# Supporting Information Appendix

## Pathogen selection drives nonoverlapping associations between HLA loci.

Bridget S. Penman[a], Ben Ashby[a], Caroline O. Buckee[b] and Sunetra Gupta[a,1]

**Author affiliations:**

a. Department of Zoology, University of Oxford, South Parks Road, Oxford, OX13PS, UK

b. Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, USA, 02115.

**Corresponding author:**

1. Sunetra Gupta, Department of Zoology, University of Oxford, South Parks Road, Oxford, OX13PS UK

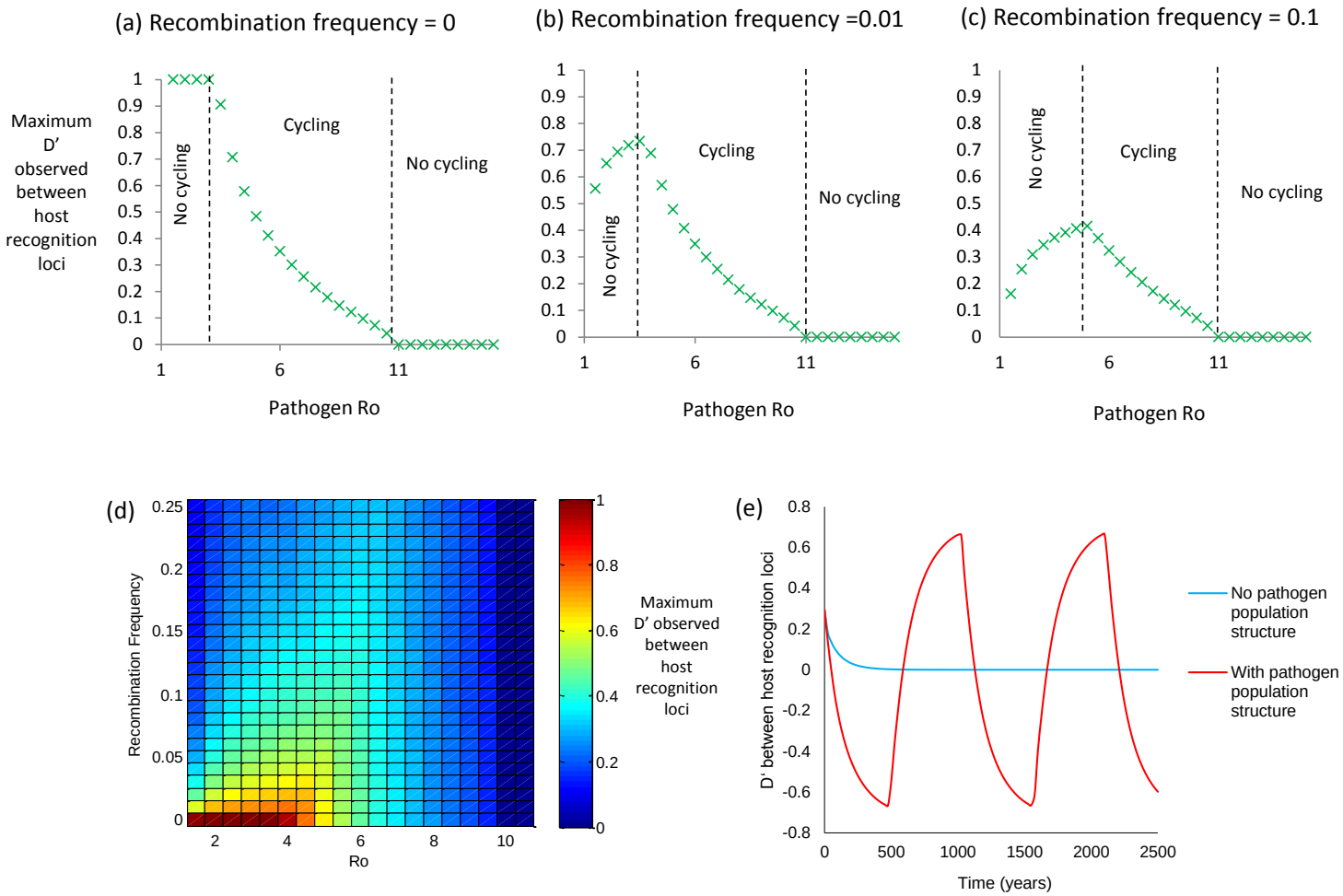
sunetra.gupta@zoo.ox.ac.uk

This appendix contains the following:

