# The Evolution of the Major Histocompatibility Complex

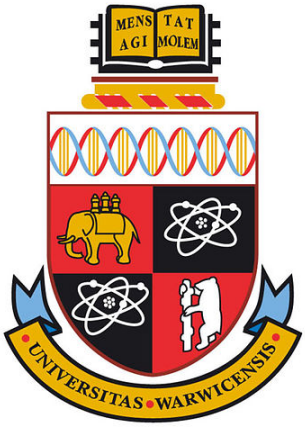
by

**Connor White**

Thesis submitted to the University of Warwick

for the degree of

**Doctor of Philosophy in Mathematics of Systems**




Warwick University, Department of Mathematics, Mathematics for Real World Systems

August 2022

# Contents

## Acknowledgements

I want to thank my Supervisor Bridget Penman who even though was on her maternity leave, gave me all the supervision I needed to finish this PhD. I thank all my friends and colleagues whether they knew it or not provided moral support. I especially thank my wife Prim who has been there for me through all the tough times.

This work would also not be possible without the funding from EPSRC.

## DECLARATION

The work presented here is my own, except where stated otherwise. This thesis has been composed by myself and has not been submitted for any other degree or professional qualification.

Chapter 2 and Appendix A have been published as:

White C, Pellis L, Keeling M and Penman B. Detecting HLA-infectious disease associations for multi-strain pathogens, Infect Genet Evol 2020, 83:104344

**Abstract**

Here I present theoretical work that explores the mechanisms underlying, and epidemiological consequences of, Major Histocompatibility Complex (MHC) genetic diversity. MHC genes encode molecules which present peptides for recognition by T cells, and are essential to the immune system of vertebrates. I explore three aspects of MHC genetic diversity and evolution:

1. I created an epidemiological model that takes into account a host population's MHC genotype. I modelled the infectious disease dynamics of a multi strain pathogen infecting that host population, in order to investigate how MHC allele frequencies affect the pathogen climate. I investigated a case control study carried out in that simulated population, in order to investigate the protectiveness of MHC variants. I found that the apparent protection against infection a specific MHC allele confers to a host is inversely proportional with the frequency of said allele in the population.

2. I created an individual based model that models the number of repeated MHC genes on chromosomes within a diploid population (i.e. copy number variation, or CNV). I simulated unequal crossing over recombination to generate repeated copies of genes, and also allowed mutations to occur which varied the properties of each gene. I tested different rules of fitness for MHC genotype and demonstrated a variety of evolutionary outcomes in terms of CNV for the MHC. I found that if the fitness of the host is equal to the mean of the possible fitness contributions of their MHC genes, then the number of copies of MHC genes is inversely related to the intensity of changing pathogen selection

3. I created an individual based model that allowed the promiscuity of MHC molecules (i.e. their MHC peptide binding repertoire) to vary. I show how pathogen climates may shape the pattern of MHC promiscuity present in a host population.

# Chapter 1

# Introduction

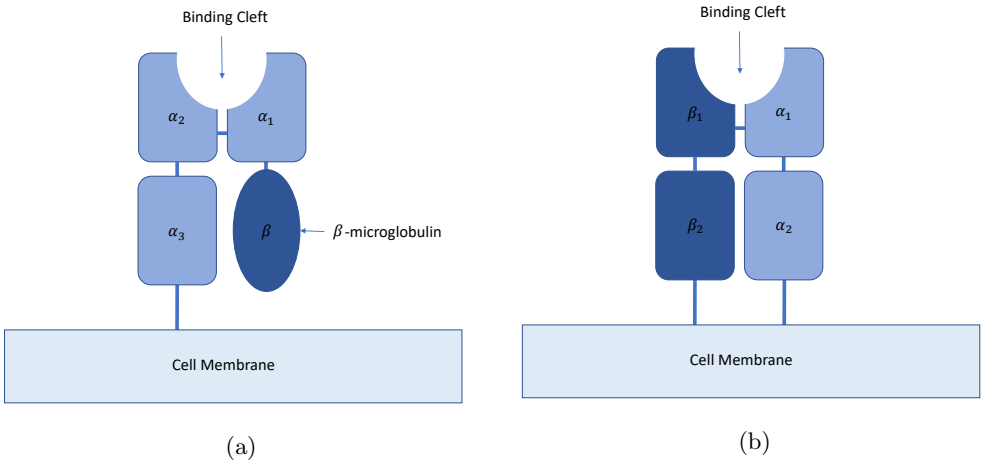
This thesis considers evolutionary and epidemiological questions relating to immune system gene diversity, specifically Major Histocompatibility Complex (MHC) diversity. In this introductory chapter, I provide a background of previous research about the MHC from both experimental and theoretical backgrounds. In chapter 2, I investigate the dynamics of a multi strain pathogen infecting an immunogenetically diverse host population, and show how these dynamics can cause difficulties in detecting the true impact of MHC genes on infection. In chapter 3 I investigate why copy number variation (CNV) of the MHC may have evolved differently for different vertebrate species. Chapter 4 concerns why MHC molecules in some species have evolved to be very fastidious or promiscuous. I explore these research questions using mathematical modelling and computational simulations.

## 1.1 The Major Histocompatibility Complex

### 1.1.1 Background

MHC molecules are found on the surface of all nucleated cells in vertebrates. They play a vital role in the adaptive immune system (Klein et al. [2007]). The main function of the MHC molecule is to present peptides to T-cells to test for foreign peptides. MHC molecules hold peptides by binding peptides to a part of the MHC molecule called a

binding cleft. MHCs are divided into two classes the MHC class I and the MHC class II. MHC class I molecules are expressed on all nucleated cells. MHC class II molecules are expressed only on the surface of antigen-presenting cells (macrophages, dendritic cells, and B cells) (Abbas et al. [2014]). MHC class I molecules are heterodimers that consist of two polypeptide chains $\alpha$ and $\beta$-microglobulin. In figure 1.1a we see that the binding cleft is formed from different parts of the $\alpha$ chain. MHC class II molecules are also heterodimers and consist of two peptide chains an $\alpha$ and $\beta$ chain. The MHC class II binding cleft is formed from both the $\alpha$ and $\beta$ chain.




(a) (b)

Figure 1.1: **Diagram showing the peptide chain layouts of MHC molecules**. (a) is of a MHC class I molecule and (b) is of a MHC class II molecule.

The shape of the binding cleft determines the shape of peptide an MHC molecule can hold. This aspect is extremely important in the recognition of foreign peptides from pathogens as the shape of binding cleft an individual has (which is determined by the MHC alleles they have) might affect whether or not that individual can present peptides from a particular pathogen.

The main function of MHC molecules is to present peptides to T-Cells. Class I MHC molecules present peptides from inside the cell. These peptides are representations of what is inside the cell. CD8+ T-cells express a T-cell receptor (TCR) which can recognise a specific antigen bound by a class I MHC molecule. If the TCR binds strongly to the antigen and the peptide held by the MHC class I molecule is then recognised as non

self the CD8+ T-cells then destroys the cell presenting the non self peptide. The CD8+ T-cell destroys the infected cell by secreting cytotoxins which will eventually lead the cell to apoptosis which is called programmed cell death. It then replicates multiple times to produce more CD8+ T-cells to detect and destroy further infected cells around the currently found infected cell. Once the infection has been dealt with, memory T cells are produced that will recognise the peptide from the pathogen that has been previously destroyed.

Class II MHC molecules present antigens that are derived from extracellular proteins. As class II MHC present extracellular peptides they are mainly concerned with helping to detect extracellular pathogens. For MHC class II molecules, it is CD4+ T-cells that the MHC molecules present peptides to. If a CD4+ T-cell recognises the peptide from an MHC class II molecule it will release cytokines which help to polarize the immune response into the appropriate kind. For instance these cytokines could help a B cell secrete antibodies or or macrophages to destroy ingested microbes.

A T-Cell recognises a peptide as non-self if its receptor binds strongly to the peptide. The shape of the receptor determines whether or not it can bind to a particular molecular motif. As there are many peptide shapes that are generated that are non self, T-Cells need to cover a wide range of receptor shapes, which in turn means T-Cell receptors need to be generated with a wide range of diversity. This is achieved through a process called somatic recombination, and a single individual has the possibility to generate T cell receptors to bind to theoretically any molecular shape. However these non self peptides need to be displayed in order for T-Cells to even attempt to recognise them. As mentioned this is the main function of MHC molecules and in particular it is an MHC molecules binding cleft which determines if a peptide can be displayed or not. As an individual is limited in what MHC genes they have, they are limited in the peptides that can be displayed, thus the MHC acts as a bottleneck in terms of how effective our T-cell responses can be.

### 1.1.2 MHC diversity

Genes encoding MHC molecules are found within the MHC region, which in humans is a 3Mbp stretch within chromosome 6. MHC genes are some of the most polymorphic gene families known. For instance the MHC for humans, known as the Human Leukocyte

Antigen (HLA) has three major class I genes: HLA-A, HLA-B and HLA-C. As of 2020 the recorded number of HLA-A alleles are 5266, HLA-B 6537 and HLA-C 5140 (Robinson et al. [2020]).

Despite this large polymorphism of class I HLA alleles, it has been argued that the functionality of these alleles may have large crossovers and the number of uniquely functioning HLA class I types is much smaller. Studies looking into the binding properties of MHC molecules that are encoded by HLA-A and HLA-B alleles have shown that for the majority of tested alleles there are common shared peptide binding motifs. From this they have defined 10 groups that HLA-A and HLA-B alleles can fall into (in terms of their binding peptide repertoire) called supertypes (Sidney et al. [2008]) which are defined around shared peptide motifs between the HLA alleles.

The MHC between different species all share the fact that it is generally one of the most polymorphic gene regions on a species genome. However an aspect of the MHC that may differ between species is the number of copies of a particular MHC gene and more particularly how that number may vary. The number of copies of a gene and how this number varies within a population is called "copy number variation" (CNV) (Freeman et al. [2006]).

CNV of the MHC is different between species. More copies of an MHC gene could potentially mean more pathogen peptides that could be presented to the immune system. However if we see varying numbers of copies of MHC genes it could suggest there is a cost to possessing too many genes. The selective forces that govern how many copies of an MHC gene that are present in a genome are not very well understood. If we look at the HLA-A gene, all humans have only one copy of this gene therefore there is no CNV for HLA-A. The equivalent gene in Indian Rhesus macaques (mamu-A) has 4 copies and the number of copies of the mamu-A gene varies between individuals of that population (Otting et al. [2007]). Cynomolgus macaque HLA-A and HLA-B gene the mafa-A and mafa-B genes also have CNV (Otting et al. [2007], Wiseman et al. [2013]) where there are up to 5 copies of mafa-A genes. If we look at HLA-B only one copy exists among humans however over 18 functional Mamu-B-like genes exist (Daza-Vamenta et al. [2004]). Comparing the configurations of mamu-A with cynomolgus macaque mafa-A Otting et al. [2007] shows that Mamu-A haplotypes or gene copy structures pre date speciation of the

two macaque species (the rhesus macaque, *Macaca mulatta* and the cynomolgus macaque *Macaca fascicularis*). This suggests haplotypes containing variable numbers of Mamu-A or Mafa-A have been around for a long time and potentially are evolutionarily stable.

MHC alleles can vary in their promiscuity in terms of their peptide binding repertoire. What this means is that some MHC alleles encode MHC molecules that are capable of binding to more varied peptide shapes than others. For example the most promiscuous HLA alleles are ones that belong to the HLA-A2 supertype (Kaufman [2020]), these molecules can accommodate two or three different amino acids in each pocket of its binding cleft (Madden et al. [1993], Chen et al. [2012]). In contrast HLA-B*57:01 which is considered fastidious for humans, has a pocket requiring a rare amino acid tryptophan (Illing et al. [2012]). A nice example of predicted binding repertoires and how they vary is shown in figure 1A of Paul et al. [2013] and in general Paul et al. [2013] shows evidence of HLA alleles varying in binding peptide repertoire. For chickens one of their most promiscuous MHC alleles is BF2*02:01 (Chappell et al. [2015]) which has binding cleft pockets that can hold six different amino acids. BF2*04:01 is a highly fastidious chicken MHC allele which requires binding of rare amino acids in each of three pockets (Wallny et al. [2006], Zhang et al. [2012]). Chickens compared to humans have the more extreme alleles on the spectrum of fastidious to promiscuous MHC alleles. MHC molecule promiscuity has had very few studies in depth in other species (Kaufman [2020]).

For humans binding peptide repertoire is not necessarily related to the defined supertype group. As mentioned earlier some of the most promiscuous HLA alleles belong to the A2 supergroup (Kaufman [2020]). Kaufman [2020] points out that it is only this supertype group (A2) that correlates to the promiscuity of HLA alleles and in fact the peptide binding range of HLA alleles is not necessarily related to the supertype group.

### 1.1.3 Pathogen selection on the MHC

Given the nature of MHC genes' role in adaptive immunity it is believed that natural selection from pathogens has impacted their evolution. (Spurgin and Richardson [2010], Jeffery and Bangham [2000], Hedrick [2002], Pierini and Lenz [2018], Prugnolle et al. [2005]).

It has been shown for a wide range of infectious diseases that HLA genotype can

influence disease outcome (Hill et al. [1991], Tian et al. [2017], Dunstan et al. [2014], McLaren et al. [2015], Sveinbjornsson et al. [2016], Thursz et al. [1995], Oliveira-Cortez et al. [2016], Carrington and O'Brien [2003]), more details in section 2.1. However, even with the numerous works on finding associations between HLA alleles and pathogens it has classically been very difficult to pin down the exact roles of different HLA types, with associations being very weak or even contradictory between case control studies on different populations.

While MHC allele associations with infectious disease outcomes are suggestive of pathogen involvement in MHC evolution, direct evidence for pathogen selection acting on the MHC is hard to come by. When discussing the potential types of selection on the MHC generally people are concerned with the type of selection that has maintained the high level of MHC polymorphism. The difficulties of finding whether or not pathogen mediated selection is the cause for maintaining MHC polymorphism is due to disentangling the effects of potentially other selective forces (e.g mating selection).

Balancing selection, meaning selective processes which maintain multiple alleles, seems to be acting on the MHC over purifying selection given the fact that the MHC is very polymorphic. Not only is it polymorphic but frequencies of individuals who are homozygous are lower than expected from under neutrality theory (Hedrick and Thomson [1983]). Further research on MHC diversity has given evidence of balancing selection (Aguilar et al. [2004], Hawley and Fleischer [2012]). Hawley and Fleischer [2012] found sequence and diversity based signatures of pathogen mediated balancing selection on a population of house finches exposed to an epidemic of Mycoplasmal Conjunctivitis which were not found in unexposed populations of house finches. Diversity studies have also shown that balancing selection seems to only occur on the allelic level and not on the level of MHC haplotypes (Alter et al. [2017]). Evidence of balancing selection does not uncover the actual mechanisms driving diversity, as balancing selection itself is a broad term and can be reduced to more fundamental mechanisms. Three main mechanisms could be maintaining the allelic diversity of (and hence balancing selection on) MHC loci: Heterozygote advantage, negative frequency dependent fitness and fluctuating selection (Spurgin and Richardson [2010]). The vast majority of theoretical work on MHC evolution is around the question of which of these mechanisms (or combinations of these mechanisms) is most

important, as I discuss in section 1.1.5.

The heterozygote advantage proposes that an individual who has two different MHC alleles at a single locus will be protected from more pathogens than an individual who is homozygous (has two copies of the same MHC allele at that locus), since more peptide shapes can be presented by the two different MHC molecules. It follows that individuals who are heterozygous would on average have an increase in fitness compared to homozygous individuals. If heterozygous individuals are surviving longer then there will be more MHC alleles within a population due to this. The heterozygote advantage can be referred as overdominant selection which means that the phenotype of a heterzygote differs to that of homozygotes. Doherty and Zinkernagel [1975] used in vivo and in vitro experiments on mice to showcase that immune responsiveness to lymphocytic choriomeningitis is enhanced with heterozygous individuals at H-2 genes (the mouse version of MHCs). However this specific experimental system does not show heterozygote advantage as increased immune responses to lymphocytic choriomeningitis actually could be lethal to the mice. Hughes and Nei [1988] analysed the pattern of nucleotide substitutions between polymorphic alleles. They find that nonsynonymous substitutions are far more frequent than synonymous substitutions in the binding cleft region of MHC alleles. However the opposite is true outside of this binding cleft region. Using these results and relying on a theoretical prediction from Maruyama and Nei [1981] (that in the presence of overdominant selection the rate of gene substitution should be more than that of neutral alleles) they reject other mechanisms for causing polymorphism and suggest overdominant selection, specifically heterozygote advantage to be the main cause for MHC polymorphism. However heterozygote disadvantage has been shown for mice, Ilmonen et al. [2007] infected mice with different strains of Salmonella. They find that resistance is recessive and that females infected and who were heterozygous produced less pups than homozygous females.

Frequency dependent selection is the idea that a pathogen may evolve to avoid the majority MHC genotype binding cleft peptide repertoire, whilst rarer MHC molecules may still be able to display peptides from an adapted pathogen climate (Slade and McCallum [1992]). Frequency dependent selection thus implies the existence of host-pathogen coevolution, since it would not occur if the pathogen were not adapting to the host. Lange-

fors et al. [2001] find associations between MHC alleles and resistance/susceptibility to furunculosis for Atlantic Salmon. They do not find any association between being heterozygous and being resistant/susceptible to the disease hence they suggest their results indicate that frequency dependence has maintained polymorphism for the MHC in Atlantic Salmon. Frequency dependent selection has been tackled theoretically numerous times and is discussed later in section 1.1.5).

Fluctuating selection in terms of pathogens is the idea that pathogen spatial and temporal climates may change (fluctuate), which may occur even in the absence of pathogens adapting to their hosts, causing the selection on MHC's to fluctuate in different spaces/ or time (Spurgin and Richardson [2010]). With multiple pathogens fluctuating in presence or not causing fluctuating selection on the MHC could maintain diversity of the MHC. Hedrick [2002] creates a model that allows the number of pathogens present in a population to vary temporally. They show this fluctuation of the number of pathogens could maintain diversity and that intrinsic heterozygote advantage is unnecessary. Hedrick [2002] also points out in the model that when multiple pathogens were present in a given generation that heterozygotes would have a higher fitness than homozygotes, without this needing to be a specific assumption of the model.

### 1.1.4 Mate choice and the MHC

An alternative hypothesis to pathogen selection is that mate choice has driven MHC diversity. The earliest evidence of such selection was found by Yamazaki et al. [1976]. Yamazaki et al. [1976] tested male mice mating choices when presented with two females of differing H-2 genotypes (Mice MHC). They found that statistically significantly more males chose to mate with female mice who have a differing H-2 type. Mating choices like this would lead to larger amounts of heterozygous genotypes of the H-2 gene. Later indirect evidence of MHC dependent sexual selection for house mice was found (Potts et al. [1991]). Potts et al. [1991] found significantly fewer MHC homozygous offspring than would be expected from random mating and no evidence of for abortional selection or nonrandom fertilisation.

Hughes and Hughes [1995] suggests that the above results could come from mate choices avoiding inbreeding and that experimental studies would need to be done in

order to account for the relatedness of mice. Direct experimental evidence of MHC-disassortative mating preferences have been mixed. Experimental evidence for MHC-dependent sexual selection has been found for both male labarotory mice and wild-derived female mice (Beauchamp et al. [1988], Yamazaki et al. [1988], Eklund [1997]). Other studies have found no evidence for MHC sexual selection (Eklund et al. [1991], Beauchamp et al. [1988], Manning et al. [1992]). More detailed reviews of MHC-sexual selection evidence exist (Penn [2002], Penn and Potts [1999]).

More recent work Schubert et al. [2021] review the role of the MHC in sexual selection such as the affect the MHC has on odour. Huang et al. [2021] use Monte Carlo simulations to investigate MHC non random mating along with diplotyped sheep population. Huang et al. [2021] find evidence of sexual selection against certain haplotypes of the MHC, but offer other suggestions for why this might be occurring for instance the avoidance of inbreeding.

The vast majority of MHC sexual selection studies have been on mice. Some studies found evidence for humans that MHC-disassortative odour and mating preferences, however evidence is mixed. MHC disassortative odour preferences were found among Swiss students (Wedekind et al. [1995], Wedekind and Füri [1997]). MHC associated marriage preferences were found in Hutterites in isolated populations of North America (Ober et al. [1997]). However other studies found no evidence for MHC associated marriage preferences in Amerindians and Japanese couples (Hedrick and Black [1997], Ihara et al. [2000]).

### 1.1.5 Theoretical models for MHC Polymorphism

The vast majority of theoretical work has been trying to disentangle the relative effects of different possible mechanisms which could maintain MHC polymorphism. Early work by Hughes and Nei [1988] used theoretical models from Maruyama and Nei [1981] which is a stochastic differential equation model applied to population genetics. While Maruyama and Nei [1981] is not directly applied to the MHC it extremely relevant. They explore how overdominant genes differ from neutral genes in terms of genetic polymorphism and heterozygosity. They show that overdominant selection is a strong mechanism for maintaining polymorphism.

Takahata and Nei [1990] is another early example of theoretical work applied directly to the MHC. They use an individual based model of a diploid population to find mechanisms to explain long-term persistence of MHC alleles using various forms of selection. They test neutral mutation, overdominant selection and frequency dependent selection. They showcase that both over-dominant selection and frequency dependent selection could maintain high levels of polymorphism, however they doubt that a rare allele advantage should exist in reality and suggest overdominant selection to be the more realistic mechanism, although they admit that experimental data would needed to be used to show this. Takahata and Nei [1990] reasoning for disbelieving in rare allele advantage for an MHC allele is very subjective and is based on a argument that an allele that becomes rarer after being dominant in a population should not gain advantage again due to pathogens having already adapted to it.

De Boer et al. [2004] create another individual based model to test the overdominant/heterozygote advantage, however they differ from Takahata and Nei [1990] due to the fact that fitness of individuals depends on the MHC alleles they have and not just only on whether they are heterozygous. Takahata and Nei [1990] gave all individuals who were homozygous the same fitness and all individuals who were heterozygous the same fitness. De Boer et al. [2004] shows here that if each allele is given individual values of fitness, that the heterozygote advantage can only maintain high levels of polymorphism if the fitness values of the MHC alleles are of extremely similar value, which of course was not a problem in Takahata and Nei [1990].

Some theoretical works also incorporate pathogen genetics into their models such as Stoffels and Spencer [2008]. They define MHC alleles in a population by the set of pathogens that an MHC molecule can recognise. Stoffels and Spencer [2008] vary the amount of overlap MHC alleles have in their recognition of pathogens. They also define an advantage an individual can have if both their MHC alleles can recognise a pathogen, which they call "Intersection advantage". They found that the more overlapping MHC molecules had in terms of pathogen recognition the more polymorphic their simulated population was. They also show that increasing intersection advantage reduced polymorphism within their model and concluded that if intersection advantage at the MHC is important for an individual's fitness then heterozygote advantage may fail to explain the

high levels of polymorphism.

Siljestam and Rueffler [2019] produces a model that does not give a direct fitness to alleles or hosts and simulates pathogens to give hosts inherent fitness. Siljestam and Rueffler [2019] produced a model where each allele is given a number of attributes which is a number between 0 and 1. These attributes are matched by attributes that a pathogen also has. The closer an allele's attributes are to a pathogen's the more chance that host has of surviving and passing on its genes. With a static pathogen climate they show a host population evolves to have and maintain high levels of alleles. These systems self-organise into generating lots of alleles that specialise around the multiple pathogens within the system.

Table 1.1 summarises key theoretical models of MHC polymorphism. The majority of these use individual based models and test the heterozygote advantage. The majority do not actually include any type of co evolving system (both host and pathogen climate changing). The conclusion of these works all depend on the assumptions made in the model and no clear consensus is reached as time has moved on.

| Study | Type of model | Heterozygote advantage | Negative frequency dependence | Fluctuating selection | Coevolution | Conclusion |
|---|---|---|---|---|---|---|
| Hughes and Nei [1988] Maruyama and Nei [1981] | stochastic differential equation | yes | no | no | no | heterozygote advantage |
| Takahata and Nei [1990] | individual based | yes | yes | no | no | heterozygote advantage |
| Hedrick [2002] | stochastic model | yes | no | yes | no | fluctuating selection |
| De Boer et al. [2004] | individual based | yes | no | no | no | not heterozygote advantage |
| Borghans et al. [2004] | individual based | yes | yes | no | yes | negative frequency dependence |
| Stoffels and Spencer [2008] | individual based | yes | no | no | no | not heterozygote advantage |
| Siljestam and Rueffler [2019] | individual based | yes | no | no | no | heterozygote advantage |

Table 1.1: **Table that categorises literature that theoretically explores mechanisms to explain MHC allelic polymorphism.** Here I list descriptions of the headers in the table. Study, is a reference to the work of interest. Type of model, refers to the mathematical tool or modelling technique used in the theoretical work. Heterozygote advantage, refers to whether or not the work tested the effects of the heterozygote advantage. Negative frequency dependence, refers to whether or not the work tested the effects of negative frequency dependence. Fluctuating selection, refers to whether or not the work tested the effects of fluctuating selection. Coevolution, refers to whether or not the work included not only a host population evolving but also a pathogen climate or varying selection that changed according to the populations genetic landscape. Conclusion, refers to the main mechanism the work suggests causes MHC allelic polymorphism.

### 1.1.6 Other works related to the MHC

The majority of theoretical work about the MHC investigates the main mechanisms maintaining the high polymorphism of individual MHC loci, however there do exist other theoretical works that try to tackle other questions about the MHC or even the MHC's effect on pathogen epidemiology. Gupta and Hill [1995] was one of the first theoretical works that used an epidemiological model and applied it to questions about the HLA. Gupta and Hill [1995] looks at the possible effect HLA type has on the epidemiology of infection in terms of maintenance of pathogen diversity. Gupta and Hill [1995] use an Ordinary differential equation (ODE) model which is commonly used in epidemiology. This work is a multi host and multiple pathogen strain model where the hosts are defined around the HLA genotype. They show that under certain conditions host heterogeneity in terms of resistance can maintain pathogen virulence.

Sambaturu et al. [2018] is another work that theoretically explores the effects of HLA genotype on the epidemics of a pathogen. It explored the effect HLA genotype has on the $R_0$ of H1N1 influenza for different populations. Sambaturu et al. [2018] classifies HLA type on the number of influenza epitopes an MHC molecule that HLA type encodes are predicted to present (Mukherjee and Chandra [2014]). They assume that the more epitopes an HLA type can present the more reduced a host's susceptibility to H1N1 influenza. They show that populations with varying susceptibilities to a strain of H1N1 influenza can actually reduce the size of an epidemic from said strain. Gupta and Hill [1995] and Sambaturu et al. [2018] are rare examples of theoretical works that model host HLA diversity to see its effect on pathogen dynamics in classic epidemiological models. Including epidemiology into theoretical works about the MHC is worth doing as if we believe pathogens have co evolved with the MHC then pathogen dynamics are a very important detail needed to be considered. Understanding how a population's MHC profile affects disease dynamics could potentially increase our effectiveness at protecting populations from pathogens.

Penman et al. [2013] also uses an epidemiological ODE model to tackle questions about the MHC but uses these tools to focus on evolutionary questions about the MHC. Penman et al. [2013] defines hosts by HLA genotype over two loci, where HLA genotype affects the susceptibility to multi-strain pathogens. Penman et al. [2013] shows that

pathogen mediated selection can cause HLA alleles at different loci to form associations that are non-overlapping and are nonrandom. This could potentially contribute to the maintenance of long range haplotypic associations between MHC loci.

Lobkovsky et al. [2019] explores mechanisms for HLA haplotype diversity. They fit multiple models to haplotype frequency data sets. Each model varies how they assign fitness to their host depending on their HLA genotype. They use additive fitness, multiplicative fitness, overdominance and hybrid fitness rules. Lobkovsky et al. [2019] fits models using each of these fitness rules to haplotype frequency datasets and concludes that multiplicative fitness appeared to be the fitness rule that best accounts for MHC haplotype patterns. They show that when using frequency dependence selection the improvement in model fit over models that did not include it were significant. More surprisingly is that the frequency dependence selection on the haplotypic level was found to be positive in contrast to the negative frequency dependent selection being found on individual MHC loci. Both Penman et al. [2013] and Lobkovsky et al. [2019] explore the effects of more than one HLA loci interacting, albeit in very different ways. Both studies showed that introducing multiple loci into a model gave insight into mechanisms (such as linkage disequilibrium Penman et al. [2013]) that would have otherwise been very difficult to predict if only considering one loci at a time.

## 1.2 Specific topics explored in this thesis

### 1.2.1 The effectiveness of case control studies to detect MHC - pathogen associations

As mentioned the peptide shape an MHC can hold depends on the binding cleft of the MHC molecule. The binding cleft shape of an MHC molecule depends on the MHC allele that encodes the MHC molecule. This has led to case control studies trying to find which MHC alleles or haplotypes confer protection to which pathogens. In large this has been done for humans i.e for HLA genes. Hill et al. [1991] is the earliest work in finding such associations between a HLA-B type and malaria. Hill et al. [1991] finds that HLA-Bw53 was associated with reduced chances of developing severe malaria in The Gambia. Class II haplotype DRB1*1302-DQB1*0501 was likewise found to offer protection against severe

malaria.

I mentioned previously that despite numerous works into finding HLA allele association with infectious disease outcome, associations tend to be weak or even inconsistent. Explanations as to why this might be happening have been theorised. One explanation could be for multi strain pathogens HLA alleles might be associated with specific strains rather than the pathogen in general. Toyo-Oka et al. [2017], Salie et al. [2013] show HLA associations being strain specific for the case of Tuberculosis (TB). It is this phenomenon I explore theoretically in chapter 2.

The theory of how to obtain accurate results in a case control study of the impact of host genotype on a pathogen has not been widely explored. One of the most relevant theoretical works is MacPherson et al. [2018], this study is not specific to the MHC but does tackle genome wide association studies with pathogen infection. They use what they call a "phenotypic-difference model" and show that if you do not take into account a pathogen's genetics, important host loci may be ignored. Although not directly related to the HLA, comparisons are obvious with what we see in contradictory results in HLA association studies with pathogens.

To my knowledge it has not been theoretically explained why, if a HLA molecule does confer protection or susceptibility to pathogen infection, it might not be possible to detect such associations in a case control study. In chapter 2 of this thesis I explore the case of a multi-strain pathogen for which the presence of a specific HLA molecule in the host is necessary to develop an effective memory immune response against a specific pathogen strain, and how well case-control studies can detect the importance of that particular HLA type.

### 1.2.2 The evolution of MHC copy number variation (CNV)

I previously noted that CNV of the MHC varies between species and some CNV of the MHC may be evolutionary stable due to evidence of CNV being present pre speciation of two macaque species (Otting et al. [2007]).

Steinmetz et al. [1986] found 11 recombination events on the MHC for the mouse, of which 7 of these recombination events were unequal crossing over recombination. Otting et al. [2005] note that variation in the Mamu-A and Mamu-B gene in the rhesus macaque

is probably due to unequal crossing over. Unequal crossing over recombination is one mechanism that can cause copies of genes and result in a species having CNV. In chapter 3 I simulate unequal crossing over recombination, as it provides a mechanism to generate new copies of genes. I investigate how natural selection, due to interaction between MHC molecules and pathogens, will affect the numbers of copies of MHC genes within a population.

As far as I am aware there is very little theoretical work into the variation of CNV of the MHC for different species. There are however theoretical works that tackle similar questions. Krüger and Vogel [1975] creates several models that theoretically explore how unequal crossing over affects the distribution of numbers of copies of a gene in a population for different types of selection. The three types of selection they explore are: No selection, the individual with the larger number of copies is fittest and finally the individual with a set number of copies of genes is the fittest. For certain versions of the model in the case where there is no selection they found that the stationary distribution of the number of alleles depended on the initial distribution of allele copies and the number of generations to converge to this stationary distribution depended on the probability of unequal crossing over events occurring. For the case where larger numbers of copies of alleles are the fittest, they show that the distribution of the number of allele copies increases continuously. When an optimum number of allele copies is the fittest they show stationary distributions vary around this optimum number of alleles where the most frequent value number of alleles is the optimum number, the width of these distributions positively correlated with the probability of unequal crossing over events occurring.

Takahata [1981] creates a similar model to Krüger and Vogel [1975] however it incorporates sister chromatid exchange. Due to boundary conditions implemented in Takahata [1981] on the number of repeated genes that are viable for an individual to survive, equilibrium distributions on the number of repeated genes can always be found. They show that when the sum of the rate of sister chromatid exchange and inter-chromosomal crossing-over were constant, that distributions of the numbers of repeated genes did not change when the relative rates of these two processes changed, implying that theoretically it would be unnecessary to distinguish the two processes. This however is only because of the assumption that sister chromatid exchange and inter-chromosomal crossing-over had

the same patterns.

Since then multiple mathematical studies exist on the distribution of the number of repeated genes due to unequal crossing over (Baake [2008], Shpak and Atteson [2002], Redner and Baake [2004]). In the second chapter I do not explore multiple versions of unequal crossing over; rather I choose specific assumptions and analyse the model numerically (see methods chapter 2 and result section chapter 2). While previous theoretical work explores unequal crossing over and the distribution of repeated genes, I am concerned with the rules of selection caused by pathogens and its effect on CNV and use unequal crossing over as a tool to merely generate new functional genes.

As mentioned earlier there is very little theoretical work on the MHC and gene copy number variation. One such piece of work (Bentkowski and Radwan [2019]) uses an individual based model that explores how the presence of multiple pathogens might affect the number of copies of MHC genes. They use a similar framework to Borghans et al. [2004] where MHC molecules and pathogens are represented by strings of bits and mutations can change both MHC molecules and pathogens by altering this string of bits however they introduce a new feature that allows the duplication and deletion of MHC genes. Bentkowski and Radwan [2019] assumes there is a cost associated with increased number of copies of MHC genes but this increased number of genes could help recognise more pathogen variants. Bentkowski and Radwan [2019] varies the diversity of pathogens present, the cost of having multiple MHC copies and the rates of mutations. Bentkowski and Radwan [2019] found that having a higher diversity of pathogens present only caused more copies of MHC genes when the cost of multiple MHC copies was low. My work in chapter 3 follows similar assumptions, such as having a cost for having too many MHC copies. However, the process by which I generate copies of genes and how I represent MHC allele fitness and therefore individual fitness is different, and offers new insights into MHC copy number variation.

Theoretical works that model the MHC with multiple loci are very sparse. However as noted in section 1.1.6, previous models have considered the haplotypic structuring of MHCs (specifically the HLA in humans) over multiple loci (Penman et al. [2013], Lobkovsky et al. [2019]). Lobkovsky et al. [2019] develop a range of possible rules for combining MHC haplotype fitnesses (e.g. a "maximum" rule, whereby an individual's

fitness is equal to the higher fitness out of the fitnesses of the two MHC haplotypes they possess). This approach is similar to the work here however I combine the fitnesses of all MHC genes encoded by an individual into an overall individual fitness. Lobkovsky et al. [2019] assigns individual haplotype fitnesses for all the possible MHC haplotypes in the population (regardless of the particular genes within each haplotype), and then each individual's fitness is based on a combination of the two MHC haplotypes they possess. The research objective here (understanding MHC CNV) is also different to the objective of Lobkovsky et al. [2019].

### 1.2.3 The evolution of MHC molecule binding promiscuity

As noted in section 1.1.1, MHC molecules possess a binding cleft which is crucial to their functions as they need to represent peptides to the immune system and do this by binding them to the binding cleft. However it has been noted that MHC alleles differ in their binding cleft peptide repertoire (Paul et al. [2013]). This gives rise to the possibility that some alleles have evolved to protect against multiple pathogens/pathogen strains while others have evolved to be more specialised. Considering MHC alleles with varying binding peptide repertoires has given rise to viewing them as specialist (i.e. fastidious - binding a small range of peptides) or generalist (i.e. promiscuous - binding a large range of peptides) (Kaufman [2018], Chappell et al. [2015], Kaufman [2020]). In chapter 4 I explore what pathogen climates give rise to MHC alleles of varying promiscuity co existing.

MHC promiscuity in terms of peptide binding cleft repertoire is another form of diversity among MHC alleles and the knowledge of MHC alleles having various promiscuity brings in ideas for different ways of modelling MHC allele finesses in a theoretical sense. A large portion of theoretical work just considers MHC alleles as having flat fitness values (De Boer et al. [2004], Lewontin et al. [1978]) as well as the work here in chapter 2. These fitness values cannot represent MHC alleles functionality if we are considering how well a MHC allele helps a host is based upon how well its MHC molecule can bind to a peptide or how many different types of peptides it can bind to. Siljestam and Rueffler [2019] represents MHC functionality by representing MHC alleles as having different functioning parts and these parts are represented by numbers. How close these numbers match

to a pathogens numbers (which also have corresponding attributes) determines how well it helps a host survive infection. My work in chapter 4 also represents MHC alleles as having multiple attributes; however, in the work here each attribute represents a type of peptide shape. How high the number is for a particular shape represents how well that MHC molecule can bind to a peptide of that shape. Using this way of representing MHC alleles I develop an individual based model that explore how pathogen climates affect the types of alleles that can co exist in terms of MHC molecule binding promiscuity.

# Chapter 2

# Detecting HLA-infectious disease associations for multi-strain pathogens

## 2.1 Introduction

An individual's ability to fend off invading pathogens is affected by their genotype. Co-evolution between humans and pathogens has generated extreme diversity in certain human genes, in particular the human Major Histocompatibility Complex loci: the Human Leokocyte antigens (HLAs) (Spurgin and Richardson [2010], Jeffery and Bangham [2000], Hedrick [2002], Pierini and Lenz [2018]) making HLA loci the most polymorphic in the human genome (Robinson et al. [2014]). Understanding which HLA genotypes are best adapted to which infectious diseases is an ongoing challenge. Here I investigate how epidemiological and population genetic factors combine to influence whether individual HLAs appear to protect against infection with multi-strain pathogens.

HLA molecules play an integral role in the human immune system. They are cell surface proteins containing a binding cleft. The binding cleft is loaded with peptides sampled from either inside (class I HLA molecules) or outside (class II HLA molecules) the cell

(Horton et al. [2004]). T cell receptors bind to the HLA/peptide complex and if the bound peptide is recognised as "non-self" by a T cell, this will trigger an immune response. HLA molecules encoded by different alleles have binding clefts with different properties. The specific peptide fragments which are displayed by HLA molecules determine an individual's T cell responses, thus HLA genotype acts as a bottleneck, which can shape adaptive immunity.

HLA genotype has been shown to affect the outcome of a wide range of infectious diseases (Tian et al. [2017], Dunstan et al. [2014], McLaren et al. [2015], Sveinbjornsson et al. [2016], Hill et al. [1991], Thursz et al. [1995], Oliveira-Cortez et al. [2016], Carrington and O'Brien [2003]). The first HLA/infectious disease association was demonstrated for malaria: HLA-Bw53 was associated with a reduced chance of developing severe malaria disease in a population in The Gambia (Hill et al. [1991]). Also in The Gambia, HLA-DRB1*1302 was associated with a reduction in the probability of developing persistent hepatitis B (HBV) infection (Thursz et al. [1995]). Various HLA alleles have been shown to affect the time to Acquired Immune Deficiency Syndrome (AIDS) for Human Immunodeficiency Virus (HIV) infected individuals (Just [1995]); and individual amino acids in the binding clefts both HLA-A and HLA-B molecules can impact HIV setpoint viral load (McLaren et al. [2015]). For some common infections including mumps, childhood ear infections and strep throat, a recent Genome Wide Association Study (GWAS) suggests that HLA genotype may affect the probability of experiencing symptomatic infection at all (Tian et al. [2017]).

Studies which identify HLA-infectious disease associations typically compare a group of individuals with an infectious disease phenotype (cases) to a group of individuals without it (controls) and examine differences in the frequencies of HLA types in the two groups. An over representation of a specific HLA type in the control group could be because it has a protective effect. However, such case control studies do not always give consistent results. A case control study of severe malaria, similar to the previously mentioned study in The Gambia (Hill et al. [1991]), was performed in Mali (Lyke et al. [2011]). The Mali study did not find HLA-Bw53 to be protective against severe malaria but they did find that HLA-A*30:01 and HLA-A*33:01 increased susceptibility to developing severe malaria. Lyke *et al* noted that this discrepancy could be due to different strains of

malaria parasite circulating in Mali as opposed to The Gambia. If specific HLA alleles are associated with better immune responses to just a subset of pathogen strains, then the effectiveness of HLA alleles will depend on which pathogen strains are circulating in a population.

