Title: Detecting signatures of past pathogen selection on human HLA loci: are there needles in the haystack?

Authors: Bridget S Penman[1] and Sunetra Gupta[2]

1. School of Life Sciences, University of Warwick, Coventry CV4 7AL

2. Department of Zoology, University of Oxford, Oxford, OX1 3PS

Running title: Understanding HLA evolution under multi-pathogen selection

Address for correspondence: Bridget Penman, School of Life Sciences, University of Warwick, Coventry CV4 7AL, Tel: 024 761 50249, Email: b.penman@warwick.ac.uk

Summary:

Human Leukocyte Antigens (HLAs) are responsible for the display of peptide fragments for recognition by T cell receptors. The gene family encoding them is thus integral to human adaptive immunity, and likely to be under strong pathogen selection. Despite this, it has proved difficult to demonstrate specific examples of pathogen-HLA coevolution. Selection from multiple pathogens simultaneously could explain why the evolutionary signatures of particular pathogens on HLAs have proved elusive. Here, we present an individual-based model of HLA evolution in the presence of two mortality-causing pathogens. We demonstrate that it is likely that individual pathogen species causing high mortality have left recognizable signatures on the HLA genomic region, despite more than one pathogen being present. Such signatures are likely to exist at the whole-population level, and involve haplotypic combinations of HLA genes rather than single loci.

Key words: Human Leukocyte Antigen (HLA), Host pathogen coevolution, Pathogen selection, Human evolution

Introduction

JBS Haldane's 1949 paper *Disease and Evolution* (Haldane 1949) presaged human host-pathogen coevolution as a field of study (Lederberg 1999). Haldane noted that the "surprising biochemical diversity" exhibited by mammalian and avian species is likely a consequence of selection from pathogens, with: "a particular race of bacteria or virus being adapted to individuals of a certain range of biochemical constitution". By way of example, Haldane pointed out that different human blood group antigens may have determined the susceptibility of our ancestors to particular strains of bacteria. At the conference where he presented this work, Haldane also introduced the hypothesis that mutations responsible for heritable human blood disorders, such as the thalassaemias, had spread in certain populations due to the protection they afforded their carriers against death from malaria. Few today would dispute any of the arguments in *Disease and Evolution*, and the malaria hypothesis has gone on to be confirmed as an example of selection from a specific pathogen leaving a detectable signature on the human genome (Allison 1954; Siniscalco et al. 1961; Flint et al. 1986; Hill et al. 1991).

While the list of malaria resistance loci continues to grow (Kwiatkowski 2005; Band et al. 2015), examples of genetic changes as a consequence of selection from any other known human pathogen are few. In a review of the host genetics of human infection (Hill 2006), Adrian Hill highlighted 6 human genes known to have a particularly strong impact on disease susceptibility. Of these, three are malaria resistance loci. The fourth is the prion protein gene, for which the selective agent is a special case of human infectious disease, consisting as it does of transmissible prion proteins themselves. The fifth is a deletion in the gene for C-C chemokine receptor type 5 (CCR5Δ32), which offers near complete resistance to HIV infection in the homozygous state (Dean et al. 1996; Huang et al. 1996; Liu et al. 1996). However, the high frequency of this mutation in Northern European populations cannot be attributed to selection from HIV since it is such a recent addition to the set of human pathogens. Smallpox and the plague have been suggested as potential selective pressures to account for the distribution of CCR5Δ32, with smallpox shown to be a theoretically more plausible candidate (Galvani and Slatkin 2003), but the cases for either are

weak (Hummel et al. 2005; Hedrick and Verrelli 2006), and others have suggested that CCR5 diversity may not be a consequence of any recent selection (Sabeti et al. 2005). The sixth gene highlighted by Hill is that encoding fucosyltransferase 2. A loss-of-function mutation in this gene affords resistance to Norwalk-like virus (Lindesmith et al. 2003; Thorven et al. 2005), and other diarrhoea-causing pathogens (Imbert-Marcille et al. 2014) - but it is not clear to what extent selection from these pathogens (or perhaps a combination of these and other infectious diseases) have determined the frequency of the mutation worldwide.

Haldane himself foresaw the difficulty of detecting human biochemical adaptations to pathogens, due to the transient nature of the advantage in most cases: "a disease such as diphtheria or tuberculosis is caused by a number of biochemically different races of pathogens … in a different epidemic a different a different type [of host] would be affected" (Haldane 1949). Our best-understood examples of *Plasmodium falciparum* malaria resistance mutations (haemoglobinopathies, glucose-6 phosphate dehydrogenase deficiency, southeast Asian ovalocytosis) all cause far reaching changes to red blood cell physiology, changes which are likely to affect multiple, even all, strains of *P. falciparum*. Were it not for the fact that these mutations all cause severe physiological problems in the homozygous state, and are thus under balancing selection, we may never have detected them as resistance mutations at all since they might have become fixed in an ancestral human population.

One set of diverse human proteins which are prime candidates for Haldane's notion of biochemical, pathogen strain specific adaptation are the human leukocyte antigens (HLAs). Class I and class II HLA molecules are responsible for the display of intracellularly derived peptide antigens (class I) and extracellularly derived peptide antigens (class II) so that they can be recognised by T cell receptors. The nature of the HLA molecule binding cleft determines the type of peptide which can be presented, creating a potential recognition bottleneck in human adaptive immunity. Hundreds of different alleles have been reported at the 3 class I loci and the 3 paired class II loci responsible for peptide display (HLA-A, B and C in class I and HLA-DRA and HLA-DRB, HLA-DPA and HLA-DPB

and HLADQA and HLADQB in class II) - polymorphism which has been attributed to selection, most likely from pathogens (Doherty and Zinkernagel 1975; Hughes and Nei 1988; Parham et al. 1989; Takahata and Nei 1990; Hedrick 2002; Borghans et al. 2004; De Boer et al. 2004; Prugnolle et al. 2005; Lenz 2011; Eizaguirre et al. 2012a; Eizaguirre et al. 2012b). It has in particular been argued that host-pathogen coevolution as opposed to heterozygote advantage is necessary to maintain such levels of polymorphism (Borghans et al. 2004). Specific HLA genotypes have been shown to confer susceptibility or resistance to different infectious disease outcomes (e.g. Hill et al. 1991; Kaslow et al. 1996; Jeffery et al. 1999; Dunstan et al. 2014). A recent study comparing the genetic diversity of an indigenous North American population before and after the arrival of European invaders (and their pathogens), found a dramatic change in the frequency of a HLA-DQA1 allele (Lindo et al. 2016), which might reflect that population's changing experience of infectious disease.

