THE STABILITY OF HOST-PATHOGEN MULTI-STRAIN MODELS



Susan Hawkins

Pembroke College

Department of Zoology

A thesis submitted for the degree of *Doctor of Philosophy*

Trinity 2017

The Stability of Host-Pathogen Multi-Strain Models

Susan Hawkins

Doctor of Philosophy, DPhil

Pembroke College, Trinity 2017

Abstract

Previous multi-strain mathematical models have elucidated that the degree of cross-protective responses between similar strains, acting as a form of immune selection, generates different behavioural states of the pathogen population. This thesis explores these multi-strain dynamic states, to examine their robustness and stability in the face of pathogenic intrinsic phenotypic variation, and the extrinsic force of immune selection. This is achieved in two main ways: Chapter 2 introduces phenotypic variation in pathogen transmissibility, testing the robustness of a stable pathogen population to the emergence of an introduced strain of higher transmission potential; and Chapter 3 introduces a new model with a possibility of immunity to both strain-specific and cross-strain (conserved) determinants, to investigate how heterogeneity in the specificity of a host immune response alters the pathogen population structure. A final investigation in Chapter 4 develops a method of reverse-pattern oriented modelling using a machine learning algorithm to determine which intrinsic properties of the pathogen, and their combinations, lead to particular disease-like population patterns. This research offers novel techniques to complement previous and ongoing work on multi-strain modelling, with direct applications to a range of infectious agents such as *Plasmodium falciparum*, influenza A, and rotavirus, but also with a wider potential for other multi-strain systems.

1 TABLE OF CONTENTS

1 INTRODUCTION 4

1.1 SINGLE STRAIN MODELS 4

1.2 MULTI-STRAIN MODELS 7

1.3 THE MULTI-LOCUS MODEL 10

1.4 THE EFFECT OF IMMUNE SELECTION ON PATHOGEN POPULATION DYNAMICS 12

1.5 EVIDENCE OF MULTI-LOCUS MODEL BEHAVIOUR 16

1.5.1 Evidence of discrete strain structure 16

1.5.2 Evidence of cyclical strain structure 17

1.6 EXPLORATION OF THE BOUNDARIES OF MULTI-LOCUS MODEL BEHAVIOUR 19

1.7 THESIS SUMMARY 20

1.7.1 Effects of phenotypic heterogeneity within the pathogen population on the stability of multi-locus systems 22

1.7.2 Effects of heterogeneity in host immune response on the stability of multi-locus systems 22

1.7.3 Qualitative model fitting through machine learning 24

2 DISRUPTION OF DISCRETE STRAIN STRUCTURE THROUGH PHENOTYPIC CHANGE 26

2.1 INTRODUCTION 26

2.2 METHODS 30

2.3 RESULTS 33

2.3.1 Analytic derivation of invasion thresholds 33

2.3.2 Numerical simulations investigating the invasion threshold 39

2.3.3 The effects of the size of the initial seed 42

2.3.4 The duration of infectiousness alters the size of the initial epidemic peak of invading strains 43

2.3.5 Mutation and migration shifts the R_0 required for invasion 45

2.3.6 Parameter variation and seasonally forced models allow invasion when R_0^{inv} is low 49

2.4 DISCUSSION 54

3. THE EFFECTS OF HETEROGENEITY IN HOST IMMUNE RESPONSES ON PATHOGEN POPULATION STRUCTURE 61

3.1 INTRODUCTION 61

3.2 METHODS 69

3.3 RESULTS 73

3.3.1 The effect of immunity to conserved determinants on strain structure 73

3.3.2 The influence of R_0 on the cross-protection thresholds 77

3.3.3 The effects of differences in R_0 between strains 79

3.3.4 The effect of immunity to conserved determinants on the oscillations within CSS 80

3.4 DISCUSSION 86

4 DETERMINING ROTAVIRUS-LIKE BEHAVIOURS WITH A RANDOM FOREST APPROACH . 92

4.1 INTRODUCTION 92

4.1.1 The biology of rotavirus 94

4.1.2 The population dynamics of rotavirus 98

4.1.3 Modelling of rotavirus population dynamics 100

4.2 METHODS 101

4.2.1 The random forest algorithm 101

4.2.2 Multi-locus modelling for the time series output 103

4.2.3 Definitions of RV-like dynamics 105

4.2.4 Predictors of RV-like dynamics used in the random forest 108

4.2.5 Measures of random forest performance 109

4.3 RESULTS 110

4.3.1 Strain heterogeneity Vs. cross-protection 110

4.3.2 Changing cross-protection without including strain heterogeneity 116

4.3.3 Changing cross-protection while including strain heterogeneity 118

4.3.4 Changing strain heterogeneity without including variation in cross-protection 122

4.4 DISCUSSION 126

4.5 APPENDIX 132

4.5.1 Perfecting the definition of RV-like dynamics 132

4.5.2 Choosing which predictors to use 134

4.5.3 Table of all models used for the RFA 135

5 DISCUSSION 136

BIBLIOGRAPHY 143

1 INTRODUCTION

Epidemiological models are useful tools to understand the dynamics of pathogens, and the effects of evolutionary and ecological pressures. Available models differ widely in their complexity, from modelling the behaviour of a single strain to incorporating multiple strains, and including a variety of epidemiological factors such as seasonality, host contact networks, and vector behaviour. There are generally two types of model: those which are data-driven and applied to particular real-world scenarios, and those used for theoretical insight, which deal more generally with qualitative rather than quantitative data properties. The multi-strain models described here are for theoretical understanding rather than predictive ability, to understand qualitative properties of multi-strain systems.

It is important to understand the stability of these multi-strain systems to comprehend the destabilising factors which may be involved, and the consequential effect on strain evolution and ecology. This is particularly important for pathogens of multiple strains, not simply because of the interactions between strains, but from a public health point of view: these pathogens, such as influenza and malaria, are typically of high concern, with a high prevalence and high mortality.

1.1 SINGLE STRAIN MODELS

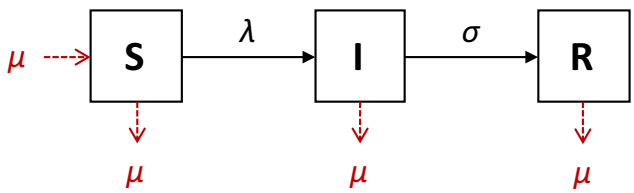
A simple compartmental model of epidemiology, first introduced by Kermack and McKendrick (Kermack & McKendrick 1927), tracks the number of susceptible (S), infected (I), and recovered (R) hosts (Figure 1.1a). It is assumed that after infection with the pathogen, the host recovers to then acquire life-long immunity.

The SIR model is deterministic, and uses ordinary differential equations (ODEs) to track the rate of change of the S, I, and R compartments over time.

Susceptible hosts are infected by the pathogen with a force of infection (λ), dependent on the number of infectious hosts and a transmission coefficient (β), or other additional constants depending on the pathogen's life history strategy (e.g. vector biting rate, host contact rate, host incubation period, etc.). After infection, a host moves into the recovered class by the recovery rate (σ), which in effect is the inverse of the number of days the host remains infectious. Demography can be included into the model (in red in Figure 1.1a) by adding a background death rate (μ) from each class, and a birth rate (b) into the susceptible class. The birth rate may be made equal to the total death rate to keep the population size constant, a simplifying assumption widely used for epidemiological models.

The transmission potential of the pathogen is determined by the basic reproduction number, $R_{0,}$ which may be interpreted as the number of secondary cases arising from a single primary case in a totally susceptible population (Anderson & May 1992). In the SIR model in Figure 1.1, this R_0 is dependent on the transmission coefficient (β) and the replenishment of susceptible hosts via births (μ) and the loss of infectiousness (σ), as so: $R_0 = \frac{\beta}{\sigma + \mu}$. If $R_0 > 1$, the disease can spread, resulting in an epidemic (Anderson & May 1992) and eventually endemic equilibrium, after an extended period where the susceptible and recovered classes display damped oscillations (Figure 1.1b). These damped oscillations are a fundamental property of an SIR model, resulting from fluctuations in the replenishment of susceptible hosts (dynamic resonance), with

a)



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

$$\frac{dI}{dt} = \lambda S - \sigma I - \mu I$$

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

b)

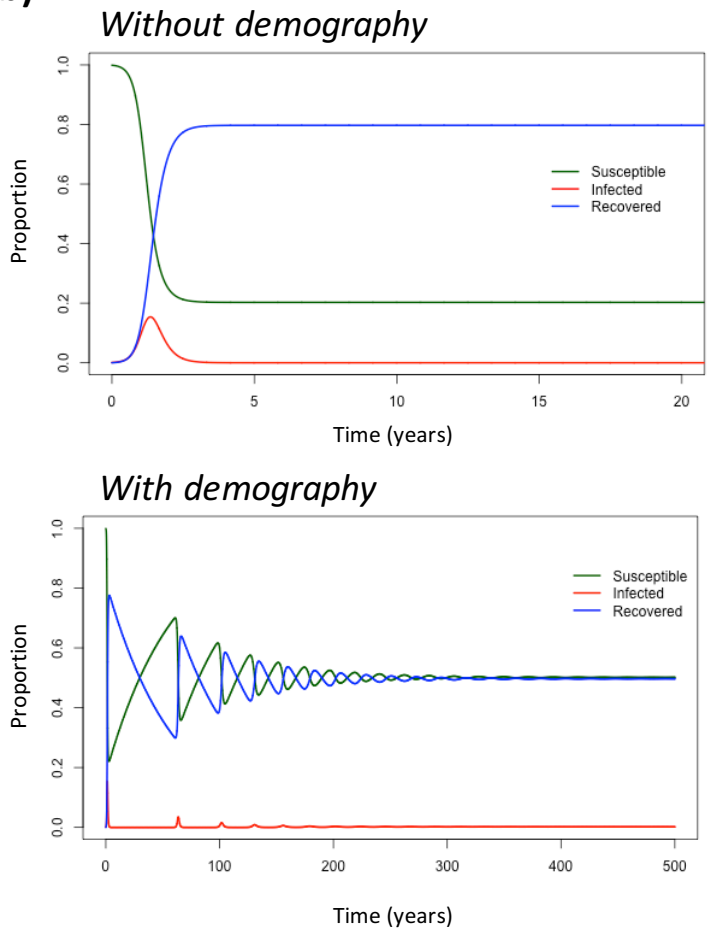


Figure 1.1 A) The basic SIR model, with demography included in red. μ = birth rate, σ = recovery rate, λ = force of infection (here defined as λ=βI where β is the transmission coefficient). **B) The output of the SIR model with and without demography.** Without demography, there is only one epidemic and the system quickly reaches equilibrium. With the inclusion of demography, the susceptible and recovered populations display damped oscillations about an equilibrium. Birth rate (μ) = 1/60 (60 years lifespan), rate of recovery (σ) = 5 (days), transmission coefficient (β) = 10. Both models have a naïve initial population (S = 0.999, R = 0) with a proportion of 0.001 infected hosts.

the oscillation period dependent on the duration of infection and the birth rate of new susceptibles (Anderson & May 1992). Oscillations can be sustained by the effects of seasonality or demographic stochasticity (Anderson & May 1992; Rohani et al. 2002; Keeling et al. 2001; Grassly & Fraser 2006).

1.2 MULTI-STRAIN MODELS

Many important diseases, such as malaria and influenza, circulate as multiple distinct strains. A strain is defined as a genetic variant that is recognised differently by the immune system as to other genetic variants. To understand the evolutionary dynamics of these pathogens, the SIR model needs to expand to incorporate multiple strains, and to include an ecological interaction between them as strains compete for susceptible hosts.

Incorporating multiple strains, however, quickly becomes problematic and computationally intensive. For one, to extend an SIR model to include n strains, the number of classes grows exponentially, with $(n + 2)2^{n-1}$ variables required (Andreasen et al. 1997). An alternative to using multiple analogous SIR equations for n strains, which quickly becomes cumbersome, is to use individual-based models, or history- or status-based models.

Individual-based models (IBMs) (Buckee et al. 2004; Bedford et al. 2012; Bedford et al. 2015) track the infection history of each individual host, which is a useful tool to incorporate host contact networks and heterogeneity between individuals. These models define rules of transmission for a single host, and this host is then multiplied, removing the need for a multitude of equations and variables. Nevertheless, this technique has the caveat of being highly computationally

intensive and difficult to track, and so history- and status-based models have been popular for multi-strain modelling.

History-based models provide the history of previous strain exposures, but not the current immune status (Castillo-Chavez et al. 1989). Status-based models (Gog & Swinton 2002; Gog & Grenfell 2002; Kryazhimskiy et al. 2007) provide the proportion of the population immune to each strain (the current immune status), but provide no history of previous infections: the status-based method allows a theoretically unlimited expansion of strains, but in doing so sacrifices its tractability.

Regardless of the method used, the inclusion of multiple strains in an epidemiological model can sustain oscillations due to the ecological interaction between strains, akin to the effects of stochasticity and seasonality. Interactions between strains can occur by: 1) transmission suppression of strains (Andreasen et al. 1997; Dawes & Gog 2002; Gupta et al. 1998; Gog & Grenfell 2002; Gomes et al. 2002); 2) enhancement of transmission of strains (e.g. antibody-dependent enhancement in dengue (Ferguson et al. 1999; Wikramaratna et al. 2010; Recker et al. 2009)); 3) interference (e.g. measles and whooping cough interact via removing the other's susceptible individuals (Rohani et al. 2003)). This work will focus on the former, where similar strains compete for susceptible hosts via cross-protection.

Cross-protection can act to either reduce transmission or susceptibility, though in regard to the model of Gupta et al (Gupta et al. 1996), used in this thesis, the method of reduction used makes little difference to the pathogen dynamics (Ferguson & Andreasen 2002). It has been argued that to have a large strain

space within a status-based model, this can only be achieved if cross-protection acts to reduce transmission (Kucharski et al. 2016): it is mathematically easier to assume cross-protection reduces transmission, as infection can be independent of immune status, though this has a caveat of being biologically unrealistic (Kryazhimskiy et al. 2007). This is because immunological evidence shows that infection is dependent on immune status, which prevents infection of pathogens which the host's immune system can recognise.

Cross-protection between strains is biologically a function of genetic similarity, and specifically by the part of the genome which determines recognition by the host, i.e. the antigenic loci. Strain similarity has been based on genetic distance on a linear scale (where cross-protection is to nearest neighbours (Andreasen et al. 1997), or using a Gaussian distribution to neighbouring strains (Gog & Grenfell 2002)), and with a circular arrangement of strains where adjacent strains elicit a cross-protective response (Gog & Swinton 2002). Gog et al utilised cross-protection differently, assuming polarised immunity: only some hosts gain full immunity to strains of shared antigenic types, whereas other hosts acquire no immunity at all (Gog & Swinton 2002; Gog & Grenfell 2002; Dawes & Gog 2002). Here, the focus will be on the model of (Gupta et al. 1998), where similarity is based on the sharing of alleles at particular antigenic loci.

The multi-locus model of Gupta et al (Gupta et al. 1996; Gupta et al. 1998) is referred to as being history-based (Kucharski et al. 2016; Gog & Swinton 2002), as hosts move to different immune compartments depending on their infection history. However, this model only tracks the immune status as a proportion of the population at each step, and lacks tractability as the full infection history of each

individual is unknown. These two characteristics are those of a status-based method, and so I will classify the Gupta model as status-based within this thesis.

1.3 THE MULTI-LOCUS MODEL

In the multi-locus model developed by Gupta et al (Gupta et al. 1996), strains are defined by their combination of alleles at each antigenic locus, so that an [n m] system has two loci of *n* and *m* alleles each. Figure 1.2 displays the strains for a [2 2] antigenic structure, with A/B alleles at one locus and X/Y at the other.

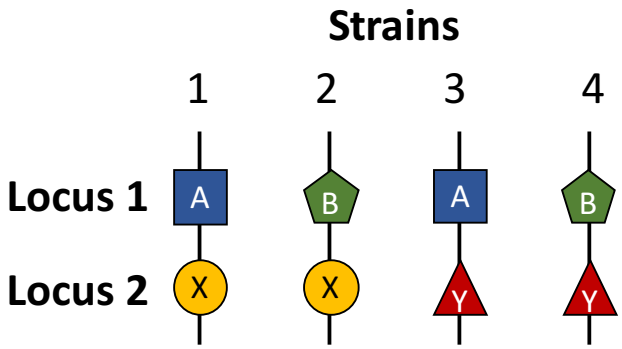


Figure 1.2 The antigenic structure of strains for a [2 2] antigenic system. Strains are defined by their combination of variable determinants at antigenic loci (shapes).

The model framework is shown in Figure 1.3, where individuals immune to a strain occupy a *Z* circle, which also provides cross-protective immunity to other strains where the circles overlap, creating the compartment *W.* Cross-protection provides a degree of immune selection by, after exposure to strain *i,* reducing the probability of infection to any strain that shares an allele with strain *i* (these strains are defined with the notation *j~i*), but not *i* itself. Recognition of a strain is assumed to provide life-long immunity.

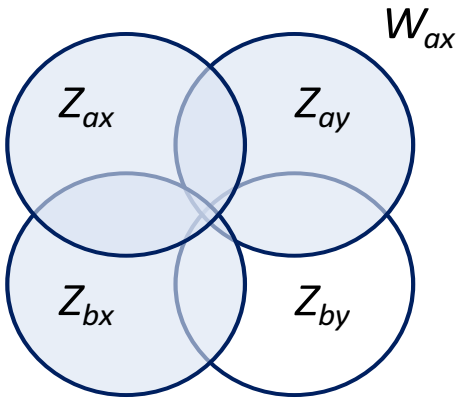


Figure 1.3 The multi-locus framework, based on (Gupta et al. 1996). The proportion immune to each strain are in the Z classes, and those immune to any strain containing the same antigens of strain *i*, including *i* itself, are in W classes. Those susceptible to strain AX are outside of Z_{ax} and W_{ax}.

Using the [2 2] antigenic structure, cross-protection specifically acts as so: after infection with strain AX (later recovering to Z_{ax}), the host will acquire cross-protection to any strain containing the antigenic alleles A (in the first locus) or X (in the second locus) by entering the cross-protected compartment W_{ax}. This reduces the chance of infection by strains AY and BX, but the host will remain completely susceptible to strain BY.

The parameters used in this model can be found in Table 1.1, with the ODEs as follows:

Proportion immune to strain i*,* Z_i*:*

$$\frac{dZ_i}{dt} = \lambda_i(1 - Z_i) - \mu Z_i \tag{1.1}$$

Proportion immune to any strain which shares alleles with i*, including* i *itself,* W_i*:*

$$\frac{dW_i}{dt} = \sum_{j\sim i} \lambda_j (1 - W_i) - \mu W_i \tag{1.2}$$

Proportion infectious with strain i*,* Y_i*:*

$$\frac{dY_i}{dt} = \lambda_i[1 - W_i + (1 - \gamma)(W_i - Z_i)] - (\sigma + \mu)Y_i \tag{1.3}$$

Parameter	Definition
μ	Background death rate and birth rate
σ	Recovery rate from infection
β	The transmission coefficient, which defines the force of infection, λ, from the equation $\lambda_i = \beta Y_i$ where Y_i is the proportion of hosts infectious with strain *i.*
γ	Degree of cross-protection, defined as the reduction of the probability of transmission to strains which share alleles with strain *i*. If γ = 1, after infection with strain *i* a host is completely protected against all strains sharing alleles, but if γ = 0 this host will only be protected against strain *i*.

Table 1.1 Parameter definitions of the multi-locus model

1.4 THE EFFECT OF IMMUNE SELECTION ON PATHOGEN POPULATION DYNAMICS

The earlier epidemiological models of two strain systems started to elucidate the effects of inter-strain competition, with the interaction between strains sustaining oscillations (Castillo-Chavez et al. 1989; Gupta et al. 1994), which are otherwise damped for one-strain deterministic SIR models. Within these two-strain systems, of only one antigenic locus, coexistence is only possible with low cross-immunity – however, where multiple antigenic loci are included and therefore a wider number of strains are considered, the population patterns of cyclical dynamics and coexistence become more complex. These population patterns are illustrated in Figure 1.4a, for varying levels of cross-protection.

When cross-protection is low or very high, the system reaches two steady states: 1) no strain structure (NSS) where all strains coexist, as low immune selection dictates that competition for susceptible individuals is weak and therefore strains can circulate independently, and 2) discrete strain structure (DSS) where a

discordant set of strains, with non-overlapping antigens, coexist whilst all other strains are removed from the pathogen population.

In DSS, the inter-strain competition is too high for coexistence of strains which share antigens – the pathogen population self-organises into discordant strains, while the other strains are suppressed by competitive exclusion. The discordant set of strains can coexist as they do not suppress each other's transmission via cross-protection, and therefore circulate independently, as in the NSS state. DSS behaviour has been demonstrated in numerous multi-strain mathematical models (Gupta et al. 1996; Gupta et al. 1998; Gog & Grenfell 2002; Artzy-Randrup et al. 2012).

The third, unstable, dynamic state is that of cyclical strain structure (CSS), at an intermediate level of cross-protection. Strains interfere with each other's transmission but not enough to permanently exclude other strains sharing alleles, creating oscillatory behaviour. Close to the NSS and DSS boundaries, strains cycle within the discordant sets of strains (Gupta et al. 1998), however in the middle of the CSS region this pattern is lost and the patterns become more chaotic and impossible to predict (the oscillatory behaviour, and boundaries of the three behaviours, can be seen in the bifurcation diagram in Figure 1.4b). Status-based models which use polarised immunity as a method of cross-protection have not found this oscillatory behaviour (Gog & Swinton 2002), analytically examined in (Dawes & Gog 2002).

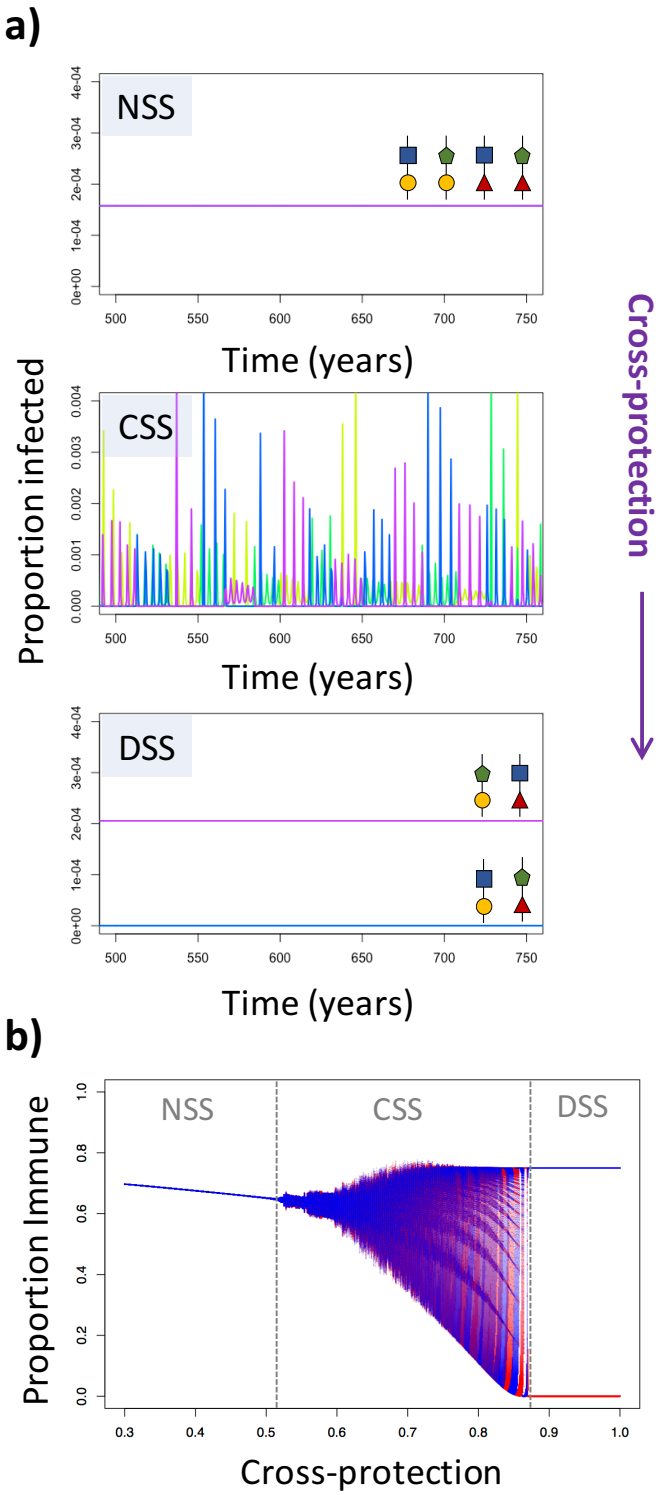


Figure 1.4 A) Time series for a [2 2] antigenic structure. As cross-protection increases, the pathogen population shifts behaviours. There are two stable states with the co-circulation and persistence of either all strains (top, NSS) or a discordant set of strains with non-overlapping antigens (bottom, DSS) at each end of the cross-protection spectrum. At intermediate levels, the population structure becomes cyclic (CSS), where strains serially replace each other for dominance. **B) Bifurcation diagram of the proportion of immune hosts.** The shift across the three behaviours can be seen, with changes to the strain amplitude in the mid-range of cross-protection. Birth rate (μ) = 1/50 (50 years lifespan), rate of recovery (σ) = 73 (equivalent to 5 days), transmission coefficient (β) = 292, and therefore R_0 = 4.) The initial conditions are a susceptible population, and all strains have an infected proportion of 10^{-5}, other than strain 4 with a proportion of 10^{-4}. Red=strain 1, blue= strain 4 (i.e. discordant pair).

The critical cross-protection threshold (γ*) for the appearance of DSS (Gupta et al. 1996), depends on the transmission potential of the dominating strains. For the [2 2] antigenic system, of two dominant strains (*i* and *i'*), γ* is defined as:

$$\gamma * = \frac{1}{1 - R_0^i R_0^{i'}}$$

The influence of cross-protection and R_0 on the appearance of CSS and DSS can be seen in Figure 1.5. As the R_0 for all strains increases, the system destabilises, and a higher level of cross-protection is necessary for DSS.

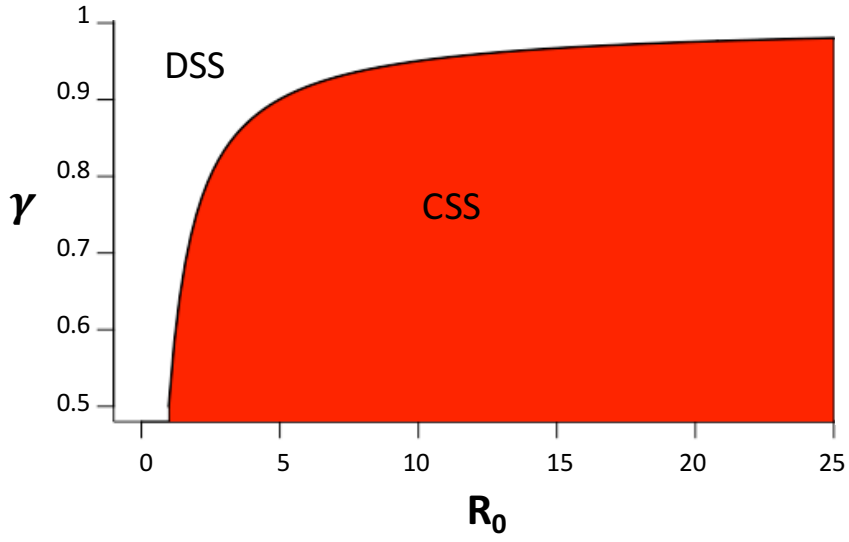


Figure 1.5 The critical cross-protection thresholds for DSS behaviour, for a range of R_0 values applied to all strains. Increasing the transmission potential of the dominant strains (assuming R_0 values are equal) increases the cross-protection threshold, γ*, lowering the parameter space where DSS occurs.

Multi-locus models based on the Gupta model have explored oscillatory dynamics using within-host models of *Plasmodium falciparum* (Recker et al. 2004) and human immunodeficiency virus (HIV) (Wikramaratna et al. 2015), and the population dynamics of influenza A (Recker et al. 2007; Wikramaratna et al. 2013). In the context of influenza, the multi-locus model provides an alternative

paradigm to antigenic drift, instead suggesting that serial replacement of influenza strains may occur due to the presence of antigens of low variability, without the need for mutation or recombination, a hypothesis coined as “antigenic thrift” (Wikramaratna et al. 2013).

1.5 Evidence of multi-locus model behaviour

Within the population structure of NSS, as all strains circulate independently due to minimal inter-strain competition, the behaviour is akin to various single-strain pathogens (i.e. each strain circulates within an independent SIR model). As NSS is not strictly displaying a multi-locus population, the evidence of DSS and CSS population dynamics will be the focus within this section, and throughout the rest of this thesis.

1.5.1 Evidence of discrete strain structure

Epidemiological studies support the results of multi-strain models, where strains self-organise into non-overlapping antigenic types (Gupta et al. 1998). Bacteria such as *Streptococcus pneumoniae* and *Neisseria meningitidis* have highly diverse capsular antigens, but some *S. pneumoniae* serotypes (Babl et al. 2001), and *N. meningitidis* surface antigens (PorA and FetA combinations) (Buckee et al. 2010; Urwin et al. 2004; Callaghan et al. 2008) are more frequent than expected – with some combinations being maintained, and dominating, for over ten years (Buckee et al. 2010; Watkins & Maiden 2012). Within another surface protein of *N. meningitidis*, Opa, the repertoires of the two hypervariable regions are non-overlapping (Callaghan et al. 2008). Particular variants of the major antigens factor H-binding protein (fHbp), Neisserial Heparin-Binding Antigen (NHBA), and Neisseria adhesion A (NadA) can also persist for decades, with

non-overlapping combinations associating with particular clonal complexes (Bambini et al. 2013). This discrete structuring of antigenic loci is not observed in housekeeping genes, suggesting immune selection plays a role (Buckee et al. 2010; Callaghan et al. 2008). Minimal overlap between antigenic repertoires has also been observed in the tick-borne bacteria *Borrelia hermsii* (Hinnebusch et al. 1998).

In protozoan populations, malarial variant surface antigen *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) repertoires do not overlap in local areas of Africa and Thailand (Fowler et al. 2002), Venezuela (Tami et al. 2003), Papua New Guinea (Barry et al. 2007), Kenya and Senegal (Chen et al. 2011). This has also been reported in rifin, another *P. falciparum* surface antigen (Tami et al. 2003). However, repertoire overlap was observed in Sudan, where some isolates were antigenically identical (Giha et al. 1999). Another antigen of *P. falciparum*, merozoite surface protein 3 (MSP-3), despite some overlap still has a very limited diversity (Polley et al. 2007). *Plasmodium vivax*, even with a higher diversity than *P. falciparum*, displays strong linkage disequilibrium at antigenic loci (Batista et al. 2014).

1.5.2 Evidence of cyclical strain structure

Cyclical dynamics have been observed for multi-strain pathogen populations, particularly for RNA viruses. Influenza A has an extremely limited antigenic diversity at points in time, with the observation of a sequential replacement of dominant strains (Fitch et al. 1997; Fitch et al. 1991; Bush et al. 1999). The antigenic loci under positive selection, due to host immunity, are the haemagglutinin (HA) and neuraminidase (NA) proteins (Wolf et al. 2006; Fitch et

al. 1991; Bush et al. 1999; Ina & Gojobori 1994), required for host cell binding and entry.

Successive lineage replacement has also been observed in respiratory syncytial virus (RSV) (Cane & Pringle 1995; Zlateva et al. 2005), with the G protein surface antigen under immune selection (Botosso et al. 2009). Unlike influenza A, RSV displays a coexistence of previous strains, even 30 years after dominance (Martinelli et al. 2014).

Rotavirus displays spatial and temporal cyclical patterns, with strains varying in frequency in the same location through time, and across different geographic regions (Santos & Hoshino 2005; Dóró et al. 2014a; Bányai et al. 2012). However, there is far more co-circulation and diversity than observed in influenza A. Within developing countries, there are five main circulating strains at any one time, with strain G1P[8] of strongest dominance (Santos & Hoshino 2005; Dóró et al. 2014b; Iturriza-Gómara et al. 2011), being replaced by strain G2P[4] every seven years (Dennis et al. 2014), and its genome has been observed to have no genetic change for 1-3 years (McDonald et al. 2012). Within developing countries, especially in Sub-Saharan Africa, there is much more diversity of strains, with unusual antigenic combinations being more common (Santos & Hoshino 2005). It is believed that novel strains emerge from a pool of globally circulating strains, such as via migration (Page & Steele 2004; Palombo et al. 1993; Doan et al. 2011), rather than punctual change from the current dominant strain as observed for influenza A (Smith et al. 2004).

Rotavirus dynamics are influenced by immune selection, evident by positive immune selection in the outer capsid genes viral protein 4 (VP4) (encoding the

G protein) and VP7 (encoding the P protein) (Gómara et al. 2001a; Page & Steele 2004; Kirkwood et al. 2004; Zeller et al. 2015), both of which serotypically define a strain. Additionally, positive selection is found in the non-surface genes VP8 and non-structural protein 4 (NSP4) (Kirkwood et al. 2004), and VP2 and VP3 (Zeller et al. 2015)).

A degree of lineage replacement has also been observed in foot and mouth disease virus (FMDV) in cattle (Subramaniam et al. 2013), with immune selection in sites A and D of the FMDV serotype C genome (Mateu et al. 1994). Lineage replacement is observed in dengue virus (Bennett 2003; Zhang et al. 2005), however this is believed to be due to enhancement of transmission of antigenically similar strains (Recker et al. 2009; Wikramaratna et al. 2010; Ferguson et al. 1999; Cummings et al. 2005), known as antibody-dependent enhancement (ADE), or via stochastic effects over an NSS population structure (Lourenço & Recker 2013), rather than cross-protection acting to reduce transmission or infectiousness. Some epidemiological models have suggested, however, that it is possible for cross-protection to create dengue-like oscillations without the inclusion of ADE (Adams et al. 2006).

1.6 Exploration of the boundaries of multi-locus model behaviour

The steady states of the multi-strain system have previously been found numerically (Gupta et al. 1996; Gupta et al. 1998; Recker & Gupta 2005; Recker et al. 2007; Wikramaratna et al. 2013), and analytically for a three (Lin et al. 1999) or four strain system (Dawes & Gog 2002). This has recently been expanded to

a method of finding these boundaries analytically for an unlimited number of strains (Blyuss 2013; Blyuss 2014; Cherif 2015).

These boundaries, and the dynamics of the behaviours themselves, have been shown to change with the inclusion of different host and pathogen characteristics into the original model by Gupta et al (summarised in Table 1.2). Overall, although the regions of the stable states can shift, they are still observed in multiple differing scenarios with different host and pathogen characteristics.

Aside from the multi-locus model of Gupta et al, Artzy-Randrup et al developed a multi-locus stochastic IBM, based on the antigenic *var* genes of *P. falciparum,* with a possibility of recombination (via mutation and migration events) between pathogen genomes (Artzy-Randrup et al. 2012). Despite the practically unlimited number of possible *var* repertoires, diversity was low and the pathogen population displayed DSS (albeit with partial overlapping of repertoires). The degree of overlap increased as transmission intensity (the number of transmission events per year) increased. The addition of recombination introduced a punctuated turnover of *var* repertoires through time, not observed in the deterministic history- or status-based models, with an increase in recombination increasing the frequency of the turnover events – though DSS was stable between these events.

1.7 THESIS SUMMARY

The work presented in this thesis expands on the current knowledge of how population dynamics and the boundaries of the stable states change, with a focus on the robustness and stability of the multi-locus system. It is so far understood

Host/Pathogen Characteristics	Effect	Reference(s)
Recombination of antigenic loci.	DSS still appears even when recombination between strains is possible.	(Gupta et al. 1996)
Different R_0 of dominant strains.	Increasing R_0 increases the degree of cross-protection required for the multi-locus model to transition from CSS to DSS (Figure 1.5).	(Gupta et al. 1998)
Cross-protection is dependent on the number of shared antigenic variants.	Changes appearance of DSS when there is a low degree of cross-protection between strains which do not share many antigens – the strains in DSS have a minimal degree of overlap (but not complete discordance at every loci).	(Recker & Gupta 2005)
An additional degree of cross-protection (δ) to strains, where an individual has been exposed to all antigenic variants after a combined exposure of similar strains.	Shifts the thresholds of DSS and NSS (lower cross-protection required as δ increases).	(Recker et al. 2007)
Host heterogeneity in the breadth of immune response, by 1) including host genotypes which respond to specific or conserved determinants, 2) hosts differ by a "polyclonality" factor.	A small host population of specific responders is enough to allow coexistence of strains (1). Increased breadth in immune response increases the range of cross-protection where CSS occurs (2). Both methods lower the cross-protection threshold for the appearance of DSS.	1.(Gupta & Galvani 1999) 2.(Cobey & Pascual 2011)
Immunodominance of antigens, so the probability of acquiring immunity to certain antigens is skewed.	No CSS or DSS if host responds to only one antigen (complete immunodominance).	(Cobey & Pascual 2011)
Heterogeneity in host birth rates.	Changes to the oscillatory dynamics seen in CSS. Shifts the thresholds of CSS (a lower cross-protection is required for a longer lifespan).	(Wikramaratna et al. 2014)

Table 1.2 A summary of the effects of different host/pathogen characteristics included in the multi-locus model of Gupta et al (Gupta et al. 1996)**.**

that DSS is highly robust, as it persists even with recombination (Gupta et al. 1996; Artzy-Randrup et al. 2012; Buckee et al. 2004). The stability of the multi-locus model developed by Gupta et al is here examined in two ways: Chapter 2) the stability of DSS in the threat of an emerging non-discordant strain of a higher fitness; Chapter 3) the shift in the stable states as immunity may be acquired to an additional, conserved determinant. A third investigation (Chapter 4) uses a machine learning algorithm for pattern-oriented modelling of rotavirus using the multi-locus framework with an application to rotavirus epidemiology.

1.7.1 Effects of phenotypic heterogeneity within the pathogen population on the stability of multi-locus systems

It is understood that within a two strain model, the strain of the highest R_0 "wins" (Anderson & May 1992), in a classic example of Darwinian fitness. Various single strain models, within both ecology (Gonzalez et al. 2013) and epidemiology (May et al. 2001; Antia et al. 2003; Alexander et al. 2010; Lloyd-Smith et al. 2013), have explored the characteristics required to allow a pathogen to invade. However, pathogen emergence has not been extended to a multi-strain system, particularly one within stable DSS equilibrium. Knowing that phenotypic variation in the fitness and transmission of strains exists, Chapter 2 explores under which conditions subdominant strains can be released from strong competitive suppression and invade a stable pathogen population.

1.7.2 Effects of heterogeneity in host immune response on the stability of multi-locus systems

The only instance where a multi-locus model did not observe discrete structuring of the pathogen population is where hosts only recognised one antigen (Cobey

& Pascual 2011), essentially acting as a conserved determinant and creating a single strain system – not a system of inter-strain competition. Chapter 3 explores a multi-locus model in which different host genotypes result in a response to either conserved or strain-specific determinants. Heterogeneity in immunity between hosts is a known phenomenon, as individuals do not have the same response when exposed to the same pathogen or strain (or vaccine), though often host homogeneity is assumed for the sake of simplicity in epidemiological models.

Host heterogeneity in immunity can be caused by a genetic predetermination. The blood disorder of sickle cell anaemia, specifically individuals who are heterozygous (known to have the “sickle cell trait”), provides malarial protection (Allison 1954), and though the mechanism is still unclear these individuals’ antibody response is broader and of a higher titre (Cabrera et al. 2005). Human leukocyte antigens (HLAs), also known as major histocompatibility complexes (MHCs), are involved in the adaptive immune response, and specific HLA alleles have been demonstrated to reduce susceptibility to certain diseases, such as malaria (Hill et al. 1991) and HIV (Carrington & O’Brien 2003).