The interaction between host genotype and pathogen strain in determining disease outcome is highlighted by recent studies of Tuberculosis (TB). A case control study in the Chiang Rai province of Thailand split TB patients into groups infected with modern strains of TB and those infected with ancient strains of TB (Toyo-Oka et al. [2017]) (the ancient/modern distinction is based on the presence/absence of a TbD1 deletion in the TB genome). They found HLA DRB1*09:01 to be associated with protection from infection with modern strain tuberculosis. They did not find this association when they grouped patients with modern and ancient tuberculosis strains together. A similar study was performed in Cape Town, South Africa (Salie et al. [2013]), distinguishing TB strains by restriction fragment length polymorphism genotyping. Salie *et al* found that HLA class I types A*01, B*08 and C2* were all associated with increased susceptibility to Beijing strain TB; B*27 and C1 were associated with lower susceptibility to the Beijing strain.

The possible impact of HLA type on the epidemiology of infection has been considered in terms of the maintenance of parasite diversity (Gupta and Hill [1995]), and more recently in terms of the effect of HLA type on pathogen $R_0$ in different populations (Sambaturu et al. [2018]). Specifically, Sambaturu *et al* addressed the impact of HLA type on the spread of H1N1 influenza (Sambaturu et al. [2018]), by first classifying HLA types by the range of influenza epitopes they are predicted to present, (Mukherjee and Chandra [2014]), and then making the assumption that the more viral epitopes a host can represent, the lower that host's susceptibility to H1N1 influenza. They show that when a population has a wide range of individual susceptibility to a strain of H1N1 influenza, it can reduce the size of an epidemic from said strain of H1N1 influenza. This previous work highlights the potential significance of HLA diversity for public health. However, no previous model has considered the epidemiological processes underlying why, if an HLA type truly does affect infection, it might not always be detected as having such an effect

*HLA-C in this study was classified into C1 or C2, a distinction based on the interaction between HLA-C and Killer-cell Immunoglobulin-like Receptor (KIR) molecules.

in a case control study.

MacPherson *et al* recently analysed the effect of pathogen diversity and host-pathogen coevolution on the ability of genome wide association studies (GWAS) to detect which host genes matter for infection (MacPherson et al. [2018]). They elegantly demonstrated that if pathogen diversity is ignored, many important host loci will go undetected by GWAS. In the case of HLA genes, and the further complication of adaptive immunity, the problems highlighted by MacPherson may be compounded.

Here I explore the case of a multi-strain pathogen for which the presence of a specific HLA molecule in the host is necessary to develop an effective memory immune response against a specific pathogen strain. I identify the different epidemiological outcomes which arise as a consequence of this HLA-strain relationship, when population HLA frequencies vary. I simulate case control studies of infection, and demonstrate the circumstances under which an HLA type offering an advantage against a specific pathogen strain is likely to appear protective or risky against the prevailing local infection.

*Neisseria meningitides, Streptococcus pneumoniae* and *Plasmodium falciparum* are examples of multi-strain pathogens where humans experience multiple infections; become immune to different strains, and where T cell responses (and hence HLA type) are implicated in the generation of protective immunity (Wiertz et al. [1996], Davenport et al. [2003], Aslam et al. [2010], Mordmüller et al. [2017], Aslam et al. [2011]). My model offers insight into how HLA/strain interactions may impact the epidemiology of such systems. My model further suggests a technique to detect the existence of HLA/strain specific associations in a system, even if the key properties distinguishing different strains are not yet known.

## 2.2 Methods

### 2.2.1 The Epidemiological Model

I consider a pathogen which exists as at least two strains (1 and 2). I assume that the diploid HLA genotype of a host (ij) determines whether or not that host will be able to develop a memory immune response against a particular pathogen strain after infection.

To model these possible immune outcomes, I use SIR and SIS models which are commonly used ordinary differential equation (ODE) models in epidemiology (for a full review of such approaches, see Keeling and Rohani [2011]). If a host's genotype includes an HLA allele which allows the recognition of strain $i$, that host will become immune to strain $i$ following infection (SIR dynamics). If a host does not have an HLA allele enabling them to recognise a pathogen strain, that host will not become immune to that pathogen strain on recovery (SIS dynamics). This is a simplifying assumption (although very stark MHC/pathogen strain relationships are observed in nature specifically in chickens, as I detail in the discussion). My aim here is to ask the question: if certain HLA types make the difference between mounting successful memory immune responses or not (as simulated in the model), will this ever be detectable by the current standard methodology (the case control study)?.

The force of infection for pathogen strain $i$ is $\lambda_i$. The rate of recovery from pathogen strain $i$ back to being susceptible is $\sigma_i$. The rate an infected individual becomes immune to pathogen strain $i$ is $\mu_i$. I simulate a population with a constant size, so births and deaths happen at the same rate $d$. Newborns are always born as susceptible and all rates have the unit $year^{-1}$. Pathogen induced mortality is not included in this model. The reasons for this decision are firstly that the rate at which pathogens would adapt to a population's genetic landscape are vastly faster than the rate a host population adapts to the pathogens. Secondly we are comparing our results to case control studies which are snap shots in time of a populations genetic landscape related to infectious disease. If we were to include pathogen induced mortality, we would see allele frequencies reach some stable equilibrium which would depend on pathogen parameters as well as initial allele frequencies of the population. For the purposes of investigating case control studies I believe fixing the allele frequencies to remain constant is a reasonable choice for simplifying the analysis. Flow chart of the model is given in figure 2.1.

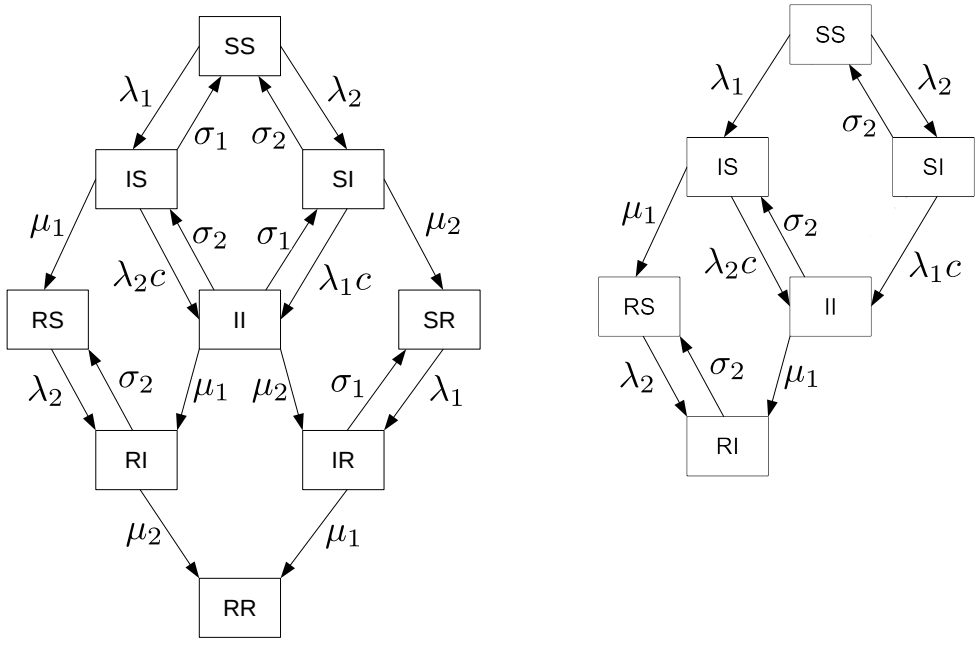



Figure 2.1: **The possible states of the population and pathways between them.** (a) is a flow chart of the possible paths a host of any genotype $ij$ can take from initially being susceptible to all pathogen strains. (b) is a flow chart specifically for a host whose genotype means they can only mount a memory immune response against pathogen strain 1.

I use a double letter notation to denote susceptibility (S), infectiousness (I), immunity (R), to pathogen strain 1 (first letter) and to pathogen strain 2 (second letter). For instance host IS is infected with pathogen strain 1 and susceptible to pathogen strain 2; a host with SR is susceptible to pathogen strain 1 and immune to pathogen strain 2. I denote the proportion of the population susceptible to pathogen strain 1 and immune to pathogen strain 2 with genotype $ij$ as $N_{ij}^{SR}$. The total proportion of a genotype in the population is

$$\begin{aligned} N_{ij} =& N_{ij}^{SS} + N_{ij}^{SI} + N_{ij}^{SR} + N_{ij}^{IS} + N_{ij}^{IR} + N_{ij}^{RS} \\ &+ N_{ij}^{II} + N_{ij}^{RI} + N_{ij}^{RR}, \end{aligned} \tag{2.1}$$

where the total size of the population is

$$N = \sum_{i,j} N_{ij} = 1. \tag{2.2}$$

I denote the proportion of hosts who are infected with pathogen strain $i$ as $I_i$ and I denote the proportion of hosts immune to pathogen strain $i$ as $R_i$.

$$\begin{aligned} I_1 &= \sum_{\{i,j\}} \left( N_{ij}^{IS} + N_{ij}^{IR} + N_{ij}^{II} \right) \\ I_2 &= \sum_{\{i,j\}} \left( N_{ij}^{SI} + N_{ij}^{RI} + N_{ij}^{II} \right) \\ R_1 &= \sum_{\{i,j\}} \left( N_{ij}^{RS} + N_{ij}^{RI} + N_{ij}^{RR} \right). \end{aligned} \tag{2.3}$$

The transmission rate for pathogen strain $i$ is notated as $\beta_i$, therefore the force of infection from pathogen strain $i$ is

$$\lambda_i = \beta_i I_i. \tag{2.4}$$

ODEs for this model can be found in section A.2 in the Appendix A.

In order to scale the degree to which a host can be infected with both pathogen strains simultaneously I introduce the parameter $c$. When $c = 0$ coinfection is impossible. When $c = 1$ infection with one pathogen strain has no effect on the rate a host becomes infected with the other pathogen strain. In order to scale the strength of strain transcending (and host-genotype-independent) immunity, I introduce the parameter $\alpha$. A proportion $\alpha$ of all recoveries enter the RR state and a proportion $(1 - \alpha)$ recover according to the host's genotype. In section 2.3.1, 2.3.2 and 2.3.3 of the results, $c = 0$ and $\alpha = 0$. In section 2.3.4 I demonstrate the impact of varying these parameters.

My host population is divided into different possible diploid HLA genotypes $(ij)$ at a single HLA locus. In my main model, I include two pathogen strains (1 and 2) and

two HLA alleles (1 and 2). I assume a 1:1 correspondence between HLA alleles and pathogen strains, meaning that the presence of HLA allele 1 is necessary to mount a memory immune response against pathogen strain 1.

I also extend the main model in two ways. I allow for three possible pathogen strains (1, 2 and 3), with 3 corresponding HLA alleles necessary for the recognition of each (see section A.3 of Appendix A). I also include a "perfect" or a "useless" HLA allele, recognising both or neither of pathogen strains 1 and 2, alongside strain specific HLA 1 and 2 alleles (see section A.4 of Appendix A).

A key parameter of all the models is HLA allele frequency. I notate this as $p_i$ for allele $i$. The birth rates of different host genotypes are in accordance with the Hardy-Weinberg principle (Relethford [2012]), implying random mating within the population. In a two allele model, the frequency of homozygotes (11 and 22) and heterozygotes (12), are as follows:

$$\begin{aligned} N_{11} &= p_1^2 \\ N_{12} &= 2p_1(1-p_1) \\ N_{22} &= (1-p_1)^2. \end{aligned} \tag{2.5}$$

Figure 2.2 illustrates how the frequencies of each genotype vary as $p_1$ varies between 0 and 1.

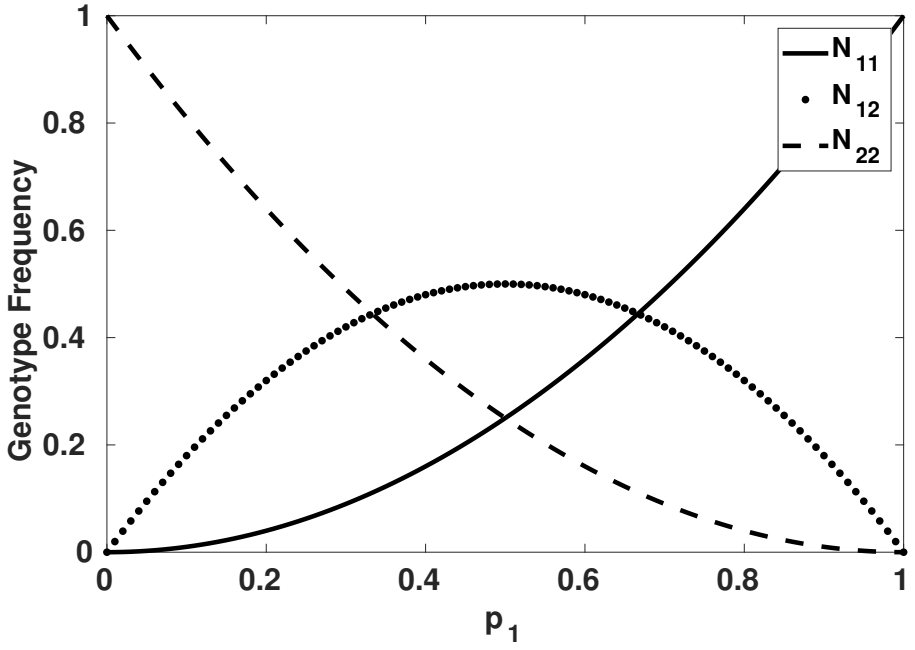



Figure 2.2: **The frequency of genotypes according to the Hardy-Weinberg principle**

Starting frequencies of each genotype were assigned in these proportions and the frequencies of each genotype remained unchanged over time. Hardy-Weinberg proportions were similarly used for three allele models, in which 3 homozygotes (11, 22 and 33) and 3 heterozygotes (12, 23, 13) are possible. In three allele models I set $p_2 = p_3 = (1 - p_1)/2$. This means that as $p_1$ is varied I allocate an equal frequency to alleles 2 and 3. This maximises HLA diversity.

All results described here are obtained by numerically solving the ODEs described above, using solver ode45 in Matlab and calculating quantities from these numerically solved solutions. Matlab code to run the two pathogen and three pathogen ODE models is available in the following repository : `https://github.com/ConnorFrancisWhite/HLA_Infection_Association_Model_White_et_al_2020`.

### 2.2.2 The Odds Ratio

The key feature of the model is that possession of a particular HLA type is necessary in order for a host to develop a specific memory immune response against a specific pathogen strain. That HLA type will always be "protective" against infection with the strain it matches. However, when assessing the protectiveness/riskiness of an allele or genotype in a case control study, it is rare that such functional strain definitions are already known. I wish to investigate how the interplay between a diverse host and pathogen makes particular alleles risky or protective against infection in general (i.e. infection with any strain), since this is the property which will most likely be captured by case control studies in practise.

The odds ratio for being infected with any strain given a host has a specific allele $i$ is calculated as follows:

$$OR_i = \left(\frac{P(I \mid i)}{1 - P(I \mid i)}\right) \Big/ \left(\frac{P(I \mid \hat{i})}{1 - P(I \mid \hat{i})}\right).$$

Here, $P(I \mid i)$ is the proportion of hosts with allele $i$ that are infected and $P(I \mid \hat{i})$ is the proportion of hosts that do not have allele $i$ that are infected. If $OR_i < 1$ allele i is protective against the prevailing local infection. This method of calculating the odds ratio uses proportions of the population at the steady state of the ODE model, which was calculated numerically.

In section 2.3.3 I investigate whether the effects I are modelling could be detected in a real world study. I assume this real world study involves 500 cases and 500 controls, and assign different numbers of genotypes to both groups by multiplying relevant steady state proportions from the model by 500. A standard method to calculate the odds ratio for studies involving finite sample sizes of cases and controls, which I denote as $OR'_i$, is as follows:

$$OR'_i = \left(\frac{n(I \mid i) + g}{n(\hat{I} \mid i) + g}\right) \Big/ \left(\frac{n(I \mid \hat{i}) + g}{n(\hat{I} \mid \hat{i}) + g}\right).$$

Where $n(I \mid i)$ is the number of people who are infected and have allele $i$. $n(\hat{I} \mid i)$ is the number of people who are not infected and have allele $i$. $n(I \mid \hat{i})$ is the number of people who are infected and do not have allele $i$. $n(\hat{I} \mid \hat{i})$ is the number of people

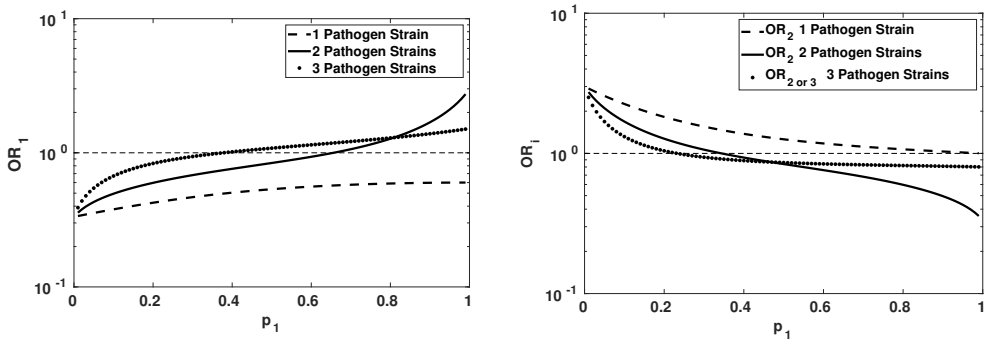
who are not infected and do not have allele $i$. $g$ is a value used so there is no undefined calculation for the odds ratio even if the number of individuals in one of the categories is 0, $g$ is always greater than 0. A commonly used value for $g$ is $g = 0.5$ (Woolf [1955], Gart [1966]). Woolf [1955] also provides a method for calculating the confidence interval for the odds ratio which I use in section 2.3.3. In figures 2.8 and 2.9 I round the sample sizes to integer values to further simulate a real world study.

## 2.3 Results

### 2.3.1 *The protectiveness of a strain-specific HLA allele against infection is related to the population frequency of that allele.*

I first consider one pathogen strain in a population that contains two HLA alleles. Only one of these alleles (allele 1) allows a host to generate a memory immune response against the pathogen strain. In figure 2.3a I illustrate how varying HLA frequencies affects $OR_1$, the odds ratio for a genotype containing allele 1 being infected with any strain.

If only pathogen strain 1 is present $OR_1$ is always below 1 for the entire range of $p_1$ values ($0 < p_1 < 1$) (figure 2.3a, dashed line). The distribution of HLA alleles within the population does not alter the fact that allele 1 is beneficial. This result is intuitive, since the only circulating pathogen strain is one to which hosts with allele 1 can develop memory immunity. No matter how many hosts have allele 1, possessing allele 1 will always make hosts less likely to be infected. In figure 2.3b for the 1 pathogen strain scenario we see $OR_2$ is always above 1 meaning you are more likely to be infected if you have HLA allele 2. $OR_2$ approaches the value 1 but never actually equals it in figure 2.3b.




(a) (b)

Figure 2.3: **The relationship between HLA allele frequencies and $OR_i$.** The three different line styles indicate scenarios with 1, 2 and 3 pathogen strains. Panel (a) illustrates how the odds ratio for allele 1 changes as $p_1$ is increased. Panel (b) illustrates how the odds ratio for every other allele other than allele 1 varies as $p_1$ is increased. The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_i = 0.02$ and $\sigma_i = 0.02$ $(i = 1, 2, 3)$.

As noted in the introduction, however, many pathogens exist as multiple strains. Let us suppose that among the pathogen variants, some express an immunogenic peptide which can be bound by one HLA type, and others express a different form of the immunogenic peptide which can be bound by a different HLA type at the same HLA locus. I simulate a two strain, two HLA type system, where HLA allele 1 is necessary to mount a memory immune response against strain 1, and HLA allele 2 is necessary to mount a memory immune response against strain 2. Heterozygotes for alleles 1 and 2 (genotype 12) can generate a memory immune response against both strains. Now as $p_1$ is increased $OR_1$ goes from below 1 to above 1 (figure 2.3a, solid line). The frequency of HLA allele 1 in the population is negatively correlated with its ability to protect a host from the prevailing local infection. The same applies to HLA allele 2 (figure 2.3b), noting that $p_2 = 1 - p_1$. To understand why this is happening I need to look at the level of infection from each strain, within the population as $p_1$ varies.

If allele 1 is more frequent than allele 2, there are more hosts who can become immune to pathogen strain 1 than hosts who can become immune to pathogen strain 2. Pathogen strain 2 will be more successful in such an environment. As $p_1$ is increased (figure 2.4),

the proportion of hosts infected with pathogen strain 1 at equilibrium decreases and the proportion of hosts infected with pathogen strain 2 increases. The distribution of HLA alleles in the population affects the pathogen strain structure in the population. This explains why $OR_i$ is positively correlated with the frequency of allele $i$.

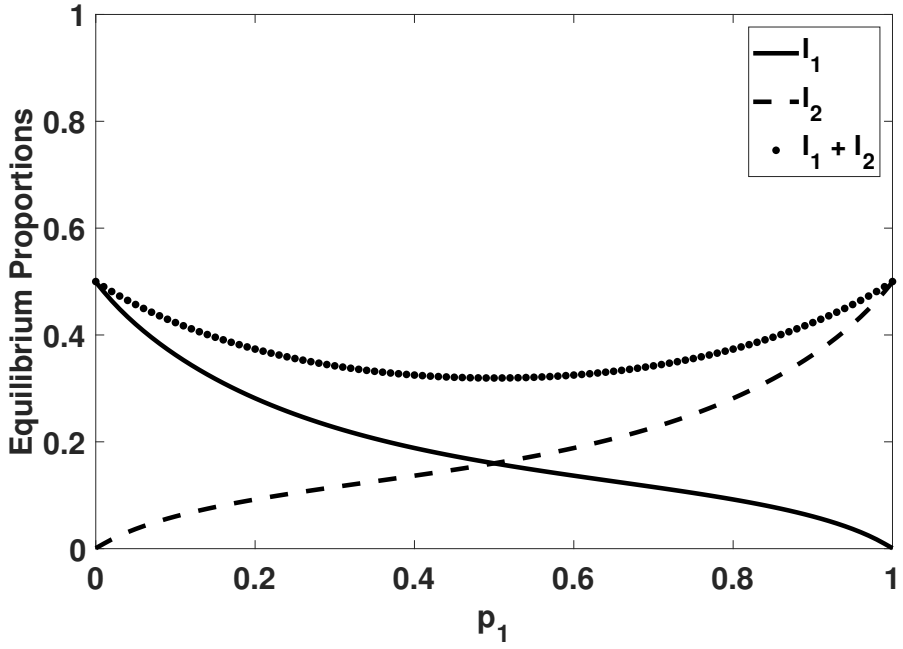



Figure 2.4: **The relationship between the steady state values of $\mathbf{I_1}$ and $\mathbf{I_2}$ and the frequency of HLA allele 1, in a 2 strain, 2 allele model where possessing allele $i$ is necessary to mount a memory immune response against strain $i$.** The equilibrium values of $I_1$ and $I_2$ against $p_1$. $I_i$ is the total proportion of the population infected with pathogen strain $i$. The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_i = 0.02$ and $\sigma_i = 0.02$ $(i = 1, 2)$.

In a 3 strain, 3 allele extension of the system, $OR_1$ follows the same trend found with the two pathogen strain model (figure 2.3a, dotted line). As a further extension of the model, in Appendix A section A.4, I add a different type of third allele to the system where HLA alleles 1 and 2 confer the ability to recognise pathogen strains 1 and 2. This third allele is either a "perfect" allele (conferring the ability to mount an immune response

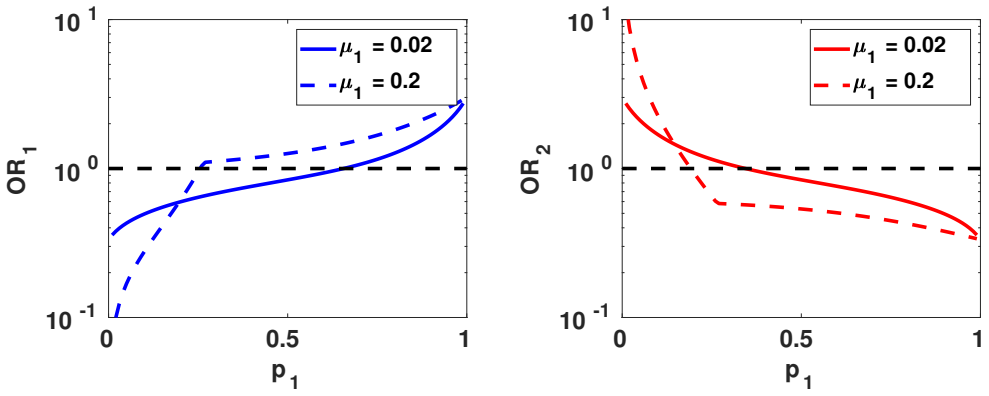
against either strain 1 or strain 2) or a "useless" allele (conferring no ability to mount any immune response).

The presence of the useless allele alongside alleles 1 and 2 and strains 1 and 2 does not substantially alter the pattern shown in figure 2.3a (see figure A.2) The presence of the perfect allele likewise does not alter the positive correlation between $p_1$ and $OR_1$, although it does reduce the set of circumstances where HLA allele 1 is protective against the prevailing local infection.

I finally tested a system including two pathogen strains, in which HLA allele 1 recognises strain 1, and only the perfect or useless allele is present alongside HLA allele 1. For both of these cases, $OR_1$ did not cross the value 1 for the entire range of $p_1$ (figure S3). It would seem a system requires at least two pathogen strains and the presence of at least two HLA alleles that differ in their strain specificity in order for $OR_1$ to go from below 1 to above 1 as $p_1$ is increased.

### 2.3.2 *The more rapid the immune response associated with a particular HLA allele against a particular pathogen strain, the less likely that HLA allele is to appear protective.*

In the results just presented, genotypes containing allele 1 become immune to strain 1 at exactly the same rate that genotypes containing allele 2 become immune to strain 2. However, it is possible that HLA molecules encoded by different HLA alleles differ in their fundamental ability to activate T cells and allow hosts to clear infection. One way to investigate this possibility within the framework is to allow each allele to be associated with a different recovery rate. I retain the strain specificity of the alleles in my main two strain, two allele model, but hosts with allele 1 recover more quickly than hosts with allele 2 after infection with their matched strain.




(a) (b)

Figure 2.5: **$OR_i$ and how its relationship with the frequency of allele 1 changes as the rate of recovery associated with HLA allele 1 becomes higher than that of associated with HLA allele 2.** $OR_1$ against $p_1$ for the two pathogen strain model (a). $OR_2$ against $p_1$ for the two pathogen strain model (b). The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_2 = 0.02$ and $\sigma_i = 0.02$ $(i = 1, 2)$.

As shown in figure 2.5, increasing $\mu_1$ relative to $\mu_2$ (enhancing the recovery rate associated with allele 1), changes the behaviour of $OR_i$ (compare dashed lines to solid lines). Lower allele frequencies are still associated with greater protection against infection in general, but $OR_1$ crosses the value 1 for a much smaller value of $p_1$. The faster recovery rate of allele 1 has caused allele 1 to be protective over a smaller region of $p_1$, and allele 2 to be protective over a greater range of values of $p_1$. This counter intuitive result is explained when I look at the level of infection of each pathogen strain at equilibrium (figure 2.6).

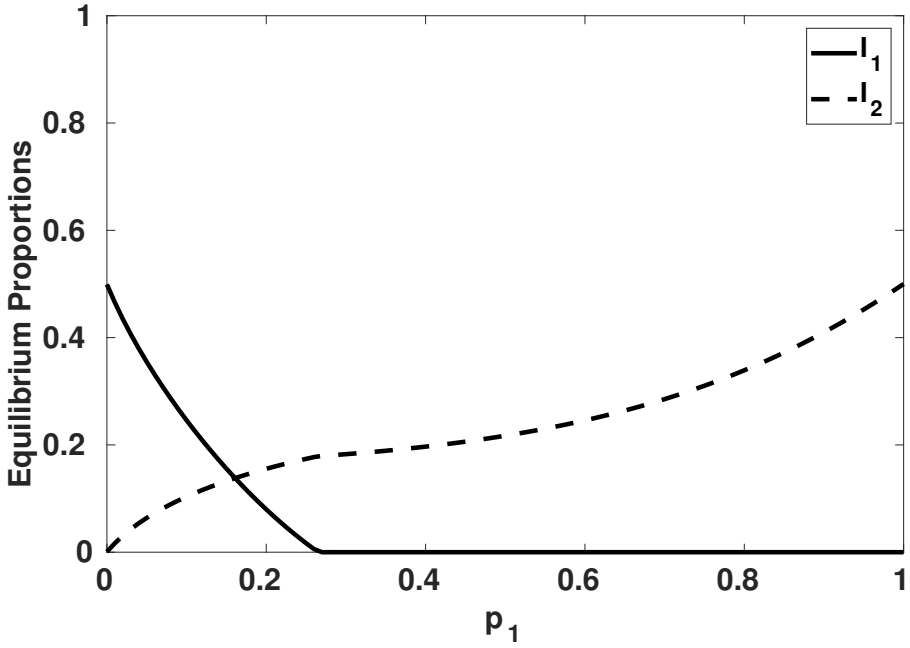



Figure 2.6: **Pathogen strain dynamics are affected by increasing the recovery rate associated with HLA allele 1 relative to that associated with HLA allele 2.** The graph is of $I_i$ against $p_1$ for the two pathogen strain model. The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_1 = 0.2$, $\mu_2 = 0.02$ and $\sigma_i = 0.02$ $(i = 1, 2)$.

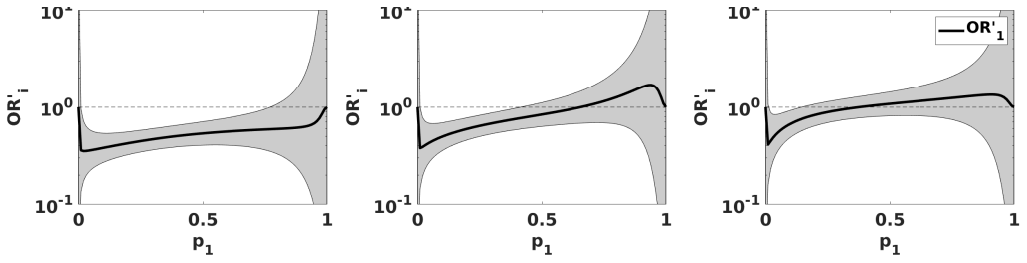
When $\mu_1 >> \mu_2$, pathogen 1 can only invade the system at low levels of $p_1$. The shorter duration of infection with pathogen strain 1 associated with allele 1 makes it harder for pathogen strain 1 to thrive. This in turn causes pathogen strain 2 to be the only circulating pathogen strain for the majority of $p_1$ values. In this environment only allele 2 will provide any protection.

These results demonstrate that, in a multi-strain pathogen system, the protective effect of an HLA allele may arise from being associated with a relatively slow recovery rate. Counter intuitively, this "poor" HLA allele could end up being protective in a population due to it helping one pathogen strain to out compete another. If we were to allow HLA allele frequencies to change we might see situations where allele 2 is selected for for a larger region of $p_1$, if $\mu_1$ is higher than $\mu_2$.

### 2.3.3 *Genotype/infection associations for HLAs which protect against specific pathogen strains are only detectable under limited circumstances.*

To investigate how strain specificity of HLA/pathogen recognition may impact our ability to detect HLA-infectious disease associations in real world case control studies, I calculated 95% confidence intervals for $OR'_1$, on the assumption of 500 infected cases and 500 disease-free controls (see Methods for further details).

I considered three different systems containing only HLA alleles recognising single pathogen strains. In a two HLA allele system where just pathogen strain 1 is present, for most of the range of $p_1$ $OR'_1$ has a confidence interval small enough that a case control study would be able to conclude that HLA type has an effect on infection (figure 2.7a). In a two allele system where both strains 1 and 2 are present a case control study would only identify that an HLA type protects against the prevailing local infection for a smaller range of $p_1$ (figure 2.7b). For the three strain, three allele system, with these parameter values, there is an even smaller range of values of $p_1$ where a case control study of 500 cases and 500 controls could declare if $OR'_1$ is below 1. This effect arises because the value of $p_1$ at which the odds ratio moves from below 1 to above 1 (henceforth $p_{crit}$), shifts to lower values of $p_1$ as the complexity of the system increases from two to three strains (examined further in section A.5.2 of the Appendix A).



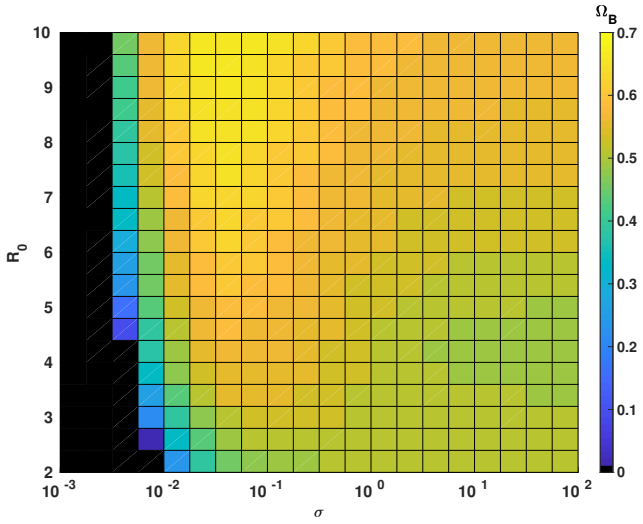

(a) (b) (c)

Figure 2.7: **$OR_1'$ and its relationship with the frequency of HLA allele 1.** The highlighted regions are the 95% confidence intervals for $OR_i'$. (a) illustrates a system with one pathogen strain and two HLA alleles in the population. (b) illustrates a system with two pathogen strains and two strain specific HLA alleles within the population. (c) illustrates a system with three pathogen strains and three strain specific HLA alleles within a population. The confidence intervals were calculated with a sample size of 500 cases and 500 controls (see Methods for further details). The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_i = 0.02$ and $\sigma_i = 0.02$ $(i = 1, 2, 3)$.

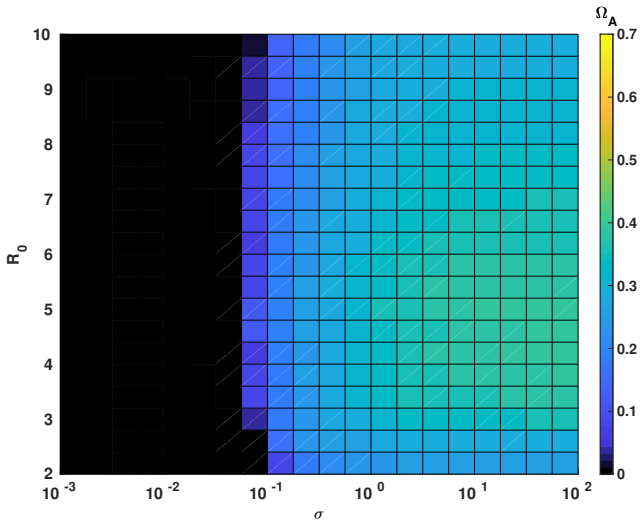
Figure A.4 in section A.5.1 of appendix A material illustrates the behaviour of the system if only 100 cases and 100 controls are used. Now, despite there still being a relationship between HLA frequency and protectiveness (i.e. $OR_1$ still correlates with $p_1$), there is no frequency of $p_1$ at which the 95% confidence interval for $OR_1$ does not encompass 1. However, if I change the properties of the pathogen by increasing $R_0$, I do find values of $p_1$ at which the 95% confidence interval for $OR_1$ does not include 1, even in the smaller case control study (figure S5).

To further explore how pathogen properties affect the ability of case control studies to detect protective or risky HLAs, I define $\Omega$ as the difference between the maximum and minimum values of $p_1$ at which the odds ratio for HLA type 1 is significantly below $(\Omega_B)$ or above $(\Omega_A)$ 1 (which is taken to be when the 95% confidence interval does not encompass 1). $\Omega$ therefore represents the ease with which a HLA protective $(\Omega_B)$ or risky $(\Omega_A)$ association against local prevailing infection might be detected in the system. If $\Omega$ is large, then the detection of the association is less dependent on a specific frequency of $p_1$. Figure 2.8 and 2.9 illustrates how $\Omega_B$ and $\Omega_A$ vary for different recovery rates and

different values of the basic reproductive number ($R_0$) of the pathogen in the 2 strain/2 allele and 3 strain/3 allele scenarios. For these simulations $\mu_i = \sigma_i = \sigma$ thus $\sigma$ refers to all recovery rates in the system.

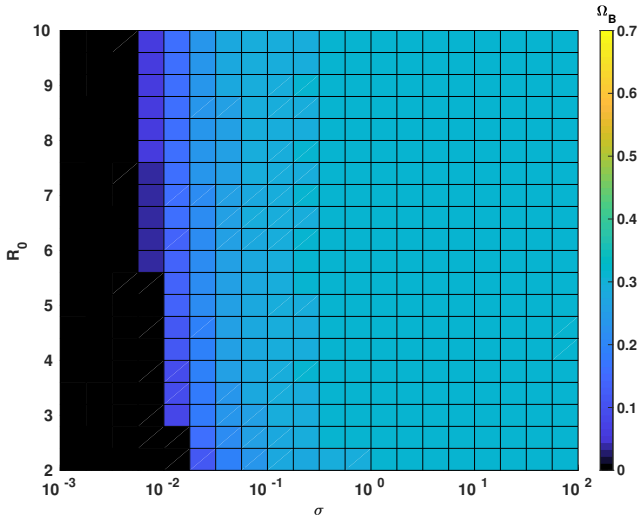



(a)

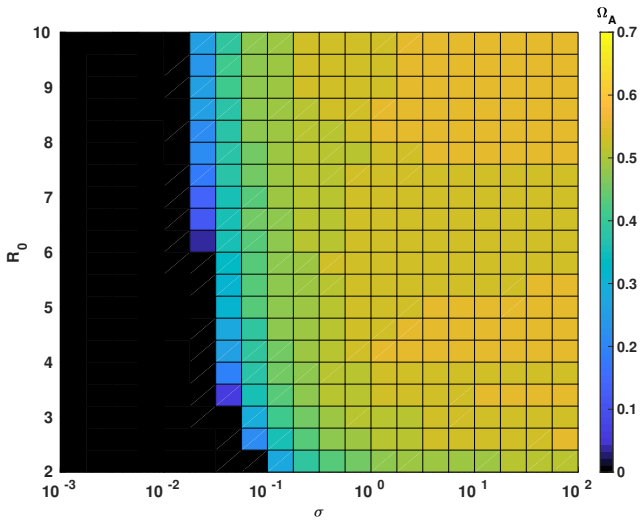



(b)

Figure 2.8: **A measure of how well protective ($\Omega_B$) or risky ($\Omega_A$) associations can be detected in simulated case control studies in the two strain, two allele system.** Panel (a) shows $\Omega_B$, panel (b) shows $\Omega_A$. $R_0$ is increased from 2 to 10 along the $y$ axis of each heat map, $\sigma$ is increased logarithmically from $10^{-3}$ to $10^2$ along the $x$ axis of each heat map. $\beta_i$ is calculated $\beta_i = R_0(d + \sigma)$ (for $i = 1, 2$) and $d = 0.01$.




(a)




(b)

Figure 2.9: **A measure of how well protective ($\Omega_B$) or risky ($\Omega_A$) associations can be detected in simulated case control studies in the three strain, three allele system.** Panel (a) shows $\Omega_B$, panel (b) shows $\Omega_A$. $R_0$ is increased from 2 to 10 along the $y$ axis of each heat map, $\sigma$ is increased logarithmically from $10^{-3}$ to $10^2$ along the $x$ axis of each heat map. $\beta_i$ is calculated $\beta_i = R_0(d+\sigma)$ (for $i = 1, 2, 3$) and $d = 0.01$.

