TISSUE ANTIGENS
IMMUNE RESPONSE GENETICS

Tissue Antigens ISSN 0001-2815

A limit to the divergent allele advantage model supported by variable pathogen recognition across HLA-DRB1 allele lineages

Q. Lau[1], Y. Yasukochi[2] & Y. Satta[1]

1 Department of Evolutionary Studies of Biosystems, Sokendai (The Graduate University for Advanced Studies), Kanagawa, Japan
2 Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan

Key words
allele divergence; human leukocyte antigen; major histocompatibility complex; pathogen recognition

Correspondence
Yoko Satta
Department of Evolutionary Studies of Biosystems
Sokendai (The Graduate University for Advanced Studies)
Kanagawa
Japan
Tel: +81 46 858 1610
Fax: +81 46 858 1544
e-mail: satta@soken.ac.jp

Received 9 March 2015; revised 24 August 2015; accepted 26 August 2015

doi: 10.1111/tan.12667

Abstract

Genetic diversity in human leukocyte antigen (HLA) molecules is thought to have arisen from the co-evolution between host and pathogen and maintained by balancing selection. Heterozygote advantage is a common proposed scenario for maintaining high levels of diversity in HLA genes, and extending from this, the divergent allele advantage (DAA) model suggests that individuals with more divergent HLA alleles bind and recognize a wider array of antigens. While the DAA model seems biologically suitable for driving HLA diversity, there is likely an upper threshold to the amount of sequence divergence. We used peptide-binding and pathogen-recognition capacity of *DRB1* alleles as a model to further explore the DAA model; within the *DRB1* locus, we examined binding predictions based on two distinct phylogenetic groups (denoted group A and B) previously identified based on non-peptide-binding region (PBR) nucleotide sequences. Predictions in this study support that group A allele and group B allele lineages have contrasting binding/recognition capacity, with only the latter supporting the DAA model. Furthermore, computer simulations revealed an inconsistency in the DAA model alone with observed extent of polymorphisms, supporting that the DAA model could only work effectively in combination with other mechanisms. Overall, we support that the mechanisms driving HLA diversity are non-exclusive. By investigating the relationships among HLA alleles, and pathogens recognized, we can provide further insights into the mechanisms on how humans have adapted to infectious diseases over time.

Introduction

Genes of the major histocompatibility complex (MHC) encode membrane-bound glycoproteins that bind and present specific self and non-self peptides to T lymphocytes, and thus have an important role in the adaptive immune system. In humans, the MHC is referred to as the human leukocyte antigen (HLA) system, with six classical loci split into class I (HLA-A, B, C) and class II (HLA-DR, DQ, DP) which present intracellular and extracellular antigens, respectively. HLA genes are one of the most polymorphic in the human genome (1), due to a highly variable peptide-binding region (PBR), and one of the primary driving forces of HLA diversity is maintenance by pathogen-mediated balancing selection (2, 3).

There are three main non-exclusive mechanisms proposed for pathogen-driven polymorphism at HLA, as reviewed by Spurgin and Richardson (4): heterozygote advantage (overdominance), negative frequency-dependent selection, and fluctuating selection. The heterozygote advantage model with overdominant selection proposes that individuals heterozygous at HLA respond to a higher diversity of antigens due to co-dominant expression of HLA genes, resulting in higher fitness than homozygotes (5–7). The negative frequency-dependent selection or rare allele advantage model suggests that selection drives pathogen resistance to common HLA alleles, and newly emerging alleles offer protection and are selected for, and the changes in allele frequencies over time results in a cyclical co-evolutionary arms race between HLA alleles and pathogens (8). The third model, fluctuating selection, proposes that HLA diversity is maintained by allelic selection across time and/or space in response to changes in pathogen type and levels (9).

In this study, we focus on examining the model of overdominant selection (heterozygote advantage) and an extension (or modification) of this called divergent allele advantage (DAA) hypothesis. The symmetrical overdominance model proposes that all heterozygotes have an equal fitness or selective advantage over homozygotes (10). The DAA hypothesis

assumes asymmetrical balancing selection which accounts for high divergence in allele sequences and suggests that heterozygotes vary in their fitness advantage. This model was first proposed by Wakeland et al. (11) suggesting that individuals with more diverse MHC alleles could respond to a larger range of antigens, resulting in maintenance of highly divergent alleles within populations. Support for the DAA has come through studies of sequences, computer-based binding prediction (12), and even free-ranging wildlife populations (13, 14).

Using computer-based peptide binding prediction algorithms for class II alleles, Lenz (12) found that an increase in sequence divergence within HLA-DRB1 allele pairs was associated with fewer overlapping ('joint') antigens and wider overall ('union') range of antigens bound, to support the DAA model. Sequence-based analysis of balancing selection in human HLA loci has also supported the DAA model, with evidence of a high proportion of divergent alleles maintained within populations (15) and deviations in allelic branch lengths (16). However, this could be attributed to heterogeneity in selection intensities between allele lineages associated with changes in PBR substitution rates over time (17).

Using the DRB1 locus as a study model, two distinct phylogenetic groups (denoted group A and B) have been identified based on non-PBR nucleotide sequences with an estimated divergence time of 28 MYA for group A and 41 MYA for group B (17, 18) (Figs. S1 and S2, Supporting Information). Group A forms a monophyletic group including allele lineages *HLA-DRB1*03*, **08*, **10*, **11*, **12*, **13*, and **14*. This group appears to be human-specific, attributed to a loss in other primates or retainment in the human lineage for a long time without expansion since species divergence (18). Group B forms a polyphyletic group with primates, suggestive of trans-species polymorphism, and includes allele lineages *HLA-DRB1*01*, **02*, **04*, **07, *09, *15* and **16*. Slow nucleotide substitution rates in the PBR, possibly important for retaining binding affinity for specific peptides, were found prominently in group B allele lineages and absent in group A allele lineages (17). These differences in selection intensities between groups are suggestive of fluctuating selection intensity across allele lineages and groups.