# 1. A sensitivity analysis of the deterministic HLA model

As noted in the main text, cyclical and stable structuring of both pathogen and host emerge within our framework as a result of immunological feedbacks. The nature of this structuring, however, is sensitive to the basic reproductive number of the pathogen ($R_o$) and the recombination frequency between host recognition loci. Figure S1 explores these effects, using a standard measure of linkage disequilibrium (Lewontin's D' [17]) to measure the degree of structuring that is generated between the host loci.

**Figure S1: The effects of varying $R_o$ and recombination frequency on host genetic structuring.** Panels (a-c) illustrate the effects of varying $R_o$ and the recombination frequency on the maximum D' observed between host loci. In these panels, host mortality rate ($\mu_1$)=0.05 ; mutation rate (m) =0.00001; pathogen mortality rate ($\mu_2$)=5 and recovery rate ($\sigma$) =10. The heatmap in panel (d) illustrates the maximum D' possible for different levels of $R_o$ combined with different recombination frequencies, when host mortality rate ($\mu_1$)=0.05; mutation rate (m) =0.00001; pathogen mortality rate ($\mu_2$)=5 and host recovery rate ($\sigma$) =7. The time series in panel (e) compares the change in D' over time for a host where the pathogen population is allowed to become structured (as described in the main text), with a host whose pathogen population is forced to remain unstructured (i.e. all possible pathogen strains are always present). For this panel, host mortality rate ($\mu_1$)=0.05; mutation rate (m) =0.0001; pathogen mortality rate ($\mu_2$)=4; host recovery rate ($\sigma$) =8; recombination frequency (r)=0.01 and $R_o$ =4.75.

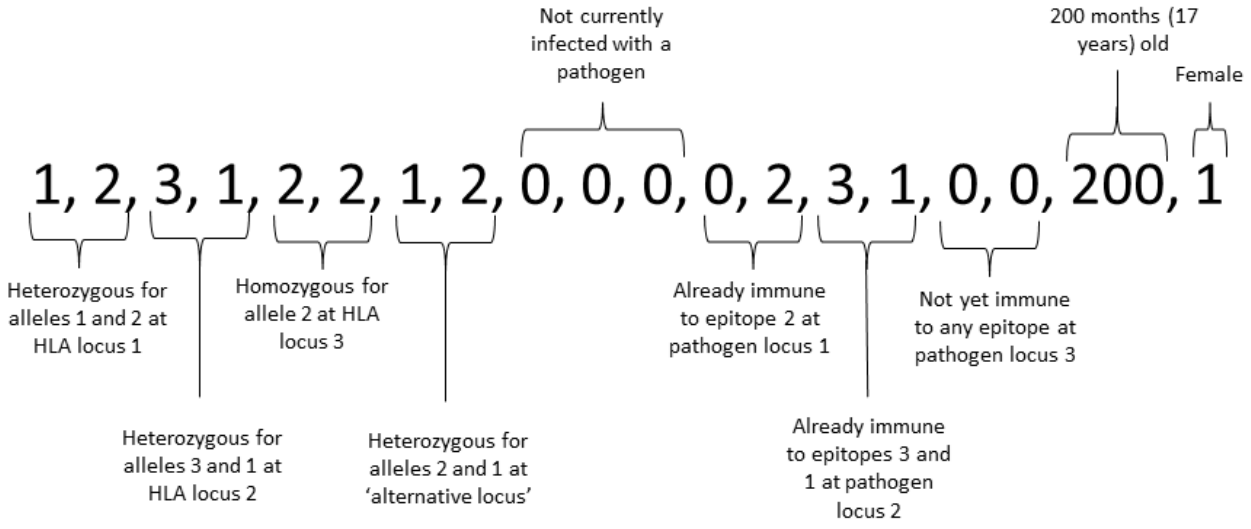


## 2. The stochastic HLA model and its behaviour

### 2.1 The stochastic model

We represent each member of the host population by a 19 element identifier code ($H_1$- $H_{19}$), recording: (i) their genotype at up to 3 different HLA loci ($H_1$- $H_6$); (ii) their genotype at an alternative locus, linked to the HLAs, that can be under non-HLA forms of selection ($H_7$-$H_8$) ; (iii) the genotype of up to one pathogen infecting them ($H_9$- $H_{11}$); (iv) whether or not they have generated a memory immune response against pathogen epitopes they have already been exposed to ($H_{12}$- $H_{17}$); (v) their age ($H_{18}$) and finally (vi) their gender ($H_{19}$).