Figures 2.8 and 2.9 shows that $\Omega_B$ and $\Omega_A$ are both 0 when the recovery rate ($\sigma$) is low (black region). Detecting any HLA-infection association is difficult if the infectious period is too long. This is unsurprising, because the key to protection against infection in the system is the ability, or not, of a host to mount a memory immune response. A pathogen which has a recovery rate similar to the mortality rate of the host cannot generate a strong HLA-dependent signal within such a system.

Figure 2.8 relates to the two strain, two allele system. $\Omega_B$ is a measure of the ability to detect whether HLA type 1 is protective against the prevailing local infection. Detecting a protective effect of HLA type 1 in a two strain-two allele system requires that homozygotes without HLA type 1 (e.g. genotype 22) bear the brunt of infections. This can only occur if pathogen strain 1 (to which genotype 22 is especially vulnerable) is circulating in the population. The ability to detect protective associations over a wide range of $p_1$ (a high value of $\Omega_B$ in figure 2.8a) implies that pathogen 1 is able to circulate at a reasonably high frequency in the population even as the proportion of allele 1 becomes relatively high. This requires there to remain a good balance between strain 2 ,which benefits from the increased frequency of genotype 11, and strain 1. The balance between the strains is affected by both $R_0$ and $\sigma$ (explored in section A.5.2 of the Appendix A). In most cases, increasing $R_0$ increases the coexistence of the strains, which is reflected in the broad relationship between $R_0$ and $\Omega_B$ seen in figure 2.8a.

Detecting that HLA type 1 is risky (i.e. increases susceptibility to the prevailing local infection) in the two strain-two allele system ($\Omega_A$) requires genotype 11 to bear the brunt of infections. This requires the presence of a high level of pathogen strain 2 among circulating strains. A high value of $\Omega_A$ implies that pathogen 2 is able to dominate the population even at relatively low frequencies of allele 1. This effect, likewise, depends on the relative values of $R_0$ and $\sigma$ (see section A.5.2 of Appendix A). Broadly, the faster the recovery rate, the easier it is for one pathogen strain to out-compete the other. $\Omega_A$, therefore, tends to increase with increasing $\sigma$, as seen in figure 2.8b.

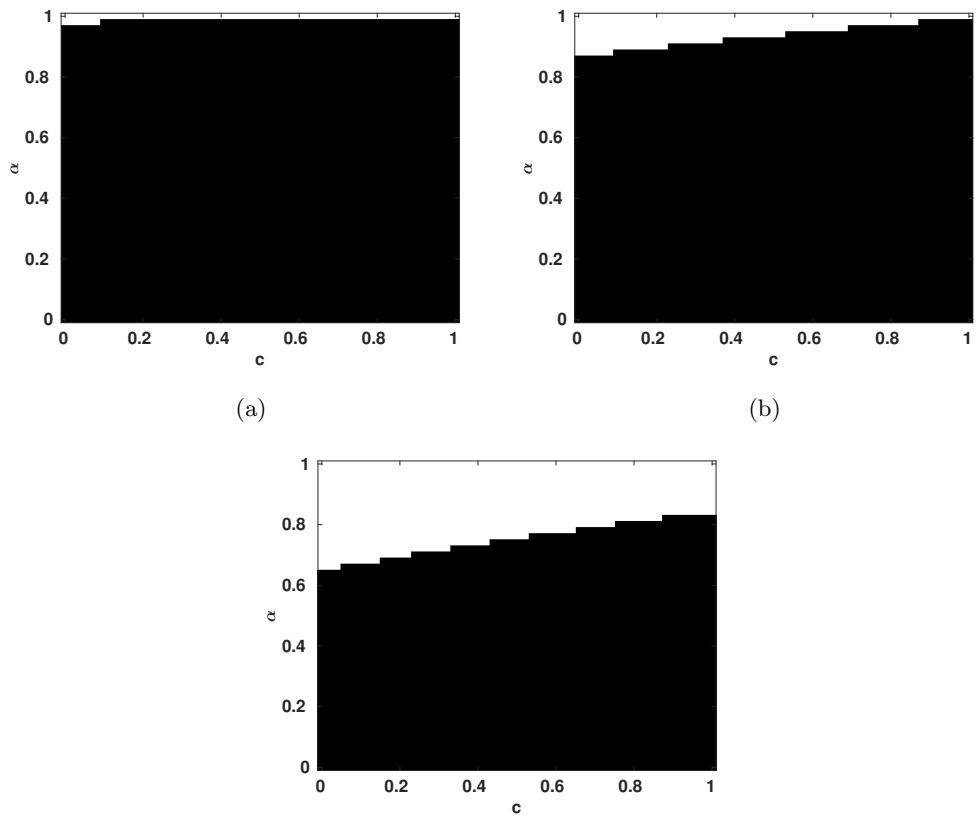
When we move from a two strain, two allele system to a three strain, three allele system (figure 2.9), $\Omega_B$ becomes smaller and $\Omega_A$ becomes larger. In other words, it has become easier to detect risky associations and harder to detect protective associations. It also appears that so long as $\sigma$ and $R_0$ are both above a certain threshold, they have

little effect on $\Omega_A$ or $\Omega_B$. When calculating an odds ratio for the effect of HLA-1 on infection in a three allele-three strain system, we are comparing the distribution of "HLA-1 containing genotypes" (11, 12, 13) and "non HLA-1 containing genotypes" (22, 33, 23) among cases and controls. Unlike in the two strain, two allele system, both sets of genotypes now include heterozygotes as well as homozygotes, and both sets of genotypes include a genotype particularly vulnerable to one of the strains (genotype 11 and genotype 33 are both equally vulnerable to strain 2). These factors increase the similarity of the two sets of genotypes being compared by the odds ratio, and mean that a lower value of $p_1$ is necessary for HLA-1 to be detected as protective against the prevailing local infection. In section A.5.2 of the Appendix A I derive an expression for $p_{crit}$, the frequency at which an HLA type switches from being protective to being risky, and show that $p_{crit}$ is more limited in the three strain, three allele model than in the two strain, two allele model, accounting for the plateauing of $\Omega$ in figure 2.9.

### 2.3.4 *Model behaviour is insensitive to co-infection, but breaks down at high levels of strain transcending immunity*

As noted in section 2.2, parameter $c$ controls the effect to which a host can be infected simultaneously with two pathogen strains and $\alpha$ controls the proportion of hosts infected with a pathogen who recover to being immune to both pathogen strains (henceforth strain transcending immunity). So far, I have displayed results where c = 0 and $\alpha = 0$.

My key result is that the protectiveness of a strain-specific HLA allele against the prevailing local infection caused by a multi-strain pathogen is correlated with its population frequency. I define the breakdown of this trend as the case where neither $OR_1$ nor $OR_2$ switch from above/below to below/above 1 over the range of $p_1$. Figure 2.10 illustrates when this breakdown occurs in the main two strain, two allele model as $c$ and $\alpha$ are varied. I also explore the impact of introducing a discrepancy in fitness between the pathogen strains in the model, achieved by varying the transmission parameter, $\beta$.




(a)

(b)

(c)

Figure 2.10: **Showcasing how co-infection and strain transcending immunity affects the outcome of the original model.** The three figures are of binary heat maps where a white indicates the trend disappearing, black indicates when the trend is still present. The parameter values are: $d = 0.01$, for (a) $\beta_1 = 0.061$, for (b) $\beta_1 = 0.065$, for (c) $\beta_1 = 0.08$, $\beta_2 = 0.06$, $\mu_i = 0.02$, $\sigma_i = 0.02$ $(i = 1, 2)$.

When there is only a small difference between $\beta_1$ and $\beta_2$ (figure 2.10a), the relationship between HLA allele frequency and protectiveness against infection exists at all levels of co-infection, and in the presence of some strain transcending immunity, but not when strain transcending immunity is complete ($\alpha = 1$). As the fitness difference between the strains is increased, the relationship between HLA allele frequency and the protectiveness of that allele against the prevailing local infection breaks down at lower levels of strain

transcending immunity, but this effect can be countered by allowing more co-infection to occur (figure 2.10b and 2.10c). Overall, the relationship I have identified requires that both strains of pathogen are able to co-circulate in the population, over at least some values of $p_1$. Both strains need to be present in order to drive the relationship between HLA frequency and protectiveness against local prevailing infection. Any process that promotes co-circulation (e.g. co-infection) makes such a relationship more likely, and any process which acts against co-circulation (e.g. fitness differences between strains; strain transcending immunity) makes such a relationship less likely.

## 2.4 Discussion

Population level differences in the protectiveness of HLAs against infection with multi-strain pathogens may be a result of HLAs being strain specific in their effects. Here I simulated such HLA-strain interactions in a epidemiological model. I showed that the adaptation of a multi strain pathogen to the HLAs of a host population will generate a negative association between the population frequency of an HLA allele which is beneficial against a specific pathogen strain and the protectiveness of that HLA type against the prevailing local infection. I also showed that if certain HLA types cause hosts to recover from infection with particular pathogen strains more quickly than other HLAs, those HLAs are less likely to protect against infection in general, since the strains against which those HLA types are particularly effective may be outcompeted by other pathogen strains in the population.

All of the simulations included HLA alleles/types which were "protective" in the sense that they conferred the ability to mount an immune response against a specific strain. However, despite these extreme HLA-strain associations, I found that under a great many circumstances no association would be detected between HLA type and infection in general in a typical case control study (figure 2.7 and figure S4). This is because the frequencies of the prevailing pathogen strains adapt to the HLA's present in the population. Previous efforts to use case control studies to identify HLA-infection associations for multi strain pathogens may have been hindered by such effects.

I presented results for two HLA allele and three HLA allele systems. Two and three

alleles are far fewer than the diversity of HLA alleles present in human populations (e.g. there are 100 HLA-B alleles observed in the German population Gonzalez-Galarza et al. [2010]). However, classical definitions of HLA alleles, based on the amino acid sequence of the HLA molecule, do not necessarily reflect functionally relevant properties. Class I HLA alleles are grouped into 10 HLA supertypes based on their binding capabilities (Sidney et al. [2008]). Associations between HLA supertypes and susceptibility to and severity of tuberculosis have been found Balamurugan et al. [2004]. Other HLA supertype associations have been found for HIV (Trachtenberg et al. [2003], MacDonald et al. [2000]). Functional grouping of HLA alleles has also been performed for specific pathogens such as the H1N1 influenza. Mukherjee *et al* classified the HLA diversity of human populations into "response types" which are HLA class I genotypes that share similar epitope binding pools to the H1N1 virus Mukherjee and Chandra [2014]. If I were to classify HLA alleles by whether or not they were able to bind a specific peptide sequence from an immmunodominant epitope in a particular pathogen, then there would only be two types of HLA: those that bind the peptide of interest and those that do not. The two allele and three allele systems (or more properly "two binding type" and "three binding type" systems) can therefore deliver insights into how HLA-strain systems operate, even if the simulated alleles are not directly equivalent to known HLA alleles.

I focused on the impact of strain-specific HLA types which I assumed to be mutually exclusive in their ability to protect against different pathogen strains (i.e. where the ability to display an immunogenic peptide from strain 1 precludes the ability to display an immunogenic peptide from strain 2). This has a precedent, albeit in a non-human system. Experiments in chickens have shown that single MHC types can be associated with completely opposite responses to different pathogen strains (McBride et al. [1981]). GB1 line chickens are susceptible to rous sarcoma virus (RSV) subgroup C PR-RSV strain, but resistant to rsv subgroup C B77 strain. GB2 line chickens (of a different MHC type Briles et al. [1982]) display exactly the opposite pattern (resistant to PR-RSV but susceptible to B77).

MHC/HLA molecules exhibit a range of binding properties. Some are specialist (binding only a narrow range of types of peptides), others are more generalist (able to bind

a wider range of peptides). It is possible that the maintenance of both generalist and specialist MHC/HLAs in populations may be because each are best adapted to respond to different types of pathogen (Chappell et al. [2015]). Chappell et al. [2015] highlight Marek's disease in chickens as an example of an infectious disease where generalist MHCs have been shown to provide the best responses, and HIV in humans as an example where specialist HLAs are associated with the most effective immune responses. Kaufman [2018] gives review of generalist and specalist MHC class I molecules. It might seem that, since generalists can present a wide range of peptides, specialist MHCs/HLAs are redundant and should not be maintained in populations. The continued existence of specialist MHCs/HLAs can be understood if we suppose that sometimes the best immune responses are associated with mounting an immune response against a very particular pathogen peptide. Even if a generalist MHC/HLA can present that peptide, it will not necessarily present it as reliably and consistently as a specialist MHC/HLA with narrower binding properties. As I described previously, my model applies to systems where it is beneficial for a host HLA to present a particular immunodominant peptide, but where not all HLAs are capable of doing this. This therefore implies the HLA types in my model are relatively narrow in their binding (although I do not mean this to imply they are specialised only to the hypothetical pathogen at hand). As illustrated in figure S2 , the inclusion of an HLA type capable of presenting peptides from any of the strains in the system (the "perfect allele", which could also be called a generalist) does not change the overall conclusions provided there are still at least 2 types of more specialist HLA in the system.

Figures 2.8 and 2.9 shows that it is plausible for a real world case control study to detect a significant effect of a strain specific HLA type on infection, driven by the processes modelled here. The only necessary conditions are that the recovery rate of the pathogen is faster than the background mortality rate of the host and the frequency of the HLA type is within a certain range. However, defining "HLA type" is problematic, since the functionally relevant "HLA type" for a particular pathogen might in fact be a set of HLAs encoded by a range of different alleles, which share the property of being able to bind a critically important (and unknown to us) pathogen peptide. Nevertheless, my model

suggests a method to identify such systems. If the odds ratio for being infected with a disease caused by a multi-strain pathogen changes for the same HLA type or supertype in different populations, and if there is a positive relationship between that odds ratio and the frequency of that HLA type or supertype, this would be highly suggestive that there are HLA -pathogen strain associations to be found.

In the introduction I noted three pathogens where my model is likely to be particularly relevant: *Neisseria meningitides, Streptococcus pneumoniae* and *Plasmodium falciparum* malaria. Existing work attempting to link the risk of disease caused by these pathogens to host genetics focuses on severe disease outcomes (invasive meningococcal or pneumococcal disease or severe malaria), rather than infection (or "carriage") per se. My model does not attempt to simulate the development of severe infection. However, if severe disease is associated with particular strains, and immunity to strains outside of this subset does not prevent severe disease, my model can be interpreted in terms of how HLAs impact the accumulation of immunity to just those strains that are capable of causing severe disease. For *Streptococcus pneumoniae* and *Neisseria meningitides* it is widely accepted that only a subset of strains cause severe disease (Enright and Spratt [1998], Peltola [1983]). The situation for *P. falciparum* is complicated by the antigenic variation *P. falciparum* exhibits during infection, but the presence of particular group A var genes in the genomes of parasites may determine their potential to cause severe disease (Bull et al. [2008]). Although the above pathogens have different life histories they all share the common attribute of being multi strained pathogens, which is represented in this model. Here we have shown, for a variety of pathogen parameters, the effects of HLA on the odds ratio. I believe the results presented here can be applied to such different pathogens. I propose that a correlation between population frequency of an HLA type and the odds ratio for severe disease associated with the presence of that HLA type would be suggestive of an HLA-strain relationship in any of these systems.

My model fixed HLA frequencies within each simulation which I believe to be reasonable given that we are looking into the mechanisms behind case control studies which in

themselves are snap shots in time in terms of a populations genetic landscape. However Over the longer term, HLA frequencies themselves must be evolving under selection from pathogens (Jeffery and Bangham [2000], Prugnolle et al. [2005], Hertz et al. [2011]). Extending the model into a co-evolutionary framework (similar to those explored by Penman Penman et al. [2013] and MacPherson MacPherson et al. [2018]) would deliver further insights, especially into long term HLA supertype dynamics. If I allowed pathogen induced mortality we could see our model tend to stable equilibrium or even perhaps cyclic behaviours. If we found that a stable equilibrium existed for both HLA alleles being present the pathogen strain dynamics would also stabilise and the odds ratio would remain constant for both alleles. If we observed cyclic behaviour in terms of the HLA frequencies then again pathogen strain dynamics would be cyclic and we would also see the odds ratio to be cyclic following observations I have seen in this model. I could also increase the complexity of the model further to include for more alleles or multiple linked HLA allele loci, which previous modelling work has shown to enable a range of co-evolutionary outcomes (Penman et al. [2013]).

I have explored one way in which epidemiology may affect our ability to detect the importance of HLA type in infectious disease. However, other processes can also cause case control studies to generate conflicting results, including within-host adaptation of chronic viral diseases and epistasis between HLAs and other loci. Within-host adaptation of HIV to escape HLA restricted immune responses has spilled over into population level adaptation that renders previously protective HLA alleles non protective (Kawashima et al. [2009], Payne et al. [2014]). HLA's are also known to interact epistatically with Killer-cell immunoglobulin-like receptors (KIRs) (Khakoo et al. [2004], Martin et al. [2002]), and failing to account for KIRs genotype could also lead to HLA-infectious disease associations being overlooked.

Multi-strain pathogens include infections of vast public health significance such as malaria, TB, influenza, and streptococcal and meningococcal bacterial disease. A deeper understanding of the immunogenetics of such infections could pave the way to develop more personalised vaccines or other control measures. However, finding associations between

HLA alleles and these infectious diseases has been an ongoing challenge. The model I present here explores the consequences of one potential mechanism of HLA-pathogen interaction and suggests a method to detect the signature of HLA-strain relationships in combined analyses of case control studies in different populations.

# Chapter 3

# The Evolution of copy number variation of the MHC

## 3.1 Introduction

As mentioned in section 1.1.2 variation in the number of MHC genes present in a particular MHC cluster is typically referred to as "copy number variation" (CNV) (Freeman et al. [2006]), even though strictly speaking, multiple copies of an MHC gene (e.g. mamu-A1 and mamu-A4 in Indian Rhesus Macaques) may not be identical copies of each other. For simplicity, and to be in keeping with other studies I too will refer to the existence of multiple copies of a MHC locus as CNV. Within humans, HLA class I genes have no CNV: if we consider HLA-A ,HLA-B, and HLA-C, all humans have one copy of each of these genes. However, equivalent genes in rhesus macaque (Mamu-A and Mamu-B) exhibit CNV (Otting et al. [2005], Otting et al. [2007]). In the case of Mamu-B up to 18 functional Mamu-B-like genes have been identified (Daza-Vamenta et al. [2004]). Otting et al. [2007] shows that CNV for mamu-A existed before the speciation of the Rhesus macaque and the Cynomolgus macaque, suggesting this state may be evolutionarily stable and has certainly persisted for an extremely long time.

In cattle there are 6 documented functional MHC class I genes that have varying degrees of presence between different haplotypes (Ellis and Ballingall [1999], Hammond

et al. [2012]). Tasmanian devils have been found to have varying numbers of MHC class I sequences ranging from 2 to 7 per individual (Siddle et al. [2010]). Swine have 6 classical class I genes of which SLA-1 and SLA-3 have shown evidence of CNV (Tanaka-Matsuda et al. [2009]). For the mouse certain functional H2 (MHC for the mouse) loci are and are not present on different haplotypes such as H2-Eb and H2-Ea (Stuart [2015]).

There are four major mechanisms that can cause a gene to be duplicated *Unequal crossing over*, *Retroposition*, *Duplicative transposition* and *Polyploidization* (Magadum et al. [2013]). Otting et al. [2005] note for the rhesus macaque the variation in gene numbers for the Mamu-A and Mamu-B is probably generated by unequal crossing over. Different haplotypes of the mouse have been found to have varying numbers of MHC class I genes (Stuart [2015]). Further work looking into recombination events on the mouse MHC found 11 independent recombination events with seven of those being unequal crossing over recombination events (Steinmetz et al. [1986]). However, unequal crossing over (or any other molecular mechanism duplicating genes) merely provides the raw material i.e. different number of copies of genes present on chromosomes. I speculate that natural selection, due to the interaction between MHC molecules and pathogens, will affect the fate of copy number variants within the MHC, and in this study I explore how such mechanisms could play out.

Recombination break points for recombination events where the resulting recombinant persists in the population normally occur in the regions between genes, however there is evidence that such break points can occur between exons of an MHC gene. Schwartz and Hammond [2015] compares intronic and exonic sequences of cattle MHC class I genes and note that the 6 MHC class I genes clade together after the $\alpha 2$ domain. They note that the binding cleft domains show no relationship and that it is likely that there have been recombination breakpoints at introns between the $\alpha 1$ and $\alpha 2$ exon domains. If unequal crossing over occurs between such intronic break points then it is conceivable that copy number variation could be generated at the same time as creating new combinations of exons (i.e. a new MHC gene with potentially different functional properties). I explore the possible impact of such a mechanism on CNV within the MHC in my analysis.

There have been other theoretical treatments of recombination and copy number variation. The earliest works I mentioned were Krüger and Vogel [1975] who created models

to explore how unequal crossing over recombination affects the number of copies of an allele. Takahata [1981] creates a similar model to Krüger and Vogel [1975] however it incorporated sister chromatid exchange. Since Krüger and Vogel [1975] multiple models of unequal crossing have been analysed (Takahata [1981], Baake [2008], Shpak and Atteson [2002], Redner and Baake [2004]). Bentkowski and Radwan [2019] is a model that does not explore unequal crossing over but does use an individual based model to analyse how the number of copies of MHC genes evolve due to varying pathogen diversity. They found that a higher diversity of pathogens only led to more copies of MHC genes if the cost of having more genes was low. Bentkowski and Radwan [2019] also found that when the mutation rates of pathogens were increased when there is a low number of pathogens present, it caused an increased number of MHC gene copies.

Here, I present an individual based model that explores the interaction of MHC molecules with pathogens and how this interaction may affect the CNV of MHC genes within a population. I explicitly simulate the process of unequal crossing over recombination to lengthen and shrink the MHC cluster. I compare two possible rules for calculating a host's fitness based on its MHC genotype. I demonstrate a variety of evolutionary outcomes and detail the reasons why these outcomes occur.

## 3.2 Methods

Here I will describe the individual based model I designed to gain insight into MHC gene number variation. We use an individual based model due to the complicated host types we anicicpate when dealing with varying numbers of loci. The model presented here is essentially a Wright-Fisher model where the fitness of each host is determined by their MHC alleles.

### 3.2.1 Population

The population is represented by a matrix of integers $G$. Each two rows in the matrix represents an organism in the population. The first row is the maternal chromosome and the second row is the paternal chromosome. An integer in the matrix represents an MHC exon. A pair of exons and their combination represents a unique MHC sequence, thus

every two columns is a potential MHC gene. The number of copies of MHC genes present can change by having zeros occupy positions in the matrix; only a pair of non-zero integers indicates the presence of an MHC gene. Note that I intend this model to simulate just one MHC gene type, e.g. a gene encoding an MHC-A type protein, where in some species there is just one MHC-A present per chromosome, but in others there are multiple copies of MHC-A (each of which may vary slightly from each other), per chromosome.

I set a maximum possible length of the MHC cluster in my simulations of $U_L$ thus the maximum number of columns in G is $2U_L$. The size of the population in each generation is a fixed size $N$, and so the number of rows of G is 2N. The number of MHC genes in the clusters in the population (i.e. the length of the clusters) can change via unequal crossing over recombination events (described in more detail in section 3.2.3). Mutations can also occur at the individual exon level, introducing extra diversity into the system (described in more detail in section 3.2.3). My representation of MHC genes as pairs of exons is necessary in order to explore the additional phenomenon of recombination events occurring between exons (e.g. as has been observed in cattle Schwartz and Hammond [2015]).

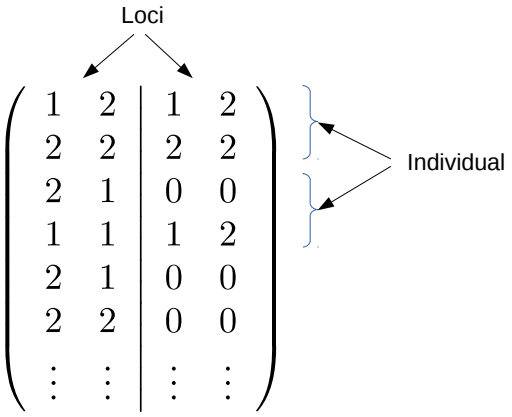



Figure 3.1: **An example of the matrix $G$ that represents the population.** The numbers 1 and 2 represent different exons and the combinations of them as pairs represent MHC genes. Within this simple example, 4 MHC sequence variants are possible (11, 12, 21, 22). Note that the identities of "left" and "right" exons are entirely independent: exon 1 on the "left" is not the same exon as exon 1 on the "right" hence a sequence of exons 1 and 2 is different to a sequence of exons 2 and 1. 0 represents an empty space, thus three of the chromosomes in this diagram contain only 1 MHC gene.

### 3.2.2 Reproduction

Each new generation in the model is generated from the previous generation, meaning individuals inherit maternal and paternal chromosomes from the previous generation. Pairs of parents are selected randomly, but the probability of being a parent (i.e. passing on a chromosome) is determined by the individual fitness of a genotype in the parental generation.

An individual's fitness is denoted as $S_n$ ($n$ denotes the $n$th individual in the population) which is a number between 0 and 1, the higher $S_n$ the more likely an individual is going to be selected to pass on its genes. The value $S_n$ takes is determined by MHC genotype, based on rules governing the fitness contributions of individual MHC sequences and how these combine to generate the contributions of a genotype as a whole (sections

3.2.2.1 and 3.2.2.2).

Figure 3.2 illustrates the parent selecting process. When we are selecting parents for the next generation I first calculate the sum of the fitnesses of all individuals in the population ($U(0, S)$, where $S = \sum_{n=1}^{N} S_n$). I then assign each individual in the population a certain range of the values between 0 and $S$ ; the size of this region is equal to that individual's fitness. I generate a uniform random number between 0 and $S$ and whichever individual's region that number corresponds to is selected to become one of the parents of the next generation.

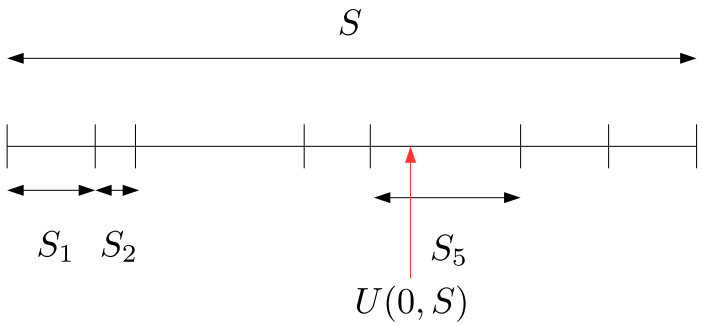



Figure 3.2: **An illustration of how $U(0, S)$ is used to select a parent.** The fitter an individual the larger its region will cover $S$ giving it a higher chance for $U(0, S)$ to be a value within that individual's region. In this example individual 5 is being selected as a parent.

#### 3.2.2.1 Gene-level Fitness Contributions

We are considering an MHC cluster containing one or more copies of a particular MHC gene. The terminology of how to describe different variants of that gene, which can exist in different combinations in chromosomal clusters of different lengths, is challenging. I will refer to individual variants of the MHC gene as "sequence variants", where a particular sequence variant in the system has the potential to appear more than once on a single chromosome, alongside different combinations of other sequence variants of the same gene, which have their origins in unequal crossing over and subsequent mutation

events. Combinations of exon pairs determine the identity of sequence variants, these are denoted as $[ij]$ where $i$ and $j$ are exon $i$ and exon $j$. For each combination of $i$ and $j$ I give the sequence variant $[ij]$ a raw fitness contribution ($A_{ij}$). $A_{ij}$ is generated uniformly between 0 and 1 ($U(0,1)$), and is meant to represent that sequence's fundamental ability to help a host survive. If $A_{ij}$ is a high number this implies $ij$ may have an inherent advantage against a wide range of pathogens; if $A_{ij}$ is a low number this implies $ij$ might lead to adverse consequences for the host (e.g. perhaps even a predisposition to severe autoimmune disease).

Various measures of MHC diversity concur with the hypothesis that negative frequency dependence occurs at the allelic level (Aguilar et al. [2004], Alter et al. [2017], Hawley and Fleischer [2012]), thus I also include negative frequency dependence in the model. A sequence variant's frequency $p_{ij}$ is calculated as the proportion of individuals who have the sequence variant. I compute the fitness contribution of an individual sequence variant as follows: $f_{ij} = A_{ij}(1 - p_{ij})$. A sequence variant makes a greater positive contribution to an individual's fitness when it has a low presence in the population.

#### 3.2.2.2 Individual Fitness

When calculating the fitness of individual members of the population, I use two different calculations to generate an individual's fitness from the fitness contributions of the MHC sequences in that individual's genotype.:

1. The fitness of an individual is equal to the fitness contribution of the most advantageous MHC sequence in the genotype as a whole (i.e. including both chromosomes). (max $f_{ij}$)

2. The fitness of an individual is equal to the mean fitness contribution of their chromosomes, where the fitness contribution of each chromosome is the mean fitness contribution of the sequences present on that chromosome. (mean $f_{ij}$)

I call the first fitness calculation the *Max* fitness rule and the second calculation the *Mean* fitness rule. The max fitness rule implies that being able to express just one well adapted variant of the MHC gene of interest will increase the fitness of an individual. The mean

fitness rule captures a case where most or all variants of the MHC gene present in the genome need to be well adapted in order for an individual to be considered fit.

These two fitness calculations view the MHC in very different ways. Exactly how the different properties of MHC alleles and genes combine to generate an overall fitness of an individual is still unknown. I regard the max rule and the mean rule as informative alternatives within a spectrum of possibilities, and will compare the results of each throughout my analysis. I also have a universal rule for all fitness calculations which is that when the number of MHC sequences in a gene cluster passes a certain number ($U_L$) the fitness of that individual is 0. This is meant to represent biological disadvantages to having too many gene sequence variants for the MHC such as auto-immune diseases (Yim et al. [2015]) or an over saturated MHC molecule pool. In the appendix B.3 I look into the affects of this boundary condition (when the number of gene sequences goes above $U_L$). I investigate the effects of the above boundary condition as well as having a gradual decrease in fitness of an individual when the number of gene copies goes above $U_L$ (which I consider a softer boundary).

### 3.2.3 Mutations and Recombination

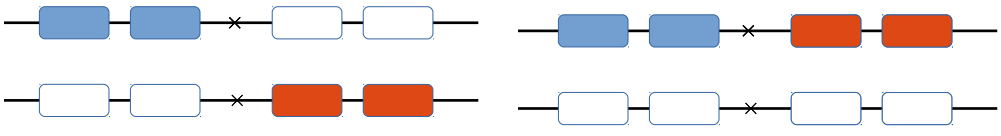
Mutations in my model occur at the level of individual exons. $p_M$ is the per individual exon probability an exon mutation will occur when a parent's chromosome is inherited.

When a mutation occurs I randomly select an exon which will mutate and I create a new exon to replace the old one. This introduces new possible combinations of exons $(ij)$ and hence new potential MHC sequences into the system. I generate new raw fitness contributions ($A_{ij}$) specific to each potential new sequence variant by selecting a number from a uniform distribution between 0 and 1 for every possible combination of the new exon with all the other exons in the system.

Recombination in my model can happen at sites located between MHC genes, or between exons. Recombination at sites between genes leads to new combinations of sequence variants on chromosomes, or to new numbers of sequence variants on chromosomes (in the case of unequal crossing over). Recombination at sites between exons leads to both of the above phenomena but can also simultaneously create a new MHC sequence by putting together a new combination of exons. The proportion of recombination break

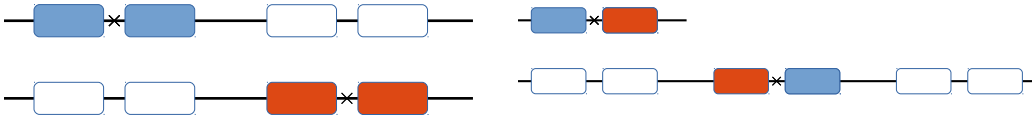
points that happen between exons compared to between genes in the model are denoted as $\beta$. When $\beta = 0$ all recombination sites happen between genes and when $\beta = 1$ all recombination sites happen between the exons of a gene. The possibility of recombination occurring between exons is motivated by the observation of this phenomenon in the cattle MHC (Schwartz and Hammond [2015]). For simplicity I assume no more than one recombination event in the gene region is occurring per parent. I assign a probability of recombination occurring in any given individual as $p_R L$. This means that the probability of recombination occurring is proportional to the number of genes in the MHC cluster (i.e. the length of the cluster). Recombination is a process involving two chromosomes, the length I use to determine the probability of recombination needs to be chosen carefully. I use the length of the shorter MHC cluster in the parent in question to determine the probability of recombination occurring in that parent (i.e. $L$ is the shorter of the two possible values). This is on the assumption that, for recombination to occur, the chromosomes must align in a way that allows crossing over between them, and the possibility of this occurring at an appropriate site will be limited by the length of the shorter MHC cluster in the pair.

Figures 3.3 and 3.4 illustrate the mechanics of between-gene recombination (Fig 3.3) and between-exon recombination (Fig 3.4) within the model. For between-gene recombination the break points are located between exons in an even column number on the left and exons with an odd column number on the right. For between-exon recombination the break points are located between exons with an odd number column on the left and an even exon column number on the right (figure 3.4). For both types of recombination I select where the break point will occur randomly for both the maternal and paternal chromosome. When the break points are decided I swap the exons on the right hand side of break point on the maternal chromosome with the exons on the right hand side of the paternal chromosome (figures 3.3 and 3.4).



(a) (b)

Figure 3.3: **Diagram illustrating allele recombination**. The black crosses represent break points and only occur in-between alleles (a) is a diagram of chromosomes before recombination and (b) is a diagram after recombination.



(a) (b)

Figure 3.4: **Diagram illustrating exon recombination**. The black crosses represent break points and only occur in-between exons (a) is a diagram of chromosomes before recombination and (b) is a diagram after recombination.

### 3.2.4 Changing Pathogen Selection

For simplicity, I do not explicitly simulate a co-evolving pathogen population. The assumption of negative frequency dependence implies some host-pathogen coevolution, whereby common MHC sequences become disadvantaged by pathogens adapting to evade their immune recognition. However, I also wanted to include additional possible fluctuations in the selective pressure from pathogens. I therefore also introduced the possibility of changing pathogen selection (CPS), by allowing the raw fitness contribution ($A_{ij}$) of each MHC sequence variant to change according to the following rules.

Each generation I take a proportion of the sequence variants within the population ($\alpha$) and assign them new raw fitness contribution values ($A_{ij}$). For each sequence variant I generate a new value of $A_{ij}$ from a truncated normal distribution which has a mean equal to that MHC sequence's $A_{ij}$ in the previous generation and a standard deviation of $\sigma$. This is meant to capture the possibility of entirely new pathogens arriving in the

population, against which certain MHC variants may now be advantageous or disadvantageous. However by making the truncated normal distributions mean equal to the MHC sequence previous value of $A_{ij}$ I allow the change to be influenced by how well adapted the MHC sequence was before the new pathogen came into the population. As an already well adapted MHC sequence is more likely to be still relatively well adapted in the new pathogen climate since the pathogens against which the MHC was originally advantageous are likely to still also be present. By changing the variance of this truncated normal distribution I can alter the probability of large changes in $A_{ij}$ occurring. The bigger the variance, the greater the chance of an MHC variant experiencing a big change in $A_{ij}$ from generation to generation.

### 3.2.5 Parameter Values

Due to computational limitations we have fixed some parameters (see table 3.1). Justification for why these value were chose can be found below. Throughout the analysis I prioritised varying other parameters (such as $\alpha$).

| Parameter | Description | Values used | Figures |
|---|---|---|---|
| $N$ | Population size | 5000 | All Figures |
| $I_L$ | Initial number of Loci | 2 | everything past section 3.3.2 |
| $I_A$ | Initial number of Alleles | 25 | everything past section 3.3.2 |
| $U_L$ | Maximum number of gene sequences allowed in a gene cluster | 5 | everything past section 3.3.2 |
| $p_R$ | The probability a recombination event will occur per individual per loci | $10^{-4}$ | everything past section 3.3.2 |
| $p_M$ | The probability a mutation event will occur per individual per loci | $10^{-6}$ | everything past section 3.3.2 |

Table 3.1: **Parameter values which were fixed throughout the results section.**

Here is a list of justifications for the value assigned to each parameter.

$N$: I needed a population which was computationally feasible to simulate but also likely to be biologically relevant.

$I_L$: see section 3.3.1

$I_A$: I wanted to start the system with a reasonable number of alleles for an MHC locus.

$U_L$: this is a plausible value for the organisms that have exhibited CNV (Otting et al. [2007]).

$p_R$: I needed the rate of recombination to be high enough for CNV to occur over a reasonable time scale as I have only so much computational power. Sperm typing studies have directly observed recombination events in hotpots, where they have seen recombination events $5 \times 10^{-6}$ to $2.4 \times 10^{-5}$ of the time (Clark et al. [2007]). The number used here is higher than this results but I needed to select a value for simplicity (and computational expedience)

$p_M$: vertebrate per-nucleotide mutation rates are of order $10^{-9}$ and $10^{-8}$ per generation. MHC class I molecule variable chains (the alpha chain -see figure 1.1) are approximately 300 amino acids long , which means they are encoded by approximately 900 nucleotides. However, mutations in some of those 900 will be silent (not change amino acids) and others will not change the functional properties of the protein. The chosen rate ($10^{-6}$) is at the higher end of what is biologically reasonable, for computational efficiency.

Other parameter values are varied throughout the results section. For a full list of parameters see section table B.1 in the Appendix B.1.

### 3.2.6 Measuring Stability

My analysis concerns the number of MHC sequences on a chromosome ( $L$). To determine whether I have run a simulation long enough for this quantity to stabilise, I look at the behaviour of $\bar{L}$ : the mean value of $L$ in the population.