In this study, we focus on the DRB1 locus, since the evolutionary manner of allele lineages have been elucidated (17, 18), and expand on the study by Lenz (12) to investigate the relationship between predicted peptide binding repertoire and sequence divergence in allele pairs. We also consider additional factors: (a) MHC or HLA molecules bind processed antigenic fragments; (b) a selection limit to sequence divergence; and (c) a simple definition of fitness to consider pathogen recognition in addition to peptide binding. Firstly, the MHC presents small processed peptides rather than entire antigenic proteins, and in MHC class II, these are formed in the cytosol by protein degradation by cathepsins (19) into fragments of 15 to 24 amino acid residues (20, 21). Secondly, optimized rather than maximum divergence of alleles within an individual has been proposed to provide greatest fitness gain, whereby too high MHC diversity would result in reduced T-cell repertoire during thymic maturation (22, 23), and this theory has been supported empirically (24, 25). While these studies consider the optimal number of alleles across multiple MHC loci in an individual, it may be applicable to sequence divergence between two alleles within a locus; whereby an increase of divergence within an allele pair beyond a threshold may denote no fitness advantage. In fact, pairs of alleles with intermediate sequence diversity is seen to be the most common empirically in adult reproductive fitness and mate choice in wildlife (14, 26) and could be applicable to antigen recognition. Finally, we consider fitness as the number of pathogens 'recognized' rather than peptides bound: in a generalized viewpoint, provided that a specific MHC allele can still 'recognize' a pathogen regardless of whether it binds one or multiple antigens within a pathogen. Therefore, while the number of peptides bound by an allele pair may increase with sequence divergence, the number of pathogens recognized could reach a limit. We also independently examined group A and group B alleles, an example group classification within DRB1, to investigate intra-locus variation in peptide binding or pathogen recognition capacity.

In addition to prediction of peptide recognition by HLA molecules, we also perform computer simulations to compare symmetrical overdominance with DAA models of HLA polymorphisms, as well as intermediate models. Under the DAA model, we assume that the number of differences at the selection target sites (PBR) of an allelic pair is a direct and only determinant of individual fitness. We hypothesize that while the DAA model may explain the mechanism of balancing selection, there is an upper limit to sequence divergence, in that there is a 'balanced' symmetrical overdominance attributed to selective or evolutionary heterogeneity across allele lineages within a single HLA locus.

Methods

Data for HLA-DRB1 binding prediction

MHC binding prediction was assessed in peptides derived from 54 extracellular pathogens that were selected as a broad representative of disease-associated pathogens that humans may be exposed to (Table S1). Examples of pathogens selected include *Plasmodium falciparum* (malaria parasite), *Mycobacterium tuberculosis, Bordetella pertussis, Helicobacter pylori*, and 10 common pathogens associated with healthcare-associated infections (27). A number of studies have identified associations between MHC class II (MHC II) with disease susceptibility, resistance or severity caused by *M. tuberculosis* (28) and *P. falciparum* (9, 29). In addition, there are reports of MHC II-restricted immune response to vaccination with peptides derived from *B. pertussis* [pertussis toxin subunit (30)] as well as increased expression of MHC II DR loci in *H. pylori*-positive gastritis patients (31).

From each pathogen, we selected one or two protein sequences (Table S1) that were either (i) surface antigens that could activate immune effector cells, or (ii) proteins with epitopes known to be presented by MHC class II, for example malarial merozoite surface protein (32). To simulate the degradation of antigenic proteins by cathepsin enzymes within antigen processing cells (19), a Biopython program was written to obtain peptides that were 15–24 amino acid residues in length and cleaved at linkages between hydrophobic amino acids (33). Following these criteria of protein cleavage resulted in only four to seven peptides per pathogen with a size of 15 to 24 amino acid residues, and a total of 265 peptides (Table S1).

Computational HLA-DRB1 binding prediction

The class II binding prediction algorithm NetMHCIIpan-3.0 (NetMHCIIpan), a pan-specific binding prediction method for all HLA loci based on artificial neural networks (34), was selected due to recommended prediction performance (35). We considered a second algorithm based on matrices for 49 HLA-DRB1 alleles, ProPred (36, 37), used previously (12) and in a preliminary study here (Fig. S3), but it was excluded due to similar results with NetMHCIIpan and limited DRB1 alleles available. For the 265 peptides, binding prediction was performed in NetMHCIIpan for 73 HLA-DRB1 alleles listed as 'common' in the catalogue of common and well-documented (CWD) HLA alleles [Table S2 (38)]. Only 'common' alleles were used because not every single allele could be used for binding prediction, and selection is more prominent in frequent alleles. An affinity threshold of half maximal inhibitory concentration (IC_{50}) = 500 nM was used for bound peptides in NetMHCIIpan, representative of weak binding, while strong binding (IC_{50} = 50 nM) was also considered (39). The number of 'peptides bound' by each of the 73 HLA-DRB1 alleles was compiled, and all peptides within a single pathogen was summed into number of 'pathogens recognized' by each allele, whereby a specific allele can 'recognize' a pathogen if it can bind one of its 4–7 peptides. We then formed a matrix of allele pairs for total combined number ('union') or proportion of common ('joint') peptides bound/ pathogens recognized by each pair, and the association with proportion of differences in amino acid residues at the PBR, referred to as p-distance at PBR or D_{PBR}. We used newly defined PBR sites that were adapted from Brown et al. (21) and based on new evidence of changes in sites inferred from crystallographic data (S. Kusano and S. Yokoyama, personal communication; residues 9, 11, 13, 26, 28, 30, 37, 78, 47, 56, 57, 60, 61, 67, 70, 71, 74, 77, 78, 81, 82, 85, 86, 88, 89, and 90); the two definitions share 80.8% (21 of 26 residues) similarity and have comparable results (Fig. S4). The relationship between D_{PBR} with either 'union' or 'joint' was assessed by quadratic regression analysis using R version 3.0.2 (http://www.r-project.org) and Spearman correlation test using GraphPad Prism 6.04 (GraphPad Software, La Jolla, CA).