Unlike using an IBM, within a status-based model differences in immune responses is challenging to include due to its intractability. Previously, Gog et al used polarised immunity, where individuals of the same immune status could respond to the same strain differently (Gog & Swinton 2002; Gog & Grenfell 2002). Heterogeneities have also been included as “hard-wired” host genotypes, where hosts respond to either conserved or strain-specific antigens (Gupta & Galvani 1999), or responses are dependent on the recognition alleles (human leukocyte antigens, HLAs) of the host (Penman et al. 2013).

The novel model presented in Chapter 3 is an extension of the work by Gupta and Galvani (Gupta & Galvani 1999), where the development of a new technique to include host heterogeneity allows both life-long strain-specific immunity and immunity towards conserved antigenic determinants to be acquired in a flexible way. The proportion of the hosts with a broad response is explored within a range of cross-protection, to test the effects of host heterogeneity in the breadth of immunity on the appearance of multi-strain pathogen population dynamics.

1.7.3 Qualitative model fitting through machine learning

Epidemiological models are often used alongside disease data to reveal the parameters (or characteristics) which may lead to observed biological patterns, which can then act to inform policies and control. Models are fitted to data sets, such as from longitudinal studies or public health records, but this has limitations: 1) under-reporting of cases must be controlled for; 2) longer term patterns such as seasonality may be missed; 3) shorter term patterns may be missed in temporally aggregated data; 4) cases may be spatiotemporally heterogeneous; and 5) the data may be prevalence data rather than specifically the incidence of strains. Models also risk being over-fitted to data when the latter is limited, either temporally, spatially, or in its details: in this way, the models may fit data-specific characteristics which are not representative of the ubiquitous properties or behaviour of the pathogen.

For pathogens for which insufficient data exists, fitting exercises can be misleading. Introduced in Chapter 4 is a technique that informs on elucidating the parameters and relationships between strains of a disease using reverse pattern-oriented modelling. General qualitative patterns observed at the population level

– such as coexistence and cyclic behaviour – rather than a specific data set, are used to infer the individual-level characteristics of the pathogen.

Rotavirus is used as a case study, specifically using its population dynamics in developed countries, for which complete and robust data (in particular, the prevalences of strain types) is missing in the literature. Based on biological insight, the two outer capsid proteins, G and P, are used to define a strain, and although there are over 70 possible combinations in humans (Matthijnssens & Van Ranst 2012), there are five main strains observed: G1P[8], G2P[4], G3P[8], G4P[8], and the recently emerged G9P[8] (Iturriza-Gómara et al. 2000; Santos & Hoshino 2005; Iturriza-Gómara et al. 2011). Of these common strains, G1P[8] is strongly dominant (Santos & Hoshino 2005; Dóró et al. 2014b; Bányai et al. 2012; Iturriza-Gómara et al. 2000), with occasional replacement by G2P[4] (Gómara et al. 2001a; Dennis et al. 2014).

Using time series generated using the Gupta model with a variety of different parameters (pathogen characteristics), simulations are classified as displaying certain rotavirus-like population dynamics. The machine learning technique known as a Random Forest algorithm is implemented to uncover which pathogen characteristics are important in the development of rotavirus-like dynamics. This technique can be applied to other disease systems, and is useful for multi-strain pathogens where there are limited epidemiological data available.

2 DISRUPTION OF DISCRETE STRAIN STRUCTURE THROUGH PHENOTYPIC CHANGE

2.1 INTRODUCTION

There are various mechanisms by which a pathogen can become introduced to a new host population, as described in Table 2.1. These events can be frequent, however, in many scenarios, novel or introduced strains are often maladapted to the new host or environment, and so may not persist. Here I examine how introducing a strain of a different transmission phenotype, via a recombination event, into a multi-strain system of a highly stable non-overlapping population structure can disrupt the stable state.

Method of Introduction	Example	References
Mutation	Seasonal human influenza	(Smith et al. 2004)
Recombination	H1N1 influenza pandemic in 2009 emerged after a triple reassortment event of swine, avian, and H3N2 human influenza strains	(Smith et al. 2009)
Spillover	Ebola, Nipah virus, Severe Acute Respiratory Syndrome (SARS) (bats)	(Leroy et al. 2005; Chua et al. 2000; Li et al. 2005)
Environmental change	Bluetongue virus spread linked to climate change	(Purse et al. 2005)
Drug resistance	Multi-drug resistant *Mycobacterium tuberculosis* (MDR-TB)	(Ford et al. 2013; Hiramatsu et al. 2001)
Post-vaccine	*Streptococcus pneumoniae* non-vaccine serotypes replaced the seven previously dominant vaccine serotypes, after the introduction of the PCV7 vaccine	(Weinberger et al. 2011; Watkins et al. 2015)

Table 2.1 Examples of emerging strains, with different methods of introduction.

The classic perspective of pathogen success arises from the transmission potential, R_0, which can be interpreted as the number of secondary infections from one primary case in a totally susceptible population (Anderson & May 1992). In a single strain system, a novel pathogen can spread and persist in a population as long as $R_0 > 1$ (Anderson & May 1992). The emergence of a novel pathogen into a wholly susceptible host population means that the number of infectious hosts initially is very small, and so stochastic effects are very important. Including demographic stochasticity provides a "probability of emergence", P_e, which is given as:

$$P_e = 1 - \left(\frac{1}{R_0}\right)^S$$

Where *S* is the number of infected seed hosts (May et al. 2001). In this instance, if a strain has an $R_0 > 1$ extinction is possible, but with enough seeded infections $(S \to \infty)$ emergence is guaranteed. A maladapted pathogen, of $R_0 < 1$, can invade but not persist, as stochastic fadeout causes stuttering transmission chains: these strains can only persist if the transmission chains are long enough to allow adaptation to the host, thereby increasing R_0 (Antia et al. 2003).

A wide variety of factors on P_e have been investigated, predominantly using SIR models with stochasticity included via an individual-based model (IBM) framework (Alonso et al. 2007) or the Gillespie algorithm (Read & Keeling 2007; Dalziel et al. 2014; Parham & Michael 2011), or by using multi-type branching of the early dynamics (Antia et al. 2003; Yates et al. 2006; Lloyd-Smith et al. 2005; Day et al. 2006; Alexander et al. 2010; Arinaminpathy et al. 2009). Properties which increase P_e are: a higher mutation rate (Antia et al. 2003; André & Day

2005) and particular forward non-reversible mutation (Alexander et al. 2010); fewer mutational steps to an adapted strain, so evolution is more punctuated than gradual (Antia et al. 2003; André & Day 2005; Alexander et al. 2010; Arinaminpathy et al. 2009); a higher R_0 of intermediate strains in the evolutionary pathway (Antia et al. 2003; Alexander et al. 2010); the presence of superspreaders (with a preference of transmission to non-superspreaders) (Yates et al. 2006; Alexander et al. 2010), though others have found the opposite is true as this creates non-transmitting individuals (Lloyd-Smith et al. 2005)); a high seed of infectious hosts (Parham & Michael 2011; Day et al. 2006; May et al. 2001); and the inclusion of seasonality in transmission (Parham & Michael 2011).

Investigating single-strain systems is useful, and certainly applicable to emerging case studies (such as models of emergence using severe acute respiratory syndrome (SARS) or Chikungunya (Alexander et al. 2010; Hartfield & Alizon 2015)). Expanding to two strains, where there is complete cross-protection between them, has previously been shown to result in out-competition and persistence by the strain of a higher R_0 (Anderson & May 1992), in a classic example of Gause's principle of competitive exclusion (Gause, 1934). However, many diseases circulate as multiple distinct strains, with a degree of cross-protection between them.

Multi-strain models investigating emergence have found that a degree of cross-protection between antigenically similar strains causes feedback (Hartfield & Alizon 2015; Lourenço & Recker 2010), where dominant pre-existing strains will deplete susceptibles and reduce the chance of an introduced strain emerging. Multi-strain models have previously demonstrated how a very high degree of

cross-protection leads to a highly stable and robust equilibrium of discrete strain structure (DSS) (Gupta et al. 1996; Gupta et al. 1998), where only strains with non-overlapping antigens are able to coexist.

The stability of DSS has been tested before (Gupta et al. 1996), with some differences to the approach used here. In this previous work, a strain with non-overlapping antigens to the one dominant strain was introduced (i.e. strain BY was added to a single strain population of endemic AX), and the ability of the recombinants of these strains (i.e. AY and BX) to invade were tested. Therefore, what was tested was the ability of a recombinant to invade a stable single strain system.

What has been lacking from research so far is to explore the introduction of a strain into a multi-locus system whilst it is in the highly stable DSS equilibrium of multiple, discordant strains. Given the stability and robustness of DSS, is it possible to break down this state by introducing an overlapping strain of a different transmission phenotype? If so, how high must the transmission advantage be for an introduced strain to not only invade, causing an epidemic, but also to persist?

Exploring the factors that increase or reduce the probability of a successful invasion of an introduced strain aids our understanding of which characteristics of a strain or pathogen are important in emergence in natural systems. This has implications for surveillance of emerging pathogens, particularly after spillover events (where a pathogen may be adapting to a new host and hence increasing fitness) or after the distribution of a new vaccine (which reduces the fitness of

vaccine-types, creating the possibility of competitive release, as observed in *Streptococcus pneumoniae* (Weinberger et al. 2011; Watkins et al. 2015)).

Using the **M**ulti-Locus **ANTI**genic **S**imulator (MANTIS) R-package (Lourenço et al. 2015) – which uses the model developed by (Gupta et al. 1996) – a function to change the transmission potential of a strain at a given time step was developed, with which I explore the resilience of a system in DSS to the introduction of a non-dominant strain.

The equilibria for the proportion of immune hosts (akin to herd immunity), and subsequently the transmission potential required by a strain to overcome this competitive suppression and proliferate (termed the invasion threshold), were derived analytically. Numerical simulations closely matched these analytical expectations, and the dynamics could be further explored with the Gupta model, focusing on the influence of: i) transmission; ii) duration of infection; iii) migration; iv) seasonality; and v) parameter variation. This research identifies properties of a pathogen which may increase the likelihood of invasion, during DSS, by a novel strain, and subsequently replace previously dominant strains.

2.2 METHODS

The MANTIS framework, based on the model of Gupta et al described in detail in Chapter 1, is used for all simulations in this work. Each strain has the same transmission ability, defined by R_0, where $R_0 = \beta/(\sigma + \mu)$. The strains dominating in DSS have the transmission potential R_0^{dom}. At a time point of 650 years (the invasion time step), while in DSS equilibria, one of the non-discordant strains (i.e. the invading strain) that has been driven to extinction due to cross-protection is then introduced back into the system with a new, higher

transmission potential (R_0^{inv}). Unless stated otherwise, the proportion infected for the invading strain (Y_{inv}) is seeded at the same order of magnitude to the infected proportions of the dominant strains (10^{-4}).

Alternatively, for a scenario defined with constant background seeding of the invading strain, a seeding factor (α) was added to the infectious host population of the introduced strain at every time point past the invasion time step, by altering the original Gupta et al equation for Y_i (infected with strain *i*):

$$\frac{dY_i}{dt} = \alpha + \lambda_i[(1 - W_i) + (1 - \gamma)(W_i - Z_i)] - (\sigma + \mu)Y_i \quad (2.1)$$

Invasion is defined as the outcome where the invading strain persists and replaces at least one of the previously dominant strains. The R_0^{inv} value where invasion is successful is defined as the "invasion threshold".

Seasonality is already included in the MANTIS framework, by generating the transmission coefficient β with a sine wave, using:

$$\beta = \beta(1 + \varepsilon \sin(\pi t)^k) \quad (2.2)$$

Where *t* is time, and ε is the degree of seasonality, determining the amplitude. The value of *k* determines the curvature of the sine wave (here *k* = 6 to provide a high transmission period of several months).

Demographic stochasticity cannot be used due to the overlapping and intractable nature of this multi-locus model, and the lack of a constant population size. Instead, stochasticity within the parameters, rather than equations, was included by adding parameter variation. This was achieved by sampling μ from a probability distribution Gaussian(μ, μ/4) and β from a probability distribution Gamma(4β, ¼). The distributions of these parameters are shown in Figure 2.1. The stochasticity in lifespan (1/μ) is applied to all strains equally, as this is a host factor and so affects strains uniformly. This is achieved by sampling the distribution once per time step (i.e. annually), and applying this value to all strains. Stochasticity in the transmission coefficient β, however, is strain-dependent and so applied to each strain separately by sampling from the distribution *x* number of times (where *x* is the number of strains) for each time step (i.e. annually), and each β value is applied to individual strains.

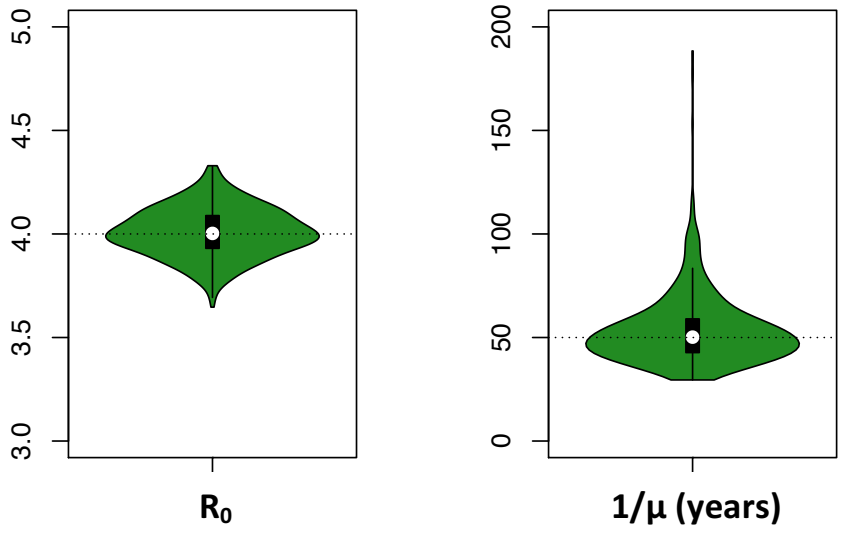


Figure 2.1 Variation in lifespan (1/μ) and R_0 when including stochasticity. The distributions follow a μ~Gaussian(μ, μ/4) and β~Gamma(4β, 1/4). The means of these parameters (black horizontal lines) are the values that would be used in the deterministic model: μ = 0.02 (lifespan of 50 years) and R_0 = 4.

When the model includes parameter variation, simulations are run for 500 trials per parameter set, and each trial is given a binary value if invasion occurred or not, to provide the proportion of successful invasion events, defined as:

$$P_{se} = \frac{No.trials\ where\ invasion\ occured}{No.trials}$$

2.3 RESULTS

2.3.1 ANALYTIC DERIVATION OF INVASION THRESHOLDS

For a system in DSS, any strain which shares alleles with the dominant strains is suppressed, since immune selection, via cross-protection, acts to reduce the availability of susceptible hosts. This availability is evident in W_i^* (derived below), the proportion of hosts at equilibrium which are immune to any strain sharing alleles with strain *i*. W_i^* is therefore a measure of herd immunity in the host population, and subsequently acts to limit the availability of hosts that can be infected by the excluded strain, influencing the possibility of emergence.

To determine the value of W_i^* at equilibrium, it is necessary to derive the values of Z_i^*, the proportion of hosts immune to strain *i*, and Y_i^*, the proportion of hosts infectious with strain *i,* also at equilibrium. Firstly, Z_i^* can be derived from Equation 1.3:

$$\frac{dY_i}{dt} = \beta Y_i^*[(1 - W_i^*) + (1 - \gamma)(W_i^* - Z_i^*)] - (\mu + \sigma)Y_i^* = 0$$

$$\beta Y_i^*[(1 - W_i^*) + (1 - \gamma)(W_i^* - Z_i^*)] = (\mu + \sigma)Y_i^*$$

$$(1 - W_i^*) + (1 - \gamma)(W_i^* - Z_i^*) = \frac{1}{R_0}$$

During DSS, circulating strains do not share alleles and so the effect of cross-protection is null, and so this equation can be reduced to:

$$Z_i^* = 1 - \frac{1}{R_0}$$

Substituting Z_i^* into Equation 1.1 the equilibrium values for Y_i can be derived:

$$\frac{dZ_i}{dt} = \beta Y_i^*(1 - Z_i^*) - \mu Z_i^* = 0$$

$$\beta Y_i^*(1 - Z_i^*) = \mu Z_i^*$$

$$\frac{\beta Y_i^*}{R_0} = \mu\left(\frac{R_0 - 1}{R_0}\right)$$

$$Y_i^* = \frac{\mu}{\beta}(R_0 - 1)$$

Finally, the equilibrium values for W_i can be derived by substituting Y_i^* into Equation 1.2. When the model is in DSS, the only strains to circulate are those which do not share alleles. This leads to the assumption that the sum of strains not sharing alleles with strain *i* is the number of dominating strains (*n*).

$$\frac{dW_i}{dt} = \sum_{j\sim i} \beta Y_j^* (1 - W_i^*) - \mu W_i^* = 0$$

$$n\beta Y_i^*(1 - W_i^*) = \mu W_i^*$$

$$n\mu(R_0 - 1)(1 - W_i^*) = \mu W_i^*$$

$$\frac{W_i^*}{(1 - W_i^*)} = n(R_0 - 1)$$

$$\frac{1}{W_i^*} = \frac{1}{n(R_0 - 1)} + 1$$

$$W_i^* = \frac{n(R_0 - 1)}{n(R_0 - 1) + 1} \qquad (2.3)$$

From Equation 2.3, the more dominant strains which share alleles with strain *i* (i.e the higher the value of *n*), the larger the proportion of immune hosts.

To derive the invasion threshold (i.e. the R_0 of the invading strain, R_0^{inv}, required to replace *n* dominant strains), the equation for infectious individuals of the invading strain is used.

Assuming an invading strain starts with no immune hosts ($Z_{inv}^* = 0$), Equation 1.3 is reduced to:

$$\frac{dY_{inv}}{dt} = \beta Y_{inv}^*[1 - W_{inv}^* + (1 - \gamma)W_{inv}^*] - (\sigma + \mu)Y_{inv}^* = 0$$

$$Y_{inv}^* = \frac{\beta Y_{inv}^*[1 - W_{inv}^* + (1 - \gamma)W_{inv}^*]}{(\sigma + \mu)}$$

$$Y_{inv}^* = R_0^{inv} Y_{inv}^*[1 - W_{inv}^* + (1 - \gamma)W_{inv}^*]$$

$$\frac{1}{R_0^{inv}} = 1 - \gamma W_{inv}^*$$

$$R_0^{inv} = \frac{1}{1 - \gamma W_{inv}^*} \quad (2.4)$$

For invasion to occur, the R_0 of the invading strain needs to be more than R_0^{inv}.

It follows that the possibility for a strain to invade a population is dependent on the proportion of immune hosts: the bigger the W_i^*, the fewer susceptibles there are available, and the higher the R_0 must be for a strain to invade. From Equation 2.4, the degree of cross-protection is also a determinant of invasion, as this reduces the transmission of strains with shared alleles to those dominating. As displayed in Figure 2.2a, which displays the solutions for Equation 2.4, an increase in cross-protection increases the invasion threshold: an invading strain requires a larger transmission advantage to overcome competitive suppression.

Under the parameter combination used throughout this work, with a recovery time of 5 days and lifespan of 50 years (where the transmission coefficient is altered to determine the R_0^{dom}), the threshold ratios for both cross-protection and *n (*the number of dominant strains) reach a peak at R_0^{dom} 2-5 (Figure 2.2). Above this range, the system starts to destabilise, making it less resilient to invading strains and therefore invasion does not require as high of a transmission advantage.

However, this peak at R_0^{dom} 2-5 is not seen where γ = 1 (Figure 2.2a), as complete cross-protection keeps the system stable in DSS at all values of R_0. This can be explained using Equation 2.3 and 2.4:

Where $R_0^{dom} \rightarrow \infty$ or $n \rightarrow \infty$ we find:

$$W_i^* \rightarrow 1$$

Where this is true, we find:

$$R_0^{inv} = \begin{cases} \frac{1}{1-\gamma} & \text{if } \gamma < 1 \\ \infty & \text{if } \gamma = 1 \end{cases}$$

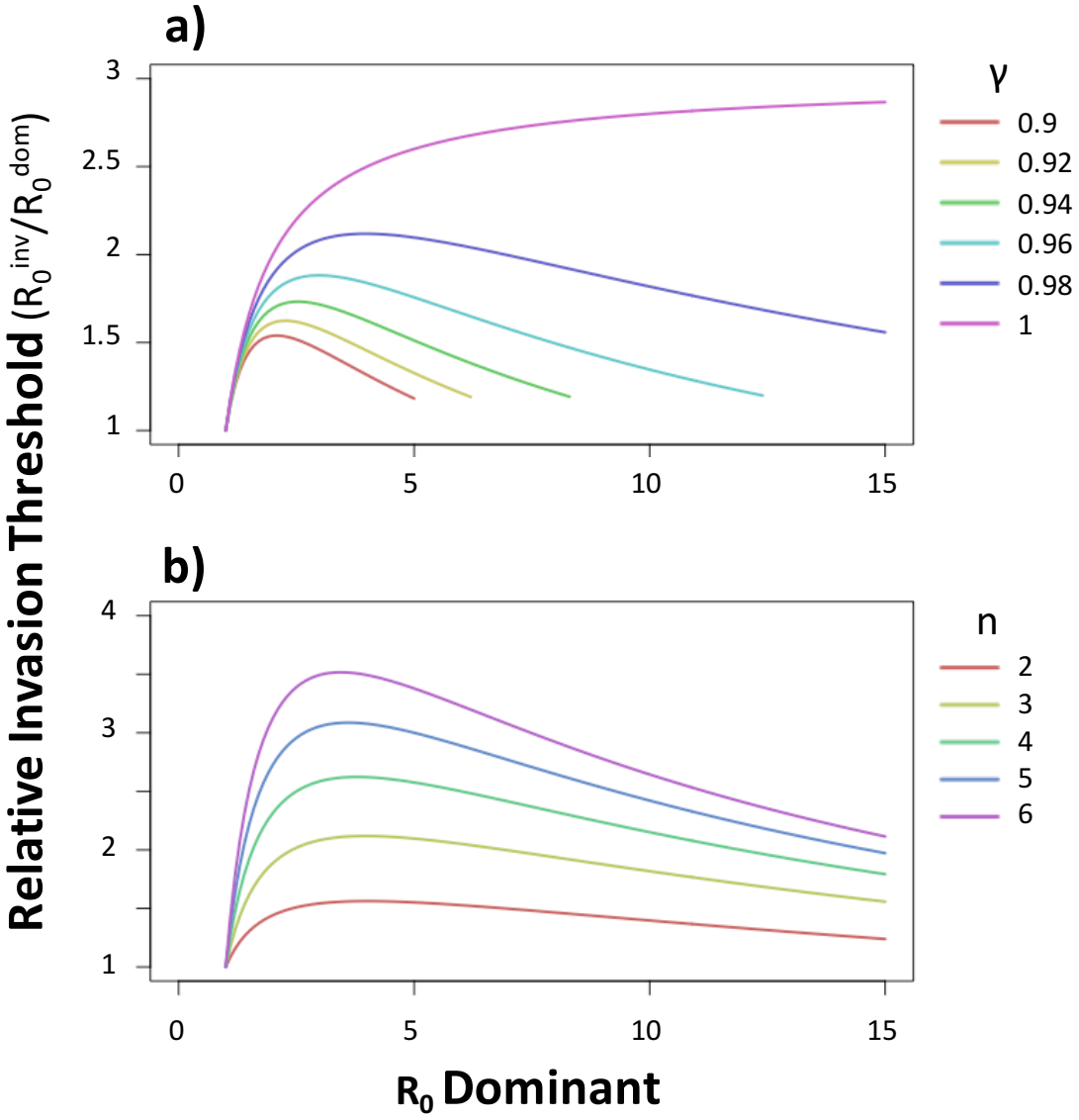


Figure 2.2 Analytical invasion threshold, relative to R_0^{dom}, for a) γ values where n = 3 and b) n values where γ = 0.98. The R_0^{inv} required for a strain to invade is higher as γ increases, as the high inter-strain competition suppresses strains of shared alleles, making the system more resistant to invasion. The peak difference between R_0^{dom} and R_0^{inv} is at R_0^{dom} 2-5: as R_0^{dom} increases the system starts to destabilise and so a smaller difference is required to have an advantage, and the relative threshold tends to 1. Where $\gamma = 1$ this pattern is not found and instead the threshold keeps increasing, as complete cross-protection keeps the system stable and resistant to invasion.

Where the antigenic structures – that is, the number of loci, and the number of alleles at each locus – are asymmetric (e.g. [3 2]), the DSS state differs to those of symmetric structures (e.g. [2 2]) by the presence of subdominant strain(s). These subdominant strain(s) do not share alleles with the dominant strain (and hence are able to circulate), however they do share alleles with the other subdominant strains (and hence are of a lower Y_i at equilibrium than the dominant strain(s), as they interfere via cross-protection).

It was assumed that because dominant strains do not share alleles, cross-protection does not have an effect. This meant that Equation 1.3, simplified as

$$(1 - W_i^*) + (1 - \gamma)(W_i^* - Z_i^*) = \frac{1}{R_0}$$

Could be reduced to

$$Z_i^* = 1 - \frac{1}{R_0}$$

However, this equation for Z_i at equilibria is not true for asymmetric antigenic structures, where some dominant strains do share alleles.

Additionally, the analytical derivations assumed that if no alleles were shared between the dominant strains (and Y_{dom} at equilibria were equal), then Equation 1.2 could be simplified as

$$n\beta Y_i^*(1 - W_i^*) = \mu W_i^*$$

Where n is the number of dominant strains. Again, this assumption is not held for an asymmetric antigenic structure. The expected thresholds from the analytical solution of Equation 2.4 is only applicable to symmetric antigenic structures, and

these structures are therefore the only ones used in the numerical simulations throughout the work presented here.

2.3.2 Numerical simulations investigating the invasion threshold

Using the MANTIS R-package, the numerical solutions for the invasion thresholds can be elucidated by increasing the transmission potential of an excluded strain in DSS, and finding the R_0 required to allow this excluded strain to become dominant. Altering R_0 can be achieved by changing the background death rate μ, transmission coefficient β, or recovery rate σ. I first consider the effects of changing β, thereby altering the force of infection of the invading strain.

Changing the transmission coefficient alone caused the numerical invasion simulation thresholds to be very high, much greater than the expected thresholds from the analytical solution of Equation 2.4. The example in Figure 2.3 is where

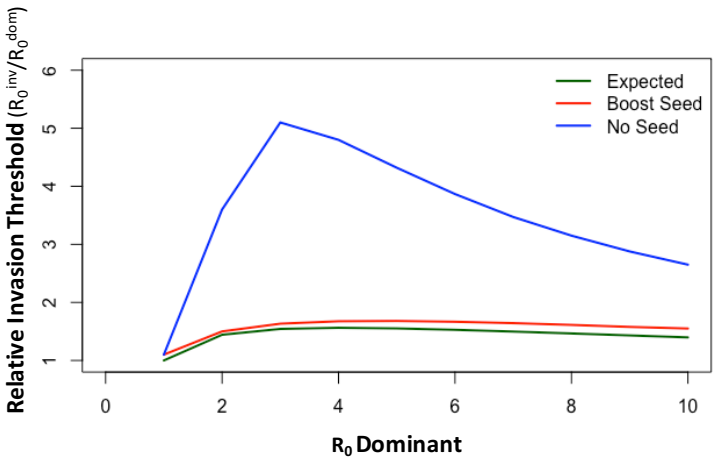


Figure 2.3 The numerical simulation results for the invasion threshold, comparing the results from the expected thresholds from the analytical solution (green), and the observed thresholds either with (red) or without (blue) a boost seed of 10^{-4} for the invading strain at the point of transmission coefficient change. Without this boost seed, the numerical simulation is a lot more robust to invasion, with the thresholds being far higher. This uses a [2 2] system with n = 2, and where γ = 0.98, 1/μ = 50 years, and 1/σ = 5 days. The population is initially susceptible with the proportion of infectious for each strain sampled as a uniformly distributed random number of magnitude 10^{-4}.

$\gamma = 0.98$, but where $\gamma = 1$ the threshold is unachievably high (for $R_0^{dom} = 2$ the invasion threshold is >100). In changing only the transmission coefficient, the introduced strain is starting from an infected population of (virtually) zero, so an epidemic is either impossible or requires a much higher R_0 advantage than expected.

To overcome this large difference in results from the analytical solution, a "boost" seed of 10^{-4} is introduced to the invading strain at the point of transmission coefficient change (invasion time step). This is a reasonable initial condition for an introduced strain, being the same order of magnitude as the proportion of infected individuals for the dominant strains, and as can be seen in Figure 2.3 this seeding provides invasion thresholds more like the analytical results. The boost seed of 10^{-4} is used for all future numerical simulations unless stated otherwise.

Using numerical simulations with a variety of antigenic structures, there is good agreement between the observed numerical simulations and expected thresholds from the analytic solution at low R_0, but certain discrepancies were found as R_0^{dom} increased (Figure 2.4). Among the antigenic structures allowing for the replacement of two dominant strains, the discrepancy between expected and observed thresholds are the same for different antigenic structures. Where three dominant strains are being replaced, there is more of a discrepancy.

The differences between expected and observed thresholds shows that R_0 is not the only parameter involved in the invasion threshold, as Equation 2.4 would suggest. Additional factors may also be involved in determining whether a strain can invade the stable system.

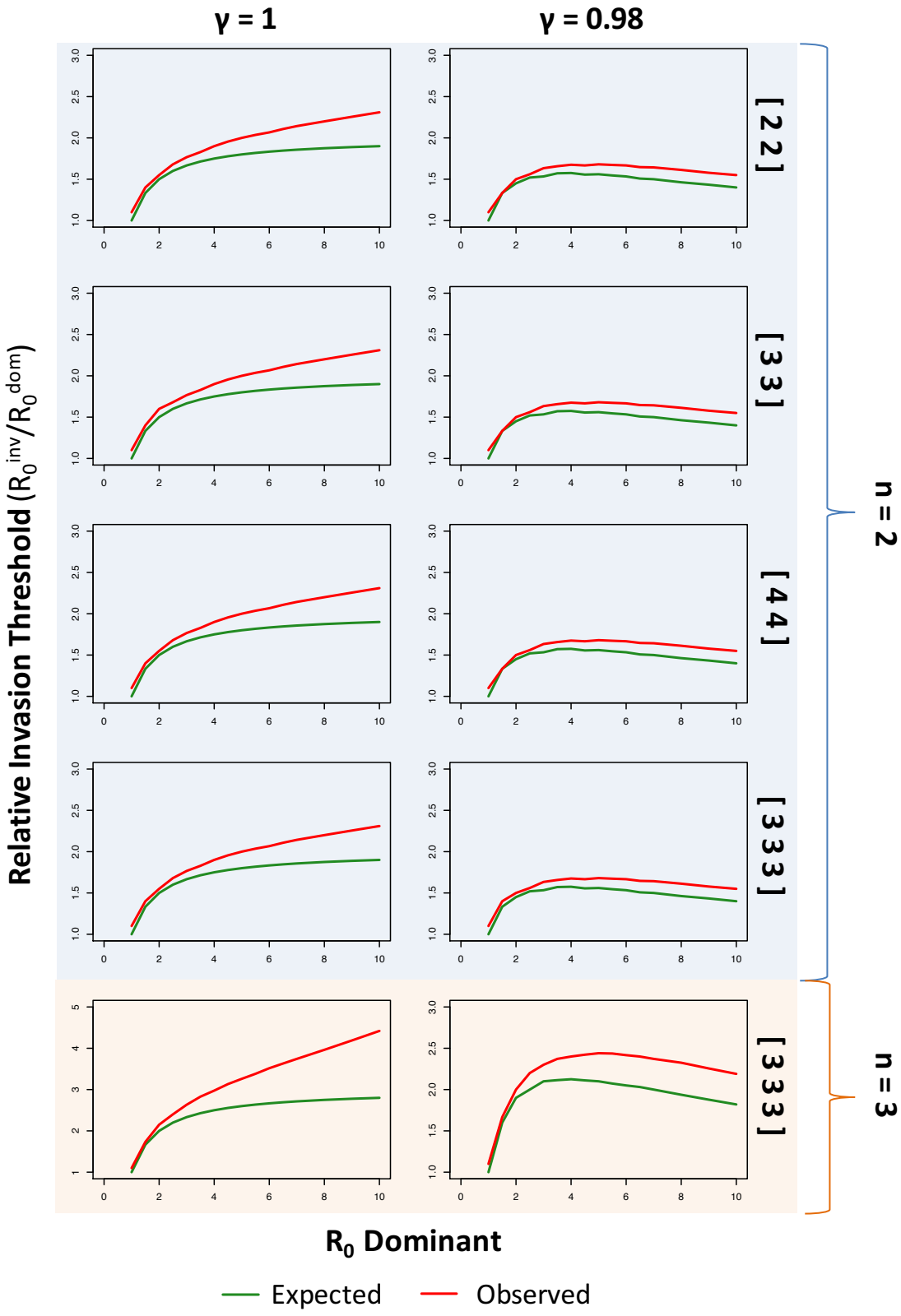


Figure 2.4 The invasion thresholds for different γ and n values, for the expected thresholds from the algebraic approximations, and observed thresholds by MANTIS. The expected and observed thresholds are similar, though differ as R_0^{dom} increases. The margin between observed and expected is wider where n=3. For the simulations, $1/\mu$ = 50 years and $1/\sigma$ = 5 days, and the proportion infectious for each strain is sampled as a uniformly distributed random number of magnitude 10^{-4}.

2.3.3 THE EFFECTS OF THE SIZE OF THE INITIAL SEED

The transmission coefficient is used as a constant in the force of infection, defined as $\lambda_i = \beta Y_i$. The proportion infected with strain *i,* Y_i, is also a determinant of λ_i. As seen in section 2.3.2, a seed of individuals infected with the invading strain (Y_{inv}), added at the invasion time step, changes the invasion threshold of the numerical simulations, and this occurs because the seed influences the force of infection for the invading strain (λ_{inv}). In this section, the effect of the size of this "boost" seed is explored.

An increase in the seed changes the size of the initial epidemic and shifts the invasion threshold to a lower R_0^{inv} (Figure 2.5). There is a threshold of $R_0^{inv} = 7$ where, at the point of introduction, Y_{inv} is virtually 0. Another threshold is observed at approximately $R_0^{inv} = 3$, matching the expected invasion threshold from the analytical solution, which occurs for a Y_{inv} (10^{-4}), which was previously used for Figure 2.4 as it is of the same order of magnitude as Y_{dom} at the point of introduction.

As the seed becomes unrealistically high (> 0.07) the invasion threshold drops below the R_0^{dom} value of 2. At this point Y_{inv} is so high it destabilises the population, allowing invasion. Overall, the invasion threshold falls as the size of the seed increases: the immediate force of infection at the invasion time step changes the chance of invasion, even if the R_0 conditions are met.

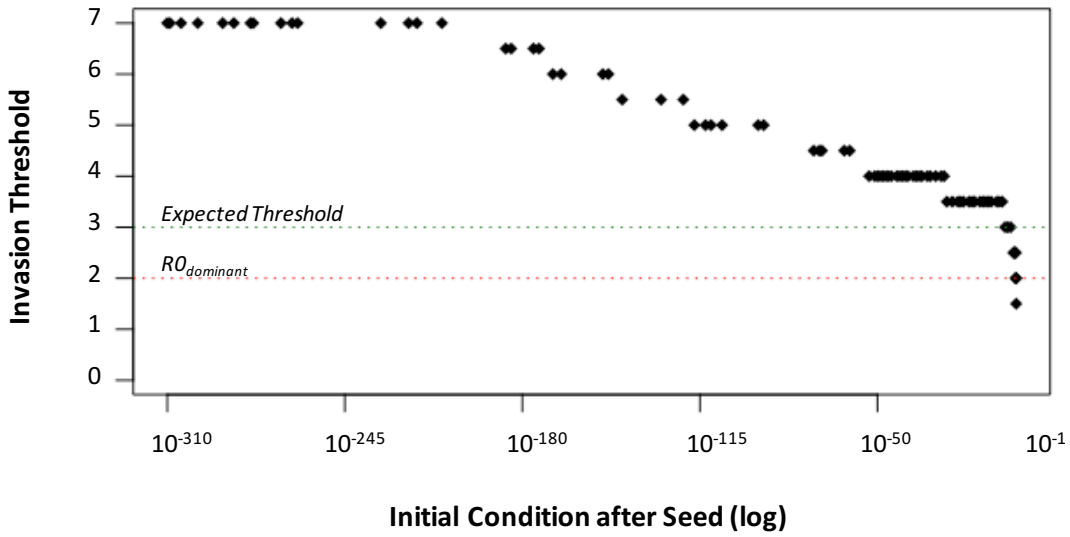


Figure 2.5 The effect of seeding the proportion of hosts infected with the invading strain (Y_{inv}) on the invasion threshold, for $R_0^{dom} = 2$, $\gamma = 0.98$, $1/\mu = 50$ years, $1/\sigma = 5$ days, with a [2 2] strain system (n = 2). Without virtually any seeding, the MANTIS framework finds an abstract threshold of 7. A second threshold is found at approximately $R_0^{inv} = 3$, as approximated from Equation 2.4, and this is found when the seed is at the same order of magnitude to the proportion infectious for the dominating strains (10^{-4}). As the seed becomes very high, the threshold drops below R_0^{dom}, but at this level of seeding the proportion infected is too high and the system destabilises.

2.3.4 The duration of infectiousness alters the size of the initial epidemic peak of invading strains

In addition to the force of infection, the duration of infectiousness (σ) contributes to the value of R_0. As the duration of infectiousness is lengthened, the invasion threshold increases above the expected threshold (Figure 2.6a). When investigating the time series of the invasion events at the same R_0 values for different durations (Figure 2.6b), the first peak in the proportion infected for the invading strain reduces in size as the duration lengthens, and the size of this initial epidemic is a determinant of invasion (here defined as the strain's ability to persist, replacing the previously dominant strains).