Population-level evolution of the pathogen HIV has been shown to occur in response to the immunological selection pressure generated by the presence of particular HLA types (Cotton et al, 2014; Payne et al. 2014), emphasising the co evolutionary potential of pathogen/HLA interactions. Although, as previously noted, HIV is unlikely to have had enough time to drive substantial changes in human allele frequencies, the relationship between HIV and HLA is worth considering in more detail, since it is the best studied pathogen-HLA interaction to date, with data drawn from cohorts of thousands of patients (reviewed in McLaren and Carrington, 2015). Certain HLA alleles are associated with better viral control and a slower progression to AIDS (e.g. HLA B*57 and HLA B*27 alleles), whilst others are associated with faster progression to AIDS (e.g. some HLA B*35 alleles). As noted by McLaren and Carrington, many individuals with a protective HLA type progress to AIDS at a similar rate to those without, thus there is no HLA allele which guarantees control of HIV. However, even though protection is not consistent across individuals, a chimpanzee MHC–B variant, Patr-B*06:03, with structural similarities to HLA B*57, is associated with lower SIV loads in chimpanzee faecal samples (Wroblewski et al 2015) – demonstrating that aspects of the way HLA/MHC molecules help combat retroviruses may be consistent across species. Recent studies have been able to identify the amino acids present at specific sites in HLA binding grooves which

account for protective effects previously identified at the allelic level – underscoring that the specific properties of the peptides that HLA molecules are capable of presenting to T cells has a critical impact on disease progression (The International HIV Controllers Study, 2010; McLaren et al 2015). However, the expression level of HLA-C has also been shown to have a protective effect, with higher expression of HLA-C associated with better viral control (Thomas *et al*, 2009; Kulkarni et al 2011; Apps et al 2013). Whether the protective effect of HLA-C expression level is due to better presentation to T cells, or to interactions with other elements of the immune system (e.g. Natural Killer cells) is unknown.

We have previously shown that a multi-strain pathogen and multi-gene host HLA haplotypes have the potential to display complex co-evolutionary cycling (Penman et al. 2013). Within the framework we proposed, at any given time only a small subset of host homozygotes would be susceptible to severe infection. The nature of that subset depended on the state of the pathogen population. However, this generated enough selection pressure to drive long lasting non-overlapping associations between alleles at separate HLA loci, even in the presence of recombination between those loci. We proposed that such non-overlapping associations could be a signature of pathogen selection and could even be harnessed as a means to functionally classify different HLAs. Our original model, however, included only a single pathogen species. HLAs must be under selection from multiple pathogens simultaneously. Here we simulate the co-evolution of two linked HLA loci with two independently circulating pathogens, where antigens from either pathogen can be displayed at either HLA locus. We show that, despite conflicting selection from a second pathogen, a pathogen that causes consistent, high mortality could theoretically leave a strong signature in HLA population genetics.

Methods

We adopted an individual-based simulation approach, extending that described in (Penman et al. 2013). We considered 10 different HLA binding types (represented by the digits 1-10), which could

be found on HLA molecules encoded by either of 2 linked HLA loci in the host genome. There was no restriction on which binding properties could be present at which HLA locus, which meant there were 100 possible HLA haplotypic combinations ([1,1]; [1,2]; [1,3] …. [10,10]) in our simulated population, arranged into diploid host genotypes. Once a host had been infected with a pathogen expressing a peptide which could be displayed by an HLA molecule encoded in that host's genome, we assumed that host to have lifelong immunity against infection with any other pathogen of that species expressing that peptide.

We assumed that two pathogen species were present (1 and 2). Each species possessed a number of antigenically variant peptides, expressed at two different sites per pathogen, and defined by the HLAs which could bind them (i.e. a digit between 1 and 10). We allowed 4 variants per antigenic site on each pathogen, thus 16 possible strains of each pathogen (K_{ij}). The distribution of variant peptides which could be displayed by particular HLA binding types for each antigenic site on each pathogen is given in table 1. A visualization of the relationship between host HLA genotype and pathogen strains is provided in Figure 1.

HLA binding types 1 and 6 only present motifs from pathogen 1. HLA binding types 5 and 10 only present motifs from pathogen 2. All other HLA binding types can present motifs from either pathogen. We assumed no cross immunity between the pathogens, thus whatever peptide from pathogen K_{ij} happened to be displayed by HLA *i* would not elicit any memory immune response against a peptide from pathogen L_{ij} that could also be displayed by HLA *i*.

Each host in the population was represented by a vector containing the host's age, sex, diploid HLA genotype, infection status and immunological status. A maximum of 2000 hosts could exist in the population, but this maximum did not have to be present at every time step. If, in a given generation, the population size ever dropped down to or below an arbitrarily chosen threshold (for the simulations shown here, 20 individuals), the population was deemed to have failed and that particular simulation ceased.

A single time step of our simulation represented one day. During each day, every host could (with probabilities defined in Table 2) become infected; recover from infection, or die from infection or die by random chance. Every time step, adult (>5400 days (~15 years) old) female hosts could also reproduce with a given probability, choosing a male partner at random, and generating an offspring genotype via Mendelian inheritance. For the simulations shown here, the age of reproducing males was not restricted to >15 years, but applying such a restriction makes no difference to the conclusions. An individual with the offspring genotype was then added to the population. If the population happened to be at its maximum possible size of 2000 then the new individual replaced a randomly chosen existing member of the population.

Every time a new infection occurred, one of that pathogen's antigenic sites could mutate so that it expressed one of the other peptides possible at that site with probability m. Every time hosts reproduced, recombination could occur between the two HLA loci, (in either maternal or paternal genotype) with probability r. For simplicity, our model does not explicitly simulate HLA mutation: over the timescale simulated, in a small population, frequency changes of existing HLA variation are likely to be more important than the spontaneous emergence of new HLA variants. However, each time step there was a fixed probability (α) of a new host individual, of randomly generated diploid genotype selected from HLA genes 1-10, replacing a randomly chosen existing member of the population. This represents migration into the population and ensured that the stochastic loss from the population of one of the pathogen species, or of a particular HLA binding specificity, was not permanent.

Parameter values and starting conditions

Our purpose was to determine whether it is possible that a specific pathogen species should leave a population genetic signature in the HLA region, despite conflicting selection pressures from other pathogens. Our analyses therefore focused on varying the probabilities of death from infection associated with the two pathogens (θ_1 and θ_2). We varied θ_1 between 0 and 0.0001 per day, and θ_2 between 0 and 0.002 per day. For most of our simulations, θ_1 and θ_2 applied to hosts of any age,

but we also explored whether our conclusions would change if infectious disease mortality only affected young children. To achieve this we carried out separate simulations where we only applied probabilities θ_1 and θ_2 of dying whilst infected to those < 1800 days (~5 years) old.

For pathogen 1, the transmission parameter and the probability of recovering from infection during any given day were always β_1= 0.3 and σ_1 = 0.02. When pathogen 2 was continuously present, we also applied β_2 = 0.3 and σ_2 = 0.02. However, we additionally sought to investigate the consequences of the periodic loss of pathogen 2 from the population. To generate this behaviour we applied a higher transmission parameter (β_2= 0.4) and a higher probability of recovering from infection on any given day (σ_2 = 0.1).

Rates of recombination in the HLA region appear to vary considerably (Carrington 1999, Cullen *et al* 2002). The results in the main text use a value of r =0, thus are more likely to apply to HLA loci that are physically very close, but we explore the effects of two higher recombination probabilities (r=0.01 and a very high probability of r=0.05) in the supplementary material.

All other parameters were fixed at values chosen to be plausible for human populations. The probability of dying from non-infectious disease causes on any given day (μ) was = 0.00007, and the probability of a female over the age of 5400 days (~15 years) giving birth on any given day (ϖ) was = 0.0015. These values ensured that the population exhibited a plausible age distribution for a human population in the absence of modern medicine: pyramidal in shape, with the greatest numbers of individuals in the youngest age groups and ~5% or less of the population over the age of 40 years (see figure S1). The probability of a new individual of a random genotype entering the population during any day was set at α = 0.000278, equivalent to assuming a migrant might arrive on average once every 3600 days, and the mutation probability of the pathogen was set at m= 0.00001 per new infection.