The relationship between the two aspects of binding/recognition capacity ('union' and 'joint') and sequence divergence at the PBR was further assessed to look for variation within a locus. We used a previously identified within-DRB1 group classification (group A and B alleles) (18) and examined binding/recognition capacity independently for the two groups. Of the 73 'common' HLA-DRB1 alleles used, 37.0% were group A alleles ($n = 27$) and 63.0% were group B alleles ($n = 46$). In addition, the allele frequencies of all HLA-DRB1 alleles used in this study were surveyed using the allele frequency net database (40) across 26 populations from Africa (northern and Sub-Saharan), Asia or Europe (Table S3).

Simulation tests to compare selection models

We performed simulations of three different models of balancing selection: symmetrical (symmetrical overdominance: SOD), asymmetrical (DAA) and modified DAA (MDAA) models. Computer simulation was conducted by considering a diploid population of effective size N ($N = 1000$) under the assumption of random mating and no recombination. Each gene consists of regions comprised of 50 target sites of selection and 1000 linked neutral sites. In each generation, mutation was introduced to both target sites and linked neutral sites and 2N genes were chosen at random after selection. Mutation was introduced at a rate of M ($M = 2N\mu = 0.04$, 0.4, or 4.0) mutations per target region per population per generation. Each site had a state of integer of 1, 2, 3, or 4, and mutation at each site was assumed to produce a new state that always differed from the existing one. Selection and sampling of genes were conducted at the same time. In each generation, we chose two genes at random with replacement from the parental gene pool to form a zygote, and this zygote was subjected to selection with a given probability (ps), which depended on the selection scheme. We first generated a uniform random number, x, and if $x < ps$, the zygote was chosen to be a member of the next generation, otherwise it was discarded. We repeated this process until N zygotes were chosen.

The value of ps was determined differently in each selection scheme or model. Under the SOD model, the ps value assigned to each homozygote was $1 - s$, where s is a selection coefficient, and the ps value for heterozygotes was 1. For the DAA model, first we counted the number of different accumulated mutations between a given pair of genes in an individual i, k_i, and the relative fitness for each individual was determined as $s(i) = (1 + s \times k_i)/S_{MAX}$ if the individual i is a heterozygote, where $S_{MAX} = 1 + s \times k_{MAX}$. k_{MAX} is the maximum number of accumulated mutations in an individual in a parent population. The value of ps for a heterozygote i is $ps = s(i)$, whereas that for a homozygote is $ps = 1/S_{MAX}$. The MDAA model is an intermediate of the SOD and DAA models, and considers mechanisms functioning non-exclusively: the DAA mode of selection functions up to a critical value of $K_{cri,}$ ($k_i \leq K_{cri}$), while SOD type of selection begins if the k_i is over the critical value ($k_i > K_{cri}$).

Within the MDAA model, we tested three critical values of $K_{cri} = 1$ (MDAA1), $K_{cri} = 2$ (MDAA2), and $K_{cri} = 3$ (MDAA3). Higher critical values led to measurements similar to the DAA model and were excluded. In all simulations we used a selection co-efficient of $N_S = 100$ and started with a monomorphic state and discarded the results before 25*N* generations had passed. The measurements based on replicates ($n = 20$) collected every 10N generations of simulations include: (a) the average number of alleles with different sequences at target sites (n_a) and at the entire region (target and neutral sites, n); (b) the average heterozygosity at the target sites (H_B) and at the neutral sites (H_N); (c) the nucleotide diversity at the target sites (π_B) and that at the linked neutral sites (π_N); (d) the relative rate of mutant substitution at target sites (π_B) and at a neutral sites (π_N); and (e) the total load (L_T) based on proportion of individuals passing genes to the next generation and the genetic load (L_G) based on fitness of individuals as defined by Crow (41). Statistical *t*-tests were performed using GraphPad Prism with a significance level of $p < 0.05$.

Results

Peptide binding and pathogen recognition by HLA-DRB1 alleles

From initial examination of peptide binding in all alleles ($n = 73$), regression analyses indicated a linear relationship between D_{PBR} and combined number of peptides bound ('union', Fig. 1A) although there was no significant correlation ($r = 0.035$, $p = 0.072$). Further analyses of 'union' of peptide binding by alleles showed that group A alleles also had no significant correlation with D_{PBR} ($r = -0.003$, $p = 0.922$, Fig. 1B). However, group B alleles had significant positive correlation between 'union' and D_{PBR} ($r = 0.202$, $p < 0.0001$, Fig. 1C), classified as 'low' ($r < 0.2$) according to Evans (42). When considering pathogens recognized instead of peptides bound, similar patterns were identified: D_{PBR} had no significant correlation with combined number of pathogens recognized by allele pairs in all alleles ('union', $r = 0.034$, $p = 0.084$, Fig. 1D) and in group A alleles ($r = -0.025$, $p = 0.427$, Fig. 1E), while group B alleles had significant 'medium' ($0.2 < r < 0.4$) positive correlation between 'union' and D_{PBR} ($r = 0.421$, $p < 0.0001$, Fig. 1F).

We found a non-linear relationship between D_{PBR} and proportion of shared peptides bound or pathogens recognized by an allele pair ('joint', regression analysis $p < 0.001$, Fig. 1G and J, respectively), suggestive of minimum overlap of shared peptides bound at intermediate sequence divergence. This non-linear relationship between 'joint' and D_{PBR} was also present when examining group A or group B alleles independently (Fig. 1h–i, k–l). Relationships between D_{PBR} and either 'union' or 'joint' peptides bound or pathogens recognized were similar when we only considered strong binders of less than 50 nM affinity (Fig. S5) compared to weak binders ($IC_{50} < 500$ nM).

Using the allele frequency net database, we surveyed HLA-DRB1 allele frequencies across Sub-Saharan Africa, northern Africa, Asia and Europe populations ($n = 26$). There is variation in allele frequencies between and within continents, and some alleles that are more frequent in specific regions include *DRB1*15:03* and *DRB1*09:01* in Sub-Saharan Africa and Asia, respectively (Fig. S6). We further examined the frequency of group A and group B allele lineages in 56 populations across Africa, Asia and Europe using the allele frequency net database, whereby frequency data for allele lineages rather than specific alleles was adequate (populations listed in Table S3). We found that populations from Sub-Saharan Africa have a significantly lower cumulative frequency of group B alleles compared with northern Africa, Asia and Europe ($p < 0.0001$, Fig. 2). Furthermore, to check that group A and B allele lineages may be under different selective constraints across regions, we divided DRB1 alleles into two random groups rather than group A or B and found different patterns of cumulative frequencies (Fig. S7).

