framework allows us to make predictions of seasonal force of infection and vaccine efficacy and durability, and then to scale up those predictions to evaluate the impact of adding an RTS,S mass

vaccination campaign to existing mass drug administration and transmission blocking (ITNs and IRS) campaigns.

There have been a wealth of studies into the magnitude and durability of protection afforded by the RTS,S vaccine, and we here propose a means by which these studies can be utilized to predict the potential population level impact of mass vaccine administration with a moderate but transiently efficacious vaccine. Our approach allows us to scale up a mechanistic model structure from single and within-host estimates of vaccine protection and durability (as generated by field trials) to make between-host, population-level predictions of the potential impact of mass vaccination. As described here, and in Chapter 4, it is likely that an effective malaria elimination campaign will comprise carefully applied combinations of new and existing interventions. We therefore provide a full exploration of possible implementation strategies of existing and hypothetical interventions in order to predict those combinatorial strategies that could have the greatest clinical impact in endemic regions.

5.2 Methods

5.2.1 Data/study design

Extensive data provided from a number of field trials of the RTS,S vaccine, provided under licence by GSK, affords an excellent opportunity to fit models of pathogen dynamics on an individual level. For each individual within the trial we have a series of time-course infection data. Our primary model fitting is conducted by analysis of data from a field trial (Bejon *et al.*, 2008) of RTS,S/AS01E in Kilifi, Kenya and Korogwe, Tanzania (ClinicalTrials.gov number NCT00380393). The study design for this trial is described in full by Bejon *et al.* 2008, and is summarised below.

At the trial onset, 1162 were assessed for eligibility, of which 809 underwent the full 'according-to-protocol' analysis. Participants were assigned to receive the RTS,S/AS01E vaccine (n=447, three doses at intervals of between 20 to 60 days), or a rabies control vaccine (n=447, three doses at intervals of between 20 to 60 days). Monitoring for clinical malaria was conducted by both active and passive case detection. For active detection, participants were visited each week by fieldworkers who measured auxillary temperature and asked parents if any fever was seen in the week previous. If an elevated temperature was found, a blood film was analysed and rapid detection test performed. If the parent or carer reported fever (but no such fever was measured) the fieldworker returned to the participant within 6-12 hours to take their temperature again: if there was fever at the second visit then a blood film was taken and rapid test performed. Testing for malaria was also available at any time if a parent or carer thought that a child was febrile, with fieldworkers stationed in study villages to allow rapid access.

This trial design, particularly the active detection of cases, provides us with a rare opportunity to study clinical malaria rates in young children in two transmission settings.

5.2.2 Individual-based survival model (IBSM)

For each individual in the trial, time-course data describes the participant's infection status; a series of time points at which data was collected, and a quantitative measure of parasitemia at each of those time points given. Our individual-based survival modelling (IBSM) approach allows us to use this data to estimate the force of infection placed upon an individual over each season, and the impact of vaccination on the infection history of an individual.

We begin by constructing a model of only the control data from the vaccine trials. In the first model we make the simplifying assumption that the force of infection does not vary over the season. This assumption is obviously an oversimplification; there is irrefutable evidence that malaria dynamics are shaped by vector dynamics, which are inextricably linked to seasonal patterns. This assumption, however, allows us to construct and test a basic model framework. By then incrementally increasing the complexity of the model we can select a model structure that represents the data most parsimoniously.

The infection data for each trial participant is simplified to a series of binary infection states (0s and 1s), and so in the time between each sample being taken an individual can transition between each possible infection status. In the malaria field trials, standard protocol is that any detected infection is treated, and so, if an individual has infection at two sequential time-points, we can assume that the second infection

comprises a distinct infection event from the first. We also assume that an infection that occurs in between time sampling points will persist until the patient is sampled and thus treated, or presents at a clinic with symptoms and is treated. Given the considerable longevity of most malaria infections and the frequency with which sampling occurs this is not an unrealistic assumption to make.

The transitions between each of an individual's disease states can thus be used to calculate a probability of infection or, conversely, the probability that they have remained uninfected between each sampling point. There are four possible transitions, going to uninfected to infected (0→1), from infected to uninfected (1→0), remaining uninfected (0→0), and remaining infected (1→1). As stated, all infections are rapidly cleared by drug treatment, and so we can, all else remaining the same, assume that the probability of transition from 1→0 is equal to transition from 0→0. Similarly, the probability of infection given previous infection is, all else the same, the same as the probability of infection given no infection in the previous sample (0→1 = 1→1).

The probability of infection in the time since last sampling (p_{inf}) can be defined as the cumulative of the Poisson distribution:

$$p_{inf} = (1 - e^{-\lambda dt})$$

where λ is the force of infection, and dt is the time since last sampling. Conversely the probability of not being infected given the last sample time is

$$p_{inf} = e^{-\lambda dt}$$

These equations can subsequently be used to generate a log likelihood of the infection status that has occurred in a single individual given a value of λ. The sum of these log likelihoods gives an overall likelihood that the cohort as a whole would experience that infection history given the force of infection input into the equation. Note the use of log likelihood: if the likelihoods were not logged then the combined likelihood would require the values to be multiplied, resulting in very small numbers and, inevitably, computational errors.

An overall equation for log likelihood (LL) can thus be expressed as

$$LL = -\sum_{i=1}^{n} \begin{pmatrix} -\log((1-I_t)\,(1-I_{t-1})\left(e^{-\lambda dt}\right) + \\ I_t(1-I_{t-1})\left(1-e^{-\lambda dt}\right) + \\ I_{t-1}I_t\left(1-e^{-\lambda dt}\right) + \\ I_{t-1}(1-I_t)\left(e^{-\lambda dt}\right)) \end{pmatrix}$$

where n is the number of individuals in the cohort, I_t is the infection state at current time, and I_{t-1} is the infection state at the previous time. Note that this equation is structured such that, of the four summands, and for a given state transition (of the four possible), only one will be non-zero. E.g., if an individual remains uninfected, $I_t = I_{t-1} = 0$, and so only the first summand $(1-I_t)(1-I_{t-1})\left(1-e^{-\lambda dt}\right)$ is not equal to 0.

This model structure is amenable to the addition of further complexity by modification, namely time dependence, of the λ (force of infection) parameter. In this, the simplest version of the model, fitting regimes (described below) can attempt to fit

a single parameter value to λ. One benefit of this approach is that, as no ODE systems need to be solved or complex routines implemented, millions of runs of the model can be completed very quickly.

5.2.2.1 IBSM – force of infection

By making λ a function of time we can incorporate seasonality into force of infection. In the first instance, seasonality will be assumed to follow a sigmoidal function, described by the equation:

$$\lambda = M * \left(A * sin\left(f * \pi * (t - \varphi)\right) + 1\right)$$

where M is the magnitude of force of infection, A is the amplitude, f is the periodicity of seasonality (fixed at 6 months so that each cycle occurs with regularity over each year), and φ is the phase of seasonality (i.e., shifting the peaks of FOI).

It may become apparent from data on incidence of infections that seasonality has a strong impact upon force of infection, and that there are multiple peaks of force of infection per season. Previous epidemiological works have noted this pattern (Roca-Feltrer *et al.*, 2009; Cairns *et al.*, 2015): peaks of infection can occur either side of the rainy season, as in those periods of moderate rainfall are the optimum conditions for vector spread.

A number of methods have been considered for allowing each peak of FOI in a given season to vary, and there to be multiple peaks per season. The first was to increase periodicity in equation 1, and consider either the magnitude or amplitude of the FOI

function to be a step function: at each peak or trough one of these factors can change. The main problem with this approach is that, as shown in Figure 5-1, it causes there to be potentially sizeable jumps in force of infection at the troughs, something that is not biologically plausible and that may introduce unpredictable behaviours into the model.

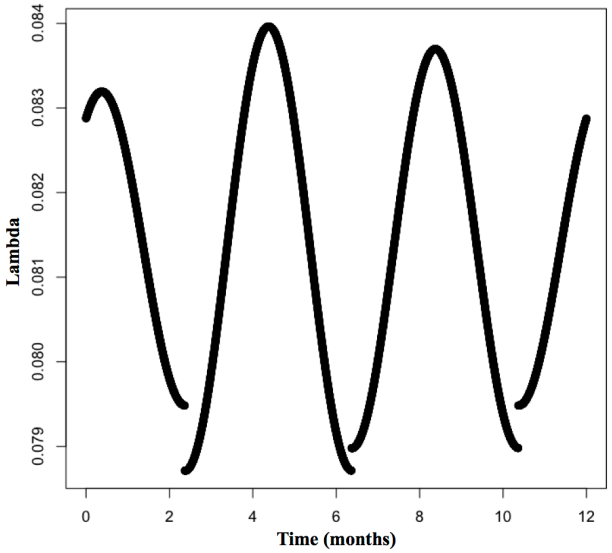
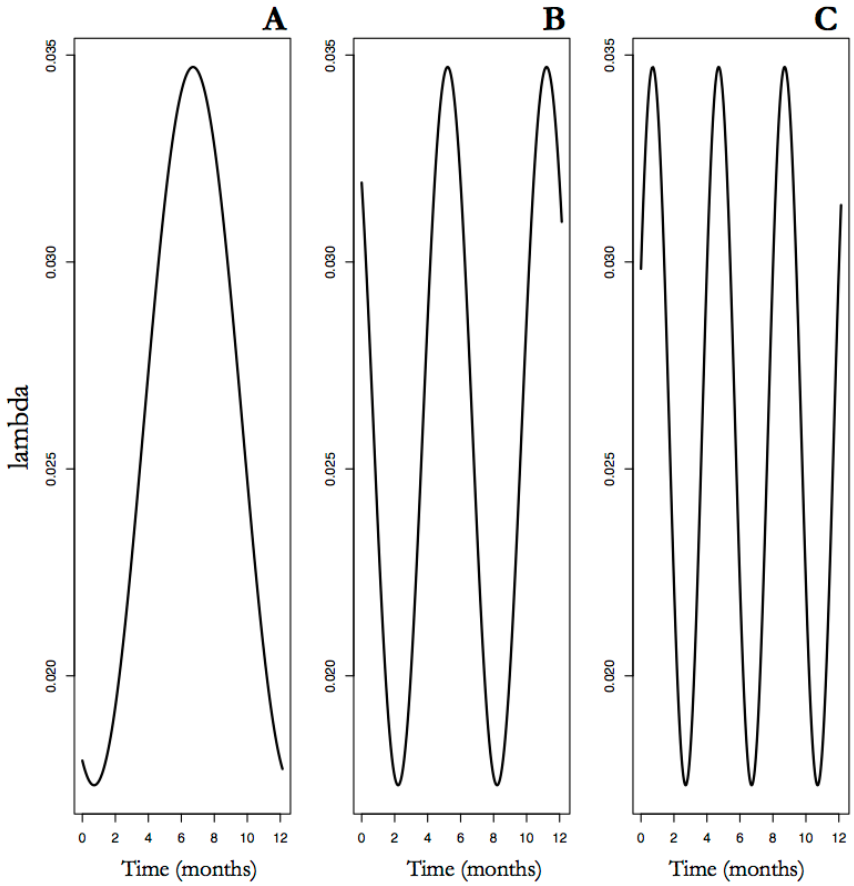


Figure 5-1 Force of Infection calculated using a simple step function for amplitude, thus causing λ to be discontinuous



A_1	1	0	0
A_2	0	1	0
A_3	0	0	1

Figure 5-2 Sinusoidal functions generated by the function in equation X where only one of the amplitude components (A_x) is not zero. **A** shows a sinusoidal function with 12 month periodicity (A_1=1, A_2=0, A_3=0), **B** shows a function with 6 month periodicity (A_1=0, A_2=1, A_3=0), and **C** shows a function with 4 month periodicity (A_1=0, A_2=0, A_3=1).

An approach devised here allows force of infection to deviate from a single sinusoidal function of fixed periodicity. Instead of a single sinusoidal curve, we consider FOI to be the sum of a number of sinusoidal functions of differing, but interlocking, periodicities (Figures 5-2 and 5-3). The first summand (λ_A) exhibits the lowest frequency possible, with one peak and one trough every 12 months. The output of this function, herein termed λ_A, is added to second function with a frequency of exactly double that of λ_A, termed λ_B, and a third sinusoidal function with a frequency of three times λ_A, termed λ_C (Figure 5-3). λ is thus the sum of λ_A, λ_B, and λ_C.

In the below equation, A_1, A_2, and A_3 are the amplitudes of each sinusoidal curve, thus they relate the absolute contributions of each sinusoidal curve to the overall FOI. M is the overall magnitude of FOI, t is the time in months, φ is the phase, and f is the cycle frequency (fixed at 1/6 to give a minimum of exactly one peak and one trough per annum).

$$\lambda = M * \begin{bmatrix} \big(A_1 * sin\big(f * \pi * (t-\varphi)\big) + 1\big) + \\ \big(A_2 * sin\big(2 * f * \pi * (t-\varphi)\big) + 1\big) + \\ \big(A_3 * sin\big(3 * f * \pi * (t-\varphi)\big) + 1\big) \end{bmatrix} \begin{matrix} \lambda_A \\ \lambda_B \\ \lambda_C \end{matrix}$$

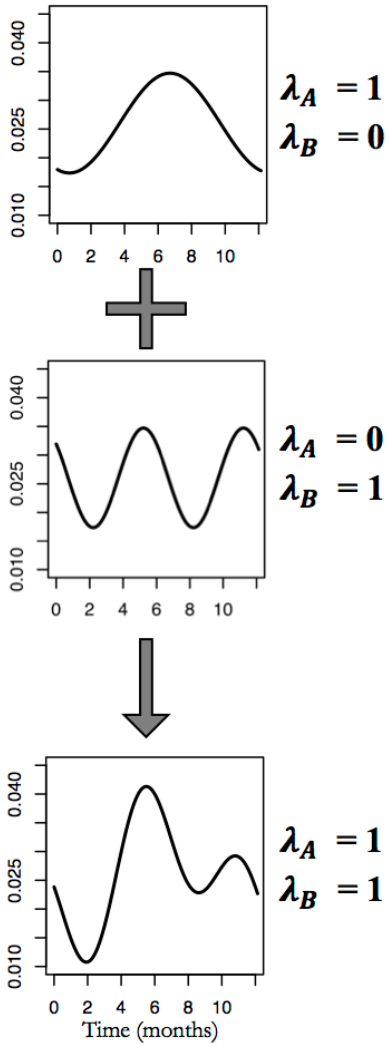


Figure 5-3 A complex sinusoidal curve (λ_A = 1, λ_B = 1; bottom) generated by summing the outputs of simple sinusoidal curves with periodicities of 12 months (λ_A = 1, λ_B = 0; top) and 6 months (λ_A = 0, λ_B = 1; middle).

Varying the absolute contributions of each λ_X thus dictates the complex periodicity of the force of infection, as demonstrated in Figure 5-4. In Figure 5-4, magnitude and phase are set to arbitrary values and the impact of varying the contributions of λ_A and λ_B to λ tested. Note in these examples that negative values (>-1) for each λ_X are also tested – by comparing the λ generated from $\{A_1 = -1; A_2 = 0\}$ and $\{A_1 = 1; A_2 = 0\}$, we can see that negative values of A_x cause the periodicity of the sine wave to be offset by

exactly 6 months. In other words, there are multiple combinations of parameters that generate identical patterns of Force of Infection. This scenario, however, does not hold true in most regions of parameter space. When there are multiple peaks of infection per annum (i.e., when $A_2 \neq 0$), and when those peaks vary in height (i.e., when $A_1 \neq 0$), this identifiability issue resolves. For example, the distribution of λ produced when $\{A_1 = -1; A_2 = 0.5\}$ cannot be reproduced by shifting the phase of the λ distrubution from $\{A_1 = -1; A_2 = -0.5\}$.

Incorporating non-zero values for A_3 into the model allows for a further level of complexity in the sinusoidal function to be achieved; there can be 3 peaks of FOI per annum (Figure 5-5). This additional level of complexity may lead to overfitting; however, we should consider that such complex dynamics may be possible in natural systems. We will take care to interpret the biological relevance of any findings indicative of such patterns.

This method allows us to a generate continuous, complex sinusoidal function of λ which can, but does not necessarily (depending on the values of A_x), exhibit multiple peaks of infection per season.

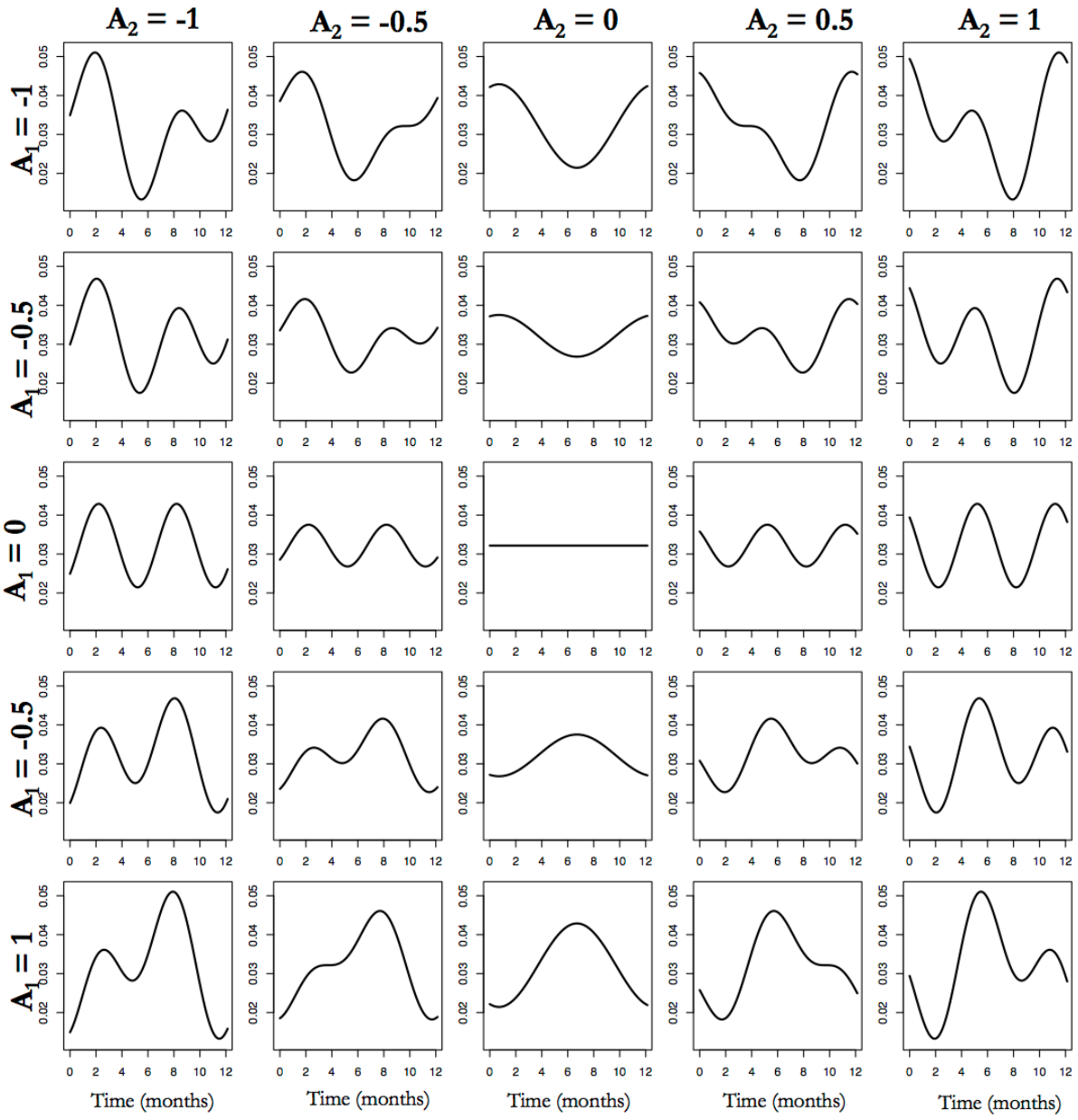


Figure 5-4 Overall λ values generated by combinations of sinusoidal curves of interlocking periodicities: non-zero values of A_1 and A_2 allow the λ_A and λ_B components to contribute to the overall λ value.

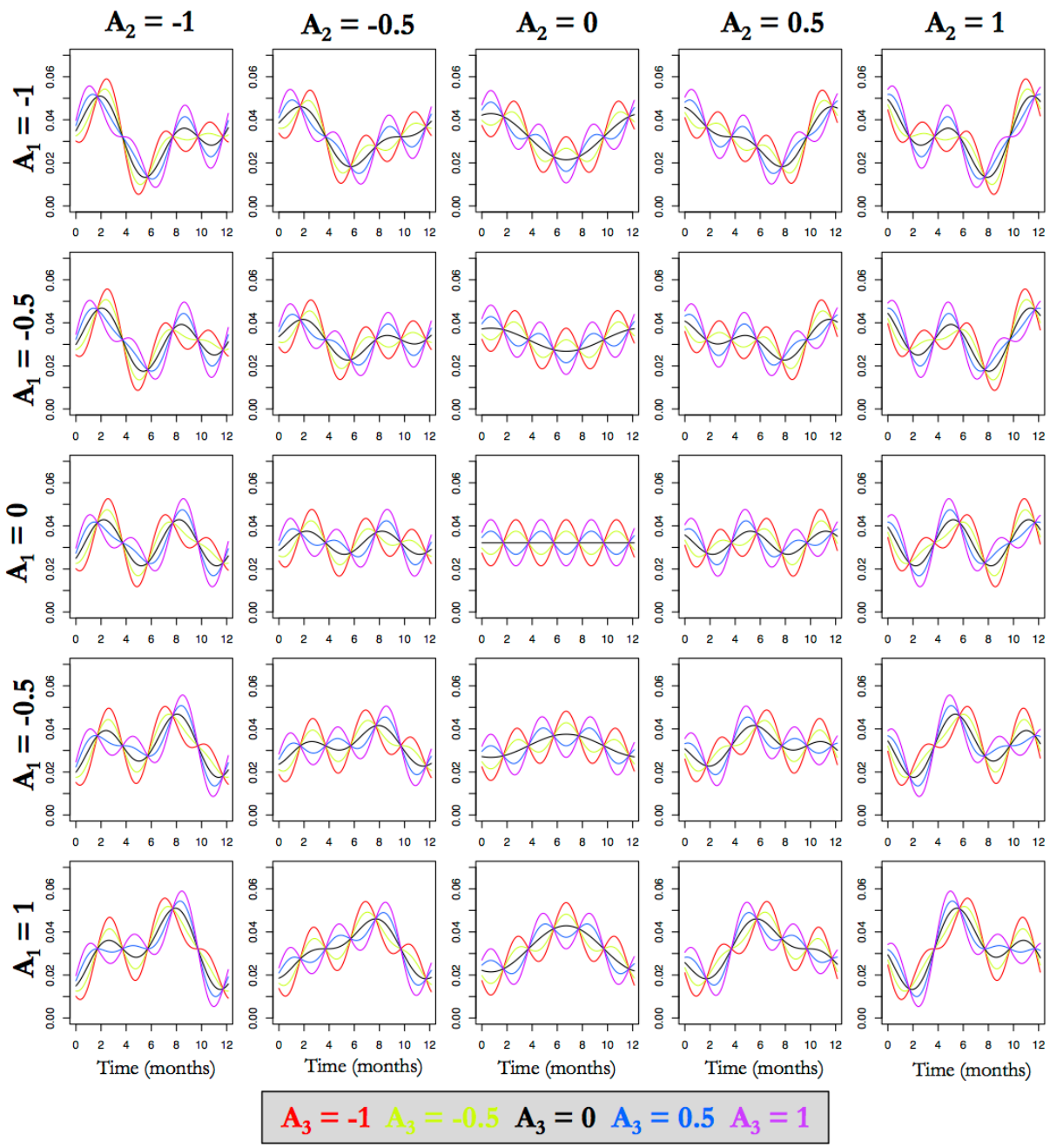


Figure 5-5 Generating complex functions for force of infection comprising simple sinusoidal functions of interlocking periodicities.

5.2.2.2 IBSM – modelling vaccination

The next step in our approach is to modify the individual based survival analysis model to incorporate vaccination. We assume here that vaccine efficacy acts independently of age and exposure. This is a reasonable assumption to make, as exposure is not thought to profoundly impact rates of infection, only rates of clinical presentation and severe disease. Vaccine efficacy is assumed to decay exponentially from a baseline level of protection to 0, as described by the following equation

$$\rho(\tau) = \rho_{base} * \mathrm{e}^{-\frac{\tau}{\rho_{dur}}}$$

where $\rho(\tau)$ is the efficacy of the vaccine at time τ since vaccination, ρ_{base} is the baseline threshold efficacy, ρ_{dur} is the rate of waning efficacy, and τ is the time since vaccination.

The force of infection incorporating vaccination (λ_V) placed upon an individual at a given point in time is thus

$$\lambda_V(t) = (1 - \rho(\tau)) * \lambda(t)$$

Independently varying the parameters (ρ_{base} and ρ_{dur}) that govern the profile of vaccine efficacy produces the profiles seen in Figure 6. With baseline efficacy fixed at 1, we explore the effect of varying the baseline level and duration of protection (Figure 6, denoted by line style and colour, respectively). We can see that as ρ_{dur} tends to 0, $\rho(\tau)$ tends to 0, and as ρ_{dur} tends to ∞, $\rho(\tau)$ tends to ρ_{base}. At low values of ρ_{dur}, i.e.,

ρ_{loss} = 1, the loss of efficacy is almost instantaneous, with more than 90% loss of efficacy within ~10 days.

Previous modelling approaches often make the simplifying assumption that all individuals are vaccinated simultaneously. In medium- to large-scale field trials this is impossible, and the modelling approach we present here allows us to take this into account: trial participants are vaccinated at different points in the transmission cycle, and so estimates of vaccine efficacy and longevity can be influenced by study design (i.e., when in the transmission cycle vaccines are administered), and study size/length. Our approach here uses precise data on when vaccines were administered and parasitemia quantified to produce estimates of efficacy that are robust to transmission setting and study design.

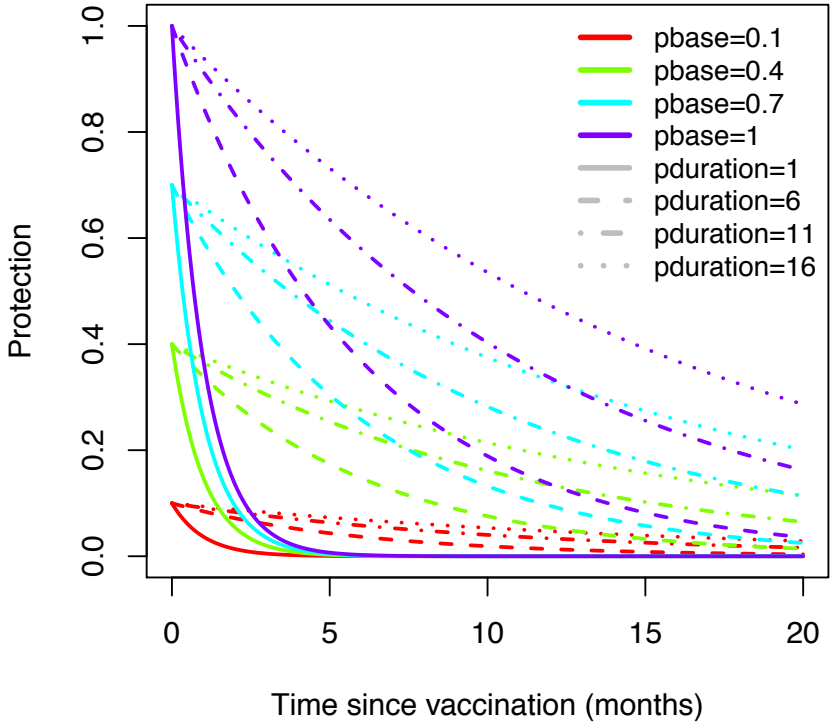


Figure 5-6 The degree of protection afforded by vaccines of various baseline efficacies (ρ_{base}) and durations of protection (ρ_{dur}).

5.2.2.3 IBSM – parameter estimation

Given the function of force of infection (λ) described previously, the following parameters will be estimated:

- M; the overall magnitude of λ, this allows for the overall FOI to be scaled
- A_1, A_2, and A_3; the amplitude of each sinusoidal component of λ, these parameters allow for the contributions of each of the periodicity of sinusoidal curves to FOI to be modified
- φ; the phase of λ, this allows the timing of the peaks of infection to be modified
- ρ_{base}; the baseline efficacy of vaccination
- ρ_{dur}; the duration of vaccine protection

An MCMC procedure (as described in Chapter Four) is used to produce parameter estimates for the model that allow the model to fit data collected from vaccine trials. Parameters that define seasonality are fit to the data from control individuals for each cohort in turn. In order to avoid overfitting of the model, and to allow for selection of the most parsimonious representation of seasonal dynamics, increasingly complex discrete fitting routines are performed for each cohort.

As demonstrated by Figure 5-5, there are multiple values of the amplitude parameters (A_x) which can produce qualitatively similar profiles for force of infection. As such, a routine is designed which allows for a number of models to be selected on the basis of both goodness of fit and parsimony – this prevents the inclusion of parameters which do not significantly improve the quality of fit, thus allowing the model to provide a parsimonious and biologically relevant representation of the data.

First, seasonality for each cohort is allowed to be simply a function of the magnitude (M) and phase (φ) of force of infection: the values of A_1, A_2, and A_3 are fixed to 0. Seven further discrete MCMC fitting procedures are then performed, testing each combination of the A_x parameters. Where A_x=0 indicates parameter A_x fixed at 0, and $\mathbf{A_x>0}$ indicates that parameter A_x is fit by an MCMC procedure, the 8 discrete routines are:

- A_1=0, A_2=0, A_3=0;
- $\mathbf{A_1>0}$, A_2=0, A_3=0;
- A_1=0, $\mathbf{A_2>0}$, A_3=0;
- A_1=0, A_2=0, $\mathbf{A_3>0}$;
- $\mathbf{A_1>0}$, $\mathbf{A_2>0}$, A_3=0;
- $\mathbf{A_1>0}$, A_2=0, $\mathbf{A_3>0}$;
- A_1=0, $\mathbf{A_2>0}$, $\mathbf{A_3>0}$;
- $\mathbf{A_1>0, A_2>0, A_3>0.}$

This protocol allows us to compare the fit of each of model variant in either cohort, test the relative contributions of each parameter to goodness of fit of the model, and identify the most parsimonious model of seasonal malaria transmission.

Strict maximum and minimum values are set for each of the parameter values in the MCMC procedure. Any parameter values selected that fall outside of these bounds are rejected, and another parameter value selected from the original distribution. For the magnitude of λ (M), the lower bound is set at 10^{-8}, and the upper bound set at 10: a non-zero lower bound prevents λ reaching 0, as in this biologically implausible region of parameter space changing the other parameters will have no effect on the goodness

of fit of the model. For the amplitudes of the three sinusoidal components, bounds are set at -1 and 1, as values outside of this range can result in a negative value of λ. The phase of λ (φ) is effectively unbounded: a value of 0 for φ (measured in months) produces a sinusoidal λ equivalent to that produced by a φ of 12, thus values sampled by the MCMC routine of $\varphi = 0 - x$ (where $x > 0$) are set to $\varphi < 12 - x$, and *vice versa*.

After estimating parameter values that describe seasonal patterns of force of infection in either cohort, we can then fix those parameters, and find the parameter pair that describe best vaccine efficacy; the baseline efficacy (ρ_{base}) and duration of protection (ρ_{dur}). As only two parameters are to be estimated, we employ a high-resolution parameter space sampling method to identify a best fitting parameter pair. This absolves any potential identifiability issues, and reduces computational load. Combinations of parameter values are sampled for ρ_{dur} between 0 and 300 (with 0 corresponding to instant loss of protection, and 300 to >80% of the initial level of protection for at least 6 years) and ρ_{base} between 0 and 1 (with 0 corresponding to no protection at time of vaccination, and 1 to total protection from infection) are tested. By analysing the goodness of fit of the model produced by each possible parameter pair, we can identify regions of parameter space in which the fit of the model is optimized.

5.2.3 Epidemiological transmission model

5.2.3.1 Transmission model structure

Next, we utilize an ODE model to produce projections of the impact on transmission of a malaria vaccine of the efficacy estimated previously, and the combined efficacy of ITN use, MDA campaigns, and mass vaccine administration. The ODE model we use is based on that described by Aguas et al., 2008, where it was utilized to test the age distribution of clinical disease in light of transmission dynamics. Here we explore transmission dynamics using a version of the model that excludes any impact of age on force of infection and incorporates an extra parameter to allow for MDA (ε). The model comprises four compartments, as described by the following set of ODEs:

$$\frac{dS}{dt} = \alpha R - \lambda S(I_C + I_A) + \mu - \mu S$$

$$\frac{dI_C}{dt} = \lambda S(I_C + I_A) - I_C(\varepsilon + \tau_C) - \mu I_C$$

$$\frac{dI_A}{dt} = \lambda R(I_C + I_A) - I_A(\varepsilon + \tau_A) - \mu I_A$$

$$\frac{dR}{dt} = I_A(\varepsilon + \tau_A) + I_C(\varepsilon + \tau_C) - \alpha R - \lambda R(I_C + I_A) - \mu R$$

where the four equations describe the changing number of susceptible (S), infected (clinical; I_C), infected (asymptomatic; I_A), and recovered (R) individuals.

In this model, individuals who have never had malaria infection are considered to be susceptible, and to progress to clinical malaria at a rate dependant on a seasonally fluctuating force of infection (λ, as described previously) and the number of infected individuals (both clinical and asymptomatic). Clinical and asymptomatic individuals

are assumed to contribute equally to transmission. Individuals are born into the susceptible compartment; maternal antibodies are assumed not to impact infection. Individuals can move from the recovered to the susceptible compartment as immunity wanes with a rate α. Those individuals whose immunity wanes are assumed to have equal susceptibility to individuals who have never seen infection.

Influx to the infected compartments is a function of force of infection, the number of infected individuals, and the number of susceptible (into the I_C compartment) or recovered (into the I_A compartment) individuals. Efflux from the infected compartments is at a rate proportion to the number infected, and the sum of the rate of recovery from clinical or asymptomatic infection (τ_C and τ_A respectively) and the proportion of individuals at that time point who have received anti-malarial drugs (ε). Anti-malarial drugs are assumed to clear all infection from those individuals who receive them.

Finally, the recovery compartment comprises those individuals who clear infection from either infected compartment. Individuals leave the recovery compartment as immunity wanes (αR), as they become asymptomatically infected $\lambda R(I_C + I_A)$, or die (μR).

Deaths occur from all classes at an equal rate, given by μ; note that the sum of all compartments at any given time is 1 and the sum of the μ components is 0, thus population size remains constant. Malaria-induced mortality is assumed to be 0 – all clinically infected individuals recover naturally or are successfully treated and clear infection.

As described previously, this model allows us to test the impact of three discrete types of intervention on malaria dynamics – mass drug administration, mass vaccination, and transmission blocking interventions such as vector killing or bed-net usage. Mass drug administration is enacted, as described, through the ε parameter. When $\varepsilon > 0$, the rate of movement of individuals from I_A and I_C to R is increased, that is, MDA induces clearance of infection and movement to the recovered class independently of clinical infection.

Transmission blocking interventions and vaccines are enacted through modification of the force of infection parameter incorporating interventions, λ_I, which is defined as:

$$\lambda_I(t) = (1 - \delta) * \lambda(t) * \left(\rho_{base} * e^{-\frac{\tau}{\rho_{dur}}}\right)$$

where δ is the proportional reduction in transmission due to transmission blocking interventions, $\lambda(t)$ is seasonal force of infection at time t, ρ_{base} and ρ_{dur} are vaccine efficacy (as described in *Methods*) and τ is the time since vaccination. If no vaccine has been administered or transmission blocking intervention enacted, then $\rho_{base} = 0$ and $\delta = 0$, thus $\lambda_I(t) = \lambda(t)$.

Transmission blocking interventions are assumed to have a constant effect, to commence at the same time as vaccination, and to continue indefinitely. The proportional reduction in transmission is represented by a single parameter, δ, however this can be adjudged to incorporate a number of factors, including the

coverage, uptake, and effectiveness of interventions such as ITNs, IRS, and gametocidal drug administration.