During the simulations I record 50 time steps for the last 25% of generations. I split these 50 recorded time steps into two groups of 25 (a first half and a second half). For each group of 25 time steps I calculate the mean $\bar{L}$ and I take the difference between these two mean values of $\bar{L}$ and notate this as $\bar{L}_{diff}$. For each parameter set I run 100 simulations giving us a distribution of a 100 samples of $\bar{L}_{diff}$. I then do a one sample t-test to determine whether or not this distribution has a mean of zero. If the resulting $p$ value is above 0.05 I assume that the simulations are not trending in any particular direction and the duration of the simulations is sufficient for the lengths of the clusters to have reached a stable distribution. As shown in table 3.2, after 1,000,000 generations all but three of the scenarios had stabilised. For the three which had not, the change in $\bar{L}$ over the final 25% of the generations is so small that I do not expect it to change much until its steady state and are satisfied with presenting the results for those two scenarios. In appendix B, figure B.1 I illustrate the time series of the distribution of $\bar{L}$ (section B.2) for selected parameter sets. In figure B.1 (a) is the only one where $\bar{L}$ has not stabilised, where as in (b), (c) and (d) $\bar{L}$ has stabilised.

| Gen | FR | $\beta$ | $\alpha$ | $\sigma$ | $\bar{L}$ $p$ value | $\bar{L}$ pass | $\langle \bar{L}_{diff} \rangle$ |
|---|---|---|---|---|---|---|---|
| 1000000 | Null | 0 | 0 | 0 | 3.57e-01 | ✓ | 4.66e-02 |
| 1000000 | mean | 0 | 0 | 0 | 7.53e-03 | ✗ | 6.07e-02 |
| 1000000 | mean | 0 | 0.01 | 0.01 | 9.358e-01 | ✓ | 1.93e-03 |
| 1000000 | mean | 0 | 0.01 | 0.1 | 3.69e-01 | ✓ | 4.66e-03 |
| 1000000 | mean | 0.0001 | 0 | 0 | 6.63e-02 | ✓ | 4.24e-02 |
| 1000000 | mean | 0.0001 | 0.01 | 0.1 | 1.45e-01 | ✓ | 5.96e-03 |
| 1000000 | mean | 0.001 | 0 | 0 | 7.35e-03 | ✗ | 6.41e-02 |
| 1000000 | mean | 0.001 | 0.01 | 0.1 | 9.47e-01 | ✓ | 2.96e-04 |
| 1000000 | mean | 0.01 | 0 | 0 | 1.06e-01 | ✓ | 3.65e-02 |
| 1000000 | mean | 0.01 | 0.01 | 0.1 | 1.67e-01 | ✓ | 6.93e-03 |
| 1000000 | mean | 0.1 | 0 | 0 | 1.64e-01 | ✓ | 3.03e-02 |
| 1000000 | mean | 0.1 | 0.01 | 0.1 | 5.47e-01 | ✓ | 2.25e-03 |
| 1000000 | mean | 0.5 | 0 | 0 | 2.41e-01 | ✓ | 2.89e-02 |
| 1000000 | mean | 0.5 | 0.01 | 0.1 | 1.49e-01 | ✓ | 4.14e-03 |
| 1000000 | mean | 1 | 0 | 0 | 2.65e-08 | ✗ | 1.61e-02 |
| 1000000 | mean | 1 | 0.01 | 0.1 | NaN | ✓ | 0 |
| 1000000 | max | 0 | 0 | 0 | 5.43e-01 | ✓ | 5.09e-03 |
| 1000000 | max | 0 | 0.01 | 0.01 | 8.80e-01 | ✓ | 1.41e-03 |
| 1000000 | max | 0 | 0.01 | 0.1 | 7.01e-01 | ✓ | 2.98e-03 |
| 1000000 | max | 0.0001 | 0 | 0 | 7.10e-01 | ✓ | 2.94e-03 |
| 1000000 | max | 0.0001 | 0.01 | 0.1 | 5.54e-02 | ✓ | 1.92e-02 |
| 1000000 | max | 0.001 | 0 | 0 | 6.90e-01 | ✓ | 3.35e-03 |
| 1000000 | max | 0.001 | 0.01 | 0.1 | 5.75e-01 | ✓ | 5.20e-03 |
| 1000000 | max | 0.01 | 0 | 0 | 2.30e-01 | ✓ | 8.90e-03 |
| 1000000 | max | 0.01 | 0.01 | 0.1 | 9.04e-01 | ✓ | 1.35e-03 |
| 1000000 | max | 0.1 | 0 | 0 | 7.47e-01 | ✓ | 2.56e-03 |
| 1000000 | max | 0.1 | 0.01 | 0.1 | 8.47e-01 | ✓ | 2.51e-03 |
| 1000000 | max | 0.5 | 0 | 0 | 1.78e-01 | ✓ | 1.24e-02 |
| 1000000 | max | 0.5 | 0.01 | 0.1 | 1.27e-01 | ✓ | 1.95e-02 |
| 1000000 | max | 1 | 0 | 0 | 6.13e-01 | ✓ | 4.68e-03 |
| 1000000 | max | 1 | 0.01 | 0.1 | 1.09e-01 | ✓ | 4.86e-02 |

Table 3.2: **Table that summarises the stability test done for parameter set used in the results section** The columns represent the following quantities: Gen is the the number of generations measured for the stability test; FR is the fitness rule I used at the individual fitness level; $\beta$ is the proportion of recombination events that are between the exons of an MHC sequence; $\alpha$ is the proportion of alleles that are affected by CPS each generation and $\sigma$ is the standard deviation of the truncated normal distribution I use to draw the new raw fitness contribution of the allele when its raw fitness contribution is changed by CPS. $\langle \bar{L}_{diff} \rangle$ is the average value of $\bar{L}_{diff}$ over the 100 simulations.

## 3.3 Results

### 3.3.1 *In the absence of pathogen mediated selection, length of cluster is determined by a combination of (i) the probability of recombination occurring in any given generation and (ii) the upper limit to the size of the cluster.*

We first explore how the mean number of sequence variants of an MHC in a cluster $\bar{L}$ behaves in the absence of pathogen mediated selection. I therefore applied no selection rules to choose each pair of parents, other than if chromosomes longer than the maximum size of $U_L$ or of length zero were created by recombination during the reproduction step, they would have no chance of being selected as a parent.

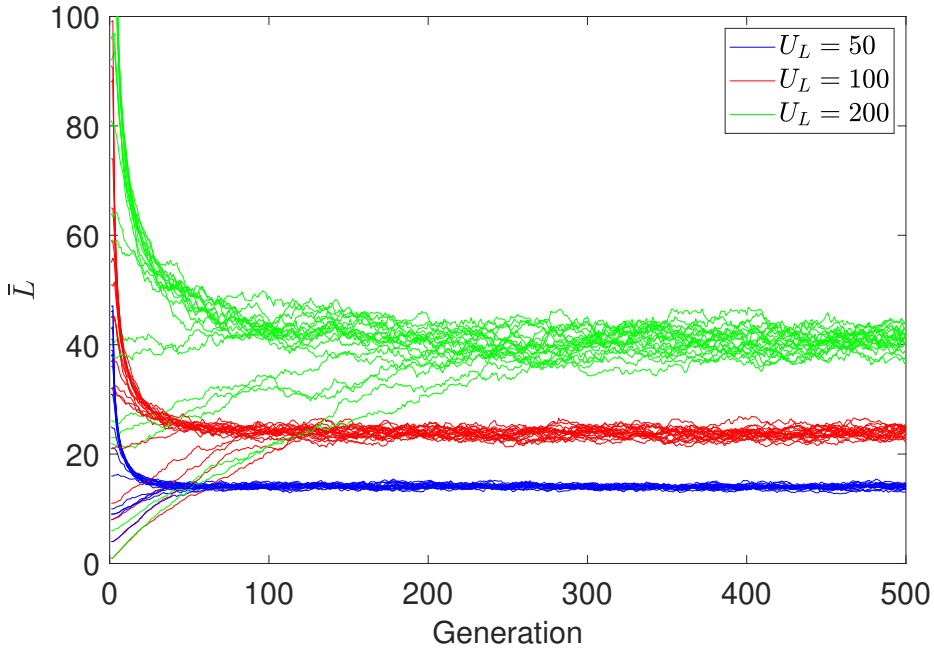



Figure 3.5: **Mean lengths of MHC clusters ($\bar{L}$) in the absence of pathogen mediated selection for each simulation.** Each individual line is the $\bar{L}$ for a single simulation. The parameters used are as follows: N = 5000, $p_R = 1$ and $\beta = 0$. For the blue lines $U_L = 50$, red lines $U_L = 100$, green lines $U_L = 200$.

The system stabilises to specific distributions of values for $\bar{L}$, depending on the upper limit for the length of the cluster and the probability of recombination happening each generation. This means that parameters : $U_L$, $p_R$ and $N$ all combine to determine the length of MHC clusters in the model. I regard this as the null state of the model, which may be altered by including different possible rules about how the presence of different MHC alleles affect the fitness of a genotype.

For the simulations shown in the remainder of the results section I have selected population size (N) of 5000, $p_R = 10^{-4}$ and an upper limit of $U_L = 5$. With these values, $\bar{L}$ stabilises on a value 2 if determined solely by recombination and $U_L$. I initiate the population with $\bar{L} = 2$ to determine if when I apply pathogen mediated and other forms of selection into the model, $\bar{L}$ is shifted from its expected null value.

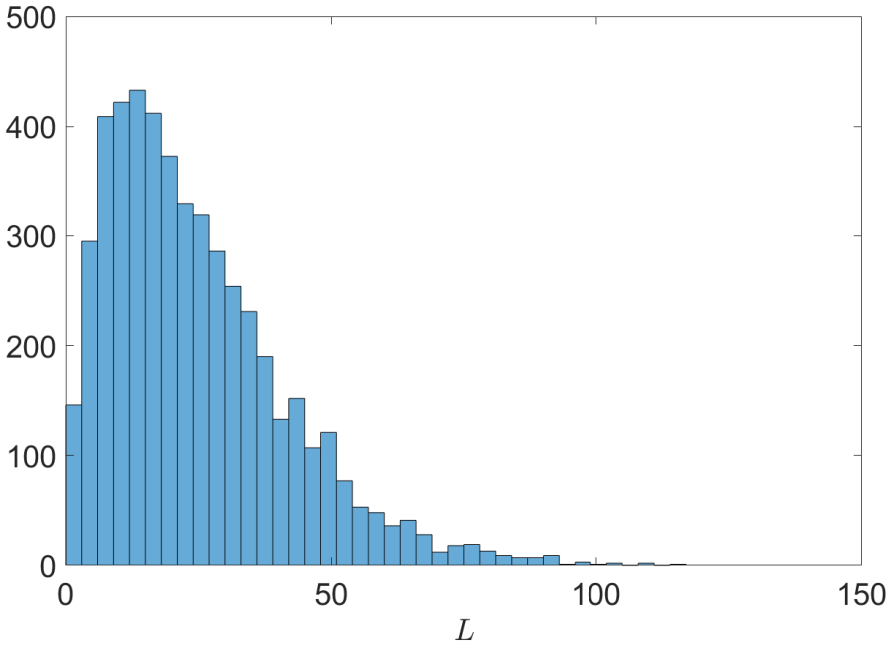



Figure 3.6: **Lengths of MHC clusters ($L$) in the absence of pathogen mediated selection for each simulation.** This is a histogram showing the distribution of $L$ for the final generation of one simulation. The parameters used are as follows: N = 5000, $p_R = 1$ and $\beta = 0$ and $U_L = 100$.

Figure 3.6 shows how $L$ was distributed in the population. We see that all values of $L$ exist within the population with a decrease towards the upper limit ($U_L = 100$). Some results do lie above $L = 100$ as recombination events can still cause this however these individuals would be given a fitness of 0. The values for $\bar{L}$ we see in figure 3.5 are the mean values of distributions like in figure 3.6. We use higher values for $U_L$ in this section than we do in the rest of the results as it gives a clearer picture of the behaviour of $L$ when there are more states of $L$ to occupy. In the rest of the results we use $U_L = 5$, which I consider to be biologically more viable.

### 3.3.2 *A "mean fitness" rule and changing pathogen selection pressure encourages short MHC clusters.*

The first fitness rule I will test is the "mean fitness" rule described in section 3.2.2.2 of the methods.

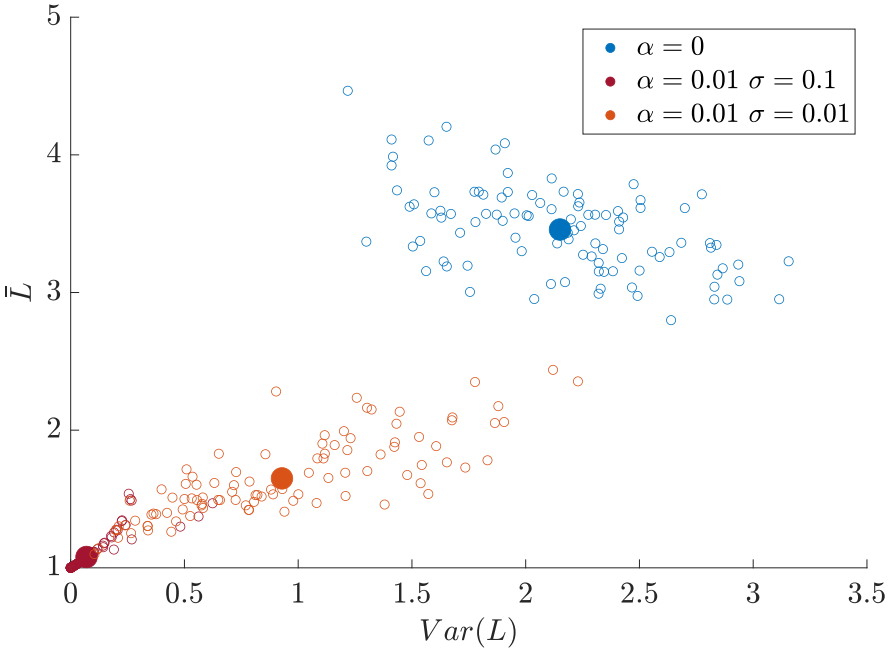



Figure 3.7: **Mean and variance of the lengths of MHC clusters under the mean fitness rule.** The empty markers are the value of $\bar{L}$ and $Var(L)$ for each individual simulation. The filled in and larger marker is the average value of $\bar{L}$ and $Var(L)$ over the 100 simulations. The parameters used are as follows: N = 5000, Initial number of exons = 10, Initial number of Loci = 2, $U_L = 5$, $p_R = 10^{-4}$ and $p_M = 10^{-6}$. For the blue markers $\alpha = 0$, for the red marker $\alpha = 0.01$ and $\sigma = 0.1$. Simulation with blue markers has not passed the stability test see table 3.2

In the absence of CPS, the mean fitness rule selects for greater numbers of sequence variants of an MHC gene in a cluster than the $\bar{L} = 2$ expected under the null model (figure 3.7, blue markers). This increase in $\bar{L}$ only occurs over very long time scales and is disrupted by any process which encourages changes in the fitness contributions of

MHC sequence variants (e.g. CPS or higher mutation rates). I therefore do not regard this result as particularly biologically plausible.

As CPS of differing levels of intensity is applied, the mean fitness rule selects for lower and lower values of $\bar{L}$ (figure 3.7 orange and red markers). The mean fitness rule allows individuals to have a high fitness only if most or all of the MHC sequences on their chromosomes are well adapted. If MHC sequence variants' fitness contributions change over time (due to the introduction of new pathogens), then longer clusters are more likely to lead to a drop in individual fitness, simply because they have more sequence variants present which could have undergone a detrimental change. Short clusters, containing whichever sequence variants happen to be best adapted in that generation, will tend to dominate the system. The value of $\bar{L}$ is strongly related to $\sigma$. $\sigma$ determines how big a change CPS tends to induce in MHC sequence variants properties. The greater the value of $\sigma$, the bigger the change in fitness contribution an MHC sequence variant could experience from generation to generation. Thus, figure 6 demonstrates that any process which tends to increase the magnitude of fitness contribution changes for MHC sequence variants from generation to generation leads to shorter chromosomes under the mean fitness rule.

$\bar{L}$ tells us about trends in the numbers of sequence variants of an MHC gene present within a single cluster. However, to consider possible variation in MHC cluster length within populations I also wish to consider the variance of $L$. Figure 3.7 demonstrates that the stronger the selection for shorter MHC clusters, the less variation observed in $L$. In Appendix B section B.3 I show that any results with lower $\bar{L}$ tend to also have low variation in $L$ due to boundary conditions on $L$.

### 3.3.3 *A "maximum fitness" rule encourages long clusters, especially in the presence of changing pathogen selection pressure*

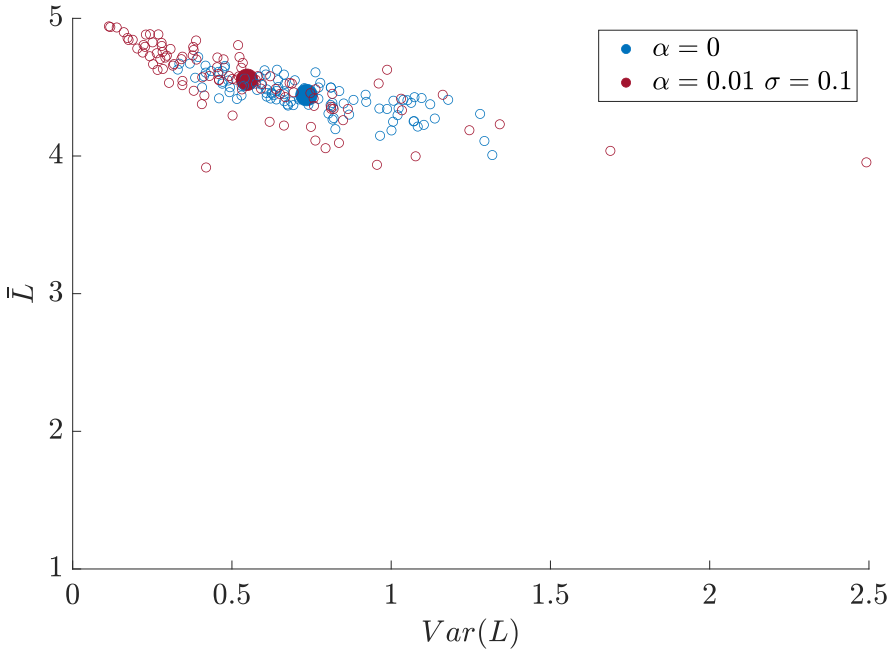



Figure 3.8: **Mean and variance of the lengths of MHC clusters under the maximum fitness rule.** The empty markers are the value of $\bar{L}$ and $Var(L)$ for each individual simulation. The filled in and larger marker is the average value of $\bar{L}$ and $Var(L)$ over the 100 simulations. The parameters used are as follows: N = 5000, Initial number of exons = 10, Initial number of Loci = 2, $U_L = 5$, $p_R = 10^{-4}$ and $p_M = 10^{-6}$. For the blue markers $\alpha = 0$, for the red marker $\alpha = 0.01$ and $\sigma = 0.1$

Under the maximum fitness rule, all scenarios are above $\bar{L} = 3$. The maximum fitness rule allows individuals to have a high fitness so long as at least one sequence is well adapted. Under the maximum fitness rule, if a biological process such as mutation, CPS or a change in allele frequency causes a reduction in the fitness contribution of a particular MHC sequence within a chromosome, the fitness of the individual will only be affected if the relevant MHC sequence had previously been the best adapted within the whole cluster.

However, if the change happens to increase the fitness contribution of an MHC sequence, under the maximum fitness rule any such change has the potential to enhance the fitness of the individual. Thus, given the max fitness rule, positive changes in individual MHC's fitness contributions are more likely to affect the overall fitness of the host than negative changes. Under these conditions, longer MHC clusters, containing more MHC sequence variants (with more potential to undergo changes in their fitness contribution) tend to be favoured.

As shown in figure 3.8, we observe longer clusters in the presence of CPS than in its absence. The inclusion of CPS offers more opportunities for changes in the fitness contributions of MHC sequence variants, thus increasing the strength of selection in favour of the longest possible lengths of cluster. The more extreme the CPS, the lower the variability in $L$ (copy number variation). In these scenarios larger gene clusters are being selected for more heavily so the populations are entirely longer gene clusters, approaching the maximum length possible within the system. Looking at both the mean and maximum fitness rules we see that CPS has drastically different outcomes for $\bar{L}$. However with both these rules a stronger CPS means certain MHC cluster sizes are selected for more strongly, therefore stronger CPS gives rise to lower variations of $L$ (i.e. lower levels of copy number variation) within populations. In appendix B section B.3 I show how systems that tend to large numbers of $L$ tend to have lower variance in $L$ due to boundary conditions.

### 3.3.4 *Real world data is more similar to the results of the mean fitness rule than the max fitness rule*

As a comparison to how CNV for the MHC behaves in closely related vertebrate species, I have plotted $\bar{L}$ and $Var(L)$ for humans at the HLA-A and HLA-B and for mauritian cynomolgus macaques at the Mafa-A and Mafa-B loci. The data was acquired from Wiseman et al. [2013] figure 2.

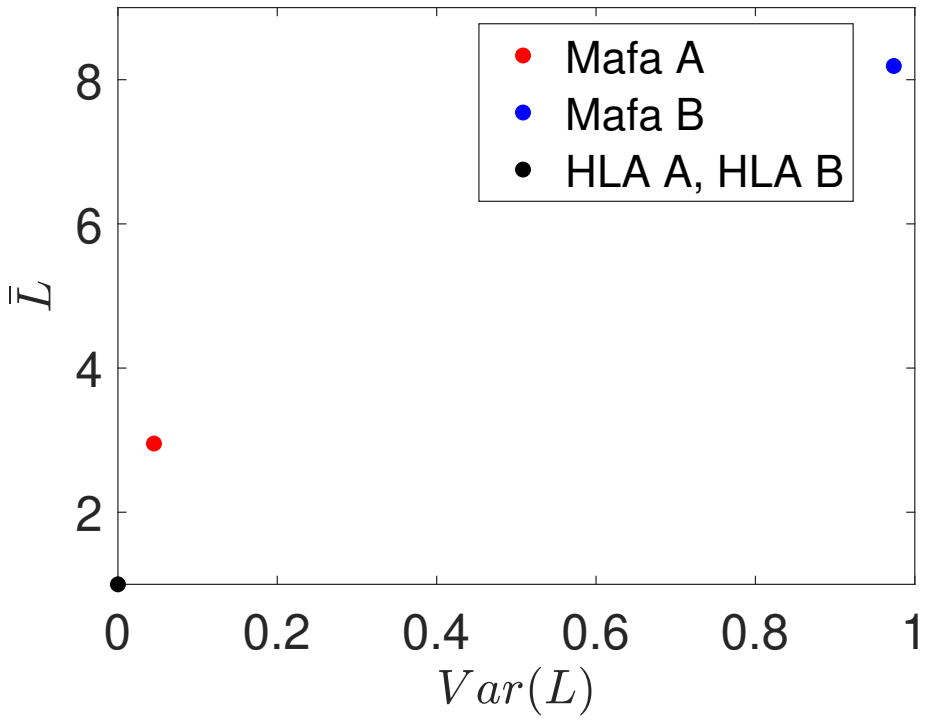



Figure 3.9: **Mean and variance of the lengths of MHC clusters from data that was acquired from Wiseman et al. [2013] figure 2.**

As we can see in figure 3.9 we have human HLA A and B which is located at the bottom left of the figure with an $\bar{L}$ if 1 and zero $Var(L)$. For mauritian cynomolgus macaques mafa-A and mafa-B have higher values for both $\bar{L}$ and $Var(L)$. This is not too far a difference from the results we have obtained in figure 3.7 where we have shown the level of CNV depends on the intensity of CPS. We can also see when comparing it to figure 3.8 that figure 3.8 does not show the same patter of results we have here in figure 3.9.

### 3.3.5 *Allowing recombination hotspots to exist between exons, as opposed to only between genes, reduces the lengths of MHC clusters*

As noted in the introduction, recombination events can occur between the exons of MHC sequences as well as between genes and this will generate new MHC sequences at the same time as changing the lengths of clusters. To explore the potential impact of this phenomenon, I look at the outcomes where I test an array of different ratios of the two types of recombination. The proportion $\beta$ is a parameter I use to decide how many recombination breakpoints occur between MHC sequences or between the exons of an MHC sequence. $\beta = 0$ means all recombination breakpoints occur between MHC sequences and $\beta = 1$ means all recombination breakpoints occur between the exons of MHC sequences.

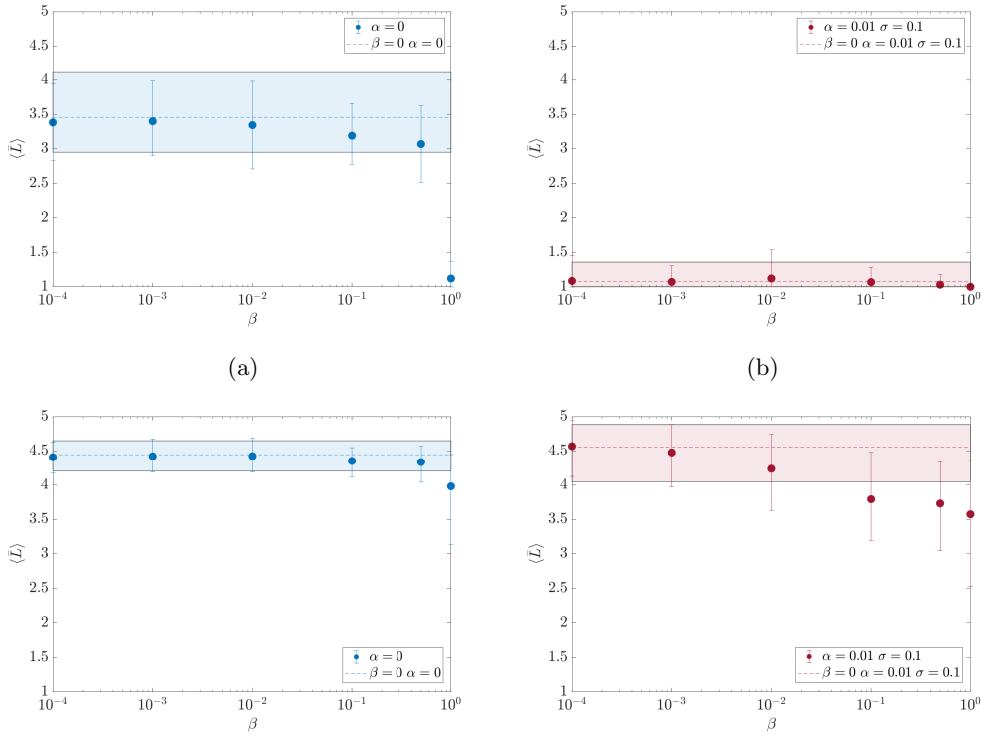



Figure 3.10: **The impact of between-exon recombination and CPS on the lengths of MHC clusters**. Each point on the graph is the average of $\bar{L}$ over 100 simulations and the error bars represent the 0.95 percentile of the distribution of $\bar{L}$. The shaded in area of the figures represents the 0.95 percentile of the distribution of $\bar{L}$ for the results where $\beta = 0$ (the dashed line result). (a) and (b) the fitness rule used is the mean fitness rule, (c) and (d) the fitness rule used is the maximum fitness rule. The parameters used are as follows: N = 5000, Initial number of exons = 10, Initial number of Loci = 2, $U_L = 5$, $p_R = 10^{-4}$ and $p_M = 10^{-6}$. For (a) and (c) $\alpha = 0$ and for (b) and (d) $\alpha = 0$ and $\sigma = 0.1$. Figure (a) $\beta = 1$ has not stabilised see table 3.2.

As can be seen in figure 3.10 we see that for both fitness rules, having a higher value for $\beta$ encourages smaller MHC clusters. In Appendix B I show a time series where no selection is present only boundary conditions on $L$ are applied and $\beta = 1$. Figure B.4 in the appendix B shows us that the system tends to an absorbing state of $L = 1$ for

every individual in the population when $\beta = 1$. The only result above that has the entire population at $L = 1$ is in figure 3.10b when $\beta = 1$. For every other scenario when $\beta = 1$ we see that $\langle \bar{L} \rangle > 1$ showing that the selection rules applied still cause some MHC CNV. It should also be noted that for the maximum fitness rule, CPS reduces $L$ when $\beta$ is higher than $10^{-2}$. This contrasts with the results presented in figure 3.8, where when $\beta = 0$, CPS increases $L$ for the maximum fitness rule.

## 3.4 Discussion

Class I MHC CNV in mammals can be characterised as highly variable. Some organisms like humans have no Class I MHC CNV while others such as the rhesus macaque have been found to have haplotypes of mamu-A that contain 2-3 copies of the mamu-A gene. Some of these differences could be due to radically different selective pressures, however among primate species it seems reasonable to speculate that the selective pressures acting on macaque class I MHCs might not have been all that radically different to those acting on humans. It would therefore be interesting to identify which conditions allow a relatively small change to give both evolutionary outcomes (no CNV, as is observed for human MHC class I and CNV, as is observed for macaque MHC class I).

My results show that when the mean fitness rule is applied, applying different levels of CPS can make the difference between no CNV and some CNV (figure 3.7 red and orange markers). I therefore propose that (i) the mean fitness rule is a better fit for the way primate MHC clusters evolve than the max fitness rule (see also figure 3.9), and (ii) differences between CNV for the MHC between humans and macaques could be due to a difference in their respective patterns of changing pathogen climates.

The mean fitness rule implies that all MHC gene sequences in a cluster need to all be functional in order to protect the host. Under these conditions, the more changes that happened to the host's pathogen climate, the more shorter gene clusters were selected resulting in lower CNV (figure 3.7 red and orange markers). Thus, my results imply humans may have less CNV than macaques because humans have had a more changing pathogen climate whereas macaques may have had a relatively more static pathogen climate.

How realistic was this CPS I implemented into the model? CPS was meant to imitate

the effect of new pathogens entering or leaving the population, be this from mutations in an existing pathogen, or a completely new pathogen entering the system. I sought to model this dynamic by changing the raw fitness contribution ($A_{ij}$) of the MHC gene sequences each generation, as the protection these genes gave in the previous pathogen climate will change when a new pathogen enters it. Examples of MHC alleles giving individuals increased susceptibility or protection to a pathogen has been a highly studied topic (Hill et al. [1991]) so saying that the advantageousness of certain MHC alleles would change due to a pathogen outbreak or disappearance is a reasonable assumption. How I applied change to $A_{ij}$ was using a truncated normal distribution where the mean of the distribution would be the previous $A_{ij}$, I also varied $\sigma$ to vary how large this change should be. The reason the distribution had the mean of the previous $A_{ij}$ was so its previous fitness contribution affected its fitness contribution in the next generation. I believed if an allele was advantageous in the previous pathogen climate it would have a higher chance of still being advantageous when a new pathogen enters. I believe this to be a reasonable assumption as MHC alleles would still convey protection to pathogens that were and are still in the population. I only change a fraction of the MHC alleles raw fitness contributions each generation ($\alpha$). It could be argued to be more realistic to change all $A_{ij}$; as all alleles' ability to defend a host might be changed if a new pathogen entered the population.

As I have shown in section B.3 in appendix B,variance in cluster length (CNV) is reduced when the average number of gene sequences in the clusters in the population is near the upper or lower boundary allowed within the system (figure B.2). A lower boundary clearly exists in real populations, and it seems reasonable that the closer the mean cluster length is to 1, the lower the variation in the number of copies. There are reasons that there should exist upper limits to how many copies of a MHC gene an organism has. For example if one had too many different MHC genes showcasing many different peptides, the important peptides from a deadly pathogen may go unnoticed due to being less represented on the surface of a cell which is a finite space. However the number is not necessarily exact and hard bounded. I have shown that even with a softer boundary condition the effect still occurs (see figure B.3 in supplementary materials) but the upper boundary in reality maybe very different to the modelled version and may vary

between individuals which would allow for potentially more variation in gene copies. It is difficult to say whether or not a evolutionary pressure which heavily encourages larger MHC gene clusters would actually exhibit lower CNV in populations due to the gene cluster size being pushed to the upper limit. However for the lower limit this effect would definitely still occur.

I have shown that between-exon recombination events in general reduce cluster length (figure 3.10) and in the appendix have shown and explained why this mechanistically has occurred (section B.1). However even with the steady state found in figure B.4 in the supplementary materials, MHC CNV was found in the vast majority of the results including when $\beta = 1$ ($\beta$ is the proportion of recombination events that happen in between the exons of a gene) especially for the maximum fitness rule. In the absence of between-exon recombination ($\beta = 0$), when the maximum fitness rule is applied, CPS increases $\bar{L}$, as explained in section 3.3.3. However I observed that different values of $\beta$ changed the effect CPS would have when using the maximum fitness rule. When $\beta$ was above $10^{-2}$ CPS reduced the value of $\bar{L}$. I am not entirely sure why $\beta > 10^{-2}$ would cause CPS to lower the value of $\bar{L}$. The results I show in section B.1 (which indicate that between-exon recombination leads to shorter MHC clusters) could suggest that CPS for the maximum fitness rule is causing individuals who have had in between exon recombination to be selected for over individuals who have had between allele-recombination.

Krüger and Vogel [1975] creates a mathematical model to analyse the stable distribution of numbers of copies of genes in the presence of unequal crossing over. They have three scenarios they look at. They have a scenario where there is no selection, a scenario where there is an optimum number of repeated genes and finally scenario where larger number of genes is advantages. The only comparable scenario to the work here is the no selection scenario and comparing it to the results in section 3.3.1. In the results we found that the stable distribution did not depend on the initial number of loci but rather the upper limit that is given to the number of MHC sequences that a gene cluster may have. Krüger and Vogel [1975] finds that the initial distribution of gene copies influences the final stable distribution. However there are multiple differences between the model outlined here and Krüger and Vogel [1975]. For instance in Krüger and Vogel [1975] no selection scenario there is no upper limit to the number of copies. They also have a lower

probability of unequal crossing happening compared with regular crossover events, the work here the probability of all break points are equal. They also cannot have unequal crossing over until there are two copies which is not the case of the work here.

Bentkowski and Radwan [2019] uses a individual based model that is similar to the model used in Borghans et al. [2004] with the added feature of allowing the duplication and deletion of MHC genes. One of the main results found in this work is that the number of MHC genes increased with a more diverse pathogen climate, but only when the cost of having many numbers of gene copies was low.

My model includes inherent differences in fitness contributions between MHC alleles, but these can change depending on the assumed pathogen climate. In the Bentkowski and Radwan [2019] model, MHC contribution to individual fitness depends only on what pathogen types (explicitly modelled) are present. Differences in MHC fitness contributions in the Bentkowski and Radwan [2019] model are greatest at low pathogen richness (when it is likely that some MHCs are able to match with pathogens and others can't). This means that the results in Bentkowski and Radwan [2019] when pathogen richness is "low" in their system are most comparable to my results. My maximum fitness rule is also most comparable to the Bentkowski and Radwan [2019] model, since within the Bentkowski and Radwan [2019] model, having just one "good" MHC will be enough to defend a host against a pathogen. In the context of low pathogen richness, Bentkowski and Radwan [2019] found that higher pathogen mutation rates would cause larger numbers of copies of MHC genes. I believe this result is comparable to my results in section 3.3.3. More intense CPS (in my model) is comparable to higher pathogen mutation rates (in Bentkowski and Radwan [2019] model), and both lead to longer gene clusters if we assume that one "good" MHC is enough to protect the host. However, I did not consider the results generated by my maximum fitness rule to be the most realistic, at least for primate class I MHCs, due to the results figure 3.9 matching figure 3.7 than figure 3.8. The effects of MHCs on fitness may not be as additive as implied by the Bentkowski and Radwan [2019] model. Whilst my fitness rules are highly abstracted, they offer a useful contrasting perspective to the results found by Bentkowski and Radwan [2019].

Bentkowski and Radwan [2019] do not present many results on within-species CNV per se, but they report they observed it the least at the lowest pathogen richnesses (which,

in terms of my model, would mean observing less CNV when fitness differences between MHCs are at their most stark). My results concur with this to some extent in that the most extreme forms of CPS (which will lead to big differences in MHC fitness contributions) tend to drive the system towards either the longest or shortest MHC clusters, and hence (as shown in appendix B) CNV is inherently lower (due to reasons discussed in section B.3). It would be interesting if the phenomenon of pushing up against a boundary also accounts for the instances of low CNV in the Bentkowski and Radwan [2019] model.

Varying CNV for the Class I MHC has been observed across many organisms. My model indicates that pathogen selection could have a critical role in determining the level of MHC CNV observed, and preliminary suggests that a "mean" fitness rule might apply to primate class I MHCs. As larger numbers of species are surveyed, improved data such as that shown in figure 3.9 will allow us to define realistic fitness scenarios with greater accuracy.

# Chapter 4

# The Evolution of MHC Promiscuity

## 4.1 Introduction

As mentioned in section 1.1.1 MHC molecules can vary in binding peptide repertoire (Paul et al. [2013]). In humans the most promiscuous HLA alleles are ones that belong to the HLA-A2 supergroup (Kaufman [2020]). Pockets in the binding grooves of A2 molecules can accommodate two or three different amino acids (Madden et al. [1993], Chen et al. [2012]). In contrast HLA-B*57:01 is considered fastidious for humans; one of its pockets requires a rare amino acid tryptophan (Illing et al. [2012]). For chickens one of their most promiscuous MHC alleles is BF2*02:01 (Chappell et al. [2015]). BF2*04:01 encodes a highly fastidious MHC molecule which requires binding of rare amino acids in each of three pockets (Wallny et al. [2006], Zhang et al. [2012]). Chickens compared to humans have the more extreme alleles on the spectrum of fastidious to promiscuous MHC alleles. Chickens only have one highly expressed MHC class I gene (as opposed to 3 in humans), and it is believed that this has caused the evolution of these alleles to be so extreme (Kaufman [2020]).

Both a fastidious and a promiscuous MHC molecule could be advantageous. The advantages of binding promiscuity, when it comes to responding to pathogens, seem fairly

obvious: having a wider variety of peptide shapes it can bind means that a promiscuous MHC could help a host to make immune responses against a greater diversity of pathogens/pathogen strains. The advantageousness of a fastidious allele, by contrast, can be explained if we suppose that focusing T cell responses against very particular pathogen peptides (not just any peptide) is the best way to respond to certain pathogens. It has been shown that fastidious HLA-B*57:01 and HLA-B*27:05 alleles are associated with better control of HIV and long HIV progression times to AIDS whereas HLA-B*35 is associated with rapid progression (Košmrlj et al. [2010], Gao et al. [2010]).

Many previous models of MHC evolution (see chapter 1) consider MHC alleles as having flat fitness values and do not consider the complexities of variable MHC attributes. Siljestam and Rueffler [2019] gives a nice example of representing MHC functionality without giving flat fitness values to alleles and individuals. They represent MHC alleles having 10 functioning parts and pathogens also have 10 corresponding parts. These parts are represented by numbers and the closer the allele parts are to the pathogen parts the more defended an individual is against that pathogen. They find that allelic polymorphism is maintained with the heterozygote advantage and alleles adapting to be more specific to the several pathogens. Here I also implement an allele trait system but model pathogen selection differently (see Discussion for a full comparison of approaches). Siljestam and Rueffler [2019] is concerned with arguing heterozygote advantage being a viable mechanism for maintaining polymorphic loci. In contrast here I ask how pathogenic climates can shape the fastidiousness or promiscuity of a population's MHC alleles and whether or not both generalist and specialist alleles can coexist.

## 4.2 Methods

Here I will describe the individual based model I designed to gain insight into MHC's alleles binding cleft promiscuity. I use an individual based model due to the wide range of allele properties I wanted to make possible within the system, and hence the wide range of possible host types.

### 4.2.1 Population

The population is represented by an integer column vector ($G$) which represents the population's MHC class I alleles at a single locus. Each row of the vector is a chromosome and every two rows represents a diploid individual in the population. I have a variable population size (N) which I limit to a size of $K$ as I wanted to see which scenarios ended in extinction. The number of elements in $G$ is $2K$, integers above zero represent the different MHC alleles and if the population is not at its limiting size of $K$, zeros represent the absence of individuals. For this work we are only considering the population from the perspective of a single MHC locus. This makes the work especially relevant to the chicken, which has only 1 highly expressed MHC class I gene (Kaufman [2018]).

Each generation I infect a proportion of the population which I notate as $p_I$. These infected individuals may die due to rules described later. The population size then increases for the next generation by an amount $r$. If I notate the proportion of individuals who die due to infection as $p_d$, the next generation population size would be according to the following equation:

$$N_{t+1} = N_t(1 - p_d)r \tag{4.1}$$

where $t$ represents generation time steps. If $N_{t+1} > K$ from equation 4.1 I then say $N_{t+1} = K$.

### 4.2.2 Allele Attributes

In this model each allele possesses a set of attributes. The number of attributes is notated as $L$. Each attribute represents a potential property of the protein encoded by that allele. In terms of the MHC each attribute could represent the capacity to bind a peptide shape, thus the value an MHC allele has for an attribute could represent how fastidiously the binding cleft of an MHC molecule binds to a peptide motif. These attributes can have a value between 0 and 1 and are notated as $a_i$ where $i$ represents the attribute in question. I constrain the properties of each allele such that the sum of all their attributes must be equal to 1. This means that, to continue the MHC analogy, if $L = 3$, then an allele with attributes 1,0,0 is a very fastidious allele that binds just one type of peptide motif. An allele with attributes 0.2,0.2,0.6 is less fastidious but has a preference for binding the third type of peptide motif. Alleles that have a value of close to 1 in an attribute could

be considered specialist in terms of this model. Alleles that do not have a high value in any attribute could be considered generalist.

This terminology gets more difficult the more attributes you have. For instance with only three attributes you have two types of generalist; the obvious type is an allele having attribute values close to a $\frac{1}{3}$, the second is an allele having two attributes close to a $\frac{1}{2}$. To distinguish between alleles which are different types of generalists I will notate them as $G_j$ where the $j$ represents the number of attributes that have some value for $a_i$ that are above 0 and are close to value $\frac{1}{j}$. Generally when the attribute values are close to the $\frac{1}{j}$ and $L - j$ attributes have values close to 0 we would call this a $G_j$ allele. Stricter definitions can be found in section 4.3.4 for a section of the results when $L = 3$. MHC alleles that are only good at one attribute, a specialist, I represent as $S$ alleles. Figure 4.1 are examples of $a_i$ values for an $S$ MHC allele (figure 4.1a) and a $G_3$ MHC allele (figure 4.1b).