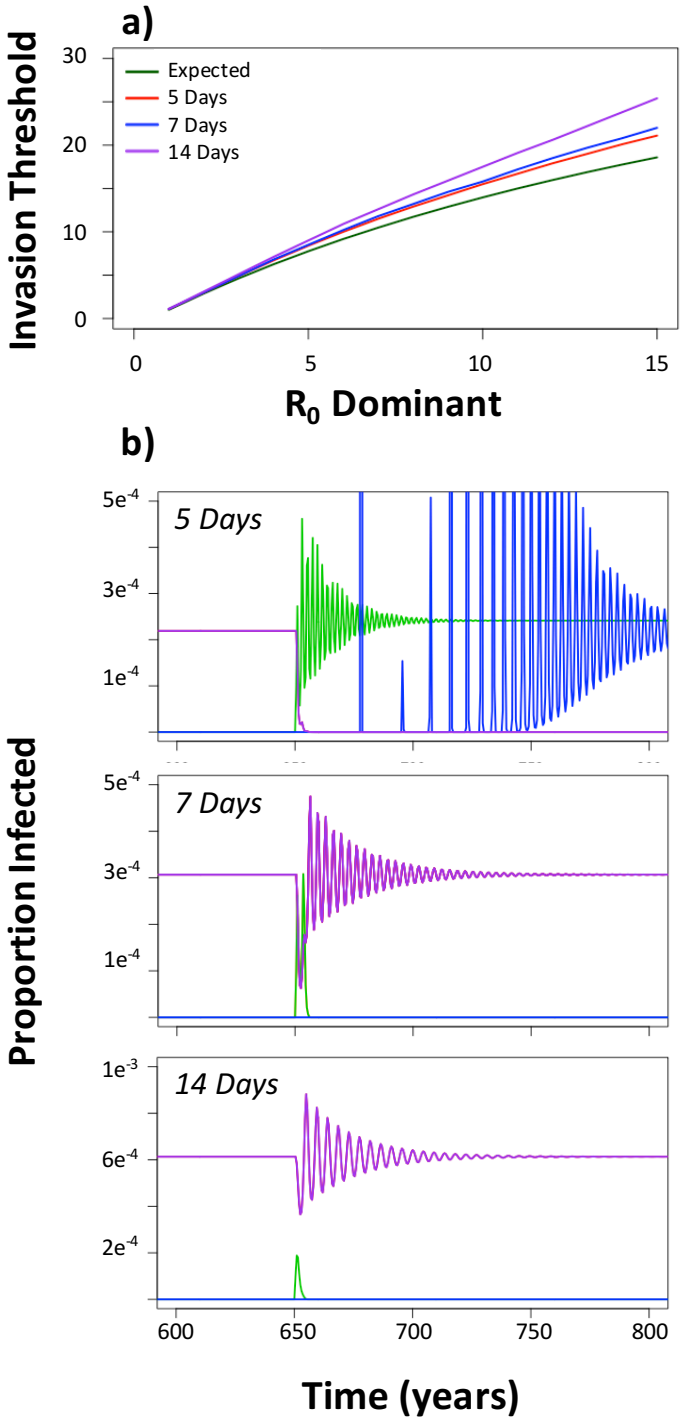


Figure 2.6 The effect of the duration of infectiousness on the invasion thresholds, in a [2 2] (n = 2) system of $\gamma = 0.98$. a) The invasion thresholds for different duration of infectiousness ($1/\sigma$). **b)** Time series for $R_0^{dom} = 5$ and $R_0^{inv} = 8.4$ (the invasion threshold for a recovery of 5 days), where the invading strain was introduced at a time point of 650 years, for different duration of infectiousness. The size of the first epidemic of an invading strain changes in size with the duration, determining whether an invasion can occur and for the two dominant strains to be replaced by the invading strain, despite equal R_0 values for both dominant and invading strains. Once the two previously dominant strains (in purple, of the same prevalence) are removed, the discordant strain (blue) to the invading strain is able to circulate and persist. For all, $1/\mu = 50$ years, and the proportion infectious for each strain is sampled as a uniformly distributed random number of magnitude 10^{-4}.

It is hypothesised that the effect of the duration of infectiousness on the invasion threshold, despite this parameter being absent from the invasion threshold from Equation 2.4, may be caused by transient effects. As the strain is introduced, previously infected individuals remain unable to be infected until recovery from their infection, or new naïve individuals enter via the birth rate. This reduces the susceptible population available at the point of introduction and therefore acts to increase the R_0^{inv} required to invade as this duration increases. A smaller infection period decreases the generation time, and allows the system to respond quicker – meaning the epidemic of the invading strain peaks faster but also clears faster. These transient effects can destabilize DSS, making invasion more likely.

Aside from the effects of the duration of infectiousness, Figure 2.6b also shows that when the invading strain is able to invade and persist, and subsequently the previously dominant strains are removed from the population as they are outcompeted, the other strain which shares no antigenic determinants (i.e. discordant) with the invading strain also emerges. The removal of the dominant strains releases the competitive suppression on the discordant strain, and enables this strain to increase prevalence and persist, despite having no change to its own transmission potential. Therefore, strains can emerge due to ecological interactions with other (evolutionary advantageous) strains, even without any intrinsic changes.

2.3.5 Mutation and migration shifts the R_0 required for invasion

To simulate migration of infected individuals into a susceptible population, a constant seed was introduced to the invasive strain from the start, before the

point of the transmission coefficient change, and this seed was added continuously throughout the simulation. The constant seed acted to boost the immune population of the invading strain, but due to its equal transmission and competitive suppression from the dominating strains, would never be able to spread in the infected population (acting as a “dead end” chain of transmission).

As discussed in section 2.3.2, without any “boost” seed at the point of transmission change, the invasion threshold is very high and far above the expected threshold. This is again demonstrated in Figure 2.7a, which compares the invasion thresholds for various modes of seeding. With a boost seed alone, there is a reasonably close fit to the expected thresholds (as already seen in Figure 2.3 and Figure 2.4). Adding constant background seeding increases the threshold greatly when R_0^{dom} is high. This constant low level of seeding, which can be seen as a proxy for background migration or mutation, makes the system more resilient to invasion than if a strain was introduced only once.

The constant seed can also act to make a difference in the proportion of immune hosts. Initially it was thought that this may reduce the effect of infectiousness duration on the invasion threshold, as described in Figure 2.6a. However, it amplifies the effect. As can be seen in Figure 2.7b, while including a constant seed of 10^{-4}, the invasion thresholds become very high, and a duration of 14 days makes the system practically too stable for invasion to ever occur. Repetitive migration or mutation events reduces the likelihood for invasion in comparison to a one-off event, and this effect is observed across different durations of infectiousness in both Figure 2.6a (boost seed only) and Figure 2.7b (with background seeding).

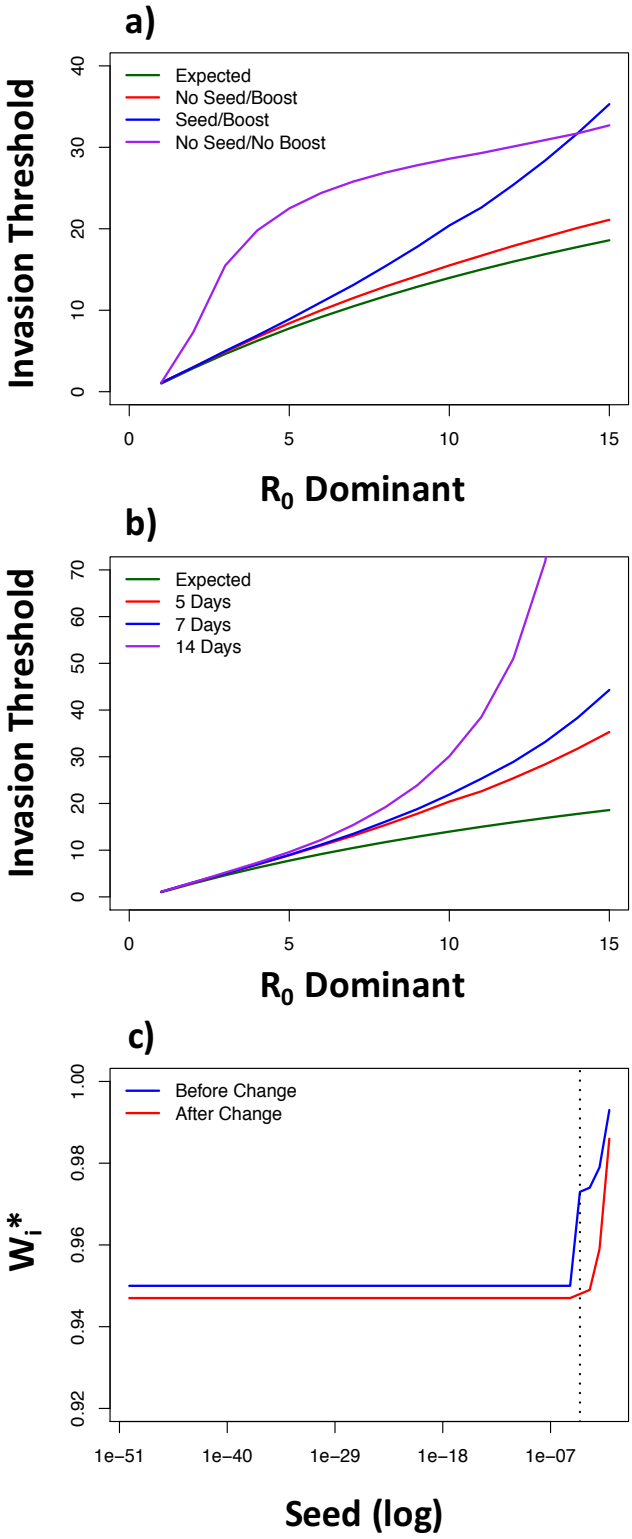


Figure 2.7 The effect of a constant seed on the invasion threshold and immune population, for a [2 2] antigenic structure where γ = 0.98. a) The invasion thresholds for expected values, constant seeding, and a boost seed for the invading strain at the transmission change ($1/\sigma$ = 5 days). **b)** The invasion thresholds for various durations of infection, where there is both a constant seed and a boost seed at transmission change. **c)** The proportion immune to the invading strain (W_{inv}^*) for different constant seeds, while there is always a boost seed of 10^{-4} for the invading strain. This shows W_{inv}^* before the transmission change (blue), after the transmission change (red), and the point at which above this seed value no invasion occurs (dotted black vertical line). R_0^{dom} = 10 and R_0^{inv} = 20. For all, $1/\mu$ = 50 years, and the proportion infectious for each strain is sampled as a uniformly distributed random number of magnitude 10^{-4}.

A reason for the strong stability of the system with constant seeding may be explained by Figure 2.7c. Across a range of constant seed values, above 10^{-4} the immune population of the invading strain (W_{inv}) increased both before and after transmission change, as the seed, which artificially boosts Y_{inv} without being able to transmit, acts to "prime" the population with herd immunity. There is a threshold of the proportion of immune individuals which does not allow any infection.

The different effects of an introduced strain only having a one-off seed at the transmission coefficient change, or with a constant background seed as well, can also be seen for a change in birth rate (μ) (Figure 2.8). For both types of seeding of the infected population of the introduced strain, a shorter lifespan increases the invasion threshold. As observed for Figure 2.7b, a constant background seed of the introduced strain prior to transmission coefficient change increases the invasion threshold, particularly where R_0^{dom} is high (Figure 2.8b). This is unusual, as it would be expected that, in the same way as reducing the duration of infection, a shorter lifespan will increase the generation time and provide more susceptible hosts for the invading strain. Perhaps, however, with a change in birth rate the newly susceptible hosts are completely naïve, and this may act to stabilise DSS and subsequently increase the invasion threshold.

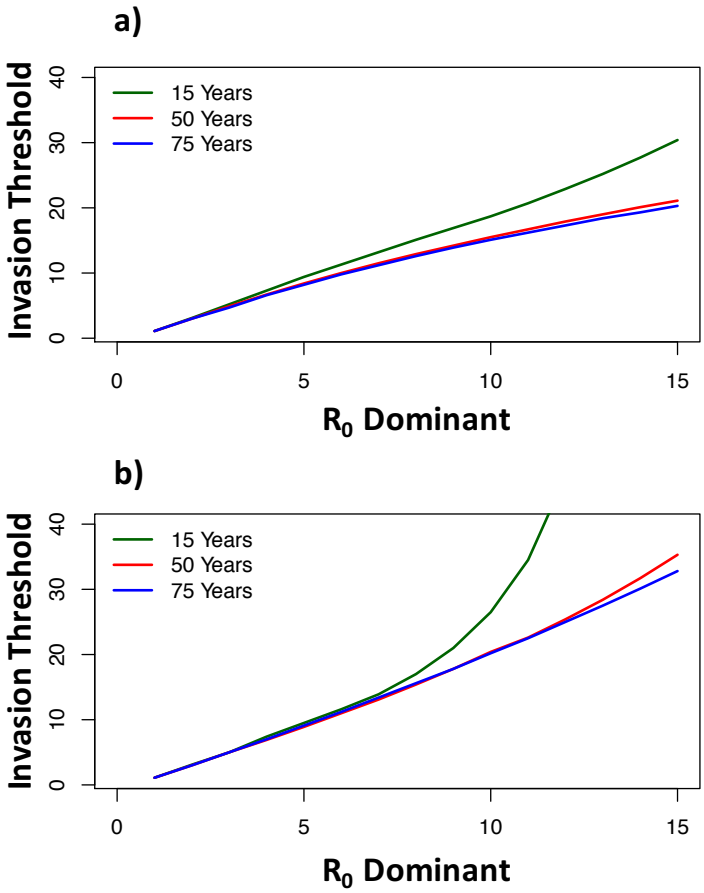


Figure 2.8 The invasion thresholds for different lifespans, where $1/\sigma$ = 5 days, for a [2 2] system, and where a 10^{-4} seed is added to the infected population of the introduced strain at the point of transmission coefficient change. a) The thresholds for where there is no constant background seed to the introduced strain. Decreasing the length of lifespan increases the invasion threshold. **b)** The thresholds for where a constant background seed is applied. The results with a constant background seed are like a), but the invasion thresholds become even higher as R_0^{dom} increases; a constant seed to the invading strain reduces the ability of that strain to invade. The proportion infectious for each strain is sampled as a uniformly distributed random number of magnitude 10^{-4}.

2.3.6 Parameter variation and seasonally forced models allow invasion when R_0^{INV} is low

When looking at invasion of an introduced strain, that strain's infected population is very small, meaning that its success is highly affected by stochasticity. It was important to include stochastic effects into the model for this reason, in line with many other pathogen emergence studies based on SIR frameworks (Alonso et al. 2007; Dalziel et al. 2014; Parham & Michael 2011; Read & Keeling 2007).

Here, stochastic effects could not be included in the typical way, by having stochastic demography, due to the fact that the population size cannot be changed in the MANTIS model (this would be possible if the model was tractable, such as by using an IBM). Instead, parameters were varied in a stochastic way, making them plastic.

Parameter variation was developed for MANTIS as an extra function, by sampling μ and β values annually from Gaussian and Gamma distributions, respectively, with the means set as the deterministic parameter values. The model could also be forced by adding an annual cycle to β (if forcing was included, parameter variation was removed so only one effect would be involved in the pathogen dynamics).

The background seed was found to cause very high invasion thresholds in the deterministic model, so when parameter variation was included it was decided that no constant seed was to be used for the analysis. Only the boost seed at the transmission coefficient change event was included in these analyses.

The proportion of successful events (P_{se}) of invasion and subsequent replacement of the dominant strains was compared for different durations of infectiousness. The threshold shifts to a higher R_0^{inv} as the length of infectiousness increases (Figure 2.9), as seen previously (Figure 2.6a and Figure 2.7b). Parameter variation makes the invasion threshold plastic. The R_0^{inv} values for P_{se} = 1 are slightly higher than the invasion thresholds of the deterministic model (shown as arrows on the x-axis in Figure 2.9). Equally, however, there is success with a lower R_0^{inv} than the deterministic results: parameter variation allows some successful invasion events at a lower R_0.

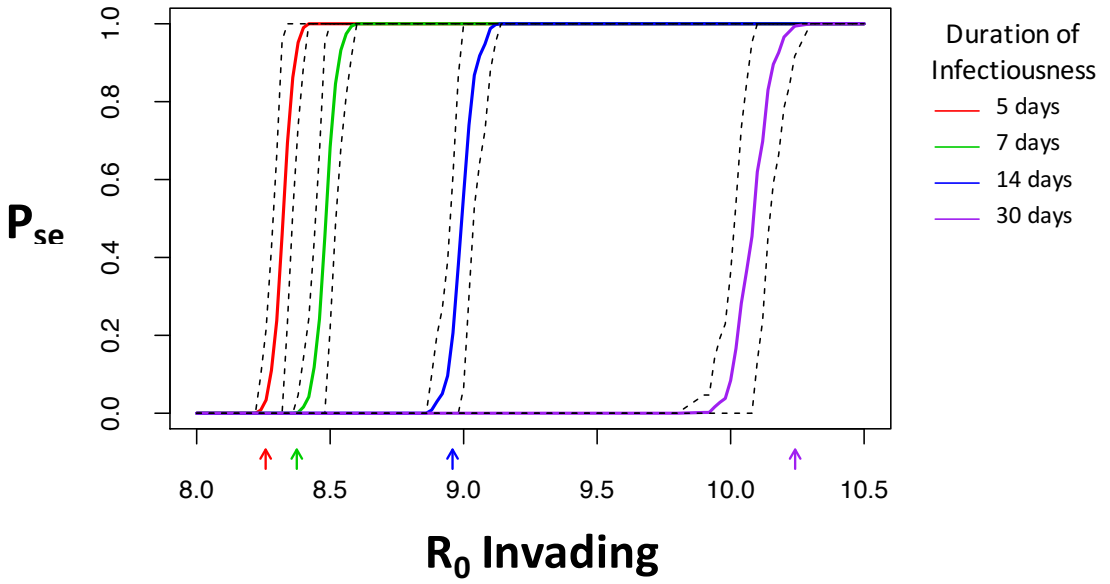


Figure 2.9 The inclusion of variation for both transmission and birth rates provides a plasticity to the invasion threshold, and plotted here is the proportion of successful invasion events (P_{se}) as R_0^{inv} increases for different durations of infectiousness, where R_0^{dom} = 5. Deterministic invasion thresholds are shown by the arrows on the x-axis. As observed in the deterministic model, increasing the duration of infectiousness increases the transmission advantage required for the introduced strain to invade. What can be seen with parameter variation is that an increase in the duration of infectiousness widens the range of transmission potential where the strain could emerge. The 95% confidence limits (black dotted lines) also widens as the duration increases. This is a [2 2] system of $\gamma = 0.98$, $1/\mu = 50$ years, and a boost seed of 10^{-4} but no constant background seeding. Each simulation is run for 500 trials. For all trials, the initial proportion infectious for each strain is sampled as a uniformly distributed random number of magnitude 10^{-4}.

Stochasticity is important in the first stages of a strain's invasion, but an important aspect for a strain's subsequent persistence is seasonality – the fluctuation of transmission rates over the year, for example by changes in temperature, rainfall, and host or vector behaviour. The model was forced by adding a sine function to the transmission coefficient β, with the strength of seasonality determined by the coefficient ε.

It was of interest to test whether introducing a strain at the peak of transmission would aid or hinder the invasion process. Surprisingly, very little to no difference was found (Figure 2.10) for the transmission coefficient change at different time frames and at either the strains' minimum (650 years) or maximum (650.5 years)

transmission potential. The transmission coefficient of the invading strain was also tested after a longer period (1000 and 1000.5 years) to allow the system to reach the DSS equilibrium point – the system can demonstrate cyclical replacement of strains (transient behaviour) for longer in the time series when under stronger seasonal forcing (i.e. equilibrium takes more time to be reached). Again, the time of transmission coefficient change of the introduced strain had little effect on P_{se}.

In addition to the time of introduction, it was also of interest to test the effect of the strength of seasonality, ε. This parameter shifts P_{se}, so that the R_0^{inv} value where emergence is definite (i.e. $P_{se} = 1$) is increased as ε increases (Figure 2.10). It was found that as seasonality effects become high (i.e. as $\varepsilon \to 1$), a lower R_0^{inv} is needed for $P_{se} > 0$. That is, as the amplitude of seasonality increases, the invasion is possible at much lower R_0 values than found with the deterministic model. More than this, though, is that even as $R_0^{inv} < R_0^{dom}$, and more surprisingly at $R_0^{inv} < 1$ (the baseline transmission potential of the invading strain), there is always a small chance of invasion in the presence of strong seasonality.

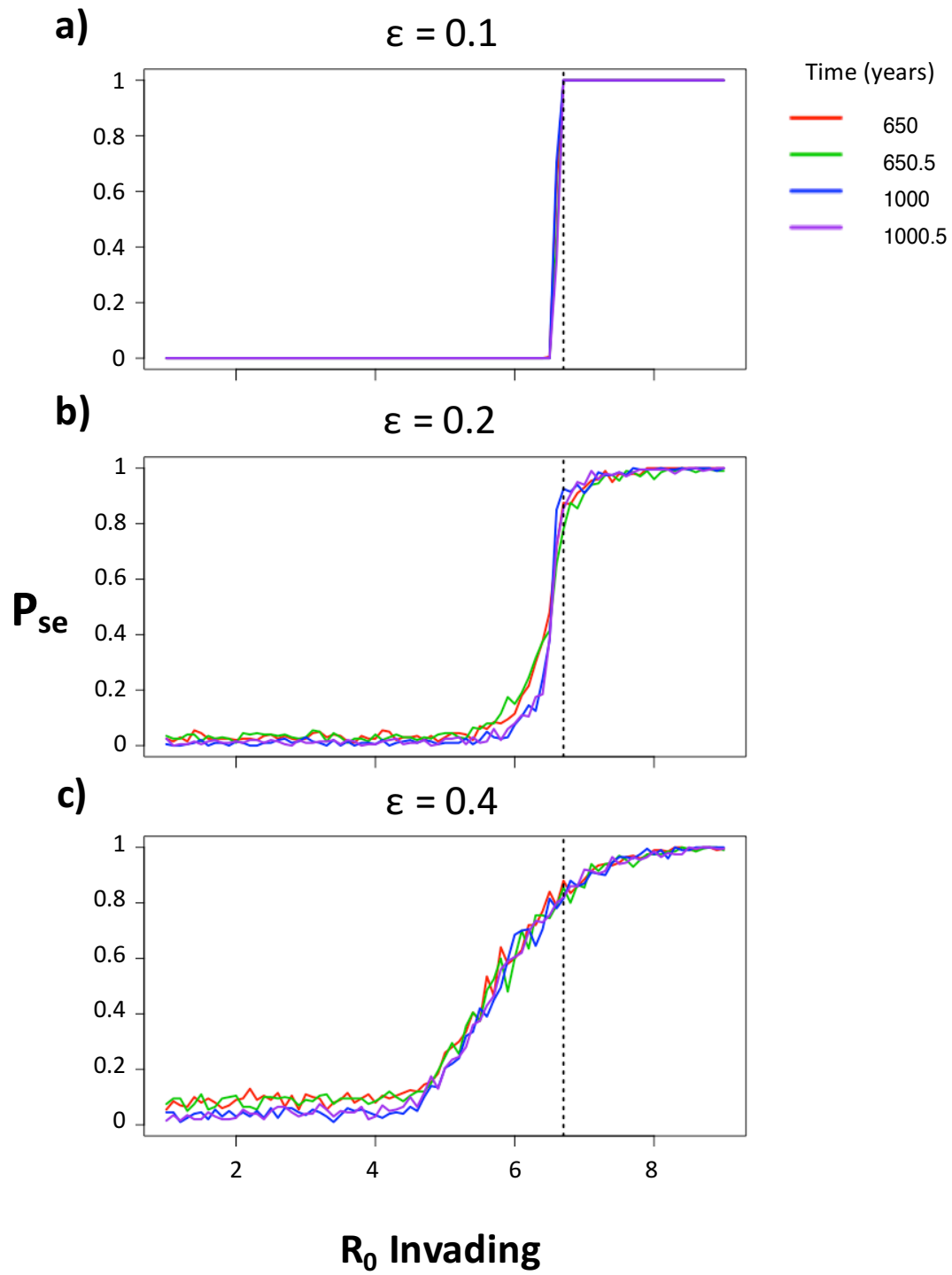


Figure 2.10 The effect of seasonality on the probability of emergence, where $R_0^{dom} = 4$, and the seasonality coefficient is a) 0.1, b) 0.2, and c) 0.4. Increasing the amplitude of seasonality increases the R_0^{inv} values required for $P_{se} = 1$. Additionally, the range that emergence can occur widens, and there is still a small chance of emergence as $R_0^{inv} < R_0^{dom}$ and even as $R_0^{inv} < 1$. The different lines are for different time points in the simulation where the transmission coefficient of the introduced strain was changed. For all, this is a [2 2] system where $\gamma = 0.98$, $1/\mu = 50$ years, and $1/\sigma = 5$ days. The proportion infectious for each strain is sampled as a uniformly distributed random number of magnitude 10^{-4}.

2.4 DISCUSSION

It is understood that strong immune selection can cause the antigenic variants of a multi-strain pathogen to sort into distinct strains with non-overlapping alleles at antigenic loci (Gupta et al. 1996; Gupta et al. 1998). I use a similar framework using the MANTIS R-package (Lourenço et al. 2015), expanded considerably for this research, to explore the system's resilience to invasion by an introduced overlapping strain of a more transmissible phenotype.

Invasion within a multi-locus model is not simply a case of "highest R_0 wins" as described previously (Anderson & May 1992): the transmission potential of an introduced strain must be sufficiently high to overcome the immunological suppression from the pre-existing dominant strains with which they share antigenic alleles. The analytical results in this research have shown that the degree of suppression can be increased by: i) higher cross-protection and therefore inter-strain competition; ii) circulation of more dominant strains which share alleles with the invading strain; iii) a moderate R_0 for the existing strains, which here are assumed to be equal and lower than that of the invading strain.

The analytical results are supported by the numerical simulations, however there are discrepancies between the expected and observed invasion thresholds. In part, this is due to the fact that the definition of invasion in the analytical results refers to the fact that the invading strain is able to circulate, whereas within the simulations this is defined as the ability of a (previously removed) strain to circulate and persist. This *persistence* requires a slight raise in the simulation invasion thresholds.

The invasion simulations were found to be sensitive to other factors, not just those analytically derived (which only include the number of dominant strains and the degree of cross-protection). The model simulations allow the investigation of the effects of these potential additional factors on the invasion threshold. Here, the duration of infectiousness, birth rates, and seeding are found to shift the invasion threshold, even if the deterministic R_0 conditions are met.

Seeding can be a proxy for migration or mutation, as it boosts the infected population for the excluded, and now invading, strain without the need for a transmission event. Consequently, the force of infection increases, and if this is sufficiently high the invading strain can displace previously dominant strains. It has been noted previously that the condition for a strain to emerge is dependent on not only R_0 but also the seed (May et al. 2001; Parham & Michael 2011), and this has also been found here. A higher seed at the point of a strains' introduction lowers the invasion threshold, and this is also a better fit to the analytical thresholds derived from the model ODEs.

Additionally, a background seed was included so that there was a constant flow of the invading strain into the population. Increasing the infected population for the introduced strain does not allow emergence due to the immunological suppression from the dominating strains (until the transmission coefficient changes), but the seed of infected individuals acts to also prime the immune population to the invading strain. This build-up of herd immunity adds an extra level of suppression to the introduced strain, increasing the invasion threshold. Where dominant strains have a high transmission potential, the invasion threshold reaches an almost unachievable level. As the R_0 of the dominant strains determines the proportion immune to the introduced strain, the

combination of both a high dominant transmission and the herd immunity priming from a constant seed creates a highly stable and robust DSS equilibria. These results suggest that a pathogen with the chance of multiple reintroductions – rather than a one-off migration or mutation event – has much greater resistance to invasion via the effect of herd immunity, particularly if that pathogen normally has a high R_0.

The duration of infectiousness, otherwise defined as the recovery rate in transmission models, temporarily delays the transition from infectious to immune and back to susceptible, which transiently reduces available hosts; the restocking of the susceptible population by birth and recovery creates oscillations (dynamical resonance) in an SIR model (Anderson & May 1992). Increasing the duration of infectiousness was found to lessen the amplitude of the initial epidemic of the invading strain, and therefore the chance of emergence and ultimately persistence. It can be suggested from this that for pathogens with longer durations of infectiousness, strains will have to achieve a higher R_0 advantage to invade, even though the conditions for invasion in terms of R_0 alone may have been met.

Including stochasticity in the model was considered important, as the early invasion process involves very small proportions of infectious hosts, which would be highly influenced by stochastic effects. Due to the limitations of the model, which assumes a constant population size, demographic stochasticity could not be included (unlike with an IBM, where this would have been possible). Instead, parameter variation by sampling at random from a distribution was added as a function. This effect was found to provide at least a small chance of a strain to

invade at a lower R_0 than expected from the simulations without parameter variation.

Many multi-strain pathogens demonstrate strong seasonality (such as *P. falciparum*, influenza virus, rotavirus, and dengue virus). Seasonal forcing of the transmission rate was included to make the model more biologically realistic, and this is known to reduce the chance of fadeout. Increasing seasonality shifted the probability of emergence higher, and allowed a small chance of emergence for very low R_0 values – even if the invading strain had a lower transmission potential than the dominating strain (against the idea that the highest R_0 wins (May & Anderson 1983)), and also if the invading strain had a transmission potential of $R_0 < 1$.

The invasion thresholds found here can be very high, and the phenotypic differences required between dominating and invading strains may not be achievable (or would require multiple mutations to be present in the same strain, which is unlikely, and a factor known to decrease the probability of emergence in evolutionary models (Antia et al. 2003; André & Day 2005; Alexander et al. 2010; Arinaminpathy et al. 2009)). There is also a drawback of this analysis in that all other strains are assumed to have equal R_0, which may be biologically unrealistic. Nevertheless, this assumption allowed a clear investigation of the effects of increasing the transmission potential of just one invading strain.

The research presented here demonstrates that multi-strain pathogens which are more prone to invasion by an introduced strain may have the characteristics, either combined or singularly, of: low transmission throughout the endemic population, a low number of competing strains, shorter duration of infection, a

longer lifespan of the host, one-off migration or mutation events (as opposed to continuous), strong seasonality, and a degree of stochasticity.

It has previously been shown, with the multi-locus model framework used here, that birth rates are understood to influence oscillatory dynamics and shift the threshold of DSS (Wikramaratna et al. 2014), and here these results have been expanded to show that a shorter lifespan makes the DSS equilibrium more resilient to infection. This is the opposite to the results of Wikramaratna et al, though the system was fundamentally different – strains were available in both virulent and non-virulent forms, and so invasion was a replacement of the exact same strain rather than a strain of shared alleles (Wikramaratna et al. 2014). This other model also examined emergence in the wake of mixed contact patterns of hosts of different lifespans, which influenced which strain type could "win", and therefore explored, fundamentally, a different epidemiological scenario.

Stochastic effects, as described widely in population genetics theory, are more important in the dynamics of small populations, where genetic drift plays a more significant part in evolutionary change: maladapted genotypes can become fixed, while fitter genotypes can be removed by chance. Where variation in parameter values is included in the model used here, as a proxy for stochasticity, there is a chance of emergence for introduced strains of a lower fitness than possible for the deterministic framework. Populations which are more stochastic in their dynamics are small or fragmented – such as, with spillover events from zoonotic reservoirs of diseases such as ebola, avian influenza, or human immunodeficiency virus (HIV). Zoonotic diseases are known to be much more likely to cause emerging diseases (Taylor et al. 2001).

Viruses, too, are believed to have a higher risk of emergence (Taylor et al. 2001). This risk can in part be due to the higher mutation rates of viruses, but also, as seen here, due to the typically shorter duration of infections (such as influenza (4-5 days (Carrat et al. 2008)) and rotavirus (1-7 days (Ruuska et al. 1990))), compared to weeks or months observed for bacterial pathogens (invasive pneumococcal disease, by *Streptococcus pneumoniae,* can last between two weeks and two months (Sutton et al. 2007)).

Antia et al found that maladapted strains can emerge, but only if evolution could enable that strain to increase its fitness, whereas here this was possible with strong seasonality alone. This suggests that pathogens with strong seasonality can be disrupted by a maladapted invading strain. Rotavirus is a highly seasonal multi-strain pathogen, which also shows cross-reactivity, and has recently had an emergence of strains G9 and G12 (Lin et al. 2006; Miles et al. 2012; Mullick et al. 2014) which are replacing the usual dominant genotypes (Iturriza-Gomara, Cubitt, et al. 2000). It is worth noting here that surveillance for well-adapted pathogens or strains may miss possible invading strains if the effects of seasonality are ignored.

The benefits of using a multi-strain model such as this was the ability to incorporate views of evolutionary fitness (using intrinsic transmission potential) with ecological interactions (from the competitive suppression from dominant strains, via immune selection). This provided the ability of a strain to emerge without a change in transmission (fitness), but simply by ecological interactions. The strains with non-overlapping antigenic alleles were also able to emerge and persist if the focal invading strain was successful, as a result of the removal of competitive suppression from the previously circulating strains. This has

implications for the surveillance of potentially emerging strains in a multi-strain system – strains which have no shared antigenic determinants may enable each other's emergence, without the requirement of those strains to increase in fitness.

This study has brought together modelling techniques from the immunological competition perspective of multi-locus models with the emergent disease perspective from evolutionary, phenomological models. I have shown that a multi-locus system under high immune selection exhibiting a stable and structured antigenic diversity, despite being robust to recombination (Gupta, 1996), can be disrupted by the introduction of an overlapping strain, which was excluded or driven to extinction by immune selection. Nevertheless, the potential for phenotypic variation must be high, as a strain of a different phenotype will not be guaranteed to displace existing strains unless above a certain invasion threshold. This threshold is dependent on a multitude of factors including baseline transmission, duration of infectiousness, seasonality, parameter variation, and the degree of mutation and migration of a pathogen.

3. The Effects of Heterogeneity in Host Immune Responses on Pathogen Population Structure

3.1 Introduction

Immune responses to pathogens differ in strength, longevity, and specificity. Within a theoretical "response space", antigens will locate at different points with these characteristics as the axes, as demonstrated on one plane for specificity and strength in Figure 3.1. Strongly protective antigenic determinants can be highly diverse, as well as conserved – and the dominance of these antigenic types determines the pathogen's population pattern. As described in the Introduction, previous multi-locus models, of which this research is based on, predict that pathogens containing conserved antigens which elicit a strong response circulate as single strains (e.g. measles), whereas multi-strain systems are typically dominated by strain-specific immune responses towards variable antigens (e.g. influenza A).

The strength of the immune response to an antigen is determined by its ability to protect against other strains which contain the same antigen, and is a combination of both the strength of the response (e.g. the neutralising antibody titre) and the longevity of that response (e.g. the lifespan of memory B cells). Diversity is a combination of the antigenic diversity, defined as the variability in the determinants at that antigenic locus, along with the specificity of the host immune response. For instance, a highly diverse antigen will contain many different antigenic determinants and immunity is prone to be strain-specific, so that infection against one strain will only protect against the determinants of that

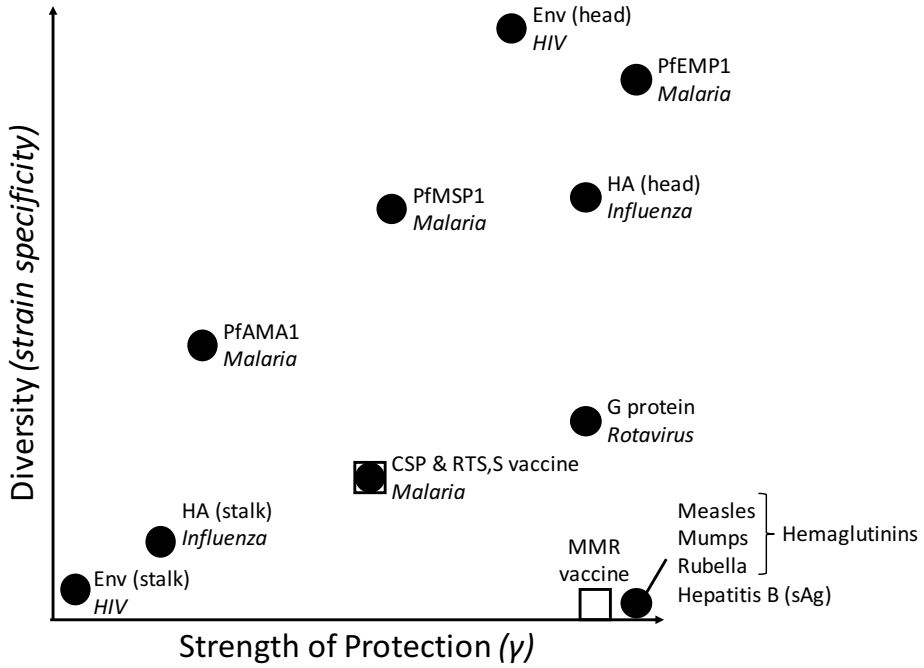


Figure 3.1: The "response space", where antigens elicit responses which vary in their diversity and strength. The example antigens are: the malarial proteins of *Plasmodium falciparum* of PfEMP1 (Newbold et al. 1992; Baruch et al. 1995), PfAMA1 (Remarque et al. 2008), PfMSP1(Fowkes et al. 2010), CSP and the CSP-based vaccine RTS,S (Chaudhury et al. 2016; Ockenhouse et al. 2015); the influenza HA head (Smith 2004) and stalk (Okuno et al. 1993; Sui et al. 2011); the human immunodeficiency virus (HIV) envelope protein head and stalk (Burton & Mascola 2015); rotavirus G protein (Velázquez et al. 1996; Franco et al. 2006); the surface antigen (sAg) of hepatitis B (Trepo et al. 2014); and hemagglutnin proteins of measles, mumps, and rubella (Amanna et al. 2007), as well as the MMR vaccine (Amanna et al. 2007).

strain and no others. Conversely, if an antigen of low diversity with conserved antigens produces a strong response, infection by the pathogen will protect against all other possible strains.

The dominant antigen of childhood infections such as measles and mumps have no diversity and elicit a strong reaction, with the immune response being non-specific and re-infection does not occur, and immunity is life-long (Amanna et al. 2007). Rotavirus is another major childhood disease (mainly infecting those <5 years of age), but circulates as multiple diverse strains. Rotavirus infection provides cross-protection to other strains of the same and heterologous genogroups (Greenberg et al. 1983; Velázquez et al. 1996), where genogroups are defined by their antigenic region in the gene VP6 (Iturriza Gomara et al.

2002). Intestinal antibodies are short-lived (Franco & Greenberg 2001), but re-infection allows an accumulation of cross-protection – complete protection to disease symptoms is acquired after two or more infections (Velázquez et al. 1996).

A strong, specific, and long-lasting immune response creates high immune pressure on the pathogen, resulting in highly variable antigens. The malarial antigen *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) is encoded by 60 variants of the *var* gene (Smith et al. 1995; Su et al. 1995), and creates a highly specific antibody response, with low or no cross-reactivity (Newbold et al. 1992; Baruch et al. 1995). Protection is of a long duration and accumulates with age after exposure to more PfEMP1 types (Chattopadhyay et al. 2003; Barry et al. 2011). Another antigen is the merozoite surface protein 1 (PfMSP1), which is of a reasonably high diversity and responses are strongly-strain specific but the immune response is far weaker than PfEMP1 (Fowkes et al. 2010). Apical membrane antigen 1 (PfAMA1) has a low diversity, and the immune system can respond to in a similar way to several different allelic variants (Remarque et al. 2008).

The pre-erythrocytic stage antigen of the circumsporozoite protein (PfCSP) has a low diversity with conserved epitopes, and elicits a protective response. This antigen is being used to develop the *P. falciparum* vaccine RTS,S to elicit a similar but longer lasting response (Chaudhury et al. 2016; Ockenhouse et al. 2015). The hepatitis B surface antigen (sAg) is being used in the RTS,S vaccine to improve the response as it is strongly protective, as well as having highly conserved epitopes (Trepo et al. 2014).

The envelope (*env*) surface antigen of human immunodeficiency virus (HIV) elicits different antibodies. These are highly strain-specific antibodies (being incapable of binding heterologous strains) to the variable loops on the head of *env,* and these antibodies are immunodominant and strongly protective (Burton & Mascola 2015). A second type of antigen is to a conserved region on the "spike" of *env*, which is broadly neutralising but poorly protective (Burton & Mascola 2015).

The haemagluttinin (HA) protein of influenza A has highly variable regions in the head of the protein, which is exposed to the immune system. This region acts as an immunodominant antigen, and responses are highly specific (Smith 2004). However, the stalk of HA is less exposed, and its function constrains variability, resulting in conserved regions – anti-stalk antibodies are rare, but are observed to be broadly neutralising (Okuno et al. 1993; Corti et al. 2010) against different influenza A groups (Wang et al. 2010; Krause et al. 2010; Sui et al. 2011).

