Heredity (2011) 106, 727–740

www.nature.com/hdy

ORIGINAL ARTICLE

Genetic structure and contrasting selection pattern at two major histocompatibility complex genes in wild house mouse populations

D Čížková[1], J Gouy de Bellocq[2], SJE Baird[3], J Piálek[1] and J Bryja[1]
[1]*Department of Population Biology, Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Brno, Czech Republic;* [2]*Evolutionary Ecology Group, University of Antwerp, Antwerp, Belgium and* [3]*CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, University of Porto, Porto, Portugal*

The mammalian major histocompatibility complex (MHC) is a tightly linked cluster of immune genes, and is often thought of as inherited as a unit. This has led to the hope that studying a single MHC gene will reveal patterns of evolution representative of the MHC as a whole. In this study we analyse a 1000-km transect of MHC variation traversing the European house mouse hybrid zone to compare signals of selection and patterns of diversification at two closely linked MHC class II genes, *H-2Aa* and *H-2Eb*. We show that although they are 0.01 cM apart (that is, recombination is expected only once in 10 000 meioses), disparate evolutionary patterns were detected. *H-2Aa* shows higher allelic polymorphism, faster allelic turnover due to higher mutation rates, stronger positive selection at antigen-binding sites and higher population structuring than *H-2Eb*. *H-2Eb* alleles are maintained in the gene pool for longer, including over separation of the subspecies, some *H-2Eb* alleles are positively and others negatively selected and some of the alleles are not expressed. We conclude that studies on MHC genes in wild-living vertebrates can give substantially different results depending on the MHC gene examined and that the level of polymorphism in a related species is a poor criterion for gene choice.
Heredity (2011) **106,** 727–740; doi:10.1038/hdy.2010.112; published online 8 September 2010

Keywords: MHC; house mouse; selection; population structure; trans-species polymorphism

Introduction

Patterns of evolution in the major histocompatibility complex (MHC) are of interest because of its function: class I and II genes encode glycoproteins that recognize and present antigens to T cells (Klein, 1986). A Red Queen arms race (Van Valen, 1973) between parasites and the immune system would lead to immune gene alleles experiencing episodes of positive selection. These episodes will tend to (1) reduce population variation in DNA tracts centred on MHC genes and (2) leave signals of positive selection (for example, elevated nonsynonymous to synonymous substitution (dN/dS) ratio) across MHC sequences. The multiplicity of parasites in natural populations may lead to maintenance of multiple MHC variants because of balancing selection based on heterozygote advantage or spatiotemporal variation in selection pressure (for example, reviewed in Garrigan and Hedrick, 2003; Piertney and Oliver, 2006). If such balancing selection is sufficiently strong and persistent, multiple variants may be maintained through speciation events, resulting in trans-species polymorphism (TSP; Klein, 1987) manifested by incongruence between gene and species trees.

The mammalian MHC is tightly linked, and often thought of as inherited as a unit (Klein *et al*., 1991). This has led to the hope that studying a single MHC gene will reveal patterns of evolution representative of the MHC as a whole. Based on this assumption, most studies are restricted to only one MHC gene (see exceptions, for example, in Edwards *et al*., 1997; Sommer, 2003; Bryja *et al*., 2007; Tollenaere *et al*., 2008; Babik *et al*., 2008.). However, there are at least two reasons for analysing more MHC genes. First, evolutionary forces may differ between MHC genes within the complex. For example, Bryja *et al*. (2007) studied selection on two MHC class II genes in vole populations with fluctuating density and found much stronger geographic heterogeneity in selection signal at the *DQA* gene than the *DRB*. Second, differences in selection may also occur between orthologues, that is, genes at the same locus in different species. Nizetiæ *et al*. (1987) demonstrated the absence of *DR* genes in the mole rat, despite *DRB* being the most polymorphic gene in humans (Janeway *et al*., 1999), and suggested that other MHC loci must replace DR function, implying a large change in role between species.