A typical host identifier might appear as follows:




At each of the loci making up the host genotype, the first element represents the maternally derived allele and the second element represents the paternally derived allele. We are thus able to keep track of haplotypes.

As in the deterministic model described in the main text, the variants found at a particular HLA locus specifically target the variants of a particular pathogen epitope locus. This stochastic framework allows us to define any number of variants at each pathogen locus, and equivalent recognition alleles in the host. Section 2.3 of this document illustrates the 3 allele, 3 locus case.

At every time step of the simulation, the following events occur:

- Individuals age by one month.
- Uninfected individuals become infected with strain *jkl*, with probability $p_{jkl}$

if $H_{12,13} \neq j$ , $H_{14,15} \neq k$ and $H_{16,17} \neq l$ $$p_{jkl} = \frac{\sum H_{x\ldots z}}{\sum H_{x\ldots z}} . \theta$$

otherwise $$p_{jkl} = 0$$

where $H_{x...z}$ = any host within the population and $\theta$ controls the likelihood of transmission per infection. $\theta$ is directly equivalent to the transmission parameter β in the deterministic model.

- For every infection that occurs, mutation of a randomly chosen pathogen epitope (to any of the alleles that are possible at that locus) occurs with probability *m*.

- Infected individuals recover with probability Ω. The average length of infection within this system is therefore $\frac{1}{\Omega}$ months.

- Upon recovery from infection with *jkl*, if $H_1 = j$ then $H_{12} = j$ ; if $H_2 = j$ then $H_{13} = j$; if $H_3 = k$ then $H_{14} = k$ ; if $H_4 = k$ then $H_{15} = k$; if $H_5 = l$ then $H_{16} = l$; if $H_6 = l$ then $H_{17} = l$.

- Infected individuals who can recognise none of the pathogen's epitopes (i.e. for whom $H_{10,11} \neq j$ , $H_{12,13} \neq k$ and $H_{14,15} \neq l$) die with probability ϖ. Upon death, individuals are removed from the population.

- If we are imposing selection at the 'alternative locus' as a substitute for HLA interacting pathogen selection (see figure 5 in the main text), a maximum of 2 out of 4 possible alleles that may occupy the alternative locus are designated 'favoured' at any one time. Every time step, there is a probability *k* that the identity of an allele occupying one of the 2 'favoured' slots will change. If a change occurs, there is an equal chance (1/5) that the newly chosen 'favoured' allele will be any one of the alleles that can exist at the locus, or 'null' – i.e. no favoured allele at that time. Any individual who lacks a favoured allele at their alternative locus has a probability *d* of dying within a given time step.

- Females over the age of 15 years (180 months) reproduce with a randomly chosen male partner (also over the age of 15 years), with probability *b*. A new host (aged 0 months, with a genotype generated from a combination of maternal and paternal HLA haplotypes) is added to the population. Intra-haplotypic recombination is assumed to occur between the HLA loci, independently in both parents, with probability *r*. If the population is already at carrying capacity (*C*), the new host displaces a randomly chosen existing host.

- *Q* new individuals, with randomly generated genotypes, are introduced to the population with probability *α*; replacing *Q* existing individuals in the population. This step simulates gene flow between the simulated population and the wider world.

- Random host death occurs with probability ϕ.

- All hosts over the age of 40 years (480 months) are removed from the population.

A programme to perform these operations was written in C. A Mex file was created so that the model could be called from within Matlab, version 7.10.0 (R2010b).