Many antigens eliciting strong immune responses are either conserved or diverse, and the dominance of these antigens in the immune response determine the competition between strains and subsequently the observed diversity of the pathogen. Responses to conserved antigens create competitive exclusion, where only one strain circulates, as in measles (Rota et al. 2011), whereas responses to variable antigens enable the circulation of multiple strains – but the degree of immunological competition between strains determines the diversity and dynamics of these strains.

Pathogens contain many different proteins, however from the perspective of the immune system the pathogen is recognised as a string of epitopes. Multi-locus

modelling uses this view of “strings” of antigenic determinants as a framework to determine what constitutes a strain, and these antigenic determinants may vary between strains (variable determinants) or be present on all possible strains (conserved determinants). Figure 3.2 describes how immune responses of different specificity and longevity have been modelled previously at the population level.

The most basic form of modelling is where immunity is acquired to a conserved antigen only (Figure 3.2A), present on all possible strains, and this leads to competitive exclusion, where the strain of the highest transmission “wins” (May & Anderson 1983). This results in the persistence of just one dominant strain, as observed for the measles virus (Rota et al. 2011).

The models represented by Figure 3.2B are typical multi-locus models where no conserved determinants are included: immunity is only acquired to variable determinants (Gupta et al. 1996; Gupta et al. 1998). Immunity to one strain’s variable determinants can allow the recognition of these determinants in other strains, and this cross-recognition leads to cross-protection. Cross-protection provides a probability of immunity to antigenically similar strains, reducing their transmission, without the need for exposure to those strains. This provides a degree of inter-strain competition, as strains which share antigens compete for susceptible hosts.

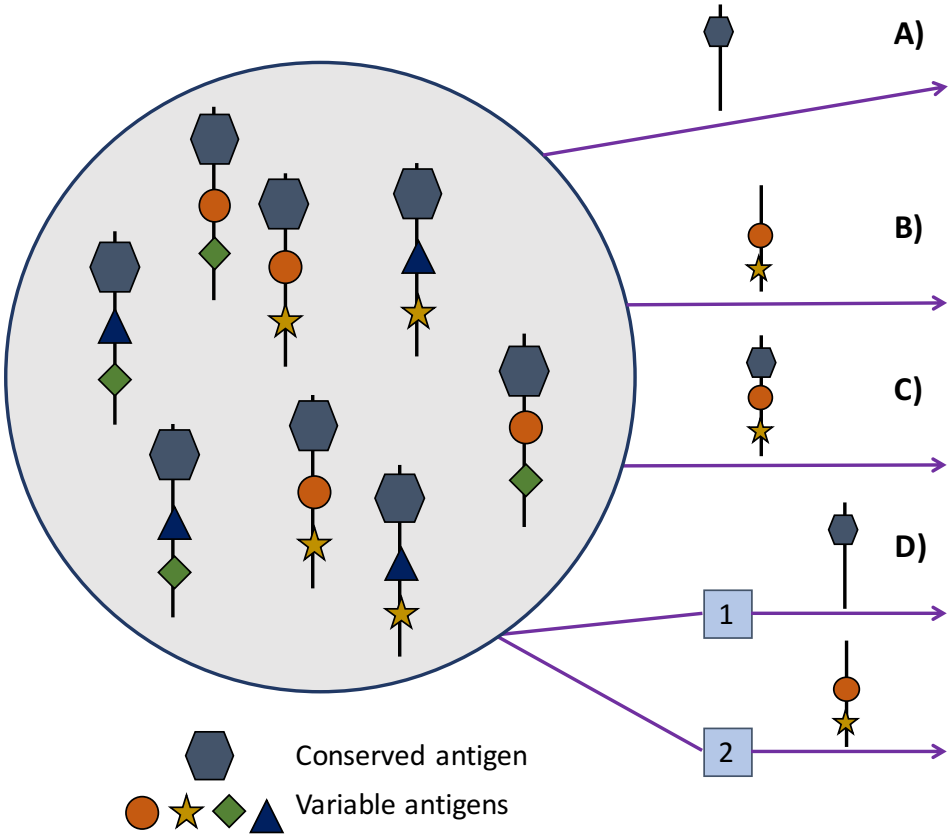


Figure 3.2 Diagram of previous multi-locus models. Strains, indicated by the groups of shapes in the circle, are defined by their combination of antigens, which here are displayed by their variable and conserved epitopes. From this pool of strains, pathogen population dynamics have been modelled with hosts acquiring immunity in different ways: **A)** Immunity is to a conserved epitope only, and so after one infection the host cannot be infected with any other strain; **B)** The host acquires immunity to variable epitopes only, termed as a "multi-locus model" (Gupta et al. 1998), and there can be cross-recognition between strains which share variable epitopes, leading to additional immunity and a degree of inter-strain competition; **C)** Immunity is acquired to both conserved and variable antigens, although immunity to conserved antigens is only temporary**; D)** Hosts acquire immunity in two different ways, either to variable epitopes (1) or to conserved epitope (2), previously this has been achieved by including hard-wired host genotype differences in the acquisition of immunity.

Cross-protection leads to three type of pathogen population dynamics (Gupta et al. 1998), explained in more detail in Chapter 1. At a low degree of cross-protection, inter-strain competition is low or non-existent and so strains circulate independently of each other. This allows all strains to coexist, and the population reaches a stable equilibrium (no strain structure, NSS). Another stable state is found where cross-protection is high. Here, strains greatly limit the transmission of other strains which share antigens, and so to avoid inter-strain competition the

strains which circulate are discordant, consisting of non-overlapping antigens (discrete strain structure, DSS). Any strain which shares an antigenic determinant with the dominating strains are competitively excluded from the pathogen population.

Discrete strain structuring has been observed in several pathogens, such as the malarial parasite *Plasmodium falciparum* (Fowler et al. 2002; Tami et al. 2003; Barry et al. 2007; Chen et al. 2011), and the meningitis-causing bacteria *Neisseria meningitidis* (Buckee et al. 2010; Urwin et al. 2004; Callaghan et al. 2008; Watkins & Maiden 2012) and *Streptococcus pneumoniae* (Babl et al. 2001).

For an intermediate level of cross-protection, where neither of the stable states occur, the system displays cyclical strain structure (CSS). Cyclical dynamics are observed in many multi-strain RNA viruses. Influenza A displays one dominant strain at any point in time (Fitch et al. 1997; Fitch et al. 1991; Bush et al. 1999), and the successive dominance of strains is also observed, albeit with more strain coexistence, in rotavirus (Page & Steele 2004; Palombo et al. 1993; Doan et al. 2011), respiratory syncytial virus (RSV) (Cane & Pringle 1995; Zlateva et al. 2005), and foot and mouth disease virus (FMDV) in cattle (Subramaniam et al. 2013).

In Figure 3.2C, models of within-host dynamics have included responses to both conserved and variable antigens concurrently in models of HIV (Nowak & May 1991; Wikramaratna et al. 2015) and influenza A (Recker et al. 2004; Ferguson et al. 2003; Minayev & Ferguson 2009), though strain-transcending immunity to the conserved determinants is temporary. Incorporating temporary strain-

transcending immunity lowers strain diversity and introduces a sequential replacement of strains (Recker et al. 2004; Ferguson et al. 2003; Minayev & Ferguson 2009), making the dynamics influenza-like.

The example of influenza A HA (and HIV *env*) in Figure 3.1 shows how immune responses to both specific and conserved antigenic determinants can be triggered by the same protein. Life-long responses to both variable and conserved determinants have been included by (Gupta & Galvani 1999), shown in Figure 3.2D, where two host genotypes are included, with one host genotype responding to a conserved determinant (1) and the other genotype to variable determinants (2) (Gupta & Galvani 1999). The results of this model of host heterogeneity suggested that discrete strain structure can be maintained for the pathogen population even if most the host population do not respond to variable, strain-specific determinants.

Heterogeneity in host response has also been explored in other ways. In a similar manner to Gupta and Galvani, a hard-wired difference in immune responses was included by making responses dependent on the recognition alleles of the host (specifically the human leucocyte antigens, or HLAs) (Penman et al. 2013). Gog et al used a different method of cross-protection to variable determinants, using polarised immunity to allow a difference in individual host responses after exposure to the same strain (Gog & Swinton 2002). Cobey et al adapted the multi-locus model of (Gupta et al. 1998) to incorporate a degree of breadth so that hosts could vary in the number of variable determinants they could become immune to after an infection (Cobey & Pascual 2011).

Immune responses may not be entirely towards variable, or conserved, determinants but across a range of the “response space” – a characteristic not captured in previous mathematical models. Equally, individuals, even when infected with the same strain, may not mount the same immune response. There remain unanswered questions: what is the effect on pathogen dynamics if a single host can acquire long-term immunity to either variable strain-specific or to conserved determinants? What if individual hosts differ in their probability of acquiring immunity to different antigenic determinants? This research aims to explore how a variation in specificity of the host immune response may influence the population dynamics of the pathogen, by developing a new multi-locus model incorporating variability in the host immune response, and then contrasting this model with the results of (Gupta & Galvani 1999) where differences in the immune response are hard-wired.

3.2 METHODS

A multi-locus model is adapted from (Gupta et al. 1996; Gupta et al. 1998), firstly developed into a model of one host genotype, where a host acquires immunity to either the variable or conserved antigenic determinants of a strain, and then a second model is developed which includes two host genotypes, to explore host heterogeneity in immunity.

This is a status-based model, a technique which enables a large strain space, with the number of loci and alleles being altered easily. Consequently, this model does not track exposure history, making it problematic to include a broad immunity to all strains via the conserved determinant. For instance, a host may become immune to all possible strains after multiple infections, however,

intractability makes this immune status impossible to deduce. To overcome this, a separate compartment for immunity to a conserved determinant is included, known as '*X*' (Figure 3.3). Individuals can enter this compartment after an infection with probability ψ, or acquire strain-specific immunity to variable determinants ('*W*', the bold line surrounding all circular compartments containing strain-specific immunity classes Z_{ax} etc.).

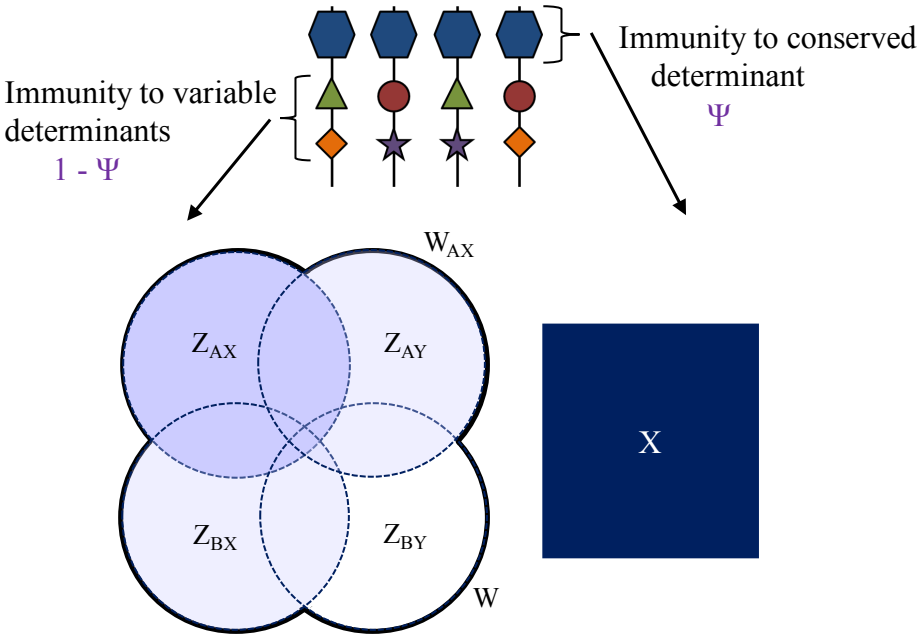


Figure 3.3 The model. Strain-specific immunity ('*W*') is shown as the dark blue bold line of all the circles, and broad immunity to a conserved antigen ('*X*') is the rectangle. This schematic diagram is based on a [2 2] system of loci A/B and X/Y. After an infection with strain *ax*, individuals can enter *X* at a probability ψ, or enter *W* (via Z_{ax}) at a probability 1 – ψ. W_{ax} is shown by the shaded circles, which is the space where individuals are immune to all strains which share any alleles with strain *ax,* including *ax* itself.

Strains are defined by their combination of the alleles at available antigenic loci, for example [2 3 5] indicates three loci of two, three, and five alleles. No mutation or recombination is included in the model.

The ordinary differential equations (ODEs) for the model described in Figure 3.3 are described below. As defined in the Introduction section 1.3, the parameters

μ and σ are the background mortality and loss of infectiousness, respectively, and γ is the degree of cross-protection to strains which share at least one allele.

Within this multi-locus model, two new parameters are added: ψ, the probability of acquiring immunity to conserved antigens (therefore entering category 'X'); and N, the host genotype.

For the one host model, there is only one host genotype which after infection can acquire immunity to conserved antigens and enter category 'X' by the probability Ψ, or acquire immunity to strain-specific antigens and enter 'W' with probability 1 – Ψ. The force of infection is defined here as $\lambda_i = \beta Y_i$ where β is the transmission coefficient. The reproductive rate, defined as the average number of secondary cases generated by one primary case in a totally susceptible population (Anderson & May 1992), is given here by $R_0 = \frac{\beta}{\mu+\sigma}$.

For the two host model, the subscript N = A or B for host genotype is used, where genotype A has no conserved immunity ($\psi^A = 0$) and genotype B has only conserved immunity ($\psi^B = 1$). The host populations are linked by their force of infection (genotype A can infect genotype B and vice versa), so that $\lambda_i = \beta[(1-P)Y_i^A + PY_i^B]$ where *P* is the proportion of host genotype B.

Proportion immune to strain i, Z_i^N:

$$\frac{dZ_i^N}{dt} = \lambda_i(1 - \psi^N)(1 - X^N - Z_i^N) - \mu Z_i^N \quad (3.1)$$

Proportion immune to any strain which shares alleles with i (notated as j~i), including i itself, W_i^N:

$$\frac{dW_i^N}{dt} = \sum_{j\sim i} \lambda_j (1 - \psi^N)(1 - X^N - W_i^N) - \mu W_i^N \quad (3.2)$$

Proportion immune to all specific antigens, W^N:

$$\frac{dW^N}{dt} = \sum_k \lambda_k (1 - \psi^N)(1 - X^N - W^N) - \mu W^N \quad (3.3)$$

Proportion immune to all strains via broad immunity to a conserved antigen, X^N:

$$\frac{dX^N}{dt} = \sum_k \lambda_k \psi^N (1 - X^N - W^N) - \mu X^N \quad (3.4)$$

Proportion infectious with strain i, Y_i^N:

$$\frac{dY_i^N}{dt} = \lambda_i[(1 - X^N - Z_i^N) + (1 - \gamma)(W_i^N - Z_i^N)] - (\sigma + \mu)Y_i^N \quad (3.5)$$

In Equation 3.3 and 3.4, the sum is over all strains *k*, since hosts enter W^N or X^N (depending on ψ^N) after acquiring immunity to any of the strains.

The classifications for NSS, CSS and DSS are the same as used in (Wikramaratna et al. 2014), and are defined as:

- NSS: individual strains do not cycle and there is no difference in prevalences (Y_i^N) between strains.
- DSS: individual strains do not cycle and prevalences between strains differ by > 10^{-4}.
- CSS: individual strains cycle with amplitude > 10^{-4}, and prevalences between strains differ by > 10^{-4}.

Within an area of CSS, there can be different levels of dominance of a set of strains, or a single strain. Single strain dominance is a quantitative measure of this dominance over the range of CSS, by comparing the prevalence of the two most dominant strains over a series of epidemics (Recker et al. 2007; Wikramaratna et al. 2013), defined as Y_i^{max} and Y_i^{sub} where 'max' is the most dominant and 'sub' the second most dominant (i.e. subdominant) strain. Averaging over *P* epidemics calculates this value, ε, defined as:

$$\varepsilon = \frac{1}{P}\sum_{i=1}^{P}\frac{Y_{max}^{i} - Y_{sub}^{i}}{Y_{max}^{i}} \quad (3.6)$$

Where a high ε value corresponds to strong single strain dominance, where a dominant strain is of much higher prevalence, and/or for longer durations of time throughout the simulation.

Throughout this chapter, the last 500 generations of a total of 1000 generations were used for analysis, unless otherwise stated. The parameters were set at 1/μ = 50 years, and σ = 10 years^{-1} (36.5 days), so β is altered depending on the desired R_0 values used.

3.3 RESULTS

3.3.1 THE EFFECT OF IMMUNITY TO CONSERVED DETERMINANTS ON STRAIN STRUCTURE

It is apparent that despite the addition to the original multi-locus model (Gupta et al. 1998) of an ability of the host to acquire immunity to a conserved determinant, which subsequently provides a broad immune response to all strains, the three strain structures (NSS, DSS, and CSS) are maintained. Example time series for the different behaviours across the range of cross-protection can be seen in Figure 3.4, for a [2 3 5] antigenic structure. For a low cross-protection, all 30 strains in the system persist at the same prevalence (NSS), and at a high cross-protection of 0.9 DSS occurs with nine discordant strains persisting at the same prevalence and all other strains being removed. Different cyclical dynamics (CSS) for the mid-range of cross-protection can be observed, which correspond to different values of single strain dominance (ε), where a higher ε corresponds to stronger dominance of strains within the epidemic cycles.

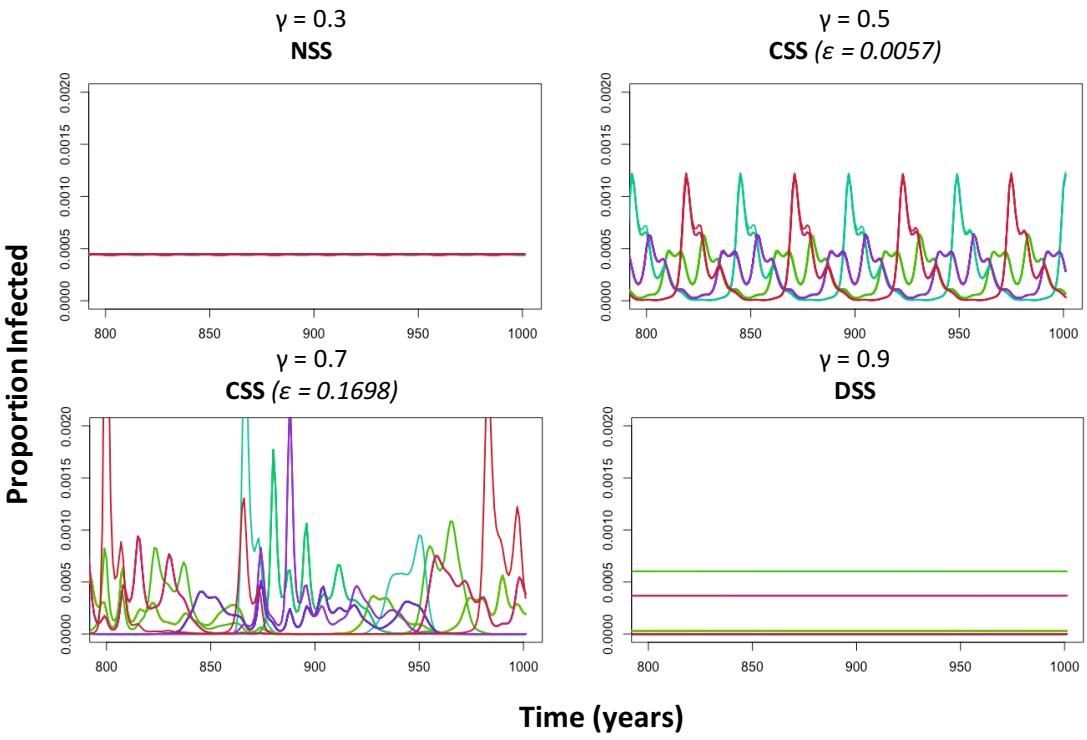


Figure 3.4 Time series for the one host model with a [2 3 5] antigenic structure, over a range of γ and where Ψ = 0.6, 1/μ = 50 years, 1/σ = 36.5 days, and β = 50 (R_0 = 4). The infected population is initially at 0.001, with one strain slightly higher at 0.00101. These series show the dynamics for all 30 strains, with the single strain dominance (ε) shown for where CSS occurs. The plots clearly show the change in dynamics from NSS (all 30 strains co-circulating), CSS (cyclical dynamics) and DSS (only a subset of 9 discordant strains co-circulating).

To understand the effects of both cross-protection and conserved immunity on the pathogen population dynamics, three behaviours are classified over a range of ψ and γ combinations, with CSS being further classified into a quantitative scale of single strain dominance (ε).

This classification is obtained for two different antigenic structures ([2 3 5] and [6 5]), both of a total of 30 strains each. The [6 5] system is analysed after a longer time, as it requires a longer time to reach equilibrium (to reduce the "undefined" region, where none of the three behaviours can clearly be classified, for instance where all strains are the same prevalence but oscillate over time, so is neither NSS nor CSS). For the one host genotype model, the probability of a host to acquire immunity to a conserved determinant (Ψ) is changed, which alters the

breadth of immune response. For the two host genotype model, where genotype A is assigned no conserved immunity ($\psi^A = 0$), and genotype B is assigned complete conserved immunity ($\psi^B = 1$), the proportion of genotype B (P) is changed. Both ψ and P influence the probability of conserved immunity – the former more directly, by altering the probability of this immunity within the host, with P altering the proportion of hosts with only conserved immunity at the population level.

Independently of antigenic structure and the number of host genotypes, DSS is still observed even with a high likelihood of conserved immunity (Figure 3.5). The general pattern here suggests that as long as there is a small chance of strain-specific immunity, and consequently inter-strain competition, strain structuring can occur. Thus, having a small probability of strain-specific immunity is enough to allow immunological competition and the suppression of subdominant strains, resulting in competitive exclusion.

For both the one host (Figure 3.5a and b) and two host (Figure 3.5c and d). models, the cross-protection space of the three strain structures shifts; Ψ and P both act to lower the cross-protection threshold for CSS and DSS to emerge.

As cross-protection increases, the metric for single strain dominance increases across the range of CSS, as the stronger cross-protection and transition into DSS allows strains to persist and dominate for longer. Interestingly, at very high levels of Ψ, CSS is not displayed in the one host model (Figure 3.5a and b). Although both cross-protection via immunity to shared variable determinants (γ) and

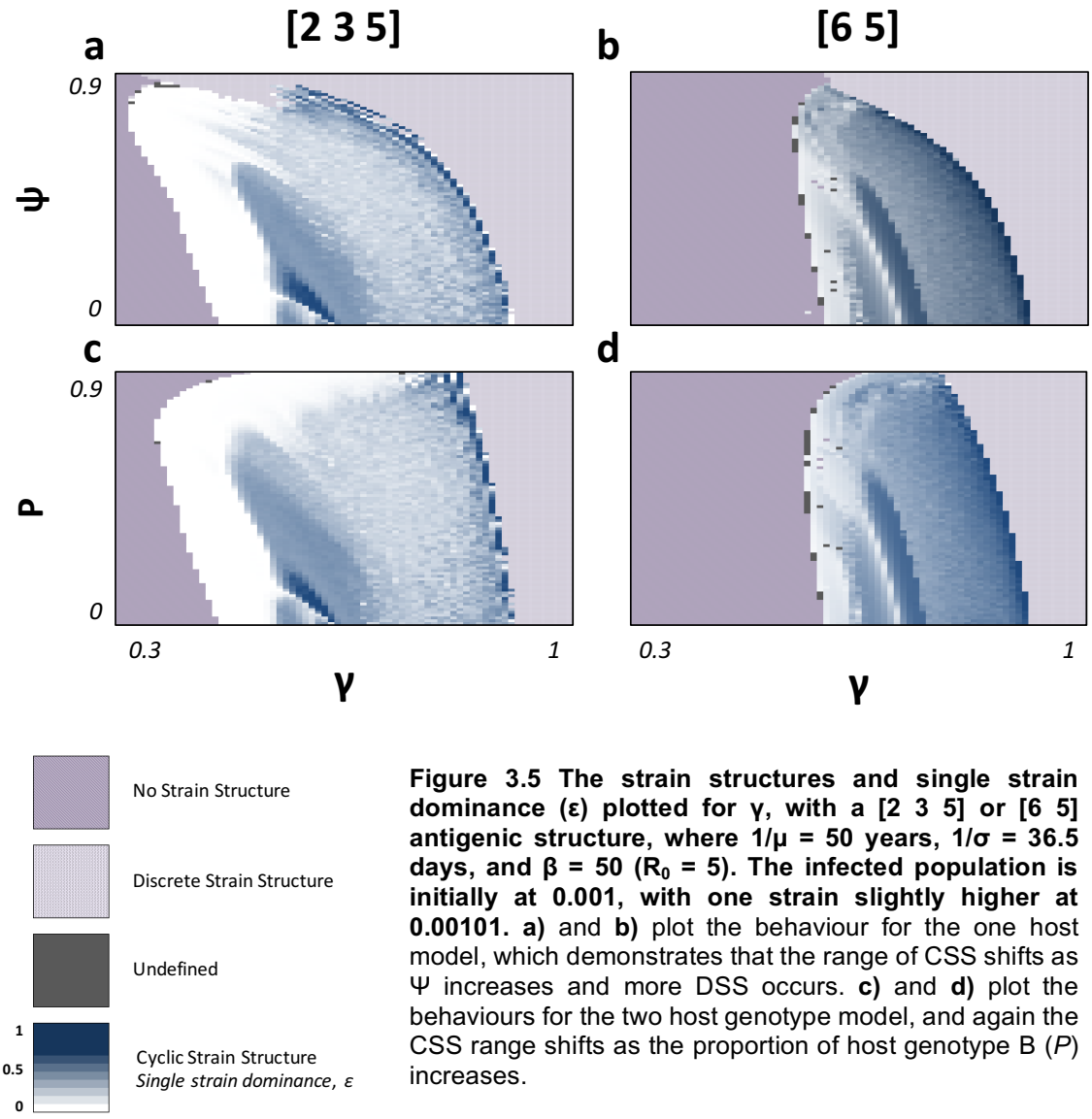


Figure 3.5 The strain structures and single strain dominance (ε) plotted for γ, with a [2 3 5] or [6 5] antigenic structure, where 1/μ = 50 years, 1/σ = 36.5 days, and β = 50 (R_0 = 5). The infected population is initially at 0.001, with one strain slightly higher at 0.00101. a) and **b)** plot the behaviour for the one host model, which demonstrates that the range of CSS shifts as Ψ increases and more DSS occurs. **c)** and **d)** plot the behaviours for the two host genotype model, and again the CSS range shifts as the proportion of host genotype B (*P*) increases.

immunity to conserved determinants (Ψ) impose immune selection on the pathogen, cross-protection imposes frequency-dependent selection, whereas the latter affects all strains equally. This may explain why the pathogen population does not display any CSS when Ψ is very high, as immune responses are less likely to be strain-specific and therefore immune selection is less frequency-dependent: any strain (or a discordant set of strains) reaching dominance is not at a disadvantage via frequency-dependent selection at any point in time, and is therefore able to persist.

By contrast, within the two host model, while a very high proportion of genotype B (up to the limit of *P* = 0.9) causes the range of CSS to contract, cycling still occurs (Figure 3.5c and d). Unlike the one host model, here there are still individuals of another genotype that acquire immunity to *only* variable and cross-reactive antigens. This provides enough strain-specific competition for cyclic behaviour to remain, even if the degree of conserved immunity is very high.

3.3.2 THE INFLUENCE OF R_0 ON THE CROSS-PROTECTION THRESHOLDS

The results displayed in Figure 3.5 are the dynamics where all strains are of R_0 = 5, but by changing R_0 values (via changing the transmission coefficient β) the effect of transmission on the strain structures can be explored (Figure 3.6). An increase in R_0 increases the γ threshold required for DSS. The dependence of R_0 on this threshold is clear from the analytically derived critical γ for DSS (Gupta et al. 1998):

$$\gamma_c = \frac{1 - 1/R_0^i}{1 - 1/R_0^i R_0^j} \tag{3.7}$$

Where *i* is the dominant discordant set and *j* is the removed set of strains.

The cross-protection thresholds for strain structures can be seen in Figure 3.7, where again it is clear how the cross-protection thresholds for stable NSS and DSS states increase with an increase in the transmission potential of the system.

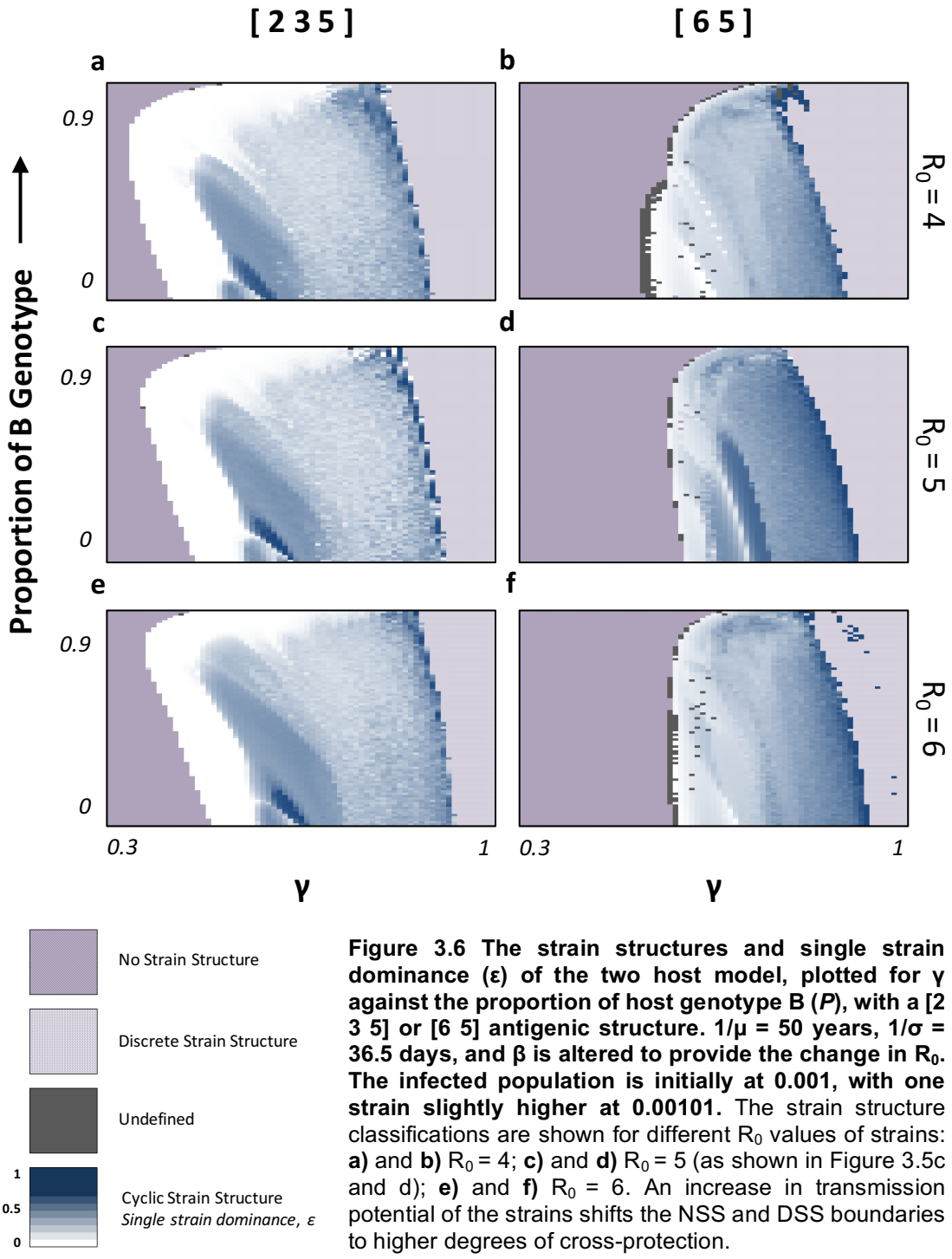


Figure 3.6 The strain structures and single strain dominance (ε) of the two host model, plotted for γ against the proportion of host genotype B (*P*), with a [2 3 5] or [6 5] antigenic structure. 1/μ = 50 years, 1/σ = 36.5 days, and β is altered to provide the change in R_0. The infected population is initially at 0.001, with one strain slightly higher at 0.00101. The strain structure classifications are shown for different R_0 values of strains: **a)** and **b)** $R_0 = 4$; **c)** and **d)** $R_0 = 5$ (as shown in Figure 3.5c and d); **e)** and **f)** $R_0 = 6$. An increase in transmission potential of the strains shifts the NSS and DSS boundaries to higher degrees of cross-protection.

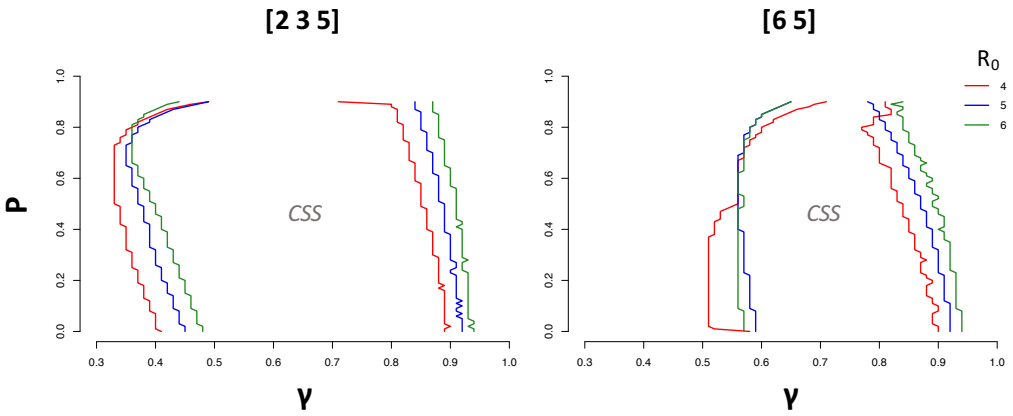


Figure 3.7 The boundaries of the strain structures for different values of R_0, for the two host model with either a [2 3 5] or a [6 5] antigenic structure. As the transmission potential of the strains increases, the degree of cross-protection required for the NSS/CSS boundary and CSS/DSS boundary increases. $1/\mu$ = 50 years, $1/\sigma$ = 36.5 days, and β is altered to provide the change in R_0. The infected population is initially at 0.001, with one strain slightly higher at 0.00101.

3.3.3 The effects of differences in R_0 between strains

To not only explore the effects of transmission on the cross-protection thresholds, but on differences of transmission *between* strains, the R_0 of just one strain was increased via a change in β (to have an R_0 = 7 instead of R_0 = 5) to create heterogeneity in transmission.

Including differences in the R_0 values of strains into the two host model widens the region of DSS even further as *P* increases (Figure 3.8a and b). A higher R_0 of one strain forces that strain's dominance and allows competitive exclusion (DSS) at a lower γ and a lower *P* (in comparison to where transmission potentials are equal, in Figure 3.8c and d), as this strain has the advantage of increased transmission success in addition to the advantage of immune selection acting to suppress competing (antigenically overlapping) strains.

The increased prevalence of the dominant strain also increases single strain dominance (ε, shown in blue gradation in Figure 3.8). This metric measures the

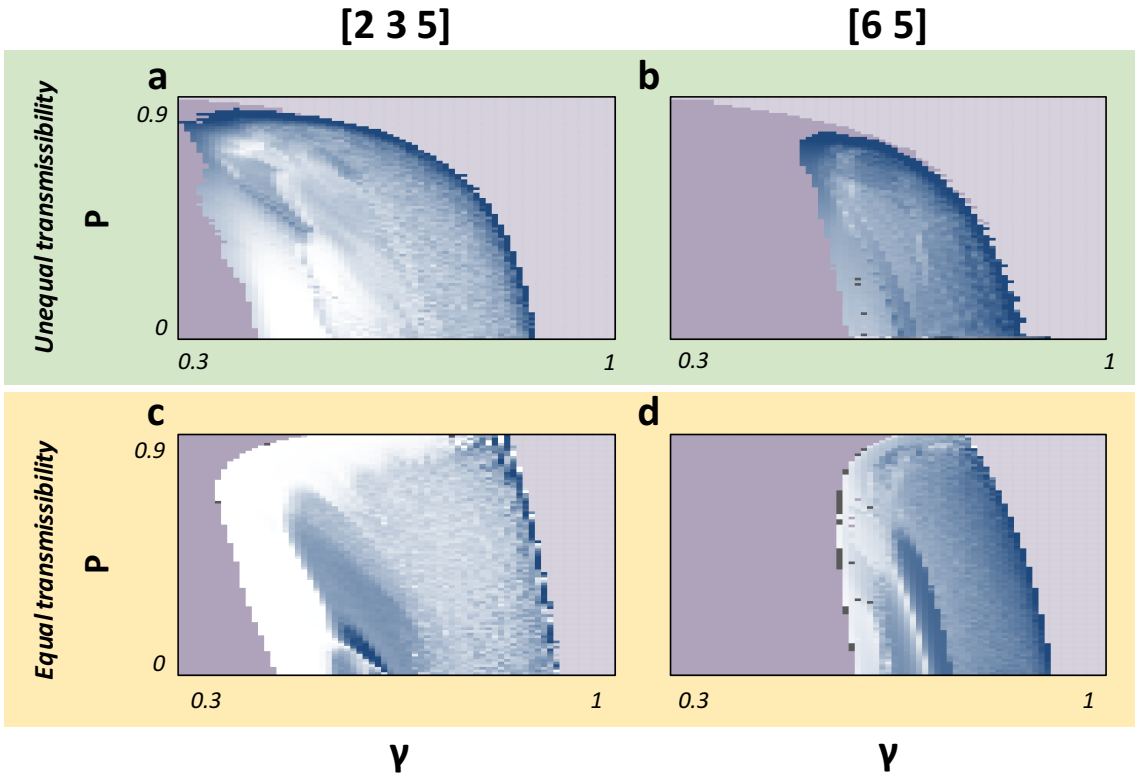


Figure 3.8 The strain structures of the two host model of either a [2 3 5] or a [6 5] strain structure, with and without equality in transmission potentials of strains. 1/μ = 50 years, 1/σ = 36.5 days, and β altered to provide a change in R_0. The infected population is initially at 0.001, with one strain slightly higher at 0.00101. **a)** and **b)** All strains have R_0 = 5, apart from one strain with an R_0 = 7. In comparison with the same system but with equal transmissibility (the same figure as in Figure 3.5c and d) in **c)** and **d)**, differences in transmission potential widens the range of DSS, especially at high *P*, and increases single strain dominance (blue).

difference in prevalence of dominant and subdominant strains and averaged over the number of epidemics; it is high with great amplitudes in the epidemics of the dominant strain, or where the dominant strain is of much higher prevalence most of the time. The latter is true with heterogeneous transmission – the more transmissible strain has a higher prevalence throughout the time series.

3.3.4 The effect of immunity to conserved determinants on the oscillations within CSS

Within the region of cyclical dynamics, there are differences in the pathogen population dynamics. The models are explored further across this CSS range to

investigate the effects of conserved immunity and host heterogeneity in more detail.

Bifurcation diagrams of both the one and two host models (Figure 3.9) demonstrate how the threshold for CSS shifts, and the size of oscillations change, for an increase in ψ (one host model) or proportion of genotype B (two host model). Overall, the proportion of strain-specific immune individuals decreases as ψ (or *P*) increases, as a higher proportion of individuals enter the conserved immunity class '*X'* and therefore cannot gain strain-specific immunity.