Mass drug administration is assumed to follow a simplified protocol: administration occurs simultaneously across all individuals, and drugs act with a constant effect for one week post-administration. Two doses of anti-malarials are given in a single MDA campaign: the first dose is given along with vaccine administration and a second is given 5 weeks later, as in other recent models (Brady *et al.*, 2017).

5.2.3.2 Transmission model parameter values

Parameters in the transmission model pertaining to force of infection are set to those found by the IBSM MCMC model fitting procedure (Table 5-2). Additional model parameters in the control scenario are set according to the literature or to those found by fitting performed by Aguas *et al*, 2008 (Table 5-2). A scaling factor is applied to the magnitude of force of infection such that infection in the population model is maintained at levels comparable with natural systems. This step recognizes that the magnitude of FOI estimated by the IBSM is that excluding submicroscopic infection, predicted in some scenarios of endemic transmission to comprise the majority of cases (Bousema *et al.*, 2014; Chourasia, Raghavendra, R. M. Bhatt, *et al.*, 2017; Chourasia, Raghavendra, Rajendra M. Bhatt, *et al.*, 2017).

In order to test the impact of combinations of MDA, MVA, and intervention-blocking intervention, the epidemiological patterns predicted under each single intervention are first tested. Interventions in the field vary in their effective coverage (a product of coverage and efficacy), and so a number of scenarios are presented for each intervention that approximately correspond to strengths seen. A “typical” MDA

regimen used in previous modelling studies comprises two rounds of treatment with an interval of several weeks at an effective coverage of ~70% (Brady *et al.*, 2017). We also test an intervention at a minimum level of coverage previously predicted by theory to be capable of driving local eradication (Newby *et al.*, 2015) (80%, "medium-intensive"), and a highly intensive MDA tested in some locations (Ali *et al.*, 2017) (95%, "intensive"). Our model comprises a system of ODEs, with drug administration (i.e., clearing of infection) enacted through the ε parameter, which is a rate at which individuals are cleared of infection; that is, moved from the infected compartments to the recovered compartment). Setting this parameter such that the aforementioned effective coverages are obtained must be performed for each cohort individually (see Results).

Estimates of bed net usage from field studies where they are provided vary wildly, from 35-40% to 68% in a cohort of school-children (Snow *et al.*, 2015). Previous theoretical works have suggested that the effective coverage with a 100% efficacious ITN required to halve prevalence is areas of moderate transmission is 80% (Smith *et al.*, 2009; Griffin *et al.*, 2010). Here, we test a trio of scenarios, with the values for each representing the effective coverage (incorporating provision, usage, and effectiveness): "high" transmission blocking (80%), "intermediate" transmission blocking (50%), and "low" transmission blocking (35%). We assume that rates of bed net usage and effectiveness of their chemical constituents do not vary over the time-scales studied herein.

We further test a number of theoretical MVA campaigns before simulating campaigns of efficacy profiles modelled by the IBSM approach. As discussed previously, vaccine efficacy can be defined as the duration and baseline efficacy of responses (Figure 5-6).

We therefore test four scenarios, vaccines with protection that is; i) high initially but quickly waning, ii) intermediate initially but durable, iii) high initially and durable, iv) intermediate initially and quickly waning.

5.3 Results

5.3.1 Malaria prevalence data

Analysis of data collected from the two study sites shows that there is clear seasonal variation in the patterns of incidence of malaria infection (Figure 5-7). Of the 224 individuals in the control group of the Kilifi cohort, 78 experienced infection during the study. In this cohort, there were two clear peaks in prevalence – one from December to January after short rains, and a second more profound peak between June and August after the longer rains (Figure 5-7).

In the Korogwe cohort, prevalence is generally lower and seasonal patterns of infection less defined (Figure 5-7, bottom left panel). As in Kilifi, there are two periods of rain (Mugalavai *et al.*, 2008): low transmission is typically observed during the October-December short rains, and moderate transmission during and immediately following a long rainy season which from March to July. Qualitatively, this pattern is not well reflected by our data, perhaps due to the sparsity of data points – in total, of 223 individuals in the Korogwe control group, only 42 experienced any clinical infection.

Overall incidence of malaria was lower in the RTS,S vaccine groups than in the recipients of the rabies control vaccine in both cohorts (Figure 5-7). Amongst the Kilifi vaccine recipients, 46 of 223 experienced infection, and, of the Korogwe vaccine recipients, 22 of 224 experienced infection.

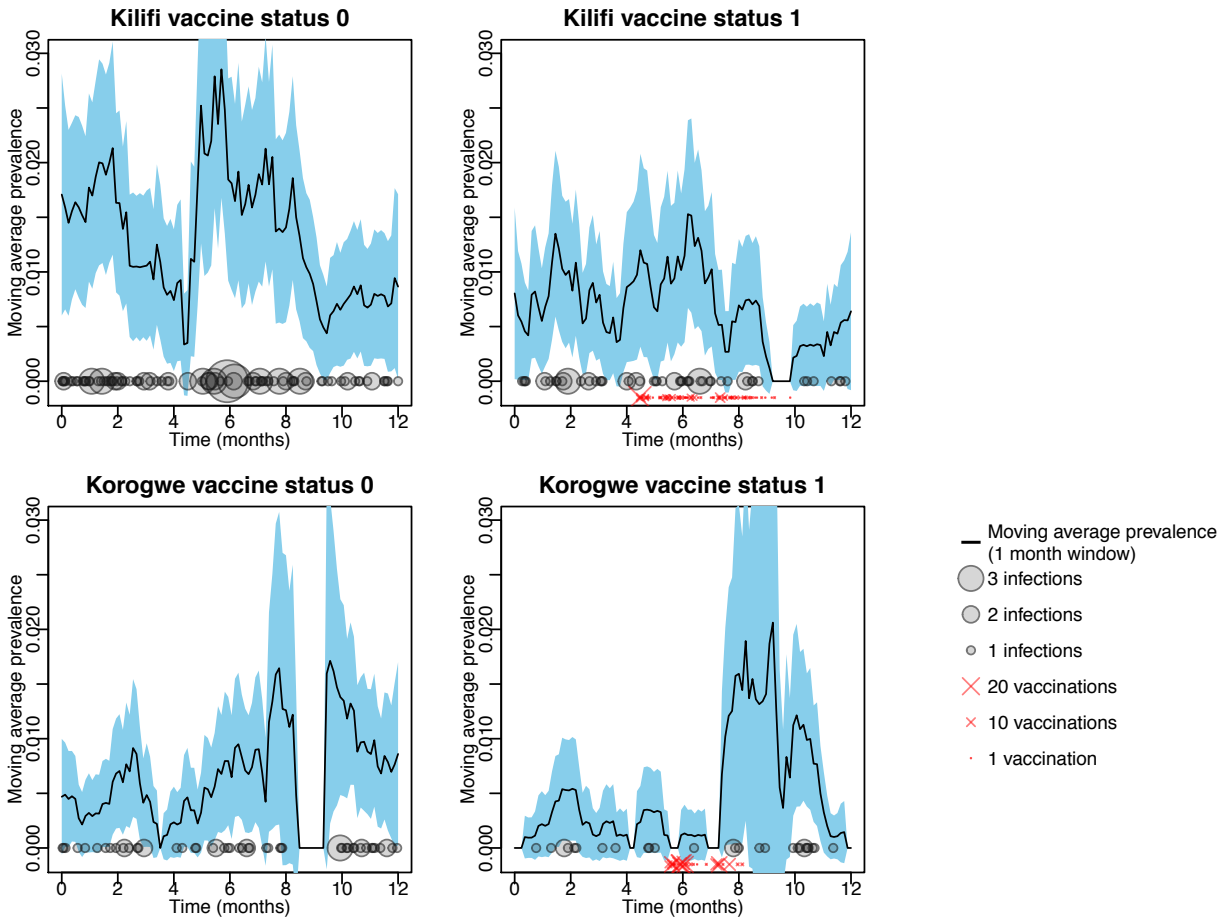


Figure 5-7 Seasonal patterns of incidence of malaria in each cohort; Kilifi (top panels) and Korogwe (bottom panels), vaccine recipients (right panels) and control individuals (left panels). Lines in each plot indicate moving average prevalences on a one month window with 95% confidence intervals indicated by blue areas. Confidence intervals are calculated by the normal approximation method, a standard method for estimating distribution of error around binomially-distributed observations. Size of the grey circles indicates the raw numbers of infections in a week-long window, and red crosses indicate the timings of vaccination (number of complete vaccination courses per week).

5.3.2 Parameterisation of the Individual Based Survival Model

5.3.2.1 Model selection by quality of fit

Here we present the results of the suite of models that were produced to fit to data from control individuals in vaccine trials conducted in Kilifi, Kenya, and Korogwe, Tanzania. The simplest model considers force of infection to not vary over time, and then we add to it components that incorporate seasonality into the model at various periodicities. As discussed in the methods, when $A_1>0$, force of infection (λ) has some 12-month periodicity; when $A_2>0$, λ has some 6-month periodicity; when $A_3>0$, λ has some 4-month periodicity. By assessing the median and minimum log-likelihoods produced by MCMC parameter fitting of each version of the model, we can identify the model that provides the most accurate and most parsimonious description of the natural disease systems in Kilifi (Figure 5-8A) and Korogwe (Figure 5-8B).

It is apparent from analysis of the model of the Kilifi cohort (Figure 5-8A) that: i) the best fitting versions of the model are those which incorporate 12-month periodicity in λ into the model ($A_1 > 0$), ii) the incorporation of a 6-month periodicity component offers little, if any improvement in the goodness of fit ($A_2 > 0$), iii) those that do corporate a 4-month periodicity component ($A_3 > 0$) seem to produce a marginally worse fit than those that do not. Given that addition of parameters beyond A_1 produce fits of equal or lesser quality, models presented herein for the Kilifi cohort will consider force of infection to vary with only a 12-month periodicity ($A_1>0$, $A_2=0$, $A_3=0$).

These findings differ from those in the Korogwe cohort (Figure 5-8B), in which the overall goodness of fit is markedly better than the Kilifi cohort. In the various models of this cohort, those that incorporate a 4-month periodicity to infection produce better

fitting models than those that do not. The 6- and 12-month periodicities seem to have little impact on goodness of fit: in fact, there is a slight trend towards better fit when those parameter values are fixed at 0 – the lowest median log-likelihood in this cohort comes from a model in which only magnitude, phase, and the contribution of the 4-month periodicity of infection (A_3) are fit. As such, models presented herein for the Korogwe cohort will consider force of infection to vary with only a 4-month periodicity ($A_1=0$, $A_2=0$, $A_3>0$).

	Kilifi		Korogwe	
	Rabies control vaccine	**RTS,S vaccine**	**Rabies control vaccine**	**RTS,S vaccine**
Participants	224	223	223	224
Participants experiencing infection	78	46	42	22
Participants experiencing clinical infection	21	12	12	6
Total number of infections	133	66	59	30
Total number of clinical infections	22	14	13	6

Table 5-1 Number of individuals with various prognoses

	Parameter	Kilifi	Korogwe	Source
Force of Infection	M, magnitude	0.01405	0.006649	MCMC fitting
	A_1, amplitude of 12-month sinusoid	-1	0	
	A_2, amplitude of 6-month sinusoid	0	0	
	A_3, amplitude of 4-month sinusoid	0	-1	
	φ, phase of function	4.7060	3.01359	
Vaccine	ρ_{base}, baseline vaccine efficacy	0.7851	0.5304	Parameter space estimation
	ρ_{dur}, duration of vaccine efficacy	14.3825	300	
Other model parameters	τ_C, recovery from asymptomatic infection	$14.12^{1/12}$		Aguas *et al.*, 2008
	τ_A, recovery from clinical infection	$2.23^{1/12}$		
	α, waning of immunity (R→S)	$1.07^{1/12}$		

Table 5-2 Parameter values used in the transmission model. All values are rates in months

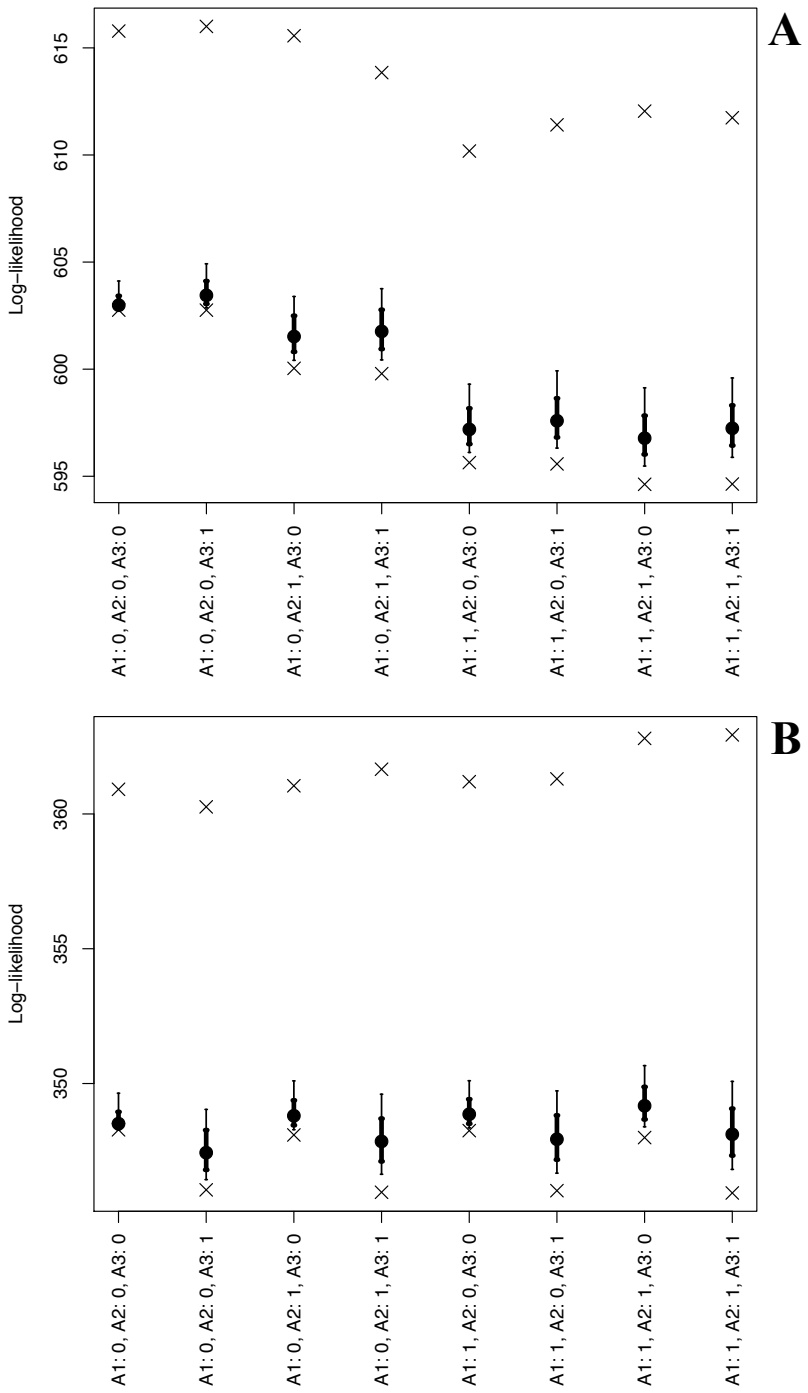


Figure 5-8 Log likelihoods describing the quality of fit of the versions of the model with and without the three periodicities (A1>0 for 12-month periodicity; A2>0 for 6 month periodicity; A3>0 for 4 month periodicity). Solid dots correspond to median log-likelihood, thick vertical bars to 1 standard deviation from the mean, and thin vertical bars to 2 standard deviations from the mean. Crosses correspond to the minimum and maximum log-likelihood for that parameter set. Quality of fit is shown for models fit independently to the Kilifi (**A**) and Korogwe (**B**) cohorts.

5.3.2.2 Estimating force of infection parameter values

Assessing the distribution of parameters selected by the MCMC procedure in the post-burn-in phase of the routine allows for parameter estimates to be generated for each cohort for the model variant that was previously selected.

In both cohorts, the distributions of magnitude and phase of force of infection selected by the MCMC routine follow approximately Gaussian distributions, indicating that parameter values have converged on an optimal solution. This is not the case, however, for the parameters that define the contribution of sinusoidality. In both optimal models (5.3.2.1) the A_x values that are fit (A_1 in the Kilifi cohort; A_3 in the Korogwe cohort) appear to converge on a value of -1 however, due to this being the lower bound allowed by the fitting procedure, the distributions comprise a heavily truncated portion of a Gaussian distribution. If this lower bound were to be relaxed, it is likely that the goodness of fit of the model would be improved. This would, however, allow negative values of force of infection to be produced: values of A_x below -1 can be therefore be deemed biologically implausible and so should not be explored. A value of A_x near the allowed bounds allows the model to take a sinusoidal shape without causing $\lambda < 0$, with the height of each peak set by the magnitude of force of infection.

The optimal parameter values in the MCMC procedure are typically defined as the median of all runs after a burn-in period has elapsed. For the A_x parameters, however, for the reasons stated above, we elect instead to select the peak of the parameter distribution. In the Kilifi cohort there is one peak value for A_1, at -1. The peaks of the normally distributed MCMC sampled offset and magnitude parameters are 4.71 and

0.014 respectively, thus these parameter values are selected for use in further analyses (Table 5-2).

In the Korogwe cohort the relevant amplitude parameter (A_3) peaks at both -1 and 1. As demonstrated by Figure 5-11, this is due to non-identifiability of A_3 and the offset parameter: there are two regions of parameter space in which the model fits the data well (as discussed later). The first is when A_3 is at -1 and the offset is at ~1.7, and the other is when A_3 is at 1 and the offset is at ~3. The sinusoidal curves produced by these two sets of parameter values are identical, and so, in interpreting the model outcomes, we must consider that one set of those parameter should be selected as representative of the data, rather than some intermediary value. Despite similar log-likelihood values, the region of parameter space that was sampled most by the MCMC routine is that in which parameters take the following values: A3 = -1, φ= 3.01359, M = 0.006649 (Table 5-2, Figure 5-9, Figure 5-10 region I), and so those parameters are accepted.

By sampling the parameter values from the MCMC chains generated during model fitting of the accepted models, I can generate force of infection curves for each cohort (with credible intervals). This demonstrate the differing patterns between the cohorts, with a single peak of infection per annum in Kilifi, and multiple (albeit lower amplitude) peaks in Korogwe (Figure 5-11).

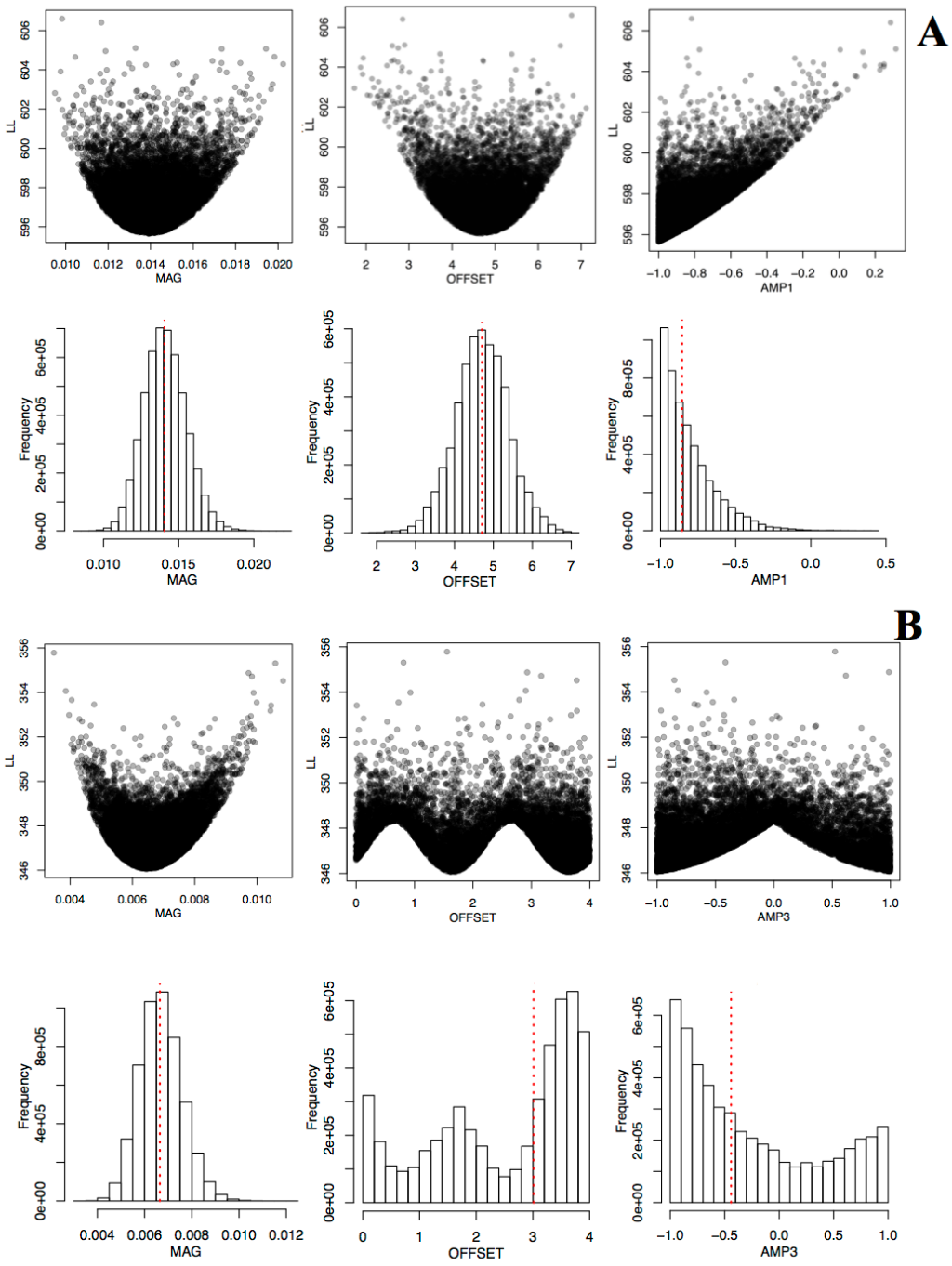


Figure 5-9 Parameter values found by MCMC procedure for FOI in Kilifi (**A**) and Korogwe (**B**). Scatter plots indicate parameter values accepted by the MCMC chain, and bar plots give frequency distributions of those values.

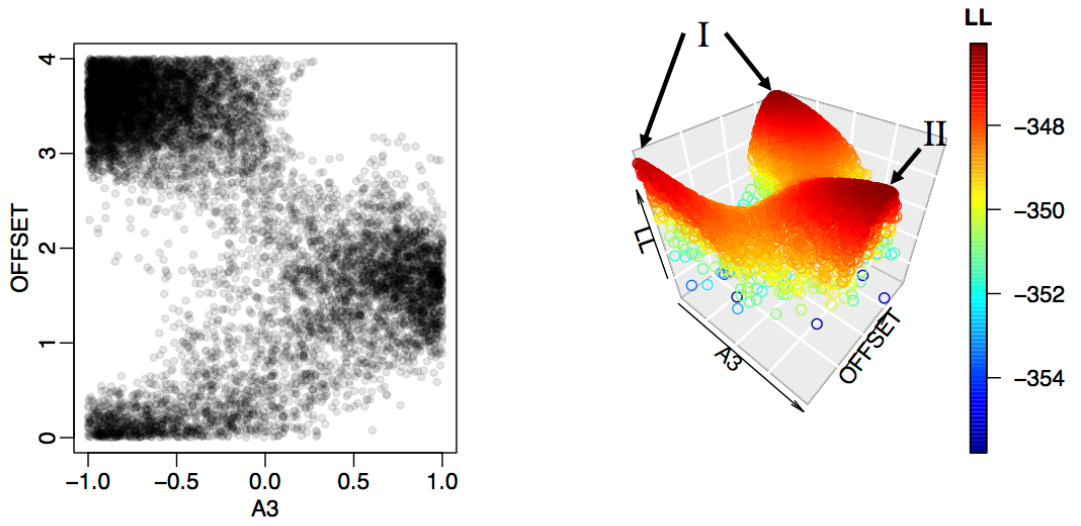


Figure 5-10 The regions of parameter space explored by the phase (OFFSET) and A_3 parameter in MCMC fitting of force of infection. There appear to be two regions (I and II, right) of parameter space that the MCMC chain converges upon, and these two regions have similarly high good quality of fit (indicated by low log likelihood, right). This serves as is evidence of identifiability issues in the model, and is the motivation for selecting one of the regions of parameter space (I, right) as producing a model that is representative of the data.

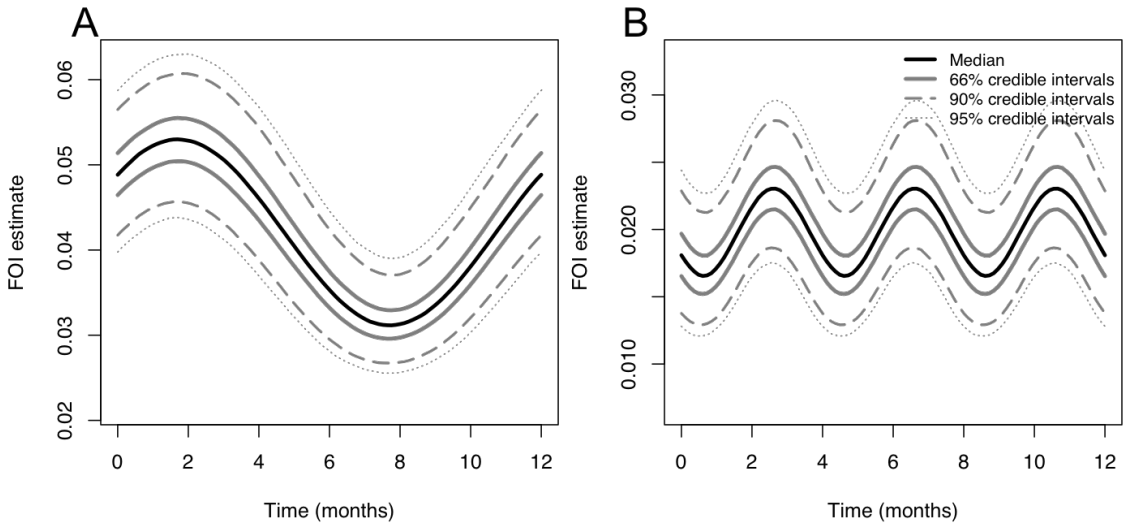


Figure 5-11 Force of infection profiles generated by random sampling of 10,000 parameters from the MCMC chains for Kilifi (A) and Korogwe (B).

5.3.2.3 Estimated vaccine efficacy parameter values

As there are only two parameters that describe the efficacy profile of the vaccine, a simple parameter space exploration was performed. The best fitting pairs of parameter values in each cohort differ, and a number of possible explanations for this are given in the *Discussion*. In the Kilifi cohort, fit of the model is optimized with a baseline efficacy (ρ_{base}) of 0.7851 and a duration of efficacy (ρ_{dur}) of 14.3825. In the Korogwe cohort the best fitting parameter pair in the allowed parameter ranges are a ρ_{base} of 0.5304 and ρ_{dur} of 300. That said, we observe that as ρ_{dur} tends to infinity, the log likelihood also tends to infinity: in the data to which the model is fit, there is no signal of waning vaccine efficacy.

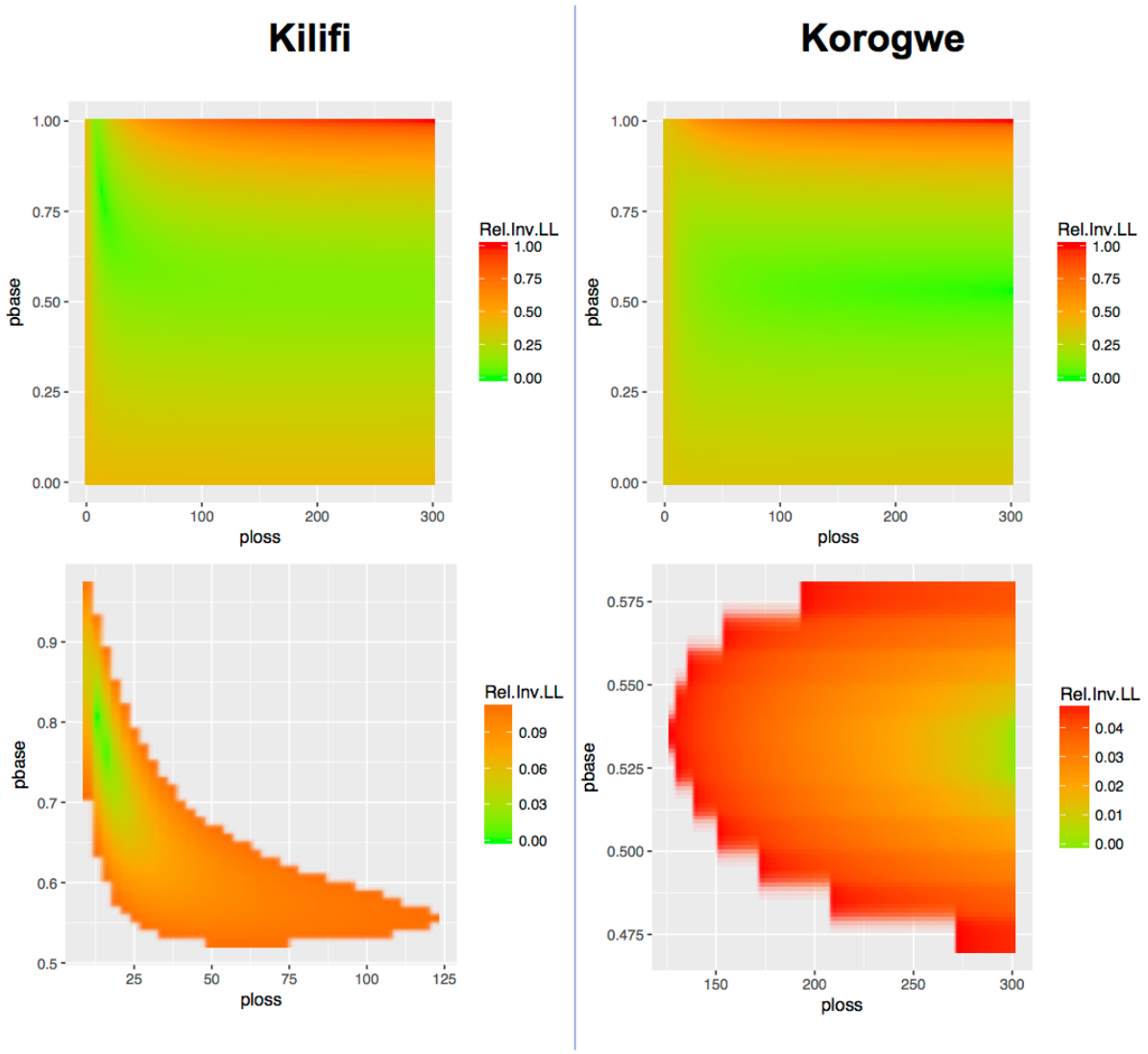


Figure 5-12 Interrogation of all regions of parameter space in Kilifi (left) and Korogwe (right). Top panels correspond to all of parameter space, bottom panels shown only the best fitting 5% of parameter pairs (N.B., there are different x and y-axis ranges in the bottom panels).

5.3.3 Simulating malaria dynamics using the transmission model

5.3.3.1 Control scenario

We next utilize an existing epidemiological model of malaria (Aguas et al., 2008) to simulation infection dynamics in a malaria population using parameters estimated in the IBSM approach and taken from the literature. The model allows us to consider various intervention scenarios, namely comparing a no intervention scenario with concurrent and individually administered Mass Drug Administration, Mass Vaccine Administration, and transmission blocking interventions such as IRS or use of ITNs.

As discussed in Methods, the model is run for an arbitrary time period to allow it to reach an equilibrium state. In the figures that follow, dynamics are presented from the end of that period. According to the model, in the control scenario in the Kilifi cohort – one in which there the interventions match those occurring in the scenarios in which the control data was taken (drug treatment only upon clinical infection, standard baseline rates of bed net usage, *etc.*) – the average proportion of the population who are infected is 0.721, fluctuating between 0.554 in the low transmission seasons, and peaking at 0.876 in the high transmission season (Table 5-2). Of those who are infected, the majority (77.3%) are asymptomatically infected. In the Korogwe cohort, the average proportion infected over the whole year is 0.411, rising from 0.321 in the low transmission season, to 0.509 in the high transmission season. According to the model, less than half of the infections were asymptomatic (48.6%).

The dynamics of infection in each cohort are shown in Figure 5-13. In each there is a clear seasonal pattern to infection, with one peak per annum in the Kilifi cohort (left), and multiple peaks per season (in line with an A3 parameter of >0) in the Korogwe cohort (right).

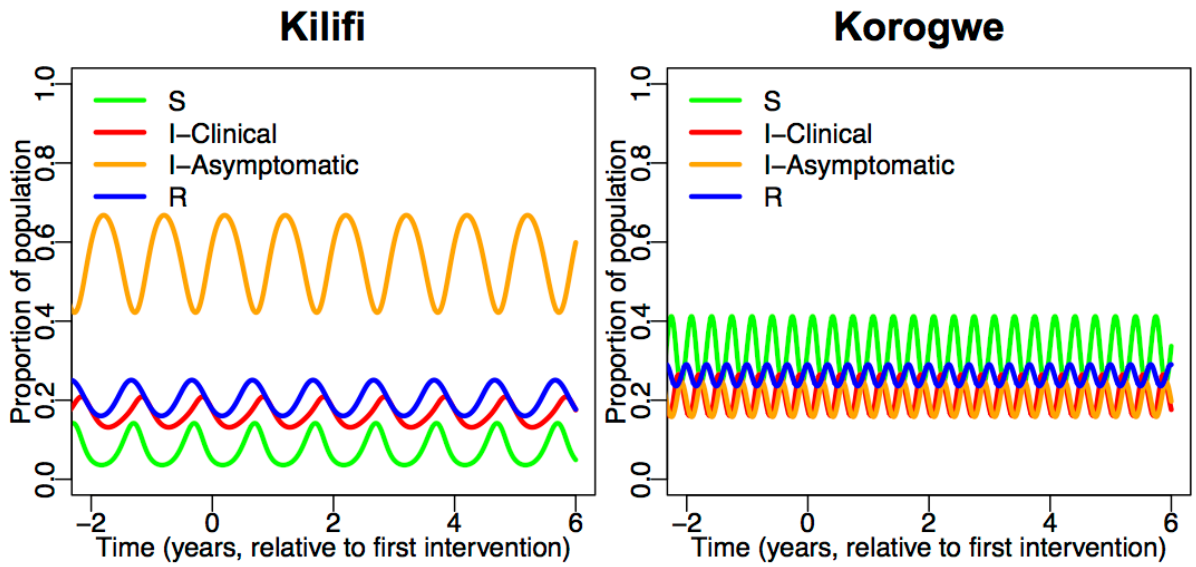


Figure 5-13 Disease dynamics in the control scenario (no intervention). In this plot, and the dynamic plots that follow, each line represents a proportion of the population who fall into each disease class (corresponding to line colour)

5.3.3.2 Disease dynamics under modelled vaccine interventions

We here utilize the parameter estimates presented previously to predict the potential impact of mass vaccination campaigns (with vaccination with a vaccine analogous to that modelled here) on disease dynamics in each cohort (Figure 5-14). In the Kilifi cohort, a vaccine analogous to that tested in the field trial would not drive local elimination of disease. There is a period of about 2 years where normal patterns are perturbed with a net decrease in disease, however normal patterns soon resume. In Korogwe, vaccination is predicted in the model to drive total local eradication of infection.