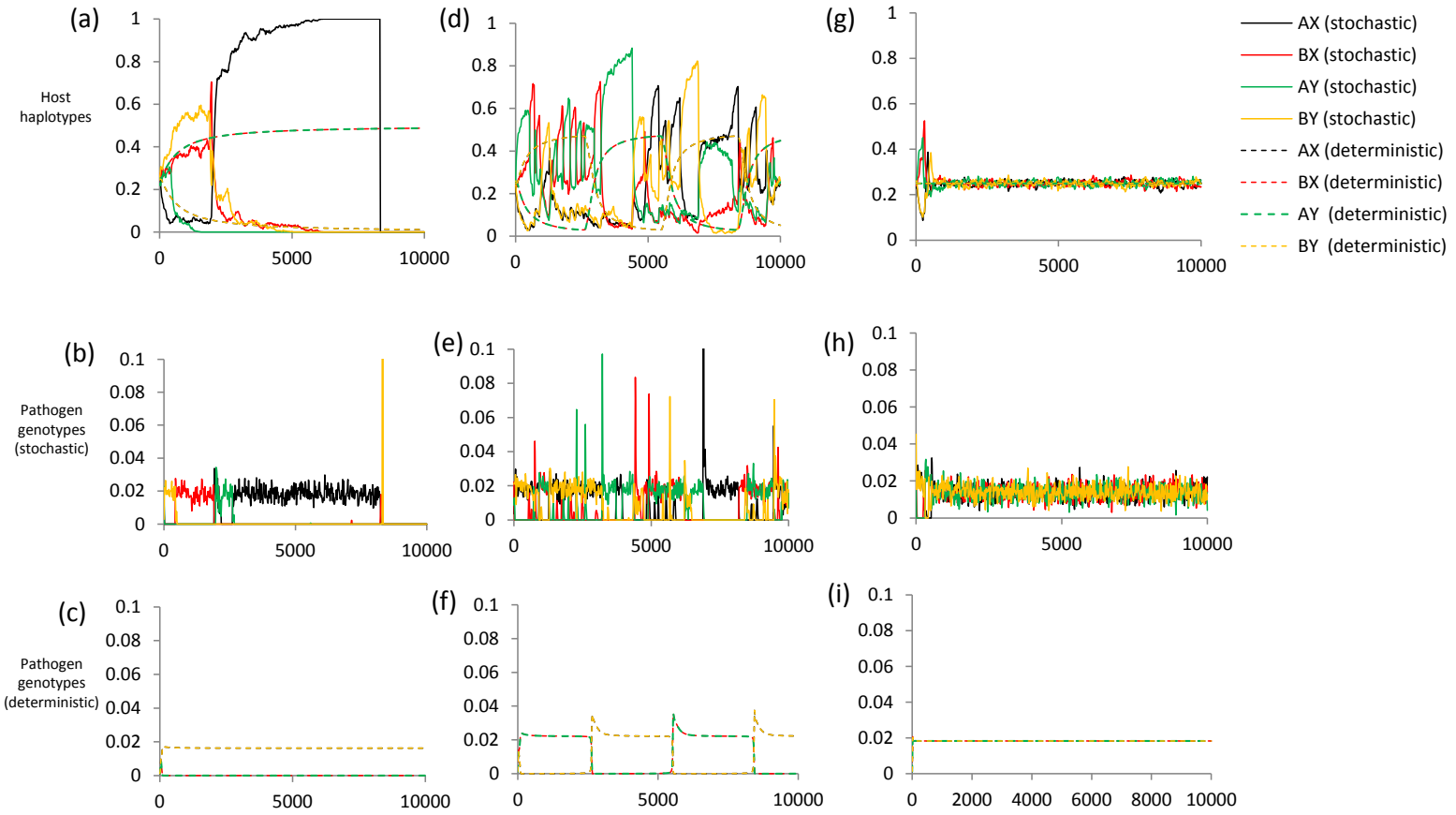
The initial conditions for each of our stochastic simulations were either (i) a population containing entirely randomly generated host and pathogen genotypes, or (ii) a population in which a certain proportion of host haplotypes were a specific 'founder' haplotype, and the rest were randomly generated. In these latter simulations, the *initial* pathogen population was such that its epitopes could definitely be recognised by the 'founder' haplotype. This condition was intended to avoid the host population being wiped out by a pathogen within a very short space of time, but the pathogen could quickly mutate to adopt alternative

structures. For figure 5 of the main text, the initial frequency of the founder haplotype was 0.505 in all simulations.