At the start of each simulation, 1000 hosts were present, with ages randomly assigned between 1 and 12600 days (~35 years). 90% of the HLA haplotypes in the population were of the combination [3,3], intended to capture the fact that a human population might be dominated by a relatively small number of founding HLA haplotypes. The remaining 10% of HLA haplotypes present were generated at random from the 10 possible HLA binding types. No hosts had any preexisting immunity to either pathogen at the start of the simulation. To seed infections, 10 hosts were chosen at random to be infected with randomly generated strains of pathogen 1, and 10 with randomly generated strains of pathogen 2. The simulations ran for 270000 days (~740 years). For each parameter combination presented in the main text or the supplementary material we carried out 300 simulations.

Results

i. High mortality from a single pathogen selects for host HLA haplotypes which recognise as many variants as possible from a single antigenic site on that pathogen.

We first considered the behaviour of the model when just one of the two pathogens caused mortality. We observed that when the mortality caused by pathogen *K* is very high, the two most frequent host haplotypes in the population after 740 years of coevolution contained within them exactly the 4 HLA binding types required to display all of the possible variants present at one of the antigenic loci belonging to pathogen *K*. Figure 2 displays the results of a simulation exhibiting such adaptation as a consequence of high levels of mortality from pathogen 2. [5,4] and [7,3] dominate the population, and between them could present any peptide that could be displayed at locus 1 of pathogen 2.

As noted in the Methods, we started each simulation with the population containing a high frequency of a single haplotype. Figure 2C shows that the high level of homozygosity (H_{obs}) associated with this state declines as pathogen selection begins, but as adaptation to pathogen

2 emerges, homozygosity rises once more – reflecting the high frequency of only a few HLA haplotypes in the adapted population.

Taking the behaviour shown in figure 2 as the most extreme form of population genetic adaptation possible, we defined 3 levels of population genetic adaptation to a multi-strain pathogen, which should occur at different pathogen mortality rates:

(i) *weak adaptation to pathogen K* : one of the two most frequent HLA haplotypes in the population can display one of pathogen *K*'s unique motifs.

(ii) *moderate adaption to pathogen K* : the two most frequent HLA haplotypes in the host population contain exactly the 4 HLA types required to display all of the possible variants present at one of pathogen *K*'s two antigenic loci, and the combined frequency of those two haplotypes is <=50%

(iii) *strong adaptation to pathogen K* : the two most frequent haplotypes in the population contain exactly the 4 HLA types required to display all of the possible variants present at one of pathogen *K*'s two antigenic loci, and the combined frequency of those two haplotypes is >50%. This is the case represented in figure 2.

ii. A high mortality pathogen can leave a strong genetic signature despite conflicting selection from a second pathogen

Figure 3 illustrates the population genetic patterns observed when both pathogens 1 and 2 cause mortality, and are continuously present in the population. The strength of adaptation to pathogen 2 increases with the probability of mortality whilst infected with pathogen 2 (indicated on the *x* axis in each graph).

Adaptation to pathogen 2 at high levels of mortality occurs despite the presence of conflicting selection from pathogen 1 (figures 3B and 3C). Mortality from pathogen 1 at a low level (figure 3B) barely disrupts adaptation to pathogen 2 at all, despite the greater pathogen burden on the

population evidenced by the reduction in population survival. A higher level of pathogen 1 mortality (figure 3C) is associated with some reductions in the probability of observing adaptation to pathogen 2, but so long as pathogen 2 has the *greater* probability of causing mortality (bars to the right of the red lines in figure 3), there is a greater probability that the population will display a form of adaptation to pathogen 2 than pathogen 1 (54% adaptation to pathogen 2 , 27% adaptation to pathogen 1 at θ_2 =0.00015; 47% adaptation to pathogen 2 , 37% adaptation to pathogen 1 at θ_2 =0.0002).

When both pathogens 1 and 2 cause a high level of mortality, we might have expected the conflicting selection pressures to lead to many simulated populations displaying no obvious adaptation. However, as seen in figure 3C, where both θ_1 =0.0001 and θ_2 =0.0001 (i.e. the pathogens have identical mortality probabilities), 89% of simulations display adaptation to one or other pathogen. At this level of pathogen 2 mortality, adding mortality from pathogen 1 simply increases the probability of observing any population adaptation at all. When θ_1 =0 and θ_2 =0.0001, 52% of simulations display no adaptation (figure 3A); when θ_1 =0.00005 and θ_2 =0.0001 , 47% of simulations display no adaptation (figure 3B), but when θ_1 =0.0001 and θ_2 =0.0001 , only 11% of simulations display no adaptation (figure 3C).

The patterns just described are largely unchanged by the addition of recombination (figures S2 and S3). However, the probability of observing strong adaptation to pathogen 2 in the presence of high mortality from pathogen 1 is reduced at 5% recombination between the HLA loci (figure S3C). Our definition of strong population adaptation involves the top two HLA haplotypes in the population having a combined frequency >50%. Recombination breaks up haplotypic associations, thus it is entirely reasonable that high levels of recombination should make strong adaptation less likely. Nevertheless, adaptation to pathogen 2 itself (when weak, moderate and strong forms are taken together) increases with increasing pathogen 2 mortality in our simulations at 5% recombination (figure S3C), despite the conflicting selection from pathogen 1.

When we limited infectious disease mortality to individuals < 5 years of age, we obtained similar patterns at higher values of θ_1 and θ_2. (figure S4). To achieve a selective pressure capable of shaping the population's HLA distribution when individuals are only vulnerable to infectious disease mortality for a short period of time requires higher individual probabilities of death from infection within that window of vulnerability.

iii. High mortality pathogens are less likely to leave a strong genetic signature if their presence in the population is not continuous

If we allow pathogen 2 to have a faster recovery rate and a higher transmission probability we can generate scenarios where pathogen 2 can become lost from the population due to burning through its available susceptible hosts. Following such stochastic loss, pathogen 2 can be re introduced by an infected host arriving in the host as a random introduction. As shown in figure 4, weak, moderate or strong adaptation to an intermittently present pathogen 2 can still be observed if pathogen 2 causes mortality. However, the greater the mortality caused by the continuously present pathogen 1, the more likely we are to observe adaptation to pathogen 1 at the expense of adaptation to pathogen 2, and the less likely we are to observe moderate or strong adaptation to pathogen 2 (compare panels 4A, 4B and 4C). A low level of mortality from the continuously present pathogen 1 (θ_1 = 0.0005) causes more loss of adaptation to the intermittently present pathogen 2 than when pathogen 2 was continuously present (compare figures 4B and 3B).

In figure 4, we allowed the intermittently present pathogen 2 much higher mortality rates than the continuously present pathogen 1, so as to maximise the selective pressure caused by pathogen 2. As shown in the left hand panels in figure 4, at the highest mortality probabilities, the pathogen load approaches that at which most populations do not survive. Interestingly, however, increasing the probability of death from infection with pathogen 2 seems to have little impact on the probability of

observing a population specifically adapted to pathogen 2 (compare the left to right trends within the graphs in figure 4 with the graphs in figure 3). It may be that for mortality-causing pathogens which are only present intermittently, the frequency of the exposure of the population to the pathogen is more important than the chance of dying whilst infected in determining whether or not the population exhibits population genetic adaptation to that pathogen. Additionally, too high a mortality rate for pathogen 2 may contribute to its rapid loss from the population during any individual epidemic, which could also reduce its ability to leave a population genetic signature.