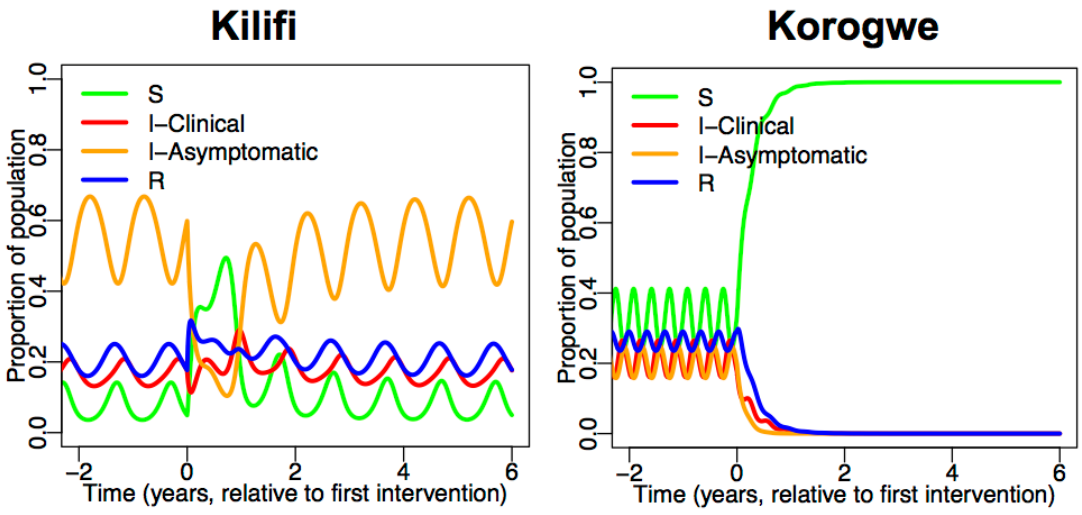
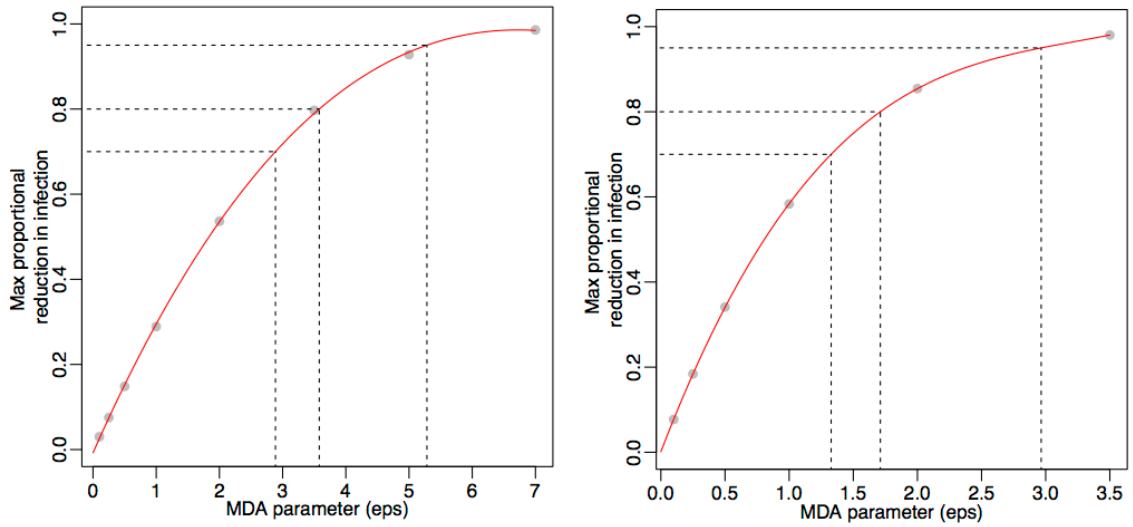


Figure 5-14 Impact of mass vaccination with vaccines of the efficacies estimated by fitting of the IBSM

5.3.3.3 Efficacy of single interventions

We can use our model framework to test the impact on disease dynamics on each of the interventions (MDA, MVA, transmission blocking) in turn, and at various effect sizes. For MDA and transmission blocking interventions we present three scenarios: "low", "medium", and "high" efficacy interventions (see *Methods*). In the transmission blocking intervention this corresponds to 25%, 50%, and 80% effective coverage respectively, thus, in the model described previously, the δ parameter is set at 0.35, 0.5, and 0.8.

When modelling MDA, these scenarios correspond to effective coverage of 70%, 80%, and 95% respectively – these scenarios assume a peak reduction in case load that corresponds to those values. In order to find parameter values that cause this reduction in infection, the model was run for a range of parameter values between 0.1 and 7. Each run was compared to the control scenario to compute the peak proportional case reduction. In each case, the timing of the peak in case reduction coincided with administration of anti-malarial drugs. A third-order polynomial fit of the relationship between the MDA parameter and peak reduction in case load was then used to estimate the parameter values that corresponded to each of the aforementioned scenarios (Figure 5-15).



	Kilifi	Korogwe
95% reduction	5.28308	2.96492
80% reduction	3.57931	1.71111
70% reduction	2.88644	1.32734
$red_{max} \approx$	$-0.0130 + 0.352\varepsilon - 0.0406\varepsilon^2 + 0.00156\varepsilon^3$	$0.00119 + 0.786\varepsilon - 0.226\varepsilon^2 + 0.0234\varepsilon^3$

Figure 5-15 Calculating the correct MDA parameter (ε) for each intervention scenario

In the MVA, four hypothetical scenarios are presented (in addition to the modelled scenarios presented previously, Figure 5-14) that comprise every combination of high and low efficacy, and very high and low duration vaccines. High and low efficacy correspond to p_{base} values of 0.8 and 0.3, and high and low duration correspond to p_{dur} values of 50 and 2, respectively.

Model outputs for each of the four MVA intervention scenarios are given in Figure 5-16. According to our model, in the Kilifi cohort, none of the MVA scenarios tested are sufficient to drive the model to a stable, disease-free equilibrium for 5 years. A high efficacy vaccine with long-lasting protection is able to significantly disrupt normal disease patterns for a number of years, however the model returns to its pre-intervention within 10 years. The low efficacy durable vaccine has some effect on

disease dynamics, however, used alone, this would not have a profound impact on incidence. Vaccines with low durability, whether high efficacy or not, are not predicted by the model to have a significant impact on disease dynamics.

In the Korogwe cohort, only the highly durable, high efficacy vaccine is predicted to drive the model to a stable, disease-free equilibrium for at least 6 years. All other vaccines, used alone, do not have a lasting impact on disease state, although the low duration, high efficacy model does reduce overall incidence for some years.

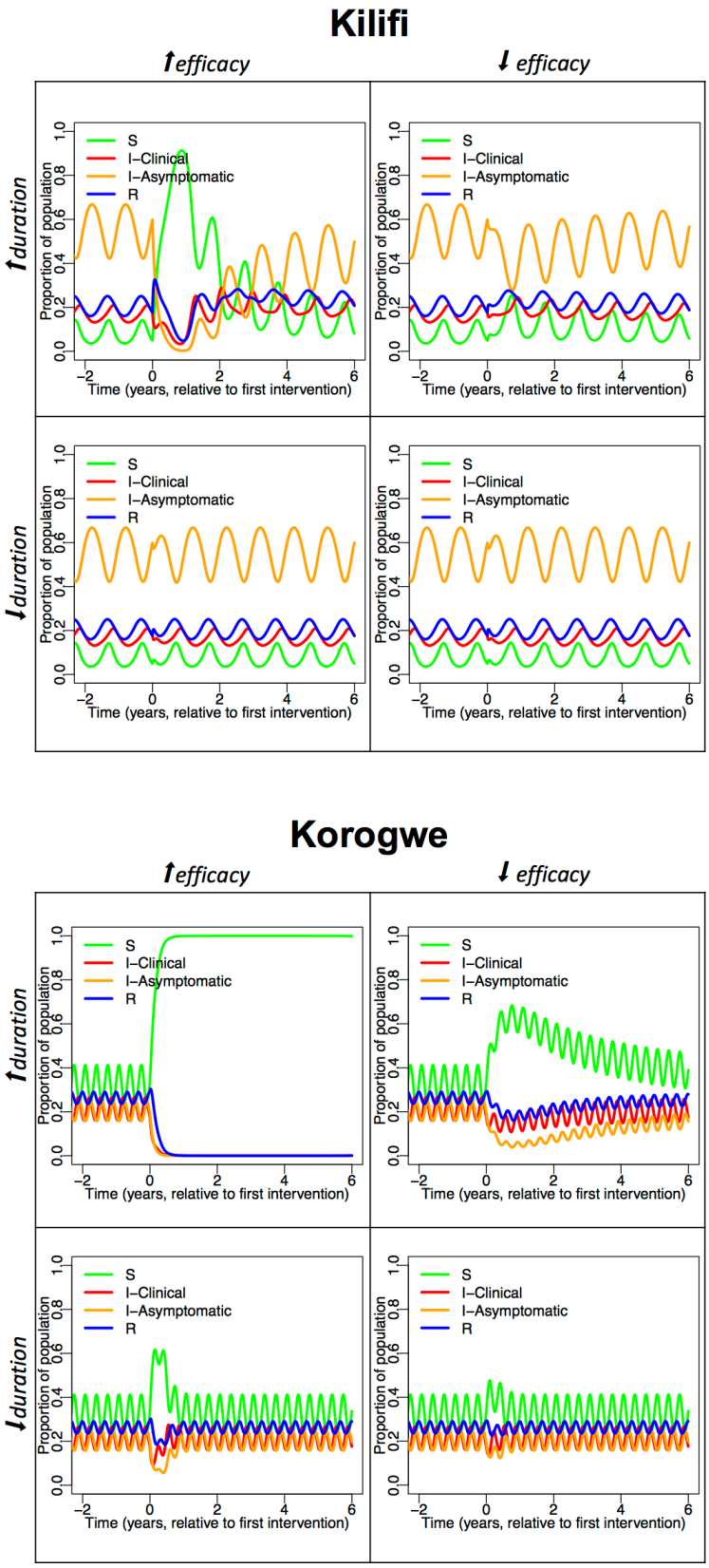


Figure 5-16 Simulations of the impact of mass vaccine administration on disease dynamics in Kilifi (top) and Korogwe (bottom). Four classes of vaccine are shown for each setting, comprising high or low baseline efficacy, with a long or short duration of protection.

Transmission blocking interventions with high effective coverage (80%) are predicted to drive local elimination in both cohorts with 1 year of implementation, however at moderate effective coverage (50%) are predicted to drive a disease-free state in only the lower transmission setting (Korogwe). Note that our model assumes that transmission blocking interventions are applied with constant effect. In the field this would require, for example, replacement of ITNs such that the effectiveness of the insecticide treatment does not wane, nor damage to them occur.

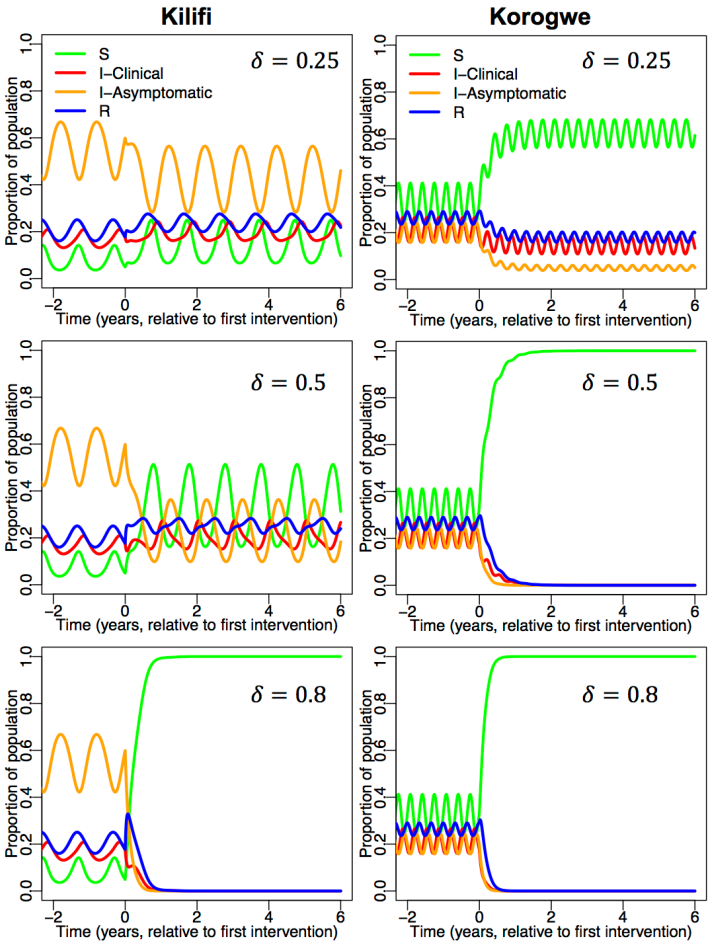


Figure 5-17 Simulations of the impact of high, medium, and low effective coverage transmission-blocking interventions in Kilifi (left three plots) and Korogwe (right three plots)

Finally, MDA in either transmission setting is predicted to only have a short-term impact on disease dynamics. In the highest coverage drug intervention applied to the Korogwe cohort, total cases of infection are reduced to less than 25% of the population for a period of approximately one year, however by year two case load has returned to pre-intervention levels. Similar, albeit less pronounced, patterns are also evident as a result of lower coverage drug administration. In the Kilifi cohort, mass drug administration results in a pair of sharp peaks in the proportion of the population who are recovered from infection, however these very quickly pass – normal infection prevalences return within 1 year in all of the MDA interventions simulated.

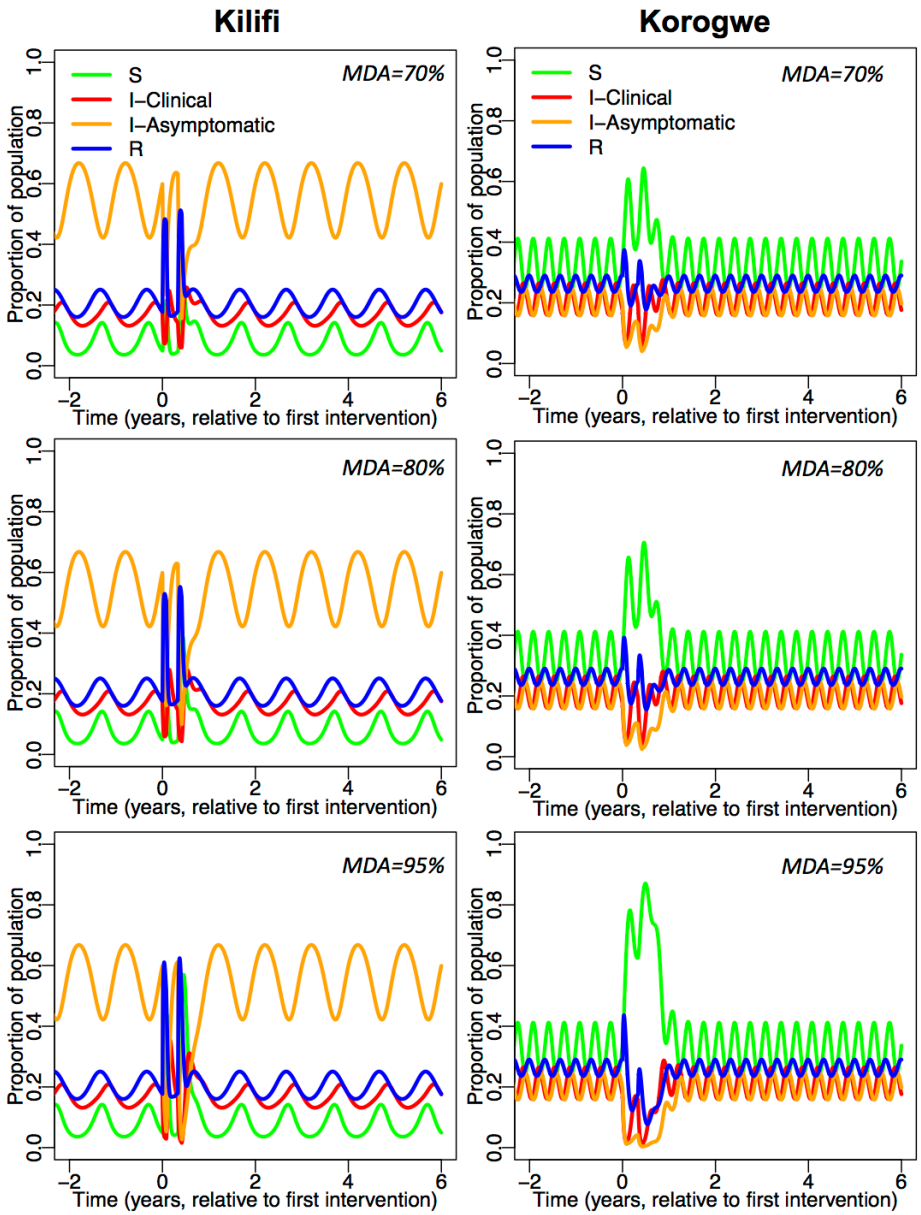


Figure 5-18 Simulations of the impact of high, medium, and low effective coverage Mass Drug Administration campaigns in Kilifi (left) and Korogwe (right)

5.3.3.4 Efficacy of hypothetical combined interventions

Next, we simulate scenarios in which the previously described interventions are applied in combination.

The first scenario models the use of transmission blocking interventions alongside a MDA campaign. As stated in the Methods, typical bed net usage is reported to range from 35-40% (Snow *et al.*, 2015), and a typical MDA may be considered to have a coverage of around 70% (Brady *et al.*, 2017).

In this scenario, our model suggests that the combination of an MDA and transmission blocking intervention may be sufficient to drive prevalence to very low levels in the lower transmission setting (Korogwe; Figure 5-19). Overall prevalence is reduced to less than 5% of the population in the years post-intervention, however, as a result of waning immunity, 87% of infections result in clinical disease (note that this still comprises a substantial net reduction on pre-intervention levels). This is a substantial reduction from a pre-intervention mean prevalence of 41%, 52% of which comprised clinical disease (Figure 5-19). In the higher transmission Kilifi cohort, however, prevalence remains reasonably high when only MDA and typical transmission blocking interventions are implemented. Before intervention total prevalence stands at 72% of the population (23% of which is clinical), and the combined interventions reduce prevalence to 57% of the population (35% of which is clinical).

We then add an MVA campaign (using parameter values found by the IBSM model fitting) to the MDA and the transmission blocking intervention (Figure 5-19). We see that the addition of an MDA campaigns clears the low-level infection seen in the Korogwe cohort under MDA and TBI and drives stable eradication. In the higher

transmission Kilifi cohort, the triple combination causes infection to be virtually cleared from the population for one year. Through years 1 and 2 post-intervention caseloads gradually increase, and from after year 3 a new, stable, disease equilibrium is reached. Over all of the years sampled (up to 6 years post-intervention), total prevalence post-intervention in Kilifi is 40%, of which 41% of infections are clinical (Figure 5-19).

In both the dual and triple combination intervention scenarios in the Kilifi transmission setting there is an increase in clinical infection rates post-intervention (Figure 5-19); this is discussed later (see *Discussion*).

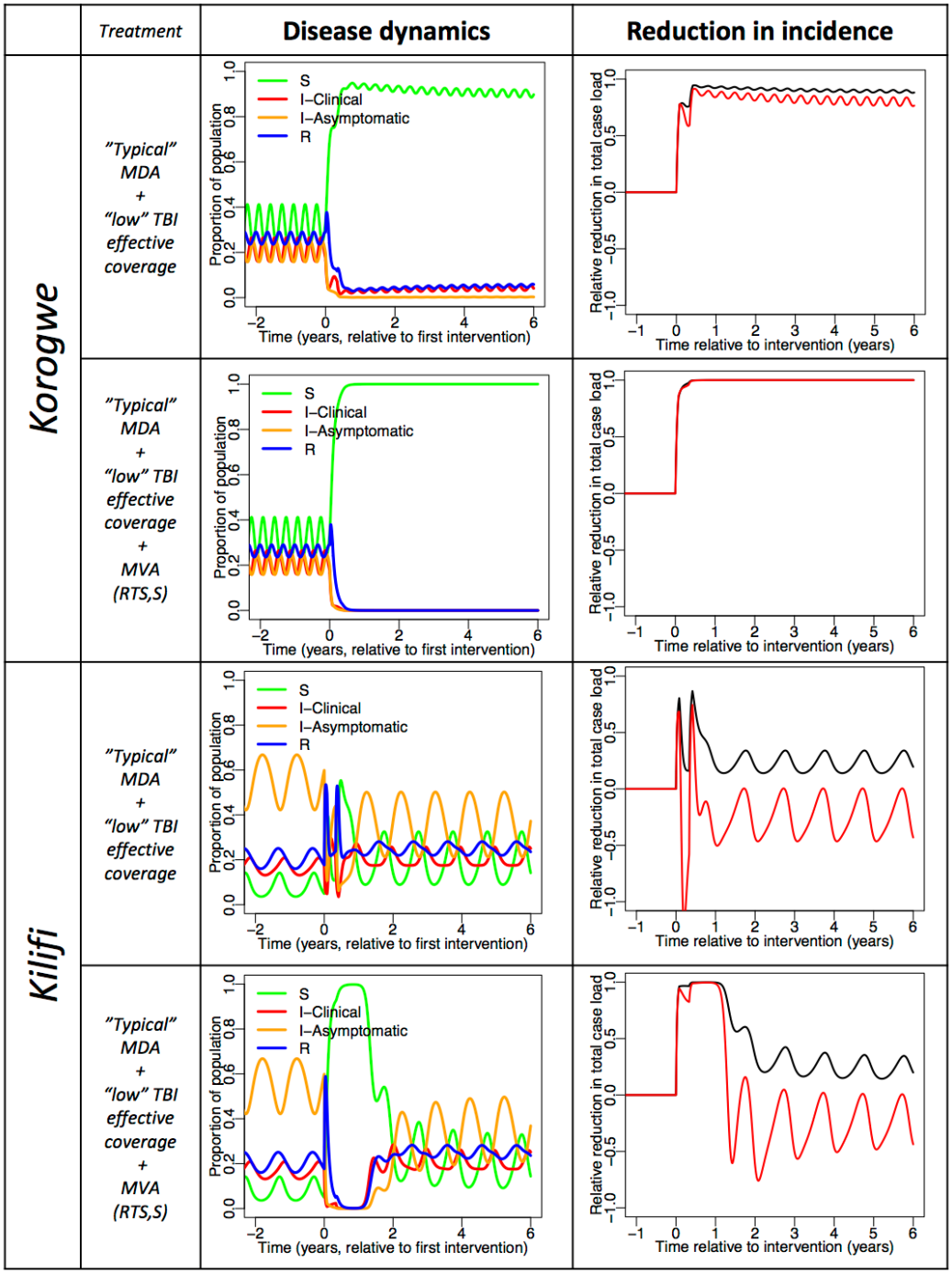


Figure 5-19 Infectious disease dynamics in a pair of intervention scenarios: i) a combination of typical MDA (70% effective coverage) and transmission blocking interventions (35% effective coverage), and ii) typical MDA and TBI with the addition of a vaccine with parameter values for duration and baseline efficacy set by the MCMC routine. The left column plots show the proportion of the population who are susceptible (S), infected with clinical presentation (I-Clinical), infected but asymptomatic (I-Asymptomatic), or recovered from infection (R). The right column gives the proportion reduces in incidence (relative to a control scenario in that cohort) that the intervention causes.

5.4 Discussion

In recent decades, considerable progress has been made in the control of malaria, with a reduction in overall incidence and malaria-associated mortality in many settings (WHO, 2016). This progress is borne of increased access to medical facilities and anti-malarial drugs, and decreased transmission through vector control and ITN usage, both of which have been enabled by increases in funding (Cibulskis *et al.*, 2016). Achievement of the 2030 WHO global malaria targets, however, will require further increases in investment (Head *et al.*, 2017), both to develop and optimize novel interventions, and to roll out those interventions on a mass scale. Maximising the utility of available resources is a crucial step in designing an effective anti-disease campaign, and mathematical modelling has and will continue to play a key role in intervention design (Keeling and Danon, 2009; Mandal, Sarkar and Sinha, 2011; Siettos and Russo, 2013; Veliov and Widder, 2016).

Malaria vaccine development efforts have thus far yielded a single licensed vaccine, RTS,S. Estimates of efficacy of the vaccine vary, with the largest scale phase 3 trial conducted to date reporting efficacy at one year post-administration of 46.3% (95% CI of 21.4% to 62.8%), but not reporting any statistically significant efficacy from 1 year of administration (Olotu *et al.*, 2016). It has, therefore, been suggested that alternative administration routines should be considered that take advantage of the vaccine's high, albeit transient, protective efficacy (Gosling and von Seidlein, 2016). We here test the hypothesis that a combination of existing interventions (ITNs and MDA) with mass administration of an RTS,S-like malaria vaccine could comprise an effective eradication solution in many contexts. This combination would clear parasitemia from enough of the population (and for a sufficient period of time) for normal transmission cycles to be interrupted.

Numerous other works have sought to use mathematical models to assess the efficacy of RTS,S in 3 or 4 dose regimens. In Penny et al., 2016, four such models are systematically compared. Each of these models were designed and analysed by a different research team and differ in their structure and complexity, however each produces broadly comparable estimates of RTS,S efficacy based on large amounts of data from clinical trials of >32 months (Penny *et al.*, 2016). Differences in model structure between the four models and the model described here means that direct parameter comparisons are impossible, however I can analyse and compare profiles of waning efficacy from each. We see that the basic efficacy profiles from each study and described here (for the Kilifi cohort, in whom there is sufficient power to detect waning efficacy) are qualitatively similar (Figure 5-20).

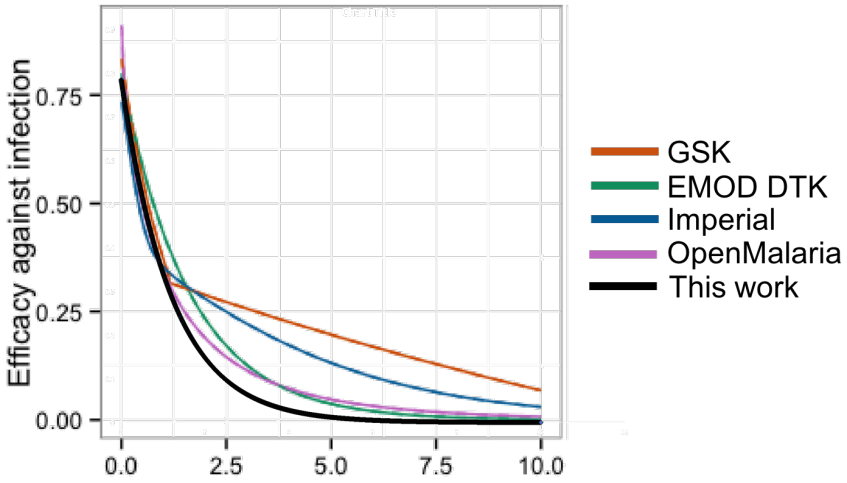


Figure 5-20 Qualitative comparison of vaccine efficacy profiles from the four models described in Penny *et al.*, 2016. Figure modified from Penny et al, with vaccine profile from the model described here added.

Our approach here allows us to utilize data from field trials of the vaccine to predict the population-level impact on transmission of mass vaccination. We estimate that MVA alone, with a vaccine with an efficacy profile analogous to that of RTS,S, may be

sufficient to fully interrupt local transmission in low transmission settings such as Korogwe, Tanzania. In higher transmission settings, transmission interruption also relies on co-implementation of existing strategies of MDA campaigns and ITN usage. We note, however, some of the simplifying assumptions that are made by our approach, and some limitations that may require further investigation.

The model that was used to test the population level effects of intervention comprises a system of ODEs (Águas *et al.*, 2008). The outputs of such models should be carefully interpreted as they do not allow for stochastic events such as extinction or reintroduction to occur. In the case of malaria this is particularly important given the role that naturally acquired immune responses play in limiting incidence rates of severe disease. As disease rates approach 0, naturally acquired immunity to infection in the population will wane, and so there is a possibility for a rebound in cases some years post-intervention (Olotu *et al.*, 2016). In most cases, and at a population level, it might be expected that this rebound may likely not fully offset the number of cases prevented by interventions, however it is important for disease surveillance operations to continue as transmission is reduced. Further work should seek to study possible case resurgence scenarios under waning immunity as local eradication is achieved.

There are some scenarios tested here in which a model that allowed for extinction may have produced different estimates of long-term disease rates. For example, in the *"Typical MDA" + "low" TBI effective coverage + MVA (RTS,S)* scenario modelled in Kilifi (Figure 5-19), all-case incidence of malaria reduces to nearly 0 for at least 1 year post-intervention. In our model, incidence then rebounds to a new (disease-positive) equilibrium state, however it is reasonable to expect that a model incorporating stochastic extinction (re-introduction aside) would predict eradication. Indeed,

previous studies conclude that malaria eradication is "sticky", that is, models often tend to a disease-free equilibrium (Smith *et al.*, 2013). The authors cite a number of possible reasons for this, including having an increased capacity to treat infections (and control transmission).

An additional consideration that should be made in interpreting these results is that the data that was used to fit our IBSMs to each cohort was collected in modestly sized phase 2 trials in areas where incidence of malaria is reasonably low. The number of individuals across both cohorts (Kilifi and Korogwe, vaccine recipients and controls) totals 894, with a total of 179 and 89 infections in the Kilifi and Korogwe cohorts respectively. The power with which we can deduce patterns in transmission and reduction in incidence as a result of vaccination is therefore reasonably low. That said, we do detect signals of seasonal variation in transmission and of vaccine efficacy, in spite of a reasonably limited dataset, thus indicating that the approach is robust to such data and that strong signals are present. The estimates of efficacy produced by fitting models to either cohort vary; in the Korogwe cohort, there is insufficient power to be able to detect any signal of waning efficacy. In Kilifi, the higher overall case load means that a signal of waning efficacy can be detected, and so we also test a population model implemented using the vaccine efficacy estimates gained from that cohort but implemented in the Korogwe transmission setting (data not shown). In this scenario, the outcomes are qualitatively similar to the triple combination shown in Figure 5-19 – eradication is achieved within 1 year of intervention.

An interesting finding from our model is that, in some transmission settings, it may be possible for there to be short-term increases in clinical cases of malaria as a result of interventions. This observation corroborates those made previously (Snow *et al.*, 1997; Snow and Marsh, 2002). Aguas *et al.* noted in their discussion of the same model used

in our population modelling approach that the peak of prevalence of clinical malaria occurred at low to intermediate transmission (Águas *et al.*, 2008). This is reported to be due to waning of clinical immunity in the population, and the authors suggest that transmission reducing interventions must be combined with active case management to limit potential increases in clinical burden. That said, as eradication is approached, the immense pressure placed upon healthcare systems will wane and so robust outbreak management solutions will be possible (Smith *et al.*, 2013).

A valuable extension to the model that may improve its accuracy and utility could be to consider that transmission of (on spatial scales) and exposure to (on an individual scale) varies. These heterogeneities mean that a small portion of the population may be responsible for a large portion of transmission of a disease, and this has important implications for control programs (Woolhouse *et al.*, 1997). In the case of malaria, numerous works, modelled and otherwise, have sought to explore the impact of transmission "hotspots" at various scales, from household-to-household to multi-national regions (Rodriguez-Barraquer *et al.*, 2016; Mugenyi, Abrams and Hens, 2017; Platt *et al.*, 2018). These areas of high transmission can persist season on season and are associated with factors including environmental characteristics such as proximity to water, availability of healthcare, and behavioural factors (discussed in Platt *et al.*, 2018). Accounting for heterogeneity in modelling transmission has been found to improve the accuracy of models, however authors have noted that accurate analysis of the degree of heterogeneity in a population is difficult, as technical variance (i.e., from sampling variation) and true biological variance (i.e., heterogeneity) are confounded (Smith, 2008).

In the work presented here, I find that eradication in high transmission cohorts may only be achieved by the combination of multiple arms of therapy: mass vaccination, mass drug administration, and transmission blocking interventions. As introduced previously, stopping an outbreak of disease requires the basic reproductive number, R_0, to fall below 1. In the version of the model presented here, R_0 is assumed not to vary across geographic locations – it only varies temporally. In reality, it has been shown that transmission within some regions areas is low (i.e., *local* R_0 can be close to or below 1) and that, in those areas, disease would not persist were it not for reintroductions from areas of higher transmission, where R_0 is much greater than 1 (Smith *et al.*, 2007).

Restructuring the model to incorporate such heterogeneities would have a two-fold effect. First, it would allow malaria to persist in the model population at a lower overall (*i.e.*, averaged across the entire population) prevalence, which may be more representative of natural systems. Secondly, such a restructuring would also allow for heterogeneities in intervention to be modelled. It is reasonable to expect that mass interventions such as those described here would be most effective if resources were focussed onto transmission hotspots. This would potentially require lower overall resource use, would prevent overspill of disease into the adjacent lower transmission regions, and thus potentially drive eradication. Further work could utilise my model framework to optimise the cost-effectiveness of mass interventions against malaria by considering targeting resources against those regions where they would be most effective.

In conclusion, we here present a novel approach for scaling up estimates of vaccine efficacy from clinical trials to predict the potential impact of population-wide

implementation of an existing anti-malaria vaccine. We generate robust estimates of seasonal force of infection and vaccine efficacy profiles in a number of contexts, and then scale-up our framework to produce population level projections of disease dynamics. Disease dynamics are modelled without intervention, under monotherapy, and in scenarios where a vaccine is used in combination with existing intervention strategies. We find that combinations of interventions are highly effective in stopping malaria transmission, and that the addition of a mass vaccination campaign with an existing malaria vaccine targeting the pre-erythrocytic stages of infection may allow local elimination across transmission settings. Further work should seek to re-evaluate RTS,S as a candidate for mass administration in adults: a combinatorial intervention including MVA would be efficacious, cost-effective, and could form a keystone of near-future eradication campaigns.

Chapter 6 Summary and conclusions

The central role that computational biology plays in designing and optimising treatments against pathogenic infections of global importance is irrefutable. For decades, scientists have been producing mathematical models of the spread of disease in order to understand the mechanisms by which a pathogen can persist and transmit in a population and to assist in designing therapies that stand the best possible chance of interrupting transmission, clearing infection, and eradicating the parasite (Kermack and McKendrick, 1927; Anderson and May, 1991). In addition, the widespread availability and greatly reduced cost of Whole Genome Sequencing stands to hail an era of genomic medicine (Land *et al.*, 2015), in which long-held host-pathogen theories, previously only testable through models, can be rigorously investigated and validated.

In this thesis, I present a number of distinct computational works centred around a common theme: the design, evaluation and optimisation of therapies against (multi-strain) infectious diseases. These works span various stages of vaccine development, from the identification of epitopes that a vaccine could raise an immune response against (Chapter 2 and 3), to within host modelling of the potential effects of combinations of malaria vaccines on parasite numbers (Chapter 4 and Appendix B), and finally to modelling, at a population level, the potential impact of implementation of novel vaccination regimens, particularly in combination with existing therapies (Chapter 5).

6.1 Immune-driven evolution

Effective vaccines against infectious diseases often seek to replicate those immune responses that an individual would mount in response to infection. For many antigenically conserved pathogens, this is relatively trivial as a single infection commonly produces a protective immune response. Successes seen in vaccine development to date have been against pathogens such as polio or measles, against which a single infection raises a protective immune response; vaccines can seek to mimic that highly efficacious response. In contrast, in more antigenically variable pathogens, effective immune responses often only accrue after multiple or prolonged infections (Chattopadhyay *et al.*, 2003; Gatton and Cheng, 2004). That is, immune responses against a single antigenic variant fail to prevent or clear infection, and so responses must accumulate to cover enough of antigenic space that the ability of the pathogen to evade or tolerate responses is limited.

Strain theory states that pathogens evolve in response to the immune pressure that they face and, in certain extreme conditions (i.e., high levels of cross-immunity, Chapters 2 and 3), segregate into non-overlapping, sets of antigenic variants (strains) (Gupta, Ferguson and Anderson, 1998). This segregation is expected to be evident in the amino acid sequences of proteins containing epitope regions under immune selection and so, by searching for patterns of non-overlap in sequence data, we are able to reverse strain theory to putatively elucidate epitopes as those which exhibit the aforementioned evolutionary patterns (Chapters 2 and 3).

As discussed previously, an extra factor to consider when designing vaccines (against pathogens where T cell responses dominate) is the ability of the host to present different variants of the same peptide (i.e. different amino acid residues at the same

site) in the peptide MHC complex. If some variants of a peptide undergo less cellular processing or are less likely, once processed, to be presented on the cell surface by the MHC, then this may affect the potential efficacy of a vaccine and may potentiate escape. As such, intracellular pathogens seek to evolve towards variants which cannot be presented by the host's MHC, however this evolution is constrained by a need to maintain protein function. The optimum fitness of a pathogen is one which maintains intrinsic fitness through having functional proteins whilst still evading host immune responses by minimising presentation or avoiding T cell recognition through antigenic change (Wu *et al.*, 2015; Wu and Wilson, 2017).