## 2.2 Equivalence between the stochastic and deterministic HLA models

When we consider a 2 locus, 2 allele recognition system and only apply HLA-interacting pathogen selection (i.e. $d$=0; $\varpi$>0), the stochastic model behaves equivalently to the deterministic model described in the main text.

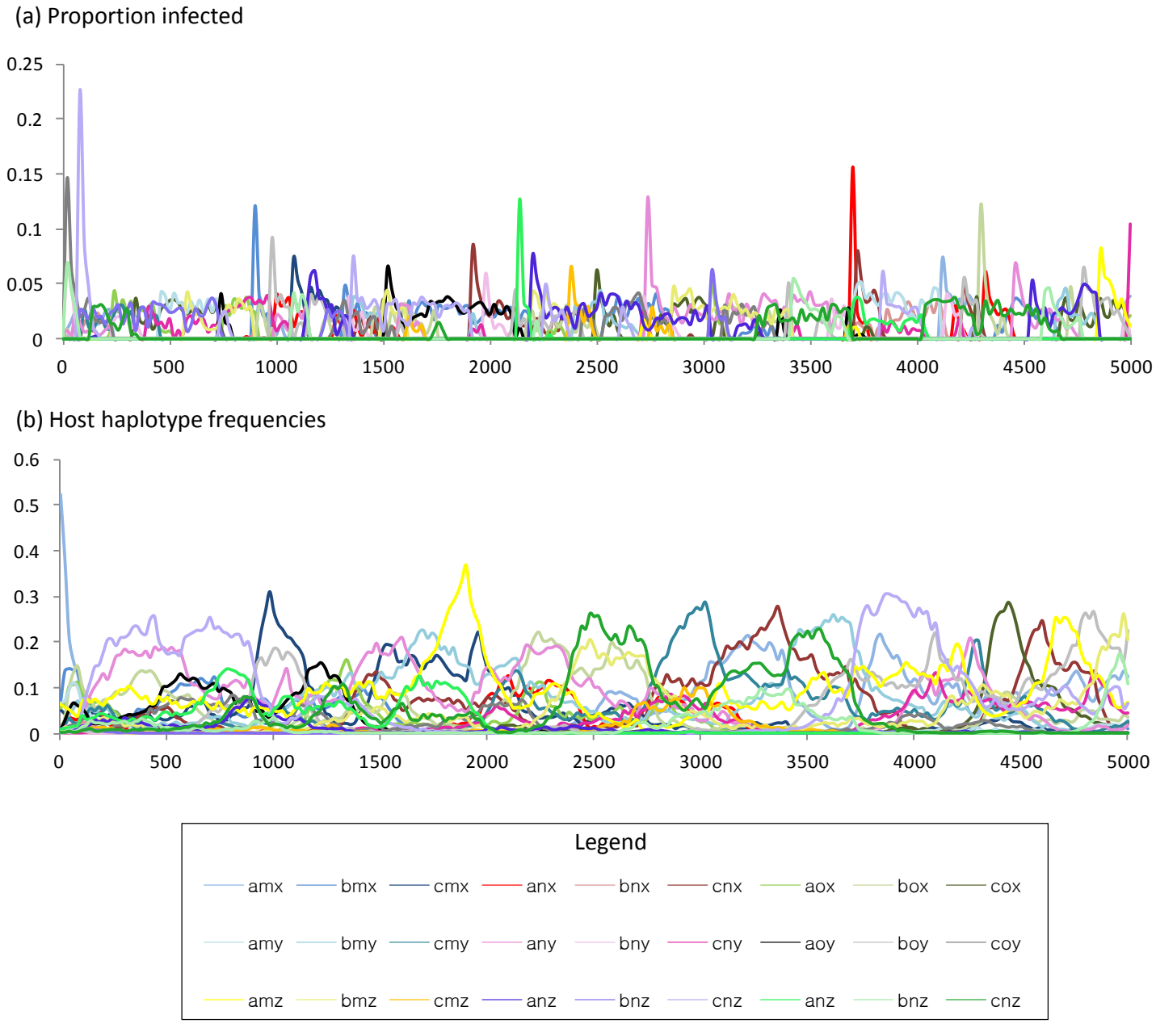
**Figure S2: Comparing the stochastic and deterministic models, in a 2 locus, 2 allele system.** The behaviour of the system is illustrated under 3 different sets of conditions: those which lead to permanent genetic structuring in the deterministic model (panels a-c); those which lead to cyclical behaviour in the deterministic model (panels d-f), and those which lead to no structuring in the deterministic model (panels g-i). For the deterministic model, $\mu_1$ = 0.04 years$^{-1}$; $\mu_2$ = 0.5 years$^{-1}$; r= 0; m= 0.0001; σ= 1.2 years$^{-1}$; β= 2.5 in (a-c); 4 in (d-f) and 10 in (g-i). For the stochastic simulations, ϕ = 0.0035; b =0.03; Q =0; Ω= 0.1; ϖ= 0.04; m= 0.0001; r = 0; C=1500; α = 0; d=0; k=0; $\theta$ = 2.5 in (a-c); 4 in (d-f) and 10 in (g-i). In all panels, the simulated populations contained randomly generated host and pathogen genotypes at time =0.