Computer simulation comparing symmetrical overdominant selection and DAA model

Of the multiple measurements collated from the simulations in Table S4, nucleotide diversity (π) and genetic load (L_G) had marked significant differences between the SOD and DAA models. The nucleotide diversity at both target ($\pi_B = 0.661–0.887$, Fig. 3A) and neutral sites ($\pi_N = 0.128–0.882$) were very high for the DAA model under any condition, and this decreases with the critical value (K_{cri}) in the MDAA models. The total load (L_T) and genetic load (L_G) is heavier in the DAA model compared to all other models ($p < 0.01$) and there is no significant difference between the SOD and MDAA1 ($K_{cri} = 1$) model (Fig. 3B, Table S4). With the exception at high mutation rate condition ($M = 4.0$), the strength of selection (γ), represented by the ratio of nucleotide diversity at target sites to neutral sites ($\gamma = \pi_B/\pi_N$), is highest at the SOD and MDAA models with low K_{cri} values (Fig. 3C). At low mutation rates, the fluctuation in selection is very large even with increased replication, indicating that strength of selection depends strongly on mutation rate. The average number of alleles at the target sites (n_a) are significantly lower in the DAA model compared with SOD and all MDAA models ($p < 0.01$), regardless of rate of mutation (Fig. 3D), and the number of alleles in the entire region (n) is significantly lower in the DAA model only at high and intermediate mutation rates ($p < 0.05$, Table S4). When comparing average heterozygosity between models, there were no strong trends with the exception of marginally lower heterozygosity at target sites (H_B) at some models of DAA ($p < 0.05$, Table S4).

Discussion

While the model of DAA seems biologically suitable for driving MHC polymorphism, there should be an upper

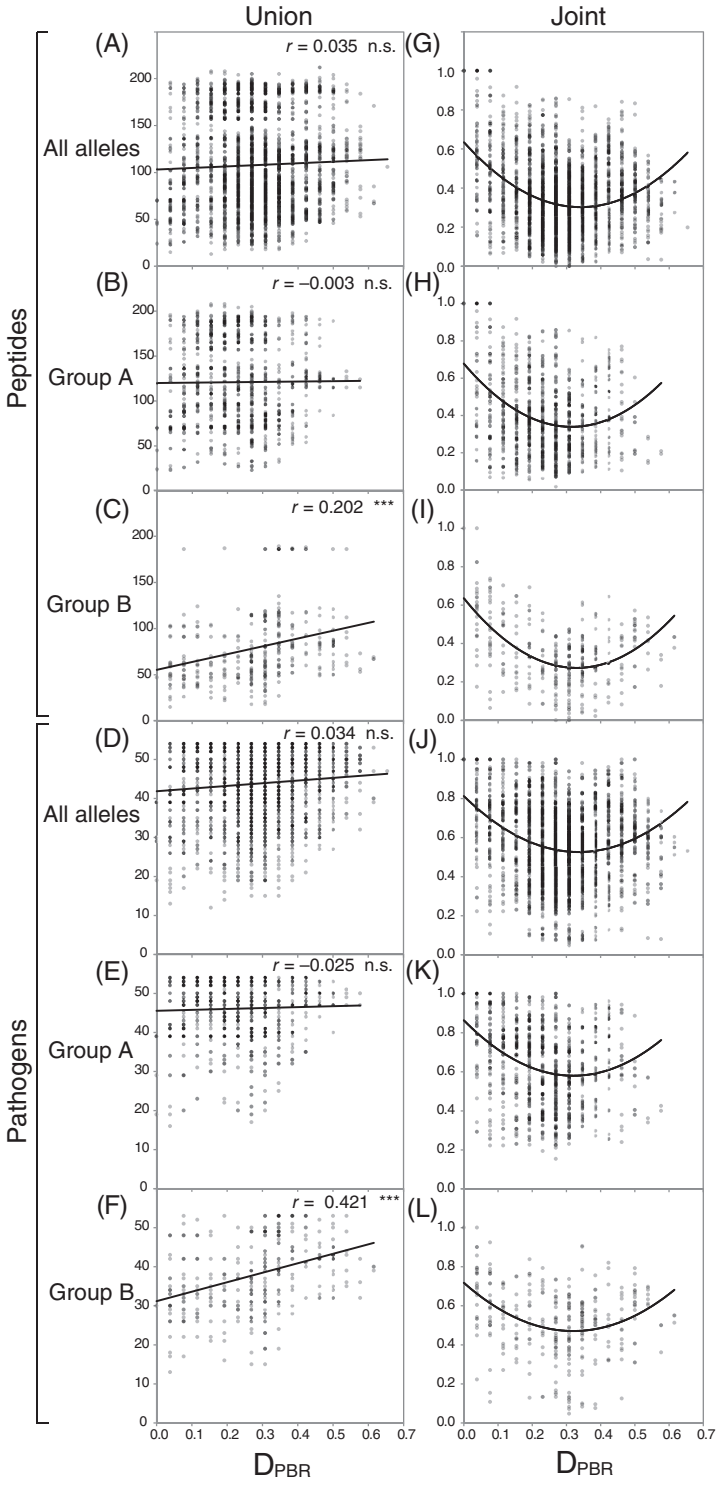

Figure 1 Relationship between major histocompatibility complex (MHC) DRB1 pairwise distance at PBR (D_{PBR}) of allele pairs with total number (union) of peptides bound in (A) all alleles ($n=73$), (B) group A alleles ($n=46$), and (C) group B alleles ($n=27$); or pathogens recognized in (D) all alleles, (E) group A alleles, and (F) group B alleles. The relationship between D_{PBR} with proportion (joint) of (G–I) peptides bound or (J–L) pathogens recognized was also examined. Spearman *r* correlation coefficients and significance are presented in each graph; $***p<0.0001$, n.s. $= p>0.05$. Darker dots represent higher frequency of allele pairs with the same D_{PBR} and union or joint.

limit to sequence divergence. Using two independent computer-based approaches, prediction of pathogen recognition and computer simulations, this study supports that the DAA model alone is not suitable for driving MHC diversity but could function non-exclusively with other mechanisms.