As considered in Chapter 3, factors other than immune selection may drive patterns of non-overlap in sequence data. Stringent functional constraints impose fixed bounds upon possible genotypic space. Outside of these regions, the protein, and thus the pathogen, would be non-viable or face severely limited fitness. Compensatory mutations can go some way in rescuing deleterious mutations however, given that each possible mutation occurs with a given probability, it follows that the probability of simultaneous co-occurrence of each additional mutation decreases non-linearly. That said, in some pathogens where mutation rates are extremely high, e.g. in retroviruses, there are multiple examples of such compensatory mutations (Brockman *et al.*, 2007; Rihn *et al.*, 2013).

These considerations, however, do not have a strong bearing on the conclusions of this thesis. Here, I postulate that it is not necessary to fully extricate potential causes of non-overlap from each other. The fitness of a pathogen is a function of both of immune evasion and intrinsic fitness; these factors are inextricably linked, and patterns of diversity borne of both. Immune selection pressure drives evolution, however under

the imperative that those evolved variants may be passed on. Evolutionary change and the emergence of strain structure can only exist within the bounds of structural-functional constrains placed upon the antigenic region.

6.2 Rethinking vaccine strategy

In Chapter 4, I expand my focus to consider the impact of vaccination on the within-host dynamics of infection with *P. falciparum* malaria. I develop and fit a mathematical model to data from Controlled Human Malaria Infection (CHMI) trials in which healthy individuals are vaccinated with either a malaria vaccine candidate or a control vaccine, and then regularly tested by qPCR or blood smear analysis for malaria positivity (Walker *et al.*, 2015). These detection points effectively correspond to thresholds of net parasitemia within each host. Hence, by analysing the time it takes for the thresholds to be reached, one can estimate the average pre-erythrocytic and erythrocytic burden in each patient, as well as the effect that each vaccine has on that burden.

By fitting the model separately to data from recipients of two different vaccines, I have projected the efficacy of their combination in a single individual. This model predicts that vaccines that act primarily against different stages of the pre-erythrocytic cycle may do so synergistically, producing efficacy estimates well in excess of the sum of the two components.

A further interesting point to arise from this work corroborates a paradox that has been noted by modelling works previously: the number of merozoites that egress from a single infected liver cell should be sufficient for blood stage infection to be detected by qPCR immediately (Bejon et al. 2005). The hepatic stage of infection is known to

last for almost exactly 1 week however, in the trial data, there is a delay to qPCR detection in some individuals. A possible explanation could be that vaccines that act against infected hepatocytes can act to reduce the net output of a single infected hepatocyte, but do not always cause its total destruction. An alternative explanation could be that newly released merozoites, encapsulated in merosomes, sequester in the deep vasculature of the liver, or the lungs (Chapter 4). Release of the merosome cargo is known to occur in the lungs in murine models of malaria (Baer *et al.*, 2007), however further experimental work is required to confirm whether or not this occurs *in vivo*.

In Chapter 5, I investigated the potential impact of interventions on a population level. Mass interventions against infectious diseases reduce net transmission and (in the cases of vaccines) confer herd immunity. In other words, individuals extraneous to the intervention can gain benefit from it as a result of net reductions in transmission. Malaria vaccine development has focussed primarily on providing treatment for children and infants, in whom the vast major of clinical presentations of malaria occur. In my approach, I generate estimates of seasonal force of infection and vaccine efficacy profiles in modestly sized field trails in a number of transmission contexts, and then utilise those estimates in a population model to predict the potential impact of mass vaccine administration. I provide evidence that mass administration of a vaccine analogous in efficacy profile to RTS,S may, especially in combination with existing interventions such as MDA and ITN use, interrupt transmission in many transmission settings.

This approach allowed me to produce estimates of seasonality and vaccine efficacy in a number of transmission settings in the context of medium-scale clinical trials. This research could be highly informative for trial design and to policy makers, and the

results encourage the widespread use of an existing medium-efficacy vaccine in careful conjunction with existing interventions, even in moderate transmission settings.

6.3 Conclusions

In coming decades, if current trends continue, incidence rates of malaria are set to increase. Ongoing control efforts face a number of challenges, particularly in ensuring access to modern anti-malarial drugs, insecticide resistance, and the lack of funding plans to ensure that current interventions are maintained (WHO, 2016). These factors, teamed with the rapidly emerging threat of drug resistance in South East Asia and Africa (Lu *et al.*, 2017; Menard and Dondorp, 2017), stand to hamper achievement of elimination goals. Overall, this thesis postulates that careful application of existing therapies may go a long way towards achieving those goals.

This thesis also suggests that the wealth of genetic sequence data that is generated for pathogens of global health important can be utilized, in combination with existing theory, to better our understanding of the immune interactions between pathogen and host. Understanding these interactions will facilitate development of vaccines that are efficacious, strain-transcending, and that elicit immune responses robust to parasites developing resistance against them.

The evolution of multi-drug resistance by pathogens is projected to cause 10 million fatalities per year by 2050 (O'Neill, 2014): rapid action is required to ensure that this worrisome prediction does not come to pass. In the fight against malaria, it is imperative that we utilize all of the tools at our disposal, both those that are already routinely implemented, and also that have not yet seen widespread use. This thesis provides evidence that careful implementation of vaccines (or combinations of

vaccines), in tandem with existing intervention strategies, may be highly effective in driving eradication. Considering infectious diseases more widely, it is imperative that the potential impact on pathogen dynamics that immune-based interventions could have is fully considered. The interactions that occur between host and pathogen are complex and often borne of millennia of co-evolution, and we would be naïve to think that pathogenic evolution has ceased.

Bibliography

Adami, C. (2004) 'Information theory in molecular biology', *Physics of Life Reviews*, 1(1), pp. 3–22.

Agnandji, S. T., Lell, B., Fernandes, J. F., Abossolo, B. P., Methogo, B. G. N. O., Vansadia, P., *et al.* (2012) 'A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants.', *The New England journal of medicine*, 367(24), pp. 2284–95.

Águas, R., White, L. J., Snow, R. W. and Gomes, M. G. M. (2008) 'Prospects for Malaria Eradication in Sub-Saharan Africa', *PLoS ONE*. Edited by C. J. Sutherland, 3(3), p. e1767.

Al-awadi, A. R., Al-kuhlani, A., Breman, J. G., Doumbo, O., Eberhard, M. L., Guiguemde, R. T., Magnussen, P., Molyneux, D. H. and Nadim, A. (2014) 'Guinea worm (dracunculiasis) eradication: Update on progress and endgame challenges', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 108(5), pp. 249–251.

Ali, A. S., Thawer, N. G., Khatib, B., Amier, H. H., Shija, J., Msellem, M., Al‐mafazy, A., Garimo, I. A., Mkali, H., Ramsan, M. M., Kafuko, J. M., Paxton, L. A., Reithinger, R. and Ngondi, J. M. (2017) 'Artemisinin combination therapy mass drug administration in a setting of low malaria endemicity: programmatic coverage and adherence during an observational study in Zanzibar', *Malaria Journal*. BioMed Central, 16(16), pp. 1–8.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) 'Basic local alignment search tool.', *Journal of molecular biology*, 215(3), pp. 403–10.

Ambinder, R. F. and Lin, L. (2005) 'Mononucleosis in the laboratory.', *The Journal of infectious diseases*, 192(9), pp. 1503–4.

Anderson, R. M. and May, R. M. (1991) *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Andreasen, V., Lin, J. and Levin, S. A. (1997) 'The dynamics of cocirculating influenza strains conferring partial cross-immunity', *Journal of Mathematical Biology*, 35(7), pp. 825–842.

Andrews, L., Andersen, R. F., Webster, D., Dunachie, S., Walther, R. M., Bejon, P., Hunt-Cooke, A., Bergson, G., Sanderson, F., Hill, A. V and Gilbert, S. C. (2005) 'Quantitative real-time polymerase chain reaction for malaria diagnosis and its use in malaria vaccine clinical trials', *Am J Trop Med Hyg*, 73, pp. 191–198.

Arama, C. and Troye-Blomberg, M. (2014) 'The path of malaria vaccine development: Challenges and perspectives', *Journal of Internal Medicine*, 275(5), pp. 456–466.

Aregawi, M., Smith, S. J., Sillah-Kanu, M., Seppeh, J., Kamara, A. R. Y., Williams, R. O., Aponte, J. J., Bosman, A. and Alonso, P. (2016) 'Impact of the Mass Drug Administration for malaria in response to the Ebola outbreak in Sierra Leone', *Malaria Journal*. BioMed Central, 15(1), p. 480.

Areschoug, T., Stålhammar-carlemalm, M., Larsson, C. and Lindahl, G. (1999) 'Group B Streptococcal Surface Proteins as Targets for Protective Antibodies : Identification of Two Novel Proteins in Strains of Serotype V Group B Streptococcal Surface Proteins as Targets for Protective Antibodies : Identification of Two Novel Proteins ', 67(12), pp. 6350–6357.

Arnott, A., Jardine, D., Wilson, K., Gorry, P. R., Merlin, K., Grey, P., Law, M. G., Dax, E. M., Kelleher, A. D., Smith, D. E. and McPhee, D. A. (2010) 'High viral fitness during acute HIV-1 infection', *PLoS ONE*, 5(9), pp. 1–12.

Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., White, N. J., *et al.* (2014) 'Spread of artemisinin resistance in Plasmodium falciparum malaria', *N.Engl.J.Med.*, 371(1533–4406 (Electronic)), pp. 411–423.

Baer, K., Klotz, C., Kappe, S. H. I., Schnieder, T. and Frevert, U. (2007) 'Release of hepatic Plasmodium yoelii merozoites into the pulmonary microvasculature.', *PLoS pathogens*, 3(11), p. e171.

Beaty, S. M. and Lee, B. (2016) 'Constraints on the genetic and antigenic variability of measles virus', *Viruses*, 8(4), pp. 1–20.

Beier, J. C., Beier, M. S., Vaughan, J. A., Pumpuni, C. B., Davis, J. R. and Noden, B. H. (1992) 'Sporozoite transmission by Anopheles freeborni and Anopheles gambiae experimentally infected with Plasmodium falciparum.', *Journal of the American Mosquito Control Association*, 8(4), pp. 404–8.

Bejon, P., Andrews, L., Andersen, R. F., Dunachie, S., Webster, D., Walther, M., Gilbert, S. C., Peto, T. and Hill, A. V. S. (2005) 'Calculation of liver-to-blood inocula, parasite growth rates, and preerythrocytic vaccine efficacy, from serial quantitative polymerase chain reaction studies of volunteers challenged with malaria sporozoites.', *The Journal of infectious diseases*, 191(4), pp. 619–26.

Bejon, P., Lusingu, J., Olotu, A., Leach, A., Lievens, M., Vekemans, J., Mshamu, S., Lang, T., Gould, J., Dubois, M.-C., Demoitié, M.-A., Stallaert, J.-F., Vansadia, P., Carter, T., Njuguna, P., Awuondo, K. O., Malabeja, A., Abdul, O., Gesase, S., Mturi, N., Drakeley, C. J., Savarese, B., Villafana, T., Ballou, W. R., Cohen, J., Riley, E. M., Lemnge, M. M., Marsh, K. and von Seidlein, L. (2008) 'Efficacy of RTS,S/AS01E Vaccine against Malaria in Children 5 to 17 Months of Age', *New England Journal of Medicine*, 359(24), pp. 2521–2532.

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E. and Gething, P. W. (2015) 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015', *Nature*, 526(7572), pp. 207–211.

Blake, N., Lee, S., Redchenko, I., Thomas, W., Steven, N., Leese, A., Steigerwald-Mullen, P., Kurilla, M. G., Frappier, L. and Rickinson, A. (1997) 'Human CD8+ T cell responses to EBV EBNA1: HLA class I presentation of the (Gly-Ala)-containing protein requires exogenous processing', *Immunity*, 7(6), pp. 791–802.

Bogedain, C., Wolf, H., Modrow, S., Stuber, G. and Jilg, W. (1995) 'Specific cytotoxic T lymphocytes recognize the immediate-early transactivator Zta of Epstein-Barr virus.', *Journal of virology*, 69(8), pp. 4872–4879.

Boisier, P., Nicolas, P., Djibo, S., Taha, M.-K., Jeanne, I., Mainassara, H. B., Tenebray, B., Kairo, K. K., Giorgini, D. and Chanteau, S. (2007) 'Meningococcal Meningitis: Unprecedented Incidence of Serogroup X--Related Cases in 2006 in Niger', *Clinical Infectious Diseases*, 44(5), pp. 657–663.

Bollard, C. M., Gottschalk, S., Torrano, V., Diouf, O., Ku, S., Hazrat, Y., Carrum, G., Ramos, C., Fayad, L., Shpall, E. J., Pro, B., Liu, H., Wu, M. F., Lee, D., Sheehan, A.

M., Zu, Y., Gee, A. P., Brenner, M. K., Heslop, H. E. and Rooney, C. M. (2014) 'Sustained complete responses in patients with lymphoma receiving autologous cytotoxic T lymphocytes targeting Epstein-Barr virus latent membrane proteins', *Journal of Clinical Oncology*, 32(8), pp. 798–808.

Bornkamm, G. W. and Hammerschmidt, W. (2001) 'Molecular virology of Epstein-Barr virus', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1408), pp. 437–459.

Bousema, T., Okell, L., Felger, I. and Drakeley, C. (2014) 'Asymptomatic malaria infections: detectability, transmissibility and public health relevance', *Nature Reviews Microbiology*. Nature Publishing Group, 12(12), pp. 833–840.

Brady, O. J., Slater, H. C., Pemberton-Ross, P., Wenger, E., Maude, R. J., Ghani, A. C., Penny, M. A., Gerardin, J., White, L. J., Chitnis, N., Aguas, R., Hay, S. I., Smith, D. L., Stuckey, E. M., Okiro, E. A., Smith, T. A. and Okell, L. C. (2017) 'Role of mass drug administration in elimination of Plasmodium falciparum malaria: a consensus modelling study', *The Lancet Global Health*, 5(7), pp. e680–e687.

Brockman, M. A., Schneidewind, A., Lahaie, M., Schmidt, A., Miura, T., DeSouza, I., Ryvkin, F., Derdeyn, C. A., Allen, S., Hunter, E., Mulenga, J., Goepfert, P. A., Walker, B. D. and Allen, T. M. (2007) 'Escape and Compensation from Early HLA-B57-Mediated Cytotoxic T-Lymphocyte Pressure on Human Immunodeficiency Virus Type 1 Gag Alter Capsid Interactions with Cyclophilin A', *Journal of Virology*, 81(22), pp. 12608–12618.

Brunham, R. C., Plummer, F. A. and Stephens, R. S. (1993) 'Bacterial antigenic variation, host immune response, and pathogen-host coevolution.', *Infection and Immunity*, 61(6), pp. 2273–2276.

Buckee, C. O., Gupta, S., Kriz, P., Maiden, M. C. J. and Jolley, K. A. (2010) 'Long-term evolution of antigen repertoires among carried meningococci.', *Proceedings. Biological sciences / The Royal Society*, 277(1688), pp. 1635–1641.

Buckee, C. O., Jolley, K. a, Recker, M., Penman, B., Kriz, P., Gupta, S. and Maiden, M. C. J. (2008) 'Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(39), pp. 15082–7.

Buckee, C. O., Recker, M., Watkins, E. R. and Gupta, S. (2011) 'Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations', *Proc Natl Acad Sci U S A*, 108(37), pp. 15504–15509.

Bull, P. C., Berriman, M., Kyes, S., Quail, M. a., Hall, N., Kortok, M. M., Marsh, K. and Newbold, C. I. (2005) 'Plasmodium falciparum variant surface antigen expression patterns during malaria', *PLoS Pathogens*, 1(3), pp. 0202–0213.

Cairns, M. E., Walker, P. G. T., Okell, L. C., Griffin, J. T., Garske, T., Asante, K. P., Owusu-Agyei, S., Diallo, D., Dicko, A., Cisse, B., Greenwood, B. M., Chandramohan, D., Ghani, A. C. and Milligan, P. J. (2015) 'Seasonality in malaria transmission: implications for case-management with long-acting artemisinin combination therapy in sub-Saharan Africa', *Malaria Journal*. BioMed Central, 14(1), p. 321.

Calis, J. J. A., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., Keşmir, C. and Peters, B. (2013) 'Properties of MHC Class I Presented Peptides That Enhance Immunogenicity', *PLoS Computational Biology*, 9(10).

Callaghan, M. J., Buckee, C. O., Jolley, K. a, Kriz, P., Maiden, M. C. J. and Gupta, S. (2008) 'The effect of immune selection on the structure of the meningococcal opa protein repertoire.', *PLoS pathogens*, 4(3), p. e1000020.

Chattopadhyay, R., Sharma, A., Srivastava, V. K., Pati, S. S., Sharma, S. K., Das, B. S. and Chitnis, C. E. (2003) 'Plasmodium falciparum Infection Elicits Both Variant-Specific and Cross-Reactive Antibodies against Variant Surface Antigens', 71(2), pp. 597–604.

Chen, J. and Deng, Y.-M. (2009) 'Influenza virus antigenic variation, host antibody production and new approach to control epidemics.', *Virology journal*. BioMed Central, 6, p. 30.

Chen, M. R. (2011) 'Epstein-barr virus, the immune system, and associated diseases', *Frontiers in Microbiology*, 2(JAN), pp. 1–5.

Cherif, A., Dyson, J., Maini, P. K. and Gupta, S. (2017) 'An age-structured multi-strain epidemic model for antigenically diverse infectious diseases: A multi-locus framework', *Nonlinear Analysis: Real World Applications*, 34, pp. 275–315.

Chia, W.-K., Teo, M., Wang, W.-W., Lee, B., Ang, S.-F., Tai, W.-M., Chee, C.-L., Ng,

J., Kan, R., Lim, W.-T., Tan, S.-H., Ong, W.-S., Cheung, Y.-B., Tan, E.-H., Connolly, J. E., Gottschalk, S. and Toh, H.-C. (2014) 'Adoptive T-cell Transfer and Chemotherapy in the First-line Treatment of Metastatic and/or Locally Recurrent Nasopharyngeal Carcinoma', *Molecular Therapy*. The American Society of Gene & Cell Therapy, 22(1), pp. 132–139.

Chia, W. K., Wang, W.-W., Teo, M., Tai, W. M., Lim, W. T., Tan, E. H., Leong, S. S., Sun, L., Chen, J. J., Gottschalk, S. and Toh, H. C. (2012) 'A phase II study evaluating the safety and efficacy of an adenovirus-ΔLMP1-LMP2 transduced dendritic cell vaccine in patients with advanced metastatic nasopharyngeal carcinoma.', *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, 23(4), pp. 997–1005.

Childs, L. M., Abuelezam, N. N., Dye, C., Gupta, S., Murray, M. B., Williams, B. G. and Buckee, C. O. (2015) 'Modelling challenges in context: Lessons from malaria, HIV, and tuberculosis', *Epidemics*, 10, pp. 102–107.

Chourasia, M. K., Raghavendra, K., Bhatt, R. M., Swain, D. K., Meshram, H. M., Meshram, J. K., Suman, S., Dubey, V., Singh, G., Prasad, K. M. and Kleinschmidt, I. (2017) 'Additional burden of asymptomatic and sub-patent malaria infections during low transmission season in forested tribal villages in Chhattisgarh, India', *Malaria Journal*, 16(1), p. 320.

Chourasia, M. K., Raghavendra, K., Bhatt, R. M., Swain, D. K., Valecha, N. and Kleinschmidt, I. (2017) 'Burden of asymptomatic malaria among a tribal population in a forested village of central India: a hidden challenge for malaria control in India', *Public Health*. Elsevier Ltd, 147, pp. 92–97.

Cibulskis, R. E., Alonso, P., Aponte, J., Aregawi, M., Barrette, A., Bergeron, L., Fergus, C. A., Knox, T., Lynch, M., Patouillard, E., Schwarte, S., Stewart, S. and Williams, R. (2016) 'Malaria: Global progress 2000 – 2015 and future challenges', *Infectious Diseases of Poverty*. Infectious Diseases of Poverty, 5(1), p. 61.

Cohen, J. I. (2000) 'Epstein–Barr Virus Infection', *New England Journal of Medicine*, 343(7), pp. 481–492.

Cohen, J. I., Fauci, A. S., Varmus, H. and Nabel, G. J. (2011) 'Epstein-Barr Virus: An

Important Vaccine Target for Cancer Prevention', *Science Translational Medicine*, 3(107), p. 107fs7-107fs7.

Cuevas, J. M., Geller, R., Garijo, R. and López-aldeguer, J. (2015) 'Extremely High Mutation Rate of HIV-1 In Vivo', pp. 1–19.

Dame, J., Williams, J., McCutchan, T., Weber, J., Wirtz, R., Hockmeyer, W., Maloy, W., Haynes, J., Schneider, I., Roberts, D. and et, al. (1984) 'Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite Plasmodium falciparum', *Science*, 225(4662), pp. 593–599.

Dayhoff, M. and Schwartz, R. (1978) 'A Model of Evolutionary Change in Proteins', *In Atlas of protein sequence and structure*, pp. 345–352.

Dicko, A., Brown, J. M., Diawara, H., Baber, I., Mahamar, A., Soumare, H. M., Sanogo, K., Koita, F., Keita, S., Traore, S. F., Chen, I., Poirot, E., Hwang, J., McCulloch, C., Lanke, K., Pett, H., Niemi, M., Nosten, F., Bousema, T. and Gosling, R. (2016) 'Primaquine to reduce transmission of Plasmodium falciparum malaria in Mali: a single-blind, dose-ranging, adaptive randomised phase 2 trial', *The Lancet Infectious Diseases*, 16(6), pp. 674–684.

Dondorp, A., Nosten, F. and Yi, P. (2009) 'Artemisinin Resistance in Plasmodium falciparum malaria', *The New England Journal of Medicine*, 361, pp. 455–467.

Douglas, A. D., Andrews, L., Draper, S. J., Bojang, K., Milligan, P., Gilbert, S. C., Imoukhuede, E. B. and Hill, a V. S. (2011) 'Substantially reduced pre-patent parasite multiplication rates are associated with naturally acquired immunity to Plasmodium falciparum.', *The Journal of infectious diseases*, 203(9), pp. 1337–40.

Drew, D. R., Hodder, A. N., Wilson, D. W., Foley, M., Mueller, I., Siba, P. M., Dent, A. E., Cowman, A. F. and Beeson, J. G. (2012) 'Defining the Antigenic Diversity of Plasmodium falciparum Apical Membrane Antigen 1 and the Requirements for a Multi-Allele Vaccine against Malaria', *PLoS ONE*, 7(12).

Dublin, L. I. and Lotka, A. J. (1925) 'On the True Rate of Natural Increase', *Journal of the American Statistical Association*, 20(151), pp. 305–339.

Duffy, P. and Kaslow, D. (1997) 'A novel malaria protein, Pfs28, and Pfs25 are genetically linked and synergistic as falciparum malaria transmission-blocking

vaccines', *Infection and Immunity*, 65(3), pp. 1109–1113.

Duncan, C. J. a and Draper, S. J. (2012) 'Review: Controlled human blood stage malaria infection: Current status and potential applications', *American Journal of Tropical Medicine and Hygiene*, 86(4), pp. 561–565.

Dutta, S., Dlugosz, L. S., Drew, D. R., Ge, X., Ababacar, D., Rovira, Y. I., Moch, J. K., Shi, M., Long, C. A., Foley, M., Beeson, J. G., Anders, R. F., Miura, K., Haynes, J. D. and Batchelor, A. H. (2013) 'Overcoming Antigenic Diversity by Enhancing the Immunogenicity of Conserved Epitopes on the Malaria Vaccine Candidate Apical Membrane Antigen-1', *PLoS Pathogens*, 9(12), p. e1003840.

Edgar, R. C., Drive, R. M. and Valley, M. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.', *Nucleic Acids Research*, 32(5), pp. 1792–7.

Elliott, S. L., Suhrbier, A., Miles, J. J., Lawrence, G., Pye, S. J., Le, T. T., Rosenstengel, A., Nguyen, T., Allworth, A., Burrows, S. R., Cox, J., Pye, D., Moss, D. J. and Bharadwaj, M. (2008) 'Phase I trial of a CD8+ T-cell peptide epitope-based vaccine for infectious mononucleosis.', *Journal of virology*, 82(3), pp. 1448–57.

Elwood, J. M. (1989) 'Smallpox and its eradication', *Journal of epidemiology and community health*, 43(1), p. 92.

Epstein, J. E. and Richie, T. L. (2013) 'The whole parasite, pre-erythrocytic stage approach to malaria vaccine development: a review.', *Current opinion in infectious diseases*, 26(5), pp. 420–8.

European Medicines Agency (2015) *First malaria vaccine receives positive scientific opinion from EMA*.

Ewer, K. J., O'Hara, G. A., Duncan, C. J. A., Collins, K. A., Sheehy, S. H., Reyes-Sandoval, A., Goodman, A. L., Edwards, N. J., Elias, S. C., Halstead, F. D., Longley, R. J., Rowland, R., Poulton, I. D., Draper, S. J., Blagborough, A. M., Berrie, E., Moyle, S., Williams, N., Siani, L., Folgori, A., Colloca, S., Sinden, R. E., Lawrie, A. M., Cortese, R., Gilbert, S. C., Nicosia, A. and Hill, A. V. S. (2013) 'Protective CD8(+) T-cell immunity to human malaria induced by chimpanzee adenovirus-MVA immunisation.', *Nature communications*, 4, p. 2836.

De Filippis, I., De Andrade, C. F., Silva, L., Prevots, D. R. and Vicente, A. C. P. (2007) 'PorA variable antigenic regions VR1, VR2, and VR3 of Neisseria meningitidis serogroups B and C isolated in Brazil from 1999 to 2004', *Infection and Immunity*, 75(7), pp. 3683–3685.

Finotello, F. and Di Camillo, B. (2015) 'Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis', *Briefings in Functional Genomics*, 14(2), pp. 130–142.

Fitch, W. M., Bush, R. M., Bender, C. A. and Cox, N. J. (1997) 'Long term trends in the evolution of H(3) HA1 human influenza type A', *Proceeding of National Academy of USA*, 94(July), pp. 7712–7718.

Fitch, W. M., Leiter, J. M., Li, X. Q. and Palese, P. (1991) 'Positive Darwinian evolution in human influenza A viruses.', *Proceedings of the National Academy of Sciences of the United States of America*, 88(May), pp. 4270–4274.

Fulton, B. O., Sachs, D., Beaty, S. M., Won, S. T., Lee, B., Palese, P. and Heaton, N. S. (2015) 'Mutational Analysis of Measles Virus Suggests Constraints on Antigenic Variation of the Glycoproteins', *Cell Reports*, 11(9), pp. 1331–1338.

Funk, S., Bansal, S., Bauch, C. T., Eames, K. T. D., Edmunds, W. J., Galvani, A. P. and Klepac, P. (2015) 'Nine challenges in incorporating the dynamics of behaviour in infectious diseases models', *Epidemics*, 10, pp. 21–25.

Futse, J. E., Brayton, K. a, Dark, M. J., Knowles, D. P. and Palmer, G. H. (2008) 'Superinfection as a driver of genomic diversification in antigenically variant pathogens.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(6), pp. 2123–2127.

Gatton, M. L. and Cheng, Q. (2004) 'Modeling the Development of Acquired Clinical Immunity to Plasmodium falciparum Malaria', *Infection and Immunity*, 72(11), pp. 6538–6545.

Gianola, D., Qanbari, S. and Simianer, H. (2013) 'An evaluation of a novel estimator of linkage disequilibrium', *Heredity*. Nature Publishing Group, 111(4), pp. 275–285.

Girard, M. P., Preziosi, M. P., Aguado, M. T. and Kieny, M. P. (2006) 'A review of vaccine research and development: Meningococcal disease', *Vaccine*, 24(22), pp.

4692–4700.

Gloor, G. B., Martin, L. C., Wahl, L. M. and Dunn, S. D. (2005) 'Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions', *Biochemistry*, 44(19), pp. 7156–7165.

Gog, J. R. and Grenfell, B. T. (2002) 'Dynamics and selection of many-strain pathogens.', *Proceedings of the National Academy of Sciences of the United States of America*, 99(26), pp. 17209–14.

Gomes, M. G. M., Medley, G. F. and Nokes, D. J. (2002) 'On the determinants of population structure in antigenically diverse pathogens.', *Proceedings. Biological sciences / The Royal Society*, 269(1488), pp. 227–233.

Gosling, R. and von Seidlein, L. (2016) 'The Future of the RTS,S/AS01 Malaria Vaccine: An Alternative Development Plan', *PLoS Medicine*, 13(4), pp. 1–6.

Grassly, N. C. and Fraser, C. (2008) 'Mathematical models of infectious disease transmission', *Nature Reviews Microbiology*, 6(6), pp. 477–487.

Griffin, J. T., Hollingsworth, T. D., Okell, L. C., Churcher, T. S., White, M., Hinsley, W., Bousema, T., Drakeley, C. J., Ferguson, N. M., Basáñez, M.-G. and Ghani, A. C. (2010) 'Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies.', *PLoS medicine*, 7(8), p. e1000324.

Groce, N. E., Banks, L. M. and Stein, M. A. (2014) 'Surviving polio in a post-polio world', *Social Science and Medicine*, 107(2014), pp. 171–178.

Gu, S. Y., Huang, T. M., Ruan, L., Miao, Y. H., Lu, H., Chu, C. M., Motz, M. and Wolf, H. (1995) 'First EBV vaccine trial in humans using recombinant vaccinia virus expressing the major membrane antigen.', *Developments in biological standardization*, 84, pp. 171–177.

Guerin, P. J., Olliaro, P., Sundar, S., Boelaert, M., Croft, S. L., Desjeux, P., Wasunna, M. K. and Bryceson, A. D. M. (2002) 'Visceral leishmaniasis: Current status of control, diagnosis, and treatment, and a proposed research and development agenda', *Lancet Infectious Diseases*, pp. 494–501.

Gupta, S., Ferguson, N. and Anderson, R. (1998) 'Chaos, persistence, and evolution

of strain structure in antigenically diverse infectious agents.', *Science (New York, N.Y.)*, 280(5365), pp. 912–915.

Gupta, S., Maiden, M., Feavers, I. and Nee, S. (1996) 'The maintenance of strain structure in populations of recombining infectious agents', *Nature medicine*, 2(4), pp. 437–442.

Hanboonkunupakarn, B. and White, N. J. (2016) 'The threat of artemisinin resistant malaria in Southeast Asia', *Travel Medicine and Infectious Disease*, 14(6), pp. 548–550.

Hartlage, A. S., Liu, T., Patton, J. T., Garman, S. L., Zhang, X., Kurt, H., Lozanski, G., Lustberg, M. E., Caligiuri, M. A. and Baiocchi, R. A. (2015) 'The Epstein-Barr virus lytic protein BZLF1 as a candidate target antigen for vaccine development.', *Cancer immunology research*, 3(7), pp. 787–795.

Head, M. G., Goss, S., Gelister, Y., Alegana, V., Brown, R. J., Clarke, S. C., Fitchett, J. R. A., Atun, R., Scott, J. A. G., Newell, M. L., Padmadas, S. S. and Tatem, A. J. (2017) 'Global funding trends for malaria research in sub-Saharan Africa: a systematic analysis', *The Lancet Global Health*, 5(8), pp. e772–e781.

Hemelaar, J. (2012) 'The origin and diversity of the HIV-1 pandemic', *Trends in Molecular Medicine*, 18(3), pp. 182–192.

Henikoff, S. and Henikoff, J. G. (1992) 'Amino acid substitution matrices from protein blocks.', *Proceedings of the National Academy of Sciences*, 89(22), pp. 10915–10919.

Hill, A. V. S. (2006) 'Pre-erythrocytic malaria vaccines: towards greater efficacy.', *Nature reviews. Immunology*, 6(1), pp. 21–32.

Hill, A. V. S. (2011) 'Vaccines against malaria.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 366(1579), pp. 2806–14.

Hinnebusch, J., Barbour, A. G., Restrepo, B. I. and Schwan, T. G. (1998) 'Population structure of the relapsing fever spirochete Borrelia hermsii as indicated by polymorphism of two multigene families that encode immunogenic outer surface lipoproteins', *Infection and Immunity*, 66(2), pp. 432–440.

Hjalgrim, H., Friborg, J. and Melbye, M. (2007) 'The epidemiology of EBV and its association with malignant disease', in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*. Cambridge: Cambridge University Press.

Hodgson, S. H., Ewer, K. J., Bliss, C. M., Edwards, N. J., Rampling, T., Anagnostou, N. a., De Barra, E., Havelock, T., Bowyer, G., Poulton, I. D., De Cassan, S., Longley, R., Illingworth, J. J., Douglas, A. D., Mange, P. B., Collins, K. A., Roberts, R., Gerry, S., Berrie, E., Moyle, S., Colloca, S., Cortese, R., Sinden, R. E., Gilbert, S. C., Bejon, P., Lawrie, A. M., Nicosia, A., Faust, S. N. and Hill, A. V. S. (2014) 'Evaluation of the Efficacy of ChAd63-MVA Vectored Vaccines Expressing Circumsporozoite Protein and ME-TRAP Against Controlled Human Malaria Infection in Malaria-Naive Individuals', *Journal of Infectious Diseases*, 211(7), pp. 1076–1086.

Hui, E. P., Taylor, G. S., Jia, H., Ma, B. B. Y., Chan, S. L., Ho, R., Wong, W. L., Wilson, S., Johnson, B. F., Edwards, C., Stocken, D. D., Rickinson, A. B., Steven, N. M. and Chan, A. T. C. (2013) 'Phase I trial of recombinant modified vaccinia ankara encoding Epstein-Barr viral tumor antigens in nasopharyngeal carcinoma patients', *Cancer Research*, 73(6), pp. 1676–1688.

Hui, E. P., Taylor, G. S., Jia, H., Ma, B. B. Y., Chan, S. L., Ho, R., Wong, W. L., Wilson, S., Johnson, B. F., Edwards, C., Stocken, D. D., Rickinson, A. B., Steven, N. M. and Chan, A. T. C. (2013) 'Phase I trial of recombinant modified vaccinia ankara encoding Epstein-Barr viral tumor antigens in nasopharyngeal carcinoma patients', *Cancer Research*, 73(6), pp. 1676–1688.

Humphris-narayanan, E., Akiva, E., Varela, R. and Kortemme, T. (2012) 'Prediction of Mutational Tolerance in HIV-1 Protease and Reverse Transcriptase Using Flexible Backbone Protein Design', 8(8).

Hutchings, C. L., Birkett, A. J., Moore, A. C. and Hill, A. V. S. (2007) 'Combination of protein and viral vaccines induces potent cellular and humoral immune responses and enhanced protection from murine malaria challenge.', *Infection and immunity*, 75(12), pp. 5819–26.

IEDB - National Institute of Allergy and Infectious Diseases (2017) *IEDB Analysis Resource References*.

Ishizuka, A. S., Lyke, K. E., DeZure, A., Berry, A. A., Richie, T. L., Seder, R. A., *et al.*

(2016) ‘Protection against malaria at 1 year and immune correlates following PfSPZ vaccination’, *Nature Medicine*, 22(6), pp. 614–623.

Israel, B. F. and Kenney, S. C. (2005) ‘EBV lytic infection’, in Robertson, E. S. (ed.) *Epstein-Barr virus*. Norfolk, England: Caister Academic Press, pp. 571–611.

Iwasa, Y., Michor, F. and Nowak, M. A. (2004) ‘Stochastic Tunnels in Evolutionary Dynamics’, *Genetics*, 166(3), pp. 1571–1579.

Jodar, L., Feavers, I. M., Salisbury, D. and Granoff, D. M. (2002) ‘Development of vaccines against meningococcal disease’, *The Lancet*, 359(9316), pp. 1499–1508.

Johnson, D. R., Kaplan, E. L., VanGheem, A., Facklam, R. R. and Beall, B. (2006) ‘Characterization of group A streptococci (Streptococcus pyogenes): Correlation of M-protein and emm-gene type with T-protein agglutination pattern and serum opacity factor’, *Journal of Medical Microbiology*, 55(2), pp. 157–164.

Jolley, K. A. and Maiden, M. C. (2010) ‘BIGSdb: Scalable analysis of bacterial genome variation at the population level’, *BMC Bioinformatics*, 11(1), p. 595.