MHC class II proteins are heterodimers consisting of α and β subunits. They are coded by linked α and β genes, originating from an ancient duplication event in mammalian evolution (Takahashi *et al*., 2000). Subsequently, different MHC class II proteins arose by tandem

Correspondence: Dr D Čížková, Department of Population Biology, Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Studenec 122, Koněšín 67502, Czech Republic.
E-mail: dejsha@seznam.cz
Received 4 March 2010; revised 6 July 2010; accepted 19 July 2010; published online 8 September 2010

duplication of α and β gene pairs (Hughes and Nei, 1990). Despite the long divergence between α and β genes, it is still possible to trace homologies based on amino acid and protein structure alignments (Lefranc *et al*., 2005). The subunits combine to perform a single function—binding of antigen for presentation to T cells. This bond is via antigen binding amino acids coded predominantly by the second exon of each subunit (see, for example, Fremont *et al*., 1998). In the house mouse there are two functional MHC class II proteins, I-E and I-A (orthologues of human DR and DQ). Their α subunits are encoded by *H2-Ea* (*DRA*) and *H2-Aa* (*DQA*) and their β subunits by *H2-Eb* (*DRB*) and *H2-Ab* (*DQB*) (Kumanovics, 2007). In this study we contrast *H2-Aa* and *H2-Eb* coding for α and β subunits, respectively. They lie only 18 kb, and 0.01 cM, apart (Mouse Genome Informatics, http://www.informatics.jax.org/) and are the most widely studied MHC class II genes in rodents.

Although house mice are excellent models for immunogenetics of wild-living vertebrates, with well-studied immune genes, most studies of mouse MHC class II variation in the wild have used serology and allozymes—phenotypic rather than genotypic measures (Götze *et al*., 1980; Arden and Klein, 1982; Nadeau *et al*., 1988). To our knowledge, the only exception was a study of microsatellite variation in the MHC region in *Mus musculus castaneus* (Huang and Yu, 2003). Other genetic studies of mouse MHC class II genes are of laboratory or wild-derived inbred strains (for example, Saha *et al*., 1993; Edwards *et al*., 1997) that are known to be developed from a few ancestral individuals or a restricted number of wild pairs (Guenet and Bonhomme, 2003, and references therein). Thus, it is likely that the vast majority of mouse MHC variability remains to be discovered.

In the wild, *Mus musculus* is polytypic: taxonomists recognize five subspecies that originated probably in the Indian subcontinent (Guenet and Bonhomme, 2003). Two of them, *M. m. musculus* and *M. m. domesticus*, colonized Europe along different routes: the former across Russia to eastern and central Europe, and the latter across middle and near-East and the Mediterranean to western and northern Europe (Cucchi *et al*., 2005). In Europe, the two subspecies meet and form a narrow zone of secondary contact 2500 km long and at most 25 km wide (Tucker *et al*., 1992; Raufaste *et al*., 2005; Macholán *et al*., 2007). This limited hybridization is probably caused by genetic incompatibilities or conflicts between the subspecies (Macholán *et al*., 2008). The *a priori* uncertainty about the scale of introgression of MHC variants in a secondary contact zone moved us to sample mice from localities over a long linear transect (cca 1000 km, from central Germany to eastern Slovakia), compared with the scale of introgression of allozyme loci (about 6–18 km, Macholán *et al*., 2007). Our wide sampling allows us to distinguish standing variation of two MHC class II genes in the subspecies from variants that have introgressed across the contact zone since secondary contact. This is vital if we are to identify TSPs dating from the earlier separation of the taxa. We are then able to compare not only evolutionary patterns for the two class II genes across European house mice, but also to compare patterns for these genes between 'pure' *musculus* and *domesticus* mice, and assess levels of TSP. Together with the insight gained into the evolutionary history of MHC in natural populations of two closely related taxa that are ancestral to the laboratory mouse, this study is a necessary preliminary to a fine-scale study of clines in MHC genes across the mouse contact zone.

Materials and methods

Sampling design

We examined 367 *Mus musculus* individuals caught between years 1997–2007 from 17 localities in Germany, Czech and Slovak Republics, with 17–22 individuals per locality (Figure 1, for global positioning system coordinates see Supplementary Table 1). The localities are situated on a 1000-km transect and include representatives of *Mus m. domesticus* and *Mus m. musculus*. The hybrid zone between the subspecies lies at the centre of the transect. Based on allozyme loci, Arzdorf, Schweben and Straas are genetically 'pure' *domesticus* localities, whereas Buskovice, Studenec and Cejkov are 'pure' *musculus* localities (Macholán *et al*., 2008; Macholán M and Piálek J, unpublished data). The other 11 localities include individuals with a range of hybrid indices.