Computer simulations comparing models of HLA polymorphism show that the DAA model could function in combination with other mechanisms of pathogen-mediated selection. The nucleotide diversity generated by the DAA model alone at target or neutral sites is exceedingly high compared with the present polymorphism found in HLA loci, which is less than

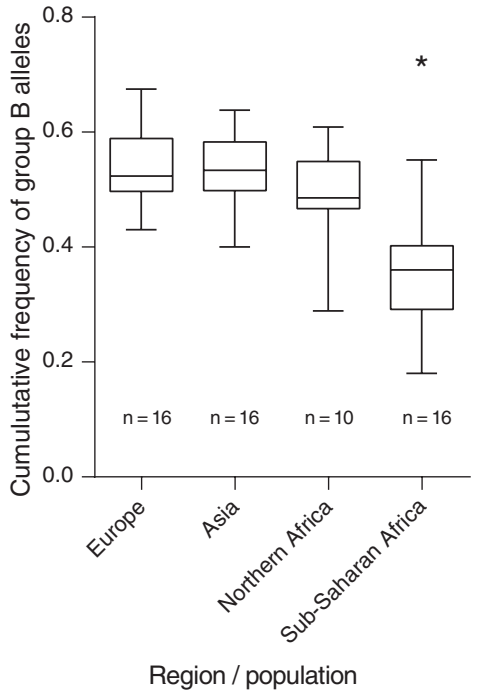

Figure 2 Boxplot of average cumulative frequency of group B alleles comparing different geographic regions: Europe, Asia, northern Africa, and Sub-Saharan Africa. *n* = number of populations from each region.

10% across all sites at various loci (15) and at most 17% at target sites at HLA-DRB1 (Table S5). In addition, the level of total and genetic load generated from the DAA model is too high, whereby it is not biologically sustainable for 10–25% (represented by load from the DAA model) of individuals to be replaced in each generation. On the other hand, the MDAA models support that an intermediate between the SOD model and DAA model is feasible for adequate generation of alleles and heterozygosity without excess of nucleotide diversity or genetic load. While we have shown that the DAA model could function non-exclusively with symmetrical overdominant selection, it may also work with other mechanisms such as negative selection to remove over-divergent alleles or haplotypes that cause too high genetic load in order to maintain an appropriate level of TCR repertoire.

Using the NetMHCIIpan MHC binding prediction algorithm with a biologically sound approach adapted from Lenz (12), we have examined the relationship between sequence divergence with two aspects: combined total ('union') and proportion of shared ('joint') peptides bound or pathogens recognized by an allele pair. Both aspects support that there is limit to the DAA model, albeit in different manners. (a) The total number of peptides bound or pathogens recognized ('union') increased with HLA-DRB1 allele sequence divergence only within a sub-group of the locus (Group B alleles), while there was no correlation when all alleles were considered. (b) For 'joint', we found that a low proportion of the same peptides or pathogens were bound/recognized by an allele pair at intermediate sequence divergence. These two aspects of binding/recognition capacity align with computer simulations in supporting that the DAA model can only function non-exclusively.

The model of optimal fitness at intermediate sequence divergence has been seen empirically in reproductive fitness and could also apply to pathogen recognition to fit the common understanding that an excessively high MHC sequence divergence could reduce the T-cell repertoire during thymic maturation. Increasing sequence divergence might be ideal for peptide binding advantage, while negative selection is required to remove alleles that recognise self-peptides, resulting in a blind-spot; thus too many diverged MHC genes will result in an excessive blind spot (43). While the NetMHCIIpan peptide binding algorithm cannot account for the possible disadvantages of MHC-based T-cell selection, the current HLA diversity may be a consequence of trade-offs in many evolutionary mechanisms such as DAA with negative selection against alleles that are too divergent, and here it appears this component of fitness supports an optimum at intermediate sequence divergence. It may be possible that at high sequence divergence, the proportion of peptides bound or pathogens recognized by an allele pair overlaps in the blind-spot. Alternatively, a balance in sequence divergence could be related to different allele lineages with heterogeneous substitution rates, which may be the case for group A and group B alleles within the HLA-DRB1 locus.

The pattern of pathogen recognition in group B alleles ('union' aspect) seems to support the DAA model, whereby there is an increase in total number of pathogens recognized by an allele pair with increasing sequence divergence. In contrast, the group A alleles do not have a significant correlation between sequence divergence and pathogen recognition. We checked for potential bias of peptide binding prediction in NetMHCIIpan for either group A or B alleles, and found similar proportion of alleles used to develop prediction tools between the groups (Table S6). The differences in recognition capacity between the two groups are consistent with the differences in substitution rates, whereby most group B allele lineages have low levels of nonsynonymous substitution rates while no group A alleles lineages have low levels of nonsynonymous rates (17). So while group B alleles follow the pattern of the DAA model, there are reduced effects of selection in this lineage, evident through low nonsynonymous substitution rates and our computer simulations demonstrating that the DAA model is less effective for the ratio of nonsynonymous to neutral substitution rates (strength of selection, Fig. 3C).

Interestingly, we found that the cumulative frequency of group B alleles is significantly lower in Sub-Saharan Africa relative to other regions and continents, including northern Africa. Modern humans dispersed from Sub-Saharan Africa about 100,000 to 50,000 years ago (44), and HLA variation in migrating populations was probably shaped by different evolutionary forces associated with novel pathogens and