When we allowed recombination to occur between the HLA loci, we observed a clear reduction in the probability of observing strong population adaptation to the intermittently present pathogen (pathogen 2) [figures S5 and S6]. At 5% recombination between the loci we never observed strong adaptation to pathogen 2 (figure S6), although weak and moderate adaptation was still possible. As noted previously, this effect is unsurprising, since our definition of strong population adaptation involves >50% of the HLA haplotypes in the population being adapted to the pathogen in question. If pathogen 2 is only intermittently present, every time it is absent from the population, recombination will act unchecked to break up the haplotypic combinations that are specifically adapted to pathogen 2 – thus maintaining combined frequencies of such haplotypes >50% is unlikely.

Discussion

Our simulations demonstrate that individual high-mortality pathogens have the potential to generate specific signatures amongst HLA genes, despite conflicting selection from other mortality-causing pathogens. These signatures take the form of population-level HLA haplotype frequency patterns. The most important implications of our two-pathogen model can be summarized as follows:

(i) The greater the overall pathogen burden, the more likely a population is to display specific adaptation to any pathogen.

(ii) For continuously-present pathogens, the higher the pathogen mortality, the more likely the pathogen is to leave a signature.

(iii) Population genetic signatures of adaptation to intermittently-present pathogens can be readily disrupted by selection from continuously present pathogens, and the lethality of an intermittently present pathogen *per se* is not a predictor of whether adaptation will occur.

Pathogens which are likely to have caused high levels of mortality for continuous periods in the history of various human populations include *Plasmodium falciparum*, *Leishmania* spp. *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, *Treponema pallidum*, Poliovirus, Smallpox Virus and Yellow Fever virus. Our simulations suggest that pathogens such as these might be more likely to have determined the array of HLA haplotypes that successfully reached high frequencies in affected populations than characteristically intermittent pathogens such as *Bacillus anthracis*, *Yersinia pestis* or *Rickettsia* spp. We demonstrated that the mortality rate of an intermittently present pathogen has little effect on the probability of observing adaptation to that pathogen, and speculated that the frequency of introduction of intermittently present pathogens may be more important. However, to make any prediction of the frequency of introduction and/or duration of epidemics necessary for any given intermittent pathogen to have left an HLA signature will require additional theoretical work, as well as improved understanding of the strain diversity present in the pathogen species of interest.

Although the population genetic signatures of pathogen selection we have identified take the form of HLA haplotype frequency patterns, our model makes no explicit assumption that selection acts at the allelic or haplotypic level. However, we do assume that the effects of being able to express different HLA molecules combine additively. This means that it is always advantageous to maximize the diversity of HLA recognition types present in a host genome, and this in turn generates a specific form of selection at the haplotypic level, for only certain combinations of haplotypes maximize HLA recognition diversity when they coexist. Maximising recognition diversity certainly

366 seems likely to be a major factor in determining the evolution of HLA alleles and haplotypes – but it
367 is possible that HLA alleles interact with one another in non-additive ways too. The most obvious
368 ways in which this could occur are (i) if HLA expression level is important (as for HIV and HLA-C),
369 or (ii) if HLA alleles differ in the breadth of types of peptide that they can display (i.e. in their binding
370 promiscuity), which certainly affects MHC based infectious disease susceptibility in chickens
371 (Chappell et al, 2015), and which is also linked to expression level of the MHC/HLA molecule in
372 question (Chappell et al, 2015). The type of population genetic pattern which may result from
373 pathogen selection where HLA expression level or binding promiscuity (or both) is crucial is beyond
374 the scope of our present model, and allowing for such effects in future models is a priority.

375

376 Most pathogens possess greater antigenic diversity than that represented in our model, and
377 humans certainly possess greater HLA diversity. Furthermore, our definitions of “weak”, “moderate”
378 and “strong” patterns of selection rely on our complete knowledge of the modelled system and
379 which antigenic variants are expressed by which pathogens. These definitions are therefore not
380 intended to be applied directly to human populations (where such complete knowledge is beyond
381 our current understanding), but rather to illustrate the principle that the highest frequency HLA
382 haplotypes present in a given human population *might* represent “moderate” or “strong” population
383 genetic signatures of specific pathogens. In other words those haplotypes might, between them,
384 maximize the capacity of the human immune system to recognize the antigenic diversity present at
385 just one variable site of a single pathogen species, despite the fact that HLA loci are under selection
386 from multiple pathogen species. We propose that it is worth considering which of the mortality-
387 causing pathogens that have coexisted with particular populations for a long time could be
388 responsible for the elevation of particular combinations of HLA haplotypes. An additional important
389 principle to emerge from our model is that selection from identical pathogens could still result in
390 completely different suites of HLA haplotypes reaching high frequency in different populations,
391 depending on the antigenic site which happened to become immunodominant (i.e. the antigenic
392 site that population’s HLAs evolved to target).

Will it ever be feasible to measure the degree to which HLA recognition capacity in a population prioritises the variants of a specific pathogen antigen? The immune epitope database (Vita et al. 2015) is an invaluable resource, collating our current knowledge of antibody and T cell epitopes. It is, however, limited by the experiments which have so far been carried out, so does not represent an unbiased sampling of epitopes that *could* be recognized. However, as epitope prediction continues to improve for different MHC molecules, and as whole genome datasets become available for more and more pathogens, it may become possible to look for correlations between the highest frequency HLA haplotypes in specific populations and their capacity, across multiple HLA loci, to recognize the variation encoded by candidate antigenic regions in high-mortality pathogen genomes. If evolutionary HLA-pathogen relationships can be identified in this way, they will help focus our attention on the most immunogenic elements of those pathogens, which will be of enormous benefit to ongoing efforts to develop effective treatments and prophylaxis.

In the introduction we noted that selection from malaria parasites has had the most easily measurable impact on human genetics. It is becoming clear, however, that understanding malaria selection by examining a single locus at a time is insufficient: interactions between protective mutations at separate loci can cancel out the malaria protective effect of both when they are co-inherited (Williams et al. 2005). Furthermore, such epistasis may have determined the particular suites of protective variants that co-exist in given populations (Penman et al. 2009; Penman et al. 2011; Penman et al. 2012). The simulations we present here demonstrate that these principles could be usefully applied to understanding human-pathogen coevolution generally: adaptation to a pathogen can take the form of the specific collection of alleles found across several loci, not the particular variants found at only a single locus.

In addition to the likely non-additive fitness consequences of particular alleles at different HLA loci, HLA alleles have been shown to interact epistatically with variants at Killer cell

Immunoglobulin like Receptor loci [KIRs] (Martin et al. 2002; Hiby et al. 2004; Khakoo et al. 2004; Seich al Basatena et al. 2011). KIRs are Natural Killer Cell receptors which are very likely to be undergoing co-evolution with pathogens (Parham and Moffett 2013; Carrillo-Bustamante et al. 2013; Carrillo-Bustamante et al. 2014; Carrillo-Bustamante et al. 2015; Penman et al. 2016), and many KIRs interact directly with HLA molecules in order to perform their function. The repertoire of KIR alleles present in a particular population may thus also shape the set of HLA haplotypes that come to dominate. Balancing selection for extremely high polymorphism in both HLAs and KIRs is evident in a detailed study of a West African population (Norman et al. 2013). Cappittini *et al* observed that HLA-A,B haplotypic combinations in an Italian population are configured so that HLA-B alleles which do not serve as KIR ligands are more likely to be found alongside HLA-A alleles which do serve as KIR ligands – maximizing the chance that at least one of HLA-A or HLA–B in an individual's genome should have an interacting KIR (Capittini et al. 2012). It has also been shown that class I HLAs tend to exist in haplotypes that either combine HLA-B and -C KIR ligands, *or* have HLA B alleles which are able to supply ligands for another Natural Killer Cell receptor, CD94:NKG2A (Horowitz et al 2016). Such effects will have acted alongside selection from specific pathogens in determining the HLA patterns that have emerged in individual populations, and incorporating them in future simulation models will assist in attempts to delineate the population genetic signatures of both.