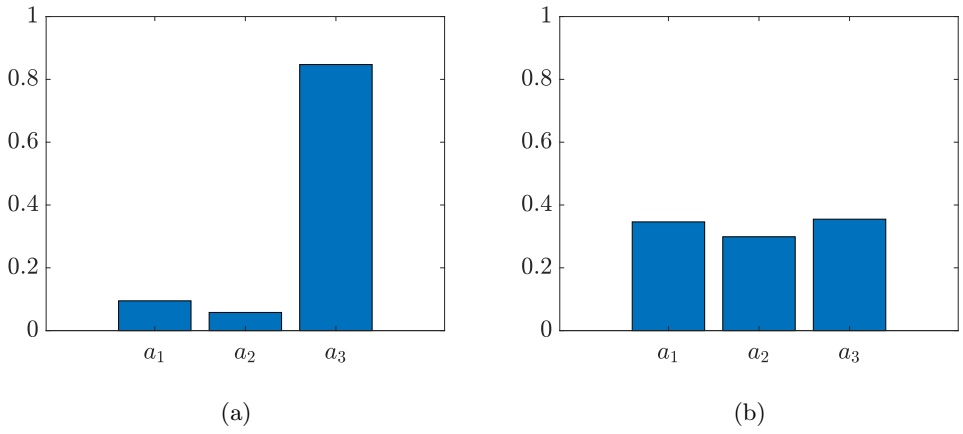



Figure 4.1: **Bar charts illustrating how an allele can be specialist or generalist in the framework described.** (a) is an example of an allele that would be considered a specialist, (b) is an example of an allele that would be considered a generalist. The parameters are as follows: $L = 3$.

In the simulations, I want different types of alleles to emerge through mutations. The rate of mutation of generalist and specialist alleles will affect the overall dynamics of how specialist and generalist alleles may co exist. I used two different approaches to

add new alleles through mutation: a deterministic method in which a fixed proportion of alleles generated through mutation were completely specialist and a fixed proportion of alleles were generalists of different properties, and a stochastic method in which mutations generated alleles along a spectrum from generalist to specialist.

#### 4.2.2.1 Deterministic

These methods are deterministic in the sense that the attribute values are decided before they are assigned, however which attribute is assigned what value is still stochastic.

When generating alleles I first have to select how promiscuous or fastidious I want the allele to be. For example if $L = 3$ there are three types of alleles: one which is good at all 3 attributes, one that is slightly better but at only 2 attributes and one that is even better but only good at 1 attribute. The number of attributes an allele will be effective at (g) depends on $L$ where $1 \leq g \leq L$. Once $g$ has been decided if $g < L$ I then randomly decide which attributes will be selected. The value of the attributes selected will then equal $\frac{1}{g}$ and the rest will equal 0.

When I want to have all allele types in the system, every time I generate an allele I randomise $g$ uniformly. This way all allele types from $S$ to $G_L$ will be generated with equal probability.

#### 4.2.2.2 Stochastic

These methods assign attributes values more randomly than the deterministic method described above. Generating alleles this way allows for more flexibility in allele properties

If I simply generated $L$ uniform random numbers between 0 and 1 ($U(0,1)$) for each attribute notated as $X_i$ and then normalised the values obtained such that they summed to 1 then I would generate generalist alleles far more often than specialist alleles (see figure 4.2).

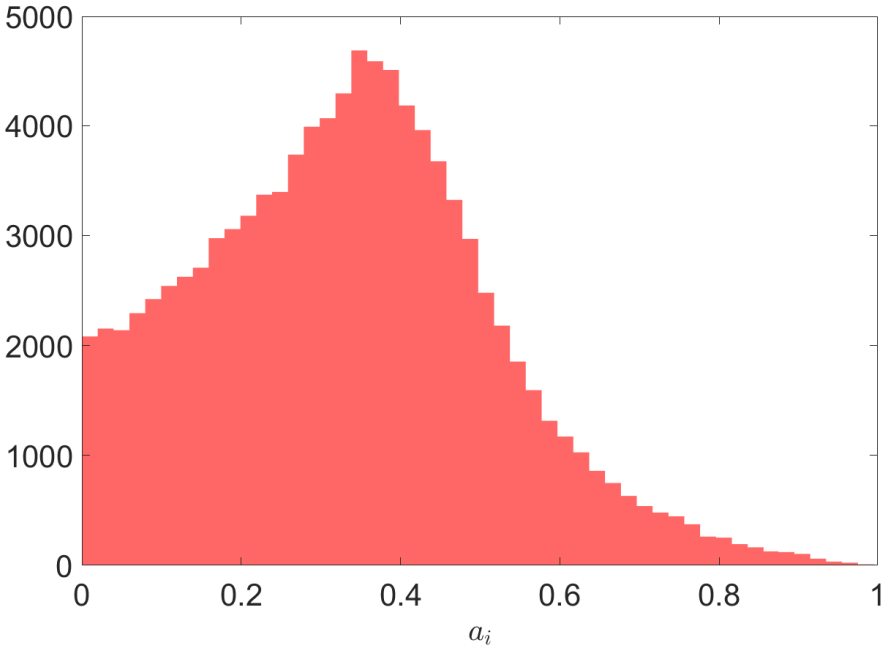



Figure 4.2: **A histogram of values $a_1$ takes when generating alleles using the normalised attributes method**. Parameters used are: $L = 3$.

As can be seen in figure 4.2 that the most likely alleles produced are just below $a_i = 0.4$ and very few examples of $a_i$ being close to 1.

I therefore developed the "one at a time" method to ensure that specialist alleles were still likely to emerge. I first select the order in which I want to assign value to the attributes of an allele. For instance if $L = 3$ I would randomly order the numbers 1 to 3 which could be 2,3 and 1. I then assign values to the attributes in this order. The first attribute would then be assigned a random number $a_2 = U(0, 1)$. The next attribute would then be assigned a value of $a_3 = U(0, 1 - a_2)$. Finally the last attribute would be assigned whatever is left of the total adding up to 1, in this example $a_1 = 1 - a_2 - a_3$. Generating alleles this way increases the chance of $a_i$ being given a value close to 1.

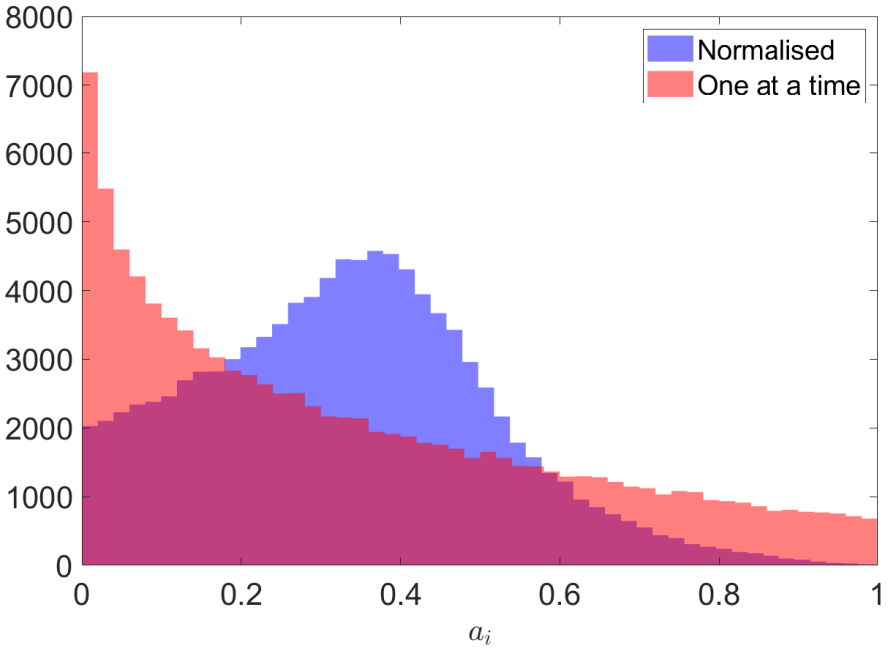



Figure 4.3: **A histogram of values $a_1$ takes when generating alleles using the normalised attributes method and the one at a time method**. Parameters used are: $L = 3$.

I use the One at a time method of generating alleles in section 4.3.4 of the results as it generates all allele types more evenly.

Alleles are introduced in two ways into the system. The first way is to initialise the population with a certain number of unique alleles. This number is 1000 for all results in section 4.3. The second way I introduce alleles into the system is through mutations. When a parent passes on its genes to the next generation there is a probability $p_M$ that a mutation will occur and create a new unique allele that will be inherited instead. When a method from above is decided for generating alleles attribute values, it is used for both initialising the alleles in the population and for creating new alleles due to mutations.

### 4.2.3 Pathogen Test

I generate a pathogen challenge each generation, in terms of a specific MHC allele attribute ($i$), and a minimum threshold value of $a_i$ (P) which a host must possess in order to survive the pathogen challenge. In biological terms, this equates to the existence of a pathogen where the ability of host MHCs to bind a particular peptide from that pathogen with a particular degree of fastidiousness is important to successfully combating the infection. The higher the value of $P$, the more fastidiousness is required (to give examples from humans, the more HIV-like the pathogen is in terms of the immune responses which control it best).

Each generation I applied this pathogen challenge as follows:

- I sum the values for each MHC attribute in the parental generation, and determine which of these MHC attributes had the lowest values. I select this attribute as the attribute to be tested. To put this in biologically relevant terms, this represents a peptide motif which the MHC molecules of the parental generation were the least likely to bind, and a pathogen bearing this motif is doing especially well in the population during the current generation. I notate the attribute that is being tested as $a_T$.
- I infect a proportion $p_I$ of the population, chosen at random.
- For each infection I generate a random number $P$ between 0 and 1. If an individual is infected and does not have an allele in their genotype for which $a_T \leq P$ that individual will die and have zero chance of passing on their genes to the next generation.
- Everyone who has survived the pathogen test or has not been infected will all be equally likely to be selected as a parent to pass on their genes to the next generation.

How $P$ is distributed determines the evolutionary outcome I get. I chose to generate $P$ using a Kumaraswamy distribution, because this distribution is capable of taking a range of contrasting relevant shapes. The Kumaraswamy distribution has the following probability density function:

$$f(P, a, b) = abP^{a-1}(1 - P^a)^{b-1} \tag{4.2}$$

where $P \in (0, 1)$. The parameters $a$ and $b$ are non negative. Examples of what the probability density function looks like for the Kumaraswamy distribution are shown in figure 4.4. If I use a distribution for $P$ like the red line ($a = 5$ and $b = 1$) it would mean that very fastidious binding is required to survive infection, most of the time however for something like the yellow line ($a = 1$ and $b = 3$) a fastidious binding is rarely required.

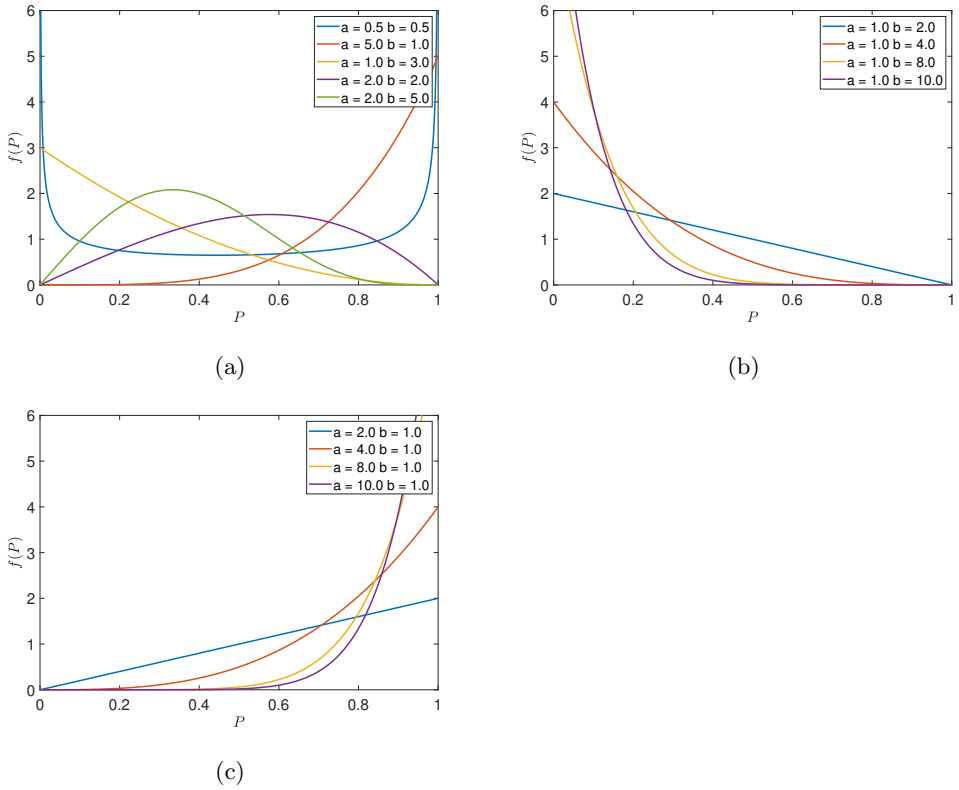



Figure 4.4: **An example of the probability density function for the Kumaraswamy distribution for varying values of $a$ and $b$.** (a) is an example of $f(P)$ when $a$ and $b$ vary. (b) is an example of $f(P)$ when $a = 1$ and $b$ varies. (c) is an example of $f(P)$ when $b = 1$ and $a$ varies.

### 4.2.4 No Co-evolution

In section 4.3.5 I investigate how a pathogen climate that does not depend on properties of the host population affects evolutionary outcomes. I do this by randomly selecting each generation which attribute will be selected for testing. This random selection is instead of selecting the attribute that was least represented in the previous generation.

### 4.2.5 Defining different evolutionary outcomes

In my results I represent an evolutionary outcome with a specific colour. First I define allele types into three groups: the first being the most fastidious alleles of type $S$, the next are alleles which are the most promiscuous of type $G_L$, are and finally alleles of which are not the most fastidious or promiscuous, alleles of type $G_2$ up to $G_{L-1}$. For each value of $a$ and $b$ I run 50 simulations. From the combined results of all 50 simulations I calculate the proportion of alleles present in the final generation that are each allele type (as just defined). According to these proportions I define each evolutionary outcomes as follows:

- If the proportion of an allele group is above 0.1 and the other two are below 0.1, I define that scenario as a population that is dominated by one allele type. For this I have three colours, red for $S$ , green for $G_2$ to $G_{L-1}$ and blue for $G_L$.
- If the proportion of 2 allele group types is above 0.1 and the other is below 0.1 I define this population as being dominated by 2 allele groups. This is again three colours yellow for populations with $S$ and $G_2$ to $G_{L-1}$. Purple for populations with $S$ and $G_L$ allele types. Turquoise for populations with $G_2$ to $G_{L-1}$ and $G_L$ allele types.
- If all allele group types are above 0.1 I represent this with the colour white.

Below is a venn diagram (figure 4.5) which illustrates the different colours used to define the evolutionary outcomes I have detailed above. These colours will be used consistently throughout the results section, to indicate the relevant combination of allele types.

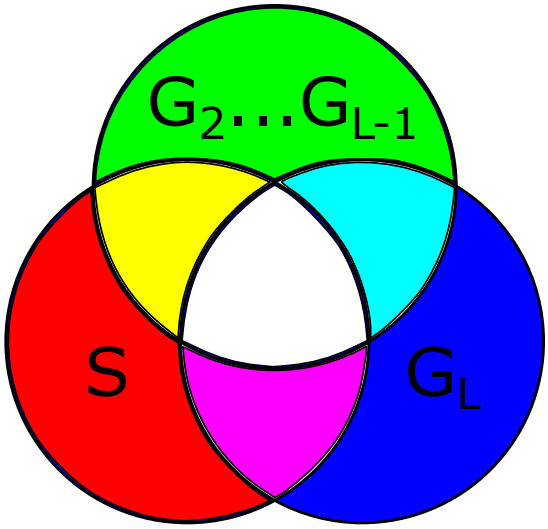


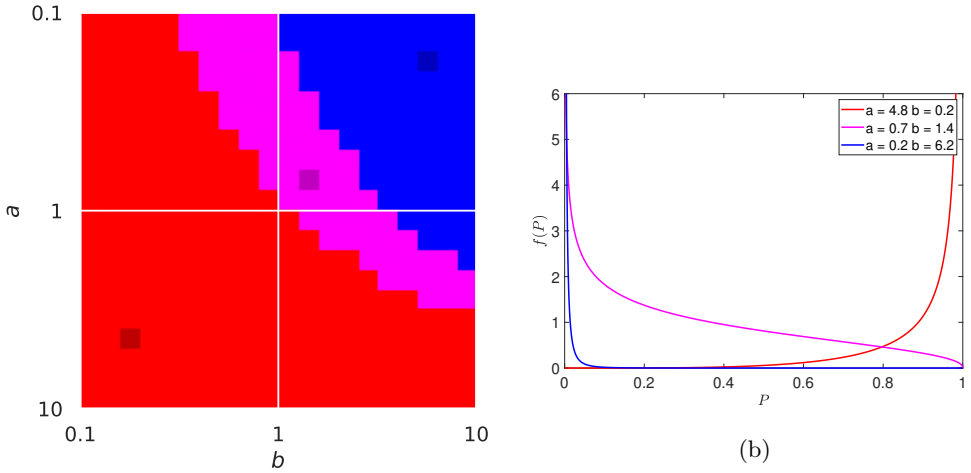

Figure 4.5: **A venn diagram illustrating the different evolutionary outcomes I have defined**.

## 4.3 Results

### 4.3.1 *Sporadic outbreaks of pathogens requiring fastidious immune responses, against a background of pathogens requiring less fastidious responses, favour the coexistence of generalist and specialist MHCs*

We first look at the case for when $L = 2$ as this is the simplest number of attributes I can have that will give us specialists ($S$) and generalists ($G_2$). To explore a wide range of $P$ distributions I use the Kumaraswamy distribution. This distribution has a bounded PDF and is defined around two parameters $a$ and $b$ (see section 4.2.3). I vary $a$ and $b$

from 0.1 to 10 on a log scale as this parameter range produces all evolutionary outcomes of interest when performing single simulations. I first consider a case where the host population never becomes extinct. In this example $p_I = 0.0909$ and $r = 1.1$. As I do not know the $p_I$ value that causes extinction for all the possible combinations of $a$ and $b$ I use $p_I = 0.0909$ because even if all the individuals in the pathogen test die $r = 1.1$ is large enough to reproduce individuals back up to the limit population size $K$ (see equation C.1 in supplementary materials).




(a) (b)

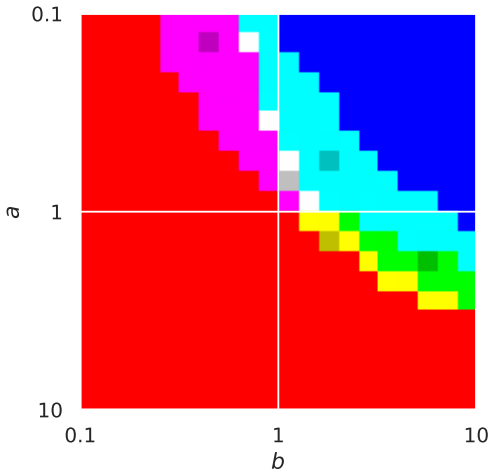
Figure 4.6: **Which allele types co-exist for varying $a - b$ parameter space, for two binding attributes.** (b) shows the distributions of $P$ that correspond to the shaded square on the heat map (a), the colour of each line corresponds to the colour of the shaded square. Different colours represent the allele type/types that are predominant. Red represents alleles of type $S$, blue of type $G_2$ and purple a mix of type $S$ and $G_2$. The parameters are as follows: $L = 2$ and $P_M = 10^{-5}$, $r = 1.1$, $P_I = 0.0909$, N = 5000 and generation simulated to 20,000.

For the case of $L = 2$ I see three evolutionary outcomes for different parameter space of $a$ and $b$ (figure 4.6). The entirely red region is where populations have predominantly $S$ MHC alleles, the entirely blue region the population has predominantly $G_2$ alleles which for the case of $L = 2$ are the most general alleles. I finally have a region where the populations tend to have a mixture of $S$ and $G_2$ MHC alleles co existing (figure 4.6

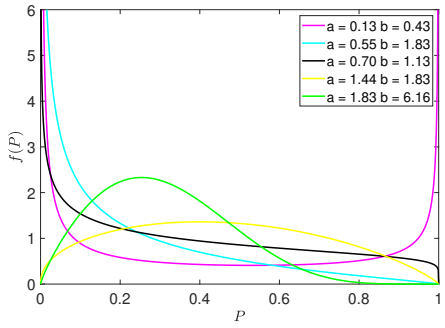
purple).

If the pathogen test is more likely to require fastidious binding (a higher $P$) then generally specialists are selected for as they are the only MHC alleles that can withstand the test (see figure 4.6b red line). If the pathogen test does not require fastidious binding (a lower $P$) then generalists are selected for given that a generalist allele can amply protect against both pathogens (see figure 4.6b blue line). In the scenario where both types of alleles co-exist the test sometimes requires high fastidiousness but most of the time the generalist will suffice, we see in figure 4.6b that the purple line is in between these two extreme distributions of $P$.

For $L = 3$ I again want to explore the parameter space of $a$ and $b$. I again set $p_I = 0.0909$ and $r = 1.1$ as I do not know the $p_I$ value that causes extinction for all the possible combinations of $a$ and $b$.




(a)




(b)

Figure 4.7: **Which allele types co-exist for varying $a-b$ parameter space, with 3 possible binding attributes.** (b) shows the distributions of $P$ that correspond to the shaded square on the heat map (a). The colour of each line corresponds to the colour of the shaded square with the exception of the black line which represents the white shaded square in (a). Different colours represent the allele type/types that are predominant. Red represents alleles of types $S$, green of type $G_2$ and blue of type $G_3$. The other colours represent populations with multiple allele types present (yellow $S$ and $G_2$, purple $S$ and $G_3$, turquoise $G_2$ and $G_3$ and white all allele types). The parameters are as follows: $L = 3$ and $P_M = 10^{-5}$, $r = 1.1$, $P_I = 0.0909$, N = 5000 and generation simulated to 20,000.

In figure 4.7 we see regions of $a$ and $b$ parameter space where each allele type for $L = 3$ dominates. We also see like we did in figure 4.6 that we get regions where $S$ and $G_2$ alleles co-exist and we get regions where $G_2$ and $G_3$ alleles co-exist. However we now see there are even regions where $S$ and $G_3$ alleles co-exist (figure 4.7a). The reason this might be surprising is because you might expect $G_2$ to thrive instead of the co-existence of $G_3$ and $S$. However this latter state is clearly being selected for over only $G_2$ alleles in the purple region of figure 4.7a.

As can be seen in figure 4.7b, purple line, the distribution has a higher chance of being very small or very large. The rise in $f(P)$ for smaller values of $P$ is larger than the rise in $f(P)$ for larger values of $P$. For $L = 3$ in order for $S$ and $G_3$ to be selected for this

needs to be the case for if the ends of the distribution were symmetrical the system is dominated by $S$ MHC alleles for the case of $L = 3$. Meaning the $P$ test needs to be more likely to have smaller values than larger ones for $S$ and $G_3$ MHC alleles to be selected.

It should be noted that in figures 4.6 and 4.7 I do not explicitly show if co-existence is occurring for each simulation just that there are multiple allele types over the 50 simulations. In section C.2 of the supplementary materials I show that these regions of multiple allele types existing over the 50 simulations correspond also to where co-existence of multiple allele types exist in single simulations.

If I increase the number of attributes past $L = 3$ we do not see any new interesting features to the heat map for the parameter space of $a$ and $b$ currently explored (see figure 4.8).

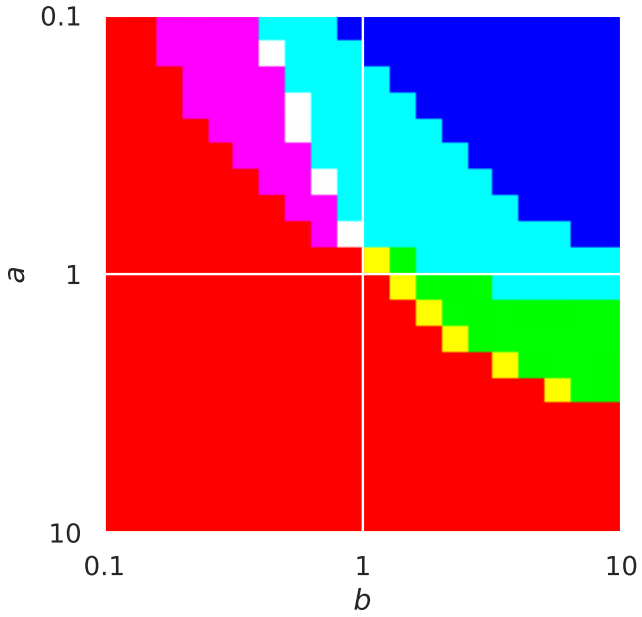



Figure 4.8: **Which allele types co-exist for varying $a - b$ parameter space, with 5 possible binding attributes.** Different colours represent the allele type/types that are predominant. Red represents alleles of type $S$, green of type $G_2$, $G_3$ and $G_4$. Blue represents alleles of type $G_5$. The other colours represent populations with multiple allele types present (yellow $S$ with $G_2$, purple $S$ with $G_5$, turquoise $G_2$, $G_3$ and $G_4$ with $G_5$ and white all allele types). The parameters are as follows: $L = 5$ and $P_M = 10^{-5}$, $r = 1.1$, $P_I = 0.0909$, N = 5000 and generation simulated to 20,000.

For $L = 5$, if I group alleles into the most specialist $S$, the most generalist $G_5$ and those that are in between $G_2$, $G_3$ and $G_4$ (figure 4.8) we get a very similar picture of evolutionary outcomes as we did for $L = 3$ (figure 4.7). If I group alleles this way it does not seem to introduce any new features to the heat map.

### 4.3.2 *If extinction can occur, the parameter space in which entirely specialist alleles emerge is severely limited*

I have shown how evolutionary outcomes are affected by the distribution of $P$ when extinction cannot occur. Now I will vary $p_I$ (proportion of population infected each generation) and allow the possibility of extinction, to get more perspective on how this affects outcomes. I again generate $P$ with a Kumaraswamy distribution. When $P$ is from a uniform distribution (when $a = b = 1$), $S$ MHC alleles are selected for when $L = 3$. From this I know then that if I make high values for $P$ more likely it will only result in populations evolving to have $S$ type alleles. I want to go from a uniform distribution of generated $P$ to distributions of $P$ where lower values are more likely to occur. In order to achieve this I keep $a = 1$ and I vary $b$ from 1 to 10. I also vary $p_I$ from 0.1 to 1. I also investigate evolutionary outcomes over this parameter space when only certain allele types are present. For instance I investigate how well a population survives with only $S$ alleles or only $G_2$ alleles or combinations of these and so on.

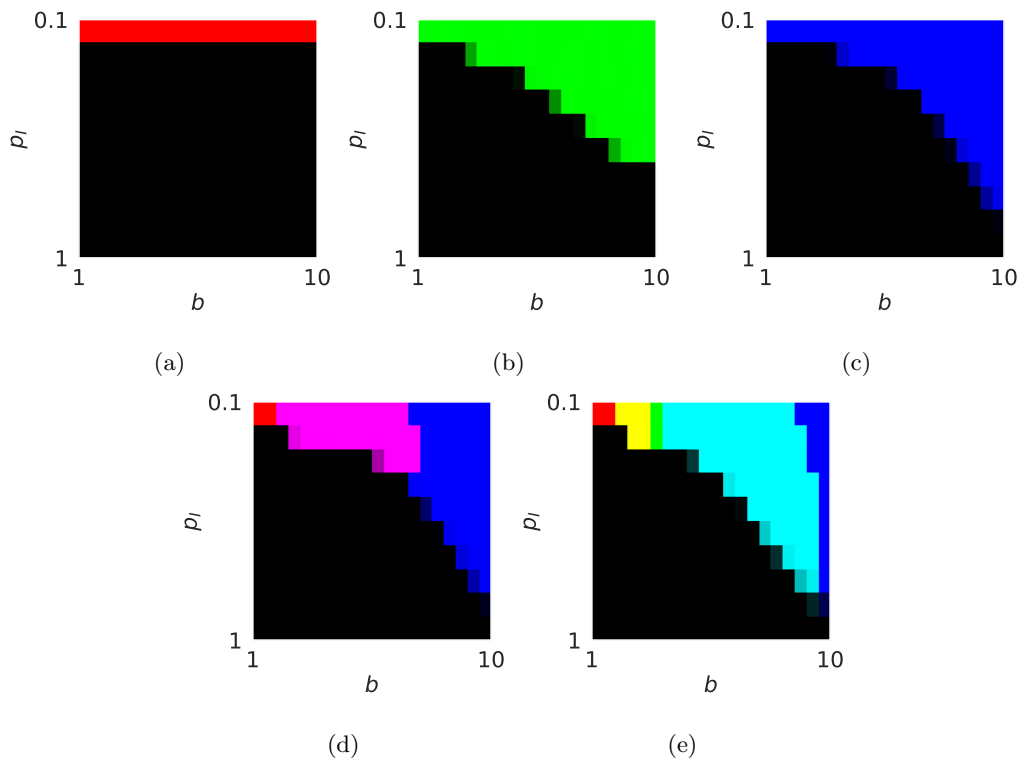



Figure 4.9: **Which allele types co-exist for varying $P_I - b$ parameter space.** Different colours represent the allele type/types that are predominant. Red represents alleles of types $S$, green of type $G_2$ and blue of type $G_3$. The other colours represent populations with multiple allele types present (yellow $S$ and $G_2$, purple $S$ and $G_3$, turquoise $G_2$ and $G_3$ and white all allele types). How dark the square is represents the size of the population (black means went extinct). For each heatmap only a limited amount of allele types were inputted in the system, (a) $S$, (b) $G_2$ and (c) $G_3$, (d) $S$ & $G_3$ and (e) $S$, $G_2$ & $G_3$. The parameters are as follows: $L = 3$, $P_M = 10^{-5}$, $r = 1.1$, a = 1, N = 5000 and generation simulated to 20,000.

A population with only specialist alleles can only survive in a very limited parameter space (figure 4.9a). We see that populations with only $S$ alleles can survive for all values of $b$ but only at the lowest values for $P_I$. Suggesting that even for $P$ distributions that mainly give lower numbers $S$ alleles are not a viable strategy if the amount of the population that gets infected is too high. Given how I have generated specialists as having no value in the

other attributes, whenever a host is infected with a pathogen that is testing an attribute they do not have a specialist for, they will die. If the population size of infected hosts whose MHC genotypes does contain the specialist allele that is being tested is consistently larger than the population increase due to reproduction then extinction will occur. This is the reason why we see population values of $S$ alleles going extinct for $P_I \geq 0.2$ for all values of $b$.

When all allele types are in the system (figure 4.9e) we see as $b$ increases from 1 to 10 that we move from the simulations tending to be dominated by $S$ type alleles to $G_3$ alleles. In between these two extremes we see regions where $G_2$ alleles thrive and regions where it is mixed which allele type dominates just as we have seen previously (figure 4.7). We can see in figure 4.9e that as $b$ increases there is a larger range of $p_I$ where populations do not go extinct. We also note that $p_I$ can vary which allele type is being selected for but only very slightly, it seems to mainly determine whether or not a population will go extinct.

### 4.3.3 *Coexistence of generalist and specialist alleles generally coincides with the greatest allelic diversity*

The allele types present in a population determine the allelic diversity within that population.

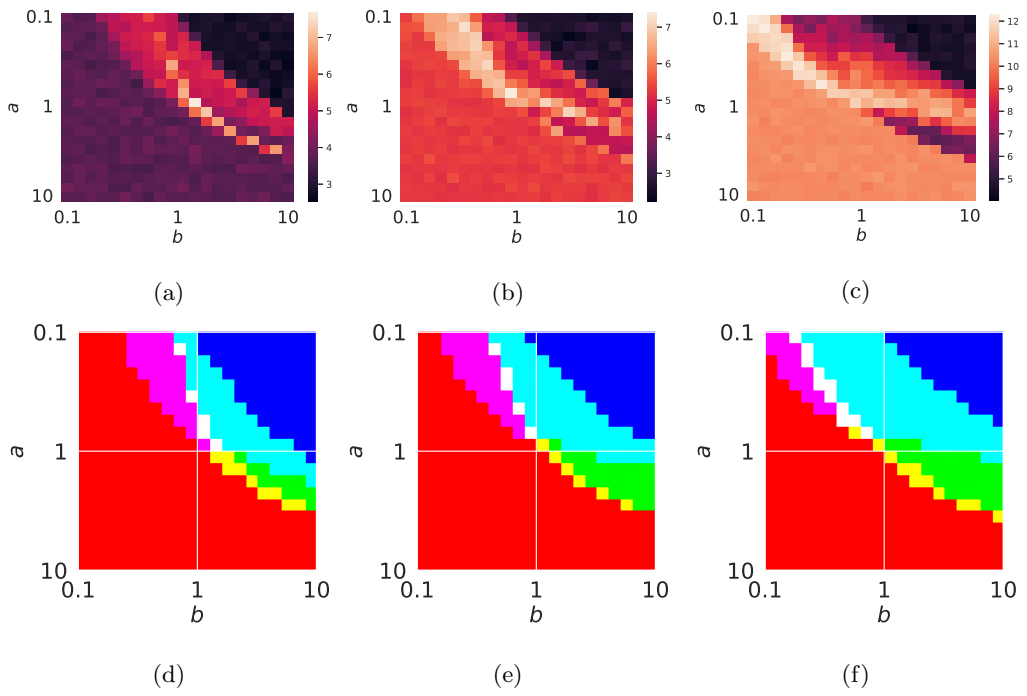



Figure 4.10: **Heat Map representing the average number of alleles over $a - b$ parameter space.** Figures (a), (b) and (c) are heat maps representing the number of unique alleles over $a$ and $b$ parameter space for $L = 3$, $L = 5$ and $L = 10$. Figures (d), (e) and (f) represent the proportion of varying allele types over $a$ and $b$ parameter space for $L = 3$, $L = 5$ and $L = 10$ respectively. Different colours represent the allele type/types that are predominant in the population. The colour red represents $S$ alleles, the colour blue represents $G_L$ alleles and green represents all allele types from $G_2$ to $G_{L-1}$. The other colours represent populations with multiple allele types present (yellow $S$ with $G_2$ to $G_{L-1}$, purple $S$ with $G_L$, turquoise $G_2$ to $G_L$ and white all allele types). For figure (a), (b), (c) and (d) $N = 5000$ and for figures (e) and (f) $N = 10,000$. The parameters are as follows: $P_M = 10^{-5}$, $P_I = 0.0909$ and generation simulated to 20,000.

The least polymorphic populations are the ones that have the most generalist MHC alleles (figure 4.10). This is likely due to the fact that if MHC alleles cover all attributes there are less available niches for other MHC alleles to co-exist. As the number of attributes increases from $L = 3$ (figure 4.10a and 4.10d) to $L = 10$ (figure 4.10c and 4.10f)

we see that populations including specialist alleles tend to have greater allelic diversity. A specialist population with a higher $L$ requires more alleles to cover all attribute types. However for all scenarios the highest allelic diversity regions seem to be where allele types (i.e. the broad types of "generalist" or "specialist") co-exist or at least close to these regions. We see this in figure 4.10c and that its highest values occur in the white regions of figure 4.10f.

### 4.3.4 *If allele properties are generated stochastically, the current classification for allele types fails to explain the range in promiscuity of MHC alleles*

In section 4.2.2.2 I outline a method of generating alleles stochastically. Generating alleles this way allows us to see more possible allele types that may form. In figure 4.11 I explore the same parameter space in figure 4.7 using the "One at a time" method of generating allele attributes.

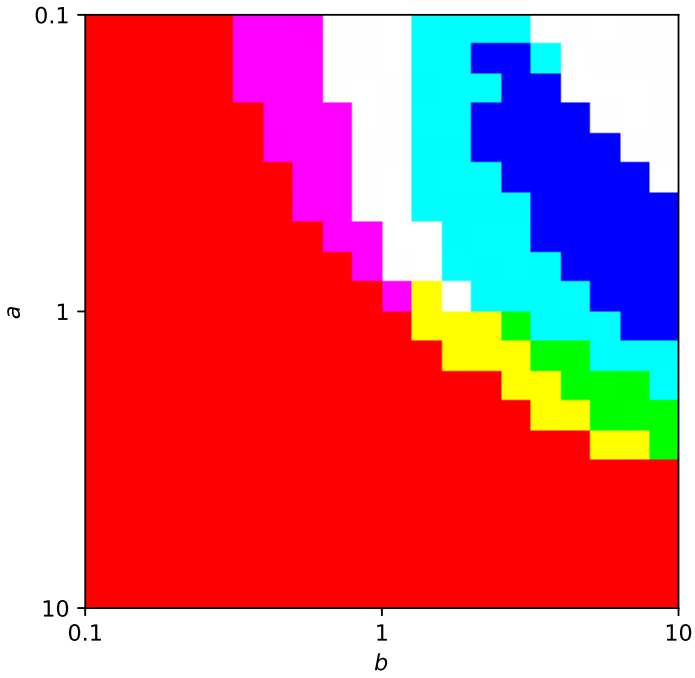



Figure 4.11: **Heat Map representing how much of the population had each allele type.** Different colours represent the allele type/types that are predominant. The colour red represents $S$ alleles, the colour blue represents $G_3$ alleles and green represents $G_2$ alleles. The other colours represent populations with multiple allele types present (yellow $S$ with $G_2$, purple $S$ with $G_3$, turquoise $G_2$ with $G_3$ and white all allele types) The parameters are as follows: $L = 3$, $P_M = 10^{-5}$, $P_I = 0.0909$, N = 5000 and generation simulated to 20,000.

We see similar regions of allele types when I compare figure 4.11 with 4.7. The major difference we see is in the top right region where all allele types seem to be somewhat equally selected for in figure 4.11. I believe this difference in the top right region happens due to $f(P)$ being large at lower values for $P$, which I believe makes it easier for all allele types to potentially exist. The reason we do not see this in the deterministic case of generating alleles is because for the case $L = 3$ there are only 3 allele types (in terms of promiscuity) and so the type of allele that dominates in the top right region is more

distinct than when the alleles are generated stochastically. However in figure 4.11 the regions depend on how I define each allele type. With a different definition of allele types we could make the top right region to be dominated by $G_3$ allele types, it all depends how we define this. Here is a list of conditions I used to define the three allele types in figure 4.11:

- Allele type is $S$ if any attributes are above 0.7. $a_i > 0.7$ for $i = 1, 2$ or 3.
- Allele type is $G_2$ if any attributes are between 0.4 and 0.7 and any attribute is below 0.1.
- Allele type is $G_3$ if all attributes are above 0.1. $a_i > 0.1$ for $i = 1, 2$ and 3.