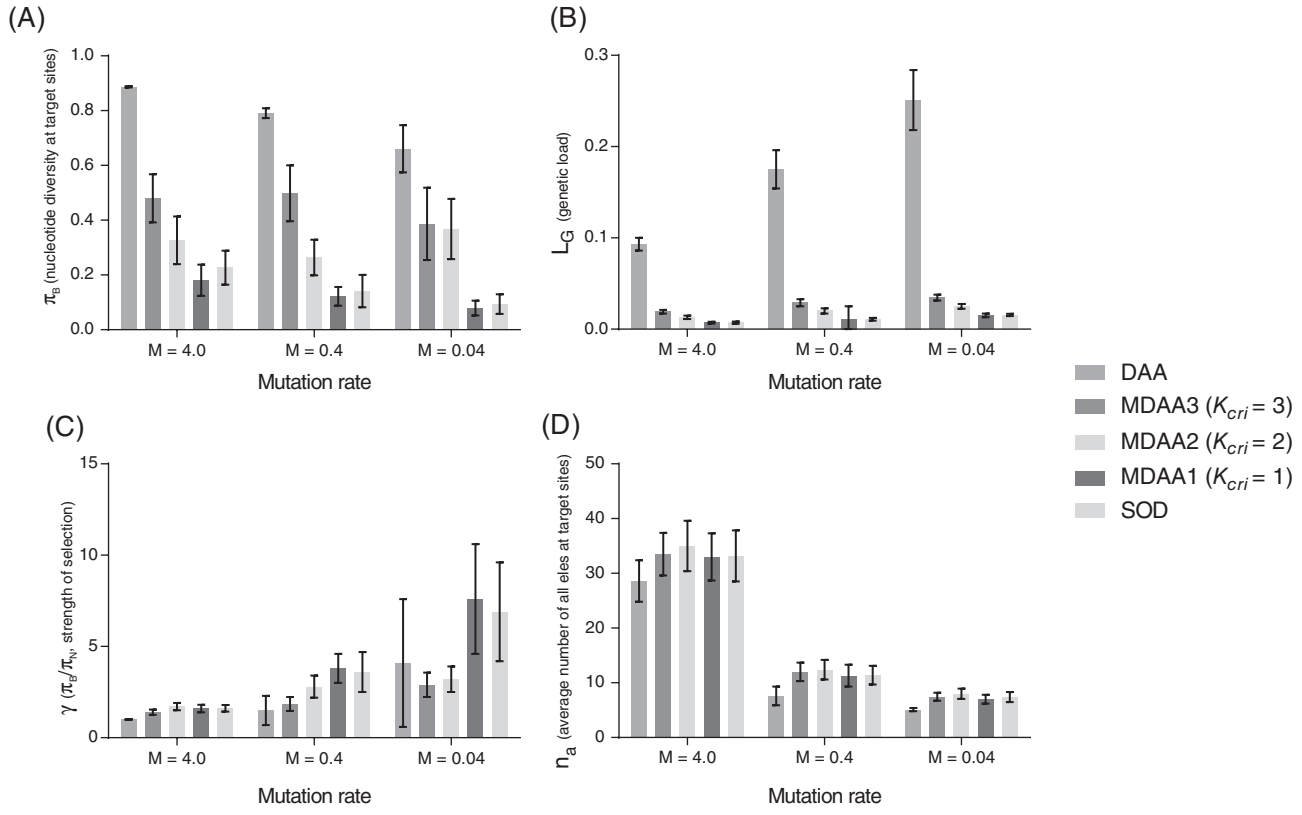

Figure 3 Computer simulations comparing MHC polymorphism models: the symmetrical overdominance (SOD) model, the divergent allele advantage (DAA) model, and modified DAA (MDAA) models where DAA selection occurs up to various critical values (K_{cri}) before SOD selection occurs. Measurements (mean, SD) include (A) nucleotide diversity at target sites (π_B), (B) genetic load (L_G), (C) strength of selection (γ) based on the ratio of nucleotide diversity at target to neutral sites, and (D) average number of alleles at target sites (n_a).

environments (45), and this change in niche (deterministic forces) may have further driven pathogen-associated selection. Meanwhile, the native human populations in Sub-Saharan Africa may have experienced continuous exposure with indigenous pathogens. Malaria, as an example, has had sufficient time (5000–10,000 years) to shape the MHC structure in Sub-Saharan Africa without major environmental changes associated with migration (46). Stronger selective pressure has targeted specific HLA loci including the HLA B gene (47), thus it is possible that novel pathogens outside of Africa have also targeted different allele lineages such as the group B allele lineage within the HLA-DRB1 locus. Apart from deterministic forces, demographic and stochastic forces are likely to have played an important role in shaping HLA diversity. It is possible that admixture with archaic hominids in Europe resulted in higher cumulative group B allele frequencies in Asia and Europe, although only class I haplotypes have been characterized to-date in ancestral species (48).

As HLA alleles may have co-evolved with multiple pathogens, we surveyed whether peptides bound by HLA alleles could be shared across different pathogens. We collated all 9-mers within the 15–24 amino acid peptides predicted to be bound by HLA-DRB1 in NetMHCIIpan, and searched for shared 9-mers between pathogens with 78–100% identity at non-anchor positions (P2, P3, P5, P8). Several shared 9-mers were identified between pathogens used in this study (Table S7), most of which were between pathogens of the same genus (*Bartonella* spp., *Clostridium* spp., and *Klebsiella* spp.), or closely related species (for example, *Enterobacter aerogenes* with *Klebsiella* spp.). The overlap in peptides bound by HLA alleles across pathogens does not exclude the possibility that HLA alleles have simultaneously evolved with multiple pathogens.

Diversity at HLA genes is driven by many non-exclusive mechanisms and not usually a single mechanism alone: for example, heterozygote advantage alone cannot explain the high HLA diversity, with the exception where fitness contributions of alleles are similar (49). Similarly, we show from two computational approaches that the DAA model alone cannot explain the observed HLA diversity and predicted binding and recognition capacity. Such diversity is probably the result of synchrony between mechanisms including heterozygote advantage, DAA, and frequency-dependent selection.

Conclusion

Using the DRB1 locus of HLA as a model for studying co-evolution of host and pathogen, we show evidence of increased peptide binding and pathogen recognition capacity of group B alleles (following the DAA model) which appear to be counterbalanced by group A alleles that have no changes with divergence. The variable pathogen recognition between lineages within a locus may provide a mechanism for preserving high HLA variation while avoiding negative selection during T-cell maturation. This is further supported by computer simulations showing that the DAA model can only function to drive allelic diversity and heterozygosity up to a certain threshold. Future studies will further explore the possible role of variable allele lineages within HLA that may maintain diversity and pathogen-recognition capacity, as well as the possible role of specific DRB1 allele lineages in minimizing negative selection during thymic maturation.