Theoretical work on generalised host-pathogen systems has shown that selection from two independent pathogens, interacting with two separate host loci, can drive the evolution of "high complementarity equilibria" whereby the host loci exhibit strong linkage disequilibrium (Kouyos et al. 2009). For that specific type of population genetic patterning to emerge, both pathogens would have to be present. Here we have focused on the confounding effects of dual pathogen selection to show that a single pathogen can still drive population genetic patterning even when a second pathogen interacts with the same loci. However, future work should also consider the situation where a subset of pathogens interact solely with a subset of HLA loci, and other pathogens interact

solely with a different subset – the overarching population genetic rules governing the associations between different sets of alleles at different HLA loci are likely to be affected by such structuring.

Conclusion

As Haldane pointed out, surviving infectious disease is on a par with the pressure to find food or successfully mate in terms of evolutionary significance. For human-pathogen coevolution, the case for malaria selection is clear, but we have few other examples of infectious diseases that can be linked directly to changes in human allele frequencies. Our simulations suggest that evolutionary signatures of specific, continuously present, high mortality human pathogens should exist in the form of particular combinations of HLA haplotypes. Identifying and understanding such patterns could ultimately pay dividends as we seek to mitigate or emulate the contributions of different genotypes to human health.

Acknowledgments

S.G. receives funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 268904-DIVERSITY

References

Allison A. C. 1954. Protection afforded by sickle-cell trait against subtertian malareal infection. British medical journal 1: 290-294.

Apps R., Qi Y., Carlson J. M., Chen H., Gao X., Thomas R., Yuki Y., Del Prete G. Q., Goulder P., Brumme Z. L., Brumme C. J., John M., Mallal S., Nelson G., Bosch R., Heckerman D., Stein J. L., Soderberg K. A., Moody M. A., Denny T. N., Zeng X., Fang J., Moffett A., Lifson J. D., Goedert J. J., Buchbinder S., Kirk G. D., Fellay J., McLaren P., Deeks S. G., Pereyra F., Walker B., Michael N. L., Weintrob A., Wolinsky S., Liao W. and Carrington M. 2013. Influence of HLA-C expression level on HIV control. Science 340: 87-91.

Band G., Rockett K. A., Spencer C. C. A., Kwiatkowski D. P., Si Le Q., Clarke G. M., Kivinen K., Leffler E. M., Cornelius V., Conway D. J., Williams T. N., Taylor T., Bojang K. A., Doumbo O., Thera M. A., Modiano D., Sirima S. B., Wilson M. D., Koram K. A., Agbenyega T., Achidi E., Marsh K., Reyburn H., Drakeley C., Riley E., Molyneux M., Jallow M., Pinder M., Toure O. B., Konate S., Sissoko S., Bougouma E. C., Mangano V. D., Amenga-Etego L. N., Ghansah A. K., Hodgson A. V. O., Wilson M. D., Ansong D., Enimil A., Evans J., Apinjoh T. O., Macharia A., Ndila C. M., Newton C., Peshu N., Uyoga S., Manjurano A., Kachala D., Nyirongo V., Mead D., Drury E., Auburn S., Campino S. G., MacInnis B., Stalker J., Gray E., Hubbart C., Jeffreys A. E., Rowlands K., Mendy A., Craik R., Fitzpatrick K., Molloy S., Hart L., Hutton R., Kerasidou A. and Johnson K. J. 2015. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature 526: 253-257.

Borghans J. A. M., Beltman J. B. and De Boer R. J. 2004. MHC polymorphism under host-pathogen coevolution. Immunogenetics 55: 732-739.

Capittini C., Tinelli C., Guarene M., Pasi A., Badulli C., Sbarsi I., Garlaschelli F., Cremaschi A. L., Pizzochero C., Monti C., Salvaneschi L. and Martinetti M. 2012. Possible KIR-driven genetic pressure on the genesis and maintenance of specific HLA-A,B haplotypes as functional genetic blocks. Genes and immunity 13: 452–457

Carrillo-Bustamante P., Kesmir C. and De Boer R. J. 2015. A coevolutionary arms race between hosts and viruses drives polymorphism and polygenicity of NK cell receptors. Molecular biology and evolution 32: 2149-2160.

Carrillo-Bustamante P., Kesmir C. and de Boer R. J. 2014. Quantifying the protection of activating and inhibiting NK cell receptors during infection with a CMV-like virus. Frontiers in Immunology 5:20

Carrillo-Bustamante P., Kesmir C. and de Boer R. J. 2013. Virus Encoded MHC-Like Decoys Diversify the Inhibitory KIR Repertoire. PLoS Computational Biology 9(10): e1003264 https://doi.org/10.1371/journal.pcbi.1003264

Carrington M. 1999. Recombination within the human MHC. Immunological reviews 167: 245-256.

Chappell P., Meziane E. K., Harrison M., Magiera L., Hermann C., Mears L., Wrobel A. G., Durant C., Nielsen L. L., Buus S., Ternette N., Mwangi W., Butter C., Nair V., Ahyee T., Duggleby R., Madrigal A., Roversi P., Lea S. M. and Kaufman J. 2015. Expression levels of mhc class i molecules are inversely correlated with promiscuity of peptide binding. eLife 2015.

Cotton L. A., Kuang X. T., Le A. Q., Carlson J. M., Chan B., Chopera D. R., Brumme C. J., Markle T. J., Martin E., Shahid A., Anmole G., Mwimanzi P., Nassab P., Penney K. A., Rahman M. A., Milloy M. -., Schechter M. T., Markowitz M., Carrington M., Walker B. D., Wagner T., Buchbinder S., Fuchs

J., Koblin B., Mayer K. H., Harrigan P. R., Brockman M. A., Poon A. F. Y. and Brumme Z. L. 2014. Genotypic and Functional Impact of HIV-1 Adaptation to Its Host Population during the North American Epidemic. PLoS Genetics 10(4): e1004295. https://doi.org/10.1371/journal.pgen.1004295

Cullen M., Perfetto S. P., Klitz W., Nelson G. and Carrington M. 2002. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. American Journal of Human Genetics 71: 759-776.

De Boer R. J., Borghans J. A. M., Van Boven M., Kesmir C. and Weissing F. J. 2004. Heterozygote advantage fails to explain the high degree of polymorphism of the MHC. Immunogenetics 55: 725-731.

Dean M., Carrington M., Winkler C., Huttley G. A., Smith M. W., Allikmets R., Goedert J. J., Buchbinder S. P., Vittinghoff E., Gomperts E., Donfield S., Vlahov D., Kaslow R., Saah A., Rinaldo C., Detels R. and O'Brien S. J. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science 273: 1856-1862.

Doherty P. C. and Zinkernagel R. M. 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 256: 50-52.

Dunstan S. J., Hue N. T., Han B., Li Z., Tram T. T. B., Sim K. S., Parry C. M., Chinh N. T., Vinh H., Lan N. P. H., Thieu N. T. V., Vinh P. V., Koirala S., Dongol S., Arjyal A., Karkey A., Shilpakar O., Dolecek C., Foo J. N., Phuong L. T., Lanh M. N., Do T., Aung T., Hon D. N., Teo Y. Y., Hibberd M. L., Anders K. L., Okada Y., Raychaudhuri S., Simmons C. P., Baker S., De Bakker P. I. W., Basnyat B., Hien T. T., Farrar J. J. and Khor C. C. 2014. Variation at HLA-DRB1 is associated with resistance to enteric fever. Nature genetics 46: 1333-1336.