DNA extraction and primer design for exon 2 of *H-2Aa* and *H-2Eb* genes

Genomic DNA was extracted from tissue samples (spleen stored in ethanol) using DNeasy 96 Blood & Tissue Kit (Qiagen, Hilden, Germany), eluted in 200 μl and 10 times diluted. We amplified the second exon of *H-2Aa* and *H-2Eb*. PCR primers were designed on the basis of the alignment of various rodent sequences retrieved from public databases. We used MusDQA primers to amplify 207 bp (excluding primers) out of 249 bp of *H-2Aa* exon 2 and MusDRB primers to amplify 270 bp out of 271 bp of *H-2Eb* exon 2 together with a part of the

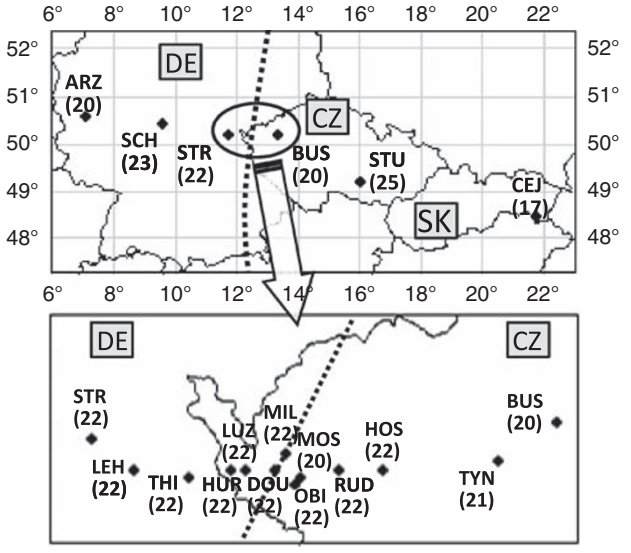


Figure 1 Position of sampled localities on the map of Europe: Arzdorf (ARZ), Schweben (SCH), Straas (STR), Buskovice (BUS), Studenec (STU) and Cejkov (CEJ), with the sample size in brackets. Dashed line schematically represents the house mouse hybrid zone. (Below) The hybrid localities between Straas and Buskovice: Tyniste (TYN), Horni Slavkov (HOSL), Rudolec (RUD), Mostov (MOS), Obilna (OBIL), Milhostov (MIL), Mostov (MOS), Doubi (DOU), Luzna (LUZ), Hurka (HUR), Thierstein (THI) and Lehsen (LEH).

adjacent intron. MusDRB primers were designed not to amplify a paralogue of *H-2Eb* called *H-2Eb*2, which is adjacent to *H-2Eb* (0.01 cM, 9.5 Kb) but has a different function from conventional class II genes (Braunstein and Germain, 1986). To control for null alleles, we used a second pair of primers for each locus. These were MusAa primers amplifying 236 bp of the *H-2Aa* exon 2 together with 5 bp of intron 1 and JS1 and JS2 primers (Schad *et al*., 2004), which are commonly used to amplify rodent *DRB* and cover 217 bp of *H-2Eb* exon 2. For positions of priming sites, primer sequences and PCR conditions, see Figure 2.

Genotyping of exon 2 from *H-2Aa* and *H-2Eb*

H-2Aa and *H-2Eb* exons 2 were amplified with fluorescently labelled MusDQA and MusDRB primers (conditions described in Figure 2). We performed single-strand conformation polymorphism (SSCP) analysis, (for principles, see for example, Bryja *et al*., 2005). PCR product (0.5–4 µl) was mixed with 0.5 µl of 500 LIZ Size Standard (Applied Biosystems, Foster City, CA, USA) with 12 µl Hi-Di formamide and denatured at 95 °C for 3 min. Electrophoresis was run at 18 °C in 5% non-denaturing conformation analysis polymer (Applied Biosystems) using ABI PRISM 3130 (Applied Biosystems). Data were analysed using GENEMAPPER v.3.7 (Applied Biosystems).