Kaplan, J. (2008) ‘The end of the adaptive landscape metaphor?’, *Biology and Philosophy*, 23(5), pp. 625–638.

Kaye, K. M., Izumi, K. M. and Kieff, E. (1993) ‘Epstein-Barr virus latent membrane protein 1 is essential for B-lymphocyte growth transformation.’, *Proceedings of the National Academy of Sciences of the United States of America*, 90(19), pp. 9150–4.

Keeling, M. J. and Danon, L. (2009) ‘Mathematical modelling of infectious diseases’, *British Medical Bulletin*, 92(1), pp. 33–42.

Kennedy, G., Komano, J. and Sugden, B. (2003) ‘Epstein-Barr virus provides a survival factor to Burkitt’s lymphomas’, *Proceedings of the National Academy of Sciences*, 100(24), pp. 14269–14274.

Kermack, W. O. and McKendrick, A. G. (1927) ‘A Contribution to the Mathematical Theory of Epidemics’, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, pp. 700–721.

Kester, K. E., Cummings, J. F., Ofori-Anyinam, O., Ockenhouse, C. F., Krzych, U., Moris, P., Schwenk, R., Nielsen, R. A., Debebe, Z., Pinelis, E., Juompan, L., Williams,

J., Dowler, M., Stewart, V. A., Wirtz, R. A., Dubois, M., Lievens, M., Cohen, J., Ballou, W. R. and Heppner, Jr., D. G. (2009) 'Randomized, Double-Blind, Phase 2a Trial of Falciparum Malaria Vaccines RTS,S/AS01B and RTS,S/AS02A in Malaria-Naive Adults: Safety, Efficacy, and Immunologic Associates of Protection', *The Journal of Infectious Diseases*, 200(3), pp. 337–346.

Kryazhimskiy, S., Dieckmann, U., Levin, S. A. and Dushoff, J. (2007) 'On State-Space Reduction in Multi-Strain Pathogen Models, with an Application to Antigenic Drift in Influenza A', *PLoS Computational Biology*, 3(8), p. e159.

Kucharski, A. J., Andreasen, V. and Gog, J. R. (2016) 'Capturing the dynamics of pathogens with many strains', *Journal of Mathematical Biology*, 72(1–2), pp. 1–24.

Kuhn, H. W. (1955) 'The Hungarian method for the assignment problem', *Naval Research Logistics Quarterly*, 2(1–2), pp. 83–97.

Kullback, S. (1959) *Information Theory and Statistics*. New York: Wiley.

Kusi, K. A., Faber, B. W., van der Eijk, M., Thomas, A. W., Kocken, C. H. and Remarque, E. J. (2011) 'Immunization with different PfAMA1 alleles in sequence induces clonal imprint humoral responses that are similar to responses induced by the same alleles as a vaccine cocktail in rabbits', *Malaria Journal*, 10(1), p. 40.

Kwok, H., Wu, C. W., Palser, A. L., Kellam, P., Sham, P. C., Kwong, D. L. W. and Chiang, A. K. S. (2014) 'Genomic Diversity of Epstein-Barr Virus Genomes Isolated from Primary Nasopharyngeal Carcinoma Biopsy Samples', *Journal of Virology*, 88(18), pp. 10662–10672.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S. and Ussery, D. W. (2015) 'Insights from 20 years of bacterial genome sequencing.', *Functional & integrative genomics*, 15(2), pp. 141–61.

Landais, E., Saulquin, X. and Houssaint, E. (2005) 'The human T cell immune response to Epstein-Barr virus.', *The International journal of developmental biology*, 49(2–3), pp. 285–292.

Lande, R. (2007) 'Expected relative fitness and the adaptive topography of fluctuating selection', *Evolution*, 61(8), pp. 1835–1846.

Laurens, M. B., Kouriba, B., Bergmann-Leitner, E., Angov, E., Coulibaly, D., Diarra, I., Daou, M., Niangaly, A., Blackwelder, W. C., Wu, Y., Cohen, J., Ballou, W. R., Vekemans, J., Lanar, D. E., Dutta, S., Diggs, C., Soisson, L., Heppner, D. G., Doumbo, O. K., Plowe, C. V. and Thera, M. A. (2017) 'Strain-specific Plasmodium falciparum growth inhibition among Malian children immunized with a blood-stage malaria vaccine', *PLoS ONE*, 12(3), pp. 1–13.

De Leenheer, P. and Pilyugin, S. S. (2008) 'Immune response to a malaria infection: Properties of a mathematical model', *Journal of Biological Dynamics*, 2(2), pp. 102–120.

Lewontin, R. C. (1988) 'On measures of gametic disequilibrium', *Genetics*, 120(3), pp. 849–852.

Lewontin, R. C. and Kojima, K. (1960) 'The Evolutionary Dynamics of Complex Polymorphisms', *Evolution*, 14(4), p. 458.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079.

Li, W., Shi, W., Qiao, H., Ho, S. Y. W., Luo, A., Zhang, Y. and Zhu, C. (2011) 'Positive selection on hemagglutinin and neuraminidase genes of H1N1 influenza viruses', *Virology Journal*, 8, pp. 1–9.

Lim, S. S., Yang, W., Krishnarjuna, B., Kannan Sivaraman, K., Chandrashekaran, I. R., Kass, I., Macraild, C. A., Devine, S. M., Debono, C. O., Anders, R. F., Scanlon, M. J., Scammells, P. J., Norton, R. S. and McGowan, S. (2014) 'Structure and dynamics of apical membrane antigen 1 from plasmodium falciparum FVO', *Biochemistry*, 53(46), pp. 7310–7320.

Lin, C. L., Lo, W. F., Lee, T. H., Ren, Y., Hwang, S. L., Cheng, Y. F., Chen, C. L., Chang, Y. Sen, Lee, S. P., Rickinson, A. B. and Tam, P. K. H. (2002) 'Immunization with Epstein-Barr Virus (EBV) peptide-pulsed dendritic cells induces functional CD8+ T-cell immunity and may lead to tumor regression in patients with EBV-positive nasopharyngeal carcinoma', *Cancer Research*, 62(23), pp. 6952–6958.

Lipsitch, M. and O'Hagan, J. J. (2007) 'Patterns of antigenic diversity and the

mechanisms that maintain them', *Journal of The Royal Society Interface*, 4(16), pp. 787–802.

Little, D. Y. and Chen, L. (2009) 'Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution', *PLoS ONE*, 4(3).

Liu, J., Yang, B., Cheung, W. K. and Yang, G. (2012) 'Malaria transmission modelling: a network perspective', *Infectious Diseases of Poverty*, 1(1), p. 11.

Lourenço, J. and Recker, M. (2014) 'The 2012 Madeira Dengue Outbreak: Epidemiological Determinants and Future Epidemic Potential.', *PLoS neglected tropical diseases*, 8(8), p. e3083.

Lourenço, J., Watkins, E. R., Obolski, U., Peacock, S. J., Morris, C., Maiden, M. C. J. and Gupta, S. (2017) 'Lineage structure of Streptococcus pneumoniae may be driven by immune selection on the groEL heat-shock protein', *Scientific Reports*, 7(1), pp. 1–11.

Lourenço, J., Wikramaratna, P. S. and Gupta, S. (2015) 'MANTIS: an R package that simulates multilocus models of pathogen evolution.', *BMC bioinformatics*, 16(1), p. 176.

Lu, F., Culleton, R., Zhang, M., Ramaprasad, A., von Seidlein, L., Zhou, H., Zhu, G., Tang, J., Liu, Y., Wang, W., Cao, Y., Xu, S., Gu, Y., Li, J., Zhang, C., Gao, Q., Menard, D., Pain, A., Yang, H., Zhang, Q. and Cao, J. (2017) 'Emergence of Indigenous Artemisinin-Resistant Plasmodium falciparum in Africa', *New England Journal of Medicine*, 376(10), pp. 991–993.

Luijkx, T. A., Van Dijken, H., Hamstra, H. J., Kuipers, B., Van Der Ley, P., Van Alphen, L. and Van den Dobbelsteen, G. (2003) 'Relative Immunogenicity of PorA Subtypes in a Multivalent Neisseria meningitidis Vaccine Is Not Dependent on Presentation Form', *Infection and Immunity*, 71(11), pp. 6367–6371.

Lyke, K. E., Ishizuka, A. S., Berry, A. A., Chakravarty, S., DeZure, A., Enama, M. E., James, E. R., Billingsley, P. F., Gunasekera, A., Manoj, A., Li, M., Ruben, A. J., Li, T., Eappen, A. G., Stafford, R. E., KC, N., Murshedkar, T., Mendoza, F. H., Gordon, I. J., Zephir, K. L., Holman, L. A., Plummer, S. H., Hendel, C. S., Novik, L., Costner, P. J.

M., Saunders, J. G., Berkowitz, N. M., Flynn, B. J., Nason, M. C., Garver, L. S., Laurens, M. B., Plowe, C. V., Richie, T. L., Graham, B. S., Roederer, M., Sim, B. K. L., Ledgerwood, J. E., Hoffman, S. L. and Seder, R. A. (2017) 'Attenuated PfSPZ Vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection', *Proceedings of the National Academy of Sciences*, 114(10), pp. 2711–2716.

Lyke, K. E., Laurens, M., Adams, M., Billingsley, P. F., Richman, A., Loyevsky, M., Chakravarty, S., Plowe, C. V, Sim, B. K. L., Edelman, R. and Hoffman, S. L. (2010) 'Plasmodium falciparum malaria challenge by the bite of aseptic Anopheles stephensi mosquitoes: results of a randomized infectivity trial.', *PloS one*, 5(10), p. e13490.

Macdonald, G. (1952) 'The analysis of equilibrium in malaria.', *Tropical diseases bulletin*, 49(9), pp. 813–829.

Mandal, S., Sarkar, R. R. and Sinha, S. (2011) 'Mathematical models of malaria--a review.', *Malaria journal*, 10(1), p. 202.

Maree, A. F., Keulen, W., Boucher, C. A. and De Boer, R. J. (2000) 'Estimating relative fitness in viral competition experiments', *J Virol*, 74(23), pp. 11067–11072.

Marsay, L., Dold, C., Green, C. A., Rollier, C. S., Norheim, G., Sadarangani, M., Shanyinde, M., Brehony, C., Thompson, A. J., Sanders, H., Chan, H., Haworth, K., Derrick, J. P., Feavers, I. M., Maiden, M. C. and Pollard, A. J. (2015) 'A novel meningococcal outer membrane vesicle vaccine with constitutive expression of FetA: A phase I clinical trial', *Journal of Infection*. Elsevier Ltd, 71(3), pp. 326–337.

Martin, L. C., Gloor, G. B., Dunn, S. D. and Wahl, L. M. (2005) 'Using information theory to search for co-evolving residues in proteins', *Bioinformatics*, 21(22), pp. 4116–4124.

Marx, A. and Alian, A. (2015) 'The Road Less Traveled: HIV's Use of Alternative Routes through Cellular Pathways', *Journal of Virology*, 89(10), pp. 5204–5212.

Materu, S., Cox, H. S., Isaakidis, P., Baruani, B., Ogaro, T. and Caugant, D. A. (2007) 'Serogroup X in meningococcal disease, Western Kenya [3]', *Emerging Infectious Diseases*, 13(6), pp. 944–945.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P.

and Lopez, R. (2013) ‘Analysis Tool Web Services from the EMBL-EBI.’, *Nucleic acids research*, 41(Web Server issue), pp. 597–600.

Medica, D. L. and Sinnis, P. (2005) ‘Quantitative dynamics of Plasmodium yoelii sporozoite transmission by infected anopheline mosquitoes.’, *Infection and immunity*, 73(7), pp. 4363–9.

Meis, J. F., Rijntjes, P. J., Verhave, J. P., Ponnudurai, T., Hollingdale, M. R., Smith, J. E., Sinden, R. E., Jap, P. H., Meuwissen, J. H. and Yap, S. H. (1986) ‘Fine structure of the malaria parasite Plasmodium falciparum in human hepatocytes in vitro.’, *Cell and tissue research*, 244(2), pp. 345–50.

Meis, J. F., Verhave, J. P., Meuwissen, J. H., Jap, P. H., Princen, H. M. and Yap, S. H. (1984) ‘Fine structure of Plasmodium berghei exoerythrocytic forms in cultured primary rat hepatocytes.’, *Cell biology international reports*, 8(9), pp. 755–65.

Menard, D. and Dondorp, A. (2017) ‘Antimalarial Drug Resistance: A Threat to Malaria Elimination’, *Cold Spring Harbor Perspectives in Medicine*, 7(7), p. a025619.

Metcalf, C. J. E., Birger, R. B., Funk, S., Kouyos, R. D., Lloyd-Smith, J. O. and Jansen, V. A. A. (2015) ‘Five challenges in evolution and infectious diseases’, *Epidemics*, 10, pp. 40–44.

Metcalf, C. J. E., Edmunds, W. J. and Lessler, J. (2015) ‘Six challenges in modelling for public health policy’, *Epidemics*, 10, pp. 93–96.

Meyohas, M. C., Maréchal, V., Desire, N., Bouillie, J., Frottier, J. and Nicolas, J. C. (1996) ‘Study of mother-to-child Epstein-Barr virus transmission by means of nested PCRs.’, *Journal of virology*, 70(10), pp. 6816–9.

Middleton, T. and Sugden, B. (1994) ‘Retention of plasmid DNA in mammalian cells is enhanced by binding of the Epstein-Barr virus replication protein EBNA1.’, *Journal of virology*, 68(6), pp. 4067–4071.

Migueles, S. A. and Connors, M. (2015) ‘Success and failure of the cellular immune response against HIV-1’, *Nat Immunol*, 16(6), pp. 563–570.

Minayev, P. and Ferguson, N. (2009) ‘Improving the realism of deterministic multi-

strain models: implications for modelling influenza A', *Journal of The Royal Society Interface*, 6(35), pp. 509–518.

Miura, K. (2016) 'Progress and prospects for blood-stage malaria vaccines.', *Expert review of vaccines*. Taylor & Francis, 15(6), pp. 765–81.

Morens, D. M., Holmes, E. C., Davis, A. S. and Taubenberger, J. K. (2011) 'Global rinderpest eradication: Lessons learned and why humans should celebrate too', *Journal of Infectious Diseases*, 204(4), pp. 502–505.

Moutschen, M., Philippe, L., Sokal, E. M., Mazzu, P., Bollen, A., Denamur, F., Peeters, P., Dubin, G. and Denis, M. (2007) 'Phase I / II studies to evaluate safety and immunogenicity of a recombinant gp350 Epstein – Barr virus vaccine in healthy adults Mich ``', 25, pp. 4697–4705.

Mugalavai, E. M., Kipkorir, E. C., Raes, D. and Rao, M. S. (2008) 'Analysis of rainfall onset, cessation and length of growing season for western Kenya', *Agricultural and Forest Meteorology*, 148(6–7), pp. 1123–1135.

Mugenyi, L., Abrams, S. and Hens, N. (2017) 'Estimating age-time-dependent malaria force of infection accounting for unobserved heterogeneity', *Epidemiology and Infection*, 145(12), pp. 2545–2562.

Mullen, G. E. D., Ellis, R. D., Miura, K., Malkin, E., Nolan, C., Hay, M., Fay, M. P., Saul, A., Zhu, D., Rausch, K., Moretz, S., Zhou, H., Long, C. A., Miller, L. H. and Treanor, J. (2008) 'Phase 1 trial of AMA1-C1/Alhydrogel plus CPG 7909: An asexual blood-stage vaccine for Plasmodium falciparum malaria', *PLoS ONE*, 3(8).

Murray, C. J. L., Rosenfeld, L. C., Lim, S. S., Andrews, K. G., Foreman, K. J., Haring, D., Fullman, N., Naghavi, M., Lozano, R. and Lopez, A. D. (2012) 'Global malaria mortality between 1980 and 2010: A systematic analysis', *The Lancet*. Elsevier Ltd, 379(9814), pp. 413–431.

Mustonen, V. and La, M. (2009) 'From fitness landscapes to seascapes : non-equilibrium dynamics of selection and adaptation', (February), pp. 111–119.

Newby, G., Hwang, J., Koita, K., Chen, I., Greenwood, B., Von Seidlein, L., Shanks, G. D., Slutsker, L., Kachur, S. P., Wegbreit, J., Ippolito, M. M., Poirot, E. and Gosling,

R. (2015) ‘Review of mass drug administration for malaria and its operational challenges’, *American Journal of Tropical Medicine and Hygiene*, 93(1), pp. 125–134.

O’Neill, J. (2014) ‘Review on Antimicrobial Resistance. Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations, 2014’, 4(December).

Ogembo, J. G., Kannan, L., Ghiran, I., Nicholson-Weller, A., Finberg, R. W., Tsokos, G. C. and Fingeroth, J. D. (2013) ‘Human Complement Receptor Type 1/CD35 Is an Epstein-Barr Virus Receptor’, *Cell Reports*. Elsevier, 3(2), pp. 371–385.

Okoye, A. A. and Picker, L. J. (2013) ‘CD4+ T-Cell Depletion In Hiv Infection: Mechanisms Of Immunological Failure’, *Immunological Reviews*, 254(1), pp. 54–64.

Old, L. J., Boyse, E. A., Oettgen, H. F., Harven, E. De, Geering, G., Williamson, B. and Clifford, P. (1966) ‘Precipitating Antibody In Human Serum To An Antigen Present In Cultured Burkitt’s Lymphoma Cells’, *Proceeding of National Academy of USA*, 56(6), pp. 1699–1704.

Olotu, A., Fegan, G., Wambua, J., Nyangweso, G., Awuondo, K. O., Leach, A., Lievens, M., Leboulleux, D., Njuguna, P., Peshu, N., Marsh, K. and Bejon, P. (2013) ‘Four-year efficacy of RTS,S/AS01E and its interaction with malaria exposure.’, *The New England journal of medicine*, 368(12), pp. 1111–20.

Olotu, A., Fegan, G., Wambua, J., Nyangweso, G., Leach, A., Lievens, M., Kaslow, D. C., Njuguna, P., Marsh, K. and Bejon, P. (2016) ‘Seven-Year Efficacy of RTS,S/AS01 Malaria Vaccine among Young African Children.’, *The New England journal of medicine*, 374(26), pp. 2519–29.

Osier, F. H., Mackinnon, M. J., Crosnier, C., Fegan, G., Kamuyu, G., Wanaguru, M., Ogada, E., McDade, B., Rayner, J. C., Wright, G. J. and Marsh, K. (2014) ‘New antigens for a multicomponent blood-stage malaria vaccine’, *Science Translational Medicine*, 6(247), p. 247ra102-247ra102.

Ouattara, A., Mu, J., Takala-Harrison, S., Saye, R., Sagara, I., Dicko, A., Niangaly, A., Duan, J., Ellis, R. D., Miller, L. H., Su, X., Plowe, C. V and Doumbo, O. K. (2010) ‘Lack of allele-specific efficacy of a bivalent AMA1 malaria vaccine.’, *Malaria journal*, 9(1), p. 175.

Penman, B. S., Ashby, B., Buckee, C. O. and Gupta, S. (2013) 'Pathogen selection drives nonoverlapping associations between HLA loci', *Proc Natl Acad Sci U S A*, 110(48), pp. 19645–50.

Penny, M. A., Verity, R., Bever, C. A., Sauboin, C., Galactionova, K., Flasche, S., White, M. T., Wenger, E. A., Van De Velde, N., Pemberton-Ross, P., Griffin, J. T., Smith, T. A., Eckhoff, P. A., Muhib, F., Jit, M. and Ghani, A. C. (2016) 'Public health impact and cost-effectiveness of the RTS,S/AS01 malaria vaccine: A systematic comparison of predictions from four mathematical models', *The Lancet*, 387(10016), pp. 367–375.

Platt, A., Obala, A. A., MacIntyre, C., Otsyula, B. and Meara, W. P. O. (2018) 'Dynamic malaria hotspots in an open cohort in western Kenya', *Scientific Reports*, 8(1), pp. 1–11.

Plotkin, J. B. and Dushoff, J. (2003) 'Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus', *Proceedings of the National Academy of Sciences*, 100(12), pp. 7152–7157.

Pollard, A. J. (2004) 'Global Epidemiology of Meningococcal Disease and Vaccine Efficacy', *The Pediatric Infectious Disease Journal*, 23(12), pp. 274–279.

Pollard, A. J., Galassini, R., Rouppe van der Voort, E. M., Hibberd, M., Booy, R., Langford, P., Nadel, S., Ison, C., Kroll, J. S., Poolman, J. and Levin, M. (1999) 'Cellular immune responses to Neisseria meningitidis in children.', *Infection and immunity*, 67(5), pp. 2452–63.

Polley, S. D., Tetteh, K. K. A., Lloyd, J. M., Akpogheneta, O. J., Greenwood, B. M., Bojang, K. A. and Conway, D. J. (2007) '*Plasmodium falciparum* Merozoite Surface Protein 3 Is a Target of Allele-Specific Immunity and Alleles Are Maintained by Natural Selection', *The Journal of Infectious Diseases*, 195(2), pp. 279–287.

Precopio, M. L., Sullivan, J. L., Willard, C., Somasundaran, M. and Luzuriaga, K. (2003) 'Differential kinetics and specificity of EBV-specific CD4+ and CD8+ T cells during primary infection.', *Journal of Immunology*, 170(5), pp. 2590–8.

Pudney, V. A., Leese, A. M., Rickinson, A. B. and Hislop, A. D. (2005) 'CD8 + immunodominance among Epstein-Barr virus lytic cycle antigens directly reflects the

efficiency of antigen presentation in lytically infected cells', *The Journal of Experimental Medicine*, 201(3), pp. 349–360.

Rambaut, A. and Grassly, N. C. (1997) 'Seq-Gen : an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees', 13(3), pp. 235–238.

Recker, M., Nee, S., Bull, P. C., Kinyanjui, S., Marsh, K., Newbold, C. and Gupta, S. (2004) 'Transient cross-reactive immune responses can orchestrate antigenic variation in malaria', *Nature*, 429(6991), pp. 555–8.

Recker, M., Pybus, O. G., Nee, S. and Gupta, S. (2007) 'The generation of influenza outbreaks by a network of host immune responses against a limited set of antigenic types.', *Proceedings of the National Academy of Sciences of the United States of America*, 104(18), pp. 7711–7716.

Redd, A. D., Mullis, C. E., Serwadda, D., Kong, X., Martens, C., Ricklefs, S. M., Tobian, A. A. R., Xiao, C., Grabowski, M. K., Nalugoda, F., Kigozi, G., Laeyendecker, O., Kagaayi, J., Sewankambo, N., Gray, R. H., Porcella, S. F., Wawer, M. J. and Quinn, T. C. (2012) 'The rates of HIV superinfection and primary HIV incidence in a general population in Rakai, Uganda', *Journal of Infectious Diseases*, 206(2), pp. 267–274.

Rees, L., Tizard, E. J., Morgan, A. J., Cubitt, W. D., Finerty, S., Oyewole-Eletu, T. a, Owen, K., Royed, C., Stevens, S. J., Shroff, R. C., Tanday, M. K., Wilson, a D., Middeldorp, J. M., Amlot, P. L. and Steven, N. M. (2009) 'A phase I trial of epstein-barr virus gp350 vaccine for children with chronic kidney disease awaiting transplantation.', *Transplantation*, 88(8), pp. 1025–1029.

Regules, J. A., Cicatelli, S. B., Bennett, J. W., Paolino, K. M., Twomey, P. S., Moon, J. E., Kathcart, A. K., Hauns, K. D., Komisar, J. L., Qabar, A. N., Davidson, S. A., Dutta, S., Griffith, M. E., Magee, C. D., Wojnarski, M., Livezey, J. R., Kress, A. T., Waterman, P. E., Jongert, E., Wille-Reece, U., Volkmuth, W., Emerling, D., Robinson, W. H., Lievens, M., Morelle, D., Lee, C. K., Yassin-Rajkumar, B., Weltzin, R., Cohen, J., Paris, R. M., Waters, N. C., Birkett, A. J., Kaslow, D. C., Ballou, W. R., Ockenhouse, C. F. and Vekemans, J. (2016) 'Fractional third and fourth dose of RTS,S/AS01 malaria candidate vaccine: A phase 2a controlled human malaria parasite infection and immunogenicity study', *Journal of Infectious Diseases*, 214(5),

pp. 762–771.

Reva, B., Antipin, Y. and Sander, C. (2011) 'Predicting the functional impact of protein mutations: Application to cancer genomics', *Nucleic Acids Research*, 39(17), pp. 37–43.

Rice, P., Longden, I. and Bleasby, A. (2000) 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics*, 16(1), pp. 276–277.

Richards, J. S. and Beeson, J. G. (2009) 'The future for blood-stage vaccines against malaria.', *Immunology and cell biology*, 87(5), pp. 377–90.

Rihn, S. J., Wilson, S. J., Loman, N. J., Alim, M., Bakker, S. E., Bhella, D., Gifford, R. J., Rixon, F. J. and Bieniasz, P. D. (2013) 'Extreme Genetic Fragility of the HIV-1 Capsid', *PLoS Pathogens*, 9(6).

Riley, E. M. and Stewart, V. A. (2013) 'Immune mechanisms in malaria: new insights in vaccine development.', *Nature medicine*, 19(2), pp. 168–78.

Roca-Feltrer, A., Schellenberg, J., Smith, L. and Carneiro, I. (2009) 'A simple method for defining malaria seasonality', *Malaria Journal*, 8(1), p. 276.

Rodriguez-Barraquer, I., Arinaitwe, E., Jagannathan, P., Boyle, M. J., Tappero, J., Muhindo, M., Kamya, M. R., Dorsey, G., Drakeley, C., Ssewanyana, I., Smith, D. L. and Greenhouse, B. (2016) 'Quantifying Heterogeneous Malaria Exposure and Clinical Protection in a Cohort of Ugandan Children', *Journal of Infectious Diseases*, 214(7), pp. 1072–1080.

Roestenberg, M., O'Hara, G. A., Duncan, C. J. A., Epstein, J. E., Edwards, N. J., Scholzen, A., van der Ven, A. J. A. M., Hermsen, C. C., Hill, A. V. S. and Sauerwein, R. W. (2012) 'Comparison of Clinical and Parasitological Data from Controlled Human Malaria Infection Trials', *PLoS ONE*, 7(6), p. e38434.

Romero, P. A. and Arnold, F. H. (2009) 'Exploring protein fitness landscapes by directed evolution', *Nature Reviews Molecular Cell Biology*, pp. 866–876.

Roskrow, M. A., Suzuki, N., Gan, Y., Sixbey, J. W., Ng, C. Y., Kimbrough, S., Hudson, M., Brenner, M. K., Heslop, H. E. and Rooney, C. M. (1998) 'Epstein-Barr virus (EBV)-specific cytotoxic T lymphocytes for the treatment of patients with EBV-

positive relapsed Hodgkin's disease', *Blood*, 91(8), pp. 2925–2934.

Rota, J. S., Heath, J. L., Rota, P. A., King, G. E., Celma, M. L., Carabaña, J., Fernandez-Muñoz, R., Brown, D., Jin, L. and Bellini, W. J. (1996) 'Molecular epidemiology of measles virus: identification of pathways of transmission and implications for measles elimination.', *The Journal of infectious diseases*, 173(1), pp. 32–7.

Rota, P. A., Brown, K., Mankertz, A., Santibanez, S., Shulga, S., Muller, C. P., Hübschen, J. M., Siqueira, M., Beirnes, J., Ahmed, H., Triki, H., Al-Busaidy, S., Dosseh, A., Byabamazima, C., Smit, S., Akoua-Koffi, C., Bwogi, J., Bukenya, H., Wairagkar, N., Ramamurty, N., Incomserb, P., Pattamadilok, S., Jee, Y., Lim, W., Xu, W., Komase, K., Takeda, M., Tran, T., Castillo-Solorzano, C., Chenoweth, P., Brown, D., Mulders, M. N., Bellini, W. J. and Featherstone, D. (2011) 'Global distribution of measles genotypes and measles molecular epidemiology', *Journal of Infectious Diseases*, 204(SUPPL. 1), pp. 514–523.

Rowe, M. and Zuo, J. (2010) 'Immune responses to Epstein-Barr virus: molecular interactions in the virus evasion of CD8+ T cell immunity', *Microbes and Infection*. russell 2004 pora, 12(3), pp. 173–181.

RTSS Clinical Trials Partnership (2014) 'Efficacy and Safety of the RTS,S/AS01 Malaria Vaccine during 18 Months after Vaccination: A Phase 3 Randomized, Controlled Trial in Children and Young Infants at 11 African Sites', *PLoS Medicine*, 11(7), p. e1001685.

RTSS Clinical Trials Partnership (2015) 'Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial', *The Lancet*. Elsevier Ltd, 386(9988), pp. 31–45.

Ruiss, R., Jochum, S., Wanner, G., Reisbach, G., Hammerschmidt, W. and Zeidler, R. (2011) 'A Virus-Like Particle-Based Epstein-Barr Virus Vaccine', *Journal of Virology*, 85(24), pp. 13105–13113.

Russell, J. E., Jolley, K. A., Feavers, I. M., Maiden, M. C. J. and Suker, J. (2004) 'PorA Variable Regions of Neisseria meningitidis', *Emerging Infectious Diseases*, 10(4), pp. 674–678.

Sæther, B. E. and Engen, S. (2015) 'The concept of fitness in fluctuating environments', *Trends in Ecology and Evolution*. Elsevier, pp. 273–281.

Sagara, I., Dicko, A., Ellis, R. D., Fay, M. P., Diawara, S. I., Assadou, M. H., Sissoko, M. S., Kone, M., I, A., Saye, R., Guindo, M. a, Kante, O., Niambele, M. B., Miura, K., Mullen, G. E. D., Pierce, M., Martin, L. B., Diallo, D. a, Doumbo, O. K., Miller, L. H. and Saul, A. (2009) 'A Randomized Controlled Phase 2 Trial of the Blood Stage AMA1- C1/Alhydrogel Malaria Vaccine in Children in Mali', *Vaccine*, 27(23), pp. 3090–3098.

Sanchez, A., Geisbert, T. and Feldmann, H. (2006) 'Filoviridae: Marburg and Ebola viruses', in Knipe, D. and Howley, P. (eds) *Fields virology*. Philadelphia: Lippincott Williams & Wilkins, pp. 1409–1448.

Sanderson, F., Andrews, L., Douglas, A. D., Hunt-Cooke, A., Bejon, P. and Hill, A. V. S. (2008) 'Blood-stage challenge for malaria vaccine efficacy trials: A pilot study with discussion of safety and potential value', *American Journal of Tropical Medicine and Hygiene*, 78(6), pp. 878–883.

Schwartz, L., Brown, G. V, Genton, B. and Moorthy, V. S. (2012) 'A review of malaria vaccine clinical projects based on the WHO rainbow table.', *Malaria journal*, 11(1), p. 11.

Seder, R. A., Chang, L.-J., Enama, M. E., Zephir, K. L., Sarwar, U. N., Gordon, I. J., Holman, L. A., James, E. R., Billingsley, P. F., Gunasekera, A., Richman, A., Chakravarty, S., Manoj, A., Velmurugan, S., Li, M., Ruben, A. J., Li, T., Eappen, A. G., Stafford, R. E., Plummer, S. H., Hendel, C. S., Novik, L., Costner, P. J. M., Mendoza, F. H., Saunders, J. G., Nason, M. C., Richardson, J. H., Murphy, J., Davidson, S. A., Richie, T. L., Sedegah, M., Sutamihardja, A., Fahle, G. A., Lyke, K. E., Laurens, M. B., Roederer, M., Tewari, K., Epstein, J. E., Sim, B. K. L., Ledgerwood, J. E., Graham, B. S. and Hoffman, S. L. (2013) 'Protection against malaria by intravenous immunization with a nonreplicating sporozoite vaccine.', *Science*, 341(6152), pp. 1359–65.

Sheehy, S. H., Spencer, A. J., Douglas, A. D., Sim, B. K. L., Longley, R. J., Edwards, N. J., Poulton, I. D., Kimani, D., Williams, A. R., Anagnostou, N. a, Roberts, R., Kerridge, S., Voysey, M., James, E. R., Billingsley, P. F., Gunasekera, A., Lawrie, A.

M., Hoffman, S. L. and Hill, A. V. S. (2013) 'Optimising Controlled Human Malaria Infection Studies Using Cryopreserved P. falciparum Parasites Administered by Needle and Syringe.', *PloS one*, 8(6), p. e65960.

Siettos, C. I. and Russo, L. (2013) 'Mathematical modeling of infectious disease dynamics', *Virulence*, 4(4), pp. 295–306.

Sivachandran, N., Dawson, C. W., Young, L. S., Liu, F.-F., Middeldorp, J. and Frappier, L. (2012) 'Contributions of the Epstein-Barr virus EBNA1 protein to gastric carcinoma.', *Journal of virology*, 86(1), pp. 60–8.

Slater, H. C., Griffin, J. T., Ghani, A. C. and Okell, L. C. (2016) 'Assessing the potential impact of artemisinin and partner drug resistance in sub-Saharan Africa', *Malaria Journal*, 15(1), pp. 1–11.

Smith, C., Tsang, J., Beagley, L., Chua, D., Lee, V., Li, V., Moss, D. J., Coman, W., Chan, K. H., Nicholls, J., Kwong, D. and Khanna, R. (2012) 'Effective treatment of metastatic forms of epstein-barr virus-associated nasopharyngeal carcinoma with a novel adenovirus-based adoptive immunotherapy', *Cancer Research*, 72(5), pp. 1116–1125.

Smith, D. L., Cohen, J. M., Chiyaka, C., Johnston, G., Gething, P. W., Gosling, R., Buckee, C. O., Laxminarayan, R., Hay, S. I. and Tatem, A. J. (2013) 'A sticky situation: the unexpected stability of malaria elimination', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1623), pp. 20120145–20120145.

Smith, D. L., Hay, S. I., Noor, A. M. and Snow, R. W. (2009) 'Predicting changing malaria risk after expanded insecticide-treated net coverage in Africa', *Trends in Parasitology*, 25(11), pp. 511–516.

Smith, D. L., McKenzie, F. E., Snow, R. W. and Hay, S. I. (2007) 'Revisiting the basic reproductive number for malaria and its implications for malaria control', *PLoS Biology*, 5(3), pp. 0531–0542.

Smith, T. A. (2008) 'Estimation of heterogeneity in malaria transmission by stochastic modelling of apparent deviations from mass action kinetics', *Malaria Journal*, 7, pp. 1–10.

Snow, R. W., Kibuchi, E., Karuri, S. W., Sang, G., Gitonga, C. W., Mwandawiro, C., Bejon, P. and Noor, A. M. (2015) 'Changing Malaria Prevalence on the Kenyan Coast since 1974: Climate, Drugs and Vector Control', *PLOS ONE*, 10(6), p. e0128792.

Snow, R. W. and Marsh, K. B. T.-A. in P. (2002) 'The consequences of reducing transmission of Plasmodium falciparum in Africa.', *Advances in Parasitology*, 52, pp. 235–264.

Snow, R. W., Omumbo, J. A., Lowe, B., Molyneux, C. S., Obiero, J.-O., Palmer, A., Weber, M. W., Pinder, M., Nahlen, B., Obonyo, C., Newbold, C., Gupta, S. and Marsh, K. (1997) 'Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa', *The Lancet*, 349(9066), pp. 1650–1654.

Sokal, E. M., Hoppenbrouwers, K., Vandermeulen, C., Moutschen, M., Léonard, P., Moreels, A., Haumont, M., Bollen, A., Smets, F. and Denis, M. (2007) 'Recombinant gp350 vaccine for infectious mononucleosis: a phase 2, randomized, double-blind, placebo-controlled trial to evaluate the safety, immunogenicity, and efficacy of an Epstein-Barr virus vaccine in healthy young adults.', *The Journal of infectious diseases*, 196(12), pp. 1749–53.

Speck, P., Haan, K. M. and Longnecker, R. (2000) 'Epstein–Barr Virus Entry into Cells', *Virology*, 277(1), pp. 1–5.

Speck, S. H. (2005) 'Regulation of EBV latency-associated gene expression', in Robertson, E. S. (ed.) *Epstein-Barr virus*. Caister Academic Press, pp. 403–427.

Spensley, K. J., Lourenco, J., Wikramaratna, P. S., Penman, B. S., Walker, A., Smith, A. L., Pybus, O. G., Jean, L. and Gupta, S. (2017) 'Reverse immunodynamics: a new method to identifying targets of protective immunity.', *bioRxiv*.