Eizaguirre C., Lenz T. L., Kalbe M. and Milinski M. 2012a. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. Nature Communications 3:621

Eizaguirre C., Lenz T. L., Kalbe M. and Milinski M. 2012b. Divergent selection on locally adapted major histocompatibility complex immune genes experimentally proven in the field. Ecology Letters 15: 723-731.

Flint J., Hill A. V. S. and Bowden D. K. 1986. High frequencies of α-thalassaemia are the result of natural selection by malaria. Nature 321: 744-750.

Galvani A. P. and Slatkin M. 2003. Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele. Proceedings of the National Academy of Sciences of the United States of America 100: 15276-15279.

Haldane J. B. S. 1949. Disease and Evolution. Ricerca Scientifica (suppl) 19: 68.

Hedrick P. W. 2002. Pathogen resistance and genetic variation at MHC loci. Evolution 56: 1902-1908.

Hedrick P. W. and Verrelli B. C. 2006. 'Ground truth' for selection on CCR5-Δ32. Trends in Genetics 22: 293-296.

Hiby S. E., Walker J. J., O'Shaughnessy K. M., Redman C. W. G., Carrington M., Trowsdale J. and Moffett A. 2004. Combinations of maternal KIR and fetal HLA-C genes influence the risk of preeclampsia and reproductive success. Journal of Experimental Medicine 200: 957-965.

544 Hill A. V. S. 2006. Aspects of genetic susceptibility to human infectious diseases. Annual Review of
545 Genetics 40: 469-486.

546 Hill A. V. S., Allsopp C. E. M., Kwiatkowski D., Anstey N. M., Twumasi P., Rowe P. A., Bennett S.,
547 Brewster D., McMichael A. J. and Greenwood B. M. 1991. Common West African HLA antigens are
548 associated with protection from severe malaria. Nature 352: 595-600.

549 Horowitz A., Djaoud Z., Nemat-Gorgani N., Blokhuis J., Hilton H. G., Béziat V., Malmberg K.,
550 Norman P. J., Guethlein L. A. and Parham P. 2016. Class I HLA haplotypes form two schools that
551 educate NK cells in different ways. Science Immunology 1: eaag1672

552 Huang Y., Paxton W. A., Wolinsky S. M., Neumann A. U., Zhang L., He T., Kang S., Ceradini D., Jin
553 Z., Yazdanbakhsh K., Kunstman K., Erickson D., Dragon E., Landau N. R., Phair J., Ho D. D. and
554 Koup R. A. 1996. The role of a mutant CCR5 allele in HIV-1 transmission and disease progression.
555 Nature medicine 2: 1240-1243.

556 Hughes A. L. and Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex
557 class I loci reveals overdominant selection. Nature 335: 167-170.

558 Hummel S., Schmidt D., Kremeyer B., Herrmann B. and Oppermann M. 2005. Detection of the
559 CCR5-Δ32 HIV resistance gene in Bronze Age skeletons. Genes and immunity 6: 371-374.

560 Imbert-Marcille B. -., Barbé L., Dupé M., Le Moullac-Vaidye B., Besse B., Peltier C., Ruvoën-Clouet
561 N. and Le Pendu J. 2014. A FUT2 gene common polymorphism determines resistance to rotavirus a
562 of the P[8] genotype. Journal of Infectious Diseases 209: 1227-1230.

563 Jeffery K. J. M., Usuku K., Hall S. E., Matsumoto W., Taylor G. P., Procter J., Bunce M., Ogg G. S.,
564 Welsh K. I., Weber J. N., Lloyd A. L., Nowak M. A., Nagai M., Kodama D., Izumo S., Osame M. and
565 Bangham C. R. M. 1999. HLA alleles determine human T-lymphotropic virus-I (HTLV-I) proviral load
566 and the risk of HTLV-I-associated myelopathy. Proceedings of the National Academy of Sciences of
567 the United States of America 96: 3848-3853.

568 Kaslow R. A., Carrington M., Apple R., Park L., Muñoz A., Saah A. J., Goedert J. J., Winkler C.,
569 O'Brien S. J., Rinaldo C., Detels R., Blattner W., Phair J., Erlich H. and Mann D. L. 1996. Influence
570 of combinations of human major histocompatibility complex genes on the course of HIV-1 infection.
571 Nature Medicine 2: 405-411.

572 Khakoo S. I., Thio C. L., Martin M. P., Brooks C. R., Gao X., Astemborski J., Cheng J., Goedert J.
573 J., Vlahov D., Hilgartner M., Cox S., Little A. -., Alexander G. J., Cramp M. E., O'Brien S. J.,
574 Rosenberg W. M. C., Thomas D. L. and Carrington M. 2004. HLA and NK cell inhibitory receptor
575 genes in resolving hepatitis C virus infection. Science 305: 872-874.

576 Kouyos R. D., Salathé M., Otto S. P. and Bonhoeffer S. 2009. The role of epistasis on the evolution
577 of recombination in host-parasite coevolution. Theoretical population biology 75: 1-13

578 Kulkarni S., Savan R., Qi Y., Gao X., Yuki Y., Bass S. E., Martin M. P., Hunt P., Deeks S. G., Telenti
579 A., Pereyra F., Goldstein D., Wolinsky S., Walker B., Young H. A. and Carrington M. 2011.
580 Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature
581 472: 495-498.

582

583 Kwiatkowski D. P. 2005. How malaria has affected the human genome and what human genetics
584 can teach us about malaria. American Journal of Human Genetics 77: 171-192.

585 Lederberg J. 1999. J. B. S. Haldane (1949) on infectious disease and evolution. Genetics 153: 1-3.

586 Lenz T. L. 2011. Computational prediction of MHC II-antigen binding supports divergent allele
587 advantage and explains trans-species polymorphism. Evolution 65: 2380-2390.

588 Lindesmith L., Moe C., Marionneau S., Ruvoen N., Jiang X., Lindblad L., Stewart P., Lependu J.
589 and Baric R. 2003. Human susceptibility and resistance to Norwalk virus infection. Nature medicine
590 9: 548-553.

591 Lindo J., Huerta-Sánchez E., Nakagome S., Rasmussen M., Petzelt B., Mitchell J., Cybulski J. S.,
592 Willerslev E., Degiorgio M. and Malhi R. S. 2016. A time transect of exomes from a Native American
593 population before and after European contact. Nature Communications 7:13175

594 Liu R., Paxton W. A., Choe S., Ceradini D., Martin S. R., Horuk R., MacDonald M. E., Stuhlmann H.,
595 Koup R. A. and Landau N. R. 1996. Homozygous defect in HIV-1 coreceptor accounts for
596 resistance of some multiply-exposed individuals to HIV-1 infection. Cell 86: 367-377.

597 Martin M. P., Gao X., Lee J. -., Nelson G. W., Detels R., Goedert J. J., Buchbinder S., Hoots K.,
598 Vlahov D., Trowsdale J., Wilson M., O'Brien S. J. and Carrington M. 2002. Epistatic interaction
599 between KIR3DS1 and HLA-B delays the progression to AIDS. Nature genetics 31: 429-434.