We selected individuals with diverse SSCP patterns to investigate sequence variation. Where homozygous individuals were detected by SSCP analysis, we directly sequenced the PCR product using BigDye terminator chemistry (Applied Biosystems). For heterozygous SSCP patterns or when peak overlap was suspected, we identified the sequences by cloning and sequencing the PCR products (Finnzymes, Espoo, Finland, proof-reading DNA polymerase was used), as described in Promerová *et al*. (2009). We accepted a sequence as a new allele only when it was sequenced at least twice from two independent PCRs. To check for PCR artefacts we always compared the SSCP pattern of the clone with that of the heterozygous individual (for individual genotypes, see Supplementary Table 1).

The sequences thus obtained were edited and aligned according to the GenBank reference sequences for mice *H-2Eb* (NM_010382) and *H-2Aa* (NM_010378) in SEQSCAPE v2.5 (Applied Biosystems). Gene and allele nomenclature follow Klein *et al*. (1990 a, b). Alleles found only in hybrid localities were named *H-2Ebmusxdom* and *H-2Aamusxdom*. For 'pure' *H-2Eb* alleles, we continued with the numbering started by Edwards *et al*. (1997). For 'pure' *H-2Aa* and the hybrid alleles we started with number 1. When the same sequence was present in both 'pure' subspecies, it was named twice, separately for each subspecies.

Transcription analysis

As Knapp (2007) clearly points out, hypothesizing selective effects on MHC requires unambiguous identification of functional, transcribed mRNA sequences. To verify transcription, we used samples from 76 individuals caught in 2006 at localities in the *Mus musculus* hybrid zone. Samples were stored both in

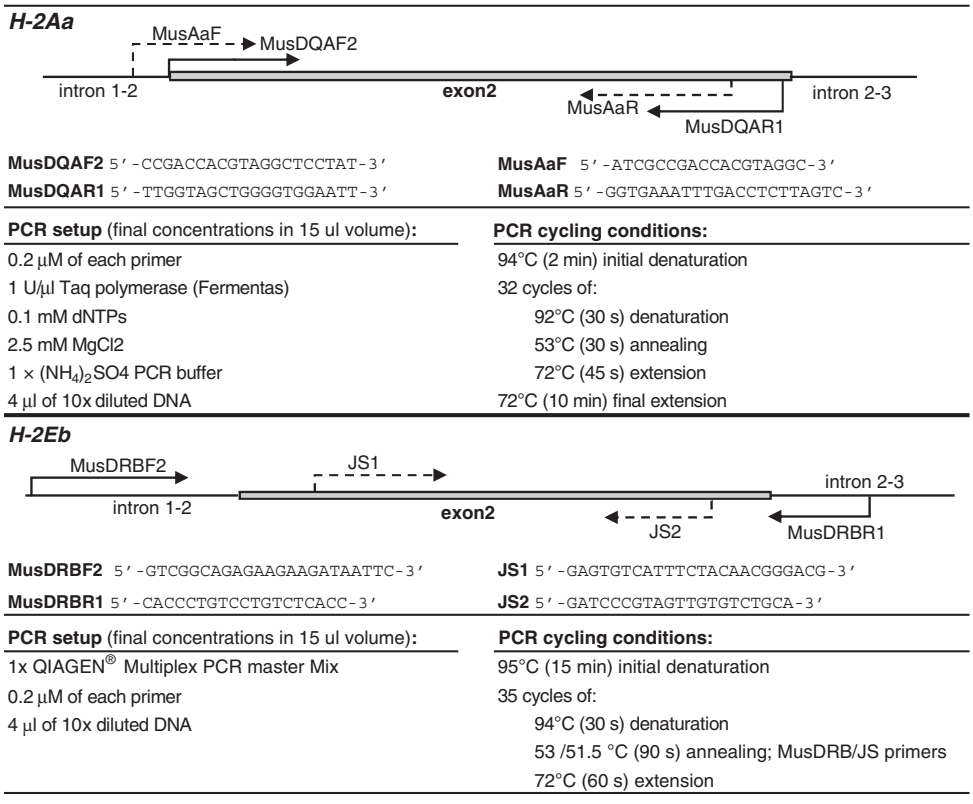


Figure 2 Amplification of *H-2Aa* and p*H-2Eb* exon 2: position and sequences of PCR primers, reaction setup and cycling conditions. Shaded boxes represent exon 2, and the lines represent adjacent introns. Arrows indicate position of primers. PCR setup and conditions are identical for the two pairs of primers.