### 2.3 Extension to a 3 locus, 3 allele system

The stochastic framework allows us to extend the principles outlined in figure 1 of the main text to a system of 3 pathogen epitope loci with 3 alleles at each (*a*,*b*,*c*; *m*,*n*,*o* and *x*,*y*,*z*), and 3 corresponding host recognition loci (A, B, C; M, N, O and X, Y, Z). The cyclical behaviour reported in the 2 locus, 2 allele model can still be observed in this higher-dimension system (figure S3).

**Figure S3: Pathogen selection and host haplotypes structuring in a 3 locus, 3 allele stochastic system.** Panel (a) illustrates the proportion of the population infected with any one of 27 pathogen strains made up of 3 epitopic loci, and panel (b) illustrates the behaviour of host haplotypes in the same population. Equivalent colours have been used for pathogen genotypes and host haplotypes, thus the colour indicated for pathogen '*anx*' also refers to host haplotype 'ANX'. In all panels, b =0.05; *m*=0.0003; *r* =0.005; ϕ=0.0035; Ω = 0.1; $\alpha$ =0.0001 *C*=1500; *Q* = 5; $\theta$ = 1.5; ϖ = 0.01. The simulated population began with a 'founder' host haplotype of 'AMX' at a frequency of 0.52, whilst all other host haplotypes at time=0 were randomly generated.



## 3. Supplementary Tables

**Table S1: Pathogen strains**

| Pathotype subscript (*p*) | Epitopes |
|---|---|
| 1 | *ax* |
| 2 | *bx* |
| 3 | *ay* |
| 4 | *by* |

**Table S2: Host haplotypes**

| Haplotype subscript (*h*) | Haplotype |
|---|---|
| 1 | AX |
| 2 | BX |
| 3 | AY |
| 4 | BY |

**Table S3: Host genotypes**

| Genotype superscript (*i*) | Genotype |
|---|---|
| 1 | AXAX |
| 2 | BXBX |
| 3 | AYAY |
| 4 | BYBY |
| 5 | AXBY |
| 6 | BXAY |
| 7 | AXBX |
| 8 | AXAY |
| 9 | BXBY |
| 10 | AYBY |

**Table S4: Possible sets of epitopes a host could be immune to**

| *j* | $E_j$ |
|---|---|
| 0 | None |
| 1 | a |
| 2 | b |
| 3 | x |
| 4 | y |
| 5 | a, x |
| 6 | b, x |
| 7 | a, y |
| 8 | b, y |
| 9 | a, x, y |
| 10 | b, x, y |
| 11 | a, b, x |
| 12 | a, b, y |
| 13 | a, b, x, y |
| 14 | a, b |
| 15 | x, y |

**Table S5 Host genotype identifiers ($c_{1-5}$) for calculating the frequency of each haplotype**

| **Haplotype subscript (*h*)** | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ |
|---|---|---|---|---|---|
| 1 | 1 | 7 | 8 | 5 | 6 |
| 2 | 2 | 7 | 9 | 6 | 5 |
| 3 | 3 | 8 | 10 | 6 | 5 |
| 4 | 4 | 9 | 10 | 5 | 6 |