600 McLaren P. J. and Carrington M. 2015. The impact of host genetic variation on infection with HIV-1.
601 Nature immunology 16: 577-583.

602 McLaren P. J., Coulonges C., Bartha I., Lenz T. L., Deutsch A. J., Bashirova A., Buchbinder S.,
603 Carrington M. N., Cossarizza A., Dalmau J., De Luca A., Goedert J. J., Gurdasani D., Haas D. W.,
604 Herbeck J. T., Johnson E. O., Kirk G. D., Lambotte O., Luo M., Mallal S., Van Manen D., Martinez-
605 Picado J., Meyer L., Miro J. M., Mullins J. I., Obel N., Poli G., Sandhu M. S., Schuitemaker H., Shea
606 P. R., Theodorou I., Walker B. D., Weintrob A. C., Winkler C. A., Wolinsky S. M., Raychaudhuri S.,
607 Goldstein D. B., Telenti A., De Bakker P. I. W., Zagury J. -. and Fellay J. 2015. Polymorphisms of
608 large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load.
609 Proceedings of the National Academy of Sciences of the United States of America 112: 14658-
610 14663.

611 Norman P. J., Hollenbach J. A., Nemat-Gorgani N., Guethlein L. A., Hilton H. G., Pando M. J.,
612 Koram K. A., Riley E. M., Abi-Rached L. and Parham P. 2013. Co-evolution of Human Leukocyte
613 Antigen (HLA) Class I Ligands with Killer-Cell Immunoglobulin-Like Receptors (KIR) in a Genetically
614 Diverse Population of Sub-Saharan Africans. PLoS Genetics 9(10):e1003938. doi:
615 10.1371/journal.pgen.1003938.
616

617 Parham P. and Moffett A. 2013. Variable NK cell receptors and their MHC class i ligands in
618 immunity, reproduction and human evolution. Nature Reviews Immunology 13: 133-144.

619 Parham P., Lawlor D. A., Lomen C. E. and Ennis P. D. 1989. Diversity and diversification of HLA-
620 A,B,C alleles. Journal of Immunology 142: 3937-3950.

621 Payne R., Muenchhoff M., Mann J., Roberts H. E., Matthews P., Adland E., Hempenstal A., Huang
622 K.-H.., Brockman M., Brumme Z., Sinclair M., Miura T., Frater J., Essex M., Shapiro R., Walker B.
623 D., Ndung'u T., McLean A. R., Carlson J. M. and Goulder P. J. R. 2014. Impact of HLA-driven HIV

624 adaptation on virulence in populations of high HIV seroprevalence. Proceedings of the National
625 Academy of Sciences of the United States of America 111: E5393-E5400.

626 Penman B. S., Pybus O. G., Weatherall D. J. and Gupta S. 2009. Epistatic interactions between
627 genetic disorders of hemoglobin can explain why the sickle-cell gene is uncommon in the
628 Mediterranean. Proceedings of the National Academy of Sciences 106: 21242-21246.

629 Penman B. S., Ashby B., Buckee C. O. and Gupta S. 2013. Pathogen selection drives
630 nonoverlapping associations between HLA loci. Proceedings of the National Academy of Sciences
631 of the United States of America 110: 19645-19650.

632 Penman B. S., Gupta S. and Buckee C. O. 2012. The emergence and maintenance of sickle cell
633 hotspots in the Mediterranean. Infection, Genetics and Evolution 12: 1543-1550.

634 Penman B. S., Habib S., Kanchan K. and Gupta S. 2011. Negative epistasis between α+
635 thalassaemia and sickle cell trait can explain interpopulation variation in South Asia. Evolution 65:
636 3625-3632.

637 Penman B. S., Moffett A., Chazara O., Gupta S. and Parham P. 2016. Reproduction, infection and
638 killer-cell immunoglobulin-like receptor haplotype evolution. Immunogenetics 68: 755-764.

639 Prugnolle F., Manica A., Charpentier M., Guégan J. F., Guernier V. and Balloux F. 2005. Pathogen-
640 driven selection and worldwide HLA class I diversity. Current Biology 15: 1022-1027.

641 Sabeti P. C., Walsh E., Schaffner S. F., Varilly P., Fry B., Hutcheson H. B., Cullen M., Mikkelsen T.
642 S., Roy J., Patterson N., Cooper R., Reich D., Altshuler D., O'Brien S. and Lander E. S. 2005. The
643 case for selection at CCR5-Delta32. PLoS biology. 3(11): e378.
644 https://doi.org/10.1371/journal.pbio.0030378

645 Seich al Basatena N., MacNamara A., Vine A. M., Thio C. L., Astemborski J., Usuku K., Osame M.,
646 Kirk G. D., Donfield S. M., Goedert J. J., Bangham C. R. M., Carrington M., Khakoo S. I. and
647 Asquith B. 2011. KIR2DL2 enhances protective and detrimental HLA class I-mediated immunity in
648 chronic viral infection. PLoS Pathogens 7(10): e1002270.
649 https://doi.org/10.1371/journal.ppat.1002270

650 Siniscalco M., Bernini L., Latte B. and Motulsky A. G. 1961. Favism and Thalassæmia in Sardinia
651 and their relationship to malaria. Nature 190: 1179-1180.

652 Takahata N. and Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent
653 selection and polymorphism of major histocompatibility complex loci. Genetics 124: 967-978.

654 The International HIV Controllers Study. 2010. The Major Genetic Determinants of HIV-1 Control
655 Affect HLA Class I Peptide Presentation. Science 330 (6010):1551-1557

656 Thorven M., Grahn A., Hedlund K. -., Johansson H., Wahlfrid C., Larson G. and Svensson L. 2005.
657 A homozygous nonsense mutation (428G→A) in the human secretor (FUT2) gene provides
658 resistance to symptomatic norovirus (GGII) infections. Journal of virology 79: 15351-15355.

659 Thomas R., Apps R., Qi Y., Gao X., Male V., O'Huigin C., O'Connor G., Ge D., Fellay J., Martin J.
660 N., Margolick J., Goedert J. J., Buchbinder S., Kirk G. D., Martin M. P., Telenti A., Deeks S. G.,
661 Walker B. D., Goldstein D., McVicar D. W., Moffett A. and Carrington M. 2009. HLA-C cell surface
662 expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. Nature genetics 41:
663 1290-1294.

Vita R., Overton J. A., Greenbaum J. A., Ponomarenko J., Clark J. D., Cantrell J. R., Wheeler D. K., Gabbard J. L., Hix D., Sette A. and Peters B. 2015. The immune epitope database (IEDB) 3.0. Nucleic acids research 43: D405-D412.

Williams T. N., Mwangi T. W., Wambua S., Peto T. E. A., Weatherall D. J., Gupta S., Recker M., Penman B. S., Uyoga S., Macharia A., Mwacharo J. K., Snow R. W. and Marsh K. 2005. Negative Epistasis between the malaria-protective effects of alpha+ thalassemia and the sickle cell trait. Nature Genetics 37: 1253-1257.