ethanol (DNA extraction) and in RNA later (Ambion, Austin, TX, USA) (RNA extraction). We extracted DNA and genotyped the samples for both *H-2Aa* and *H-2Eb* by amplification with MusDQA and MusDRB primers followed by SSCP (as described above). According to SSCP patterns we identified alleles that were identical with the alleles present in our locality samples. We chose 15 heterozygous individuals whose genotypes covered the identified alleles and extracted total RNA using the RNeasy Plus Mini Kit (Qiagen). Reverse transcription of 5 μg of total mRNA was performed with SuperScript III (Invitrogen) and poly(T) primer. For *H-2Eb* complementary DNA (cDNA) amplification, we multiplexed JS and MusDRB primers. JS primers were used to detect expression of particular alleles, as their priming sites are inside exon 2, whereas the MusDRB primers, which have priming sites in adjacent introns, served to control for genomic DNA contamination. We followed the conditions described above, but we added 0.2 μM of each primer and 0.5 μl of cDNA as a template. To amplify *H-2Aa* cDNA we used MusDQA primers and the same conditions as above, apart from adding 0.5 μl of cDNA as template. For *H-2Aa* we did not check for contamination with genomic DNA, as the cDNA template was identical to the *H-2Eb* amplification template. As a final check, we randomly chose three samples and performed PCR directly from the isolated RNA.

Data analysis

Analysis of sequence evolution and selection: We estimated substitution distances between the sequences as a measure of their evolutionary divergence. We used Kimura's two-parameter model (Kimura, 1980), adjusted to account for varying substitution rates between sites ($\gamma = 0.5$) as implemented in molecular evolutionary genetics analysis (MEGA; Tamura *et al*., 2007).

The assumptions of many methods of DNA sequence analysis (especially distance based) are violated if recombination is present. To check for the presence of recombination in the alignment of our sequences, we used PERMUTE (included in the software package OMEGAMAP; Wilson and McVean, 2006), which computes three commonly used statistics (r^2, D and G4) based on the correlation between physical distance of sites and their linkage disequilibrium.

To estimate levels of positive selection, often thought to drive the evolution of exon 2 of MHC class II sequences (reviewed in Garrigan and Hedrick, 2003), we used two approaches based on different principles. The classical approach is based on assessing the average ratio of dN/dS separately for the antigen-binding sites (ABS; supposed to be under positive selection) and the remaining sites (non-ABS) and comparing the values. We used two different substitution models as recommended in Zhang and Yu (2006). The modified Nei and Gojobori method with Jukes–Cantor correction for multiple substitutions (Nei and Gojobori, 1986) is commonly used in similar studies (see, for example, Sommer, 2003; Bryja *et al*., 2007; Babik *et al*., 2008), whereas the Kumar method (Tamura *et al*., 2007) incorporates more parameters and should give more precise estimates, especially in the case of 'complicated datasets', although risking overparameterization. The codons participating in antigen binding were identified according to the review of Bondinas *et al*. (2007) and are listed in Figure 3.