The number of circulating strains, where a strain is defined as circulating if it either has an absolute prevalence $>10^{-4}$ or a relative prevalence >1%, changes across the range of cross-protection, as the population shifts from NSS (all strains circulating, in this case 30 strains) to DSS (only discordant sets of strains circulate, so nine strains), which can be seen in Figure 3.10. Across the CSS range, the number of circulating strains gradually reduces. This is due to an increase in immune selection narrowing diversity and suppressing strains which share alleles with the dominant strain. As *P* increases, there is a wider range in the number of circulating strains in this CSS region, and this pattern is similar when circulation is defined by either a strains' relative or absolute prevalence.

The results of the two host model is mirrored in the one host model for an increase in the probability of conserved immunity (Figure 3.11), albeit with a much lower variation in the number of circulating strains at Ψ = 0.8. At this high probability of conserved immunity there is far less CSS observed (Figure 3.5a), with the majority of the range of cross-protection either exhibiting NSS or DSS behaviour (of no variance in circulating strains).

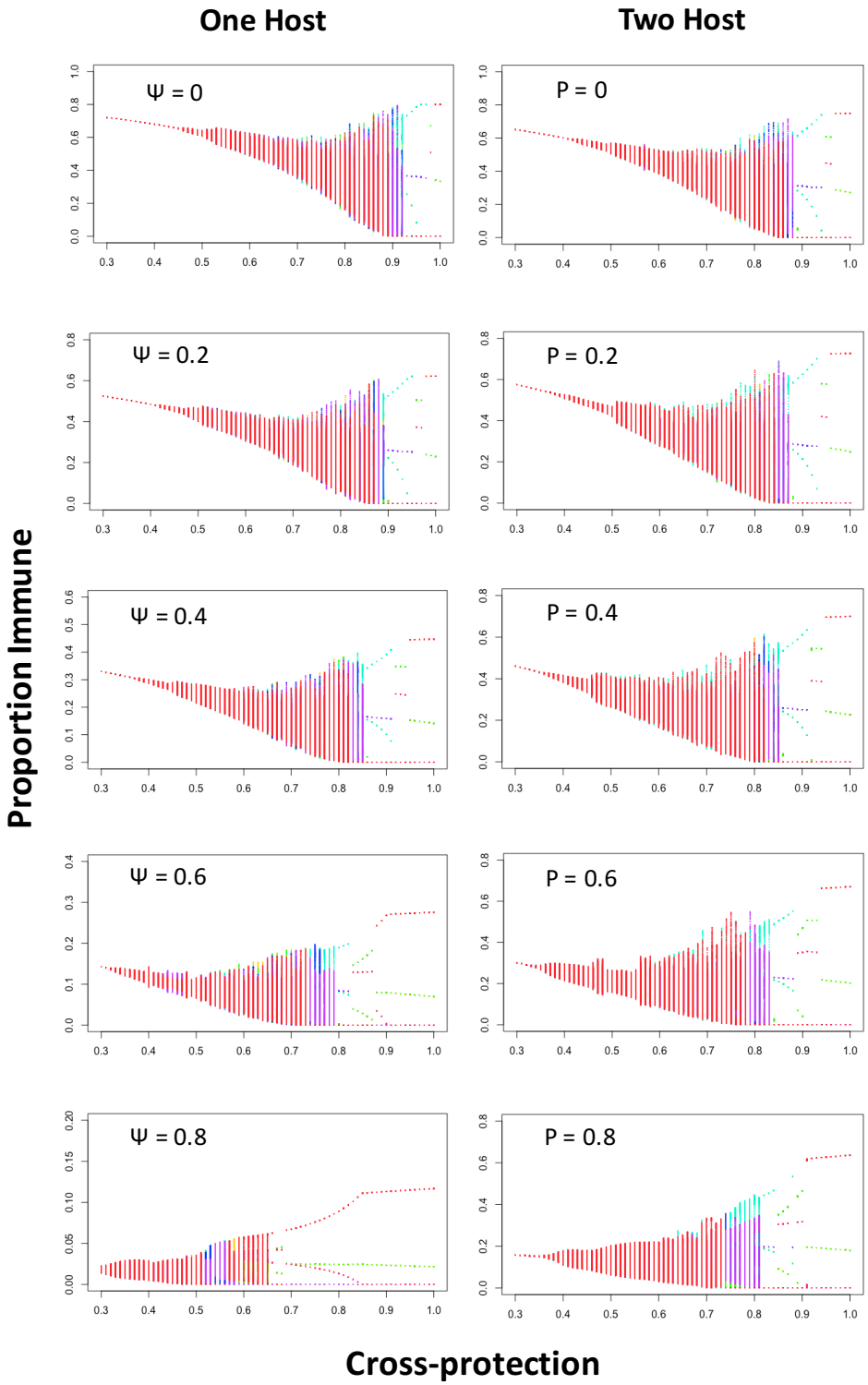


Figure 3.9 Bifurcation diagrams for the immune population in a [2 3 5] antigenic structure where R_0 = 4, for the whole range of γ for different values of Ψ for the one host model (left) and proportion of genotype B for the two host model (right). The region of CSS is clear in the mid-range of γ, and this region shifts as Ψ or P increases. Each colour is a different strain in the system, which overlap in the plot as sets of discordant strains have the same dynamics. 1/μ = 50 years, 1/σ = 36.5 days, and β = 40. The infected population is initially at 0.001, with one strain slightly higher at 0.00101.

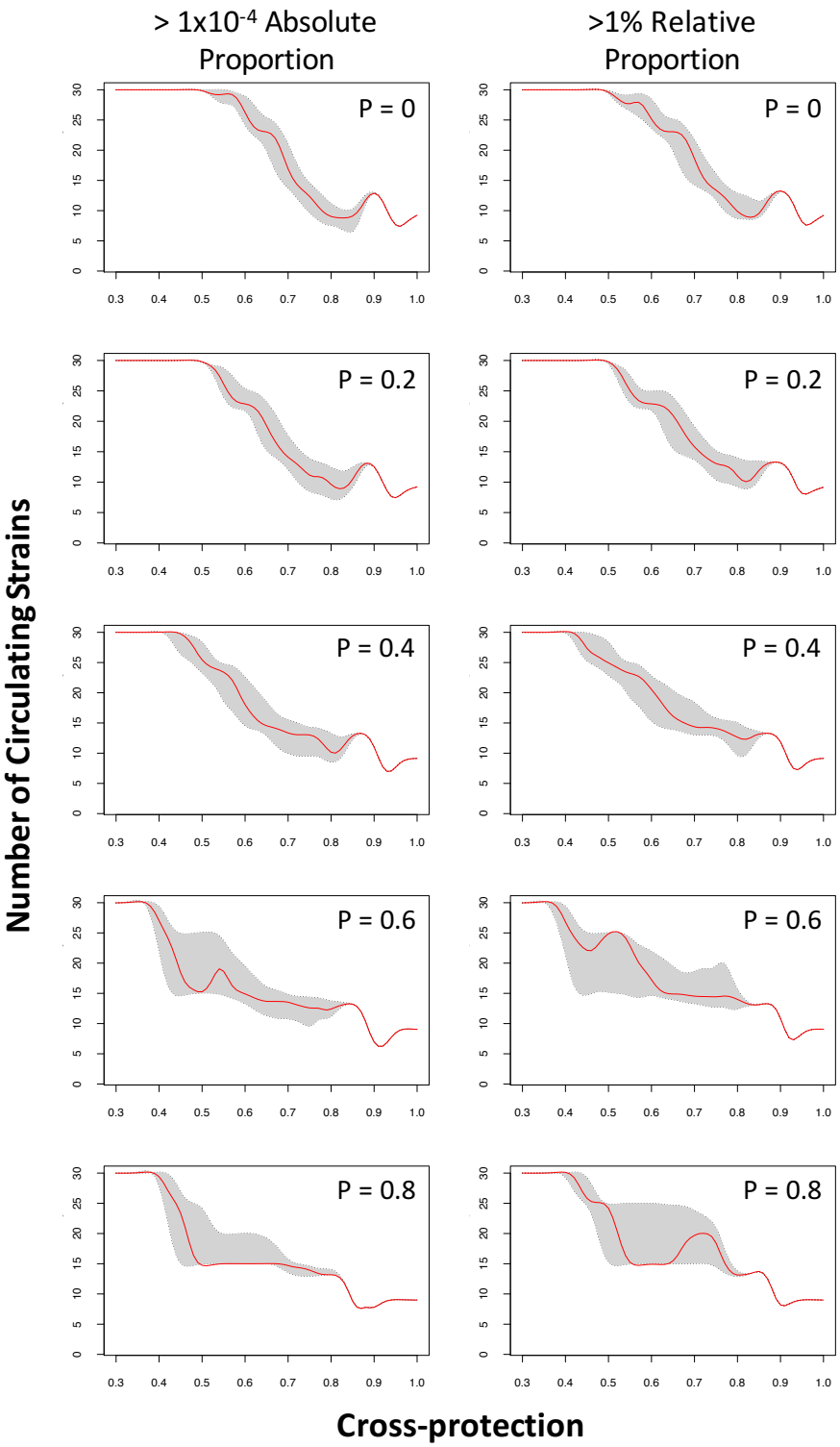


Figure 3.10 The number of circulating strains across γ for different proportions of genotype B (P) for the two host model, with a [2 3 5] antigenic structure, where R_0 = 4. A circulating strain is defined as having a proportion infected > 10^{-4} (left) or relative proportion >1% (right), and the median (solid red line) and inter-quartile range (shaded grey area) of the circulating strains are plotted here, with a spline smooth. Where 30 strains are circulating, the system is in NSS, and in DSS where 9 strains are circulating. Across the range of CSS, the number of circulating strains decreases. As the proportion of genotype B increases and more individuals have conserved immunity, the range of circulating strains widens over the CSS region. **1**/μ = 50 years, 1/σ = 36.5 days, and β = 40. The infected population is initially at 0.001, with one strain slightly higher at 0.00101.

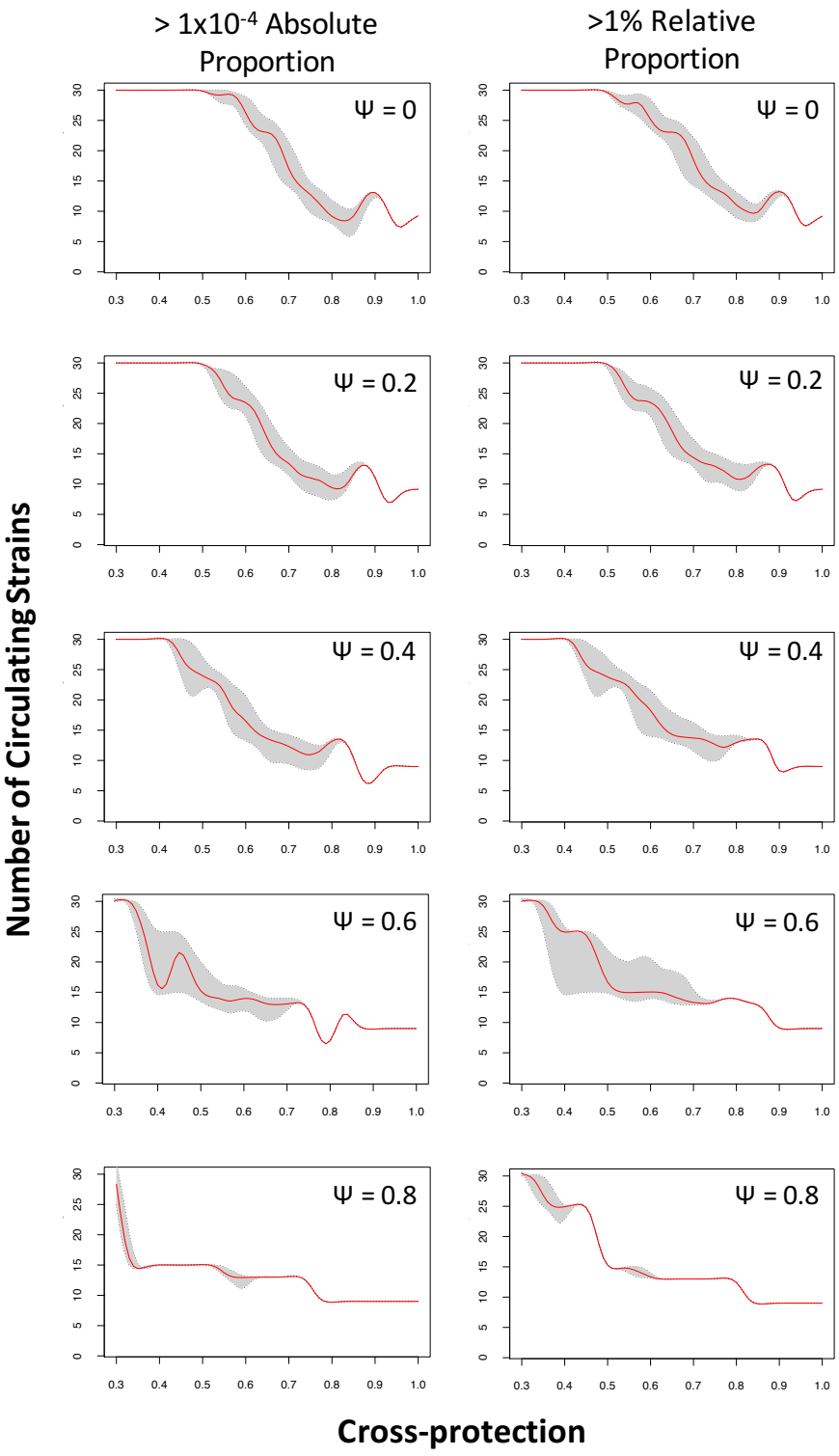


Figure 3.11 The number of circulating strains across γ for different probabilities of conserved immunity (Ψ) for the one host model, with a [2 3 5] antigenic structure, where R_0 = 4. These results mirror those of the two host model in Figure 3.10, with the variation in the number of circulating strains widening as conserved immunity increases. Where Ψ = 0.8, there is considerably less variation. As can be seen from the regimes in Figure 3.5a, CSS has a much smaller range at this degree of conserved immunity, and SSD is very low. 1/μ = 50 years, 1/σ = 36.5 days, and β = 40. The infected population is initially at 0.001, with one strain slightly higher at 0.00101.

When investigating the persistence of strains over time, by taking the time periods in which a strain is present (defined here as absolute prevalence $>10^{-4}$), an increase in *P* increases strain persistence (Figure 3.12). As more of the population acquires immunity to conserved antigens, the reduction in strain-specific immunity reduces selection pressure, allowing strains to circulate in the population for a longer time.

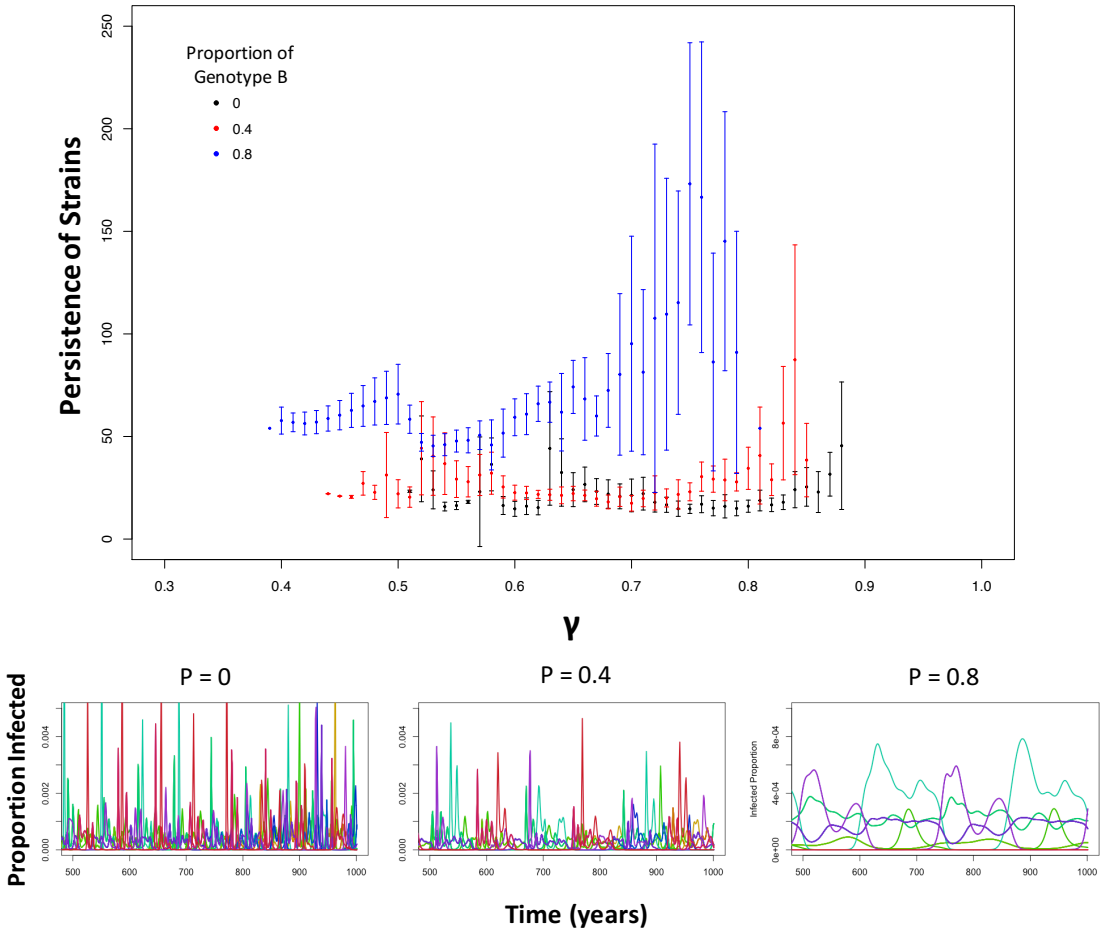


Figure 3.12 The persistence of strains across the range of γ for different proportions of genotype B, for a [2 3 5], where 1/μ = 50 years, 1/σ = 36.5 days, and β = 40 (R_0 = 4). The infected population is initially at 0.001, with one strain slightly higher at 0.00101. Top) As the proportion of hosts with conserved immunity increases, the persistence (period in which a strain's prevalence is $>10^{-4}$) increases, and the variation in persistence increases as well. Persistence is not measured for time series in NSS and DSS as there is no variation in the proportion infected. **Bottom)** The time series for the different proportion of genotype B (where γ = 0.75) shows how the cyclic dynamics changes as there is more conserved immunity in the host population. Other than the overall prevalence being lower (note: change in y-axis scale where *P* = 0.8), the amount of time a strain is dominant or at least present until being replaced by another strain is longer as P increases.

3.4 DISCUSSION

The effects of cross-reactivity and longevity of immune responses on pathogen population structure has been explored in various ways, however differences in response specificity was yet to be studied. Although previously published within-host models have included transient concurrent specific and conserved responses (Minayev & Ferguson 2009; Recker et al. 2004; Nowak & May 1991; Antia et al. 1996), it was unknown what effect life-long responses to either specific or conserved antigens may have on population dynamics. Furthermore, heterogeneity in host responses, while analysed by 'hard-wired' differences in immunity (Gupta & Galvani 1999; Penman et al. 2013), had not been investigated for a range of responses to conserved antigens. This research sought to answer these pending questions, using models with and without host heterogeneity, and including both cross-protection between strain-specific antigens and a probability of acquiring immunity to conserved antigens.

Understanding the relationship between antigenic diversity, population structure, and host immune responses is key to disease control and vaccine design. Here it has been demonstrated that both variable and conserved antigens influence pathogen structure, and heterogeneity in host immunity can alter the appearance of pathogen population dynamics by changing stability, periodicity, and persistence time. Nevertheless, the multi-locus system used here is robust to different routes of acquiring immunity: the typical behaviours, as observed in previous multi-locus instances of the model, still appear, even with only a small degree of strain-specific immunity. The often-observed pattern of DSS in previously published multi-strain models is still observed in the model presented

here, suggesting that DSS is robust to differing acquisitions of immunity, as long as there is a small probability of developing protective immune responses against variable antigenic determinants.

Although the typical behaviours usually observed for multi-locus models are also observed here, the range of cross-protection for which these are seen shifts as the probability of conserved immunity increases. There is a threshold for conserved immunity, not just cross-protection, whereby the pathogen population will not oscillate and instead reach the stable DSS equilibrium; DSS becomes more likely as conserved immunity increases. Unlike cross-protection, conserved immunity does not impose frequency-dependent selection on the strains, so when this type of immunity is high, immune selection is less frequency-dependent and subsequently strains are able to persist.

A similar pattern of shifting strain dynamics was observed in (Recker et al. 2007) with an additional level of protection for individuals who had accumulated immunity to all antigens of a strain, without having been exposed to that strain itself. It could be inferred that this acts very similarly to the model presented here as both increase immunity breadth without the need for exposure.

The new results of the two host model are similar to the one host model, which is to be expected as both models change the degree of conserved immunity in the host population, although in different ways: an increase in the probability of conserved immunity in all individuals is similar to increasing the proportion of hosts which always respond to conserved determinants only.

The results of the two host model mirrors the multi-locus host heterogeneity model of (Gupta & Galvani 1999), which established that an increase in a host

genotype responding to only conserved determinants could still maintain DSS but lowered the cross-protection threshold. This result was also seen when increasing the polyclonality (or breadth) of an individual's response to allow recognition of all possible antigens of a strain (Cobey & Pascual 2011). Conserved immunity can be seen acting to stabilise the pathogen population structure and decrease diversity.

This new model of host heterogeneity in immune responses was explored further, to investigate the effects of the transmission rates on population dynamics. Changing the transmission rates of all strains shifts the thresholds for which different strain dynamics are observed, which is to be expected as the cross-protection thresholds for NSS and DSS are known to be highly dependent on the transmission of the dominant strains (Gupta et al. 1998). Where there is heterogeneity in the transmission of strains, so that one strain has a transmission advantage, DSS is favoured as this strain has both an advantage of immune selection and of transmission. This widens the parameter space for which competitive exclusion, via DSS, occurs. Intrinsic transmission heterogeneities between strains therefore reduces the diversity and increases stability of the pathogen population further than with conserved immunity alone.

Within the region of parameter space where these models exhibit cyclical dynamics, the effect of increasing the proportion of host genotypes with conserved immunity is to i) create a higher variation in the number of co-circulating strains, in general leading to more co-existence and lower single strain dominance, and ii) allow strains to persist for a longer period.

Arinaminpathy et al also investigated the effects of a conserved response to a multi-strain pathogen, using influenza vaccinations as a case study (Arinaminpathy et al. 2012). This used an epochal evolution model to allow the molecular evolution of novel antigenic “clusters” of influenza HA by episodic selective sweeps, developed in (Koelle et al. 2010), to represent antigenic drift (that is, the antigenic changes to HA through time, in a clustered and linear manner (Smith et al. 2004)). The authors found the same results as the model presented here, but the crucial difference is that the explicit forcing of antigenic drift within the model here is not required. Both models found that increasing the number of individuals with a broad immune response acts to reduce the effects of strain-specific immunity, which lowers the advantage of being immunologically different to the dominant strain, making co-existence possible and persistence more likely, as well as lengthening the residence time of strains (or “clusters” of similar HA). As strains take a longer time to switch dominance, it is tempting to speculate that pathogens (or vaccines) eliciting a conserved immune response to protect against all strains may evince a slower rate of antigenic change.

There are vaccines in development for multi-strain pathogens where conserved determinants are being used, for instance the RTS,S vaccine for *P. falciparum* which uses the conserved region of PfCSP for a broadly protecting vaccine, which has reached Phase 3 clinical trials (White et al. 2015), as well as a “universal vaccine” being developed for influenza A which exploits conserved epitopes of various influenza antigens (Antrobus et al. 2014). The work presented here suggests that the development and implementation of universal vaccines may have different effects on pathogens depending on the behavioural states they naturally display. For pathogens within DSS, such as *P. falciparum,* there

would be no change to the regime under a universal vaccine, other than a desired reduction in prevalence. Pathogens exhibiting CSS, such as influenza A, would have a slower turnover of strains and perhaps higher diversity, with a high potential for a switch to a DSS state. On the other hand, pathogens with strains which circulate independently in NSS could start displaying cyclical dynamics, even if weak. Finally, these predictions mandate that pathogen dynamics should be investigated and monitored pre- and post-vaccine implementation, if we are to take advantage of vaccine-mediated shifts in behaviour.

Changes to the specificity of a host response may also be possible after an invasion of a cross-reactive pathogen species. For instance, the introduction of the Zika virus (ZIKV) into areas with a high prevalence of Dengue virus (DENV) has been found to trigger a response to the EDE1 epitope of the E surface protein of DENV, which is highly cross-reactive and protective against both DENV and ZIKV (Barba-Spaeth et al. 2016; Dejnirattisai et al. 2016). Individuals responding to EDE1 present a conserved immunity-like response, so in areas where both pathogens circulate, the selective force emerging from these individuals could alter the dynamics of DENV.

Indeed, any mixing of populations can create these changes to the pathogen population dynamics. Age is a factor in the breadth of immunity due to a history of previous exposures, such as in *P. falciparum* where antibody titres (Dodoo et al. 2001; Ofori et al. 2002) and breadth (Bull et al. 1998; Ofori et al. 2002) increase with age. Migration could also act to alter the balance in the host population as some introduced individuals may have a broader response due to either a different acquisition of immunity or previous exposures to similar or even distant pathogens.

For future work, it would be beneficial to not only look at changes to the pathogen population dynamics, but also to changes in the virulence of the pathogen. Virulence has not been included here, but could be added by including death induced by infection, in addition to background mortality. It has previously been suggested that imperfect vaccines (of efficacy <100%), when the coverage is below the eradiation threshold, can select for increased virulence of the pathogen if those vaccines are anti-growth (and subsequently anti-transmission), whereas anti-infection vaccines will lower virulence (Gandon et al. 2001; Gandon et al. 2003). The conserved immunity included here could act as an anti-growth vaccine as it prevents transmission rather than infection, so how might the evolutionary dynamics of virulence alter for an imperfect vaccine within this multi-strain system?

This chapter presented a novel model which included variation in the specificity of the host immune response. The population dynamics for multi-strain models of strain-specific responses were robust to the inclusion of life-long conserved immunity, suggesting that only a small degree of strain-specific immunity is required for inter-strain competition and competitive exclusion. Conserved immunity was found to increase stability and decrease strain diversity, while during oscillatory dynamics it slowed the turnover of dominant strains. The implications of the changes in pathogen population dynamics could be paralleled to the effects of heterogeneities in the immune response of the host population, such as after the introduction of a universal vaccine.

4 Determining Rotavirus-Like Behaviours with a Random Forest Approach

4.1 Introduction

The stability of the multi-locus model has previously been examined in several ways. The boundaries of the typical population dynamics observed in the multi-strain model (NSS, DSS, and CSS) have been found numerically both in Chapter 3 and in published studies (Recker & Gupta 2005). The steady states of the system have also been derived analytically, for systems of only three or four strains (Lin et al. 1999; Andreasen et al. 1997; Dawes & Gog 2002), or recently for a system of multiple strains (Blyuss 2014; Blyuss 2013; Cherif 2015). Here, instead of finding the boundaries of the different steady states, I use defining features of a particular type of CSS behaviour, in the context of rotavirus population dynamics. Using a Random Forest decision tree model, the parameters of most importance for this cyclical behaviour to emerge are elucidated. The framework developed here could be applied to other infectious disease systems which may present different dynamic behaviour.

The Random Forest algorithm (RFA) has been used on large datasets with many hundreds or thousands of variables, in a wide variety of fields from ecology to genetics. For instance, the technique has been used to find genes which may predict particular diseases (Chen & Ishwaran 2012), protein-protein interactions (Lin et al. 2004), predicting glycosylation sites (Hamby & Hirst 2008), variables leading to forest fires (Oliveira et al. 2012), the effect of climate change on

extinction (Pearson et al. 2014), and the selective pressures shaping the population structures of *S. pneumoniae* (Lourenço et al. 2017).

Pattern-oriented modelling (POM) is a term used for selecting the best model types and/or variables to include in model frameworks to better represent observed patterns, and therefore have predictive and explanatory power on the biological system (Grimm 1994). This style of modelling can use community-level patterns (such as cycling and co-existence) and/or individual-level characteristics (such as transmission rates) (Grimm & Railsback 2012). Reverse POM, which is used in this research, finds individual-level characteristics (model parameters) directly from higher level patterns, such as cycling, to inform on the possible biological drivers of the observed population dynamics.

The benefit of using reverse POM rather than fitting to a data set is that qualitative patterns can be utilised, and that model overfitting is avoided. This is particularly important for systems which do not have sufficient data available, being, for instance, spatially and/or temporally limited. Inferences based solely on fitting a model to scarce data can be specific to that data set, and can potentially mislead our understanding of the general behaviour of the host-pathogen system.

Here I apply reverse POM using an RFA model to the population dynamics of rotavirus, for which limited data on the prevalence of genotyped strains is available. Using the MANTIS model, described in Chapter 1, to generate thousands of time series while changing both cross-protection and the transmission phenotypes of strains, definitions of strain structure and rotavirus-like dynamics are used to classify whether the time series outputs qualitatively reproduce rotavirus observations. The binary classification (i.e. rotavirus-like or

not rotavirus-like) and the variables used for each simulation are inputs for the RFA, rather than utilising the time series data itself. This novel approach is used to explore a vast parameter space, detecting which parameters and characteristics of the pathogen are important for the observed cyclical and co-circulating patterns, to understand in a qualitative way the main factors influencing rotavirus population dynamics.

4.1.1 The biology of rotavirus

Rotavirus (RV) is the causing agent of gastrointestinal disease, and the most common cause of severe diarrhoea. Globally, the most recent estimate for RV mortality in children <5 years of age is 215,000 deaths in 2013 (Tate et al. 2016), and since 2009 two vaccines have been recommended for use in national immunisation programmes by the World Health Organisation (WHO).

RV is in the RNA virus family *Reoviridae,* and the main type infecting humans is RV type A. The viral genome has 11 segments of dsRNA, and being segmental provides a great ability to reassort during co-infection (McDonald et al. 2016). The genome contains six structural proteins (viral proteins, VP) and six non-structural proteins (NSP), contained in a non-enveloped protein capsule of three layers (Greenberg & Estes 2009). The outer layer consists of the two VPs of importance for neutralising immunity (Offit & Blavat 1986): VP7 (G protein), a glycoprotein used for virion assembly which forms the capsid shell, and VP4 (P protein), a protease-sensitive protein for host cell attachment and entry which forms spikes on the capsid. On cell attachment, VP4 is cleaved into VP5*, for translocation of viral particle into the host cell, and VP8*, for membrane attachment. The middle shell consists of VP6, and the inner layer of VP2.

Infection of an individual is tested by using an enzyme-linked immunosorbent assay (ELISA) with the sera against VP7 for serotyping, and VP6 for subgrouping (the latter, being on the inner shell rather than exposed, has a lower diversity and only differs between genogroups). In addition to ELISA, serotyping can also be done by using a plaque reduction neutralisation assay (Offit & Blavat 1986), where serum is mixed with the virus and spread over a layer of host cells in culture to form plaques. The amount of serum required to reduce the number of plaques by 50% provides a measure of antibody effectivity.

Serotypes identified from neutralisation techniques are concordant with the genotypes for the G protein, however there are mismatches of serotype and genotype for the P protein (Matthijnssens et al. 2008). Genotyping of all 11 segments of the genome via real-time polymerase chain reaction (RT-PCR), to provide a “genotype constellation” (Matthijnssens et al. 2011; Matthijnssens & Van Ranst 2012), is used more commonly now as it is more accurate, the technology has improved, and unusual strains can be typed more easily. At present, there have been 27 G types and 35 P types identified (Matthijnssens et al. 2011).

It has been observed that after two RV infections, individuals only develop mild or asymptomatic disease, independently of whether those first RV infections were symptomatic or asymptomatic (Velázquez 2009; Velázquez et al. 1996). There is evidence of cross-reactivity between RV strains after the initial infection: the first infection generates a homotypic IgA response, but subsequent infections are heterotypic (Velázquez et al. 2000; Velázquez 2009; Velázquez et al. 1996; Chiba et al. 1986; O’Ryan et al. 1994), and consequently individuals are highly unlikely to be reinfected with the same G genotype (Velázquez et al. 1996).

Typically, the main G and P types that circulate are G1P[8], G3P[8], G4P[8], and G2P[4] (depicted in Figure 4.1). G1P[8] is more common in developed countries, such as the United Kingdom (UK), with G2P[4] occasionally dominating. Strains of P[4] are of the genogroup DS-1, and P[8] of genogroup Wa-like.

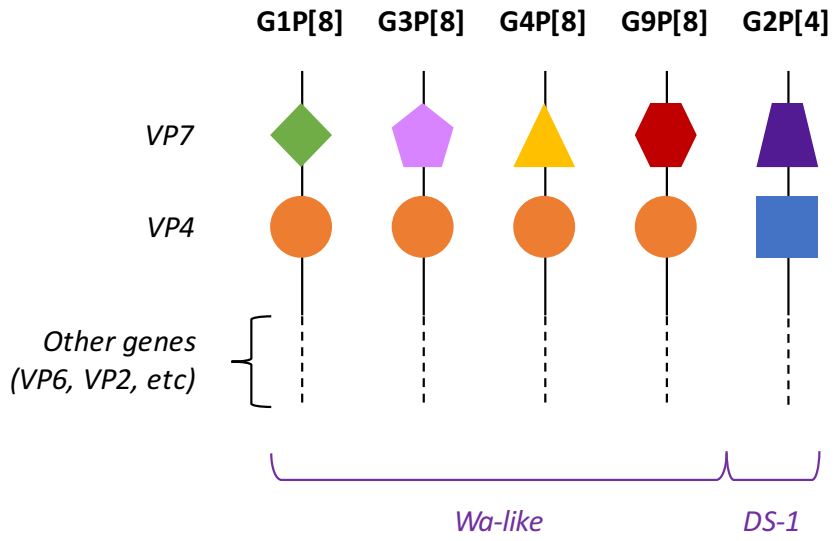


Figure 4.1 A schematic diagram of the most common rotavirus strains.

The strain G9P[8] has emerged recently, after first appearing as G9P[6], and it is believed to have reassorted with common P[8] strains to circulate as G9P[8]. In the UK, as G9P[8] spread, G9P[6] became less frequent – it is thought that G9P[8] may have a replicative advantage to allow this replacement (Iturriza-Gómara et al. 2000). Additionally, G9 strains are more common in older age groups, and are found to be increasing in incidence, suggesting that there is low cross-protection from G1-4 infections (Iturriza-Gómara et al. 2000).

The two vaccines in use are: Rotarix, or 'RV1', a monovalent vaccine from attenuated human G1P[8] strain; and RotaTeq, or 'RV5', a pentavalent vaccine from human-bovine reassortments of G types 1-4 and P[8]. Despite differences in valency, both have some cross-reactivity, having been shown to elicit

protection against G types 1-4 and emerging G9 (in European infants) (Vesikari et al. 2006; Vesikari et al. 2007). Both of these vaccines target the main circulating strains in developed countries, however they do not match dominance elsewhere, since unusual combinations are more common in Africa, Asia, and South America, perhaps due to more human-animal reassortments (Santos & Hoshino 2005). The increased diversity in developing countries may explain why the vaccines are less effective in these areas (Leshem et al. 2014). For instance, IgA neutralisation of G1 and P[8] after the pentavalent vaccine is reported to be four-fold lower in Africa when compared to Europe or the US (Armah et al. 2012).

RV, like norovirus (Huang et al. 2003), has an interesting interaction with host genetics via blood types. When the viral particle attaches to the host cell membrane via VP7, some strains recognise and bind the hosts' sialic acid. However, the most common strains (of Wa-like and DS-1 genogroups) are sialidase-insensitive. Instead, the VP8* of the P protein in these genogroups binds to human blood group antigens (HBGAs), in both *in vitro* cell cultures and *in vivo* stool samples (Huang et al. 2012). The P protein specifically interacts with these HBGAs: P[6] only binds to H type I (Lewis antigen negative individuals), whereas P[4] and P[8] can bind to both H type I and Lewis B antigens (Huang et al. 2012), a result also observed in (Nordgren et al. 2014) with the addition of P[4] and P[8] also binding Lewis A antigen. Having H type I is insufficient for a symptomatic P[4] or P[8] rotavirus infection: an additional fucose is required, to form a Lewis antigen, to be susceptible. P[6] strains are more common in Africa, where unusual strains and recombinants are also observed – the high proportion of P[6] may be due to the high proportion of Lewis negative individuals in African populations (Nordgren et al. 2014).

4.1.2 THE POPULATION DYNAMICS OF ROTAVIRUS

RV genotypes vary temporally in the same region, and between regions at the same time. Frequently, there is also a co-circulation of multiple genotypes at the same time and location. The most common strains (G1-4 and G9 with P[8], and G2P[4]) account for 91.2% of the global prevalence (in 1989-2004) (Santos & Hoshino 2005). G1P[8] is almost always dominant by a wide margin, but other strains circulate at a low level and are cyclic in their dynamics, and G2P[4] gains dominance approximately every seven years (in the US) (Dennis et al. 2014).

Not only does the switching in dominance of G1P[8] and G2P[4] occur between non-overlapping G and P types, but between two different genogroups (Wa-like and DS-1, respectively, due to their different genotype constellations (Matthijnssens & Van Ranst 2012)). G2 strains sometimes do not serotype in ELISA, which is thought to be due to antigenic drift (Gómara et al. 2001b). Analysing the antigenic regions A-F in UK G2 isolates from 1995-6, the arginine to asparagine mutation at position 96 in antigenic region A is responsible for serotyping failure (Gómara et al. 2001b). In South African strains 1984-98, these same mutations in undetectable G2 strains were found (Page & Steele 2004). The re-emergence of G2P[4], as it replaces the dominant G1P[8], may be a result of the altered antigenicity in this antigenic region.

G1P[8] is observed to have strains of the same lineage cause outbreaks at different global locations concurrently; there is little spatial or temporal heterogeneity (Arista et al. 2006). Unlike other RV strains, which disappear from the population after outbreaks, G1P[8] can persist with no allelic change for 1-3 years (McDonald et al. 2012). This persistence has been hypothesised to be due

to the fact that G1 circulates as a heterogeneous population, with high variation in the antigenic regions of VP7, which is not observed for other RV strains (Arista et al. 2006). This antigenic variation may contribute to persistence, as some G1 variants may circulate and others disappear – for instance, over a twenty year period in Italy, it was observed that, while G1 persisted, the peak in G1 infections of infants coincided with the introduction of novel G1 variants (Arista et al. 2006).

G1 has also been observed to have a higher rate of antigenic and genetic change than other RV strains, at $2x10^{-3}$ substitutions/site/year (Arista et al. 2006), like influenza A, with $3.7x10^{-3}$ and $1.8x10^{-3}$ fixation rates for H1 and H3, respectively (Ferguson et al. 2003), allowing immune evasion. Other suggestions for its persistence have been a possible replicative advantage (Santos & Hoshino 2005), or a limited possibility of reassortment (a lack of diversity and infection with only single strains reduces the likelihood of reassortment with heterologous strains) (McDonald et al. 2012).

However, unlike what is observed for influenza A, there is far more strain co-circulation, and strains which are removed can return in the same location years later without any antigenic change. This co-circulatory behaviour and higher strain diversity is more akin to avian influenza, and influenza B and C (Yamashita et al. 1988; Nelson & Holmes 2007), which could be in part due to slower evolutionary rates of these influenza types. There has been a suggested evolutionary stasis in avian flu (Webster et al. 1992) due to a long history of infecting aquatic birds; wild aquatic birds have the lowest substitution rate of influenza (Chen & Holmes 2006), but the rate is highest in epidemics in poultry and swine (Nelson & Holmes 2007). Of note is that increased diversity and lower single strain dominance within pathogen populations of multiple strains was

found in Chapter 3 where the probability of immunity towards conserved antigenic determinants was increased: strain diversity and rates of evolution are linked to the acquisition of immunity of the host.