These definitions I have made to include every type of allele produced stochastically. However my definitions of $S$, $G_2$ and $G_3$ allele types make less sense here as now any possible combination of trait fitnesses adding up to 1 ($a_i$ where $i$ is the trait) are now possible. To explore the full properties of the alleles, I plot the allele attributes on ternary plots for all 7 possible combinations of allele types (see Venn diagram figure 4.5)

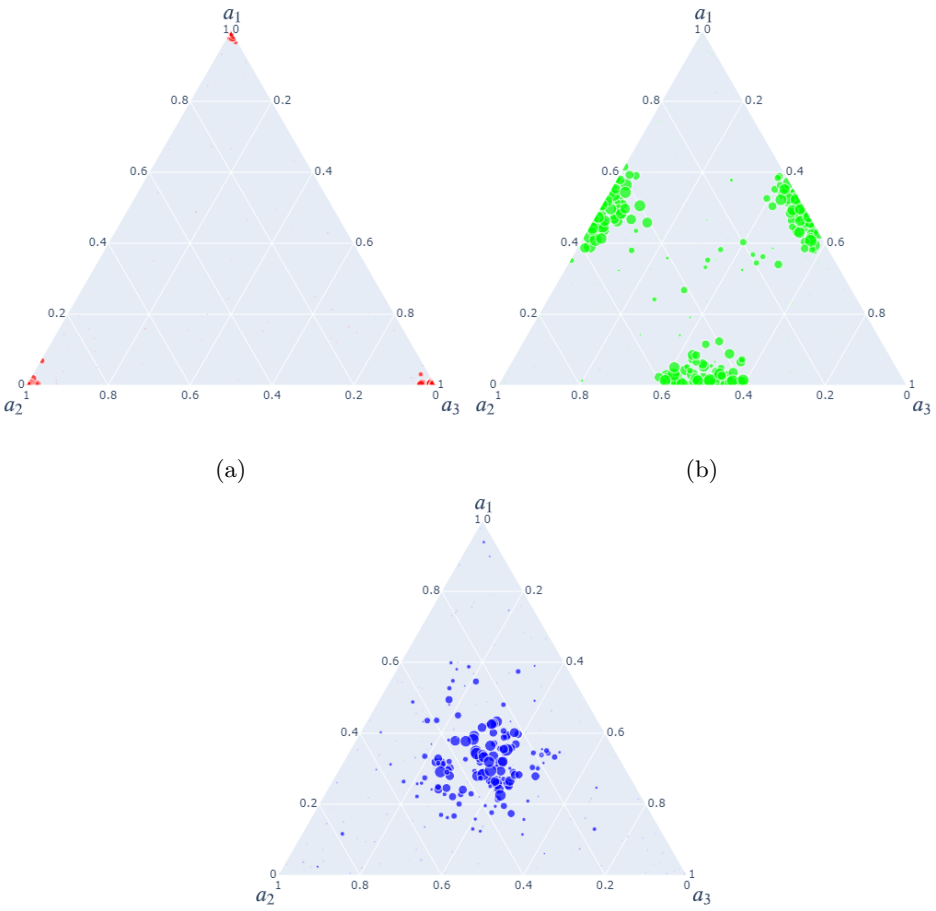



Figure 4.12: **Ternary Plots showing the attributes of stochastically generated alleles that are within the system for varying values of $a$ and $b$.** Different colours represent the allele type/types that are predominant in the population as defined above (see section 4.2.5 and figure 4.5). The size of each marker represents the frequency that allele was in the population. The parameters are as follows: $L = 3$, $P_M = 10^{-5}$, $P_I = 0.0909$, N = 5000 and generation simulated to 20,000. (a) $a = 3.79$ and $b = 0.34$, (b) $a = 1.83$ and $b = 7.85$, (c) $a = 0.7$ and $b = 7.85$. Each plot shows all alleles present at the end of each of 50 simulations.

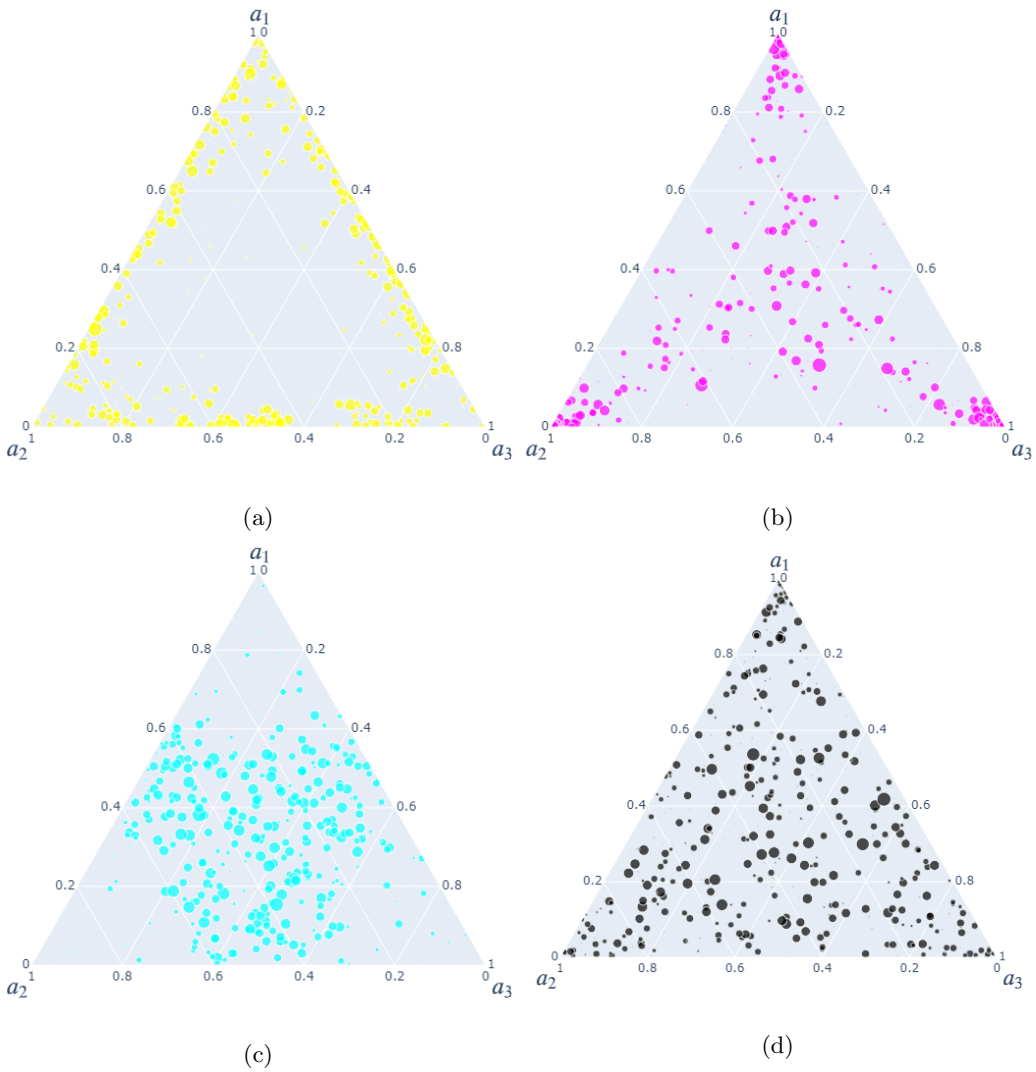



Figure 4.13: **Ternary Plots showing the attributes of stochastically generated alleles that are within the system for varying values of $a$ and $b$.** Different colours represent the allele type/types that are predominant in the population as defined above (see section 4.2.5 and figure 4.5), Except for figure (d) where we use black to make the figure clearer however this represents the white evolutionary outcome . The size of each marker represents the frequency that allele was in the population. The parameters are as follows: $L = 3$, $P_M = 10^{-5}$, $P_I = 0.0909$, N = 5000 and generation simulated to 20,000. (a) $a = 1.13$ and $b = 1.83$, (b) $a = 0.13$ and $b = 0.43$, (c) $a = 0.55$ and $b = 2.34$, (d) $a = 0.21$ and $b = 0.89$. Each plot shows all alleles present at the end of each of 50 simulations.

In figure 4.12 we see for figures 4.12a, 4.12b and 4.12c that we get alleles of attribute values that converge quite well on my definitions of allele types for $S$ $G_2$ and $G_3$. However when we look at the remaining cases where multiple allele types co-exist (figure 4.13) we see that the alleles present are also in areas of the plot that are blended between the allele types I have defined. This is particularly true for figure 4.13d, which corresponds to the value of parameters $a$ and $b$ that generates the maximum possible coexistence between specialists and different types of generalists. In this panel, we see alleles present that cover the whole attribute landscape and not just the areas covered by figures 4.12a, 4.12b and 4.12c.

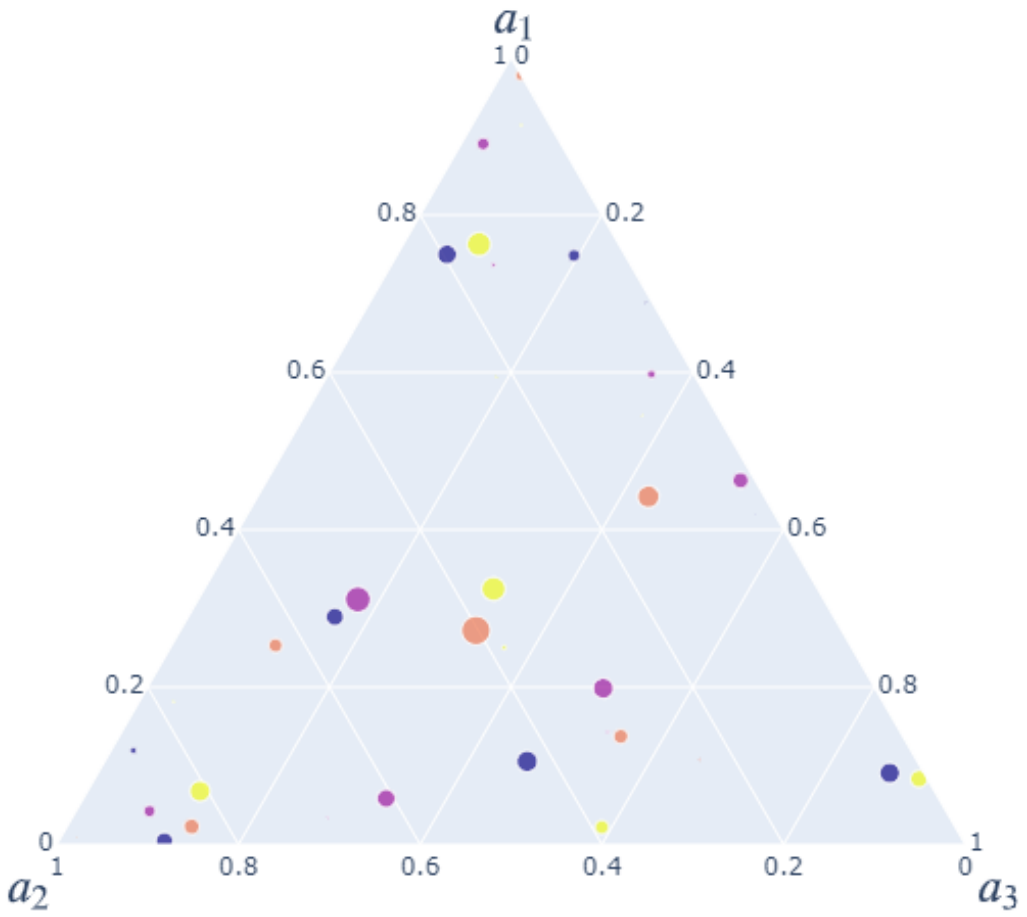



Figure 4.14: **Ternary Plot showing the properties of stochastically generated alleles in a scenario where the maximum possible range of different allele types coexist.** Different colours represent different simulations, each unique colour belongs to one simulation. The parameters are as follows: $L = 3$, $P_M = 10^{-5}$, $P_I = 0.0909$, N = 5000, generation simulated to 20,000, $a = 0.21$ and $b = 0.89$. Here we are looking at 4 simulations.

In figure 4.14 I have used the same values for $a$ and $b$ that produced the plot 4.13d, but this time focus on 4 individual simulations. For any given simulation, allele attributes are spread across the ternary plot. All attributes are covered, but how these attributes are covered can vary from simulation to simulation. For example the yellow dots show

us that this simulation ended with three alleles: a specialist ($S$) , a generalist ($G_3$) and a single $G_2$ allele type. The blue dots however do not seem to have as clear of a $G_3$ allele as the yellow dots. When looking at figure 4.13d it gives the impression that alleles attribute can belong anywhere, which is true but it needs corresponding alleles in the same population to cover other attributes as shown in figure 4.14.

### 4.3.5 *When the pathogen test is independent of the host's genetic landscape, there is less MHC allelic diversity as well as more unstable MHC allele lifetimes.*

So far, all the results that have been displayed used the co evolutionary mechanism, that the pathogen attribute that is tested is the least represented attribute in the previous generation of the host population. I wanted to see how the evolutionary outcomes might differ if I instead uniformly randomly select the pathogen attribute to be tested each generation.

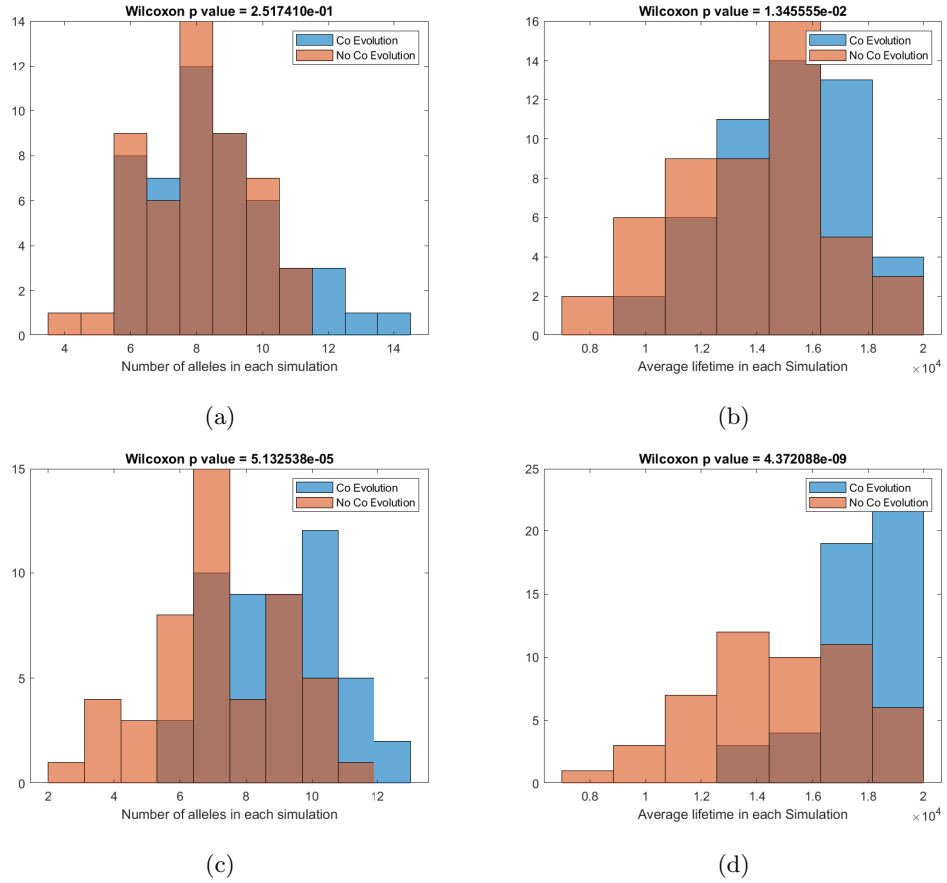
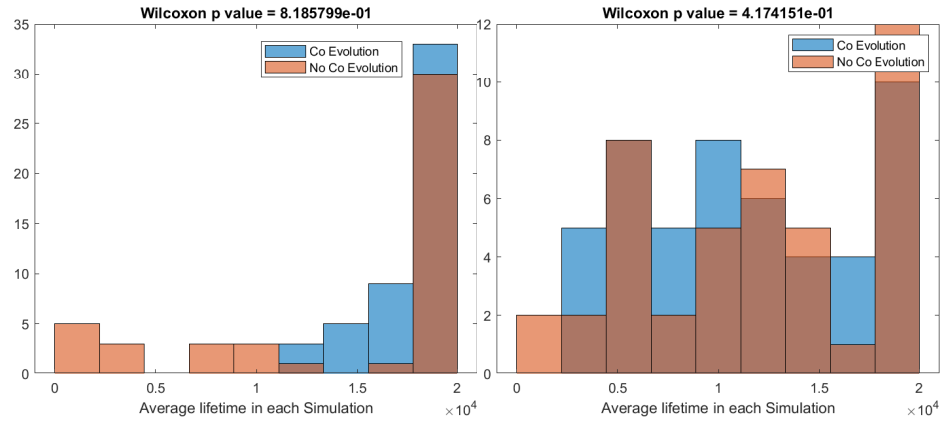


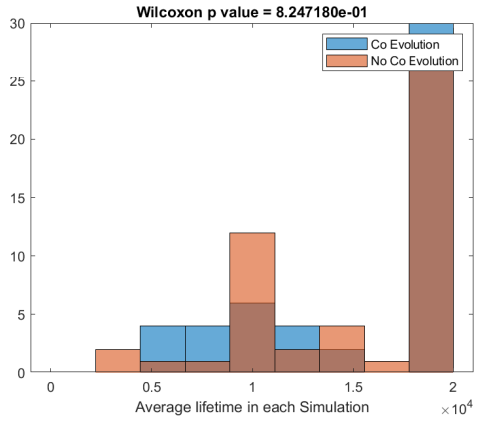

Figure 4.15: **Histograms highlighting impact of coevolution in the model**. Panels (a) and (c) compare the difference in the number of alleles in the final generation of the model. panels (b) and (d) compare the average ages of alleles present in the final generation (the mean of the number of generations that each had been present in the population). Each panel shows the results from 50 simulations. The parameters are as follows: $P_M = 10^{-5}$, N = 5000 and generation simulated to 20,000. For (a) and (b), $a = 0.545559$, $b = 1.12884$, $L = 3$, and $P_I = 0.2$. For (c) and (d), $a = 0.336$, $b = 0.546$, $L = 5$ and $P_I = 0.14$. $a$ and $b$ values were chosen to coincide with a parameter space where the maximum range of allele types coexisted.

It should be noted I use different values for $P_I$ for when $L = 3$ and $L = 5$. I do this as depending on the value of $L$, how high $P_I$ can be before populations go extinct

decreases as $L$ increases. Therefore different values of $P_I$ are needed if I want to pressure the population with pathogen selection but also not let the population go extinct. When $L = 3$ (top row of fig 27), there is not a big difference in the number of alleles present between the two types of simulation. However when I increase $L$ to 5 we start to see there is a difference in the number of alleles in the final generation. The same trend of difference is observed for the average allele lifetimes. As I increase $L$ we see that co evolution mechanisms in the model seem to increase the lifetime of the alleles and also increase the number of alleles present. When $L$ is small the cyclic effects of which attribute is being tested (due to the co evolutionary process of picking the least represented attribute) may not differ mechanically too much from just randomly selecting which attribute is tested. As I increase $L$ the co-evolutionary process is more likely to test a different attribute than one selected at random.

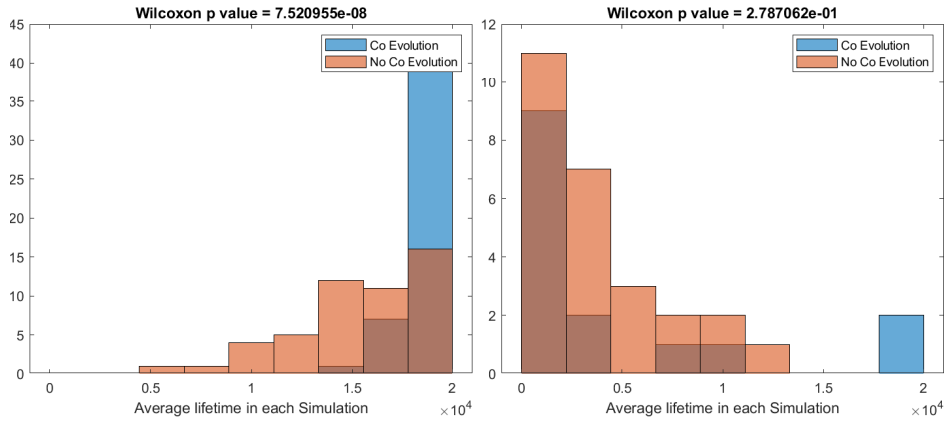



(a) (b)

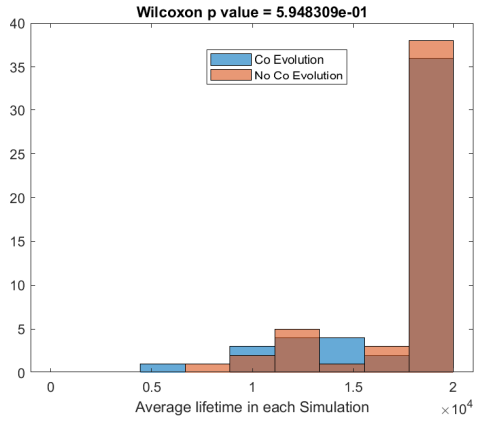



(c)

Figure 4.16: **Histograms highlighting the impact of coevolution on allele lifetime, for different allele types**. All figures are histograms of the mean lifetime of the alleles present in each simulation (how many generations they have been in the population). Figure (a) is the average lifetimes of $S$ alleles, figure (b) is the average lifetimes of $G_2...G_{L-1}$ alleles and figures (c) is the average lifetimes of $G_L$ alleles. The parameters are as follows: $P_M = 10^{-5}$, $a = 0.545559$, $b = 1.12884$, $L = 3$, $P_I = 0.2$, N = 5000 and generation simulated to 20,000. $a$ and $b$ values were chosen to coincide with a parameter space where the maximum range of allele types coexisted.




(a) (b)




(c)

Figure 4.17: **Histograms highlighting the impact of coevolution on allele lifetime, for different allele types**. All figures are histograms of the mean lifetime of the alleles present in each simulation (how many generations they have been in the population). Figure (a) is the average lifetimes of $S$ alleles, figure (b) is the average lifetimes of $G_2...G_{L-1}$ alleles and figures (c) is the average lifetimes of $G_L$ alleles. The parameters are as follows: $P_M = 10^{-5}$, $a = 0.336$, $b = 0.546$, $L = 5$, $P_I = 0.14$, N = 5000 and generation simulated to 20,000. $a$ and $b$ values were chosen to coincide with a parameter space where the maximum range of allele types coexisted.

As can be seen in figure 4.16 and 4.17 , the biggest impact of co-evolution seems to be on the lifetime of specialists ($S$) alleles, and this difference is most striking for $L = 5$

(figure 4.17a).

It seems possible that the differences observed when considering the average lifetime of all alleles in figure 4.15d was mainly being driven by differences in the lifespan of specialist alleles.

$p$ values have been presented for each panel, and they do indicate that the difference seen in figure 4.17a is significant; however, since the distributions of allele lifespans display such widely different patterns in all the panels, interpreting and comparing $p$ values between the panels is not very meaningful.

Since the numbers of simulations were the same for each panel, $p$ values provide a way of ranking how much the distributions differ within each panel (with higher $p$ values indicating small or no difference and smaller $p$ values indicating a bigger difference). I acknowledge however that using a cutoff of 5% is not especially meaningful when dealing with results of simulations, and so defining some of the differences as "significant" is not strictly valid.

### 4.3.6 Discussion

I have presented a model which illustrates how pathogen selection can drive the evolution of specialist or generalist MHC alleles, or the co-existence of the two. My model focuses on the probability of a host being challenged by a pathogen which requires a particular level of MHC specialisation for that host to survive (i.e. the probability of meeting a pathogen where the fastidiousness of MHC molecules matters). I have shown that different distributions of pathogen properties ($P$) can produce all possible combinations of co-existing alleles within my model. However, the specific (and allelic diversity maximising) co-existence of specialist and generalist alleles requires a distribution of $P$ with a particular shape (see figure 4.7). Kaufman [2018] proposes that chickens possess MHC alleles which encompass a more extreme spectrum than those of humans, thus chicken MHC class I alleles range from very fastidious to very promiscuous, whilst human MHC class I alleles occupy a narrower range. Within my model, the white region (which captures the greatest range of coexisting allele types) could thus be compared to the chicken MHC. Humans could be compared to the green regions or any region that involved $G_2...G_{L-1}$.

I have shown that different distributions of $P$ give varying results when it comes to

MHC promiscuity in our model. As I have mentioned the white regions could represent chicken populations while green could represent human populations. The shape of $f(P)$ for the green regions of figure 4.7, has highest point around the value 0.3 for $P$ whereas for the white areas in figure 4.7 $f(P)$ has a much flatter shape across all values of $P$. I have plotted the cumulative distribution of $P$ ($F(P)$) in the supplementary materials (figure C.2) as well as a table showing the values of $F(P)$. These results show us that maybe for white regions to occur, high values of $P$ can only occur every 20th generation and for green regions high values of $P$ cannot occur in order for the middle valued alleles to exist. Our results suggest that perhaps chickens have been subjected to pathogens that require a more fastidious MHC allele recognition, more often than humans have been subjected to such pathogens. The exact shape $P$ would be for organisms is not really possible to know as it is an abstract quantity but it may be possible to deduce, for certain/host pathogen systems, the likely frequency with which a highly fastidious immune response may be required.

Chimpanzees have an MHC allele Patr-B*06:03 that has been found to reduce the viral load of SIV (Wroblewski et al. [2015]). Patr-B*06:03 is structurally similar to HLA*B57, which we know is also associated with a reduced viral load for HIV as mentioned in section 4.1 (Košmrlj et al. [2010], Gao et al. [2010]). As mentioned HLA-B*57 is a fastidious MHC molecule in terms of its peptide repertoire and it is not unreasonable to assume Patr-B*06:03 is also fastidious given the similarity in structure. Given HIV is a relatively new pathogen for humans we will not witness much evolutionary effects of the pathogen however SIV may have been present with chimpanzees for much longer. It would be fascinating to study the population genetic pattern of chimpanzee MHC promiscuity, and relate this to whether SIV tends to be present in every generation, or tends to die out/re occur sporadically in chimpanzee populations.

Is it reasonable to assign a finite number of attributes an MHC allele can be good at? For HLA-A and B there have been categorised 10 HLA supertypes which are groups of binding cleft specificities (Sette and Sidney [1999]). These different groups can account for the vast majority of HLA alleles for A and B. These groups are defined by which alleles share common binding motifs however they do not necessarily represent the peptide binding repertoire each allele has as pointed out by Kaufman [2020]. When I use attributes

in the work here I imagine each attribute could represent a specific peptide shape from a pathogen and the value an allele has in that attribute is how well that shape can be held by the binding cleft of that MHC molecule. When a mutation occurred in our model I generated a new allele and had to assign new values to each attribute. For the majority of the results I did this by making allele types using the "deterministic" method. This method guaranteed I would get alleles of all types equally. However it could easily be argued as not correct as some allele types have more variations than others. For instance if $L = 3$, $S$ type alleles have three variations but $G_3$ alleles only have 1 variation. $G_3$ alleles are just as likely to be generated as $S$ alleles even though there are more variations of $S$ alleles. It could be argued I should have generated every possible variation of alleles equally. How does MHC alleles promiscuity change with mutations in reality? A mutation is only likely to change the properties of one pocket of a binding cleft. I do not know whether the change is likely to make this pocket more or less promiscuous.

In this model we assume a very simplistic infection mechanism. Each generation I infect a proportion $p_I$ of the population randomly. For a single simulation this value of $p_I$ is fixed meaning there is no epidemiological feedback. Another assumption is that if a host did not have a high enough value for $a_i$ when attribute $i$ was being tested they would die if they were infected which could be considered a very strict and fast cut off for survival as it is binary whether you survive or do not survive. The reason for the lack of explicit epidemiological modelling and drastic survival conditions is I wanted a simple way to induce selection on my hosts according to their MHC alleles. Wanting to see how the hosts evolve and also modelling infectious disease would make analyses much more difficult in a already complicated system.

Kaufman [2018] suggests that humans have evolved to have a smaller range of MHC binding promiscuities than chickens due to having multiple MHC class I molecules expressed. An obvious improvement to this study would be the addition of multiple loci in the model. We could see the effects this would have on the evolutionary outcomes over $a$, $b$ parameter space. Given that it is believed chickens have more promiscuous and fastidious alleles due to having only 1 MHC molecule it would be interesting to see if more loci would in fact increase the size of the region of the middle allele types ($G_2...G_{L-1}$).

We observe that when using co evolutionary mechanisms we have higher numbers of

alleles and alleles that have longer lifetimes. This result is in line with other previous theoretical works about co evolution and allelic diversity (Borghans et al. [2004]). Borghans et al. [2004] developed a model where both host and pathogen are selected for. Borghans et al. [2004] compares the stable number of alleles in a simulation when pathogens are selected for as well (coevolution) and when they are not. They find similar results as I have here that coevolution causes there to be a greater number of alleles and that individual alleles stably remain in the population.

Varying MHC molecule peptide repertoire has been observed for humans and for chickens in detail (Paul et al. [2013], Kaufman [2020]). The theoretical study of why MHC molecules would evolve to have varying peptide repertoires has not been studied. Here I present a model that represents MHC alleles as having different attributes that can have finite amounts of fitness in the total of these attributes. I show how a pathogen climate can shape the promiscuity of MCH alleles and what type of pathogen climates give what type of evolutionary outcome.

# Chapter 5

# Conclusion and Future work

In this thesis I explored the evolutionary dynamics and epidemiological consequences of MHC genetic diversity. This work was split into three distinct parts.

In chapter 2 I used an epidemiological ODE model to model a multi strain pathogen in a host population where hosts are defined by their MHC (or, to use the human specific nomenclature HLA) genotype at a single locus. I showed that varying the frequency of HLA alleles in a population changed the levels of which pathogen strain in the system was most prevalent. This in turn changed how an odds ratio measurement would change, if the odds ratio measurement only took into account infection by a pathogen and not a specific pathogen strain. I showed that the perceived protection a HLA allele gave a host from infection was negatively correlated with the presence of said HLA allele.

My results suggest that if case control studies are trying to find associations between HLA alleles and a multi strain pathogen disease outcome, that the associations should be done with specific pathogen strains rather than with infection in general by a pathogen of that species. Some work which has taken this into account has already done this (Toyo-Oka et al. [2017], Salie et al. [2013]). Toyo-Oka et al. [2017] shows that when not grouping their cases into pathogen strains they find no associations but do find associations when splitting the cases into pathogen strain groups. One of the major difficulties (besides having to sequence the pathogen in every infected individual) for future work to take advantage of this insight is how pathogen strains would need to be defined and grouped.

My results suggest a potential solution to this challenge. I observed that if an HLA

allele does confer protection to a specific pathogen strain, the apparent measured protection that HLA allele gives a host against the pathogen in general is inversely correlated with the frequency of said allele. If multiple case control studies of HLA associations with infectious diseases have been carried out in different populations, then it would be possible to measure if a particular HLA allele's association with infection with a multi strain pathogen is correlated with the frequency of said allele within a population. Future work could be to conduct a specific form of meta analysis of HLA case control studies. This would involve plotting the odds ratio for HLA protectiveness against infection against the frequency of the HLA allele of interest and see if any correlation of the two is occurring. If a correlation is occurring it could be a sign that the allele is associated with a particular strain rather than the pathogen itself.

Case control studies to find associations between MHC alleles and disease outcome are still being performed (Ben Shachar et al. [2021]). Ben Shachar et al. [2021] perfomed a case control study where the cases were individuals who tested positve for SARS-CoV-2 by PCR and the controls were individuals who tested negative. Ben Shachar et al. [2021] found no associations between any HLA alleles and testing positive or negative for SARS-CoV-2 by PCR. SARS-Cov-2 is still evolving and it would be interesting to see if HLA alleles are more or less protective in association with new emerging strains.

In chapter 3 I used an individual based model to model copy number variation (CNV) among MHC genes in a population. I used unequal crossing over as a mechanism for generating more copies of an MHC gene on a chromosome, and allowed mutations to generate variation amongst MHC genes. I tested different rules for how MHC genotypes (with variable numbers of copies of genes) would interact with pathogens. I found that a "mean fitness rule" (where the fitness of a host is equal to the average fitness of said hosts allele fitnesses) meant that the number of copies of MHC genes on chromosomes was inversely correlated with the intensity of changing pathogen selection. Such a mechanism could explain why species such as macaque and humans have varying CNV for the MHC. My results imply that humans could have less CNV in the MHC than macaques due to having a more changing pathogen climate compared to macaques.

I further show that when recombination happens in between exons of a gene (where

the gene itself gets new properties at the same time as possibly increasing or decreasing the number of copies of a gene in a cluster) we see a general trend to observing fewer copies of MHC genes in clusters. I showed the steady state for a scenario including only between-exon recombination, with no pathogen selection present, to be a single copy of a gene for the entire population. Even with this steady state, however, interactions with pathogens (e.g. obeying the maximum fitness rule ) can generate systems which have multiple copies of MHC genes.

To fully understand the mechanisms occurring in chapter 3, a more analytical approach could be taken that gives more insight than using an individual based model. Analytical work on unequal crossing over and the distribution of repeated genes has had some attention (Krüger and Vogel [1975], Takahata [1981], Baake [2008], Shpak and Atteson [2002]). Extensions of works like this to include selection that is more fitting to MHC genes could give clearer insights why some species exhibit CNV but others do not, that are difficult to extract from an individual based model.

In chapter 4 I used an individual based model in which MHC alleles were given different attribute values representing how well the MHC molecule could bind to a particular type of peptide shape. I showed that varying pathogen climate, in terms of how frequently the recognition of a particular peptide with a particular level of binding fastidiousness is needed, could generate differing MHC genetic landscapes in terms of MHC peptide binding repertoire. The frequency of requiring recognition of a pathogen peptide with a particular degree of fastidiousness was captured by the distribution of parameter $P$. I compared evolutionary outcomes from the model to the examples of humans and chicken MHC peptide repertoires and showed what $f(P)$ distributions caused these outcomes. I noted that the cases where all allele types were present (i.e where specialist and all potential forms of generalists co-exist) were for particular shapes of $f(P)$. My model also demonstrated a previously identified behaviour (Borghans et al. [2004]), that the inclusion of co evolutionary mechanisms produced systems with greater numbers of MHC alleles.

A useful extension of this work would be to use the same model principles but extend it to have multiple numbers of loci. As mentioned previously Kaufman [2018] suggests that the reasons humans do not display such stark differences between very fastidious

and very specialised MHC alleles could be due to the fact they have multiple MHC class I loci. With the model presented here and increasing the number of alleles one could see how it affects the promiscuity of alleles in the system.

Current research is investigating the potential of deep neural networks to predict binding affinities of MHC molecules (Jin et al. [2021], Jiang et al. [2021], Cheng et al. [2021]). Such work could also be used like Paul et al. [2013] to analyse predicted binding repertoires as well as peptide binding affinities and could extend to see if such quantities correlate. This could also be extended to see how such quantities like the peptide binding repertoire or peptide binding affinities vary across different species or populations. Any patterns observed could be the subject of further study using evolutionary simulations building on the work of chapter 4.

The use of theoretical models for examining MHC evolution is a difficult process as for every model huge assumptions have to be made and decisions on model mechanisms have to be made when the reality is certain mechanisms are just unknown. For example in chapter 3 I assume that MHC alleles experience negative frequency dependent selection, when the topic of whether or not frequency dependent selection is occurring for the MHC is still debated (Peng et al. [2021]). Until greater experimental evidence is available, such assumptions will always be the weak points of any theoretical modelling of the MHC. In the absence of experimental evidence, variations of model types on the same questions is a great way to explore mechanically why things are occurring. One of the best ways models can help is when they suggest a new way of measuring something that might give insight into underlying mechanisms in the MHC's evolution. For example in chapter 2 I give an example of how we might observe frequency dependence in the measured ability a HLA allele offers a host using meta analysis of case control studies see section 2.4.

In the work here I have used individual based models and ODE models. Individual based models have the advantage of being able to model more complicated mechanisms but at the same time are very difficult to analyse. ODE models are useful and are easy to model infectious disease mechanics as well as being easier to analyse. However for complicated host types as I have explored in chapter 3 with varying numbers of loci, an individual model seemed like a fitting type of model.

The MHC is an extremely interesting gene region being one of the most polymorphic in all vertebrates. Given the role of MHC molecules in the immune system it is highly likely that MHC loci have co evolved with pathogens. This very relationship is why it has been very difficult to determine the exact mechanisms that causes the MHC's many interesting features. This thesis has contributed to three specific, yet relatively-little studied questions about the MHC:

- Why do case control studies that investigate the association between HLA alleles and infectious disease outcomes get contradicting results?
- Why have different species evolved to have varying degrees of CNV for the MHC?
- Why have different species evolved to have varying MHC promiscuity?

I hope my models and the results of my analyses presented here have gone some way towards answering these questions, and that my work will help contextualise and explain MHC properties and behaviours as they continue to be uncovered.