Straathof, K. C. M., Bollard, C. M., Popat, U., Huls, M. H., Lopez, T., Morriss, M. C., Gresik, M. V, Gee, A. P., Russell, H. V, Brenner, M. K., Rooney, C. M. and Heslop, H. E. (2005) 'Treatment of nasopharyngeal carcinoma with Epstein-Barr virus - specific T lymphocytes', *Therapy*, 105(5), pp. 1898–1904.

Sturm, A., Amino, R., van de Sand, C., Regen, T., Retzlaff, S., Rennenberg, A., Krueger, A., Pollok, J.-M., Menard, R. and Heussler, V. T. (2006) 'Manipulation of

host hepatocytes by the malaria parasite for delivery into liver sinusoids.', *Science*, 313(5791), pp. 1287–90.

Swigris, J. J., Berry, G. J., Raffin, T. A. and Kuschner, W. G. (2002) 'Lymphoid interstitial pneumonia: A narrative review', *Chest*, 122(6), pp. 2150–2164.

Takala-Harrison, S. and Laufer, M. K. (2015) 'Antimalarial drug resistance in Africa: key lessons for the future', *Annals of the New York Academy of Sciences*, 1342(1), pp. 62–67.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) 'MEGA6: Molecular evolutionary genetics analysis version 6.0', *Molecular Biology and Evolution*, 30(12), pp. 2725–2729.

Tan, L. C., Gudgeon, N., Annels, N. E., Hansasuta, P., O'Callaghan, C. a, Rowland-Jones, S., McMichael, a J., Rickinson, a B. and Callan, M. F. (1999) 'A Re-Evaluation of the Frequency of CD8+ T Cells Specific for EBV in Healthy Virus Carriers', *Journal of Immunology*, 162(3), pp. 1827–35.

Tappero, J. W., Lagos, R., Maldonado Ballesteros, A., Plikaytis, B., Williams, D., Dykes, J., Gheesling, L. L., Carlone, G. M., H?iby, E. A., Holst, J., N?kleby, H., Rosenqvist, E., Sierra, G., Campa, C., Sotolongo, F., Vega, J., Garcia, J., Herrera, P., Poolman, J. T. and Perkins, B. A. (1999) 'Immunogenicity of 2 Serogroup B Outer-Membrane Protein Meningococcal Vaccines', *JAMA*, 281(16), pp. 1520–27.

Taubenberger, J. K. and Morens, D. M. (2008) 'The Pathology of Influenza Virus Infections', *Annual Review of Pathology: Mechanisms of Disease*, 3(1), pp. 499–522.

Taylor, G. S., Jia, H., Harrington, K., Lee, L. W., Turner, J., Ladell, K., Price, D. A., Tanday, M., Matthews, J., Roberts, C., Edwards, C., McGuigan, L., Hartley, A., Wilson, S., Hui, E. P., Chan, A. T. C., Rickinson, A. B. and Steven, N. M. (2014) 'A recombinant modified vaccinia ankara vaccine encoding Epstein-Barr Virus (EBV) target antigens: A phase I trial in UK patients with EBV-positive cancer', *Clinical Cancer Research*, 20(19), pp. 5009–5022.

Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P. M., Rammensee, H. G., Schild, H. and Holzhütter, H. G. (2005) 'Modeling the MHC

class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding', *Cellular and Molecular Life Sciences*, 62(9), pp. 1025–1037.

The RTSS Clinical Trials Partnership (2011) 'First Results of Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Children', *New England Journal of Medicine*, 365(20), pp. 1863–1875.

Thera, M. A., Doumbo, O. K., Coulibaly, D., Laurens, M. B., Ouattara, A., Kone, A. K., Guindo, A. B., Traore, K., Traore, I., Kouriba, B., Diallo, D. A., Diarra, I., Daou, M., Dolo, A., Tolo, Y., Sissoko, M. S. M., Niangaly, A., Sissoko, M. S. M., Takala-Harrison, S., Lyke, K. E., Wu, Y., Blackwelder, W. C., Godeaux, O., Vekemans, J., Dubois, M.-C., Ballou, W. R., Cohen, J., Thompson, D., Dube, T., Soisson, L., Diggs, C. L., House, B., Lanar, D. E., Dutta, S., Heppner, D. G. and Plowe, C. V (2011) 'A field trial to assess a blood-stage malaria vaccine.', *The New England journal of medicine*, 365(11), pp. 1004–13.

Thorley-Lawson, D. A. and Gross, A. (2004) 'Persistence of the Epstein–Barr Virus and the Origins of Associated Lymphomas', *New England Journal of Medicine*, 350(13), pp. 1328–1337.

Tibayrenc, M. and Ayala, F. J. (2015) 'How clonal are Neisseria species? The epidemic clonality model revisited', *Proceedings of the National Academy of Sciences*, 112(29), pp. 8909–8913.

Tomkinson, B., Robertson, E. and Kieff, E. (1993) 'Epstein-Barr virus nuclear proteins EBNA-3A and EBNA-3C are essential for B-lymphocyte growth transformation.', *Journal of virology*, 67(4), pp. 2014–2025.

Traoré, B., Sangaré, B. and Traoré, S. (2017) 'A Mathematical Model of Malaria Transmission with Structured Vector Population and Seasonality', *Journal of Applied Mathematics*, 2017.

Tsurumi, T., Fujita, M. and Kudoh, A. (2005) 'Latent and lytic Epstein-Barr virus replication strategies', *Reviews in Medical Virology*, 15(1), pp. 3–15.

Tugizov, S. M., Berline, J. W. and Palefsky, J. M. (2003) 'Epstein-Barr virus infection of polarized tongue and nasopharyngeal epithelial cells.', *Nature medicine*, 9(3), pp.

307–314.

Veliov, V. M. and Widder, A. (2016) ‘Modelling and estimation of infectious diseases in a population with heterogeneous dynamic immunity’, *Journal of Biological Dynamics*, 10(1), pp. 457–476.

Viriyavejakul, P., Khachonsaksumet, V. and Punsawad, C. (2014) ‘Liver changes in severe Plasmodium falciparum malaria: histopathology, apoptosis and nuclear factor kappa B expression.’, *Malaria journal*, 13(1), p. 106.

De Visser, J. A. G. M. and Krug, J. (2014) ‘Empirical fitness landscapes and the predictability of evolution’, *Nature Reviews Genetics*, pp. 480–490.

Vreden, S., Verhave, J., Oettinger, T., Sauerwein, R. W. and Meuwissen, J. H. (1991) ‘Phase I clinical trial of a recombinant malaria vaccine consisting of the circumsporozoite repeat region of Plasmodium falciparum coupled to hepatitis B surface antigen.’, *The American journal of tropical medicine and hygiene*, 45(5), pp. 533–538.

Walker, A. S., Lourenço, J., Hill, A. V. S. and Gupta, S. (2015) ‘Modeling combinations of pre-erythrocytic plasmodium falciparum malaria vaccines’, *American Journal of Tropical Medicine and Hygiene*, 93(6), pp. 1254–1259.

Wang, F., Kikutani, H., Tsang, S. F., Kishimoto, T. and Kieff, E. (1991) ‘Epstein-Barr virus nuclear protein 2 transactivates a cis-acting CD23 DNA element.’, *Journal of virology*, 65(8), pp. 4101–6.

Wang, L., Tian, W.-D., Xu, X., Nie, B., Lu, J., Liu, X., Zhang, B., Dong, Q., Sunwoo, J. B., Li, G. and Li, X.-P. (2014) ‘Epstein-Barr virus nuclear antigen 1 (EBNA1) protein induction of epithelial-mesenchymal transition in nasopharyngeal carcinoma cells.’, *Cancer*, 120(3), pp. 363–72.

Watkins, E. R. and Maiden, M. C. J. (2012) ‘Persistence of hyperinvasive meningococcal strain types during global spread as recorded in the PubMLST database.’, *PloS one*, 7(9), p. e45349.

Wekerle, H. and Hohlfeld, R. (2003) ‘Molecular Mimicry in Multiple Sclerosis’, *NEJM*, 349(2), pp. 185–186.

White, M. T., Bejon, P., Olotu, A., Griffin, J. T., Riley, E. M., Kester, K. E., Ockenhouse, C. F. and Ghani, A. C. (2013) 'The relationship between RTS,S vaccine-induced antibodies, $CD4^+$ T cell responses and protection against Plasmodium falciparum infection.', *PloS one*. Public Library of Science, 8(4), p. e61395.

White, M. T., Griffin, J. T., Drakeley, C. J. and Ghani, A. C. (2010) 'Heterogeneity in malaria exposure and vaccine response: implications for the interpretation of vaccine efficacy trials.', *Malaria journal*, 9, p. 82.

White, M. T. and Smith, D. L. (2013) 'Synergism from combinations of infection-blocking malaria vaccines.', *Malaria journal*, 12(1), p. 280.

WHO (2014) *Malaria Vaccine Rainbow Tables*.

WHO (2016) *World Malaria Report*, *World Health Organization*.

Wikramaratna, P. S., Kucharski, A., Gupta, S., Andreasen, V., McLean, A. R. and Gog, J. R. (2015) 'Five challenges in modelling interacting strain dynamics', *Epidemics*, 10, pp. 31–34.

Wilson, J. B., Bell, J. L. and Levine, A. J. (1996) 'Expression of Epstein-Barr virus nuclear antigen-1 induces B cell neoplasia in transgenic mice.', *The EMBO journal*, 15(12), pp. 3117–26.

Wisniewski-Dyé, F. and Vial, L. (2008) 'Phase and antigenic variation mediated by genome modifications', *Antonie van Leeuwenhoek*, 94(4), pp. 493–515.

Woolhouse, M. E., Dye, C., Etard, J. F., Smith, T., Charlwood, J. D., Garnett, G. P., Hagan, P., Hii, J. L., Ndhlovu, P. D., Quinnell, R. J., Watts, C. H., Chandiwana, S. K. and Anderson, R. M. (1997) 'Heterogeneities in the transmission of infectious agents: implications for the design of control programs.', *Proceedings of the National Academy of Sciences of the United States of America*, 94(1), pp. 338–42.

World Health Assembly (2016) 'Global technical strategy and targets for malaria 2016 – 2030', (January 2015), pp. 1–4.

World Health Organization (2011) 'Meningococcal vaccines: WHO position paper', *Weekly Epidemiological Record*, 86(47), pp. 521–540.

World Health Organization (2015) *World Malaria Report*. Geneva, Switzerland.

Wright, S. (1932) 'The roles of mutation, inbreeding, crossbreeding and selection in evolution', *Proceedings of the Sixth International Congress on Genetics*, pp. 356–366.

Wu, N. C., Olson, C. A., Du, Y., Le, S., Tran, K., Remenyi, R., Gong, D., Al-mawsawi, L. Q., Qi, H. and Wu, T. (2015) 'Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality', pp. 1–27.

Wu, N. C. and Wilson, I. A. (2017) 'A Perspective on the Structural and Functional Constraints for Immune Evasion: Insights from Influenza Virus', *Journal of Molecular Biology*, 429(17), pp. 2694–2709.

Xiao, J., Palefsky, J. M., Herrera, R., Berline, J. and Tugizov, S. M. (2009) 'EBV BMRF-2 facilitates cell-to-cell spread of virus within polarized oral epithelial cells', *Virology*, 388(2), pp. 335–343.

Xiao, J., Palefsky, J. M., Herrera, R. and Tugizov, S. M. (2007) 'Characterization of the Epstein-Barr virus glycoprotein BMRF-2', *Virology*, 359(2), pp. 382–396.

Xu, J., Davis, C. T., Christman, M. C., Rivailler, P., Zhong, H., Donis, R. O. and Lu, G. (2012) 'Evolutionary history and phylodynamics of influenza A and B neuraminidase (NA) genes inferred from large-scale sequence analyses', *PLoS ONE*, 7(7).

Yang, Y., Jia, Y., Wang, Y., Wang, X., Sun, Z. and Luo, B. (2014) 'Sequence analysis of EBV immediate-early gene BZLF1 and BRLF1 in lymphomas', *Journal of Medical Virology*, 86(10), pp. 1788–1795.

Yates, J. L., Warren, N. and Sugden, B. (1985) 'Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells.', *Nature*, 313(6005), pp. 812–5.

Yewhalaw, D., Wassie, F., Steurbaut, W., Spanoghe, P., Van Bortel, W., Denis, L., Tessema, D. a, Getachew, Y., Coosemans, M., Duchateau, L. and Speybroeck, N. (2011) 'Multiple insecticide resistance: an impediment to insecticide-based malaria vector control program.', *PloS one*, 6(1), p. e16066.

Young, L. S., Lau, R., Rowe, M., Niedobitek, G., Packham, G., Shanahan, F., Rowe, D. T., Greenspan, D., Greenspan, J. S. and Rickinson, A. B. (1991) 'Differentiation-associated expression of the Epstein-Barr virus BZLF1 transactivator protein in oral hairy leukoplakia.', *Journal of virology*, 65(6), pp. 2868–74.

Appendices

Appendix A: Supplementary Methods for Chapter 3

The following appendix describes a method by which amino acid sequence alignments for key malaria proteins were generated from whole genome sequencing reads published by Pk3k Malaria Genetic Epidemiology Network (MalariaGEN).

In addition to the sequences analysed in the main body of this thesis, a pipeline to generate further set of sequences for malaria antigens was devised and implemented. Analysis of this set of antigens fell outside of the scope of this work, however will be revisited at a later date.

The Malaria Genetic Epidemiology Network (MalariaGEN) is an international community of researchers undertaking a number of projects that aim to character and understand the role of genetic variation in malaria infection at the level of host, vector, and parasite. The project began in 2005 and has, to date, generated in excess of 130TB of data. A key aim of the consortium is to sequence and publish 3,000 *P. falciparum* (Pf3k) whole genome sequences: as of [insert date] 2512 have been published and are readily available for academic use through *ftp* servers linked from the MalGen website (https://www.malariagen.net/projects/pf3k).

The sequence information made available by the Pf3k project comprises the raw overlapping reads (contigs) that are generated by next-generation sequencing approaches. We thus devised a pipeline that uses existing bioinformatic tools to assemble the sequence reads to generate consensus sequences for each isolate, find certain genes of interest within the whole genome, and then extract, translate, and align the relevant nucleotide sequences.

As mentioned previously, the sequence data from Pf3k is made available on an ftp server (ftp://ngs.sanger.ac.uk/production/pf3k/release_5), with the sequences for each isolate contained within a Binary Alignment/Map (BAM) file named by the ID of the isolate. The first step in the pipeline is thus to scan this folder for a full list of the isolate IDs that are available and, if the pipeline has already been run for some

time, to generate a list of the isolates which have yet to be downloaded and assembled.

Each BAM file is downloaded onto a local Linux server and processed – this step can be conducted in parallel given the high bandwidth connection, RAM capabilities, and processing capacity of the local server, significantly increasing with which final sequence alignments can be generated and processed. The first processing steps use a set of SAMtools utilities (Li *et al.*, 2009) to generate fastq files from the BAM files:

1. "sort" – the BAM file, as is downloaded, is not necessarily sorted by genomic position, and so this step ensures that the overlapping contigs present in the correct order and orientation
2. "index" – indexing the BAM file allows for fast random access, generating a Binary Alignment Index (BAI) file
3. "mpileup" – this step maps the BAI onto a reference sequence, in this case the standard *P. falciparum* 3D7 reference sequence, generating a file in the binary call format (BCF)
4. "bcftools" – this tool receives the BCF file and marks any low quality sites in the genome, that is, any sites that fall below some pre-defined threshold read depth, generating a VCF file
5. "vcfutils vcf2fq" – this final step converts the VCF file into a FASTQ file, which is a standard file type used to save raw sequence information as well as any other important metadata that may be saved with the file such as chromosome location, read depth, and isolate information.

A final tool (fq2fa, from the seqtk toolkit, https://github.com/lh3/seqtk) is then used to strip the metadata from the FASTQ file, generating a FASTA file which can then be read by alignment tools and by the Reverse Immunodynamic pipeline.

The methodology described above generates, for each isolate, a single FASTA file of sequence information, with a length of around 22.9Mbp. A pipeline similar to that described above for EBV was then used to generate amino acid sequence alignments from this whole genome information. In short, the whole sequence for each isolate was interrogated by BLAST for nucleotide sequences similar to a set of reference sequences for known *P. falciparum* antigens. Antigenic nucleotide sequences were then removed, translated into amino acid sequences, and aligned using MUSCLE, ready for analysis using the Reverse Immunodynamic pipeline. As the isolate information is maintained in the headers of each FASTA sequence, this data could be used to identify strain structure that may occur between, as well as between, malaria antigens.

The *P. falciparum* antigens for which multiple sequence alignments were generated are as follows: AMA-1, ASP, CSP, EBA-140, EBA-175, EBA-181, ETRAMP10, GAMA, H101, MAEBL, MSP-10, MSP-1, MSP-11, MSP-4, MSP-5, MSP-6, MSP-7, MSP-9, MSRP1, MSRP2, MSRP3, MTRAP, P113, P38, P41, P39, Pf34, PTEX-150, RH5, RhopH3, RON6, TLP.

A further analysis may also be to compare the patterns of non-overlap between known antigenic sequences, and proteins not known to be targeted by any immune response. One might expect non-immunogenic proteins to exhibit weaker patterns of

selection that antigenic ones, and this would again stand as validation of the Reverse Immunodynamic approach to epitope identification.

Appendix B: Supplementary Information for Chapter 3

Gene	Details	Known T cell epitope	Source	Class	In	Out	Mean entropy	SOI
BALF4	Envelope glycoprotein (gp110)	ILIYNGWYA	See Hislop et al.	CD8	106	114	0	0
BALF4	Envelope glycoprotein (gp110)	APGWLIWTY	See Hislop et al.	CD8	190	198	0	0
BALF4	Envelope glycoprotein (gp110)	FLDKGTYTL	See Hislop et al.	CD8	276	284	0	0
BALF4	Envelope glycoprotein (gp110)	AWCLEQKRQNMVLRE	See Hislop et al.	CD4	482	496	0	0
BALF4	Envelope glycoprotein (gp110)	VPGSETMCY	See Hislop et al.	CD8	552	560	0	0
BALF4	Envelope glycoprotein (gp110)	DNEIFLTKKMTEVCQ	See Hislop et al.	CD4	583	597	0	0
BLLF1	Envelope glycoprotein (gp350)	VLQWASLAV	See Rajçáni et al.	CD8	877	885	0	0
BLLF1	Envelope glycoprotein (gp350)	VLQWASLAV	See Hislop et al.	CD8	877	885	0	0
BXLF2	Envelope glycoprotein (gp85)	SLVIVTTFV	See Rajçáni et al.	CD8	225	233	0	0
BXLF2	Envelope glycoprotein (gp85)	TLFIGSHVV	See Hislop et al.	CD8	225	233	0	0
BXLF2	Envelope glycoprotein (gp85)	TLFIGSHVV	See Rajçáni et al.	CD8	420	428	0	0
BXLF2	Envelope glycoprotein (gp85)	LMIIPLINV	See Hislop et al.	CD8	420	428	0	0
BXLF2	Envelope glycoprotein (gp85)	LMIIPLINV	See Rajçáni et al.	CD8	542	550	0	0
BXLF2	Envelope glycoprotein (gp85)	SLVIVTTFV	See Hislop et al.	CD8	542	550	0	0
EBNA1	Latency associated	RRPQKRPSCIGCKGT	See Hislop et al.	CD4	75	89	0.040005019	0
EBNA1	Latency associated	RPQKRPSCI	See Hislop et al.	CD8	76	84	0.032826577	0
EBNA1	Latency associated	RPFFHPVGEADYFEY	See Hislop et al.	CD4	407	421	0.062504792	1
EBNA1	Latency associated	HPVGEADDYFEY	See Rajçáni et al.	CD8	411	421	0.085233807	0
EBNA1	Latency associated	HPVGEADYF	See Hislop et al.	CD8	411	419	0.104174653	1
EBNA1	Latency associated	HPVGEADYFEY	See Hislop et al.	CD8	411	421	0.085233807	1
EBNA1	Latency associated	VPPGAIEQGPADDPGEGPST	See Hislop et al.	CD4	433	452	0.066371923	2
EBNA1	Latency associated	IEQGPTDDPGEGPSTGPRGQGDGGR	See Hislop et al.	CD4	438	462	0.026216637	1
EBNA1	Latency associated	DGGRRKKGGWFGRHR	See Hislop et al.	CD4	459	470	0	0
EBNA1	Latency associated	SNPKFENIAEGLRVLLARSH	See Hislop et al.	CD4	478	497	0.139679287	3
EBNA1	Latency associated	NPKFENIAEGLRALL	See Hislop et al.	CD4	479	493	0.144185652	2
EBNA1	Latency associated	ENIAEGLRVLLARSHVERTT	See Hislop et al.	CD4	483	502	0.095163217	2
EBNA1	Latency associated	IAEGLRALLARSHVERTTDE	See Hislop et al.	CD4	485	504	0.146596876	0
EBNA1	Latency associated	LRALLARSHVERTTD	See Hislop et al.	CD4	489	503	0.163364931	3
EBNA1	Latency associated	EEGNWVAGVFVYGGSKTSLYNLRRG	See Hislop et al.	CD4	503	527	0.069925322	2
EBNA1	Latency associated	VYGGSKTSLYNLRRGTALAI	See Hislop et al.	CD4	513	532	0.066643884	1
EBNA1	Latency associated	TSLYNLRRGTALA	See Rajçáni et al.	CD8	519	531	0.060053022	0
EBNA1	Latency associated	TSLYNLRRGTALAI	See Hislop et al.	CD4	519	532	0.095205549	1
EBNA1	Latency associated	YNLRRGTALAIPQ	See Hislop et al.	CD4	522	534	0.102529053	1
EBNA1	Latency associated	NLRRGRTALAIPQCRL	See Hislop et al.	CD4	528	537	0.188506609	2
EBNA1	Latency associated	AIPQCRLTPLSRLPF	See Hislop et al.	CD4	531	545	0.07362512	2
EBNA1	Latency associated	IPQCRLTPL	See Hislop et al.	CD8	532	540	0.122708533	1
EBNA1	Latency associated	PQCRLTPLSRLPFGM	See Hislop et al.	CD4	533	547	0.03681256	1
EBNA1	Latency associated	PQPGPLRESIVCYFMVFLQT	See Hislop et al.	CD4	553	572	0.031540048	1
EBNA1	Latency associated	PGPLRESIVCYFMVFLQTHI	See Hislop et al.	CD4	555	574	0.031540048	1
EBNA1	Latency associated	APGPGPQPLRESIVCYFM	See Hislop et al.	CD4	557	567	0.057345543	1
EBNA1	Latency associated	LRESIVCYFMVFLQTHIFAE	See Hislop et al.	CD4	558	577	0.031540048	1
EBNA1	Latency associated	LRESIVCYFMVFLQTHIFAEVLKDA	See Hislop et al.	CD4	558	582	0.050464077	2
EBNA1	Latency associated	YFMVFLQTHIFAE	See Hislop et al.	CD4	565	577	0.048523151	1
EBNA1	Latency associated	MVFLQTHIFAEVLKD	See Hislop et al.	CD4	567	581	0.084106796	1
EBNA1	Latency associated	VFLQTHIFAEVLKDAIKDL	See Hislop et al.	CD4	568	586	0.053787852	1
EBNA1	Latency associated	VLKDAIKDL	See Rajçáni et al.	CD8	578	586	0.113552132	1
EBNA1	Latency associated	VLKDAIKDL	See Hislop et al.	CD8	578	586	0.113552132	1
EBNA1	Latency associated	VLKDAIKDLVMTKPAPTCNI	See Hislop et al.	CD4	578	597	0.117841387	2
EBNA1	Latency associated	PTCNIKVTVCSFDDGVDLPPWFPPM	See Hislop et al.	CD4	593	617	0.04017298	1
EBNA1	Latency associated	RVTVCSFDDGVDLPPWFPPM	See Hislop et al.	CD4	598	617	0.050216225	1
EBNA1	Latency associated	PPWFPPMVEGAAA	See Hislop et al.	CD4	611	623	0	0
EBNA2	Latency associated	GQTYHLIVDTLALHGGQTYH	See Hislop et al.	CD4	-5	15	0	0
EBNA2	Latency associated	YHLIVDTDSL	See Hislop et al.	CD8	14	23	0.071973516	0
EBNA2	Latency associated	QLSDTPLIPLTIFVGNTGV	See Rajçáni et al.	CD8	39	53	0.029946915	0
EBNA2	Latency associated	DTPLIPLTIF	See Rajçáni et al.	CD8	42	51	0.044920373	0
EBNA2	Latency associated	DTPLIPLTIF	See Hislop et al.	CD8	42	51	0.044920373	0
EBNA2	Latency associated	IPLTIFVGENTGVPPPLPPP	See Hislop et al.	CD4	46	57	0	0

EBNA2	Latency associated	LPPATLVPPRPTRPTTLPP	See Hislop et al.	CD4	218	230	0.00917284	0
EBNA2	Latency associated	PRSPTVFYNIPPMPL	See Rajçáni et al.	CD8	283	297	0.031241155	0
EBNA2	Latency associated	TVFYNIPPMPL	See Rajçáni et al.	CD8	287	297	0.03197368	0
EBNA2	Latency associated	PRSTVFYNIPPMPLPPSQL	See Hislop et al.	CD4	287	302	0.02618171	0
EBNA2	Latency associated	TVFYNIPPMPL	See Hislop et al.	CD4	287	297	0.03197368	0
EBNA2	Latency associated	PAQPPPGVINDQQLHHLPSG	See Hislop et al.	CD4	308	327	0.045955293	0
EBNA3A	Latency associated	QAKWRLQTL	See Rajçáni et al.	CD8	158	166	0	0
EBNA3A	Latency associated	QAKWRLQTL	See Hislop et al.	CD8	158	166	0	0
EBNA3A	Latency associated	AYSSWMYSY	See Rajçáni et al.	CD8	176	184	0.01986598	0
EBNA3A	Latency associated	AYSSWMYSY	See Hislop et al.	CD8	176	184	0.01986598	0
EBNA3A	Latency associated	RYSIFFDY	See Rajçáni et al.	CD8	246	253	0.008151795	0
EBNA3A	Latency associated	RYSIFFDY	See Hislop et al.	CD8	246	253	0.008151795	0
EBNA3A	Latency associated	AWNAGFLRGRA	See Rajçáni et al.	CD8	320	330	0.017785736	0
EBNA3A	Latency associated	FLRGRAYGL	See Rajçáni et al.	CD8	325	333	0.046211419	0
EBNA3A	Latency associated	FLRGRAYGL	See Hislop et al.	CD8	325	333	0.046211419	0
EBNA3A	Latency associated	EDLPCIVSRGGPKVKRPPIF	See Hislop et al.	CD4	364	383	0.003260718	0
EBNA3A	Latency associated	RPPIFIRRL	See Hislop et al.	CD8	377	387	0	0
EBNA3A	Latency associated	KRPPIFIRRL	See Hislop et al.	CD8	378	387	0	0
EBNA3A	Latency associated	RPPIFIRRL	See Rajçáni et al.	CD8	379	387	0	0
EBNA3A	Latency associated	LEKARGSTY	See Rajçáni et al.	CD8	406	414	0.014492081	0
EBNA3A	Latency associated	LEKARGSTY	See Hislop et al.	CD8	406	414	0.014492081	0
EBNA3A	Latency associated	HLAAQGMAY	See Hislop et al.	CD8	450	459	0.106599225	0
EBNA3A	Latency associated	YPLHEQHGM	See Rajçáni et al.	CD8	458	466	0.202251048	1
EBNA3A	Latency associated	YPLHEQHGM	See Hislop et al.	CD8	458	466	0.202251048	1
EBNA3A	Latency associated	VFSDGRVAC	See Rajçáni et al.	CD8	491	499	0.114001697	0
EBNA3A	Latency associated	VFSDGRVAC	See Hislop et al.	CD8	491	499	0.114001697	0
EBNA3A	Latency associated	VPAPAGPIV	See Rajçáni et al.	CD8	502	510	0	0
EBNA3A	Latency associated	VPAPAGPIV	See Hislop et al.	CD8	502	510	0	0
EBNA3A	Latency associated	SVRDRLARL	See Rajçáni et al.	CD8	596	608	0	0
EBNA3A	Latency associated	SVRDRLARL	See Hislop et al.	CD8	596	608	0	0
EBNA3A	Latency associated	RLRAEAQVK	See Rajçáni et al.	CD8	607	615	0	0
EBNA3A	Latency associated	RLRAEAQVK	See Hislop et al.	CD8	607	615	0	0
EBNA3A	Latency associated	VQPPQLTLQV	See Hislop et al.	CD8	621	627	0.139607417	1
EBNA3A	Latency associated	QVADVVRAPGVPAMQPQYF	See Hislop et al.	CD4	653	671	0.075772189	0
EBNA3A	Latency associated	GPWVPEQWMFQGAPPSQGTP	See Hislop et al.	CD4	784	796	0	0
EBNA3B	Latency associated	HRCQAIRKK	See Rajçáni et al.	CD8	149	157	0	0
EBNA3B	Latency associated	HRCQAIRKK	See Hislop et al.	CD8	149	157	0	0
EBNA3B	Latency associated	TYSAGIVQ	See Rajçáni et al.	CD8	218	225	0	0
EBNA3B	Latency associated	TYSAGIVQI	See Hislop et al.	CD8	218	225	0	0
EBNA3B	Latency associated	RRARSLSAERY	See Rajçáni et al.	CD8	244	254	0	0
EBNA3B	Latency associated	RRARSLSAERY	See Hislop et al.	CD8	244	254	0	0
EBNA3B	Latency associated	VSFIEFVGW	See Hislop et al.	CD8	280	288	0	0
EBNA3B	Latency associated	AVFDRKSDAK	See Rajçáni et al.	CD8	400	409	0.159577462	0
EBNA3B	Latency associated	AVFDRKSDAK	See Hislop et al.	CD8	400	409	0.159577462	0
EBNA3B	Latency associated	IVTDFSVIK	See Rajçáni et al.	CD8	417	425	0.146561159	0
EBNA3B	Latency associated	IVTDFSVIK	See Hislop et al.	CD8	417	425	0.146561159	0
EBNA3B	Latency associated	AVLLHEESM	See Hislop et al.	CD8	489	497	0.074079136	0
EBNA3B	Latency associated	VEITPYKPTW	See Rajçáni et al.	CD8	658	667	0.098771319	1
EBNA3B	Latency associated	VEITPYKPTW	See Hislop et al.	CD8	658	667	0.098771319	1
EBNA3B	Latency associated	GOGGSPTAMJ	See Rajçáni et al.	CD8	840	849	0.01310753	0
EBNA3C	Latency associated	NRGWMQRIRRRRRR	See Hislop et al.	CD4	66	79	0.009316338	0
EBNA3C	Latency associated	PHDITYPYTARNIRDAACRAV	See Hislop et al.	CD4	105	119	0.068395866	0
EBNA3C	Latency associated	ILCFVMAARQRLQDI	See Hislop et al.	CD4	141	155	0.03681256	0
EBNA3C	Latency associated	EGGVGWRHW	See Rajçáni et al.	CD8	163	171	0	0
EBNA3C	Latency associated	EGGVGWRHW	See Hislop et al.	CD8	163	171	0	0
EBNA3C	Latency associated	QNGALAINTF	See Rajçáni et al.	CD8	213	222	0.18335564	0
EBNA3C	Latency associated	QNGALAINTF	See Hislop et al.	CD8	213	222	0.18335564	0
EBNA3C	Latency associated	LRGKWQRRYR	See Hislop et al.	CD8	249	258	0.039116822	0
EBNA3C	Latency associated	RRIYDLIEL	See Hislop et al.	CD8	258	266	0.056334579	0

EBNA3C	Latency associated	HHIWQNLL	See Rajçáni et al.	CD8	271	278	0.039633478	0
EBNA3C	Latency associated	HHIWQNLL	See Hislop et al.	CD8	271	278	0.039633478	0
EBNA3C	Latency associated	EENLLDFVRF	See Rajçáni et al.	CD8	281	290	0	0
EBNA3C	Latency associated	EENLLDFVRF	See Hislop et al.	CD8	281	290	0	0
EBNA3C	Latency associated	LLDFVRFMGV	See Rajçáni et al.	CD8	284	293	0	0
EBNA3C	Latency associated	LLDFVRFMGV	See Hislop et al.	CD8	284	293	0	0
EBNA3C	Latency associated	LDFVRFMGV	See Rajçáni et al.	CD8	285	293	0	0
EBNA3C	Latency associated	LDFVRFMGV	See Hislop et al.	CD8	285	293	0	0
EBNA3C	Latency associated	RGIKEHVIQNARFKH	See Rajçáni et al.	CD8	332	342	0.068978916	0
EBNA3C	Latency associated	KEHVIQNAF	See Rajçáni et al.	CD8	335	343	0.084307564	0
EBNA3C	Latency associated	KEHVIQNAF	See Hislop et al.	CD8	335	343	0.084307564	0
EBNA3C	Latency associated	FRKAQIQGL	See Hislop et al.	CD8	335	343	0.084307564	0
EBNA3C	Latency associated	FRKAQIQGL	See Rajçáni et al.	CD8	343	351	0.091689034	0
EBNA3C	Latency associated	SDDELPYIDPNMEPV	See Hislop et al.	CD4	386	400	0.004347624	0
EBNA3C	Latency associated	QQRPVMFVSRVPAKK	See Hislop et al.	CD4	401	415	0	0
EBNA3C	Latency associated	QKRAAPPTVSPSDTG	See Hislop et al.	CD4	547	560	0.018537174	0
EBNA3C	Latency associated	PPVVRMFMRERQLPQ	See Hislop et al.	CD4	627	641	0.004347624	0
EBNA3C	Latency associated	PQCFWEMRAGREITQ	See Hislop et al.	CD4	647	661	0.045427238	0
EBNA3C	Latency associated	QPRAPIRPI	See Rajçáni et al.	CD8	889	897	0.00724604	0
EBNA3C	Latency associated	QPRAPIRPI	See Hislop et al.	CD8	889	897	0.00724604	0
EBNA3C	Latency associated	PSMPFASDYSQGAFT	See Hislop et al.	CD4	924	938	0.033283176	0
EBNA3C	Latency associated	AQEILSDNSEISVFPK	See Hislop et al.	CD4	969	984	0.008151795	0
LMP1	Latency associated	GPPRPPLGPPLSSSIGLALL	See Hislop et al.	CD4	11	30	0.170376166	2
LMP1	Latency associated	FWLYIVMSD	See Hislop et al.	CD8	38	46	0.169053312	0
LMP1	Latency associated	ALLVLYSFAALL	See Rajçáni et al.	CD8	51	59	0.035954213	0
LMP1	Latency associated	FRRDLLCPLGA	See Hislop et al.	CD8	72	83	0.062462795	1
LMP1	Latency associated	LLLIALWNL	See Rajçáni et al.	CD8	92	100	0.014492081	0
LMP1	Latency associated	YLLEMLWRL	See Rajçáni et al.	CD8	125	133	0.197210435	0
LMP1	Latency associated	YLLEMLWRL	See Hislop et al.	CD8	125	133	0.197210435	0
LMP1	Latency associated	LWRLGATIWQLLAFF	See Hislop et al.	CD4	130	144	0.087121686	1
LMP1	Latency associated	RLGATIWQL	See Rajçáni et al.	CD8	132	140	0.055062885	0
LMP1	Latency associated	YLQQNWWTL	See Rajçáni et al.	CD8	159	167	0.021738121	0
LMP1	Latency associated	TLLVDLLWL	See Rajçáni et al.	CD8	166	174	0.00724604	0
LMP1	Latency associated	TLLVDLLWL	See Hislop et al.	CD8	166	174	0.00724604	0
LMP1	Latency associated	LLVDLLWLL	See Rajçáni et al.	CD8	167	175	0	0
LMP1	Latency associated	YLQQNWWTL	See Hislop et al.	CD8	167	175	0	0
LMP1	Latency associated	LIWMYYHGPRHTDEHHHDDS	See Hislop et al.	CD4	181	200	0.089833285	1
LMP1	Latency associated	QATDDSSHESDSNSNEGRHH	See Hislop et al.	CD4	206	225	0.111723146	0
LMP1	Latency associated	SSHESDSNSNEGRHHLLVSG	See Hislop et al.	CD4	211	230	0.120923656	1
LMP1	Latency associated	SGHESDSNSNEGRHHH	See Hislop et al.	CD4	211	225	0.118093346	0
LMP1	Latency associated	TDGGGGHSHDSGHGG	See Hislop et al.	CD4	385	399	0.081015463	0
LMP1	Latency associated	DPHGPVQLSYYD	See Hislop et al.	CD8	420	431	0.022347742	0
LMP2	Latency associated	MGSLEMVPMMGS	See Rajçáni et al.	CD8	1	9	0.012619939	0
LMP2	Latency associated	MGSLEMVPM	See Hislop et al.	CD8	1	9	0.012619939	0
LMP2	Latency associated	EDPYWGNGDRHSDYQ	See Hislop et al.	CD8	61	75	0.089983965	0
LMP2	Latency associated	DYQPLGTQDQSLYLG	See Hislop et al.	CD4	73	87	0.089869717	2
LMP2	Latency associated	NPVCLPVIVAPYLF	See Hislop et al.	CD8	121	134	0	0
LMP2	Latency associated	ASCFTASVSTVVTA	See Hislop et al.	CD8	125	133	0	0
LMP2	Latency associated	APYLFWLAA	See Rajçáni et al.	CD8	130	138	0	0
LMP2	Latency associated	PYLFWLAAI	See Rajçáni et al.	CD8	131	139	0	0
LMP2	Latency associated	PYLFWLAAI	See Hislop et al.	CD8	131	139	0	0
LMP2	Latency associated	LPVIVAPYL	See Hislop et al.	CD8	144	152	0.00724604	0
LMP2	Latency associated	FTASVSTVV	See Hislop et al.	CD8	144	152	0.00724604	0
LMP2	Latency associated	STVVTATGLALSLLL	See Hislop et al.	CD4	149	163	0.032681081	1
LMP2	Latency associated	SSYAAAQRKLLTPV	See Hislop et al.	CD4	169	182	0.080546868	2
LMP2	Latency associated	VTFFAICLTWRIEDPPFNSI	See Hislop et al.	CD4	189	208	0.03950757	1
LMP2	Latency associated	ICLTWRIEDPPFNSILFALL	See Hislop et al.	CD4	194	213	0.036246852	1
LMP2	Latency associated	IEDPPFNSL	See Rajçáni et al.	CD8	200	208	0.063271715	0
LMP2	Latency associated	IEDPPFNSL	See Hislop et al.	CD8	200	208	0.063271715	0