Phylogenetic analysis of the VP7 antigenic regions (G protein) of RV found three lineages: pre-1995, from a DS-1 ancestor; post-1995, with two sub-lineages from common circulating strains and newly emerging isolates; and intermediate, containing strains of both pre- and post-1995. This intermediate lineage suggests that there can be some coexistence of previously and newly dominant strains, and new isolates emerge from a pool of related strains (perhaps from introductions) rather than recycled from a previous epidemic (Page & Steele 2004).

4.1.3 Modelling of rotavirus population dynamics

Rotavirus dynamics have previously been explored with mathematical modelling using age-structured SIR models (Pitzer et al. 2009; Pitzer et al. 2011a; Pitzer et al. 2011b; Atchison et al. 2010; Van Effelterre et al. 2010). Pitzer et al explored the spatiotemporal dynamics of rotavirus to explain the timings of epidemics and local persistence across the US by the heterogeneity in birth rates (Pitzer et al. 2009; Pitzer et al. 2011a), and used this to predict that the vaccines (which have a similar effect to birth rates in altering the proportion of susceptibles) may delay the timings of epidemic onset, subsequently increasing the age of first infection. The introduction of a vaccine was also suggested to increase the periods of oscillations, changing the epidemics from a seasonal to biennial pattern (Pitzer et al. 2011b), as also observed in post-vaccination measles dynamics (Anderson et al. 1984; Grenfell et al. 2002; Finkenstadt & Grenfell 2000). These predicted

post-vaccine effects have been observed with rotavirus surveillance networks, as the average age of first infection is increasing in the US (CDC surveillance) (Payne et al. 2011). However these results were not observed in Portugal (Rodrigues et al. 2013).

The current models fit dynamics to data obtained either from samples or collected data from the literature. However, fitting models to complex dynamics to a relatively small number of observations is difficult, and there is a risk of over-fitting. Here I suggest a different method of analysing the patterns of rotavirus. Using the MANTIS framework to generate time series of a multi-strain pathogen, reverse POM is used to investigate the different combinations of variables which lead to RV-like dynamics.

4.2 METHODS

4.2.1 THE RANDOM FOREST ALGORITHM

To my knowledge, using the RFA on the outcomes of a multi-strain model has never been done before, and it is novel to have a binary RFA classifier (i.e., a classification of whether the time series are RV-like or not) that is defined using the model itself, rather than population data.

The RFA is a “black box” method of machine learning, using classification and/or regression trees (CART), which act to split the data into binary trees such that each split node leads to daughter nodes that are more homogeneous in their output than the parent nodes. Data are defined as having one variable to be predicted (classification) given *N* predictor variables.

The forest (set of CART) is randomised in two ways: i) each tree is grown with a random sample of the data by bootstrapping, and ii) at each node on each tree, a random set of predictor variables are used to split the data. This randomisation acts to disrupt correlations in the data (Chen & Ishwaran 2012), especially as the random set of predictor variables at each node allows weaker variables to be chosen as often as stronger ones, so that the stronger predictor variables are not disproportionately used (Breiman 2001).

The forest, composed of independent trees, is built like so (Chen & Ishwaran 2012; Jones & Linder 2015):

1. For each tree, a random subset (bootstrap) of the original data is used. This is the bootstrap aggregating (bagging) step.

2. As each tree grows, the nodes are binary split using a random subset of predictor variables.

3. The subset of data that was not used to grow each tree (out-of-bag, OOB, data) is dropped down the tree created by the bagged data. This is to make a prediction of the outcome, and the predictions for all trees are combined to create an “ensemble estimate”. As each prediction uses data that were not used to build the tree, Random Forests predict without the risk of overfitting.

4. An OOB error rate and the importance of each variable for the predictions is calculated, using the OOB data for each tree. Using this OOB data at this step avoids the need for test data and cross-validation, unlike other machine learning tools.

The importance of each predictor variable is calculated in Step 4, and this is the degree to which each variable of the data (predictor) influences the outcome of the ensemble prediction: the higher the importance of a variable, the better they are at being used to predict the outcome of the classifier. Being able to find the importance of variables in a quantifiable way, and being able to measure the error rates and performance of the model, is an incredibly useful tool and an advantage of using the RFA.

There are two main ways to measure predictor importance, using Gini importance (mean decrease in impurity, MDI) or permutation importance (mean decrease in accuracy, MDA). The latter is most commonly used and the type of importance used in this research. Permutation importance is the difference in the error rate between the original and permuted OOB data for that variable/predictor.

The importance is calculated by these steps (Chen & Ishwaran 2012): 1) randomly permute the values of each predictor variable of the OOB data, for each tree; 2) drop this permuted OOB data down the tree; 3) find the prediction error; and 4) compare permuted and non-permuted prediction errors. The higher the error from the permuted OOB data compared to the non-permuted OOB data, the more important/informative the predictor variable is.

4.2.2 Multi-locus modelling for the time series output

The **M**ulti-locus **Anti**genic **S**imulator (MANTIS) was used to model the pathogen population dynamics, as described in Chapter 3. A [5 2] antigenic structure (a system of two loci of five and two alleles each) applies to all simulations, chosen

to represent the five main circulating G types (G1-4 and G9) and the two main P types (P[4] and P[8]) of RV.

The transmission potential of each strain is defined as $R_0 = \frac{\beta}{\sigma + \mu}$ where β is the transmission coefficient, σ is the rate of recovery, and μ is the background death rate. The transmission of the strains was varied by changing β, and the other parameters were kept constant at 1/σ = 5 days (duration of infection suggested for RV in (Heymann 2004)) and 1/μ = 50 years. The baseline β was 292 (therefore R_0 = 4). This combination of parameters have been used previously for this multi-strain model and is well understood (Wikramaratna et al. 2013; Wikramaratna et al. 2014), and despite the lack of reliable R_0 estimates from empirical studies, the R_0 chosen here is within the range predicted from a mathematical model of RV dynamics in Europe (Van Effelterre et al. 2010). Strains were defined as either "high" or "low" depending on whether their transmission was higher or lower than the baseline value. Strains could also be removed from the population by having a transmission potential of zero. Other than β, the only parameter that varies across simulations is the degree of cross-protection, γ. As described in the previous chapters, cross-protection is the reduction in transmission for strains which share alleles with those a host has already been exposed to.

The simulations were run for 20 trials of each set of R_0 combinations with cross-protection values taken from a probability distribution (the table of these simulation runs can be seen in Appendix 4.5.3). At the end of each individual simulation, the RFA predictors (see section 4.2.4) are defined, and the MANTIS output dynamics classified into the three classic population dynamics (NSS, DSS, or CSS) as well as the binary RV-like dynamics classification (see section

4.2.3). The table of RFA predictors and the classification of RV-like dynamics for each MANTIS simulation are then used as an input for the RFA.

4.2.3 Definitions of RV-like dynamics

The final 10% of the time series from MANTIS was used to classify the population dynamics. The definitions for both strain structuring and RV-like dynamics use characteristics of the time series, described in Table 4.1. The definitions for RV-like dynamics and strain structuring (Figure 4.2) both follow an if/else method of classification. A simulation was classified as “yes” for its RV-like properties if it was classified as both in CSS and of RV-like behaviour (though of course, these properties are correlated: it could not be classified as RV-like without cyclical properties). For the development of the definition of RV-like dynamics, see Appendix 4.5.1.

Measure	Explanation
Equality *How different are the strains' prevalences?*	Average prevalence is calculated as the mean proportion infected for each strain, ignoring any strain where $R_0 = 0$ (these removed strains are not due to immune selection). *Equality* takes the range of the average prevalences, and strains are deemed equal if this is less than the difference threshold (10^{-6}).
Test_cycle *Are strains cycling?*	The difference between the maximum and minimum prevalence for all individual strains is calculated, and if this difference is above the cycling threshold (10^{-6}), the time series is defined as cycling. This can also test cycling while excluding strains where $R_0 = 0$, as removed strains will not cycle. *test_cycle_dom* is used for the RV definitions, and tests the cycling of the most dominant strain only.
no_domstrains *How many strains become dominant? (i.e. are there changes in dominance across time?)*	The strain of highest prevalence (i.e. dominant strain) at each time point is taken, within the difference threshold to account for multiple strains at that prevalence, i.e. in NSS/DSS. The frequency that each strain dominates is calculated. *no_domstrains* is the number of strains that dominate for more than ten time points over the period (of 200 time points). If *no_domstrains* > 1 then there is a change in dominance over the period.
PrevalentStrains_High/Low and Low_obs *How many strains have a relative prevalence above certain thresholds?*	The relative prevalences of the strains are used instead of absolute values as before. Classify "low" prevalent strains as between 10-15% (*PrevalentStrains_Low*), and "high" prevalent strains as >15%. (*PrevalentStrains_High*). If there are no high prevalent strains, any low prevalent strains above the observable threshold (10^{-4}) are counted to provide *Low_obs*.

Table 4.1 The characteristics of time series data used to classify the strain structure and RV-like dynamics.

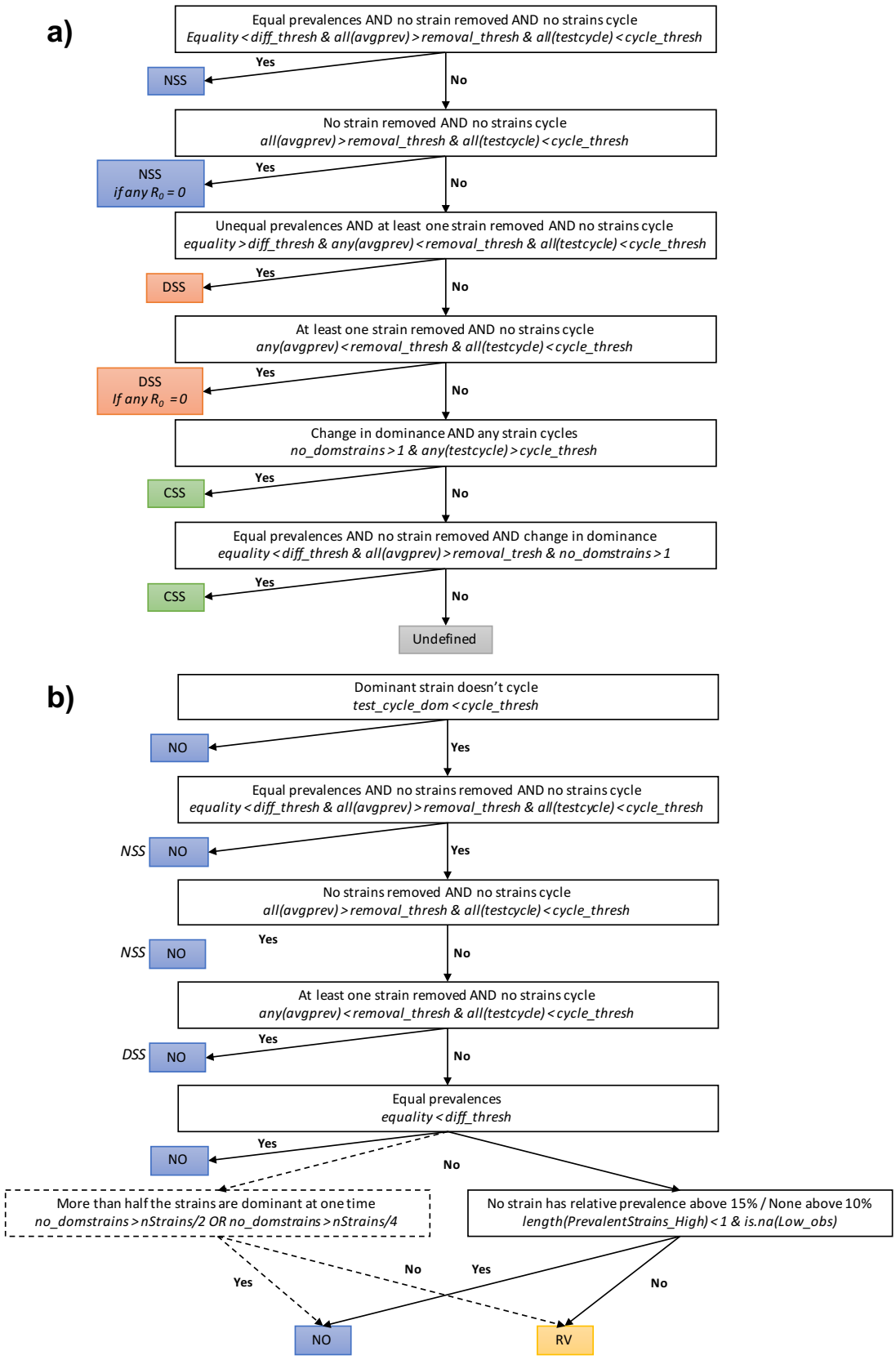


Figure 4.2 a) Diagram of definitions for NSS/DSS/CSS. If a time series does not classify into any of these categories, it is classed as undefined. **b) Definitions used for RV-like dynamics of a time series.** In dashed lines is another option for these dynamics, which was too broad and included almost all the CSS region, and so the alternative definition on the right was used for the rest of the RFA analysis. The threshold for differences (equality and cycling) is 10^{-6}, for removal 10^{-10}, and for observable 10^{-4}.

4.2.4 PREDICTORS OF RV-LIKE DYNAMICS USED IN THE RANDOM FOREST

Each MANTIS simulation was summarised using a variety of characteristics which could then be used as predictors in the RFA, rather than using the time series data itself. Along with the value for cross-protection, these characteristics were used to summarise the intrinsic heterogeneity of the strains (Table 4.2). For an explanation of which model parameters were excluded from the RFA as predictors, see Appendix 0.

Predictor	Explanation
Gamma (γ)	Cross-protection.
Quantile_0 / 25 / 50 / 75 / 100	The 0%, 25%, 50%, 75%, and 100% quantiles of the set of R_0 values of all strains.
Max/MinR_0	The maximum or minimum of the set of R_0 values of all strains.
With_Low/High/ZeroR_0	Number of strains with an R_0 value lower (excluding where R_0 = 0) or higher than the baseline $R_{0,}$ or where R_0 = 0.
Similar_to_Low/High/Zero	Taking the strains classified as either low/high/zero R_0, count how many *other* strains share alleles. This predictor is the degree to which all strains are related to those with low/high/zero R_0 values.
Similar_Within_Low/High/Zero	If there are >1 strains classified as low/high/zero R_0, the number of strains sharing alleles *within* these classifications are counted, taken as a proportion of the total possible shared combinations. This predictor is the degree to which strains of low/high/zero R_0 values are related to each other.

Table 4.2 Predictors used for the RFA

4.2.5 MEASURES OF RANDOM FOREST PERFORMANCE

There are several methods of measuring the performance of the RFA. The three types of performance measures used in this research are as follows:

Prediction scores

Percent correct of RFA predictions for “yes” (true positive classifications of RV-like dynamics) and “no” (true negatives) as a measure of the sensitivity of the RFA. These percentages are taken from the confusion matrix:

	Predicted No *(Not RV-like)*	Predicted Yes *(RV-like)*
Actual No *(Not RV-like)*	**True negative**	**False positive**
Actual Yes *(RV-like)*	**False negative**	**True positive**

Area Under the ROC Curve (AUROC)

The AUC (area under the curve) is a useful single metric to compare between models. In the case of RFA, it is beneficial to use the AUC for a ROC (receiver operating characteristic) curve. The ROC plots the true positive rate (sensitivity) against the false positive rate (specificity) of the RFA, and compares to the null hypothesis of a random predictor. The AUROC is the area under the ROC curve, and is measured between 0 and 1 (with the closest to 1 being the most accurate model, of highest specificity and sensitivity).

Root Mean Square Error (RMSE)

The RMSE calculates the residuals between actual and predicted outcomes of the RFA, which varies between 0 and 1. The lower the value, the lower the error and therefore the higher the accuracy of the model.

4.3 RESULTS

4.3.1 STRAIN HETEROGENEITY VS. CROSS-PROTECTION

The RFA results for simulations which vary in cross-protection, heterogeneity, or both are shown in Figure 4.3, where the importance (mean decrease accuracy, MDA) of the models are shown as boxplots to display the variation of the scores for each predictor amongst all simulations. Clearly for where only cross-protection is varied, this is the only important factor and all other predictors have no effect, as they are constant and subsequently irrelevant (Figure 4.3). However, it is a very similar result to where strain heterogeneity is also included. The importance of cross-protection is expected: it is already known that cross-protection plays a large part in the strain structure and cyclical dynamics (Gupta et al. 1998), and so it is logical that this parameter is important for the appearance of RV-like behaviour, which by definition must cycle.

Looking at the performance of the RFA results using a ROC curve (Figure 4.3b), the models including cross-protection variation have very similar, and high, performances (AUROC of 0.95 and 0.97) in contrast to heterogeneity alone (AUROC of 0.66). In conclusion, including variation in cross-protection improves the RFA model, and consequently the capacity to generate and identify RV-like dynamics.

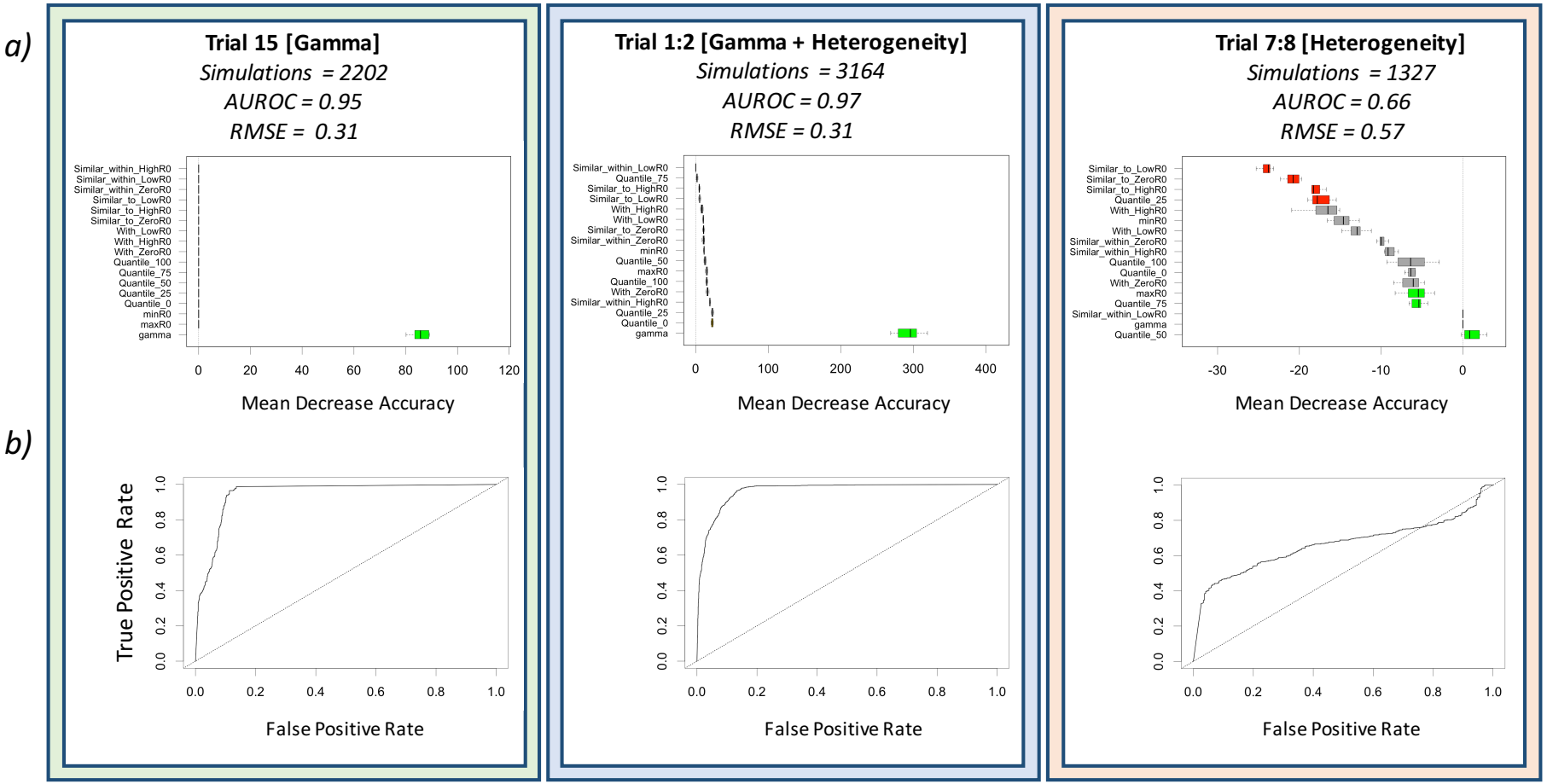


Figure 4.3 RFA results for cases where only γ is varied, only strain transmission is varied, and both. a) The importance of the predictors from the RFA. Expectedly where only γ is varied, this is the one and only predictor and all other variables are at MDA=0 as they are constant. Adding heterogeneity to this provides some variation in MDA, however γ is the only significant predictor for RV-like dynamics. Including only strain heterogeneity reduces the performance of the RFA, as almost all variables have a negative MDA. Green predictors are in the top 2.5% of importance, whereas red are the bottom 2.5%. **b**) The ROC curves for the corresponding RFA models show how changing γ with and without strain heterogeneity has a similar performance, whereas using strain heterogeneity alone has a very poor performance.

Investigating RFA performance further, in Figure 4.4, the number of “yes” and “no” classifications of RV-like dynamics (Figure 4.4a and b), the percentage correct of RFA predictions for these outcomes (i.e. the true positives and true negatives from the confusion matrix) (Figure 4.4c and d), and the AUROC and RMSE results are shown (Figure 4.4e and f). In general, including only heterogeneity greatly reduces the performance of the RFA, by reducing AUROC and increasing RMSE. Notably, there are much fewer “no” simulations in these cases, and this could be expected to bias the RFA classifier, reducing performance. However, I did not find a correlation between AUROC or RMSE and the number of “no” simulations across all models (r^2 = 0.062 and 0.065, respectively).

Combining all trials of the simulations into one large RFA model, it can again clearly be seen how cross-protection is the one and only predictor of RV-like dynamics (Figure 4.5), and this model has a high predictive performance with an AUROC of 0.83 and RMSE 0.51.

The importance of cross-protection is clear, but from mean decrease accuracy plots alone it is unclear what role cross-protection plays in the RFA. For instance, from Figure 4.4a there are differences in the number of “yes” classifications, with trials of only variation in cross-protection having a smaller number of RV-like simulations. Expanding on the result that cross-protection is important for the RFA, it is of interest to find the range of γ which determines RV-like dynamics.

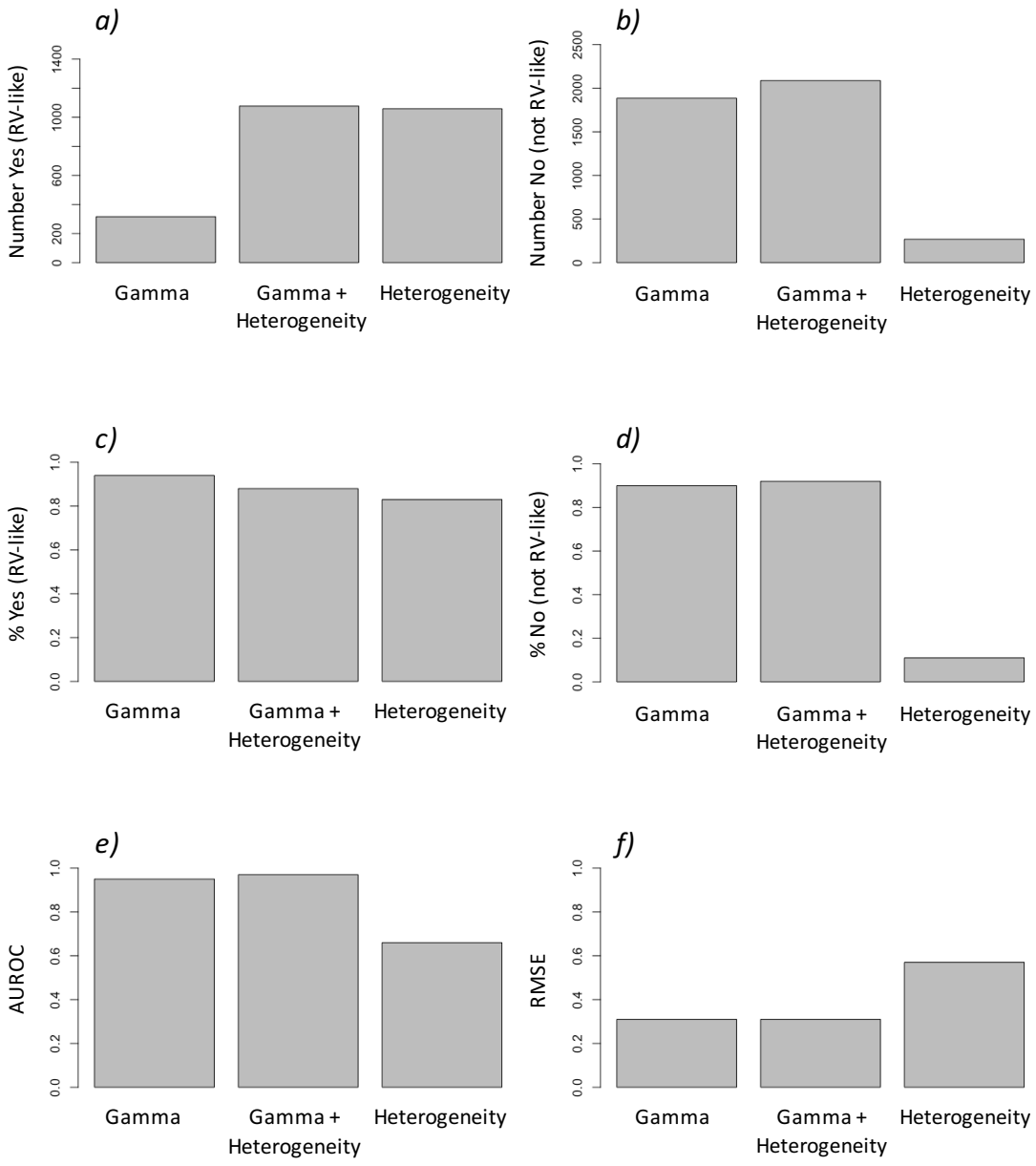


Figure 4.4 Performance measures for trials with γ, strain heterogeneity, or both. a) and b) plots the number of "yes" or "no" classifications from the simulations. Where only γ is varied, there are fewer simulations classified as "yes" – from Figure 4.3a it is known that this is the most important predictor, and critical in determining cyclical dynamics. Cycling only occurs over a relatively small region of cross-protection, which explains the low number of "yes". Where only heterogeneity is included, there are fewer "no" simulations, in part due to the constant γ chosen for these simulations being of the most appropriate value for acceptance. **c) and d)** show the percentage of correct predictions from the RFA for "yes" and "no". All scenarios have a high "yes" prediction, but the heterogeneity only case has a poor prediction rate of "no", due to the low number of simulations of that classification for the RFA to use as a reference for predictions. **e) and f)** show performance measures AUROC and RMSE, which show performances drop across the three scenarios (AUROC lowers, and RMSE increases).

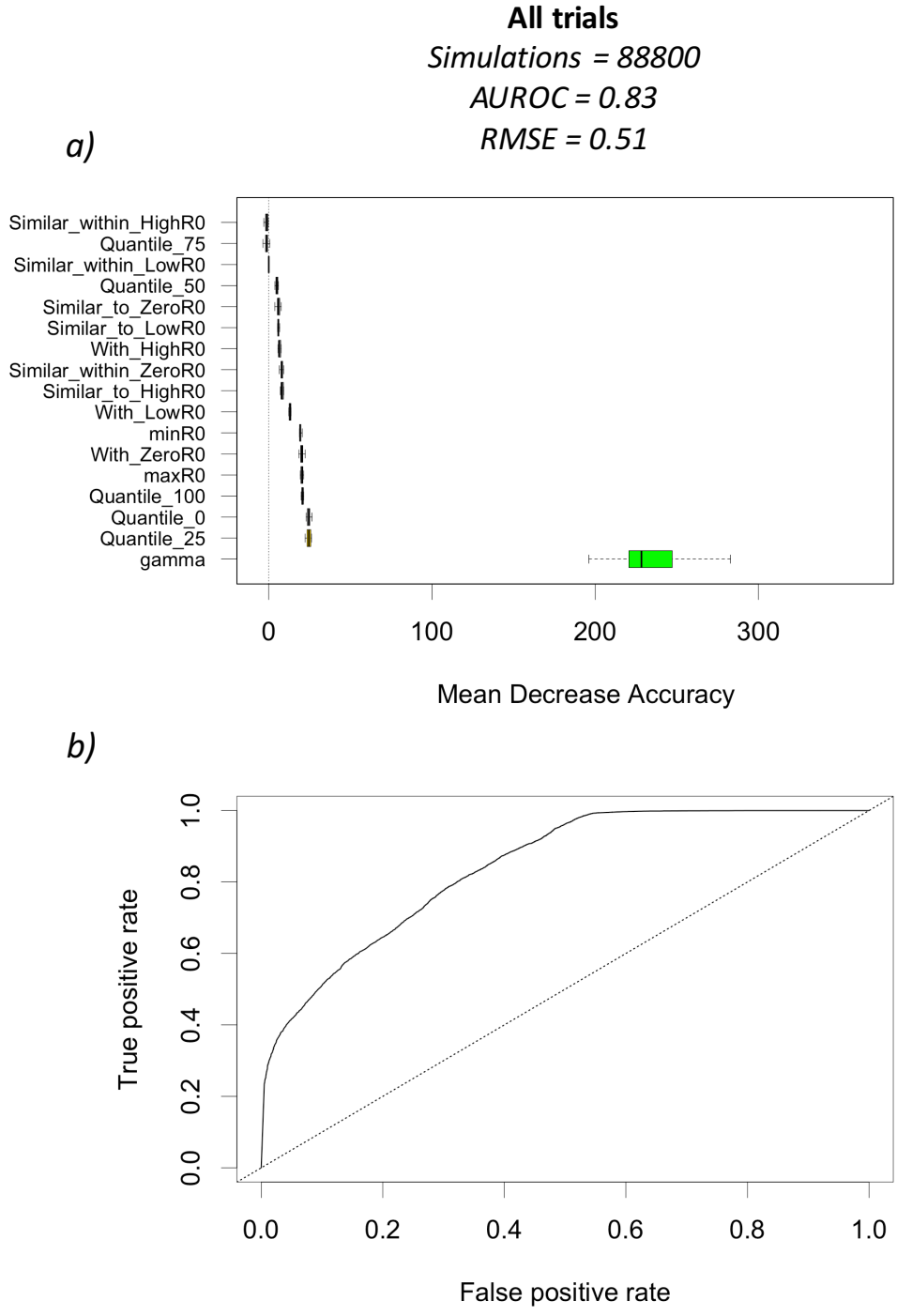


Figure 4.5 All simulations combined were run with the RFA, again showing the clear importance of cross-protection as a predictor of RV-like dynamics. a) The importance plot from the RFA. There is some variation in the MDA of the other predictors relating to strain heterogeneity in transmission, though none of these predictors are of significantly high importance in comparison to cross-protection. **b) The ROC plot from the RFA.** The RFA performance is very good, demonstrated by the good ROC curve, being close to the top and left axes. The performance measures of AUROC = 0.83 and RMSE = 0.51.

Exploring the distribution of “yes” classifications of RV-like dynamics for cross-protection with and without strain heterogeneity (Figure 4.6a) shows that the inclusion of heterogeneity in transmission widens the range of cross-protection permissible for RV-like dynamics.

Partial dependence plots deduce the marginal effect of cross-protection on the RFA results (Figure 4.6b). Using partial dependence plots is a useful technique to investigate the relative contribution of an RFA predictor variable, by giving the marginal effect of the predictor and its value on the RFA prediction ensemble of

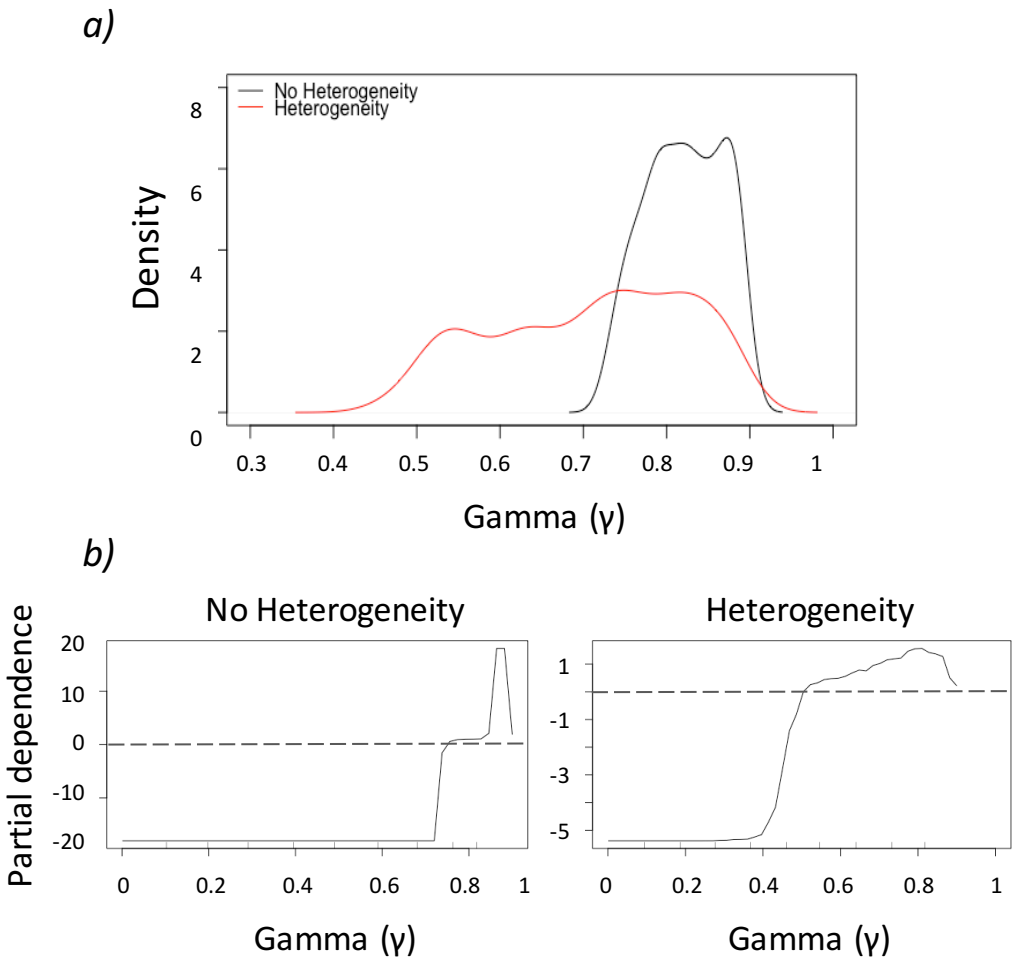


Figure 4.6 a) The distributions of γ values where the RV-like dynamics are observed for simulations with (red) and without (black) strain heterogeneity in transmission, for trials where the whole γ range is used (uniform distribution of range 0-0.9). Heterogeneity in transmission widens the range of cross-protection which leads to RV-like behaviour. **b) The same pattern is observed in the partial dependence plots for γ where heterogeneity is excluded (left) and included (right).**

“yes” and “no” classifications (while ignoring all other predictors). Here (Figure 4.6b), a positive partial dependence value (y-axis) is where the predictor can better predict a “yes” classification. These partial dependence plots also demonstrate that a lower, and wider, range of cross-protection has an influence on the appearance of RV-like dynamics where heterogeneity is included.

The results presented in Figure 4.6 suggest that differences in strain transmission makes cyclical dynamics more likely, by widening the threshold of γ with which they may occur. This explains why the number of “yes” classifications for RV-like dynamics is higher for trials with strain heterogeneity in Figure 4.4. This result was expected, since strain heterogeneity increases single strain dominance and lowers the cross-protection threshold for CSS dynamics, shifting the region of cross-protection for which CSS occurs. This result was presented in Chapter 3, albeit with a weaker effect of heterogeneity, where only one strain differed in transmission.

4.3.2 Changing cross-protection without including strain heterogeneity

Using the reverse POM method, now that it is understood that cross-protection is the most important predictor of RV-like dynamics, it is important to find the best solution of this predictor variable by discriminating the range of “yes” classifications for RV-like behaviour from the “no” classifications.

To discriminate the cross-protection range for “yes” and “no” classifications, the range of γ is narrowed on the distribution of “yes” classifications for the γ predictor variable, while keeping the other variables constant. Narrowing the γ distribution reduces performance measures, and reduces the percent of correct “no” predictions, the false positive rate, to 0% (Table 4.3). This means that as γ is set to a range that maximizes the number of “yes” classifications for RV-like dynamics, the RFA cannot predict “no” results due to the smaller number of “no” classifications to use as a reference for predictions.

Model	γ Range	“Yes” γ Distribution		Number of Simulations		% Correct		AUROC	RMSE
		Mean	SD	Yes	No	Yes	No		
I	U(0,0.9)	0.82	0.05	316	1886	94	90	0.95	0.31
II	N(0.82,0.05)	0.82	0.04	1798	1402	55	59	0.63	0.65
III	N(0.83,0.04)	0.83	0.04	1917	1265	100	0	0.62	0.63

Table 4.3. Performances of models including only variation in cross-protection, where the range of cross-protection is narrowed on the best solution. The performances are colour coded for excellent (green), satisfactory (orange), and poor (red).

The best solution for cross-protection, to maximise the classification of RV-like dynamics, is a distribution of mean 0.83 and standard deviation of 0.04. This is the higher end of CSS, towards the CSS/DSS transition (see previous chapters and (Gupta et al. 1998): the DSS threshold where all strains have an $R_0 = 4$ is equal to $\gamma = 0.875$). In this region, it is already understood that there is cycling but single strain dominance – characteristic of typical CSS dynamics – is already much lower (Gupta et al. 1998; Wikramaratna et al. 2013). As described in the introduction of this Chapter, such properties are common observations of RV population dynamics, and it is therefore not a surprise that the number of simulations achieving RV-like dynamics (“yes”) is maximised.

4.3.3 CHANGING CROSS-PROTECTION WHILE INCLUDING STRAIN HETEROGENEITY

As already discussed in section 4.3.1, including variation in both cross-protection and strain transmission in the POM RFA-based method shows that cross-protection is the most, and only, important predictor of RV-like dynamics (Figure 4.7a). Excluding cross-protection as a predictor in this model (i.e. cross-protection is varied, but not used as a predictor for the RFA) causes the failure of the RFA to predict RV-like dynamics: almost all predictors are of a negative importance (Figure 4.7b), and the performance measures of AUROC and RMSE are not as strong, at 0.58 and 0.73, respectively.

As in section 4.3.2, the distribution of cross-protection is narrowed to find the best solution, but this time while strain heterogeneity is included. A comparison of cross-protection distributions and model performance can be seen in Table 4.4.

Model	γ Range	"Yes" γ Distribution		Number of Simulations		% Correct		AUROC	RMSE
		Mean	SD	Yes	No	Yes	No		
I	U(0,0.9)	0.71	0.12	834	1570	84	88	0.94	0.37
II	N(0.71,0.12)	0.72	0.1	1777	775	85	53	0.82	0.5
III	N(0.77,0.05)	0.77	0.05	2135	434	90	22	0.79	0.47

Table 4.4 Performances of models including variation in both cross-protection and strain heterogeneity, where the range of cross-protection is narrowed on the best solution. The performances are colour coded for excellent (green), satisfactory (orange), and poor (red).