# Appendix A

# Chapter 2 supplementary material

## A.1 Table of Notation

| Notation | Definition |
|---|---|
| $\sigma_i$ | Recovery rate for hosts who do not have the correct allele for memory immune responses. |
| $\mu_i$ | Recovery rate for hosts who do have the correct allele for memory immune responses. |
| $\beta_i$ | Transmission rate. |
| $c$ | A proportion $c$ of the force of infection from a pathogen strain will be applied to hosts who are already infected with another pathogen strain. |
| $\alpha$ | A proportion $\alpha$ of recoveries will recover to being immune to all pathogen strains. |
| $p_i$ | Gene frequency of allele $i$. |
| $N_{ij}^{nm}$ | Proportion of hosts of genotype $ij$ in the n state with pathogen strain 1 and in the m state with pathogen strain 2. |
| $S_i$ $I_i$ $R_i$ | The proportion of the population that are Susceptible, Infected and Recovered respectively with pathogen strain $i$. |
| $P(I \mid i)$<br>$P(I \mid \hat{i})$ | The proportion of hosts with allele $i$ that are infected and the the proportion of hosts without allele $i$ that are infected respectively. |
| $OR_i$ | The odds ratio of being infected given you have HLA allele $i$. |
| $n(I \mid i)$<br>$n(I \mid \hat{i})$<br>$n(\hat{I} \mid i)$<br>$n(\hat{I} \mid \hat{i})$ | The number of hosts who are infected and have allele $i$. The number of hosts who are infected and do not have allele $i$. The number of people who are not infected and have allele $i$. The number of people who are not infected and do not have allele $i$ respectively. |
| $\Omega_B$ $\Omega_A$ | The difference between the minimum and maximum values of $p_1$ at which the odds ratio is significantly below and above 1 respectively. |
| $I_{ii}$ $I_{i\hat{i}}$ $I_{\hat{i}\hat{i}}$ | The proportion of hosts who are infected in the genotype group who are homozygous with HLA allele $i$, who are heterozygous with HLA allele $i$ and who do not have HLA allele $i$ respectively |

Table A.1: **Table of notations used in chapter 2**

## A.2 Two Strain, Two Allele Model

The ODEs for the main two strain, two allele model in which HLA allele 1 confers the ability to develop a memory immune response against pathogen strain 1, and HLA allele 2 confers the ability to develop a memory immune response against allele 2, are as follows:

$$\begin{aligned}
\frac{dN_{ij}^{SS}}{dt} &= dp_i p_j(2-\delta_{i,j}) - \sum_k \lambda_k N_{ij}^{SS} + ((1-\delta_{1,ij})\sigma_1 N_{ij}^{IS} \\
&\quad + (1-\delta_{2,ij})\sigma_2 N_{ij}^{SI})(1-\alpha) - dN_{ij}^{SS} \\
\frac{dN_{ij}^{SI}}{dt} &= \lambda_2 N_{ij}^{SS} - \lambda_1 N_{ij}^{SI} c - (1-\delta_{2,ij})\sigma_2 N_{ij}^{SI} - \delta_{2,ij}\mu_2 N_{ij}^{SI} \\
&\quad + (1-\delta_{1,ij})\sigma_1 N_{ij}^{II})(1-\alpha) - dN_{ij}^{SI} \\
\frac{dN_{ij}^{IS}}{dt} &= \lambda_1 N_{ij}^{SS} - \lambda_2 N_{ij}^{IS} c - (1-\delta_{1,ij})\sigma_1 N_{ij}^{IS} - \delta_{1,ij}\mu_1 N_{ij}^{IS} \\
&\quad + (1-\delta_{2,ij})\sigma_2 N_{ij}^{II})(1-\alpha) - dN_{ij}^{IS} \\
\frac{dN_{ij}^{II}}{dt} &= \lambda_1 N_{ij}^{SI} c + \lambda_2 N_{ij}^{IS} c - (1-\delta_{1,ij})\sigma_1 N_{ij}^{II} - (1-\delta_{2,ij})\sigma_2 N_{ij}^{II} \\
&\quad - \delta_{1,ij}\mu_1 N_{ij}^{II} - \delta_{2,ij}\mu_2 N_{ij}^{II} - dN_{ij}^{II} \\
\frac{dN_{ij}^{SR}}{dt} &= \delta_{2,ij}\mu_2 N_{ij}^{SI} - \lambda_1 N_{ij}^{SR} + (1-\delta_{1,ij})\sigma_1 N_{ij}^{IR}(1-\alpha) - dN_{ij}^{SR} \\
\frac{dN_{ij}^{IR}}{dt} &= \lambda_1 N_{ij}^{SR} - (1-\delta_{1,ij})\sigma_1 N_{ij}^{IR} - \delta_{1,ij}\mu_1 N_{ij}^{IR} + \delta_{2,ij}\mu_2 N_{ij}^{II}(1-\alpha) - dN_{ij}^{IR} \\
\frac{dN_{ij}^{RS}}{dt} &= \delta_{1,ij}\mu_1 N_{ij}^{IS} - \lambda_2 N_{ij}^{RS} + (1-\delta_{2,ij})\sigma_2 N_{ij}^{RI}(1-\alpha) - dN_{ij}^{RS} \\
\frac{dN_{ij}^{RI}}{dt} &= \lambda_2 N_{ij}^{RS} - (1-\delta_{2,ij})\sigma_2 N_{ij}^{RI} - \delta_{2,ij}\mu_2 N_{ij}^{RI} + \delta_{1,ij}\mu_1 N_{ij}^{II}(1-\alpha) - dN_{ij}^{IR} \\
\frac{dN_{ij}^{RR}}{dt} &= \delta_{1,ij}\mu_1 N_{ij}^{IR} + \delta_{2,ij}\mu_2 N_{ij}^{RI} + \alpha((1-\delta_{1,ij})\sigma_1(N_{ij}^{IS} + N_{ij}^{IR} + N_{ij}^{II}) \\
&\quad + (1-\delta_{2,ij})\sigma_2(N_{ij}^{SI} + N_{ij}^{RI} + N_{ij}^{II}) + \delta_{1,ij}\mu_1 N_{ij}^{II} + \delta_{2,ij}\mu_2 N_{ij}^{II}) - dN_{ij}^{RR},
\end{aligned} \tag{A.1}$$

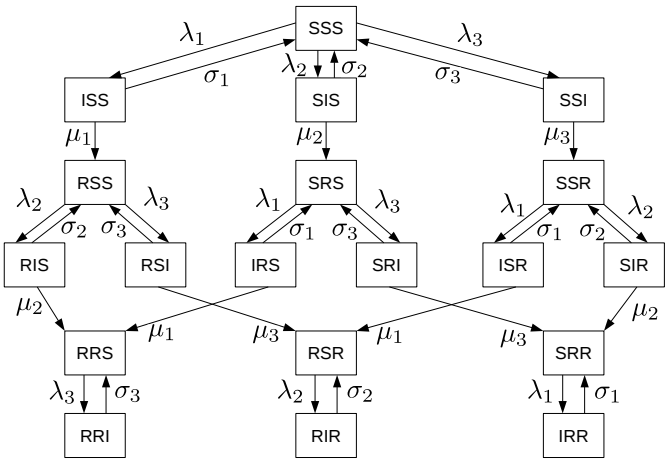
where

$$\delta_{1,ij} = \delta_{1,i} + \delta_{1,j} - \delta_{1,i}\delta_{1,j}, \tag{A.2}$$

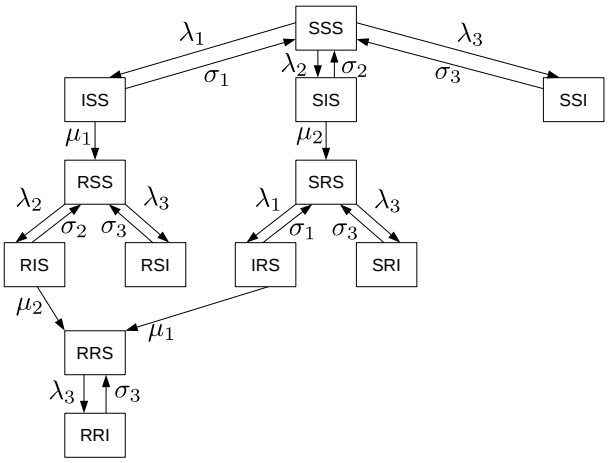
where $\delta_{1,ij} = 1$ if at least one index is 1 and 0 otherwise i.e. terms preceded by $(1-\delta_{1,ij})$ contribute only to individuals of genotype 22.

## A.3 Three Strain, Three Allele Model

In order to investigate the impact of increasing complexity I introduced a third pathogen strain and a third HLA allele at the considered locus.




(a)




(b)

Figure A.1: **The possible infectious states of the population and pathways between them in the three strain, three allele model.** (a) is a flow chart of the possible paths a host of any genotype $ij$ can take from initially being susceptible to all pathogen strains. (b) is a flow chart specifically for a host of genotype 11.

Figure A.1 illustrates the different states possible for hosts in the 3 strain model. The the notation in figure A.1 is the same as for two strain, two allele model described in the Methods of the Main text. SIR indicates a host is susceptible to pathogen strain 1, infected with pathogen strain 2 and immune to pathogen strain 3. There are six genotypes in this model ($ij = 11, 12, 22, 23, 33, 13$). Since a host can only posses two HLA alleles, no host can become immune to all three pathogen strains, so there is no RRR class of hosts. Representative ODEs for the three strain, three allele model are as follows:

$$\begin{aligned}
\frac{N_{ij}^{SSS}}{dt} =& dp_i p_j(2-\delta_{i,j}) - \sum_k \lambda_k N_{ij}^{SSS} \\
&+ (1-\delta_{1,ij})\sigma_1 N_{ij}^{ISS} + (1-\delta_{2,ij})\sigma_2 N_{ij}^{SIS} \\
&+ (1-\delta_{3,ij})\sigma_3 N_{ij}^{SSI} - dN_{ij}^{SSS} \\
\frac{N_{ij}^{ISS}}{dt} =& \lambda_1 N_{ij}^{SSS} - (1-\delta_{1,ij})\sigma_1 N_{ij}^{ISS} - \delta_{1,ij}\mu_1 N_{ij}^{ISS} - dN_{ij}^{ISS} \\
\frac{N_{ij}^{RSS}}{dt} =& \delta_{1,ij}\mu_1 N_{ij}^{ISS} - \lambda_2 N_{ij}^{RSS} - \lambda_3 N_{ij}^{RSS} \\
&+ (1-\delta_{2,ij})\sigma_2 N_{ij}^{RIS} + (1-\delta_{3,ij})\sigma_3 N_{ij}^{RSI} - dN_{ij}^{RSS}
\end{aligned} \tag{A.3}$$

There is no coinfection or strain transcending immunity included in the three strain model.

## A.4 Further analyses including HLA alleles which are not strain specific in their effects.

In order to investigate how the presence of HLA alleles with non-strain-specific effects affect the outcomes of my model, I adapted the three allele system so that allele 1 and 2 remain strain specific, but allele 3 confers no ability to recognise any strain (a useless allele) or allele 3 confers the ability to recognise strain 1 and strain 2 (a perfect allele). I only allowed pathogen strains 1 and 2 to circulate. Tables A.2 and A.3 illustrate which epidemiological rules different host genotypes follow for each pathogen strain.

| **Genotype** | **Strain 1 behaviours** | **Strain 2 behaviours** |
|---|---|---|
| 11 | SIR | SIS |
| 12 | SIR | SIR |
| 13 | SIR | SIS |
| 22 | SIS | SIR |
| 23 | SIS | SIR |
| 33 | SIS | SIS |

Table A.2: Host epidemiology when HLA allele 3 recognises neither pathogen strain 1 and 2.

| **Genotype** | **Strain 1 behaviours** | **Strain 2 behaviours** |
|---|---|---|
| 11 | SIR | SIS |
| 12 | SIR | SIR |
| 13 | SIR | SIR |
| 22 | SIS | SIR |
| 23 | SIR | SIR |
| 33 | SIR | SIR |

Table A.3: Host epidemiology when HLA allele 3 recognises both pathogen strains 1 and 2.

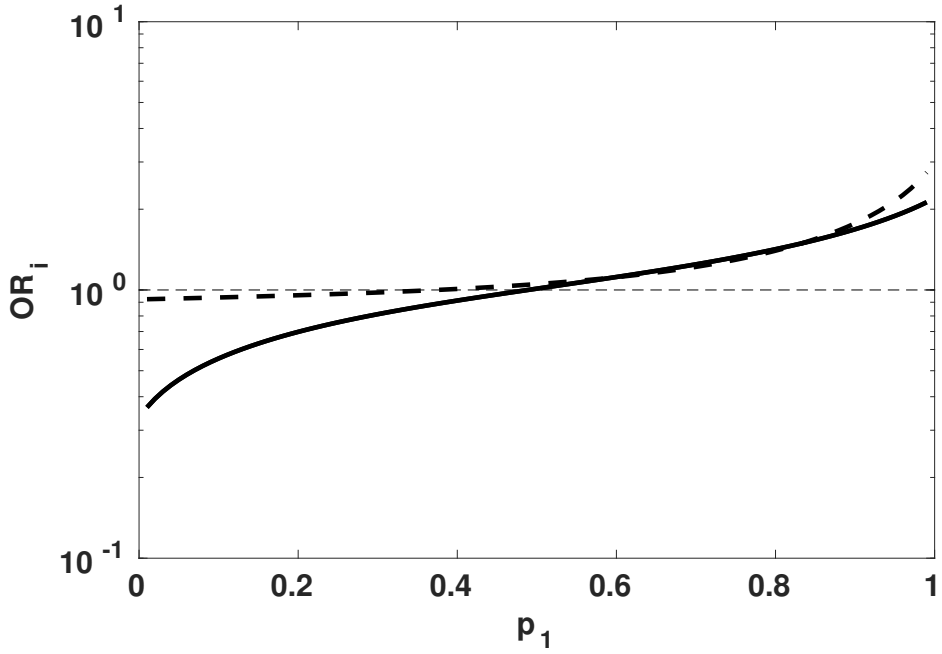



Figure A.2: $OR_1$ **and its relationship with the frequency of HLA allele 1 in a 3 allele system containing a useless or perfect allele.** The solid line is $OR_1$ when allele 3 is useless (does not confer the ability to mount a memory immune response against any strain). The dashed line is $OR_1$ when allele 3 is perfect (confers the ability to mount a memory immune response against all strains). The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_i = 0.02$ and $\sigma_i == 0.02$ $(i = 1, 2)$. For the solid line $\sigma_3 = 0.02$ for the dashed line $\mu_3 = 0.02$.

When allele 3 is useless we still see that $OR_1$ is positively correlated with $p_1$, and switches from below 1 to above 1 as $p_1$ increases (figure A.2, solid line). When allele 3 is a perfect allele (figure A.2, dashed line) it is harder for allele 1 to exhibit strong protectiveness against the prevailing local infection, due to allele 3 being protective against both pathogens strains. However, despite this, even when allele 3 is a perfect allele $OR_1$ is below 1 for low values of $p_1$ and $OR_1$ follows the same general trends we see in the multi pathogen strain systems in figure 2.

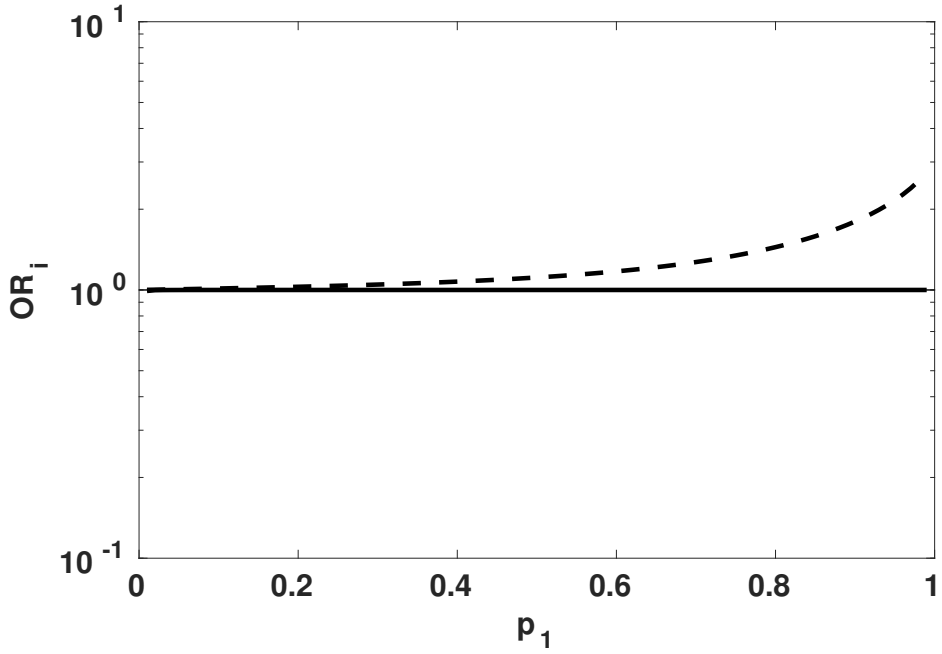



Figure A.3: $OR_1$ **and its relationship with the frequency of HLA allele 1 in a 2 allele system containing a useless or perfect allele.** The solid line is $OR_1$ when allele 2 is useless (does not confer the ability to mount a memory immune response against any strain). The dashed line is $OR_1$ when allele 2 is perfect (confers the ability to mount a memory immune response against all strains). The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_1 = 0.02$ and $\sigma_1 = 0.02$ $(i = 1, 2)$. For the solid line $\sigma_2 = 0.02$ for the dashed line $\mu_2 = 0.02$.

In a system containing only alleles 1 and 2, where allele 2 is a useless allele, $OR_1 = 1$ for almost all of $p_1$ (figure A.3, solid line). This is due to pathogen strain 2 out competing pathogen strain 1 in this system at all frequencies of $p_1$. In this system, no host can ever become immune to strain 2, so strain 2 always has an advantage. Once strain 2 has out competed strain 1, allele 1 is just as useless as allele 2, hence $OR_1 = 1$.

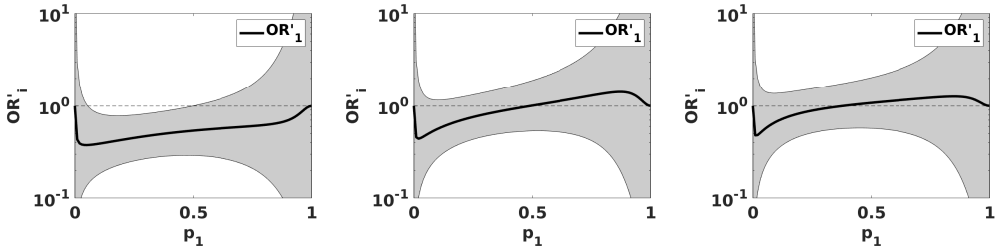
In the case where HLA allele 2 is a perfect allele (figure A.3, dashed line), hosts who have HLA allele 1 are always found to be more at risk of being infected with the prevailing local infection. Genotypes containing HLA allele 2 will be able to become immune to both

strains 1 and 2, whilst genotypes containing only HLA allele 1 will only ever be able to become immune to strain 1, hence the constant advantage to allele 2.

## A.5 Further analyses of how protective or risky associations may arise and be detected.

### A.5.1 Lower sample sizes

In figure 7 of the main text I show how $OR_1$ and its 95% confidence intervals vary when I introduce more pathogen strains into the system. Here I reproduce the same plot but for the case where we have 100 cases and 100 controls as the sample size.

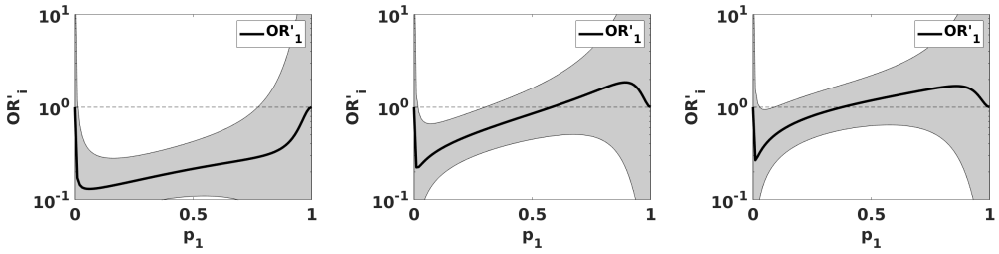



(a) (b) (c)

Figure A.4: **$OR'_1$ and its relationship with the frequency of HLA allele 1.** The highlighted regions are the 95% confidence intervals for $OR'_i$. (a) illustrates a system with one pathogen strain and two HLA alleles in the population. (b) illustrates a system with two pathogen strains and two strain specific HLA alleles within the population. (c) illustrates a system with three pathogen strains and three strain specific HLA alleles within a population. The confidence intervals were calculated with a sample size of 100 cases and 100 controls (see Methods for further details). The parameter values are: $d = 0.01$, $\beta_i = 0.06$, $\mu_i = 0.02$ and $\sigma_i = 0.02$ $(i = 1, 2, 3)$.

We see in figures A.4b and A.4c (where more than one pathogen strain is present) that the confidence interval for $OR_1$ overlaps 1 at all allele frequencies. The trend of $OR_1$ is identical to that of figure 7 in the main text and we see in figure A.4b that the upper boundary of the confidence interval is closer to the value 1 than it is in figure A.4c.

However if I increase $R_0$ we do get frequencies of $p_1$ at which $OR_1$ is significantly below 1 even with the smaller sample size (figure A.5).




(a) (b) (c)

Figure A.5: **$\mathbf{OR_1'}$ and its relationship with the frequency of HLA allele 1.** The highlighted regions are the 95% confidence intervals for $OR_i'$. (a) illustrates a system with one pathogen strain and two HLA alleles in the population. (b) illustrates a system with two pathogen strains and two strain specific HLA alleles within the population. (c) illustrates a system with three pathogen strains and three strain specific HLA alleles within a population. The confidence intervals were calculated with a sample size of 100 cases and 100 controls (see Methods for further details). The parameter values are: $d = 0.01$, $\beta_i = 0.15$, $\mu_i = 0.02$ and $\sigma_i = 0.02$ $(i = 1, 2, 3)$.

### A.5.2 Further analyses of how pathogen properties affect $OR$ and $\Omega$.

In the Methods I introduced the odds ratio for allele $i$ ($OR_i$):

$$OR_i = \left(\frac{P(I \mid i)}{1 - P(I \mid i)}\right) \Big/ \left(\frac{P(I \mid \hat{i})}{1 - P(I \mid \hat{i})}\right). \tag{A.4}$$

In the results of the main text I defined $p_{crit}$ as the $p_i$ value at which $OR_i$ crosses the value 1. To find an analytical formula for $p_{crit}$ I first write out the terms in equation (A.4) as functions of $p_i$.

$$\begin{aligned} P(I \mid i) =& \frac{I_{ii}p_i^2 + I_{i\hat{i}}2p_i(1-p_i)}{p_i^2 + 2p_i(1-p_i)}, \\ P(I \mid \hat{i}) =& I_{\hat{i}\hat{i}}. \end{aligned} \tag{A.5}$$

Here $I_{ii}$ is the proportion of the homozygotes who have allele $i$ that are infected. $I_{i\hat{i}}$ is the proportion of the hetrozygotes who have allele $i$ and are infected. $I_{\hat{i}\hat{i}}$ is the proportion of hosts who do not have allele $i$ and are infected. Substituting (A.5) into the formula for $OR_i$ and applying the conditions, $p_i = p_{crit}$ when $OR_i = 1$ I can solve for $p_{crit}$ and acquire an expression for $p_{crit}$.

$$p_{crit} = \frac{2I_{\hat{i}\hat{i}} - 2I_{i\hat{i}}}{I_{ii} + I_{\hat{i}\hat{i}} - 2I_{i\hat{i}}}. \tag{A.6}$$

I calculated values for $I_{ii}$, $I_{i\hat{i}}$ and $I_{\hat{i}\hat{i}}$ numerically, and explored how $p_{crit}$ varies when $\sigma$ and $R_0$ are varied, for both the two strain and three strain systems (figure A.6).

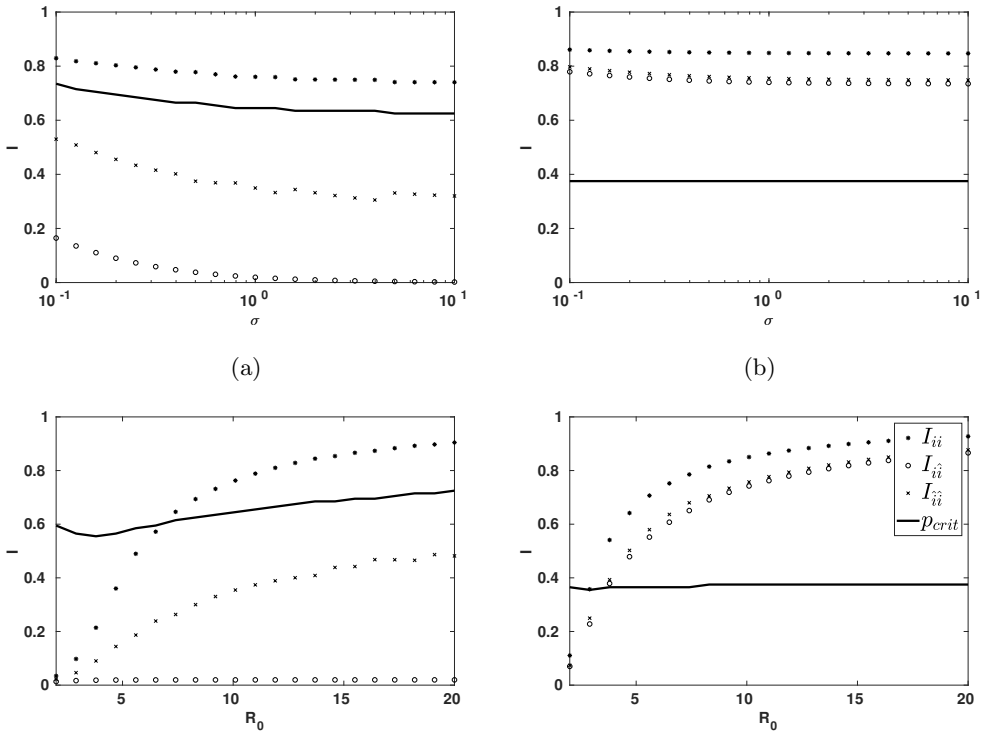



Figure A.6: **The behaviour of $\mathbf{p_{crit}}$, and key quantities affecting $\mathbf{p_{crit}}$, as $\mathbf{R_0}$ and $\sigma$ are varied in the two strain, two allele system and the three strain, three allele system**. (a) and (c) are of the two pathogen strain model. (b) and (d) are of the three pathogens strain model. The parameter values are: $d = 0.01$. For (a) and (b) $R_0 = 5$, $\mu_i$ and $\sigma_i$ are increased from $10^{-1}$ to $10^1$. For (c) and (d) $\mu_i$ and $\sigma_i$ are $10^{-1}$ and $R_0$ is increased from 2 to 20.

In figures 7 and 8 of the main text, I introduced $\Omega$, the difference between the maximum and minimum values of $p_1$ at which the odds ratio is significantly below ($\Omega_B$ ) or above ($\Omega_A$). $p_{crit}$ is the maximum possible value of $\Omega_B$, and $(1 - p_{crit})$ is the maximum possible value of $\Omega_A$. The behaviour of $p_{crit}$ under different values of $R_0$ and $\sigma$, in the two strain, two allele and three strain, three allele systems (figure A.6) helps to explain the patterns observed in figures 7 and 8 of the main text. The value $p_{crit}$ takes is determined by the balance between $I_{ii}$, $I_{i\hat{i}}$ and $I_{\hat{i}\hat{i}}$, and the relationship between these quantities is

ultimately determined by competition between the pathogen strains. Anything which promotes coexistence of the pathogen strains at higher values of $p_1$ increases $p_{crit}$, and anything which discourages that coexistence decreases $p_{crit}$.

In the two strain, two allele system there is a negative relationship between $\sigma$ and $p_{crit}$ (figure A.6a). Increasing $\sigma$ it increases the competition between the two strains. Figures A.7 and A.8 show that as $\sigma$ is increased the transition at which one pathogen strain dominates over other occurs over a shorter range of $p_1$. Increasing $\sigma$ allows pathogen strain 2 to dominate at lower values of $p_1$, hence decreasing $p_{crit}$.

In the two strain, two allele system a particular value of $R_0$ is associated with a minimum value of $p_{crit}$. As $R_0$ is increased or decreased from this value, $p_{crit}$ increases (figure A.6c). As shown in figure A.8 the fastest relative change between the frequencies of the two pathogen strains occurs when $R_0 = 5$ for all values of $\sigma$ shown. This is the value of $R_0$ where, $p_{crit}$ is at its lowest. At values of $R_0 < 5$, the two pathogen strains coexist, both at a low level, for a wide range of values of $p_1$ (see figure A.7). When both pathogen strains are only present at a low level it is harder for one to out-compete the other. At values of $R_0 > 5$, the two pathogens are also more likely to coexist, since they both occur at generally higher frequencies (figure A.7). The value of $R_0$ which minimises coexistence (here approximately 5), also minimises $p_{crit}$. Changes away from this value of $R_0$ in either direction increase strain coexistence and increase $p_{crit}$.

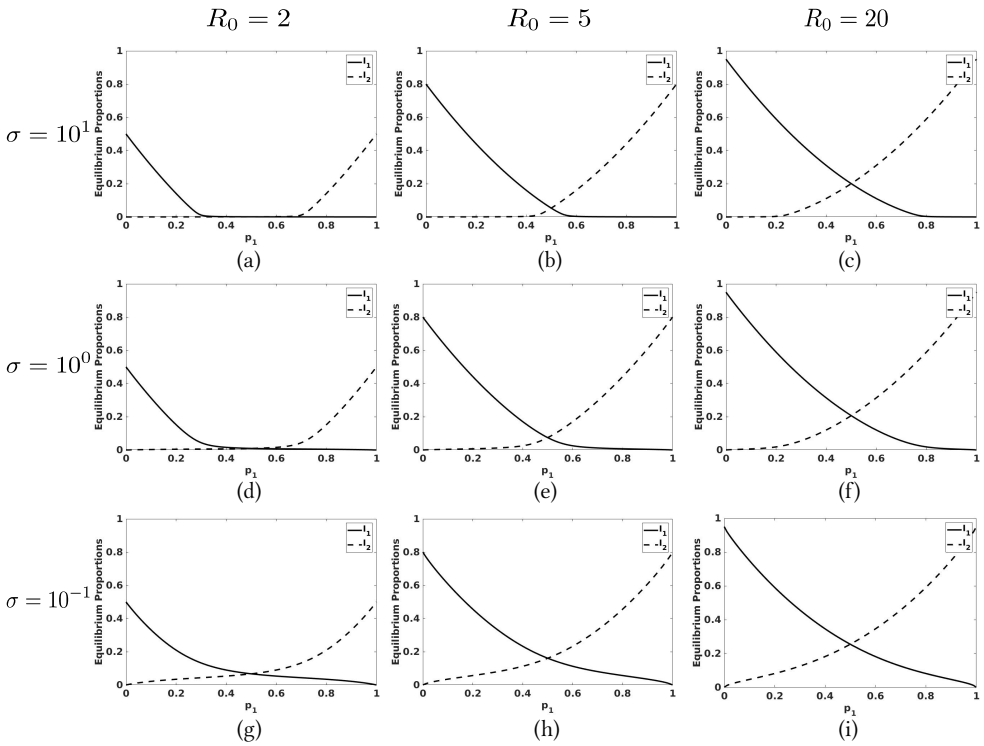



Figure A.7: **The effect of the frequency of allele 1 on the balance between the pathogen strains, and how this changes for different values of $\mathbf{R_0}$ and $\sigma$, in the two strain, two allele system.** The parameter values are: $d = 0.01$. For (a), (b) and (c) $\sigma = 10^1$, for (d), (e) and (f) $\sigma = 10^0$ and for (g), (h) and (i) $\sigma = 10^{-1}$. For (a), (d) and (g) $R_0 = 2$, for (b), (e) and (h) $R_0 = 5$ and for (c), (f) and (i) $R_0 = 20$. $\beta_i$ is calculated as $\beta_i = R_0(d + \sigma)$.

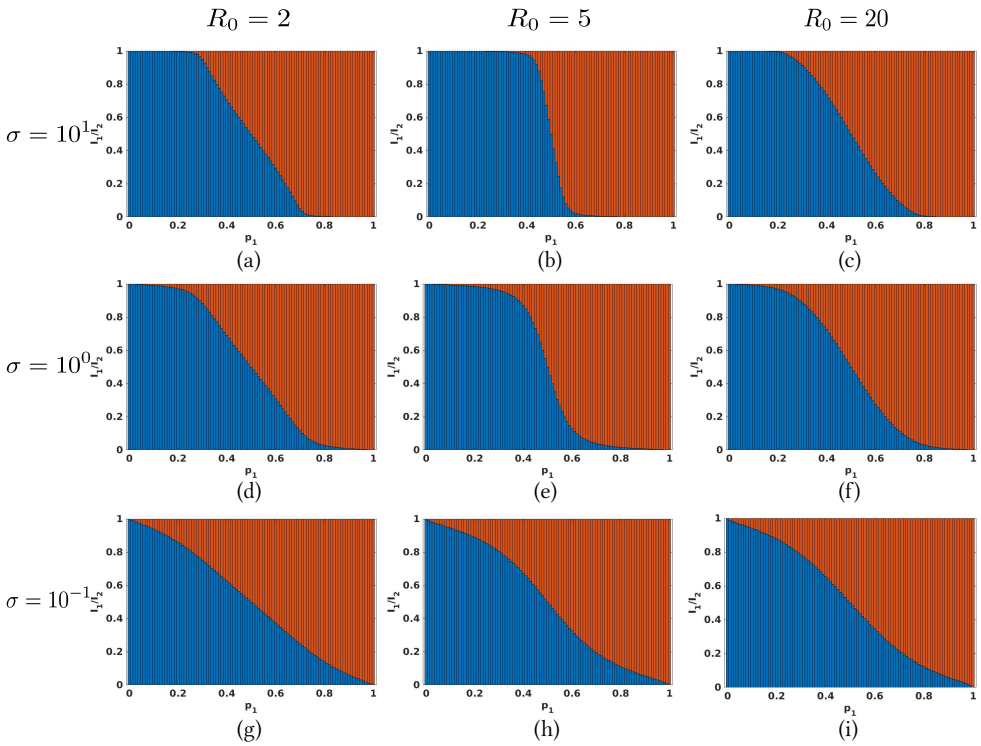



Figure A.8: **The effect of the frequency of allele 1 on the balance between the pathogen strains, and how this changes for different values of $\mathbf{R_0}$ and $\sigma$, in the two strain, two allele system.** The parameter values are: $d = 0.01$. For (a), (b) and (c) $\sigma = 10^1$, for (d), (e) and (f) $\sigma = 10^0$ and for (g), (h) and (i) $\sigma = 10^{-1}$. For (a), (d) and (g) $R_0 = 2$, for (b), (e) and (h) $R_0 = 5$ and for (c), (f) and (i) $R_0 = 20$. $\beta_i$ is calculated as $\beta_i = R_0(d + \sigma)$.

In the three strain, three allele system, we observed very little relationship between $R_0$ and $\sigma$ and $\Omega_A$ and $\Omega_B$, other than a minimal threshold value of $R_0$ and $\sigma$ required to generate a statistically significant pattern. As shown in figure A.6, in the three strain, three allele system, changes in $R_0$ and $\sigma$ have a much smaller impact on $p_{crit}$ than in the two strain, two allele system. This is due to the different characteristics of $I_{i\hat{i}}$ in the two strain and three strain systems, which places different constraints on $p_{crit}$. In the two strain model, $I_{i\hat{i}}$ contains only genotype 12 hosts who can become immune to either strain in the population. In the three strain model, $I_{i\hat{i}}$ contains genotype 12 and 13 hosts,

which can each only become immune to two out of the three strains in the system. Hence, in the three strain model, $I_{i\hat{i}}$ takes higher values, and behaves far more similarly to $I_{\hat{i}\hat{i}}$ than in the two strain model. The relationship between $I_{ii}$, $I_{i\hat{i}}$ and $I_{\hat{i}\hat{i}}$ determines what values $p_{crit}$ can take. In the 3 strain model, $I_{ii}$, $I_{i\hat{i}}$ and $I_{\hat{i}\hat{i}}$ are more similar than in the 2 strain model, and this limits the range of values of $p_{crit}$. The limited range of values of $p_{crit}$ leads to the lower variability in $\Omega_B$ and $\Omega_A$ observed for the three strain, three allele system.

# Appendix B

# Chapter 3 supplementary material

## B.1 Table of Notation

| Notation | Definition |
|---|---|
| $G$ | Matrix of integers that represent the populations MHC alleles. |
| $I_L$ | Initial number of loci in the system. |
| $I_A$ | Initial number of unique MHC alleles in the system. |
| $N$ | Population size which is constant during simulation. |
| $U_L$ | Upper limit on the number of copies of a gene that can be present on a chromosome. |
| $S = \sum_{n=1}^{N} S_n$ | Individual $n$ has a fitness of $S_n$ and the sum of the populations fitness is $S$. |
| $[ij]$ | exon $i$ and $j$ form MHC allele $[ij]$. |
| $A_{ij}$ | the raw fitness value of MHC allele $[ij]$ . |
| $p_{ij}$ | the allele frequency of MHC allele $[ij]$ . |
| $f_{ij} = A_{ij}(1 - p_{ij})$ | the actual fitness of MHC allele $[ij]$ after we account for negative frequency dependence. |
| $p_M$ | The probability a mutation event will occur per individual loci |
| $p_R$ | The probability a recombination event will occur per individual loci |
| $L$ | The number of copies of a gene in a MHC cluster |
| $\bar{L}$ | The average number of copies of a gene in a MHC cluster in the population |
| $\beta$ | The proportion of recombination events where the break point occurs between the exons of a gene |
| $\alpha$ | The proportion of MHC alleles that has its raw allele fitness changed due to changing pathogen selection in the model |
| $\sigma$ | The standard deviation of the truncated normal distribution to assign a alleles raw fitness after changing pathogen selection has occurred |

Table B.1: **Table of notations used in chapter 2**

## B.2 Time Series

To better understand how the simulations behave here I show a time series of the mean of $\bar{L}$ and the error bars are the 95 percentile of the distribution of $\bar{L}$ (figure B.1).

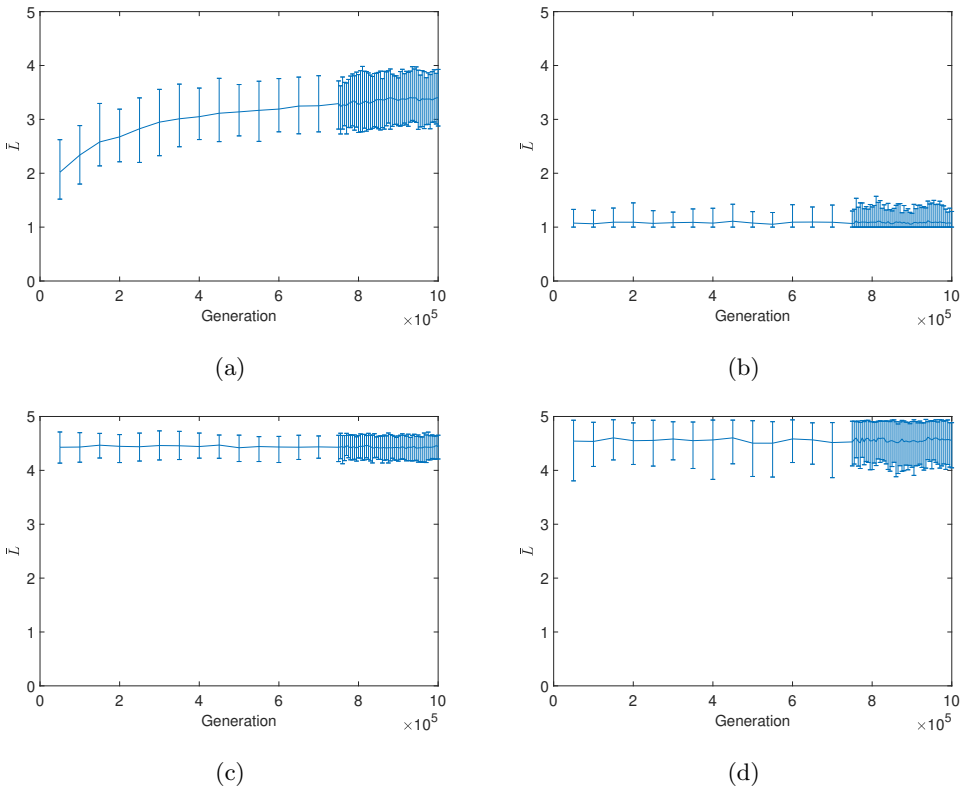



Figure B.1: **Showing a time series of $\bar{L}$.** (a) and (b) use the mean fitness rule and (c) and (d) use the max fitness rule. The parameters used are as follows: N = 5000, Initial number of exons = 10, Initial number of Loci = 2, $U_L = 5$, $p_R = 10^{-4}$ and $p_M = 10^{-6}$. For (a) and (c) $\alpha = 0$, for (b) and (d) $\alpha = 0.01$ and $\sigma = 0.1$

The reason there are more data points towards the end of 1,000,000 generations in figure B.1 is due to this being the data points I use to determine if the simulations have reached stability in terms of $\bar{L}$.

## B.3 The closer $\bar{L}$ is to the boundaries the lower the variance of $L$

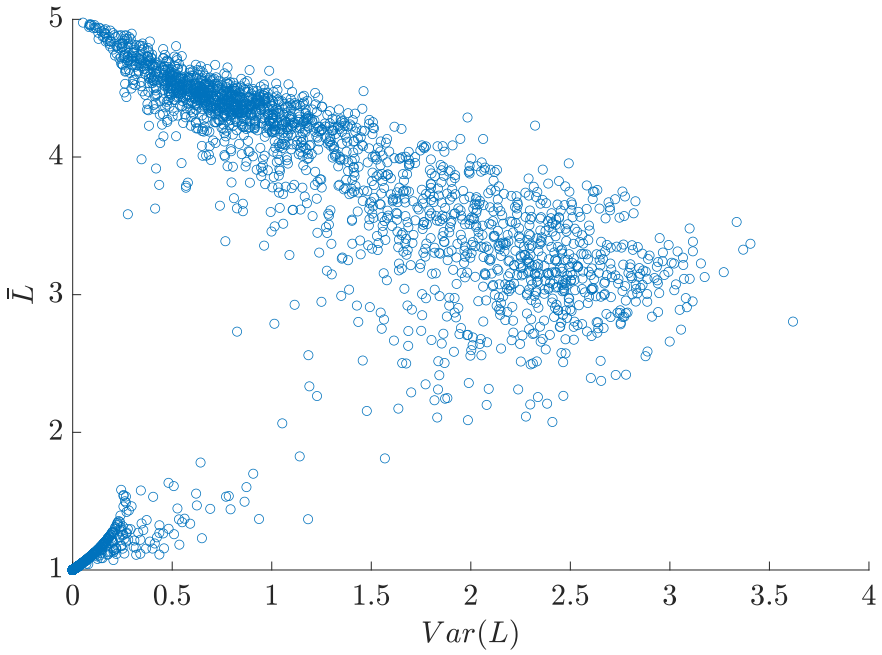



Figure B.2: **Highlighting the differences in outcome by comparing $\bar{L}$ and $Var(L)$ for varying degrees of CPS when using the max fitness rule.** The parameters used are as follows: N = 5000, Initial number of exons = 10, Initial number of Loci = 2, $U_L = 5$, $p_R = 10^{-4}$ and $p_M = 10^{-6}$. For the blue markers $\alpha = 0$, for the red marker $\alpha = 0.01$ and $\sigma = 0.1$

Here I have plotted all simulations from every scenario run in this work. I note that whenever $\bar{L}$ is close to the upper and lower boundary the lower $Var(L)$ is. This is expected as I have defined hard boundaries and if either short or long gene sequences are selected for then the system would tend to the edge of the boundaries defined. In terms of trying to replicate reality this is fine for the lower limit as there would be a lower limit in reality for an MHC gene. If an organism went below one gene copy (zero copied) then they have lost all means of that gene's functionality. However the upper limit is less likely

to be such a hard boundary. Here I have tested how a softer boundary affects $Var(L)$ as longer genes are being selected for.