LMP2	Latency associated	VLVMLVLLILAYRRRWRRLT	See Hislop et al.	CD4	224	243	0.005678973	0
LMP2	Latency associated	RRRWRRLTV	See Rajçáni et al.	CD8	236	244	0.05349595	0
LMP2	Latency associated	RRRWRRLTV	See Hislop et al.	CD8	236	244	0.05349595	0
LMP2	Latency associated	RRWRRLTVC	See Hislop et al.	CD8	237	245	0.05349595	0
LMP2	Latency associated	RRLTVCGGIMF	See Hislop et al.	CD8	240	250	0.10297684	1
LMP2	Latency associated	TVCGGIMFL	See Hislop et al.	CD8	243	251	0.125860582	1
LMP2	Latency associated	MFLACVLVLIVDAV	See Hislop et al.	CD8	249	262	0.080269254	1
LMP2	Latency associated	LIVDAVLQL	See Hislop et al.	CD8	257	265	0	0
LMP2	Latency associated	QLSPLLGAV	See Rajçáni et al.	CD8	264	272	0	0
LMP2	Latency associated	GLGTLGAAI	See Hislop et al.	CD8	293	300	0	0
LMP2	Latency associated	LLWTLVVLL	See Rajçáni et al.	CD8	329	337	0	0
LMP2	Latency associated	LLWTLVVLL	See Hislop et al.	CD8	329	337	0	0
LMP2	Latency associated	SSCSSCPLSKI	See Rajçáni et al.	CD8	340	350	0.101794511	1
LMP2	Latency associated	SSCSSCPLSKI	See Hislop et al.	CD8	340	350	0.101794511	0
LMP2	Latency associated	ILLARLFLY	See Hislop et al.	CD8	350	358	0.075578479	1
LMP2	Latency associated	FLYALALL	See Rajçáni et al.	CD8	356	363	0.008151795	0
LMP2	Latency associated	FLYALALLL	See Hislop et al.	CD8	356	363	0.008151795	0
LMP2	Latency associated	STEFIPNLFCMLLL	See Hislop et al.	CD4	385	398	0.052474374	1
LMP2	Latency associated	TYGPVFMCL	See Rajçáni et al.	CD8	419	427	0.054468468	0
LMP2	Latency associated	TYGPVFMCL	See Hislop et al.	CD8	419	427	0.054468468	0
LMP2	Latency associated	LLSAWILTA	See Hislop et al.	CD8	419	427	0.054468468	0
LMP2	Latency associated	TYGPVFMSLGGLLTMVAGAV	See Hislop et al.	CD4	419	438	0.040690206	1
LMP2	Latency associated	TZGPVFMCL	See Rajçáni et al.	CD8	421	427	0.070030887	1
LMP2	Latency associated	CLGGLLTMV	See Rajçáni et al.	CD8	426	434	0.090422681	1
LMP2	Latency associated	CLGGLLTMV	See Hislop et al.	CD8	426	434	0.090422681	1
LMP2	Latency associated	GLGGLLTMV	See Rajçáni et al.	CD8	427	434	0.040448489	0
LMP2	Latency associated	VMSNTLLSAW	See Rajçáni et al.	CD8	442	451	0.074527663	0
LMP2	Latency associated	VMSNTLLSAW	See Hislop et al.	CD8	442	451	0.074527663	0
LMP2	Latency associated	LLSAWILTA	See Rajçáni et al.	CD8	447	455	0	0
LMP2	Latency associated	LTAGFIFL	See Rajçáni et al.	CD8	453	461	0.00724604	0
LMP2	Latency associated	LTAGFLIFL	See Rajçáni et al.	CD8	453	461	0.00724604	0
LMP2	Latency associated	LTAGFLIFL	See Hislop et al.	CD8	453	461	0.00724604	0
BALF2	Non-structural in lytic infection	ARYAYYLQF	See Hislop et al.	CD8	418	426	0	0
BHRF1	Non-structural in lytic infection	TVVLRYHVLLEEI	See Hislop et al.	CD4	45	57	0	0
BHRF1	Non-structural in lytic infection	PYYVVDLSVRGM	See Hislop et al.	CD4	122	133	0	0
BHRF1	Non-structural in lytic infection	AGLTLSLLVICSYLFISRG	See Hislop et al.	CD4	171	189	0	0
BMLF1	Non-structural in lytic infection	KDTWLDARM	See Rajçáni et al.	CD8	247	255	0	0
BMLF1	Non-structural in lytic infection	KDTWLDARM	See Hislop et al.	CD8	247	255	0	0
BMLF1	Non-structural in lytic infection	*GLTCLVAML*	See Rajçáni et al.	CD8	262	270	0	0
BMLF1	Non-structural in lytic infection	GLCTLVAML	See Hislop et al.	CD8	262	270	0	0
BMLF1	Non-structural in lytic infection	DEVEFLGHY	See Rajçáni et al.	CD8	379	387	0.00724604	0
BMLF1	Non-structural in lytic infection	DEVEFLGHY	See Hislop et al.	CD8	379	387	0.00724604	0
BMLF1	Non-structural in lytic infection	SRLVRAILSP	See Hislop et al.	CD8	417	426	0	0
BMRF1	Non-structural in lytic infection	CYDHAQTHL	See Hislop et al.	CD8	20	28	0	0
BMRF1	Non-structural in lytic infection	FRNLAYGRTCVLGKE	See Rajçáni et al.	CD8	86	100	0	0
BMRF1	Non-structural in lytic infection	FRNLAYGRTCVLGKE	See Hislop et al.	CD8	86	100	0	0
BMRF1	Non-structural in lytic infection	RPQGGSRPEFVKL	See Hislop et al.	CD8	116	128	0	0
BMRF1	Non-structural in lytic infection	TLDYKPLSV	See Hislop et al.	CD8	208	216	0	0
BMRF1	Non-structural in lytic infection	YRSGIIAVV	See Rajçáni et al.	CD8	268	276	0	0
BMRF1	Non-structural in lytic infection	YRSGIIAVV	See Hislop et al.	CD8	268	276	0	0
BMRF1	Non-structural in lytic infection	YRSGIIAVV	See Hislop et al.	CD8	268	276	0	0
BMRF1	Non-structural in lytic infection	LPLDLSVILF	See Hislop et al.	CD8	286	295	0	0
BRLF1	Non-structural in lytic infection	LVSDYCNVLNKEFTA	See Rajçáni et al.	CD8	25	39	0	0
BRLF1	Non-structural in lytic infection	LVSDYCNVLNKEFT	See Hislop et al.	CD8	25	38	0	0
BRLF1	Non-structural in lytic infection	LVSDYCNVL	See Hislop et al.	CD8	25	33	0	0
BRLF1	Non-structural in lytic infection	DYCNVLNKEF	See Rajçáni et al.	CD8	28	37	0	0
BRLF1	Non-structural in lytic infection	DYCNVLNKEF	See Hislop et al.	CD8	28	37	0	0
BRLF1	Non-structural in lytic infection	AENAGNDAC	See Hislop et al.	CD8	91	99	0	0
BRLF1	Non-structural in lytic infection	IACPIVMRYYVLDHLI	See Hislop et al.	CD8	101	115	0	0

BRLF1	Non-structural in lytic infection	YVLDHLIVV	See Hislop et al.	CD8	109	117	0	0
BRLF1	Non-structural in lytic infection	FFIQAPSNRVMIPAT	See Hislop et al.	CD8	121	135	0	0
BRLF1	Non-structural in lytic infection	ATIGTAMYK	See Rajçáni et al.	CD8	134	142	0	0
BRLF1	Non-structural in lytic infection	ATIGTAMYK	See Hislop et al.	CD8	134	142	0	0
BRLF1	Non-structural in lytic infection	KHSRVRAYTYSKVLG	See Rajçáni et al.	CD8	145	159	0	0
BRLF1	Non-structural in lytic infection	KHSRVRAYTYSKVLG	See Hislop et al.	CD8	145	159	0	0
BRLF1	Non-structural in lytic infection	RALIKTLPRASYSSH	See Hislop et al.	CD8	225	239	0	0
BRLF1	Non-structural in lytic infection	TLPRASYSSH	See Rajçáni et al.	CD8	230	239	0	0
BRLF1	Non-structural in lytic infection	ERPIFPHPSKPTLFLP	See Rajçáni et al.	CD8	393	404	0.009464954	0
BRLF1	Non-structural in lytic infection	ERPIFPHPSKPTFLP	See Hislop et al.	CD8	393	407	0.015143927	0
BRLF1	Non-structural in lytic infection	EVCQPKRIRPFHPPG	See Hislop et al.	CD8	441	455	0	0
BRLF1	Non-structural in lytic infection	QKEEAAICGQMDLS	See Hislop et al.	CD8	529	542	0.049380828	0
BRLF1	Non-structural in lytic infection	QKEEIAAICGQMDLSH	See Rajçáni et al.	CD8	533	543	0.062848327	0
BZLF1	Non-structural in lytic infection	LPEPLPQGQLTAY	See Hislop et al.	CD8	52	64	0.073630756	0
BZLF1	Non-structural in lytic infection	EPLPQGQLTAY	See Hislop et al.	CD8	54	64	0.087018167	0
BZLF1	Non-structural in lytic infection	APENAYQAY	See Hislop et al.	CD8	81	89	0	0
BZLF1	Non-structural in lytic infection	DSELEIKRYKNR	See Hislop et al.	CD8	172	183	0	0
BZLF1	Non-structural in lytic infection	ELEIKRYKNRVASRK	See Hislop et al.	CD4	174	188	0	0
BZLF1	Non-structural in lytic infection	RKCCRAKFK**Q**LLQHYR	See Hislop et al.	CD8	185	201	0.027793404	1
BZLF1	Non-structural in lytic infection	CRAKFKQLLQHYR	See Rajçáni et al.	CD8	189	201	0.036345221	0
BZLF1	Non-structural in lytic infection	RAKFKQLL	See Rajçáni et al.	CD8	190	197	0.059060984	0
BZLF1	Non-structural in lytic infection	RAKFK**Q**LL	See Hislop et al.	CD8	190	197	0.059060984	1
BZLF1	Non-structural in lytic infection	LQHYREVAA	See Hislop et al.	CD8	197	205	0.072693946	0
BZLF1	Non-structural in lytic infection	KSSENDRLRLLLKQM	See Hislop et al.	CD4	207	221	0.004347624	0
BZLF1	Non-structural in lytic infection	SENDRLRLL	See Hislop et al.	CD8	209	217	0	0
BILF2	Probable envelope glycoprotein	RRRKGWIPL	See Hislop et al.	CD8	240	248	0	0

Table S1 Known T cell epitopes in Epstein Barr Virus. Columns show the protein in which the epitope is found ("Gene"), the class of the protein ("Details"), the sequence of the epitope ("Known T cell epitope", SoI marked in bold), a reference to a review in which experimental references to the epitope may be found ("Source"), the class of T cell that the epitope is recognized by ("Class"), the start and end position of the epitope in my sequence alignment ("In" and "Out"), the mean entropy of the epitope ("Mean entropy"), and the number of Sites of Interest (in non-overlap with some other site) that are found in the epitope ("SOI")

		Number of interactive amino acid residues within and between genes			Number of interactive amino acid residues between genes		
Gene Name	**Phase**	**MI_S > 0.5**	**MI_S > 0.7**	**MI_S > 0.9**	**MI_S > 0.5**	**MI_S > 0.7**	**MI_S > 0.9**
BALF2	E	4	2	1	4	2	1
BALF3	E	4	1	1	4	1	1
BALF4	Glycoprotein	8	6	0	5	3	0
BALF5	E	2	0	0	2	0	0
BBLF1	E	2	1	0	2	1	0
BBLF2BBLF3	E	13	0	0	9	0	0
BBLF4	E	7	0	0	7	0	0
BBRF1	E	2	0	0	2	0	0
BBRF2	E	3	0	0	3	0	0
BcLF1	E	3	1	1	3	1	1
BcRF1	E	15	11	10	13	11	1
BDLF2	Glycoprotein	5	0	0	3	0	0
BDLF3	Glycoprotein	4	3	2	4	0	0
BDLF4	E	2	2	1	2	2	1
BdRF1	E	4	3	0	4	3	0
BFLF1	E	4	2	0	3	0	0
BFLF2	E	4	1	0	1	1	0
BFRF1	E	4	2	0	4	0	0
BFRF1A	E	2	0	0	2	0	0
BGLF1	E	5	3	1	5	1	1
BGLF5	E	4	0	0	4	0	0
BGRF1BDRF1	E	1	0	0	1	0	0
BKRF2	Glycoprotein	2	2	0	2	2	0
BKRF3	E	1	1	0	1	1	0
BKRF4	E	3	3	0	3	0	0
BLLF1	Glycoprotein	17	11	1	17	6	1
BLLF2	E	4	4	1	4	4	1
BLLF3	E	1	1	0	1	1	0
BMLF1	E	3	2	2	3	2	2
BNLF2A	E	2	2	0	1	0	0
BNLF2B	E	1	1	0	1	1	0
BNRF1	E	7	1	0	7	1	0
BOLF1	E	15	2	1	13	2	1
BORF1	E	2	0	0	2	0	0

BORF2	E	2	1	0	2	1	0
BPLF1	E	35	10	2	11	3	0
BRLF1	IE	9	2	2	3	0	0
BRRF1	E	1	0	0	1	0	0
BRRF2	E	27	6	2	20	2	0
BSLF1	E	4	0	0	3	0	0
BSLF2BMLF1	E	3	2	2	3	2	2
BTRF1	E	4	1	0	4	1	0
BVLF1	E	1	1	0	1	1	0
BVRF1	E	2	2	0	2	0	0
BVRF2	E	6	3	0	4	3	0
BXLF1	E	1	1	0	1	1	0
BXLF2	Glycoprotein	2	2	0	2	2	0
BXRF1	E	1	1	1	1	1	1
BZLF1	IE	10	5	0	10	0	0
EBNA1	Latency	22	21	8	20	14	0
EBNA2	Latency	14	7	5	9	1	0
EBNA3A	Latency	28	12	7	11	0	0
EBNA3B	Latency	35	4	0	12	1	0
EBNA3C	Latency	33	9	0	25	1	0
LF1	E	4	0	0	2	0	0
LMP1	Latency	21	16	4	12	3	0
LMP2	Latency	16	14	2	16	12	0
RPMS1	E	2	1	0	2	1	0
SUM		**443**	**189**	**57**	**319**	**96**	**14**
Total N		58	45	21	58	36	12
N EBNA		132	53	20	77	17	0
N LMP		37	30	6	28	15	0
N Latency		169	83	26	105	32	0

Table S2 Numbers of Sites of Interest found (at each cut-off of MIS score) in each gene from EBV. Stage of infection primarily associated with the genes are shown ("Phase"), as are total number of genes in each stage (or that are in glycoproteins).

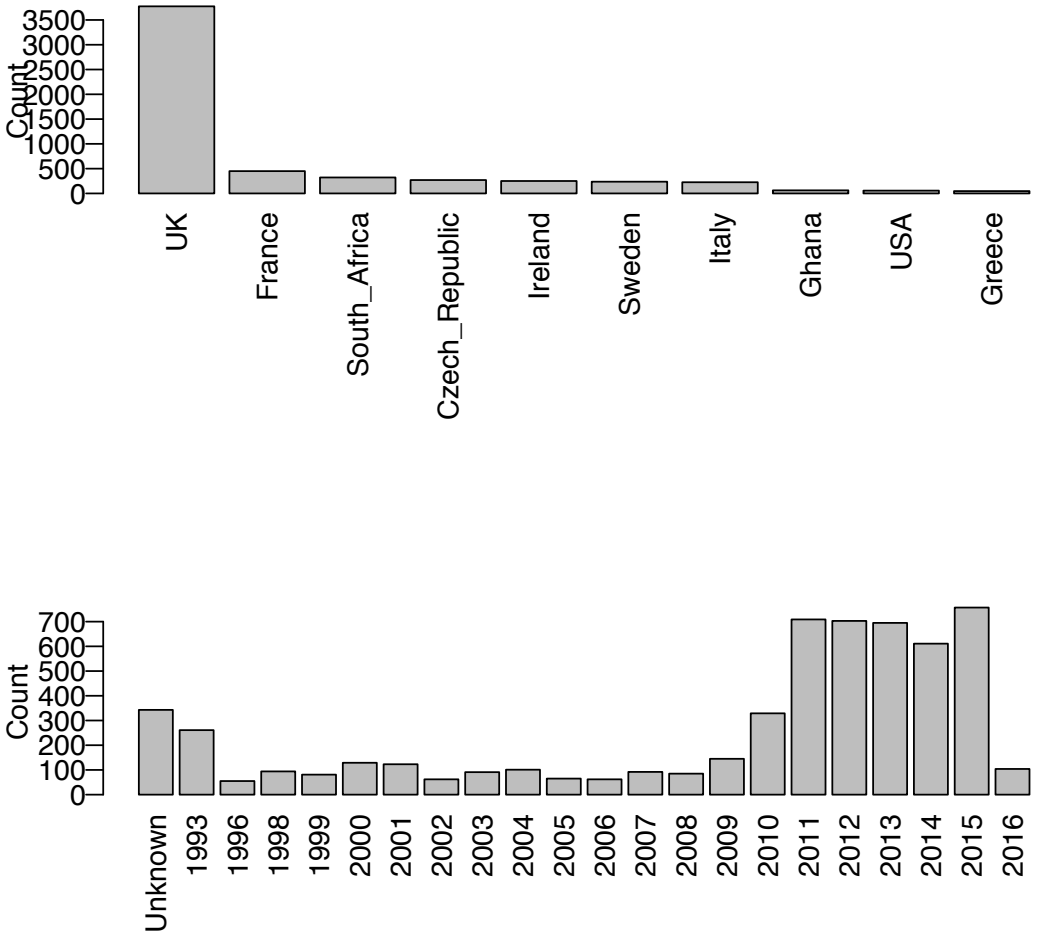


Figure S1 Geographic origin and year of collection of *Neisseria meningitidis* sequences from the PubMLST database.

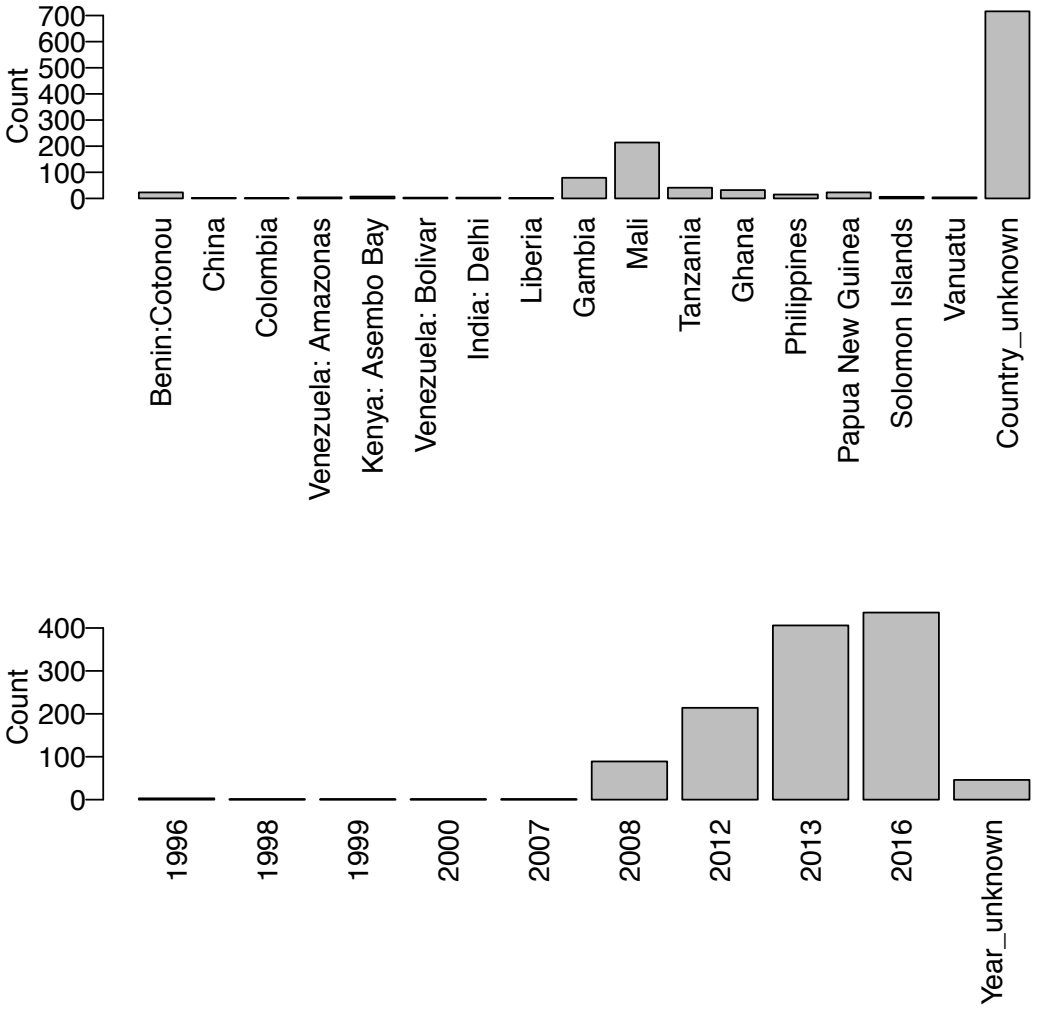


Figure S2 Geographic origin and year of collection of *Plasmodium falciparum* AMA-1 sequences from GenBank.

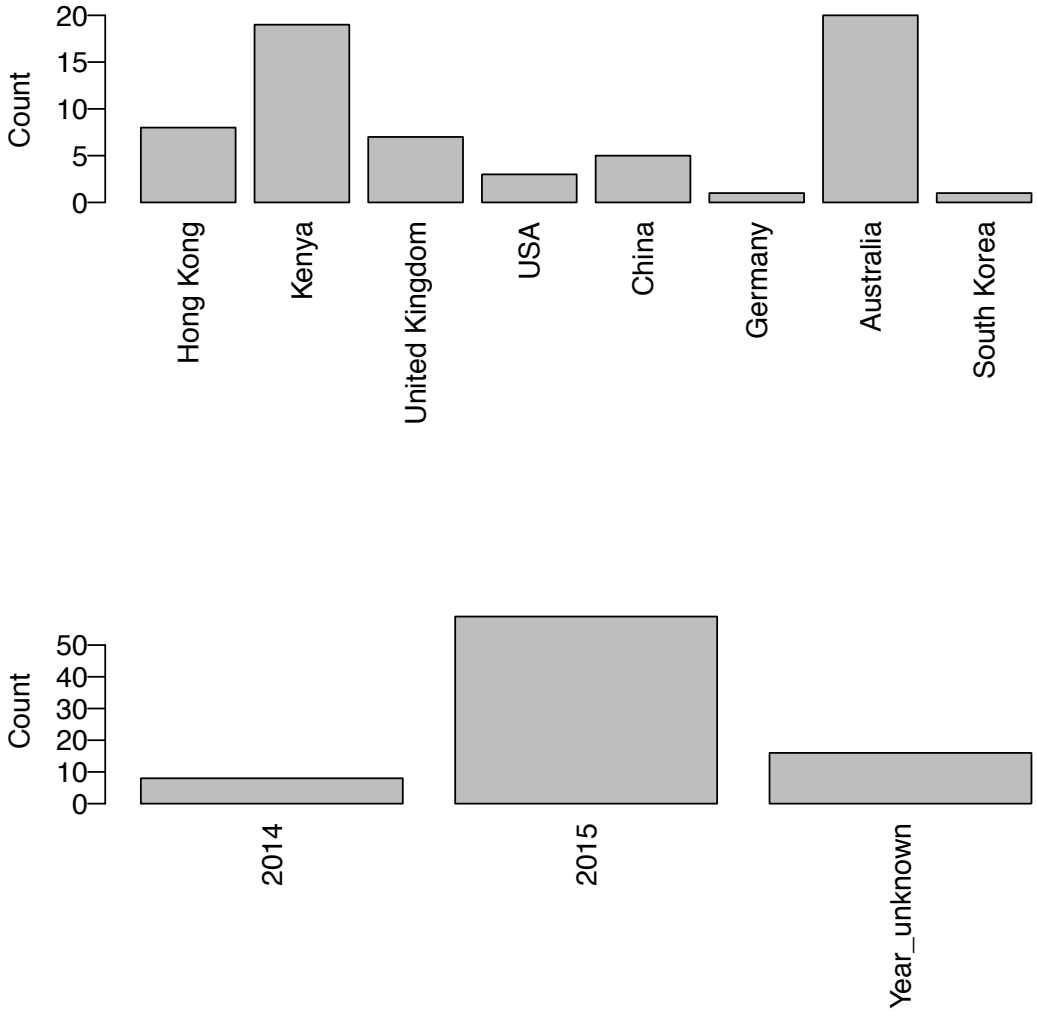


Figure S3 Geographic origin and year of collection of *EBV* sequences from GenBank.

Appendix C: Supplementary Information for Chapter 4

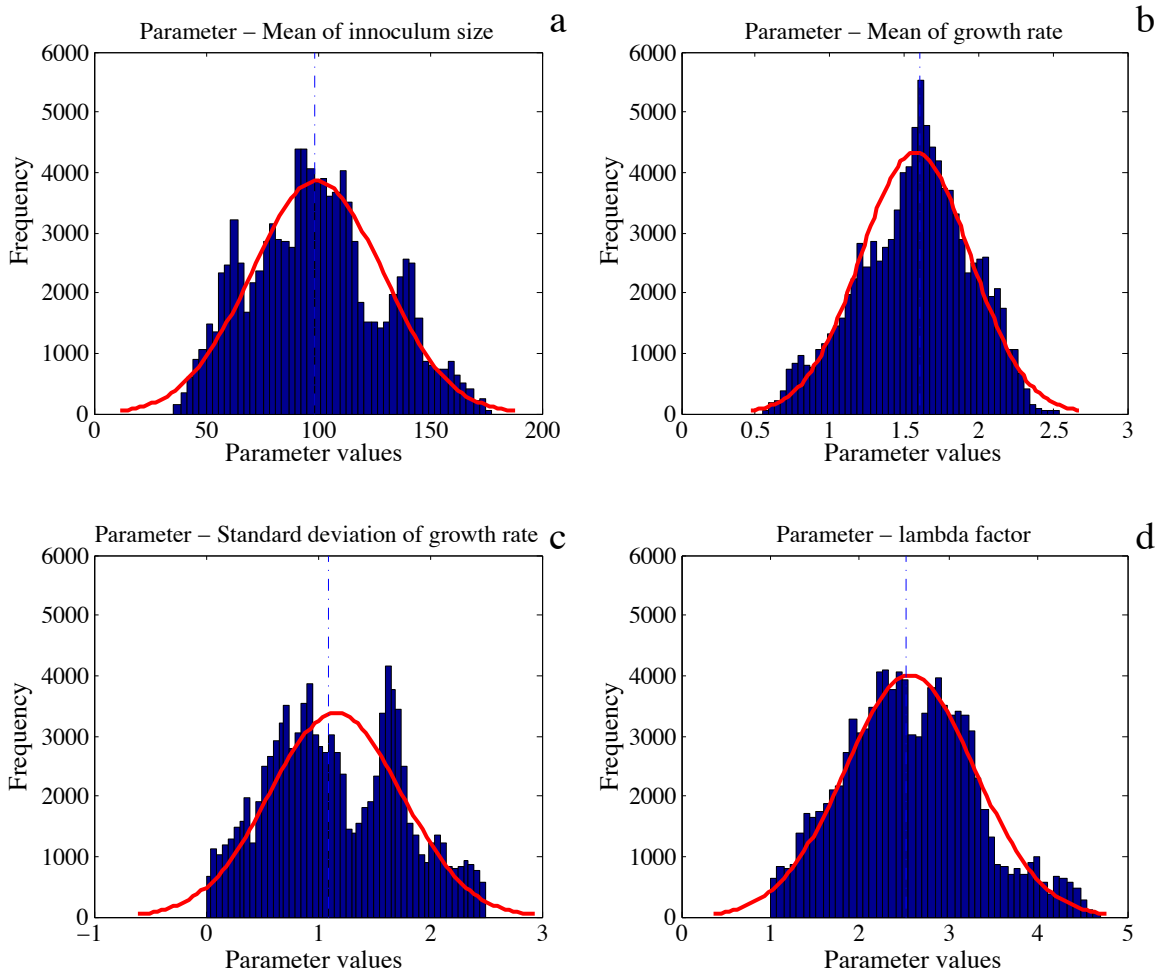


Figure S1 Parameter distributions found by the MCMC fitting protocol applied to controls from the CHMI studies. Red line shows fitting of a Gaussian curve, dashed line shows median of all values.

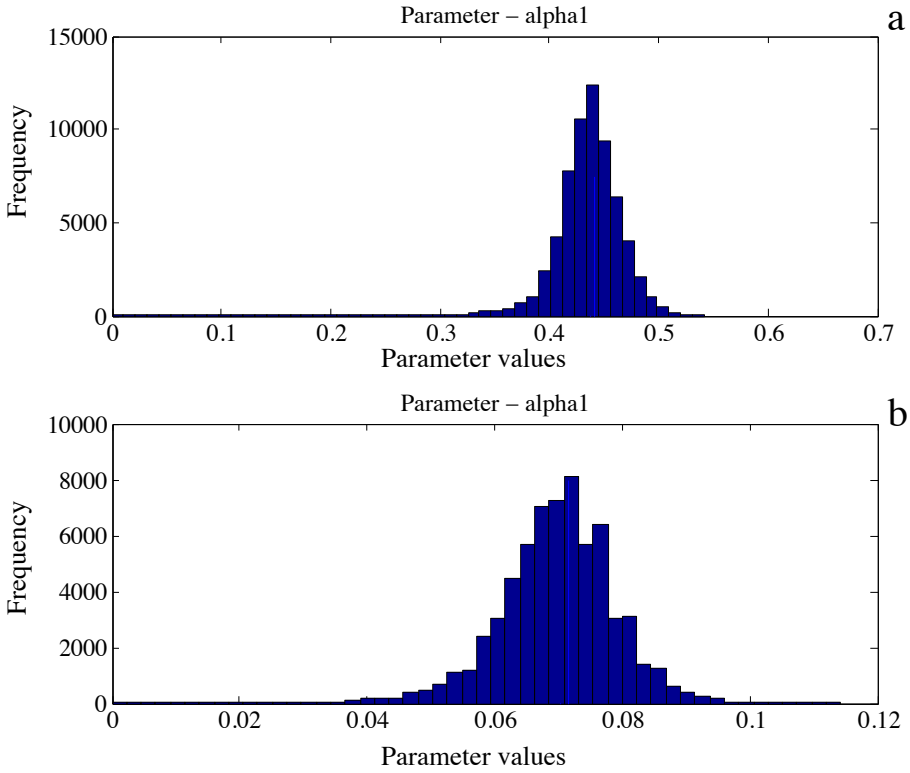


Figure S2 Parameter distributions of α_1 (a) and α_2 (b) found by MCMC fitting of the model to CHMI data for RTS,S (α_1) and ME-TRAP (α_2) vaccine recipients.

Appendix D: Published research paper (within-host modelling)

The following original research article was published in the American Journal of Hygiene and Tropical Medicine volume 93(6) in 2015.

Walker, A.S. et al., 2015. Modelling combinations of pre-erythrocytic plasmodium falciparum malaria vaccines. American Journal of Tropical Medicine and Hygiene, 93(6), pp.1254–1259.

Am. J. Trop. Med. Hyg., 93(6), 2015, pp. 1254–1259
doi:10.4269/ajtmh.14-0767
Copyright © 2015 by The American Society of Tropical Medicine and Hygiene

Modeling Combinations of Pre-erythrocytic *Plasmodium falciparum* Malaria Vaccines

Andrew S. Walker,* José Lourenço, Adrian V. S. Hill, and Sunetra Gupta

Department of Zoology, University of Oxford, Oxford, United Kingdom; The Jenner Institute, University of Oxford, Oxford, United Kingdom

Abstract. Despite substantial progress in the control of *Plasmodium falciparum* infection due to the widespread deployment of insecticide-treated bed nets and artemisinin combination therapies, malaria remains a prolific killer, with over half a million deaths estimated to have occurred in 2013 alone. Recent evidence of the development of resistance to treatments in both parasites and their mosquito vectors has underscored the need for a vaccine. Here, we use a mathematical model of the within-host dynamics of *P. falciparum* infection, fit to data from controlled human malaria infection clinical trials, to predict the efficacy of co-administering the two most promising subunit vaccines, RTS,S/AS01 and ChAd63-MVA ME-TRAP. We conclude that currently available technologies could be combined to induce very high levels of sterile efficacy, even in immune-naive individuals.

INTRODUCTION

It is estimated that, since 2000, global malaria-associated mortality has declined by 47%, largely due to increased distribution of insecticide-treated bed nets and artemisinin-based combination therapies.[1] However, it is widely accepted that vaccine research must continue as the international community pushes for elimination,[2] especially in light of recently emerged artemisinin resistance.[3] The malaria parasite has a complex lifecycle, each stage of which is the target of current vaccine research, from the initial sporozoite inoculum and subsequent liver stages to erythrocytic infection; vaccines are also being developed against the sexual stages with the aim of blocking transmission.