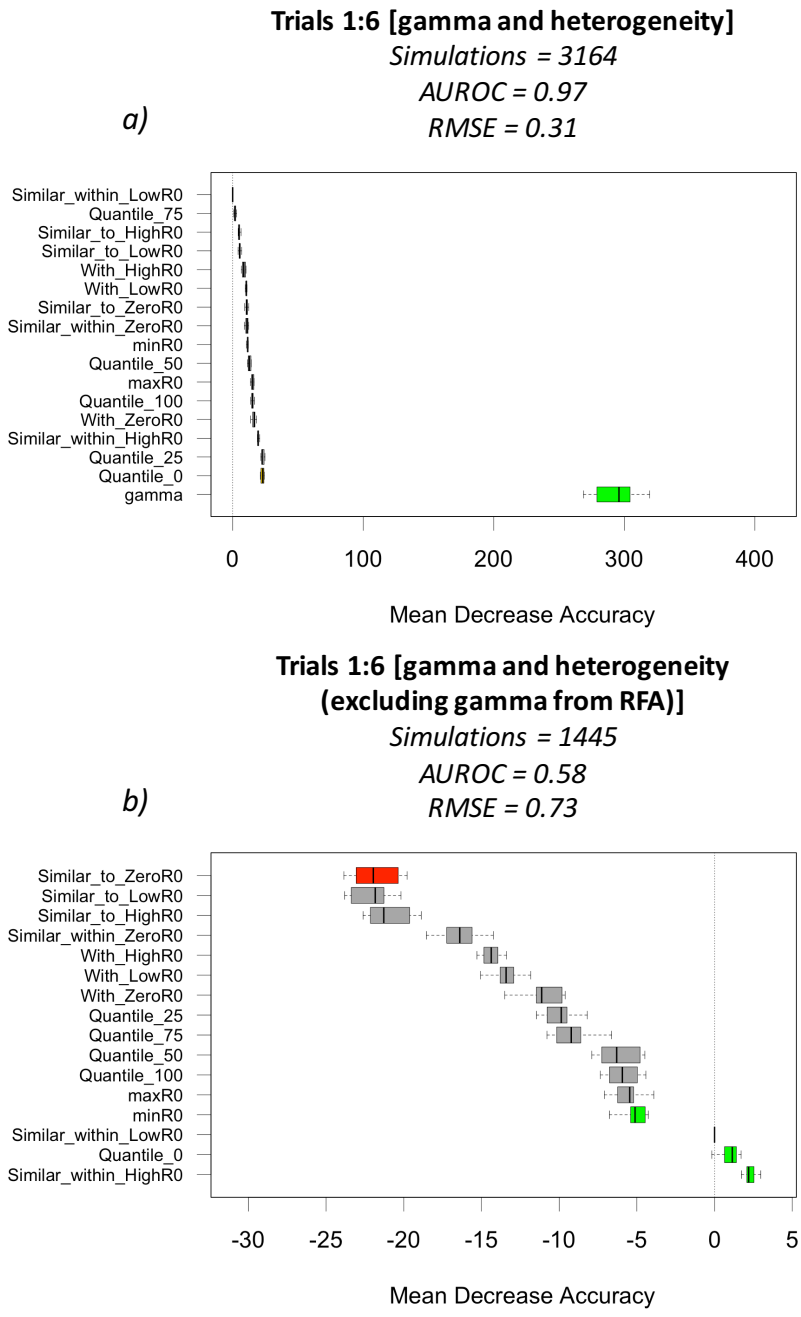


Figure 4.7 The importance of the predictors in the RFA, where both the full range of cross-protection and strain heterogeneities are included, where the variable of cross-prediction is included (a) and excluded (b) as a predictor. When included in the RFA, γ is the most important predictor, and other predictors relating to strain heterogeneity are of similar and very low importance. Ignoring γ as a predictor leads to almost all variables being of negative importance. The AUROC and RMSE for including γ are very good at 0.97 and 0.31, respectively, but these are worse when ignoring γ, at 0.58 and 0.83. Green predictors are the top 2.5% of MDA scores, red predictors are the bottom 2.5%.

As seen before, as the range of cross-protection narrows on the best region to maximise the “yes” classifications of RV-like dynamics, the performance of the model decreases, especially as the predictive ability for “no” classifications decreases – in part due to the lower number of “no” simulations available to the RFA.

As shown in Figure 4.8a, the densities of acceptances for RV-like dynamics across the range of cross-protection for Model I show a distinction between “yes” and “no” simulations. It is of note that these distributions overlap – it is this overlap where heterogeneity in strain transmission potential is playing a role in RV-like dynamics, as at these values of cross-protection the system can be either “yes” or “no” on the sole basis of cross-protection.

The range of cross-protection applied to Models II and III are narrowed and move towards the peak in “yes” for Model I (where the whole range was applied) (Figure 4.8b). This increases the proportion of “yes” simulations in these models, and so these are better ranges for the RV-like dynamics to occur. However, this simultaneously reduces the number of “no” simulations, which reduces the performance measures of the RFA due to having fewer “no” classifications as a reference for predictions.

Given the fixed antigenic structure of a [5 2] system, chosen for biological reasons given the structure of RV as decribed in the Introduction and in the diagram of Figure 4.1, it can be concluded that the influence of cross-protection alone is enough to create RV-like dynamics. This in itself is a good result, and ratifies previous research using multi-strain models that cross-protection drives cyclical dynamics. For completeness, however, it is important to analyse the

effects of strain heterogenity via changes in transmission phenotypes, without the effects of cross-protection.

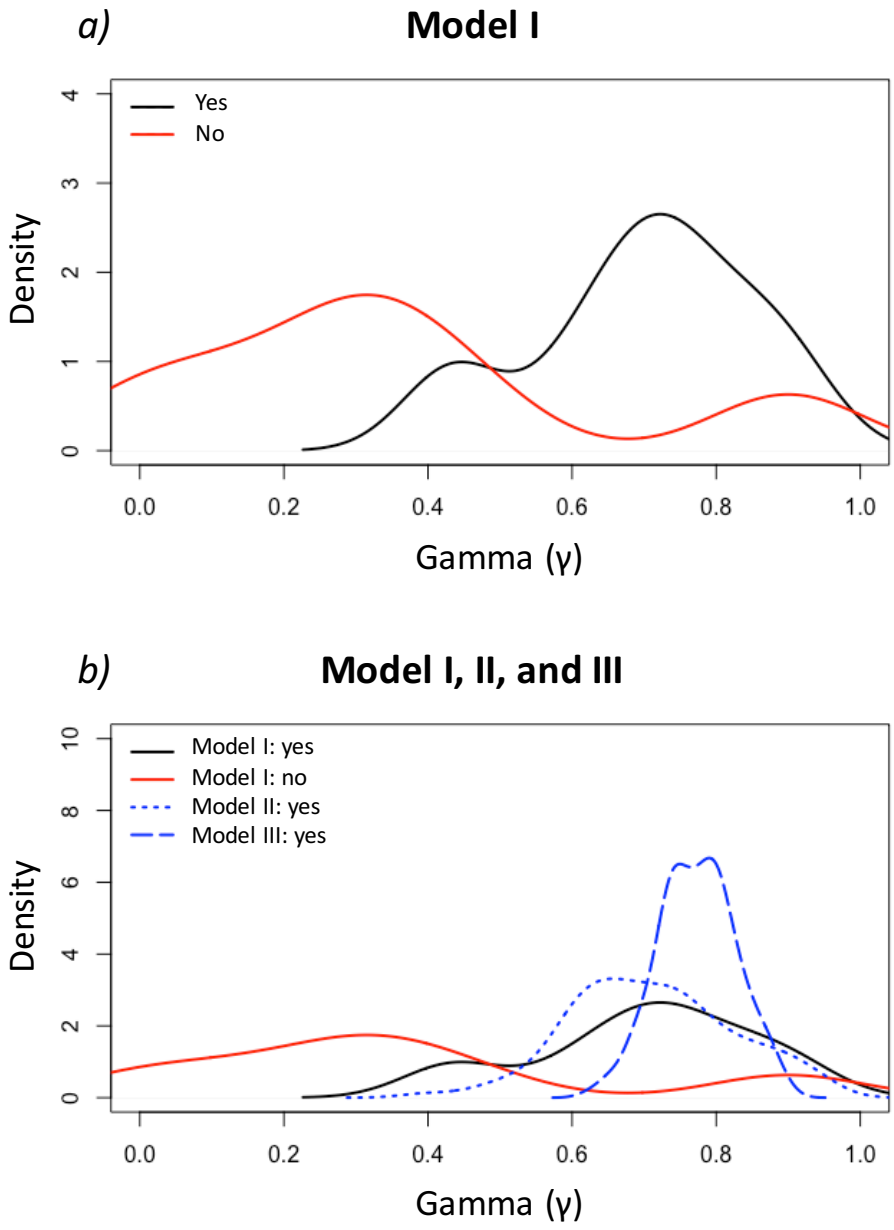


Figure 4.8 a) The density of "yes" and "no" for RV-like dynamics for the range of cross-protection in Model I are plotted. RV-like dynamics are accepted at a high degree of cross-protection, but there is also an overlap with the "no" simulations. **b) The same "yes" and "no" densities are plotted as in a) but here the distributions of "yes" for both Model II and Model III are shown.** As the range of cross-protection used for the simulations are narrowed on the best range for acceptance (moving closer to the peak of "yes"), the performance of the RFA changes.

4.3.4 CHANGING STRAIN HETEROGENEITY WITHOUT INCLUDING VARIATION IN CROSS-PROTECTION

To find the best solution for the transmission values of the strains, the effects of cross-protection is removed by keeping it a constant parameter of the best value found for RV-like dynamics where strain heterogeneity is included (0.77), from section 4.3.3. This was done to find which RFA predictors – other than cross-protection – are of importance, without the effects being blurred by the predictor already known to have the most importance in the RFA.

Figure 4.9a displays the RFA predictor importance for a model with only strain heterogeneity, where γ is kept constant at 0.77. Cross-protection, being constant, has no importance whatsoever (MDA = 0). The RFA has a poor predictive performance, with most predictors of negative MDA and therefore uninformative.

Finding the exact combinations of R_0 between strains which appear to lead to RV-like dynamics is difficult to deduce from the RFA output alone, being a “black box” method. Despite this, transmission homogeneity can be ruled out, as there is at least some variation in importance with predictors relating to strain transmission. The predictor variable of the value of the 50% quantile of R_0 of the strains is of positive importance.

Tracking the proportion of “yes” classifications for RV-like dynamics for simulations of each possible combination of strain transmission (as a measure of importance of each R_0 combination) was plotted in Figure 4.9b. If homogeneity was a good predictor of RV-like dynamics, the proportion of simulations which are accepted for RV-like dynamics should be high where there are no strains with transmission higher or lower than the baseline transmission, and of low

acceptance in other cases. However, there are simulations with heterogeneous combinations that obtain a high proportion of acceptance for RV-like behaviour.

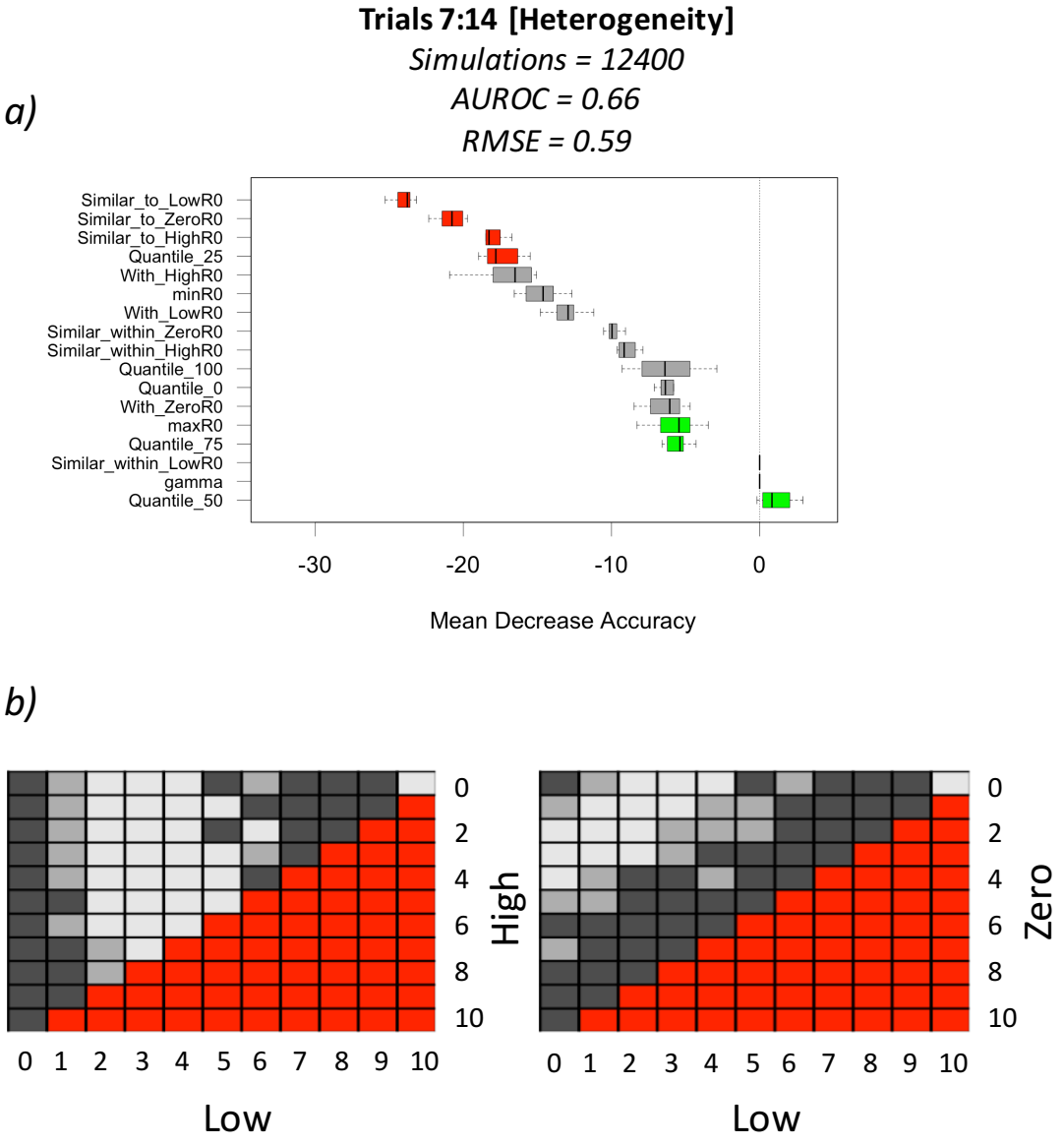


Figure 4.9 a) Importance for simulations where only heterogeneity in strain transmission is included. The only predictor with a positive mean decrease accuracy is the 50% R_0 quantile. **b) The proportion of acceptances for each set of simulations with that combination of low/high/zero strain R_0 values.** The acceptance proportions range from 0.5-1 (i.e. more than random chance), and the red area are the R_0 combinations that were not tested, as there is a maximum of 10 strains in the [2 5] antigenic structure system used here.

It is worth noting that, although the RFA performance was poor, the *only* predictor of positive importance in Figure 4.9a was the 50% quantile for R_0. With partial dependence plots, shown in Figure 4.10a, it can be deduced that the number of strains with baseline transmission potential ($R_0 = 4$), given to all strains which are neither high, low, or removed, are important in determining "yes" classifications. This pattern indicative of homogeneity is of importance for the RFA, suggesting most strains should be of equal transmission.

If homogeneity was an important factor in discerning "yes" and "no" classifications, *all* R_0 quantiles would present a strong signal at $R_0 = 4$ (the baseline transmission potential), which is indicative of homogeneity in strain transmission. Conversely, the partial dependence plots (Figure 4.10a) show three revealing patterns: 1) the 0% quantile peaks at $R_0 = 0$-1, suggesting that at least one strain needs to be removed for RV-like dynamics to occur; 2) the peak of $R_0 = 4$ for 50% quantile suggests that most of the strains should not change from the baseline; 3) the 100% quantile is above $R_0 = 4$ suggesting that there should be at least one strain of a higher R_0.

The density of all R_0 values used for each simulation in the trials used to generate these partial dependence plots is shown in Figure 4.10b. Most strains do have the baseline transmission ($R_0 = 4$), with the other strains being either removed, or double or half the baseline value.

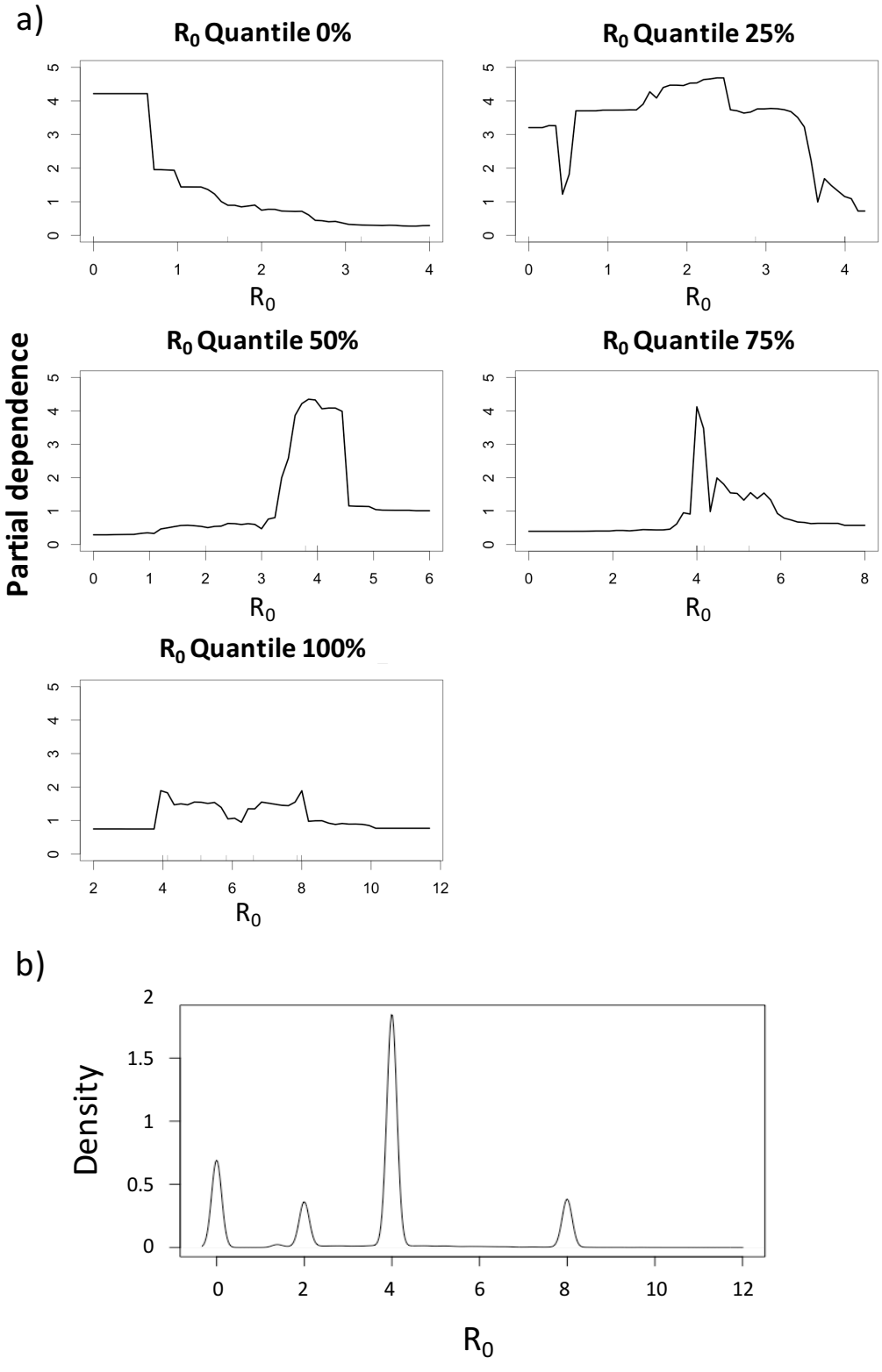


Figure 4.10 a) Partial dependence plots for the R_0 quantiles used as predictors for the RFA (where only strain heterogeneities, not γ, are included). The higher the partial dependence, the greater the ability to predict "yes" for the RFA at that value of R_0. **b) The density of R_0 values used for the simulations in a).** The highest peak, at R_0 = 4, is the baseline and most strains have this transmission potential. There are also small peaks at 2 (the "min" value; half the baseline), 8 ("max", double the baseline), and 0 (removed strains).

4.4 DISCUSSION

I have presented a novel approach for analysing the drivers of patterns of the multi-strain model developed by Gupta et al (Gupta et al. 1996; Gupta et al. 1998). Using Random Forest as a machine learning tool, the important variables, such as cross-protection and heterogeneity in strain transmission, leading to a particular type of oscillatory behaviour can be found. This technique has been developed to find the pathogen properties required for typical rotavirus population dynamics in developed countries, but could be applied to other pathogen dynamics, such as influenza or dengue.

Random Forest is an incredibly useful tool, with the ability to use very large datasets and a high number of predictor variables. Understanding which individual-level variables are of most importance to the higher-level patterns of the system (such as cyclic behaviour and co-circulation) is useful to then focus on those variables in further analyses. For instance, RFA have been used in many genomic analyses to find which genes may be associated with certain diseases (Chen & Ishwaran 2012).

Here, it has been found that the variable of cross-protection is the most important factor contributing to RV-like dynamics. This was expected, as the immunological interactions between strains of shared alleles are known to lead to cyclical dynamics (Gupta et al. 1998; Gupta et al. 1996; Lourenço & Recker 2013; Wikramaratna et al. 2013; Recker & Gupta 2005; Recker et al. 2004), a major feature of RV population patterns. However, although cross-protection was of most importance and needed to be included in the RFA for a good model

performance, heterogeneity of strain phenotypes, by changing transmission rates, was found to play a complementary role.

Heterogeneity in strain transmission was found to widen the range for which cross-protection can take to allow RV-like dynamics. It is known that differences in strain phenotype disrupts the multi-locus system, and can create oscillations: this has been demonstrated in Chapters 2 and 3. This disruption of the stable NSS and DSS states, and differences in the prevalence of strains due to transmission variation, subsequently creates RV-like dynamics – in part, however, because these are the characteristics that I defined RV populations should display according to the information available in the literature.

An issue with using the RFA is that, although it is excellent at finding key variables from large datasets, solutions are in their nature difficult to track, given the dimensionality of the analysis, seemingly appearing as a “black box”. It is therefore difficult to fully interpret which combinations of the strains’ transmission potential may lead to RV-like dynamics. From investigating the distribution of acceptances of RV-like behaviour and using partial dependence, I suggest that RV-like dynamics are more likely where: 1) there exists at least one strain of low or zero transmission; 2) there exists at least one strain of high transmission; 3) the relationships between strains of low, high, or zero transmission, and between those strains and the normal “baseline” transmission, do not make a difference.

There is limited research on the transmission rate of rotavirus and possible heterogeneities between strains. However, a multi-strain SIR model was only able to fit to rotavirus data when G1P[8] was given a 10% higher transmission rate (Pitzer et al. 2011b). Here, variation of the transmission potential of the

strains was not crucial for the appearance of RV-like dynamics, but still contributed towards these dynamics. Having one strain of double the transmission rate was an advantage, but not required, and certainly not as influential as the degree of cross-protection. Having cross-protection as the unique predictor for RV-like dynamics is biologically unreasonable: disease systems do exhibit natural variation in intrinsic properties, so it would be expected that this variation plays a role in population dynamics.

Unlike other models of rotavirus (Pitzer et al. 2009; Pitzer et al. 2011b; Atchison et al. 2010; Van Effelterre et al. 2010), and suggested from empirical data (Velázquez et al. 2000; Velázquez et al. 1996; Chiba et al. 1986; O'Ryan et al. 1994), this work did not include differences in the cross-protection (or severity of disease, as mortality or cost of infection has not been included at all) with the number of infections. This is not possible in the MANTIS framework used here, as the ability to add many strains removes the ability for the model to be tractable, and so the number of prior infections is unknown. Without this factor, though, the same patterns still emerge – the RFA suggests that with transmission differences the cross-protection range for rotavirus dynamics is 0.71-0.73 (and 0.82-0.83 without transmission differences), which is relatively high. However, empirical evidence for how cross-protective the anti-rotavirus response is limited, but the high degree found here could potentially reflect the epidemiological observation of broadening of the immune response with the number of RV infections.

The technique and definition of RV dynamics used in this research is aimed at the dynamics observed in developed countries (Iturriza-Gomara et al. 2000; Santos & Hoshino 2005), whereas these dynamics differ in developing countries. The latter differ by: more unusual strain combinations, more P[6] strains, and

high strain diversity (Santos & Hoshino 2005). It is not fully understood the mechanisms behind the differing dynamics – higher diversity and more reassortments could be a result of more animal-human contact in developing countries (Santos & Hoshino 2005), different cyclical dynamics are suggested to be a result of differences in birth rates (Pitzer et al. 2009), and more P[6] due to differences in blood types (Sub-Saharan Africa has a much higher prevalence of Lewis negative types) (Nordgren et al. 2014). It would be of interest to apply the different population dynamics of developing countries to the RFA again (this is relatively simple, achieved by changing the definition of RV, and altering the antigenic structure of the multi-locus model), and compare with the results presented here.

The research presented here shows that for a given antigenic structure (of two loci of five and two alleles each, [5 2]), particular value ranges of cross-protection are sufficient for generating RV-like dynamics. This implies two things:

1) The patterns observed in developed countries (for which two decisions were based on: the antigenic structure and the RV-like behaviour rules) can be explained by a current antigenic structure, the behaviour at the CSS-DSS boundary, and a network of hard-wired cross-protection between existing strains. According to theory, and the previous chapters within this thesis, this explains why the population structure and behaviour of RV is so stable in developed countries: inter-strain competition maintains the existing strains in a particular behaviour. In addition, a system that is locked in a behaviour with a few DSS properties must be highly stable – as Chapter 2 concluded, even with the possibility for the invasion of new, fitter strains, DSS is always maintained. Even though RV is an RNA virus with a segmented genome, with a large potential for

mutation and recombination, diversity is limited due to cross-protection and the subsequent semi-DSS state of the population structure.

2) It does not necessarily mean that because developing countries are more unstable and have a higher number of (and different) dominating strains that the population behaviours are different between developing and developed countries. The higher number of unusual strains (suggested to be due to animal reservoirs, for instance) alters the typical antigenic structure than used in this study. This would subsequently alter the degree of cross-protection required for RV-like dynamics, meaning the strains in developing countries may have cross-protection that is too low, in which case the population would display more CSS-like dynamics. However, the full potential for diversity will be kept at bay by the antigenic structure and cross-protection, in the same way as for developed countries.

The diversity and dominance of different strains in developing countries needs to be understood, as it has important implications in rotavirus control. The two vaccines currently in use target the main strains circulating in the US and Europe. It is known that the vaccine is less effective in developing countries in Africa (Armah et al. 2012; Armah et al. 2010) and Asia (Zaman et al. 2010) compared to developed countries in Europe, and the US. Understanding the dynamics may help towards future rotavirus vaccine development, to include a variety of strains of different G and P types but also different genogroups, to widen cover by cross-protection.

It has previously been suggested that differences in HBGA distributions (and subsequently, differences in susceptibility to different rotavirus P types (Nordgren

et al. 2014; Huang et al. 2012)) may be influencing the distribution and diversity of RV strains (Santos & Hoshino 2005), though evidence for this is lacking. Further expansion of the work presented here could be used to test this hypothesis. Heterogeneity in immune responses, such as those presented in Chapter 2, can be included into the multi-locus framework, to mirror the differences in rotavirus susceptibility dependent on host HBGA genotype. Using definitions for RV-like dynamics for both developing and developed countries, the importance of host genotype proportion (along with the variables tracked here) could potentially be elucidated. For instance, host heterogeneity is found in this thesis to reduce diversity (via a decrease in the cross-protection required for DSS); adding this factor could lower the cross-protection for RV-like dynamics, or perhaps change the variation in R_0 required.

The flexibility and scope of this technique to use machine learning alongside multi-locus modelling provides an excellent framework to investigate the mechanisms and variables behind pathogen dynamics. This work can be expanded for more complex rotavirus dynamics, as suggested here, but could also be altered to explore important characteristics of other pathogens, such as influenza or dengue.

4.5 APPENDIX

4.5.1 PERFECTING THE DEFINITION OF RV-LIKE DYNAMICS

The definition for RV-like dynamics has had several versions, but it was important to analyse which time series fit the RV-like best, and not simply CSS of any form. The final stage in the definition of RV-like dynamics (Figure 4.2b) is the RV-specific definition, after removing cases of NSS and DSS. This definition has been edited many times to target the more realistic dynamics, and to provide a good acceptance rate for the RFA (approximately 60% of simulations being accepted leads to better accuracy). This final definition has changed as follows:

1. **no_domstrains > 1**

   If there were more than one dominant strain in a time series, it was not accepted as RV-like. This was too strict, as no time series were accepted.

2. **no_domstrains > nStrains/2**

   The number of dominant strains was loosened to be more than one, to allow a change in dominance. There still needed to be a threshold on the number of dominant strains as regular CSS usually shows most, or all, of the strains cycling, so it was decided that only half of the possible strains could dominate over the period. This was too loose - time series which didn't look RV-like (instead, classic CSS behaviour) were always being accepted.

3. **no_domstrains> nStrains/4**

   The number of dominant strains was reduced to a quarter of possible strains, but now the number of accepted time series was far too low (leading to the prediction scores for yes being <1%).

4. **1 < no_RVobs < 2**

   The method of defining RV dominance was changed to use the number of observable strains (*no_RVobs*), i.e. a strain's prevalence being above an 'observable' threshold of 10^{-4} rather than simply being of higher prevalence than other strains. RV-like dynamics were defined as <2 observable strains for a majority (90%) of time points. However, with classic CSS all strains are cycling so the prevalence is lower, and therefore *no_RVobs* in these cases are zero. To avoid classic CSS, the definition was narrowed to 1 < no_RVobs < 2.

5. **Length(PrevalentStrains_High) > 1 | is.na(Low_obs)**

   As cross-protection and the number of cycling strains influences the absolute prevalence, using "observable" strains is difficult. Instead, relative prevalence was then used in conjunction with the number of observable strains. RV-like dynamics were defined as more than one strain with relative prevalence >15% (*PrevalentStrains_High*) and at least one strain of relative prevalence 10-15% (*PrevalentStrains_Low*). In cases where there were no strains of >15% (for instance, when there were many strains cycling overall prevalence was lower), the number of

PrevalentStrains_Low which were still above the observable threshold were counted (*Low_Obs*).

This final definition provided a good acceptance ratio and prediction scores in the RFA, and is used for all RFA results seen here.

4.5.2 CHOOSING WHICH PREDICTORS TO USE

Not all possible predictor variables should be used for an RFA analysis: pairs of predictors which are correlated or directly related to the outcome should be ignored. For this reason, the outcomes for strain structure and RV-like behaviour were ignored, as these defined the classification of RV-like dynamics which was used as the RFA classifier. The predictors “SSDMean” (single strain dominance) and “divmeanZ/Y” (mean Shannon diversity of immune and infected proportions, respectively) were also ignored due to their strong correlation with strain structuring and therefore CSS. Additionally, the number of strains and the antigenic structure were not included as predictors as here these were kept constant (10 strains, for a [2 5] system).

4.5.3 TABLE OF ALL MODELS USED FOR THE RFA

Trial	No. sims	No. repeats	γ	Lower R0	Higher R0	Zero R0	R0 Trials	No. yes	No. no	% Correct yes	% Correct no	AUROC	RMSE	OOB error
CROSS-PROTECTION AND HETEROGENEITY														
1	3600	1049	U(0,0.9)	random 0:5		0:5	20	327	722	0.81	0.91	0.95	0.35	0.12
2	8000	2115	U(0,0.9)	0:3	0:3	0:4	20	750	1365	0.89	0.92	0.97	0.3	0.09
SUMMARY 1:2	**11600**	3164	**whole γ range**	**heterogeneity**				1077	2087	0.88	0.92	0.97	0.31	0.10
3	3600	1090	N(0.71,0.12)	random 0:5		0:5	20	689	401	0.82	0.6	0.81	0.51	0.27
4	8000	2193	N(0.71,0.12)	0:3	0:3	0:4	20	1557	636	0.93	0.59	0.89	0.41	0.17
SUMMARY 3:4	**11600**	3283	**narrow γ range**	**heterogeneity**				2246	1037	0.88	0.6	0.87	0.45	0.21
5	3600	1107	N(0.77,0.05)	random 0:5		0:5	20	875	232	0.88	0.38	0.79	0.48	0.23
6	8000	2191	N(0.77,0.05)	0:3	0:3	0:4	20	1856	335	0.95	0.33	0.85	0.38	0.14
SUMMARY 5:6	**11600**	3298	**narrow γ range**	**heterogeneity**				2731	567	0.92	0.36	0.84	0.42	0.17
HETEROGENEITY														
7	3600	907	0.77	random 0:5		0:5	20	737	170	0.86	0.14	0.66	0.53	0.28
8	8000	451	0.77	0:3	0:3	0:4	20	346	105	0.77	0.055	0.66	0.63	0.39
SUMMARY 7:8	**11600**	1327	**constant γ**	**heterogeneity**				1059	268	0.83	0.11	0.67	0.55	0.31
9	900	31	0.77	0	0	0:8	20	16	15	0.12	0.04	0.95	0.97	0.91
10	900	293	0.77	random 0:8		0	20	191	102	0.53	0.12	0.65	0.79	0.61
11	2500	141	0.77	0:4	0:4	0	20	79	62	0.31	0.019	0.83	0.91	0.81
12	900	252	0.77	random 0:8		1	20	213	39	0.88	0	0.53	0.5	0.25
13	3600	130	0.77	0	0:5	0:5	20	94	36	0.66	0	0.56	0.72	0.51
14	3600	130	0.77	0:5	0	0:5	20	90	40	0.63	0.015	0.49	0.75	0.55
SUMMARY 7:14	**12400**	935	**constant γ**	**heterogeneity**				661	274	0.65	0.047	0.53	0.73	0.52
CROSS-PROTECTION														
15	**10000**	**2202**	**U(0,0.9)**	**0**	**0**	**0**	**200**	316	1886	0.94	0.9	0.95	0.31	0.10
16	**10000**	**3200**	**N(0.82,0.05)**	**0**	**0**	**0**	**200**	1798	1402	0.55	0.59	0.63	0.65	0.42
17	**10000**	**3182**	**N(0.83,0.04)**	**0**	**0**	**0**	**200**	1917	1265	1	0	0.62	0.63	0.40
ALL MODELS														
All	**88800**	**18040**						11678	8748	0.9	0.57	0.84	0.49	0.24

5 DISCUSSION

Many pathogens circulate as multiple distinct strains, such as observed for influenza A, *Plasmodium falciparum*, and *Neisseria meningitidis*. Mathematical models of multi-strain pathogens are a useful tool to explore the evolutionary and ecological interactions between strains, and the effect this has on pathogen population dynamics. The aim of this thesis was to investigate the stability of the equilibrium states of a multi-strain model, and how these states shift under different conditions; namely, the effects of intrinsic transmission of the pathogen and of heterogeneity in host immune responses.

Immune pressure from the host creates strain-specific selection, as strains which are antigenically similar can be cross-recognised by the host immune response, reducing the chance of infection or transmission via the effect of cross-protection. A high degree of cross-protection limits antigenic diversity, as only strains of non-overlapping antigenic repertoires can co-exist, and this behaviour is known as discrete strain structure (DSS) (Gupta et al. 1998).

Previous models have explored the stability of DSS in the face of recombination (Gupta et al. 1996; Artzy-Randrup et al. 2012; Buckee et al. 2004) and different host immune responses (Cobey & Pascual 2011; Recker & Gupta 2005; Recker et al. 2007; Gupta & Galvani 1999). This was further explored in Chapter 2 and Chapter 3, which both show how DSS is highly robust to strain transmission variation and different host immunity responses, respectively.

In Chapter 2, it was found that DSS is robust to invasion by strains which are antigenically similar to the dominant strains, with the invading strain requiring a

large transmission advantage to cause an epidemic and subsequently persist. Even after invasion, the pathogen population still displayed DSS, only with a change in the dominant set of strains. Although replacement was triggered by a transmission change in one previously suppressed strain, immune selection alone was seen to re-enable suppressed strains to return the system to DSS – due to the exclusion of the previously dominant strains, other strains which were previously suppressed and virtually extinct underwent competitive release, despite not changing transmission phenotype. This is an important result, and a result of using a multi-locus model with immune selection, which demonstrates that a strain's emergence is not only determined by its transmission potential but also by inter-strain competition. Exploring emergence in a multi-strain model, rather than focussing on models of one or two strains, has implications for diseases which circulate as distinct strains, such as in the meningitis-causing bacteria *N. meningitidis* and *Streptococcus pneumoniae,* and influenza A.

Analytical derivations of the multi-strain model of Gupta et al found that the system is more resistant to invasion when there is/are: 1) a higher degree of cross-protection, as a higher transmission advantage of the invading strain is required to overcome competitive suppression from the antigenically overlapping dominant strains; 2) more dominant strains, as this increases the effect of herd immunity, reducing the availability of susceptible hosts for the invading strain.

The results of the numerical simulations showed discrepancies with the analytical results, in part due to the fact that the definition of invasion in the derivations and previous work (such as in (May et al. 2001)) refers to the presence of the invading strain, whereas the simulations here explored the switch in dominance during

DSS and the subsequent persistence of the invading strain. This difference in the definition would cause differences in the results.

The discrepancies also suggested that other factors might be involved in a strain's ability to invade. These factors, which influence oscillatory behaviour and the degree of herd immunity, ultimately influence the probability of emergence. Emergence was found to be more likely in pathogens displaying the characteristics of: a shorter duration of infectiousness (as a quicker replenishment of susceptible hosts increases the size of the initial epidemic of the invading strain, making successful invasion more likely); a larger seed of infected individuals at the time of introduction; parameter variation, as this makes the invasion threshold plastic around a deterministic expectation, allowing less fit (i.e. lower R_0) strains to invade, while fit strains may be unsuccessful; and finally the inclusion of strong seasonality, which allows a small probability of emergence at low R_0 values. In addition, it was found that having continuous background seeding of infected individuals "primes" herd immunity, further suppressing the invading strain – suggesting that a pathogen with multiple re-introductions via migration or mutation events is less likely to successfully invade than if its introduction is driven by sporadic one-off events.

A limitation of the invasion model presented in Chapter 2 was that demographic stochasticity could not be included due to the intractability of the multi-locus model. This work could be extended into an IBM to incorporate stochasticity, and also including different host structures and networks, akin to a previous multi-locus model by Buckee et al (Buckee et al. 2004).

In Chapter 3, a new multi-strain model was developed which included the possibility of a host recognising a conserved determinant, thereby being immune to all possible strains. Despite a reduction in strain-specific immune selection when conserved immunity is high, DSS was still observed as long as there was a small probability of some hosts acquiring strain-specific immunity.

Akin to other multi-strain models investigating changes to the equilibrium boundaries (Recker & Gupta 2005; Gog & Grenfell 2002; Blyuss 2013; Cherif 2015; Lin et al. 1999)**,** there was a shift in the cross-protection threshold for DSS as immunity in the host population was varied. Here, it was found that an increase in the probability of a host, or proportion of hosts in a population, responding to conserved determinants reduces the cross-protection range for which cyclic dynamics occur, and decreases the cross-protection threshold for DSS. Therefore, it can be concluded that a proportion of hosts acquiring conserved immunity can reduce overall antigenic diversity and stabilise the pathogen population, a pattern also observed in (Gupta & Galvani 1999). Additional work could be carried out in future for a more in-depth analysis of the equilibrium boundaries and a comparison of the results of this model with other multi-strain theoretical frameworks.