Acknowledgements

This project was supported by JSPS KAKENHI grant number 22133007. We wish to thank Naoyuki Takahata for checking of the manuscript, Hisashi Ohtsuki for assistance with statistical analyses, and Ken Nagata for assistance with writing Biopython programs. QL wrote the manuscript, performed binding predictions and analyses. YY designed the binding prediction experiment and assisted in writing the manuscript. YS designed and performed the computer simulations and assisted in writing the manuscript.

Conflict of interest

The authors have declared no conflicting interests.

References

1. Weitzman JB. Sequence of the major histocompatibility complex. *Genome Biol* 2000: **1**: reports021. DOI: 10.1186/gb-2000-1-1-reports021
2. Bernatchez L, Landry C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J Evol Biol* 2003: **16**: 363–77.
3. Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet* 1998: **32**: 415–35.
4. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc R Soc Biol Sci Ser B* 2010: **277**: 979–88.
5. Hughes AL, Nei M. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci USA* 1989: **86**: 958–62.
6. Doherty PC, Zinkernagel RM. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* 1975: **256**: 50–2.
7. Takahata N, Nei M. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 1990: **124**: 967–78.
8. Slade RW, McCallum HI. Overdominant vs. frequency-dependent selection at MHC loci. *Genetics* 1992: **132**: 861–4.
9. Hill AVS. HLA associations with malaria in Africa: some implications for MHC evolution. In: Klein J, Klein D, eds. *Molecular Evolution of the Major Histocompatibility Complex. Berlin Heidelberg*: *Springer*, *1991*, *403–20*.
10. Takahata N. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc Natl Acad Sci USA* 1990: **87**: 2419–23.
11. Wakeland EK, Boehme S, She JX et al. Ancestral polymorphisms of MHC class II genes: divergent allele advantage. *Immunol Res* 1990: **9**: 115–22.
12. Lenz TL. Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution* 2011: **65**: 2380–90.
13. Richman AD, Herrera LG, Nash D. MHC class II beta sequence diversity in the deer mouse (*Peromyscus maniculatus*): implications for models of balancing selection. *Mol Ecol* 2001: **10**: 2765–73.
14. Lenz TL, Mueller B, Trillmich F, Wolf JB. Divergent allele advantage at MHC-DRB through direct and maternal genotypic effects and its consequences for allele pool composition and mating. *Proc R Soc Biol Sci Ser B* 2013: **280**: 20130714.
15. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PLoS One* 2011: **6**: e14643.
16. Bronson PG, Mack SJ, Erlich HA, Slatkin M. A sequence-based approach demonstrates that balancing selection in classical human leukocyte antigen (HLA) loci is asymmetric. *Hum Mol Genet* 2013: **22**: 252–61.
17. Yasukochi Y, Satta Y. Nonsynonymous substitution rate heterogeneity in the peptide-binding region among different HLA-DRB1 lineages in humans. *G3 Genes Genom Genet* 2014: **4**: 1217–26.
18. Yasukochi Y, Satta Y. A human-specific allelic group of the MHC DRB1 gene in primates. *J Physiol Anthropol* 2014: **33**: 14.
19. Bennett K, Levine T, Ellis JS et al. Antigen processing for presentation by class II major histocompatibility complex requires cleavage by cathepsin E. *Eur J Immunol* 1992: **22**: 1519–24.
20. Chicz RM, Urban RG, Lane WS et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* 1992: **358**: 764–8.
21. Brown J, Jardetzky T, Gorga J et al. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 1993: **364**: 33–9.
22. Nowak MA, Tarczy-Hornoch K, Austyn JM. The optimal number of major histocompatibility complex molecules in an individual. *Proc Natl Acad Sci USA Biol Sci* 1992: **89**: 10896–9.
23. Woelfing B, Traulsen A, Milinski M, Boehm T. Does intra-individual major histocompatibility complex diversity keep a golden mean? *Proc R Soc Biol Sci Ser B* 2009: **364**: 117–28.

24. Wegner KM, Reusch TBH, Kalbe M. Multiple parasites are driving major histocompatibility complex polymorphism in the wild. *J Evol Biol* 2003: **16**: 224–32.
25. Madsen T, Ujvari B. MHC class I variation associates with parasite resistance and longevity in tropical pythons. *J Evol Biol* 2006: **19**: 1973–8.
26. Forsberg LA, Dannewitz J, Petersson E, Grahn M. Influence of genetic dissimilarity in the reproductive success and mate choice of brown trout – females fishing for optimal MHC dissimilarity. *J Evol Biol* 2007: **20**: 1859–69.
27. Hidron AI, Edwards JR, Patel J et al. NHSN annual update: antimicrobial-resistant pathogens associated with healthcare-associated infections: annual summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2006–2007. *Infect Control Hosp Epidemiol* 2008: **29**: 996–1011.
28. Kettaneh A, Seng L, Tiev KP, Tolédano C, Fabre B, Cabane J. Human leukocyte antigens and susceptibility to tuberculosis: a meta-analysis of case–control studies. *Int J Tuberc Lung Dis* 2006: **10**: 717–25.
29. Lyke KE, Fernandez-Vina MA, Cao K et al. Association of HLA alleles with *Plasmodium falciparum* severity in Malian children. *Tissue Antigens* 2011: **77**: 562–71.
30. Fagerberg J, Askelof P, Wigzell H, Mellstedt H. Induction of CD4(+) and CD8(+) Bordetella pertussis toxin subunit S1 specific T cells by immunization with synthetic peptides. *Cell Immunol* 1999: **196**: 110–21.
31. Wee A, Teh M, Kang JY. Association of *Helicobacter pylori* with HLA-DR antigen expression in gastritis. *J Clin Pathol* 1992: **45**: 30–3.
32. Vargas LE, Parra CA, Salazar LM, Guzman F, Pinto M, Patarroyo ME. MHC allele-specific binding of a malaria peptide makes it become promiscuous on fitting a glycine residue into pocket 6. *Biochem Biophys Res Commun* 2003: **307**: 148–56.
33. Impens F, Colaert N, Helsens K et al. A quantitative proteomics design for systematic identification of protease cleavage events. *Mol Cell Proteomics* 2010: **9**: 2327–33.
34. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 2013: **65**: 711–24.
35. Huang Lin H, Lan Zhang G, Tongchusak S, Reinherz E, Brusic V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformat* 2008: **9**: 1–10.
36. Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 2001: **17**: 1236–7.
37. Sturniolo T, Bono E, Ding J et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 1999: **17**: 555–61.
38. Mack SJ, Cano P, Hollenbach JA et al. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* 2013: **81**: 194–203.
39. Sette A, Vitiello A, Reherman B et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J Immunol* 1994: **153**: 5586–92.
40. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res* 2011: **39**: D913–9.
41. Crow JF. Some possibilities for measuring selection intensities in man. *Hum Biol* 1958: **30**: 1–13.
42. Evans JD. *Straightforward Statistics for the Behavioral Sciences*. *Pacific Grove*: *Brooks/Cole Publishing*, *1996*.
43. Takahata N. MHC diversity and selection. *Immunol Rev* 1995: **143**: 225–47.
44. White TD, Asfaw B, DeGusta D et al. Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* 2003: **423**: 742–7.
45. Mboowa G. Genetics of Sub-Saharan African human population: implications for HIV/AIDS, tuberculosis, and malaria. *Int J Evol Biol* 2014: **2014**: 8.
46. Parikh S, Rosenthal PJ. Human genetics and malaria: relevance for the design of clinical trials. *J Infect Dis* 2008: **198**: 1255–7.
47. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol* 2005: **15**: 1022–7.
48. Abi-Rached L, Jobin MJ, Kulkarni S et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science (New York, NY)* 2011: **334**: 89–94.
49. De Boer R, Borghans JM, van Boven M, Keşmir C, Weissing F. Heterozygote advantage fails to explain the high degree of polymorphism of the MHC. *Immunogenetics* 2004: **55**: 725–31.