I created a soft boundary that gradually decreases linearly the fitness of an individual when they pass the upper boundary of 5. This decrease will eventually hit 0 when the gene length is 10 or more.

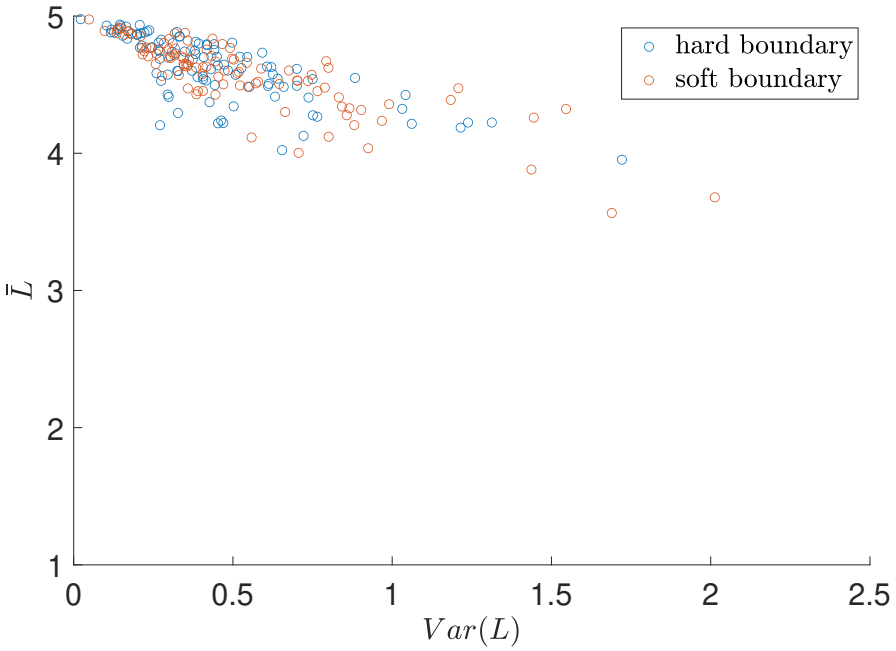
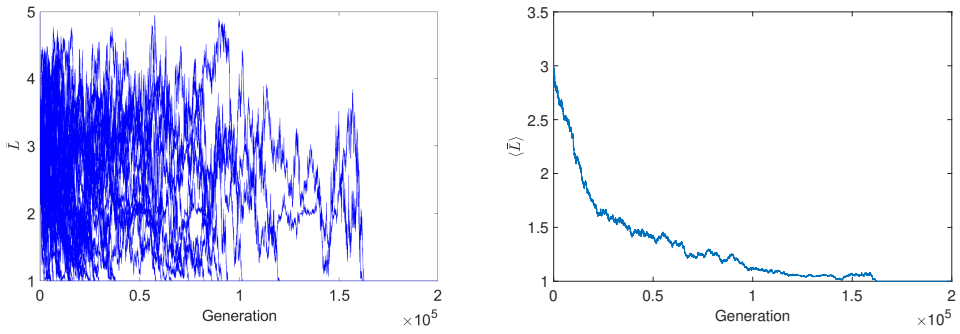



Figure B.3: **Mean and variance of MHC clusters when applying soft and hard boundary conditions.** The empty markers are the value of $\bar{L}$ and $Var(L)$ for each individual simulation. The parameters used are as follows: N = 5000, Initial number of exons = 10, Initial number of Loci = 2, $U_L = 5$, $p_R = 10^{-4}$, $p_M = 10^{-6}$, $\alpha = 0.01$ and $\sigma = 0.1$

In figure B.3 we see that the soft boundary results do not seem to differ from the hard boundary results. This was tested for the maximum fitness rule and using CPS due to that scenario selecting for longer gene clusters.

## B.4 Absorbing state when all recombination events occur between exons

Here I show how $\bar{L}$ changes when $\beta = 1$ and no selection rules are present but the boundary conditions on $L$.




(a) (b)

Figure B.4: **Mean lengths of MHC clusters ($\bar{L}$) in the absence of pathogen mediated selection.** Both (a) and (b) are of 50 simulations, (a) is a time series of $\bar{L}$ for each simulation. (b) is a time series of the average of $\bar{l}$. The parameters used are as follows: N = 5000, $p_R = 1$ and $\beta = 1$. For the blue lines $U_L = 50$, red lines $U_L = 100$, green lines $U_L = 200$.

In figure B.4 we see that before we reach generation 200,000 that every simulation has reached the absorbing state of every individual having an $L = 1$. This occurs due to when $\beta = 1$, when an individual has a chromosome with $L = 1$ and a chromosome with $L = n$ ($n > 1$), recombination between these two gene clusters can only produce gene clusters with $n$ or less gene sequences in them. So systems that have $\beta = 1$ are more drawn to $L = 1$ and when everyone having chromosomes of $L = 1$ is an absorbing state.

# Appendix C

# Chapter 4 supplementary material

## C.1 Table of Notation

| Notation | Definition |
|---|---|
| $G$ | Column vector of integers that represent the populations MHC alleles. |
| $N_t$ | The number of individuals in the population at generation $t$. |
| $K$ | The carrying capacity. |
| $L$ | The number of attributes alleles are represented by. |
| $a_i$ | The fitness value used for attribute $i$. |
| $r$ | The population increases by the proportion $r$ each generation. |
| $r$ | The population increases by the proportion $r$ each generation. |
| $p_I$ | Each generation a proportion $p_I$ of the population is infected. |
| $p_M$ | The probability a mutation event will |
| $P$ | Is the random number generated from a Kumaraswamy distribution which represents the pathogen test for that generation. |
| $f(P; a, b)$ | The probability density function of $P$. |
| $F(P; a, b)$ | The cumulative distribution function of $P$. |
| $a$ and $b$ | Parameters of the Kumaraswamy distribution, $a > 0$ and $b > 0$. |
| $S$ | Denotes an allele that has a high fitness value in one attribute (specialist). |
| $G_i$ | Denotes an allele that has a high fitness value in $i$ attributes (generalist). |

Table C.1: **Table of notations used in chapter 4**

## C.2 Proportion of simulations that had co-existing allele types

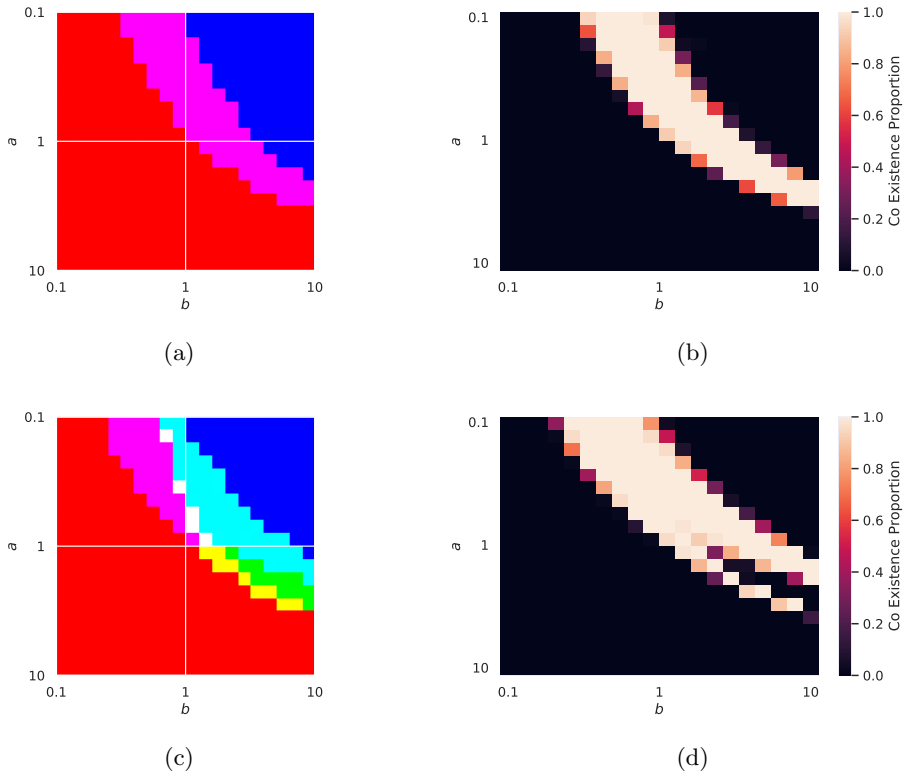



Figure C.1: **(a) Heat Map representing how much of the population had each allele type and (b) represents the proportion of simulations that have multiple allele types in each simulation.** Different colours represent the allele type/types that are predominant in the population. Red represents alleles of types $S$, green of type $G_2$ and blue of type $G_3$. The other colours represent populations with multiple allele types present (yellow $S$ and $G_2$, purple $S$ and $G_3$, turquoise $G_2$ and $G_3$ and white all allele types). (b) shows the proportion of simulations that had multiple allele types which had a frequency above 0.1. The parameters are as follows: $L = 3$ and $P_M = 10^{-5}$, $r = 1.1$, $P_I = 0.0909$, N = 5000 and generation simulated to 20,000.

In section 4.3.1 I show figures 4.6a and 4.7a which are also shown above in figures C.1a and C.1c. These figures show which allele types are present in the 50 simulations for the parameter space $a$ and $b$. It does not show if co-existence is actually occurring for allele types in a single simulation, it just shows that there are multiple allele types over the 50 simulations. As we are concerned with whether or not co existence is occurring between allele types, I have produced figure C.1b and C.1d which shows us the proportion of the 50 simulations which had multiple allele types over 0.1 allele frequency. As can be seen in figure C.1 the regions in figuresC.1b and C.1d coincide with the regions in figures C.1a and C.1c. This showcases when we observe multiple allele types over the 50 simulations we are seeing coexistence of allele types within each single simulation as well.

## C.3 No Extinction limit

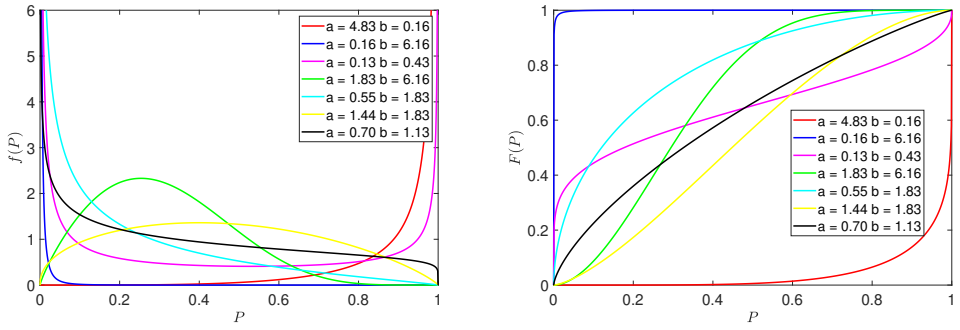
If

$$P_I < 1 - \frac{1}{r} \tag{C.1}$$

then no extinction can occur.

## C.4 Cumulative distribution of $P$

Here I have plotted the cumulative distribution of $P$ ($F(P)$) for the various probability density functions of $P$ ($f(P)$) that I have plotted throughout the result.




(a) (b)

Figure C.2: **Plots of the probability density function and cumulative distribution of $P$ for various values of $a$ and $b$.** (a) is a plot of the probability density function of $P$ (b) is a plot of the cumulative distribution of $P$. The colours match the scenarios for the varying allele promiscuity co existence for the case of $L = 3$.

Here is a table that shows the value of $F(P)$ for the same values of $a$ and $b$ in figure C.2a.

| $a$ | $b$ | $P$ | $F(P)$ | $1 - F(P)$ |
|---|---|---|---|---|
| 4.83 | 0.16 | 0.9 | 0.1385 | 0.8615 |
| 0.16 | 6.16 | 0.9 | 1.0000 | 0.0000 |
| 0.13 | 0.43 | 0.9 | 0.8424 | 0.1576 |
| 1.83 | 6.16 | 0.9 | 1.0000 | 0.0000 |
| 0.55 | 1.83 | 0.9 | 0.9949 | 0.0051 |
| 1.44 | 1.83 | 0.9 | 0.9725 | 0.0275 |
| 0.7 | 1.13 | 0.9 | 0.9497 | 0.0503 |

Table C.2: **Table that shows the value of $F(P)$ for varying values of $a$ and $b$.**

# Appendix D

# References


A. K. Abbas, A. H. Lichtman, and S. Pillai. *Cellular and molecular immunology E-book*. Elsevier Health Sciences, 2014.

A. Aguilar, G. Roemer, S. Debenham, M. Binns, D. Garcelon, and R. K. Wayne. High mhc diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences*, 101(10):3490–3494, 2004.

I. Alter, L. Gragert, S. Fingerson, M. Maiers, and Y. Louzoun. Hla class i haplotype diversity is consistent with selection for frequent existing haplotypes. *PLoS computational biology*, 13(8):e1005693, 2017.

A. Aslam, A. Mason, S. Zemenides, H. Chan, L. Nováková, P. Branny, A. Finn, H. Chapel, and G. S. Ogg. Rapid effector function of circulating cd4+ t cells specific for immunodominant regions of the conserved serine/threonine kinase found in streptococcus pneumoniae (stkp) in healthy adults. *FEMS Immunology & Medical Microbiology*, 60 (2):113–122, 2010.

A. Aslam, H. Chapel, and G. Ogg. Direct ex-vivo evaluation of pneumococcal specific t-cells in healthy adults. *PloS one*, 6(10):e25367, 2011.

M. Baake. Repeat distributions from unequal crossovers. *arXiv preprint arXiv:0803.1270*, 2008.

A. Balamurugan, S. K. Sharma, and N. K. Mehra. Human leukocyte antigen class i supertypes influence susceptibility and severity of tuberculosis. *Journal of Infectious Diseases*, 189(5):805–811, 2004.

G. K. Beauchamp, K. Yamazaki, J. Bard, and E. A. Boyse. Preweaning experience in the control of mating preferences by genes in the major histocompatibility complex of the mouse. *Behavior Genetics*, 18(4):537–547, 1988.

S. Ben Shachar, N. Barda, S. Manor, S. Israeli, N. Dagan, S. Carmi, R. Balicer, B. Zisser, and Y. Louzoun. Mhc haplotyping of sars-cov-2 patients: Hla subtypes are not associated with the presence and severity of covid-19 in the israeli population. *Journal of clinical immunology*, 41(6):1154–1161, 2021.

P. Bentkowski and J. Radwan. Evolution of major histocompatibility complex gene copy number. *PLOS Computational Biology*, 15(5):e1007015, 2019.

J. A. Borghans, J. B. Beltman, and R. J. De Boer. Mhc polymorphism under host-pathogen coevolution. *Immunogenetics*, 55(11):732–739, 2004.

W. Briles, N. Bumstead, D. Ewert, D. Gilmour, J. Gogusev, K. Hala, C. Koch, B. Longenecker, A. Nordskog, J. Pink, et al. Nomenclature for chicken major histocompatibility (b) complex. *Immunogenetics*, 15(5):441–447, 1982.

P. C. Bull, C. O. Buckee, S. Kyes, M. M. Kortok, V. Thathy, B. Guyah, J. A. Stoute, C. I. Newbold, and K. Marsh. Plasmodium falciparum antigenic variation. mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Molecular microbiology*, 68(6):1519–1534, 2008.

M. Carrington and S. J. O'Brien. The influence of hla genotype on aids. *Annual review of medicine*, 54(1):535–551, 2003.

P. E. Chappell, E. K. Meziane, M. Harrison, Ł. Magiera, C. Hermann, L. Mears, A. G. Wrobel, C. Durant, L. L. Nielsen, S. Buus, et al. Expression levels of mhc class i molecules are inversely correlated with promiscuity of peptide binding. *elife*, 4:e05345, 2015.

K. Y. Chen, J. Liu, and E. C. Ren. Structural and functional distinctiveness of hla-a2 allelic variants. *Immunologic research*, 53(1):182–190, 2012.

J. Cheng, K. Bendjama, K. Rittner, and B. Malone. Bertmhc: improved mhc–peptide class ii interaction prediction with transformer and multiple instance learning. *Bioinformatics*, 37(22):4172–4179, 2021.

V. J. Clark, S. E. Ptak, I. Tiemann, Y. Qian, G. Coop, A. C. Stone, M. Przeworski, N. Arnheim, and A. D. Rienzo. Combining sperm typing and linkage disequilibrium analyses reveals differences in selective pressures or recombination rates across human populations. *Genetics*, 175(2):795–804, 2007.

V. Davenport, T. Guthrie, J. Findlow, R. Borrow, N. A. Williams, and R. S. Heyderman. Evidence for naturally acquired t cell-mediated mucosal immunity to neisseria meningitidis. *The Journal of Immunology*, 171(8):4263–4270, 2003.

R. Daza-Vamenta, G. Glusman, L. Rowen, B. Guthrie, and D. E. Geraghty. Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Research*, 14(8):1501–1515, 2004.

R. J. De Boer, J. A. Borghans, M. van Boven, C. Keşmir, and F. J. Weissing. Heterozygote advantage fails to explain the high degree of polymorphism of the mhc. *Immunogenetics*, 55(11):725–731, 2004.

P. C. Doherty and R. M. Zinkernagel. Enhanced immunological surveillance in mice heterozygous at the h-2 gene complex. *Nature*, 256(5512):50–52, 1975.

S. J. Dunstan, N. T. Hue, B. Han, Z. Li, T. T. B. Tram, K. S. Sim, C. M. Parry, N. T. Chinh, H. Vinh, N. P. H. Lan, et al. Variation at hla-drb1 is associated with resistance to enteric fever. *Nature genetics*, 46(12):1333, 2014.

A. Eklund. The major histocompatibility complex and mating preferences in wild house mice (mus domesticus). *Behavioral Ecology*, 8(6):630–634, 1997.

A. Eklund, K. Egid, and J. L. Brown. The major histocompatibility complex and mating preferences of male mice. *Animal Behaviour*, 42(4):693–694, 1991.

S. A. Ellis and K. T. Ballingall. Cattle mhc: evolution in action? *Immunological reviews*, 167(1):159–168, 1999.

M. C. Enright and B. G. Spratt. A multilocus sequence typing scheme for streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology*, 144(11):3049–3060, 1998.

J. L. Freeman, G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith, M. E. Hurles, et al. Copy number variation: new insights in genome diversity. *Genome research*, 16(8):949–961, 2006.

X. Gao, T. R. O'Brien, T. M. Welzel, D. Marti, Y. Qi, J. J. Goedert, J. Phair, R. Pfeiffer, and M. Carrington. Hla-b alleles associate consistently with hiv heterosexual transmission, viral load and progression to aids, but not susceptibility to infection. *AIDS (London, England)*, 24(12):1835, 2010.

J. J. Gart. Alternative analyses of contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 164–179, 1966.

F. F. Gonzalez-Galarza, S. Christmas, D. Middleton, and A. R. Jones. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic acids research*, 39(suppl_1):D913–D919, 2010.

S. Gupta and A. V. Hill. Dynamic interactions in malaria: host heterogeneity meets parasite polymorphism. *Proc. R. Soc. Lond. B*, 261(1362):271–277, 1995.

J. A. Hammond, S. G. Marsh, J. Robinson, C. J. Davies, M. J. Stear, and S. A. Ellis. Cattle mhc nomenclature: is it possible to assign sequences to discrete class i genes? *Immunogenetics*, 64(6):475–480, 2012.

D. M. Hawley and R. C. Fleischer. Contrasting epidemic histories reveal pathogen-mediated balancing selection on class ii mhc diversity in a wild songbird. *PLoS One*, 7 (1):e30222, 2012.

P. W. Hedrick. Pathogen resistance and genetic variation at mhc loci. *Evolution*, 56(10): 1902–1908, 2002.

P. W. Hedrick and F. L. Black. Hla and mate selection: no evidence in south amerindians. *The American Journal of Human Genetics*, 61(3):505–511, 1997.

P. W. Hedrick and G. Thomson. Evidence for balancing selection at hla. *Genetics*, 104 (3):449–456, 1983.

T. Hertz, D. Nolan, I. James, M. John, S. Gaudieri, E. Phillips, J. C. Huang, G. Riadi, S. Mallal, and N. Jojic. Mapping the landscape of host-pathogen coevolution: Hla class i binding and its relationship with evolutionary conservation in human and viral proteins. *Journal of virology*, 85(3):1310–1321, 2011.

A. V. Hill, C. E. Allsopp, D. Kwiatkowski, N. M. Anstey, P. Twumasi, P. A. Rowe, S. Bennett, D. Brewster, A. J. McMichael, and B. M. Greenwood. Common west african hla antigens are associated with protection from severe malaria. *Nature*, 352 (6336):595–600, 1991.

R. Horton, L. Wilming, V. Rand, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. C. Talbot, M. W. Wright, et al. Gene map of the extended human mhc. *Nature Reviews Genetics*, 5(12):889–899, 2004.

W. Huang, J. G. Pilkington, and J. M. Pemberton. Patterns of mhc-dependent sexual selection in a free-living population of sheep. *Molecular Ecology*, 30(24):6733–6742, 2021.

A. L. Hughes and M. K. Hughes. Natural selection on the peptide-binding regions of major histocompatibility complex molecules. *Immunogenetics*, 42(4):233–243, 1995.

A. L. Hughes and M. Nei. Pattern of nucleotide substitution at major histocompatibility complex class i loci reveals overdominant selection. *Nature*, 335(6186):167–170, 1988.

Y. Ihara, K. Aoki, K. Tokunaga, K. Takahashi, and T. Juji. Hla and human mate choice: tests on japanese couples. *Anthropological Science*, 108(2):199–214, 2000.

P. T. Illing, J. P. Vivian, N. L. Dudek, L. Kostenko, Z. Chen, M. Bharadwaj, J. J. Miles, L. Kjer-Nielsen, S. Gras, N. A. Williamson, et al. Immune self-reactivity triggered by drug-modified hla-peptide repertoire. *Nature*, 486(7404):554–558, 2012.

P. Ilmonen, D. J. Penn, K. Damjanovich, L. Morrison, L. Ghotbi, and W. K. Potts. Major histocompatibility complex heterozygosity reduces fitness in experimentally infected mice. *Genetics*, 176(4):2501–2508, 2007.

K. J. Jeffery and C. R. Bangham. Do infectious diseases drive mhc diversity? *Microbes and infection*, 2(11):1335–1341, 2000.

L. Jiang, H. Yu, J. Li, J. Tang, Y. Guo, and F. Guo. Predicting mhc class i binder: existing approaches and a novel recurrent neural network solution. *Briefings in Bioinformatics*, 22(6):bbab216, 2021.

J. Jin, Z. Liu, A. Nasiri, Y. Cui, S.-Y. Louis, A. Zhang, Y. Zhao, and J. Hu. Deep learning pan-specific model for interpretable mhc-i peptide binding prediction with improved attention mechanism. *Proteins: Structure, Function, and Bioinformatics*, 89 (7):866–883, 2021.

J. J. Just. Genetic predisposition to hiv-1 infection and acquired immune deficiency virus syndrome: a review of the literature examining associations with hla. *Human immunology*, 44(3):156–169, 1995.

J. Kaufman. Generalists and specialists: a new view of how mhc class i molecules fight infectious pathogens. *Trends in immunology*, 39(5):367–379, 2018.

J. Kaufman. From chickens to humans: the importance of peptide repertoires for mhc class i alleles. *Frontiers in Immunology*, 11, 2020.

Y. Kawashima, K. Pfafferott, J. Frater, P. Matthews, R. Payne, M. Addo, H. Gatanaga, M. Fujiwara, A. Hachiya, H. Koizumi, et al. Adaptation of hiv-1 to human leukocyte antigen class i. *Nature*, 458(7238):641, 2009.

M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2011.

S. I. Khakoo, C. L. Thio, M. P. Martin, C. R. Brooks, X. Gao, J. Astemborski, J. Cheng, J. J. Goedert, D. Vlahov, M. Hilgartner, et al. Hla and nk cell inhibitory receptor genes in resolving hepatitis c virus infection. *Science*, 305(5685):872–874, 2004.

J. Klein, A. Sato, and N. Nikolaidis. Mhc, tsp, and the origin of species: from immunogenetics to evolutionary genetics. *Annu. Rev. Genet.*, 41:281–304, 2007.

A. Košmrlj, E. L. Read, Y. Qi, T. M. Allen, M. Altfeld, S. G. Deeks, F. Pereyra, M. Carrington, B. D. Walker, and A. K. Chakraborty. Effects of thymic selection of the t-cell repertoire on hla class i-associated control of hiv infection. *Nature*, 465(7296):350–354, 2010.

J. Krüger and F. Vogel. Population genetics of unequal crossing over. *Journal of Molecular Evolution*, 4(3):201–247, 1975.

Å. Langefors, J. Lohm, M. Grahn, Ø. Andersen, and T. v. Schantz. Association between major histocompatibility complex class iib alleles and resistance to aeromonas salmonicida in atlantic salmon. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1466):479–485, 2001.

R. Lewontin, L. Ginzburg, and S. Tuljapurkar. Heterosis as an explanation for large amounts of genic polymorphism. *Genetics*, 88(1):149–169, 1978.

A. E. Lobkovsky, L. Levi, Y. I. Wolf, M. Maiers, L. Gragert, I. Alter, Y. Louzoun, and E. V. Koonin. Multiplicative fitness, rapid haplotype discovery, and fitness decay explain evolution of human mhc. *Proceedings of the National Academy of Sciences*, 116 (28):14098–14104, 2019.

K. E. Lyke, M. A. Fernández-Viña, K. Cao, J. Hollenbach, D. Coulibaly, A. K. Kone, A. Guindo, L. A. Burdett, R. J. Hartzman, A. R. Wahl, et al. Association of hla alleles with plasmodium falciparum severity in malian children. *Tissue Antigens*, 77 (6):562–571, 2011.

K. S. MacDonald, K. R. Fowke, J. Kimani, V. A. Dunand, N. J. Nagelkerke, T. Blake Ball, J. Oyugi, E. Njagi, L. K. Gaur, R. Brunham, et al. Influence of hla supertypes on susceptibility and resistance to human immunodeficiency virus type 1 infection. *The Journal of infectious diseases*, 181(5):1581–1589, 2000.

A. MacPherson, S. P. Otto, and S. L. Nuismer. Keeping pace with the red queen: Identifying the genetic basis of susceptibility to infectious disease. *Genetics*, 208(2):779–789, 2018.

D. R. Madden, D. N. Garboczi, and D. C. Wiley. The antigenic identity of peptide-mhc complexes: a comparison of the conformations of five viral peptides presented by hla-a2. *Cell*, 75(4):693–708, 1993.

S. Magadum, U. Banerjee, P. Murugan, D. Gangapur, and R. Ravikesavan. Gene duplication as a major force in evolution. *Journal of genetics*, 92(1):155–161, 2013.

C. J. Manning, W. K. Potts, E. K. Wakeland, and D. A. Dewsbury. What's wrong with mhc mate choice experiments? In *Chemical Signals in Vertebrates 6*, pages 229–235. Springer, 1992.

M. P. Martin, X. Gao, J.-H. Lee, G. W. Nelson, R. Detels, J. J. Goedert, S. Buchbinder, K. Hoots, D. Vlahov, J. Trowsdale, et al. Epistatic interaction between kir3ds1 and hla-b delays the progression to aids. *Nature genetics*, 31(4):429, 2002.

T. Maruyama and M. Nei. Genetic variability maintained by mutation and overdominant selection in finite populations. *Genetics*, 98(2):441–459, 1981.

R. McBride, J. Cutting, L. Schierman, F. Strebel, and D. Watanabe. Mhc gene control of growth of avian sarcoma virus-induced tumours in chickens: a study on the role of virus strain. *International Journal of Immunogenetics*, 8(3):207–214, 1981.

P. J. McLaren, C. Coulonges, I. Bartha, T. L. Lenz, A. J. Deutsch, A. Bashirova, S. Buchbinder, M. N. Carrington, A. Cossarizza, J. Dalmau, et al. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of hiv-1 virus load. *Proceedings of the National Academy of Sciences*, 112(47):14658–14663, 2015.

B. Mordmüller, G. Surat, H. Lagler, S. Chakravarty, A. S. Ishizuka, A. Lalremruata, M. Gmeiner, J. J. Campo, M. Esen, A. J. Ruben, et al. Sterile protection against human malaria by chemoattenuated pfspz vaccine. *Nature*, 542(7642):445, 2017.

S. Mukherjee and N. Chandra. Grouping of large populations into few ctl immune 'response-types' from influenza h1n1 genome analysis. *Clinical & translational immunology*, 3(8):e24, 2014.

C. Ober, L. R. Weitkamp, N. Cox, H. Dytch, D. Kostyu, and S. Elias. Hla and mate choice in humans. *The American Journal of Human Genetics*, 61(3):497–504, 1997.

A. Oliveira-Cortez, A. Melo, V. Chaves, A. Condino-Neto, and P. Camargos. Do hla class ii genes protect against pulmonary tuberculosis? a systematic review and meta-analysis. *European Journal of Clinical Microbiology & Infectious Diseases*, 35(10): 1567–1580, 2016.

N. Otting, C. M. Heijmans, R. C. Noort, N. G. De Groot, G. G. Doxiadis, J. J. Van Rood, D. I. Watkins, and R. E. Bontrop. Unparalleled complexity of the mhc class i region in rhesus macaques. *Proceedings of the National Academy of Sciences*, 102(5):1626–1631, 2005.

N. Otting, A. J. de Vos-Rouweler, C. M. Heijmans, N. G. de Groot, G. G. Doxiadis, and R. E. Bontrop. Mhc class ia region diversity and polymorphism in macaque species. *Immunogenetics*, 59(5):367–375, 2007.

S. Paul, D. Weiskopf, M. A. Angelo, J. Sidney, B. Peters, and A. Sette. Hla class i alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *The Journal of Immunology*, 191(12):5831–5839, 2013.

R. Payne, M. Muenchhoff, J. Mann, H. E. Roberts, P. Matthews, E. Adland, A. Hempenstall, K.-H. Huang, M. Brockman, Z. Brumme, et al. Impact of hla-driven hiv adaptation on virulence in populations of high hiv seroprevalence. *Proceedings of the National Academy of Sciences*, 111(50):E5393–E5400, 2014.

H. Peltola. Meningococcal disease: still with us. *Reviews of infectious diseases*, 5(1): 71–91, 1983.

F. Peng, K. M. Ballare, S. Hollis Woodard, S. den Haan, and D. I. Bolnick. What evolutionary processes maintain mhc ii diversity within and among populations of stickleback? *Molecular ecology*, 30(7):1659–1671, 2021.

B. S. Penman, B. Ashby, C. O. Buckee, and S. Gupta. Pathogen selection drives nonoverlapping associations between hla loci. *Proceedings of the National Academy of Sciences*, 110(48):19645–19650, 2013.

D. J. Penn. The scent of genetic compatibility: sexual selection and the major histocompatibility complex. *Ethology*, 108(1):1–21, 2002.

D. J. Penn and W. K. Potts. The evolution of mating preferences and major histocompatibility complex genes. *The American Naturalist*, 153(2):145–164, 1999.

F. Pierini and T. L. Lenz. Divergent allele advantage at human mhc genes: signatures of past and ongoing selection. *Molecular biology and evolution*, 35(9):2145–2158, 2018.

W. K. Potts, C. J. Manning, and E. K. Wakeland. Mating patterns in seminatural populations of mice influenced by mhc genotype. *Nature*, 352(6336):619–621, 1991.

F. Prugnolle, A. Manica, M. Charpentier, J. F. Guégan, V. Guernier, and F. Balloux. Pathogen-driven selection and worldwide hla class i diversity. *Current biology*, 15(11): 1022–1027, 2005.

O. Redner and M. Baake. Unequal crossover dynamics in discrete and continuous time. *Journal of mathematical biology*, 49(2):201–226, 2004.

J. H. Relethford. *Human population genetics*, volume 7. John Wiley & Sons, 2012.

J. Robinson, J. A. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, and S. G. Marsh. The ipd and imgt/hla database: allele variant databases. *Nucleic acids research*, 43(D1): D423–D431, 2014.

J. Robinson, D. J. Barker, X. Georgiou, M. A. Cooper, P. Flicek, and S. G. Marsh. Ipd-imgt/hla database. *Nucleic acids research*, 48(D1):D948–D955, 2020.

M. Salie, L. van der Merwe, M. Möller, M. Daya, G. D. van der Spuy, P. D. van Helden, M. P. Martin, X.-j. Gao, R. M. Warren, M. Carrington, et al. Associations between human leukocyte antigen class i variants and the mycobacterium tuberculosis subtypes causing disease. *The Journal of infectious diseases*, 209(2):216–223, 2013.

N. Sambaturu, S. Mukherjee, M. López-García, C. Molina-París, G. I. Menon, and N. Chandra. Role of genetic heterogeneity in determining the epidemiological severity of h1n1 influenza. *PLoS computational biology*, 14(3):e1006069, 2018.

N. Schubert, H. J. Nichols, and J. C. Winternitz. How can the mhc mediate social odor via the microbiota community? a deep dive into mechanisms. *Behavioral Ecology*, 32 (3):359–373, 2021.

J. C. Schwartz and J. A. Hammond. The assembly and characterisation of two structurally distinct cattle mhc class i haplotypes point to the mechanisms driving diversity. *Immunogenetics*, 67(9):539–544, 2015.

A. Sette and J. Sidney. Nine major hla class i supertypes account for the vast preponderance of hla-a and-b polymorphism. *Immunogenetics*, 50(3):201–212, 1999.

M. Shpak and K. Atteson. A survey of unequal crossover systems and their mathematical properties. *Bulletin of mathematical biology*, 64(4):703–746, 2002.

H. V. Siddle, J. Marzec, Y. Cheng, M. Jones, and K. Belov. Mhc gene copy number variation in tasmanian devils: implications for the spread of a contagious cancer. *Proceedings of the Royal Society B: Biological Sciences*, 277(1690):2001–2006, 2010.

J. Sidney, B. Peters, N. Frahm, C. Brander, and A. Sette. Hla class i supertypes: a revised and updated classification. *BMC immunology*, 9(1):1, 2008.

M. Siljestam and C. Rueffler. Heterozygote advantage can explain the extraordinary diversity of immune genes. *bioRxiv*, page 347344, 2019.

R. Slade and H. McCallum. Overdominant vs. frequency-dependent selection at mhc loci. *Genetics*, 132(3):861, 1992.

L. G. Spurgin and D. S. Richardson. How pathogens drive genetic diversity: Mhc, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences*, 277(1684):979–988, 2010.

M. Steinmetz, D. Stephan, and K. F. Lindahl. Gene organization and recombinational hotspots in the murine major histocompatibility complex. *Cell*, 44(6):895–904, 1986.

R. J. Stoffels and H. G. Spencer. An asymmetric model of heterozygote advantage at major histocompatibility complex genes: degenerate pathogen recognition and intersection advantage. *Genetics*, 178(3):1473–1489, 2008.

P. M. Stuart. Major histocompatibility complex (mhc): Mouse. pages 1–7, 2015. https://doi.org/10.1002/9780470015902.a0000921.pub4. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000921.pub4`.

G. Sveinbjornsson, D. F. Gudbjartsson, B. V. Halldorsson, K. G. Kristinsson, M. Gottfredsson, J. C. Barrett, L. J. Gudmundsson, K. Blondal, A. Gylfason, S. A. Gudjonsson, et al. Hla class ii sequence variants influence tuberculosis risk in populations of european ancestry. *Nature genetics*, 48(3):318, 2016.

N. Takahata. A mathematical study on the distribution of the number of repeated genes per chromosome. *Genetics Research*, 38(1):97–102, 1981.

N. Takahata and M. Nei. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, 124 (4):967–978, 1990.

M. Tanaka-Matsuda, A. Ando, C. Rogel-Gaillard, P. Chardon, and H. Uenishi. Difference in number of loci of swine leukocyte antigen classical class i genes among haplotypes. *Genomics*, 93(3):261–273, 2009.

M. R. Thursz, D. Kwiatkowski, C. E. Allsopp, B. M. Greenwood, H. C. Thomas, and A. V. Hill. Association between an mhc class ii allele and clearance of hepatitis b virus in the gambia. *New England Journal of Medicine*, 332(16):1065–1069, 1995.

C. Tian, B. S. Hromatka, A. K. Kiefer, N. Eriksson, S. M. Noble, J. Y. Tung, and D. A. Hinds. Genome-wide association and hla region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature communications*, 8(1):599, 2017.

L. Toyo-Oka, S. Mahasirimongkol, H. Yanai, T. Mushiroda, S. Wattanapokayakit, N. Wichukchinda, N. Yamada, N. Smittipat, T. Juthayothin, P. Palittapongarnpim, et al. Strain-based hla association analysis identified hla-drb1* 09: 01 associated with modern strain tuberculosis. *Hla*, 90(3):149–156, 2017.

E. Trachtenberg, B. Korber, C. Sollars, T. B. Kepler, P. T. Hraber, E. Hayes, R. Funkhouser, M. Fugate, J. Theiler, Y. S. Hsu, et al. Advantage of rare hla supertype in hiv disease progression. *Nature medicine*, 9(7):928, 2003.

H.-J. Wallny, D. Avila, L. G. Hunt, T. J. Powell, P. Riegert, J. Salomonsen, K. Skjødt, O. Vainio, F. Vilbois, M. V. Wiles, et al. Peptide motifs of the single dominantly

expressed class i molecule explain the striking mhc-determined response to rous sarcoma virus in chickens. *Proceedings of the National Academy of Sciences*, 103(5):1434–1439, 2006.

C. Wedekind and S. Füri. Body odour preferences in men and women: do they aim for specific mhc combinations or simply heterozygosity? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1387):1471–1479, 1997.

C. Wedekind, T. Seebeck, F. Bettens, and A. J. Paepke. Mhc-dependent mate preferences in humans. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 260(1359):245–249, 1995.

E. Wiertz, A. Delvig, E. Donders, H. Brugghe, L. Van Unen, H. Timmermans, M. Achtman, P. Hoogerhout, and J. Poolman. T-cell responses to outer membrane proteins of neisseria meningitidis: comparative study of the opa, opc, and pora proteins. *Infection and immunity*, 64(1):298–304, 1996.

R. W. Wiseman, J. A. Karl, P. S. Bohn, F. A. Nimityongskul, G. J. Starrett, and D. H. O'Connor. Haplessly hoping: macaque major histocompatibility complex made easy. *ILAR journal*, 54(2):196–210, 2013.

B. Woolf. On estimating the relation between blood group and disease. *Annals of human genetics*, 19(4):251–253, 1955.

E. E. Wroblewski, P. J. Norman, L. A. Guethlein, R. S. Rudicell, M. A. Ramirez, Y. Li, B. H. Hahn, A. E. Pusey, and P. Parham. Signature patterns of mhc diversity in three gombe communities of wild chimpanzees reflect fitness in reproduction and immune defense against sivcpz. *PLoS Biology*, 13(5):e1002144, 2015.

K. Yamazaki, E. Boyse, V. Mike, H. Thaler, B. Mathieson, J. Abbott, J. Boyse, Z. Zayas, and L. Thomas. Control of mating preferences in mice by genes in the major histocompatibility complex. *The Journal of experimental medicine*, 144(5):1324–1335, 1976.

K. Yamazaki, G. K. Beauchamp, D. Kupniewski, J. Bard, L. Thomas, and E. Boyse. Familial imprinting determines h-2 selective mating preferences. *Science*, 240(4857): 1331–1332, 1988.

S.-H. Yim, S.-H. Jung, B. Chung, and Y.-J. Chung. Clinical implications of copy number variations in autoimmune disorders. *The Korean journal of internal medicine*, 30(3): 294, 2015.

J. Zhang, Y. Chen, J. Qi, F. Gao, Y. Liu, J. Liu, X. Zhou, J. Kaufman, C. Xia, and G. F. Gao. Narrow groove and restricted anchors of mhc class i molecule bf2* 0401 plus peptide transporter restriction can explain disease susceptibility of b4 chickens. *The Journal of Immunology*, 189(9):4478–4487, 2012.