Conserved immunity not only reduced the parameter range for cyclic dynamics, but within this range lengthened the persistence time of a strain; a conserved immune response reduced the period between strain replacement events. This suggests that immunity to conserved determinants not only changes population dynamics, but may slow the observed rate of antigenic change of the pathogen as strain-specific immune selection is reduced. This could have impacts on methods of control, for instance universal vaccines against conserved

determinants of a pathogen (such as the RTS,S vaccine for *P. falciparum*, and a universal influenza vaccine) may have wider effects on the pathogen's oscillatory dynamics and evolution.

The evolutionary impacts of conserved immunity and host heterogeneity could be explored further by developing the model into a co-evolutionary model, to investigate the effects of host heterogeneity on pathogen population dynamics and conversely how the pathogen influences the proportion of host genotypes. This could be achieved by incorporating host genotype-specific mortality rates, with the targeting of epitopes, similar to the work of (Penman et al. 2013) for HLA genotypes – where an infected host of the incorrect genotype for recognising the epitopes of that strain would have a risk of mortality. Chapter 3 could be extended so that hosts not only differed in how they acquire immunity to either conserved or variable antigens, but also differed in their risk of mortality.

The appearance of different population dynamics and boundaries has been investigated for cross-protection and a variety of other additional factors. However, phenotypic variation had not been explored in a multi-strain model in depth. In Chapter 4, a technique of reverse pattern-oriented modelling was applied using a Random Forest algorithm (RFA), which enables the discovery of multiple pathogen factors – alone and in combination – leading to a set of desired and pre-defined structures (i.e. strain dominance) and patterns (i.e. cycles) of the pathogen population (e.g. akin to those observed in data).

The reverse pattern-oriented modelling approach presented here focused on the dynamics of rotavirus (RV), specifically within developed countries where lower diversity is observed in comparison to developing countries. RV displays similar

cyclic dynamics to influenza A, however with more co-circulation and less frequent changes in strain dominance.

It was found that cross-protection is of most importance for the appearance of RV-like dynamics, which is to be expected for two reasons: 1) cross-protection is already known to be a key determinant in the appearance of cyclical dynamics of multi-strain models, and oscillatory behaviour is defined as RV-like; 2) RV is known to display cross-reactivity and cross-protection between strains, and the parameter region found here matches those of clinical studies (Velázquez et al. 1996) and other data-based models (Atchison et al. 2010; Atchison et al. 2010; Pitzer et al. 2009; Pitzer et al. 2011a).

Along with cross-protection, variation in transmission phenotypes between strains was explored. Finding the combinations of these transmission properties was a challenge, and this is a drawback of using a “black box” algorithm such as the RFA. Despite this challenge, heterogeneity in strain transmission potential appeared to be of some importance to generate RV-like dynamics using a multi-locus model, albeit to a lesser extent than cross-protection.

The antigenic structure and cross-protection parameter range was sufficient to elicit RV-like dynamics, with the pre-definition based on the observed patterns in developed countries. The different patterns in developing countries (e.g. higher diversity and more unusual strain combinations) could therefore be suggested to be caused by differing possible antigenic structures, and the strength of cross-protection, leading to more CSS-like dynamics. For both developing and developed countries, cross-protection maintains the population in a semi-DSS state, limiting the possibilities for diversity. Further work would benefit this

research, by altering the RFA definitions to reflect developing countries dynamic properties, and more rigorously investigating the differences in the RFA results to reflect how and why differences in the dynamics (and possibly the vaccine efficacy) between these countries exists.

This technique could be applied to other multi-strain pathogens by altering the pre-defined population dynamic characteristics of interest, and the parameters of the model, to make it disease-specific (for instance, diseases such as respiratory syncytial virus and hand, foot, and mouth disease which are becoming better understood but gain less research attention than influenza). This is a useful tool to understand the individual-level properties which may lead to particular pathogen population dynamics, where robust and detailed epidemiological data from clinical trials or public health records may be limited temporally and/or spatially.

Mathematical models are a useful and essential tool in epidemiology, for the understanding of the evolutionary and ecological effects of host immunity, intrinsic factors, and disease control at the pathogen population level. The use of mathematical models is of particular significance when epidemiological data is limited (such as with RV) or not yet available (such as with emerging diseases at the start of an epidemic, e.g. Ebola or Zika), providing insights which may not be possible any other way. As many globally important diseases circulate as multiple distinct strains, the multi-strain models described and developed in this thesis are imperative to explore the interactions and feedback between strains. The research presented here focused on facets of multi-strain population dynamics never explored before, and introduced novel approaches for future investigation.

BIBLIOGRAPHY

Adams, B. et al., 2006. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38), pp.14234–9.

Alexander, H.K. et al., 2010. Risk factors for the evolutionary emergence of pathogens. *Journal of the Royal Society, Interface / the Royal Society*, 7(51), pp.1455–74.

Allison, A.C., 1954. Protection afforded by sickle-cell trait against subtertian malareal infection. *British Medical Journal*, 1(4857), pp.290–4.

Alonso, D., McKane, A.J. & Pascual, M., 2007. Stochastic amplification in epidemics. *Journal of the Royal Society, Interface / the Royal Society*, 4(14), pp.575–82.

Amanna, I.J., Carlson, N.E. & Slifka, M.K., 2007. Duration of Humoral Immunity to Common Viral and Vaccine Antigens. *New England Journal of Medicine*, 357(19), pp.1903–1915.

Anderson, R.M., Grenfell, B.T. & May, R.M., 1984. Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. *The Journal of Hygiene*, 93(3), pp.587–608.

Anderson, R.M. & May, R.M., 1992. *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press.

André, J.B. & Day, T., 2005. The effect of disease life history on the evolutionary emergence of novel pathogens. *Proceedings. Biological Sciences / The Royal Society*, 272(1575), pp.1949–56.

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

Antia, R. et al., 2003. The role of evolution in the emergence of infectious diseases. *Nature*, 426(6967), pp.658–661.

Antia, R., Nowak, M.A. & Anderson, R.M., 1996. Antigenic variation and the within-host dynamics of parasites. *Proceedings of the National Academy of Sciences*, 93(3), pp.985–989.

Antrobus, R.D. et al., 2014. Coadministration of seasonal influenza vaccine and MVA-NP+M1 simultaneously achieves potent humoral and cell-mediated responses. *Molecular Therapy: The Journal of the American Society of Gene Therapy*, 22(1), pp.233–8.

Arinaminpathy, N. et al., 2009. Evolution and emergence of novel human infections. *Proceedings. Biological Sciences / The Royal Society*, 276(1675), pp.3937–43.

Arinaminpathy, N. et al., 2012. Impact of cross-protective vaccines on epidemiological and evolutionary dynamics of influenza. *Proceedings of the National Academy of Sciences*, 109(8), pp.3173–3177.

Arista, S. et al., 2006. Heterogeneity and temporal dynamics of evolution of G1 human rotaviruses in a settled population. *Journal of Virology*, 80(21), pp.10724–33.

Armah, G.E. et al., 2010. Efficacy of pentavalent rotavirus vaccine against severe rotavirus gastroenteritis in infants in developing countries in sub-Saharan Africa: a randomised, double-blind, placebo-controlled trial. *Lancet*, 376(9741), pp.606–14.

Armah, G.E. et al., 2012. Immunogenicity of the pentavalent rotavirus vaccine in African infants. *Vaccine*, 30 Suppl 1, pp.A86-93.

Artzy-Randrup, Y. et al., 2012. Population structuring of multi-copy, antigen-encoding genes in Plasmodium falciparum. *eLife*, 1, p.e00093.

Atchison, C., Lopman, B. & Edmunds, W.J., 2010. Modelling the seasonality of rotavirus disease and the impact of vaccination in England and Wales. *Vaccine*, 28(18), pp.3118–3126.

Atchison, C.J. et al., 2010. Temperature-dependent transmission of rotavirus in Great Britain and The Netherlands. *Proceedings of the Royal Society B: Biological Sciences*, 277(1683), pp.933–942.

Babl, F.E. et al., 2001. Constancy of distribution of serogroups of invasive pneumococcal isolates among children: experience during 4 decades. *Clinical Infectious Diseases : an official publication of the Infectious Diseases Society of America*, 32(8), pp.1155–61.

Bambini, S. et al., 2013. An analysis of the sequence variability of meningococcal fHbp, NadA and NHBA over a 50-year period in the Netherlands. *PloS one*, 8(5), p.e65043.

Bányai, K. et al., 2012. Systematic review of regional and temporal trends in global rotavirus strain diversity in the pre rotavirus vaccine era: Insights for understanding the impact of rotavirus vaccination programs. *Vaccine*, 30(Suppl. 1), pp.A122–A130.

Barba-Spaeth, G. et al., 2016. Structural basis of potent Zika–dengue virus antibody cross-neutralization. *Nature*, 536(7614), pp.48–53.

Barry, A.E. et al., 2007. Population genomics of the immune evasion (var) genes of Plasmodium falciparum. *PLoS pathogens*, 3(3), p.e34.

Barry, A.E. et al., 2011. The stability and complexity of antibody responses to the major surface antigen of Plasmodium falciparum are associated with age in a malaria endemic area. *Molecular & Cellular Proteomics : MCP*, 10(11), p.M111.008326.

Baruch, D.I. et al., 1995. Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, 82(1), pp.77–87.

Batista, C.L. et al., 2014. Genetic diversity of Plasmodium vivax over time and space: a community-based study in rural Amazonia. *Parasitology*, pp.1–11.

Bedford, T. et al., 2015. Global circulation patterns of seasonal influenza viruses

vary with antigenic drift. *Nature*, 523(7559), pp.217–20.

Bedford, T., Rambaut, A. & Pascual, M., 2012. Canalization of the evolutionary trajectory of the human influenza virus. *BMC biology*, 10(1), p.38.

Bennett, S.N., 2003. Selection-Driven Evolution of Emergent Dengue Virus. *Molecular Biology and Evolution*, 20(10), pp.1650–1658.

Blyuss, K.B., 2014. Analysis of symmetries in models of multi-strain infections. *Journal of Mathematical Biology*, 69(6–7), pp.1431–1459.

Blyuss, K.B., 2013. The effects of symmetry on the dynamics of antigenic variation. *Journal of Mathematical Biology*, 66(1–2), pp.115–137.

Botosso, V.F. et al., 2009. Positive selection results in frequent reversible amino acid replacements in the G protein gene of human respiratory syncytial virus. *PLoS pathogens*, 5(1), p.e1000254.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45, pp.5–32.

Buckee, C.O. et al., 2010. Long-term evolution of antigen repertoires among carried meningococci. *Proceedings. Biological Sciences / The Royal Society*, 277(1688), pp.1635–41.

Buckee, C.O. et al., 2004. The effects of host contact network structure on pathogen diversity and strain structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(29), pp.10839–44.

Bull, P.C. et al., 1998. Parasite antigens on the infected red cell surface are

targets for naturally acquired immunity to malaria. *Nature Medicine*, 4(3), pp.358–60.

Burton, D.R. & Mascola, J.R., 2015. Antibody responses to envelope glycoproteins in HIV-1 infection. *Nature Immunology*, 16(6), pp.571–576.

Bush, R.M. et al., 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution*, 16(11), pp.1457–1465.

Cabrera, G. et al., 2005. The sickle cell trait is associated with enhanced immunoglobulin G antibody responses to Plasmodium falciparum variant surface antigens. *The Journal of Infectious Diseases*, 191(10), pp.1631–8.

Callaghan, M.J. et al., 2008. The effect of immune selection on the structure of the meningococcal opa protein repertoire. D. S. Guttman, ed. *PLoS Pathogens*, 4(3), p.e1000020.

Cane, P.A. & Pringle, C.R., 1995. Evolution of subgroup A respiratory syncytial virus: evidence for progressive accumulation of amino acid changes in the attachment protein. *Journal of Virology*, 69(5), pp.2918–25.

Carrat, F. et al., 2008. Time Lines of Infection and Disease in Human Influenza: A Review of Volunteer Challenge Studies. *American Journal of Epidemiology*, 167(7), pp.775–785.

Carrington, M. & O'Brien, S.J., 2003. The influence of HLA genotype on AIDS. *Annual review of medicine*, 54(1), pp.535–551.

Castillo-Chavez, C. et al., 1989. Epidemiological models with age structure, proportionate mixing, and cross-immunity. *Journal of Mathematical Biology*, 27(3), pp.233–258.

Chattopadhyay, R. et al., 2003. Plasmodium falciparum Infection Elicits Both Variant-Specific and Cross-Reactive Antibodies against Variant Surface Antigens. *Infection and Immunity*, 71(2), pp.597–604.

Chaudhury, S. et al., 2016. The biological function of antibodies induced by the RTS,S/AS01 malaria vaccine candidate is determined by their fine specificity. *Malaria Journal*, 15(1), p.301.

Chen, D.S. et al., 2011. A molecular epidemiological study of var gene diversity to characterize the reservoir of Plasmodium falciparum in humans in Africa. A. C. Gruner, ed. *PloS one*, 6(2), p.e16629.

Chen, R. & Holmes, E.C., 2006. Avian Influenza Virus Exhibits Rapid Evolutionary Dynamics. *Molecular Biology and Evolution*, 23(12), pp.2336–2341.

Chen, X. & Ishwaran, H., 2012. Random forests for genomic data analysis. *Genomics*, 99(6), pp.323–329.

Cherif, A., 2015. Mathematical analysis of a multiple strain, multi-locus-allele system for antigenically variable infectious diseases revisited. *Mathematical Biosciences*, 267, pp.24–40.

Chiba, S. et al., 1986. Protective effect of naturally acquired homotypic and

heterotypic rotavirus antibodies. *The Lancet*, 328(8504), pp.417–421.

Chua, K.B. et al., 2000. Nipah virus: a recently emergent deadly Paramyxovirus. *Science*, 288(5470).

Cobey, S. & Pascual, M., 2011. Consequences of host heterogeneity, epitope immunodominance, and immune breadth for strain competition. *Journal of Theoretical Biology*, 270(1), pp.80–7.

Corti, D. et al., 2010. Heterosubtypic neutralizing antibodies are produced by individuals immunized with a seasonal influenza vaccine. *The Journal of Clinical Investigation*, 120(5), pp.1663–73.

Cummings, D.A.T. et al., 2005. Dynamic effects of antibody-dependent enhancement on the fitness of viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 102(42), pp.15259–64.

Dalziel, B.D. et al., 2014. Contact Heterogeneity, Rather Than Transmission Efficiency, Limits the Emergence and Spread of Canine Influenza Virus, *PLoS Pathogens*, 10(10), p.e1004455.

Dawes, J.H.P. & Gog, J.R., 2002. The onset of oscillatory dynamics in models of multiple disease strains. *Journal of Mathematical Biology*, 45(6), pp.471–510.

Day, T. et al., 2006. The evolutionary emergence of pandemic influenza. *Proceedings. Biological Sciences / The Royal Society*, 273(1604), pp.2945–53.

Dejnirattisai, W. et al., 2016. Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. *Nature*, 17.

Dennis, A.F. et al., 2014. Molecular epidemiology of contemporary G2P[4] human rotaviruses cocirculating in a single U.S. community: footprints of a globally transitioning genotype. *Journal of Virology*, 88(7), pp.3789–801.

Doan, Y.H. et al., 2011. The occurrence of amino acid substitutions D96N and S242N in VP7 of emergent G2P[4] rotaviruses in Nepal in 2004-2005: a global and evolutionary perspective. *Archives of Virology*, 156(11), pp.1969–78.

Dodoo, D. et al., 2001. Antibodies to variant antigens on the surfaces of infected erythrocytes are associated with protection from malaria in Ghanaian children. *Infection and Immunity*, 69(6), pp.3713–8.

Dóró, R. et al., 2014a. Review of global rotavirus strain prevalence data from six years post vaccine licensure surveillance: is there evidence of strain selection from vaccine pressure? *Infection, Genetics and Evolution*, 28, pp.446–61.

Dóró, R. et al., 2014b. Review of global rotavirus strain prevalence data from six years post vaccine licensure surveillance: Is there evidence of strain selection from vaccine pressure? *Infection, Genetics and Evolution*, 28, pp.446–461.

Van Effelterre, T. et al., 2010. A mathematical model of the indirect effects of rotavirus vaccination. *Epidemiology and Infection*, 138(6), p.884.

Ferguson, N. & Andreasen, V., 2002. The influence of different forms of cross-protective immunity on the population dynamics of antigenically diverse pathogens. In: Castillo-Chavez C., Blower S., van den Driessche P., Kirschner

D., Yakubu AA. (eds) Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory. The IMA Volumes in Mathematics and its Applications, vol 126. Springer, New York, NY

Ferguson, N.M., Donnelly, C.A. & Anderson, R.M., 1999. Transmission dynamics and epidemiology of dengue: insights from age-stratified sero-prevalence surveys. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 354(1384), pp.757–768.

Ferguson, N.M., Galvani, A.P. & Bush, R.M., 2003. Ecological and immunological determinants of influenza evolution. *Nature*, 422(6930), pp.428–433.

Finkenstadt, B.F. & Grenfell, B.T., 2000. Time series modelling of childhood diseases : a dynamical systems approach. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 49(2), pp.187–205.

Fitch, W.M. et al., 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15), pp.7712–8.

Fitch, W.M. et al., 1991. Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 88(10), pp.4270–4.

Ford, C.B. et al., 2013. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics*, 45(7), pp.784–90.

Fowkes, F.J.I. et al., 2010. The relationship between anti-merozoite antibodies and incidence of plasmodium falciparum malaria: a systematic review and meta-analysis. *PLoS Medicine*, 7(1), p.e1000218.

Fowler, E. V. et al., 2002. Genetic diversity of the DBLα region in Plasmodium falciparum var genes among Asia-Pacific isolates. *Molecular and Biochemical Parasitology*, 120(1), pp.117–126.

Franco, M.A., Angel, J. & Greenberg, H.B., 2006. Immunity and correlates of protection for rotavirus vaccines. *Vaccine*, 24(15), pp.2718–2731.

Franco, M.A. & Greenberg, H.B., 2001. Challenges for Rotavirus Vaccines. *Virology*, 281(2), pp.153–155.

Gandon, S. et al., 2003. Imperfect vaccination: some epidemiological and evolutionary consequences. *Proceedings of the Royal Society B: Biological Sciences*, 270(1520), pp.1129–1136.

Gandon, S. et al., 2001. Imperfect vaccines and the evolution of pathogen virulence. *Nature*, 414(6865), pp.751–756.

Giha, H.A. et al., 1999. Nine-year longitudinal study of antibodies to variant antigens on the surface of plasmodium falciparum -infected erythrocytes nine-year longitudinal study of antibodies to variant antigens on the surface of plasmodium falciparum -infected erythrocytes. *Infection and Immunity*, 67(8), pp. 4092-4098.

Gog, J.R. & 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.

Gog, J.R. & Swinton, J., 2002. A status-based approach to multiple strain dynamics. *Journal of Mathematical Biology*, 44(2), pp.169–184.

Gómara, M.I. et al., 2001. Amino acid substitution within the VP7 protein of G2 rotavirus strains associated with failure to serotype. *Journal of Clinical Microbiology*, 39(10), pp.3796–8.

Gomes, M.G.M., Medley, G.F. & Nokes, D.J., 2002. On the determinants of population structure in antigenically diverse pathogens. *Proceedings. Biological Sciences / The Royal Society*, 269(1488), pp.227–33.

Gonzalez, A. et al., 2013. Evolutionary rescue: an emerging focus at the intersection between ecology and evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1610), p.20120404.

Grassly, N.C. & Fraser, C., 2006. Seasonal infectious disease epidemiology. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1600).

Greenberg, H. et al., 1983. Serological analysis of the subgroup protein of rotavirus, using monoclonal antibodies. *Infection and Immunity*, 39(1), pp.91–9.

Greenberg, H.B. & Estes, M.K., 2009. Rotaviruses: from pathogenesis to

vaccination. *Gastroenterology*, 136(6), pp.1939–51.

Grenfell, B.T., Bjørnstad, O.N. & Finkenstädt, B.F., 2002. Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs*, 72(2), pp.185–202.

Grimm, V., 1994. Mathematical models and understanding in ecology. *Ecological Modelling*, 75–76, pp.641–651.

Grimm, V. & Railsback, S.F., 2012. Pattern-oriented modelling: a “multi-scope” for predictive systems ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586), pp.298–310.

Gupta, S. et al., 1996. The maintenance of strain structure in populations of recombining infectious agents. *Nature Medicine*, 2(4), pp.437–442.

Gupta, S., Ferguson, N. & Anderson, R., 1998. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science*, 280(5365), pp.912–915.

Gupta, S. & Galvani, A., 1999. The effects of host heterogeneity on pathogen population structure. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 354(1384), pp.711–9.

Gupta, S., Swinton, J. & Anderson, R.M., 1994. Theoretical studies of the effects of heterogeneity in the parasite population on the transmission dynamics of malaria. *Proceedings. Biological Sciences / The Royal Society*, 256(1347), pp.231–8.

Hamby, S.E. & Hirst, J.D., 2008. Prediction of glycosylation sites using random forests. *BMC Bioinformatics*, 9(1), p.500.

Hartfield, M. & Alizon, S., 2015. Epidemiological feedbacks affect evolutionary emergence of pathogens. *The American Naturalist,* 183(4), pp. 105-117.

Heymann, D.L., 2004. Gastroenteritis, acute viral. In *Control of Communicable Disease Manual*. American Public Health Association, pp. 224–227.

Hill, A. V et al., 1991. Common west African HLA antigens are associated with protection from severe malaria. *Nature*, 352(6336), pp.595–600.

Hinnebusch, B.J. et al., 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.

Hiramatsu, K. et al., 2001. The emergence and evolution of methicillin-resistant Staphylococcus aureus. *Trends in Microbiology*, 9(10), pp.486–93.

Huang, P. et al., 2003. Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: identification of 4 distinct strain-specific patterns. *The Journal of Infectious Diseases*, 188(1), pp.19–31.

Huang, P. et al., 2012. Spike protein VP8* of human rotavirus recognizes histo-blood group antigens in a type-specific manner. *Journal of Virology*, 86(9), pp.4833–43.

Ina, Y. & Gojobori, T., 1994. Statistical analysis of nucleotide sequences of the

hemagglutinin gene of human influenza A viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 91(18), pp.8388–92.

Iturriza-Gomara, M., et al., 2000. Characterisation of rotavirus G9 strains isolated in the UK between 1995 and 1998. *Journal of Medical Virology*, 61(4), pp.510–517.

Iturriza-Gomara, M. et al., 2000. Molecular Epidemiology of human group A rotavirus infections in the United Kingdom between 1995 and 1998. *Journal of Clinical Microbiology.* 38(12), pp.4394–4401.

Iturriza-Gómara, M. et al., 2011. Rotavirus genotypes co-circulating in Europe between 2006 and 2009 as determined by EuroRotaNet, a pan-European collaborative strain surveillance network. *Epidemiology and Infection*, 139(6), pp.895–909.

Iturriza Gomara, M. et al., 2002. Molecular characterization of VP6 genes of human rotavirus isolates: correlation of genogroups with subgroups and evidence of independent segregation. *Journal of Virology*, 76(13), pp.6596–6601.

Jones, Z. & Linder, F., 2015. Exploratory data analysis using Random Forests. In *73rd Annual MPSA Conference*. pp. 1–31.

Keeling, M.J., Rohani, P. & Grenfell, B.T., 2001. Seasonally forced disease dynamics explored as switching between attractors. *Physica D: Nonlinear Phenomena*, 148(3), pp.317–335.

Kermack, W.O. & McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772), pp.700–721.

Kirkwood, C. et al., 2004. Rotavirus serotype G9P[8] and acute gastroenteritis outbreak in children, Northern Australia. *Emerging Infectious Diseases*, 10(9), pp.1593–600.

Koelle, K. et al., 2010. A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. *Journal of the Royal Society, Interface / the Royal Society*, 7(50), pp.1257–74.

Krause, J.C. et al., 2010. Naturally occurring human monoclonal antibodies neutralize both 1918 and 2009 pandemic influenza A (H1N1) viruses. *Journal of Virology*, 84(6), pp.3127–30.

Kryazhimskiy, S. et al., 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. & Gog, J.R., 2016. Capturing the dynamics of pathogens with many strains. *Journal of Mathematical Biology*, 72(1–2), pp.1–24.

Leroy, E.M. et al., 2005. Fruit bats as reservoirs of Ebola virus. *Nature*, 438(7068), pp.575–576.

Leshem, E. et al., 2014. Distribution of rotavirus strains and strain-specific effectiveness of the rotavirus vaccine after its introduction: a systematic review and meta-analysis. *The Lancet. Infectious diseases*, 14(9), pp.847–56.

Li, W. et al., 2005. Bats are natural reservoirs of SARS-like coronaviruses. *Science*, 310(5748).

Lin, J., Andreasen, V. & Levin, S.A., 1999. Dynamics of influenza A drift: the linear three-strain model. *Mathematical Biosciences*, 162(1–2), pp.33–51.

Lin, N. et al., 2004. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5, p.154.

Lin, Y.P. et al., 2006. Molecular epidemiology of G9 rotaviruses in Taiwan between 2000 and 2002. *Journal of Clinical Microbiology*, 44(10), pp.3686–94.

Lloyd-Smith, J.O. et al., 2005. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066), pp.355–359.

Lloyd-Smith, J.O. et al., 2013. Vacated niches, competitive release and the community ecology of pathogen eradication. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1623), p.20120150.

Lourenço, J. et al., 2017. Lineage structure of Streptococcus pneumoniae is driven by immune selection on the groEL heat-shock protein. *bioRxiv*.

Lourenço, J. et al., 2015. MANTIS: an R package that simulates multilocus models of pathogen evolution. *BMC Bioinformatics*, 16(1), p.176.

Lourenço, J. & Recker, M., 2013. Natural, persistent oscillations in a spatial multi-strain disease system with application to dengue. *PLoS Computational Biology*, 9(10), p.e1003308.

Lourenço, J. & Recker, M., 2010. Viral and epidemiological determinants of the invasion dynamics of novel dengue genotypes. *PLoS Neglected Tropical Diseases*, 4(11), p.e894.

Martinelli, M. et al., 2014. Phylogeny and population dynamics of respiratory syncytial virus (RSV) A and B. *Virus Research*, 189, pp.293–302.

Mateu, M.G. et al., 1994. Antigenic heterogeneity of a foot-and-mouth disease virus serotype in the field is mediated by very limited sequence variation at several antigenic sites. *Journal of Virology*, 68(3), pp.1407–17.

Matthijnssens, J. et al., 2008. Recommendations for the classification of group A rotaviruses using all 11 genomic RNA segments. *Archives of Virology*, 153(8), pp.1621–1629.

Matthijnssens, J. et al., 2011. Uniformity of rotavirus strain nomenclature proposed by the Rotavirus Classification Working Group (RCWG). *Archives of Virology*, 156(8), pp.1397–1413.

Matthijnssens, J. & Van Ranst, M., 2012. Genotype constellation and evolution of group A rotaviruses infecting humans. *Current Opinion in Virology*, 2(4), pp.426–33.

May, R.M. & Anderson, R.M., 1983. Epidemiology and genetics in the

coevolution of parasites and hosts. *Proceedings of the Royal Society B: Biological Sciences*, 219(1216), pp.281–313.

May, R.M., Gupta, S. & McLean, A.R., 2001. Infectious disease dynamics: what characterizes a successful invader? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1410), pp.901–910.

McDonald, S.M. et al., 2012. Diversity and relationships of cocirculating modern human rotaviruses revealed using large-scale comparative genomics. *Journal of Virology*, 86(17), pp.9148–9162.

McDonald, S.M. et al., 2016. Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nature Reviews Microbiology*, 14(7), pp.448–460.

Miles, M.G. et al., 2012. A systematic review of rotavirus strain diversity in India, Bangladesh, and Pakistan. *Vaccine*, 30 Suppl 1, pp.A131-9.

Minayev, P. & Ferguson, N., 2009. Improving the realism of deterministic multi-strain models: implications for modelling influenza A. *Journal of the Royal Society, Interface / the Royal Society*, 6(35), pp.509–18.

Mullick, S. et al., 2014. Community based case-control study of rotavirus gastroenteritis among young children during 2008-2010 reveals vast genetic diversity and increased prevalence of G9 strains in Kolkata. *PloS one*, 9(11), p.e112970.

Nelson, M.I. & Holmes, E.C., 2007. The evolution of epidemic influenza. *Nature*

Reviews Genetics, 8(3), pp.196–205.

Newbold, C.I. et al., 1992. Plasmodium falciparum: The human agglutinating antibody response to the infected red cell surface is predominantly variant specific. *Experimental Parasitology*, 75(3), pp.281–292.

Nordgren, J. et al., 2014. Both lewis and secretor status mediate susceptibility to rotavirus infections in a rotavirus genotype-dependent manner. *Clinical Infectious Diseases*, 59(11), pp.1567–73.

Nowak, M.A. & May, R.M., 1991. Mathematical biology of HIV infections: antigenic variation and diversity threshold. *Mathematical Biosciences*, 106(1), pp.1–21.

O'Ryan, M. et al., 1994. Anti-Rotavirus G type-specific and isotype-specific antibodies in children with natural rotavirus infections. *Journal of Infectious Diseases*, 169, pp.504–11.

Ockenhouse, C.F. et al., 2015. Ad35.CS.01 - RTS,S/AS01 heterologous prime boost vaccine efficacy against sporozoite challenge in healthy malaria-naïve adults. *PLoS ONE*, 10(7), p.e0131571.

Offit, P.A. & Blavat, G., 1986. Identification of the two rotavirus genes determining neutralization specificities. *Journal of Virology*, 57(1), pp.376–8.

Ofori, M.F. et al, 2002. Malaria-induced acquisition of antibodies to plasmodium falciparum variant surface antigens. *Infection and Immunity*, 70(6), pp.2982–2988.

Okuno, Y. et al., 1993. A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. *Journal of Virology*, 67(5), pp.2552–8.

Oliveira, S. et al., 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *Forest Ecology and Management*, 275, pp.117–129.

Page, N.A. & Steele, A.D., 2004. Antigenic and genetic characterization of serotype G2 human rotavirus strains from South Africa from 1984 to 1998. *Journal of Medical Virology*, 72(2), pp.320–7.

Palombo, E.A., Bishop, R.F. & Cotton, R.G., 1993. Intra- and inter-season genetic variability in the VP7 gene of serotype 1 (monotype 1 a) rotavirus clinical isolates. *Archives of Virology*, 130(1–2), pp.57–69.

Parham, P.E. & Michael, E., 2011. Outbreak properties of epidemic models: The roles of temporal forcing and stochasticity on pathogen invasion dynamics. *Journal of Theoretical Biology*, 271(1), pp.1–9.

Payne, D.C. et al., 2011. Direct and indirect effects of rotavirus vaccination upon childhood hospitalizations in 3 US Counties, 2006-2009. *Clinical Infectious Diseases*, 53(3), pp.245–53.

Pearson, R.G. et al., 2014. Life history and spatial traits predict extinction risk due to climate change. *Nature Climate Change*, 4(3), pp.217–221.

Penman, B.S. et al., 2013. Pathogen selection drives nonoverlapping

associations between HLA loci. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48), pp.19645–50.

Pitzer, V.E. et al., 2009. Demographic variability, vaccination, and the spatiotemporal dynamics of rotavirus epidemics. *Science*, 325(5938), pp.290–294.

Pitzer, V.E. et al., 2011a. Influence of birth rates and transmission rates on the global seasonality of rotavirus incidence. *Journal of The Royal Society Interface*, 8(64), pp.1584–1593.

Pitzer, V.E. et al., 2011b. Modeling rotavirus strain dynamics in developed countries to understand the potential impact of vaccination on genotype distributions. *Proceedings of the National Academy of Sciences of the United States of America*, 108(48), pp.19353–8.

Polley, S.D. et al., 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.279–87.

Purse, B. V. et al., 2005. Opinion: Climate change and the recent emergence of bluetongue in Europe. *Nature Reviews Microbiology*, 3(2), pp.171–181.

Read, J.M. & Keeling, M.J., 2007. Stochasticity generates an evolutionary instability for infectious disease. *Ecology Letters*, 10(9), pp.818–827.

Recker, M. et al., 2009. Immunological serotype interactions and their effect on the epidemiological pattern of dengue. *Proceedings of the Royal Society of*

London B: Biological Sciences, 276(1667).

Recker, M. et al., 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.7711–6.

Recker, M. et al., 2004. Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. *Nature*, 429(6991), pp.555–8.

Recker, M. & Gupta, S., 2005. A model for pathogen population structure with cross-protection depending on the extent of overlap in antigenic variant repertoires. *Journal of Theoretical Biology*, 232(3), pp.363–373.

Remarque, E.J. et al., 2008. Apical membrane antigen 1: a malaria vaccine candidate in review. *Trends in Parasitology*, 24(2), pp.74–84.

Rodrigues, F. et al., 2013. The evolving epidemiology of rotavirus gastroenteritis in central Portugal with modest vaccine coverage. *Journal of Clinical Virology*, 56(2), pp.129–134.

Rohani, P. et al., 2003. Ecological interference between fatal diseases. *Nature*, 422(6934), pp.885–888.

Rohani, P., Keeling, M.J. & Grenfell, B.T., 2002. The interplay between determinism and stochasticity in childhood diseases. *The American Naturalist*, 159(5), pp.469–81.

Rota, P.A. et al., 2011. Global distribution of measles genotypes and measles

molecular epidemiology. *The Journal of Infectious Diseases*, 204 Suppl(suppl_1), pp.S514-23.

Ruuska, T., Vesikari, T. & Ruuska, T., 1990. Rotavirus disease in Finnish children: use of numerical scores for clinical severity of diarrhoeal episodes. *Scandinavian Journal of Infectious Diseases*, 223(22), pp.259–267.

Santos, N. & Hoshino, Y., 2005. Global distribution of rotavirus serotypes/genotypes and its implication for the development and implementation of an effective rotavirus vaccine. *Reviews in Medical Virology*, 15(1), pp.29–56.

Smith, D.J., 2004. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682), pp.371–376.

Smith, G.J.D. et al., 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459(7250), pp.1122–1125.

Smith, J.D. et al., 1995. Switches in expression of plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell*, 82(1), pp.101–110.

Su, X. et al., 1995. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. *Cell*, 82(1), pp.89–100.

Subramaniam, S. et al., 2013. Phylogeny and genetic diversity of foot and mouth disease virus serotype Asia1 in India during 1964-2012. *Veterinary*

Microbiology, 167(3–4), pp.280–8.

Sui, J. et al., 2011. Wide prevalence of heterosubtypic broadly neutralizing human anti-influenza A antibodies. *Clinical Infectious Diseases*, 52(8), pp.1003–9.

Sutton, K.L., Banks, H.T. & Castillo-Chávez, C., 2007. Estimation of invasive pneumococcal disease dynamics parameters and the impact of conjugate vaccination in australia. *Mathematical Biosciences and Engineering,* 5(1), pp. 175-204.

Tami, A. et al., 2003. Sympatric Plasmodium falciparum isolates from Venezuela have structured var gene repertoires. *Malaria Journal*, 2(1), p.7.

Tate, J.E. et al., 2016. Global, regional, and national estimates of rotavirus mortality in children. *Clinical Infectious Diseases*, 62, pp.96–105.

Taylor, L.H., Latham, S.M. & Woolhouse, M.E., 2001. Risk factors for human disease emergence. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 356(1411), pp.983–9.

Trepo, C., Chan, H.L.Y. & Lok, A., 2014. Hepatitis B virus infection. *The Lancet*, 384(9959), pp.2053–2063.

Urwin, R. et al., 2004. Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. *Infection and Immunity*, 72(10), pp.5955–62.

Velázquez, F.R., 2009. Protective effects of natural rotavirus infection. *The*

Pediatric Infectious Disease Journal, 28(Supplement), pp.S54–S56.

Velázquez, F.R. et al., 1996. Rotavirus infections in infants as protection against subsequent infections. *The New England Journal of Medicine*, 335(14), pp.1022–8.

Velázquez, F.R. et al., 2000. Serum antibody as a marker of protection against natural rotavirus infection and disease. *The Journal of Infectious Diseases*, 182(6), pp.1602–1609.

Vesikari, T. et al., 2007. Efficacy of human rotavirus vaccine against rotavirus gastroenteritis during the first 2 years of life in European infants: randomised, double-blind controlled study. *The Lancet*, 370(9601), pp.1757–1763.

Vesikari, T. et al., 2006. Safety and Efficacy of a Pentavalent Human–Bovine (WC3) Reassortant Rotavirus Vaccine. *New England Journal of Medicine*, 354(1), pp.23–33.

Wang, T.T. et al., 2010. Broadly protective monoclonal antibodies against H3 influenza viruses following sequential immunization with different hemagglutinins. *PLoS Pathogens*, 6(2), p.e1000796.

Watkins, E.R. et al., 2015. Vaccination drives changes in metabolic and virulence profiles of Streptococcus pneumoniae. *PLoS Pathogens*, 11(7), p.e1005034.

Watkins, E.R. & 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.

Webster, R.G. et al., 1992. Evolution and ecology of influenza A viruses. *Microbiological reviews*, 56(1), pp.152–79.

Weinberger, D.M., Malley, R. & Lipsitch, M., 2011. Serotype replacement in disease after pneumococcal vaccination. *Lancet*, 378(9807), pp.1962–73.

White, M.T. et al., 2015. Immunogenicity of the RTS,S/AS01 malaria vaccine and implications for duration of vaccine efficacy: secondary analysis of data from a phase 3 randomised controlled trial. *The Lancet Infectious Diseases*, 15(12), pp.1450–1458.

Wikramaratna, P.S. et al., 2015. Effects of neutralizing antibodies on escape from CD8+ T-cell responses in HIV-1 infection. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 370(1675).

Wikramaratna, P.S. et al., 2013. The antigenic evolution of influenza: drift or thrift? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1614), p.20120200.

Wikramaratna, P.S. et al., 2010. The effects of tertiary and quaternary infections on the epidemiology of dengue. *PLoS ONE*, 5(8), p.e12347.

Wikramaratna, P.S., Pybus, O.G. & Gupta, S., 2014. Contact between bird species of different lifespans can promote the emergence of highly pathogenic avian influenza strains. *Proceedings of the National Academy of Sciences of the United States of America*, 11(29), pp. 10767-10772.

Wolf, Y.I. et al., 2006. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct*, 1, p.34.

Yamashita, M. et al., 1988. Influenza B virus evolution: Co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. *Virology*, 163(1), pp.112–122.

Yates, A. et al., 2006. How do pathogen evolution and host heterogeneity interact in disease emergence? *Proceedings. Biological sciences / The Royal Society*, 273(1605), pp.3075–83.

Zaman, K. et al., 2010. Efficacy of pentavalent rotavirus vaccine against severe rotavirus gastroenteritis in infants in developing countries in Asia: a randomised, double-blind, placebo-controlled trial. *Lancet (London, England)*, 376(9741), pp.615–23.

Zeller, M. et al., 2015. Genome-wide evolutionary analyses of G1P[8] strains isolated before and after rotavirus vaccine introduction. *Genome Biology and Evolution*, 7(9), pp.2473–83.

Zhang, C. et al., 2005. Clade replacements in dengue virus serotypes 1 and 3 are associated with changing serotype prevalence. *Journal of Virology*, 79(24), pp.15123–30.

Zlateva, K.T. et al., 2005. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. *Journal of Virology*, 79(14), pp.9157–67.