Supporting Information

The following supporting information is available for this article:

Table S1. Antigens selected from 54 pathogens with a total of 265 peptides with 15–24 a.a. residues which were subsequently used for HLA-DRB1 binding prediction.

Table S2. Common HLA-DRB1 alleles ($n = 73$) used in NetMHCIIpan-3.0 recognition prediction programs. Common alleles are those listed as 'common in the CWD catalogue (Mack et al. 2013).

Table S3. Populations surveyed for frequencies of individual alleles or allele lineages (and total group A or B allele frequencies) from the allele frequency net database. Frequencies for individual alleles were available for a subset of populations ($n = 26$, denoted by *).

Table S4. Results from computer simulations for five different major histocompatibility complex (MHC) polymorphism models (SOD, DAA, and MDAA at three levels).

Table S5. Nucleotide diversity at target sites (πB) and non-target sites (π) in human leukocyte antigen (HLA) loci.

Table S6. To account for any prediction bias towards either group A or group B, we compared the number of alleles with experimental affinity data used for developing NetMHCIIpan compared to the number of alleles used in this study.

Table S7. Of the 265 peptides of 15–24 amino acid length derived from 54 pathogens, all 9-mers recognised by DRB1 alleles were compared between pathogens to identify core

9-mers that were similar (78-100%) at non-anchor positions (P2, P3, P5, P8).

Fig. S1. Re-published Fig. 3 from Yasukochi and Satta (17): phylogenetic analysis showing group A and group B allele lineages within the DRB1 locus, and evidence of trans-species polymorphism in group B alleles. Source: Yasukochi and Satta (17).

Fig. S2. Re-published Fig. 5 from Yasukochi and Satta (17): phylogenetic analysis showing group A and group B allele lineages within the DRB1 locus, and differences in nonsynonymous substitution rates between groups. Source: Yasukochi and Satta (17).

Fig. S3. In a preliminary study, we compared binding prediction using two algorithms, ProPred and NetMHCIIpan3.0 using 49 HLA-DRB1 alleles (51% of alleles in common between algorithms). We found comparable results between the two algorithms (results here for recognition of ten pathogens), similar to Lenz (12), and thus all subsequent studies used only NetMHCIIpan3.0 which allows more alleles to be assessed.

Fig. S4. In this study, we used PBR sites defined by Kusano and Yokoyama (personal communication) which yielded similar results to when we used the PBR sites defined by Brown et al (21). An example here is the association between DPBR with 'union' or 'joint' pathogens bound.

Fig. S5. Strong binders only (binding threshold 50 nM) show similar results to when 500 nM is used (as in Fig. 1 of manuscript): relationship between MHC DRB1 pairwise distance at PBR (DPBR) of allele pairs with total number (union) of peptides bound in (a) all alleles (n =73), (b) group A alleles (n = 46), and (c) group B alleles (n = 27); or pathogens recognised in (d) all alleles,(e) group A alleles, and (f) group B alleles. The relationship between DPBR with proportion (joint) of (g–i) peptides bound or (j–l) pathogens recognised was also examined. Spearman r correlation coefficients and significance are presented in each graph; $***p < 0.0001$, n.s. = $p > 0.05$.

Fig. S6. Boxplot of allele frequencies of individual DRB1 alleles from populations from Sub-Saharan Africa (n = 5), northern Africa (n = 7), Europe (n = 7) and Asia (n = 7). Frequencies were obtained from the allele frequency net database from populations listed in Table S3 that have individual allele frequencies available.

Fig. S7. Boxplot of cumulative allele frequencies of group A and group B DRB1 alleles]defined by Yoshiki and Satta (17)] from populations from Sub-Saharan Africa (n = 16), northern Africa (n = 10), Europe (n = 16) and Asia (n = 16). We randomly divided alleles into alternative groups and found different patterns of cumulative allele frequencies: Group E (*DRB1*01, 03, 04, 10, 12, 13, 16*) *vs* Group F (*DRB1*07, 09, 08, 11, 14, 15*) and Group G (*DRB1*01, 04, 08, 09, 12, 14, 15*) *vs* Group H (*DRB1*04, 04, 07, 10, 11, 13, 16*). These random divisions resulted in different allele frequency patterns compared to the groups A and B.