Pre-erythrocytic (PE) parasite stages have been targeted using either whole parasite approaches or recombinant vaccines.[4,5] Among the latter, the RTS,S subunit vaccine has shown moderate, short-term efficacy, with a large-scale phase III clinical trial completed in 2014.[6–8] Protection is conferred mainly by anti-circumsporozoite protein (CSP) antibodies, although a contribution from CSP-specific $CD4^+$ T cells that help antibody production cannot be ruled out.[9] An efficacy of around 50% (95% confidence interval [CI]: 32.9–67.1%) was shown in a phase IIa controlled human malaria infection (CHMI) trial when combined with AS01B, a liposome-based adjuvant system.[10] Another PE subunit vaccine ME-TRAP, which induces $CD8^+$ T-cell responses against infected hepatocytes, has been shown to provide 21% sterile protective efficacy in a phase IIa CHMI trial when administered by a prime-boost regimen (ChAd-MVA).[11] Both of these vaccines produce a substantial delay in the time to blood-stage infection among those subjects who do not show sterile protection, and levels of reduction in parasite numbers to achieve such a delay have been estimated to be in excess of 95% for both RTS,S[12] and ME-TRAP.[11]

It has been proposed that combinations of vaccines acting via distinct biological mechanisms could act synergistically.[13–18] The aforementioned vaccines, each with well in excess of a 90% reduction in parasite numbers[11,12] and discrete methods of action, could prove a potent combination. Such an effect has previously been noted in a murine model, in which a pair of T-cell- and antibody-inducing vaccines, each with around 30–35% sterile efficacy when administered alone, elicited 90% sterile efficacy upon their combination.[16] Here, we use a mathematical model to investigate the effects of combining anti-malarial vaccines acting at different PE stages of the life cycle. We used data from control subjects within phase IIa CHMI trials[10,11] to parameterize the within-host dynamics of the PE stages of *Plasmodium falciparum* and derive measures of the effects of RTS,S and ME-TRAP on infection using data from subjects who have received these vaccines. We show that high levels of sterile protection may be obtained by two vaccines which each show far lower efficacy when administered alone.

* Address correspondence to Andrew S. Walker, Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom. E-mail: andrew.walker@zoo.ox.ac.uk

METHODS

The model. Within-host parasite dynamics are modeled by the following system of equations describing the rates of change in numbers of infected hepatocytes (H) and merozoites (M), with parameters as described in Table 1:

$$\frac{dH}{dt} = -\alpha_2 H \tag{1}$$

$$\frac{dM}{dt} = \gamma M \tag{2}$$

$$H_0 = (1-\alpha_1) f S_0 \tag{3}$$

$$M_T = r H_T \tag{4}$$

Here, H_0 and M_T specify starting conditions for the respective populations, with $M = 0$ when $t < T$, where T = incubation period within the liver. There are several stochastic events in the development of malaria infection, made apparent in CHMI trials by differences between individuals in the time taken to reach defined thresholds of parasitemia. The model captures this behavior by incorporating inter-individual variation in the sporozoite inoculum (S_0), probability of liver invasion (f), incubation period (T), and erythrocytic growth rate (γ). We assume that an individual receives n bites, where the size of each sporozoite inoculum, S_i (such that the total initial inoculum $S_0 = \Sigma S_i$), is sampled from a negative binomial distribution obtained by fitting a negative binomial model to counts of parasites inoculated per-bite by anopheline mosquitoes.[19] The proportion of sporozoites invading liver cells (f) is drawn from a Poisson distribution with shape parameter

TABLE 1
Parameters and initial conditions in the model

Parameter	Explanation	Value
α_1	Vaccine-induced modifier of sporozoite invasion probability: $\alpha_{1,min} = 0$, $\alpha_{1,max} = 1$	MCMC fitted
α_2	Vaccine-induced rate of removal of infected hepatocytes: $\alpha_{2,min} = 0$, $\alpha_{2,max} = 1$	MCMC fitted
μ	Rate of sporozoite loss, set such that sporozoites are removed from system at a realistic rate	20
T	Liver incubation time: time from sporozoite inoculum to merozoite release	Selected from Gaussian distribution, mean = 7 days, SD = 0.5 days
r	Successful merozoites per hepatocyte	10,000
f	Proportion of sporozoites that successfully invade hepatocytes: if, $f > 1$, $f = 1$	Stochastically selected from Poisson distribution (see Methods)
λ	Shape determinant for Poisson cumulative density function from which sporozoite success rate is stochastically sampled: $\lambda_{max} = 10$; $\lambda_{min} = 1$	MCMC fitted
S_0	Initial inoculum size, based on sum of five samples (five bites) from negative binomial distribution	–
P	Success parameter for sampling from negative binomial to give per bite inoculum size	MCMC fitted
R	Shape parameter of inoculum size negative binomial[19]	0.246
γ_{mean}	Mean of EGR: $\gamma_{mean,max} = 5$; $\gamma_{mean,max} = 0$	MCMC fitted
γ_{sd}	SD of EGR: $\gamma_{sd,max} = 1.25$; $\gamma_{sd,max} = 0$	MCMC fitted
γ	EGR: if $\gamma < 0$ then $\gamma = 0$	Sampled from Gaussian of mean γ_{mean} and SD γ_{sd}

EGR = erythrocytic growth rate; MCMC = Markov chain Monte Carlo; SD = standard deviation.

$\varepsilon = S_0/\lambda$, where λ is a modifier of sporozoite success. Vaccine-induced anti-sporozoite immune responses reduce the probability of invasion by $(1 - \alpha_1)$, while infected hepatocytes are removed by a liver-stage vaccine at a rate α_2. Intrahepatocytic development takes approximately 7 days, and at the end of this period (T), the merozoite population size is set to rH_T, where r is the number of merozoites released per hepatocyte. Erythrocytic growth occurs at a rate γ, sampled from a Gaussian distribution with a given mean and standard deviation (SD; γ_{mean} and γ_{sd}, respectively).

The model was implemented in MATLAB R2013b (The MathWorks, Inc., Natick, MA), and the ODE system solved using a non-stiff Runge-Kutta solver, ode45.

Model fitting to trial data. The two studies used to parameterize the model[10,11] utilized similar challenge protocols, and there were no significant differences in time to patency ($t_{df} = 1.5409_{34}$, $P = 0.1326$) between the two trials, nor was there a difference between either study and an additional study investigating time to patency following *Anopheles* bites.[20] Similarly, there was no significant difference between the times to quantitative real-time polymerase chain reaction (qPCR) detection in the Ewer and Lyke studies ($t_{df} = 1.320_{22.755}$, $P = 0.2002$); qPCR data was not available for the RTS,S CHMI study.[10]

The combination of parameters that provides the best fit between the ODE model and the CHMI trial data was determined using a stepwise Markov chain Monte Carlo (MCMC) methodology,[21] detailed in Supplemental Methods. Each run of the model outputs the time from sporozoite inoculum to qPCR and blood slide detection of merozoites; these values are compared with CHMI data. Threshold levels of parasitemia were computed on the basis that 1) qPCR detection methods can reliably detect 20 parasites/mL of blood[22] and, given ~4.7 L of blood in an adult human, this equates to 94,000 parasites in a single host and 2) blood slide methods generally detect parasitemia when there are between 20 and 50 parasites/μL[23]: an estimated mean of 35 was therefore used, equating to 1.645×10^8 parasites across the human host.

First, a baseline model was parameterized (Supplemental Methods) using CHMI data from non-vaccinated individuals. The effects of administering two vaccines in combination were assessed by running the model 500 times under different parameters describing the proportional reduction in successful sporozoite infection and maturation of infected hepatocytes. These particular parameters for RTS,S and ME-TRAP were obtained by fitting the baseline model to CHMI data on vaccinated individuals.

The model was subjected to a sensitivity analysis by fixing (baseline model and vaccine) parameters at the values found by the MCMC procedure (Supplemental Table 2), and then varying each parameter independently about a Gaussian distribution with arbitrary SD. For each parameter value, the least squares distance of the model output to the trial data was calculated.

RESULTS

Baseline within-host dynamics. We obtained a mean time to patency by blood slide analysis of 11.046 days (SD = 1.205) as compared with 11.037 (SD = 1.512, $N = 53$) among the controls in the CHMI trials used in this study (Figures 1A and 2A). The mean time to the qPCR detection threshold of parasitemia in the model was 7.219 days (SD = 0.592), compared with 7.759 days (SD = 1.115, $N = 29$) in the CHMI trials (Figures 1A and 2B and Supplemental Table 1).

For completeness, we performed an additional sensitivity analysis (Supplemental Figure 1A–D) by resampling each parameter independently from a normal distribution with a mean of the value found by fitting. We found the goodness of fit of the model to be dependent on all parameters and, as expected, to maximize around the values found by the MCMC.

Within-host dynamics in single vaccine recipients. We obtained a sterile protective efficacy of 53.35% (SD = 5.20%) for RTS,S/AS01 compared with 50% (95% CI = 32.9–67.1%) reported by Kester and others[10] (Figure 1B). In those individuals who were not protected, the mean time to patency by blood slide analysis in the model was 11.898 days (SD = 1.222), compared with the CHMI result of 13.567 days (95% CI = 9.78–17.37; Supplemental Figure 2A and Supplemental Table 1). Time to qPCR detection in the model was 7.896 days (SD = 0.603); this cannot be compared with

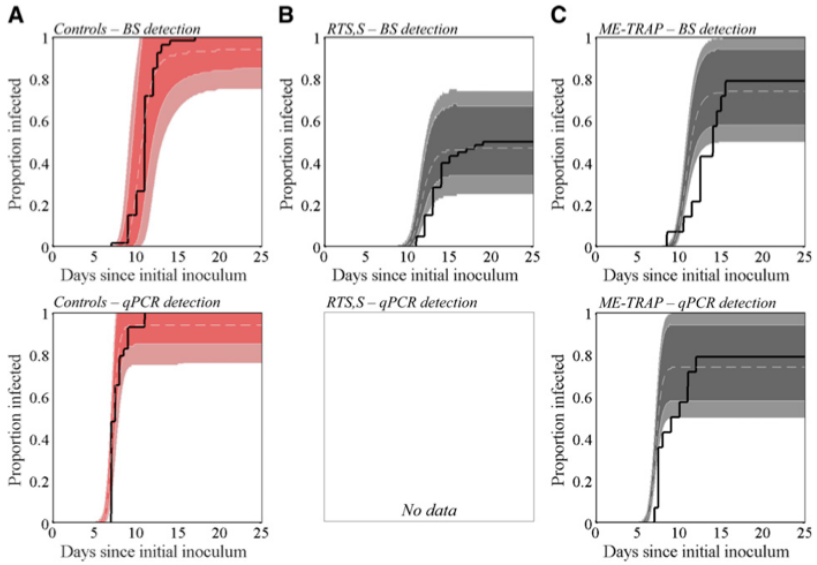
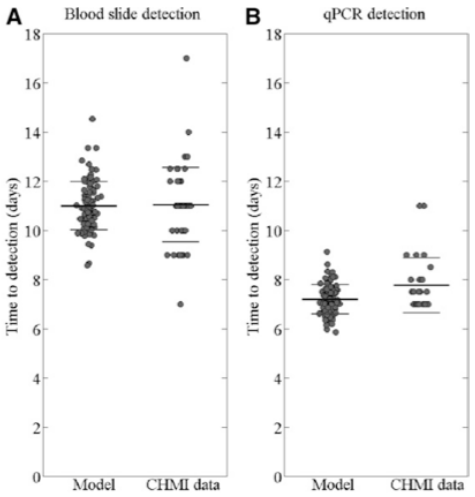


FIGURE 1. Survival curves showing the proportion of individuals who have reached the thresholds of blood-stage parasitemia required for detection by either blood slide (BS) or quantitative real-time polymerase chain reaction (qPCR) analysis—no qPCR data were available for the RTS,S/AS01B study analyzed. Solid lines represent data from CHMI trials that was used to fit (**A**) the baseline (no vaccine), (**B**) RTS,S/AS01B,[10] and (**C**) ChAd63-MVA ME-TRAP[11] models. The dashed white lines represent the median of all accepted chain steps of the Markov chain Monte Carlo (MCMC) protocol (post burn-in time), and dark and light shaded areas represent 90% and 99% credible intervals, respectively.



	Blood slide detection		qPCR detection	
	Model	CHMI data	Model	CHMI data
Mean time (days)	11.024	11.038	7.219	7.759
SD	0.985	1.512	0.592	1.115

FIGURE 2. Time to (**A**) blood slide patency and (**B**) detection by quantitative real-time polymerase chain reaction (qPCR) in CHMI trials[10,11,20] and the model for control (non-vaccine recipient) individuals. Model data is based on 500 runs with parameters fixed at the values found by Markov chain Monte Carlo (MCMC) fitting, 100 randomly selected runs were plotted. Mean and ± 1 standard deviation (SD) are shown (black/gray lines, respectively).

trial output, as qPCR data were not available in this trial (Figure 1B and Supplemental Figure 2B). This corresponds to a proportionate reduction in successful sporozoite numbers (α_1) of 0.901 (Supplemental Table 2). Overall, the fit of the model to the RTS,S (α_1) data is better than to ME-TRAP (α_2) data (Figure 4), which may in part be due to only blood slide patency data available for RTS,S, whereas both blood slide and qPCR data were available for ME-TRAP.

ChAd-63-MVA ME-TRAP elicited 22.72% efficacy (SD = 4.26%) in the model as compared with 21.4% (95% CI = 3.2–46.0%) in the study by Ewer and others.[11] The mean time to blood slide and qPCR detection in non-protected individuals in the model were 11.008 days (SD = 1.054) and 7.208 days (SD = 0.76), respectively, compared with 12.82 days (95% CI = 8.72–6.91) and 8.91 days (95% CI = 5.41–12.41) in the vaccine trial (Figure 1C, Supplemental Figure 2C, D, and Supplemental Table 1). This corresponds to a proportionate reduction in infected hepatocytes ($1 - \exp(-\alpha_2 \times T)$) of 0.780 (Supplemental Table 2).

Effects of combining vaccines. Sterile protective efficacies of combinations of vaccines acting at sporozoite and liver stages were evaluated by varying α_1 and α_2 to reflect a proportionate reduction in parasite numbers ranging from 0.5 to 1 (Figure 3). We found that the parameters identified for RTS,S and ME-TRAP from the CHMI data conferred very high levels of sterile protection when the vaccines were combined (Figure 3, intersection of double lines): 97.51% (SD = 1.52%) of infections were prevented in 500 runs of the model with α values fixed at the values found by MCMC fitting. A small increase in percentage killed for either vaccine would effectively

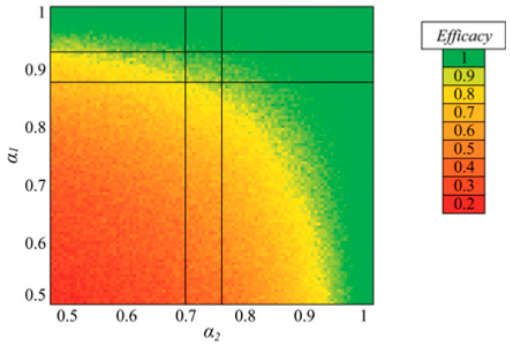


FIGURE 3. Proportional sterile protective efficacies of combinations of vaccines, with color indicating proportional sterile efficacy from green (total protection) to red (zero protection). The sets of intersecting lines indicate the calculated potencies of RTS,S/AS01B (horizontal line[10]) and ChAd63-MVA ME-TRAP (vertical line[11]) vaccines when used in monotherapy in CHMI trials: modeling their combination gives an efficacy between 93% and 99%.

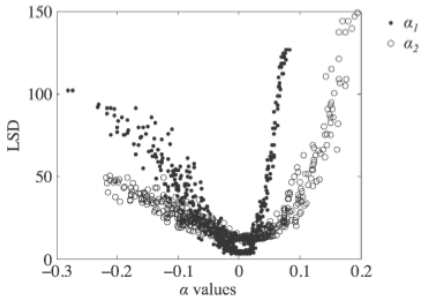


FIGURE 4. The effect of independently varying α_1 (•) and α_2 (○) on the least squares distance (LSD) between the model and the data, with all other parameters fixed. Each point (N = 500) represents the mean of 100 runs of the ODE model.

yield sterile protection in combination with the other, but only a very modest rate if used on its own.

DISCUSSION

An effective vaccine against *P. falciparum* malaria is widely accepted as an essential step toward eradication of the disease. In this study, we explore the possibilities for synergistic efficacy in combinations of PE vaccines. Underpinning this is the requirement of vaccine-induced immune responses to elicit sterile protection, since even in individuals in whom all but a few parasites are eradicated, infection is likely to occur, albeit with a slight delay. Previous studies have estimated two PE subunit vaccines currently in development, ChAd63-MVA ME-TRAP and RTS,S/AS01, to independently eradicate in excess of 90% of parasites,[11,12,] and yet sterile efficacies remain moderate to modest.[8,10,11,24,25] We suggest that combinations of vaccines with such strong parasite-killing effects will be significantly more efficacious than the aforementioned, although highly promising, single-vaccine approaches.

Our estimates of the combined efficacy of ChAd63-MVA ME-TRAP and RTS,S/AS01 would be sufficient, at high coverage, to eliminate malaria in areas where the transmission potential (R_0) of malaria is moderate[26] (~5), and would extend to areas of much higher transmission by even very slightly increasing the rate of parasite killing for either vaccine. There is considerable debate concerning the measurement of R_0 for malaria,[27,28] but methods that account for antigenic diversity and low rates of development of natural immunity[28] suggest maximum values that fall well within the range of possible elimination associated with the combined efficacy of ChAd63-MVA ME-TRAP and RTS,S/AS01. The duration of efficacy remains a problem in achieving the goal of elimination: to date, 13 PE and nine erythrocytic vaccines have entered clinical trials,[5] but very few have been found to demonstrate lasting protective efficacy in humans, despite high antibody titers or T-cell levels.[4] Thus, although our model supports the notion that a highly efficacious PE subunit vaccine may well be within reach, control of malaria is likely to be sustained only when used in conjunction with other tools[5] such as bednets[29] and fungal biopesticides.[30]

A paradox arises when trying to fit the model to data on infection dynamics among recipients of ME-TRAP (Figure 1C) in that the recorded delay to thresholds of parasitemia require less than one liver cell's worth of merozoites to be released. This suggests that ME-TRAP, as well as inducing total destruction of liver cells, may cause partial disruption of others, allowing them to release a partial load of merozoites. Partial release of merozoite load may also be a feature of natural infection, as noted by Bejon and others in trying to replicate infection dynamics among CHMI controls.[12] There are several potential mechanistic explanations for this phenomenon. It appears that hepatocytes do not "burst" per se, rather, vesicles called merosomes, each containing 100–200 merozoites, bud from the infected cell,[31] and it has been shown that merosome exit from the liver can be affected by inflammatory immune responses restricting blood flow.[32] Furthermore, the inherent stochasticity in the erythrocytic invasion process[12] may be partially responsible for the range in times to qPCR detection and patency by blood slide analysis seen in CHMI studies.[33] Finally, it is likely that there is heterogeneity in the spread of parasites across the body. For example, release of the merosome cargo has been shown in a murine model to occur mostly in the lungs[34]: parasites are released into deep vasculature, whereas blood for qPCR detection and thick smear analysis is taken from more peripheral sources.

We have assumed that there is no interaction between the T-cell- and antibody-mediated responses to ChAd63-MVA ME-TRAP and RTS,S/AS01. An important consideration in future work will be, when combining vaccines, whether interactivity exists between the mechanisms by which they act. Furthermore, certain candidate vaccines may affect more than one stage in the life cycle: for example, the blood-stage candidate apical membrane antigen 1 (AMA-1), also plays a role in hepatocyte invasion[35] and has demonstrated promising efficacy correlated with cell-mediated immunity when combined with a PE antigen.[36] Our model provides a platform to investigate the non-additive effects of the various combinations of vaccines that are likely to be tested in the near future in our ongoing battle against malaria.

Received December 1, 2014. Accepted for publication August 6, 2015.
Published online October 26, 2015.

Note: Supplemental methods, table, and figures appear at www.ajtmh.org.

Acknowledgments: We would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. Andrew S. Walker, José Lourenço and Sunetra Gupta are funded by the European Research Council (ERC Advanced Grant—Diversity). AVSH is a Jenner Investigator and Wellcome Trust and NIHR Senior Investigator.

Authors' addresses: Andrew S. Walker, José Lourenço, and Sunetra Gupta, Department of Zoology, University of Oxford, Oxford, United Kingdom, E-mails: andrew.walker@zoo.ox.ac.uk, jose.lourenco@zoo.ox.ac.uk, and sunetra.gupta@zoo.ox.ac.uk. Adrian V. S. Hill, The Jenner Institute, University of Oxford, Oxford, United Kingdom, E-mail: adrian.hill@ndm.ox.ac.uk.

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.

REFERENCES

1. World Health Organization, 2014. *World Malaria Report 2014*. Geneva, Switzerland: WHO.
2. Griffin JT, Hollingsworth TD, Okell LC, Churcher TS, White M, Hinsley W, Bousema T, Drakeley CJ, Ferguson NM, Basáñez M-G, Ghani AC, 2010. Reducing *Plasmodium falciparum* malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med 7:* e1000324.
3. Dondorp A, Nosten F, Yi P, 2009. Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med 361:* 455–467.
4. Schwartz L, Brown GV, Genton B, Moorthy VS, 2012. A review of malaria vaccine clinical projects based on the WHO rainbow table. *Malar J 11:* 11.
5. WHO, 2014. *Malaria Vaccine Rainbow Tables*. Available at: http://www.who.int/vaccine_research/links/Rainbow/en/index.html. Accessed January 6, 2015.
6. The RTSS Clinical Trials Partnership, 2011. First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children. *N Engl J Med 365:* 1863–1875.
7. The RTSS Clinical Trials Partnership, 2014. Efficacy and safety of the RTS,S/AS01 malaria vaccine during 18 months after vaccination: a phase 3 randomized, controlled trial in children and young infants at 11 African sites. *PLoS Med 11:* e1001685.
8. The RTSS Clinical Trials Partnership, 2015. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet 386:* 31–45.
9. White MT, Bejon P, Olotu A, Griffin JT, Riley EM, Kester KE, Ockenhouse CF, Ghani AC, 2013. The relationship between RTS,S vaccine-induced antibodies, $CD4^+$ T cell responses and protection against *Plasmodium falciparum* infection. *PLoS One 8:* e61395.
10. Kester KE, Cummings JF, Ofori-Anyinam O, Ockenhouse CF, Krzych U, Moris P, Schwenk R, Nielsen RA, Debebe Z, Pinelis E, Juompan L, Williams J, Dowler M, Stewart VA, Wirtz RA, Dubois M, Lievens M, Cohen J, Ballou WR, Heppner DG Jr, 2009. Randomized, double-blind, phase 2a trial of falciparum malaria vaccines RTS,S/AS01B and RTS,S/AS02A in malaria-naive adults: safety, efficacy, and immunologic associates of protection. *J Infect Dis 200:* 337–346.
11. Ewer KJ, O'Hara GA, Duncan CJA, Collins KA, Sheehy SH, Reyes-Sandoval A, Goodman AL, Edwards NJ, Elias SC, Halstead FD, Longley RJ, Rowland R, Poulton ID, Draper SJ, Blagborough AM, Berrie E, Moyle S, Williams N, Siani L, Folgori A, Colloca S, Sinden RE, Lawrie AM, Cortese R, Gilbert SC, Nicosia A, Hill AVS, 2013. Protective $CD8^+$ T-cell immunity to human malaria induced by chimpanzee adenovirus-MVA immunisation. *Nat Commun 4:* 2836.
12. Bejon P, Andrews L, Andersen RF, Dunachie S, Webster D, Walther M, Gilbert SC, Peto T, Hill AVS, 2005. Calculation of liver-to-blood inocula, parasite growth rates, and preerythrocytic vaccine efficacy, from serial quantitative polymerase chain reaction studies of volunteers challenged with malaria sporozoites. *J Infect Dis 191:* 619–626.
13. Ndungu FM, Mwacharo J, Kimani D, Kai O, Moris P, Jongert E, Vekemans J, Olotu A, Bejon P, 2012. A statistical interaction between circumsporozoite protein-specific T cell and antibody responses and risk of clinical malaria episodes following vaccination with RTS,S/AS01E. *PLoS One 7:* e52870.
14. Duffy P, Kaslow D, 1997. A novel malaria protein, Pfs28, and Pfs25 are genetically linked and synergistic as falciparum malaria transmission-blocking vaccines. *Infect Immun 65:* 1109–1113.
15. Hill AVS, 2006. Pre-erythrocytic malaria vaccines: towards greater efficacy. *Nat Rev Immunol 6:* 21–32.
16. Hutchings CL, Birkett AJ, Moore AC, Hill AVS, 2007. Combination of protein and viral vaccines induces potent cellular and humoral immune responses and enhanced protection from murine malaria challenge. *Infect Immun 75:* 5819–5826.
17. Moore AC, Hutchings CL, 2007. Combination vaccines: synergistic simultaneous induction of antibody and T-cell immunity. *Expert Rev Vaccines 6:* 111–121.
18. White MT, Smith DL, 2013. Synergism from combinations of infection-blocking malaria vaccines. *Malar J 12:* 280.
19. Medica DL, Sinnis P, 2005. Quantitative dynamics of *Plasmodium yoelii* sporozoite transmission by infected anopheline mosquitoes. *Infect Immun 73:* 4363–4369.
20. Lyke KE, Laurens M, Adams M, Billingsley PF, Richman A, Loyevsky M, Chakravarty S, Plowe CV, Sim BKL, Edelman R, Hoffman SL, 2010. *Plasmodium falciparum* malaria challenge by the bite of aseptic *Anopheles stephensi* mosquitoes: results of a randomized infectivity trial. *PLoS One 5:* e13490.
21. Lourenço J, Recker M, 2014. The 2012 Madeira Dengue Outbreak: epidemiological determinants and future epidemic potential. *PLoS Negl Trop Dis 8:* e3083.
22. Andrews L, Andersen RF, Webster D, Dunachie S, Walther RM, Bejon P, Hunt-Cooke A, Bergson G, Sanderson F, Hill AV, Gilbert SC, 2005. Quantitative real-time polymerase chain reaction for malaria diagnosis and its use in malaria vaccine clinical trials. *Am J Trop Med Hyg 73:* 191–198.
23. Guerin P, Olliaro P, Nosten F, 2002. Malaria: current status of control, diagnosis, treatment, and a proposed agenda for research and development. *Lancet Infect Dis 2:* 564–573.
24. Hodgson SH, Ewer KJ, Bliss CM, Edwards NJ, Rampling T, Anagnostou NA, de Barra E, Havelock T, Bowyer G, Poulton ID, de Cassan S, Longley R, Illingworth JJ, Douglas AD, Mange PB, Collins KA, Roberts R, Gerry S, Berrie E, Moyle S, Colloca S, Cortese R, Sinden RE, Gilbert SC, Bejon P, Lawrie AM, Nicosia A, Faust SN, Hill AVS, 2014. Evaluation of the efficacy of ChAd63-MVA vectored vaccines expressing circumsporozoite protein and ME-TRAP against controlled human malaria infection in malaria-naive individuals. *J Infect Dis 211:* 1076–1086.
25. Ogwang C, Kimani D, Edwards NJ, Roberts R, Mwacharo J, Bowyer G, Bliss C, Hodgson SH, Njuguna P, Viebig NK, Nicosia A, Gitau E, Douglas S, Illingworth J, Marsh K, Lawrie A, Imoukhuede EB, Ewer K, Urban BC, Hill AVS, 2015. Prime-boost vaccination with chimpanzee adenovirus and modified vaccinia Ankara encoding TRAP provides partial protection against *Plasmodium falciparum* infection in Kenyan adults. *Sci Transl Med 7:* 1–10.
26. Anderson RM, May RM, 1991. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, UK: Oxford University Press.
27. Smith DL, McKenzie FE, Snow RW, Hay SI, 2007. Revisiting the basic reproductive number for malaria and its implications for malaria control. *PLoS Biol 5:* e42.
28. Gupta S, Trenholme K, Anderson RM, Day KP, 1994. Antigenic diversity and the transmission dynamics of *Plasmodium falciparum*. *Science 263:* 961–963.
29. Artzy-Randrup Y, Dobson AP, Pascual M, 2015. Synergistic and antagonistic interactions between bednets and vaccines in the control of malaria. *Proc Natl Acad Sci USA 112:* 3014–3019.
30. Lynch PA, Grimm U, Thomas MB, Read AF, 2012. Prospective malaria control using entomopathogenic fungi: comparative evaluation of impact on transmission and selection for resistance. *Malar J 11:* 383.

31. Sturm A, Amino R, van de Sand C, Regen T, Retzlaff S, Rennenberg A, Krueger A, Pollok J-M, Menard R, Heussler VT, 2006. Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science 313:* 1287–1290.
32. Viriyavejakul P, Khachonsaksumet V, Punsawad C, 2014. Liver changes in severe *Plasmodium falciparum* malaria: histopathology, apoptosis and nuclear factor kappa B expression. *Malar J 13:* 106.
33. Roestenberg M, O'Hara GA, Duncan CJA, Epstein JE, Edwards NJ, Scholzen A, van der Ven AJAM, Hermsen CC, Hill AVS, Sauerwein RW, 2012. Comparison of clinical and parasitological data from controlled human malaria infection trials. *PLoS One 7:* e38434.
34. Baer K, Klotz C, Kappe SHI, Schnieder T, Frevert U, 2007. Release of hepatic *Plasmodium yoelii* merozoites into the pulmonary microvasculature. *PLoS Pathog 3:* e171.
35. Silvie O, Franetich JF, Charrin S, Mueller MS, Siau A, Bodescot M, Rubinstein E, Hannoun L, Charoenvit Y, Kocken CH, Thomas AW, Van Gemert GJ, Sauerwein RW, Blackman MJ, Anders RF, Pluschke G, Mazier D, 2004. A role for apical membrane antigen 1 during invasion of hepatocytes by *Plasmodium falciparum* sporozoites. *J Biol Chem 279:* 9490–9496.
36. Chuang I, Sedegah M, Cicatelli S, Spring M, Polhemus M, Tamminga C, Patterson N, Guerrero M, Bennett JW, McGrath S, Ganeshan H, Belmonte M, Farooq F, Abot E, Banania JG, Huang J, Newcomer R, Rein L, Litilit D, Richie NO, Wood C, Murphy J, Sauerwein R, Hermsen CC, McCoy AJ, Kamau E, Cummings J, Komisar J, Sutamihardja A, Shi M, Epstein JE, Maiolatesi S, Tosh D, Limbach K, Angov E, Bergmann-Leitner E, Bruder JT, Doolan DL, King CR, Carucci D, Dutta S, Soisson L, Diggs C, Hollingdale MR, Ockenhouse CF, Richie TL, 2013. DNA prime/Adenovirus boost malaria vaccine encoding *P. falciparum* CSP and AMA1 induces sterile protection associated with cell-mediated immunity. *PLoS One 8:* e55571.

SUPPLEMENTAL METHODS

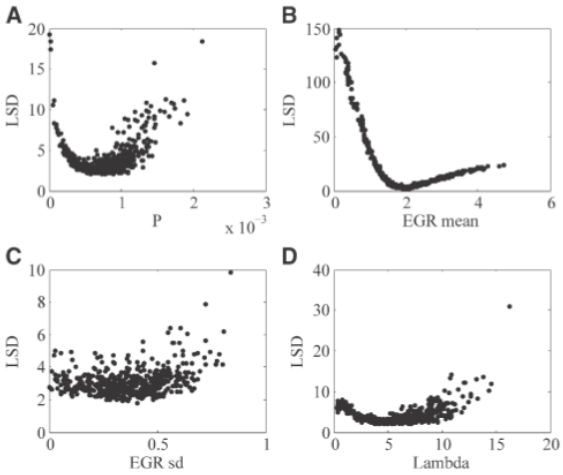
Each step of the Markov chain Monte Carlo (MCMC) routine effectively represents one clinical trial, with 100 runs of the ODE model completed with parameters at values set to the current MCMC state: the proportions of these runs that reach the two parasite density thresholds (time to blood slide/quantitative real-time polymerase chain reaction [qPCR]) comprise the main output of the model. The least squares distance (LSD) between this output ($\chi^2(O(Y))$ and the CHMI trial data ($\chi^2(O(M^t))$ was calculated, and the likelihood that the given parameter set describes the observed trial data was computed using the expression:

$$\rho = e^{\chi^2(O(Y)) - \chi^2(O(M^t))}$$

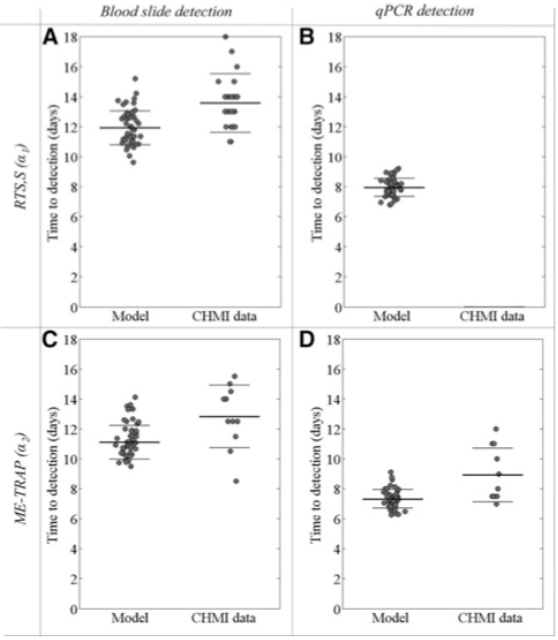
where ρ is the MCMC probability of acceptance, O the ODE model output from either Y, the proposed parameter set, or M^t, the accepted parameter set at time t. Increases in LSD are assumed to exponentially decrease the probability of acceptance. A value between 0 and 1 is selected from a uniform distribution, and if this value is less than ρ, the proposed parameter set is accepted, thus making $M^{t+1} = Y$.

To establish a baseline parameter set, the ODE model was fit to data[10,11,20] for individuals who did not receive a vaccine (parameters marked "MCMC fitted" in Table 1). First, to establish a primary parameter set, a set of independent MCMC chains were run with random starting conditions. After convergence, the parameter distributions from each chain were used to inform an initial state for a single MCMC chain of 500,000 steps, with all accepted parameter sets gathered for further analysis. From this single chain an optimum parameter set was found by calculating the median values of the 100,000 best fitting chains (by lowest LSD).

The previously sought parameters were then fixed, and the model rerun for two independent sets of 500,000 MCMC steps, now fit to data from vaccine recipients in each of the two CHMI trials: α_1 was fit to data from the RTS,S recipients and α_2 subsequently fit for ME-TRAP recipients. As only a single parameter (α_1 or α_2) was varied at a time, these chains converged more quickly than the baseline MCMC model fitting (four parameters).



SUPPLEMENTAL FIGURE 1. Results of a sensitivity analysis conducted on the model, demonstrating change in least squares distance (LSD) between the model and the controlled human malaria infection (CHMI) trial data as a result of varying single parameters about a mean set by the Markov chain Monte Carlo (MCMC) procedure. The probability parameter of the negatively binomial inoculum distribution (**A**), γ_{mean} (**B**), γ_{sd} (**C**), and λ factor (**D**) are varied in a system with no vaccine.



SUPPLEMENTAL FIGURE 2. Distributions of time to (**A** and **B**) blood slide patency and (**B** and **D**) quantitative real-time polymerase chain reaction (qPCR) thresholds of blood-stage parasitemia under vaccination with (**A** and **B**) RTS,S and (**C** and **D**) ME-TRAP, comparing the controlled human malaria infection (CHMI) trial data and the model outputs. Model data show 50 points randomly selected from 500 runs of the model. Horizontal lines indicate means and standard deviations of all outputs/CHMI trial groups. No qPCR data were available for the RTS,S CHMI trial—the model was solely fit to blood slide patency data.

SUPPLEMENTAL TABLE 1

Times to blood slide detection and qPCR detection thresholds under treatment by RTS,S or ME-TRAP in both the model and CHMI data.

		Blood slide detection		qPCR detection	
		Model	CHMI	Model	CHMI
RTS,S	Mean	11.921	13.567	7.949	n/a
	SD	1.276	1.942	0.605	n/a
ME-TRAP	Mean	11.100	12.818	7.317	8.909
	SD	1.266	2.089	0.627	1.786

SUPPLEMENTAL TABLE 2

Parameter values estimated by the MCMC fitting procedure

			Quantile		
		Parameter	0.05	0.5	0.95
Baseline model	P	Inoculum negative binomial parameter	0.000567	0.00111	0.00555
	λ	Sporozoite success modifier	1.246	5.763	9.003
	γ_{mean}	Mean of EGR	1.569	2.175	3.194
	γ_{sd}	SD of EGR	0.0652	0.546	1.143
Under vaccination	α_1	Vaccine-induced modifier of sporozoite invasion probability	0.854	0.907	0.935
	α_2	Vaccine-induced rate of removal of infected hepatocytes	0.116	0.216	0.305

EGR = erythrocytic growth rate; MCMC = Markov chain Monte Carlo; SD = standard deviation.