Wroblewski E. E., Norman P. J., Guethlein L. A., Rudicell R. S., Ramirez M. A., Li Y., Hahn B. H., Pusey A. E. and Parham P. 2015. 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. https://doi.org/10.1371/journal.pbio.1002144

678 **Figure legends**

679 **Figure 1: A schematic representation of model assumptions.** As noted in the Methods, we
680 allowed there to exist 10 HLA types with different binding properties (represented by the numbers 1-
681 10), which could be encoded by genes found at either locus of a 2 locus HLA haplotype. Pathogen
682 species 1 and 2 each possess two antigenic sites (represented here by different colours), at which
683 antigens containing peptide fragments which could be bound by specific HLA molecules can be
684 expressed. 4 different antigens can be expressed at each antigenic site (see table 1 for a
685 description of which antigenic variants are present on which site in which pathogen species).The
686 combination of HLA binding types which can present peptides from a particular pathogen defines its
687 strain, e.g. a possible strain of pathogen 1 is [2,8]. As illustrated in this figure, certain HLA
688 molecules are capable of presenting a peptide from either pathogen 1 or pathogen 2. Note that this
689 figure does not display the entire range of possible host or pathogen genotypes.

690 **Figure 2: Changing frequencies of HLA haplotypes over time, under selection from**
691 **pathogen 2.** Panel (A) illustrates the frequencies of different HLA haplotypes over the course of a
692 single simulation, panel (B) illustrates the frequencies of different strains of pathogen 2 during the
693 same simulation, and panel (C) indicates the proportion of the population which is homozygous for
694 any HLA haplotype (homozygosity, *H*), and the ratio of the observed homozygosity in the simulation
695 (H_{obs}) to that expected under Hardy Weinberg proportions (H_{exp}). Each shade of grey in panels (A)
696 and (B) represents a different haplotype or pathogen strain. There are too many HLA haplotypes
697 and pathogen strains to label individually, but 2 host haplotypes have been highlighted in red and
698 blue. Between them, these haplotypes cover all 4 possible variants at antigenic site 1 of pathogen 2.
699 Parameter values as follows: $r = 0.01$, $\beta_2 = 0.3$, $\sigma_2 = 0.02$, $\theta_1 = 0$ and $\theta_2 = 0.002$; other parameters
700 were as detailed in the Methods.

701

702 **Figure 3: The adaptation of populations under continuous selection from pathogens 1 and 2.**
703 The bar chart on the left hand side of each panel illustrates the proportion of simulated populations

surviving, out of 300 simulations at each parameter combination. The bar chart on the right hand side of each panel illustrates the proportion of the surviving populations displaying adaptation to one or other pathogen, or no adaptation signal (see legend, and see text for definition of different types of adaptation). Within each graph the mortality caused by pathogen 2 (θ_2) increases along the x axis. The mortality caused by pathogen 1 is zero in panel A ($\theta_1 = 0$), and increases in value in panels B and C (B: $\theta_1 = 0.00005$, C: $\theta_1 = 0.0001$). Pathogen 2 has a higher probability of causing death during infection than pathogen 1 ($\theta_2 > \theta_1$) in the regions to the right hand side of the vertical red line in each panel. $\beta_2 = 0.3$, $\sigma_2 = 0.02$ and r=0. All other parameter values were as detailed in the Methods.

Figure 4: The adaptation of populations under continuous selection from pathogen 1 and intermittent selection from pathogen 2. This figure uses the same layout as figure 3. Unlike in figure 3, however, the transmission parameter and recovery rate for pathogen 2 have been given values that lead to pathogen 2 being lost and re introduced into the population ($\beta_2 = 0.4$ and $\sigma_2 = 0.1$). The range of mortality rates affecting to those infected with pathogen 2 (θ_2) are also higher than in figure 3, as indicated by the x axis of each graph. Just as in figure 3, the mortality caused by pathogen 1 is zero in panel A ($\theta_1 = 0$), and increases in value in panels B and C (B: $\theta_1 = 0.00005$, C: $\theta_1 = 0.0001$). All other parameters are as given in the Methods.

Table 1: The antigenic properties of the two pathogens. We conceptualise pathogen antigenic variation in terms of which host HLA binding sites are capable of presenting a peptide from any particular pathogen antigenic site. As shown in Table 1, for pathogen 1, site *i*, we assume that there are 4 possible peptide variants the pathogen can express, which can be displayed by HLA molecules 1, 2, 3 and 4 respectively. It therefore becomes possible to define a pathogen strain in terms of which HLA types are capable of displaying the particular motifs found at its two antigenic sites (e.g. "strain 2,8 of pathogen 1" – which expresses peptides that can be bound by HLA molecules 2 and 8). We restrict the number of possible variants at each of the antigenic sites in the model to 4. Our model must allow for the possibility that the molecular properties which allow a pathogen peptide to be displayed by a particular HLA molecule might be shared by peptides from a different species of pathogen, since the fact that a particular HLA molecule might be involved in making an effective response to more than one pathogen species is the focus of this investigation. Thus, as shown in the table, HLA molecule 2 is capable of displaying a peptide from pathogen 1 antigenic site *i*, and from pathogen 2 antigenic site *j*. However, crucially, our model does not assume that these two peptides are identical – merely that they can both be displayed by HLA molecule 2. An adaptive immune response to a peptide from pathogen 1 displayed by HLA molecule 2 therefore only confers lifelong protection against infection with other strains of pathogen 1 which display the peptide that can be displayed by HLA molecule 2. HLA molecules 2,3,4,8 and 9 can all display peptides from either pathogen. HLA molecules 1,6,5 and 7 can only display a peptide from one or other pathogen species (see underlining in the third column).

Pathogen	**Antigenic site**	**HLA molecules which can present peptides from different variants at this site.**
1_{ij}	*i*	<u>1</u>,2,3,4
	j	<u>6</u>,7,8,9
2_{ij}	*i*	3,4,<u>5</u>,7
	j	2,8,9,<u>10</u>

747 **Table 2: Different events that could take place within each time step and the probability of**

748 **each.**

Event within the model	Probability	Notes
Any host not already infected with pathogen K_{ij}, and not already immune to either K_i or K_j, becomes infected with pathogen K_{ij}.	$\frac{\beta_K H_{K_{ij}}}{N}$	Where $H_{K_{ij}}$ = the total number of hosts that were already infected with pathogen K_{ij} as the population entered that timestep; N= the total number of hosts in the population and β_K = a transmission parameter such that in a population where no hosts have a genetic susceptibility to death from infection the basic reproductive number of pathogen K would be equal to $\frac{\beta_K}{\sigma_K}$.
Any host already infected with pathogen K_{ij} recovers from infection with that pathogen.	σ_K	For simplicity, recovery rate depends only on the pathogen species (K), not the strain (ij)
Any host already infected with pathogen K_{ij}, for which none of the HLAs in that host's genome can display either i or j, dies from the infection	θ_k	
Any host dies from a random cause	μ	This term represents all other causes of death, including death from old age.
Adult female host reproduces	ϖ	If a female host reproduces, a male partner is chosen at random and an offspring genotype is generated via Mendelian inheritance. A new individual with this genotype is then added to the population. If the population size is already 2000 then the new member of the population replaces a randomly chosen pre-existing member.
Migration of a new individual into the population	α	When this event occurs a single new individual (with a randomly generated HLA genotype and infected with randomly generated genotypes of both pathogens) replaces an existing member of the population.
As a new infection takes place, the variant at one of the two antigenic sites on the pathogen strain in question is replaced by a randomly chosen variant from the four which are allowed to exist at that site.	m	This simulates pathogen mutation, but implicitly assumes that the 4 peptide variants allowed at each pathogen antigenic site are limited by fitness constraints – so mutation to variants other than these is impossible.

Recombination occurs between the two HLA loci in the host	*r*	Each individual's genotype is explicitly simulated, making it possible to simulate recombination between maternal and paternal chromosomes in either the mother or the father when determining the chromosome that gets passed on to an offspring genotype during reproduction.

749

750