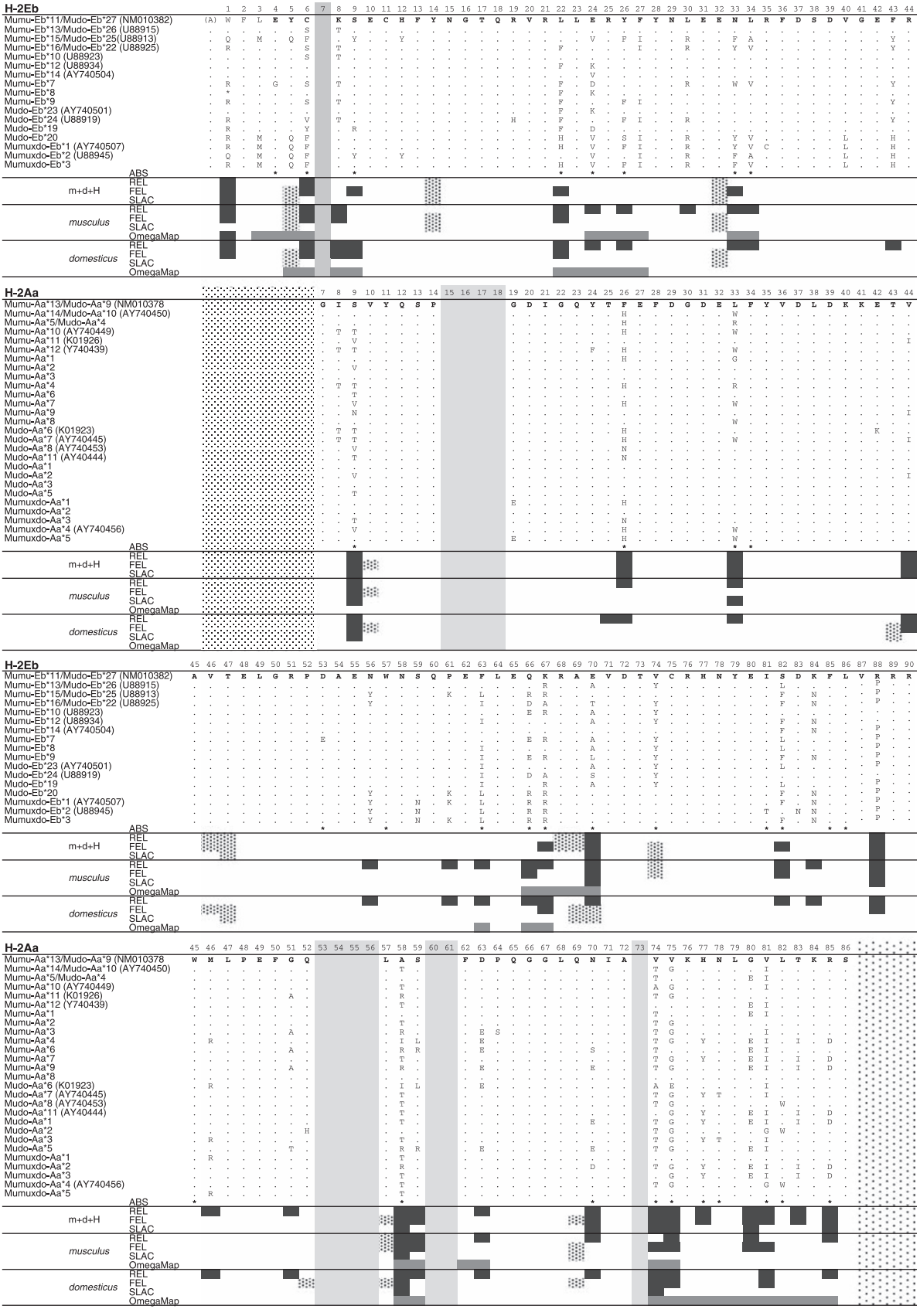
The second approach for estimating positive selection compares the rates of synonymous and non-synonymous substitutions separately for every single codon with no previous assumptions on antigen binding. We chose to use HYPHY (Pond *et al*., 2005) and OMEGAMAP (Wilson and McVean, 2006) software packages, both of which have the important advantage of taking recombination into account. The basic difference between these approaches is that HYPHY employs a 'classical phylogenetic' approach whereas OMEGAMAP is based on coalescent theory, and therefore assumes population-level sampling. The analyses were performed separately for three groups of sequences: *musculus* (that is, all the sequences that were present in three 'pure' *musculus* localities, regardless if they were present in hybrid or *domesticus* localities at the same time), *domesticus* and, thirdly, all the sequences together (that is, *domesticus*, *musculus* and alleles restricted to the hybrid zone). The last group was not analysed with the OMEGAMAP software, as samples drawn from across a taxon barrier do not represent a population sample. In HYPHY, we first identified recombination break points with genetic algorithm recombination detection (GARD; Pond *et al*., 2006), which infers phylogenies for each putative nonrecombinant fragment. The output was used to run three different maximum likelihood methods for detection of selection: single likelihood ancestral counting, fixed effects likelihood and random effects likelihood.

The single likelihood ancestral counting and fixed effects likelihood are conservative tests, and lack of information will result in low power (the data set comprises 17 and 27 relatively short sequences—the unique *DRB* and *DQA* alleles).

The random effects likelihood tests, in contrast, must be treated with caution, as they tend to return false positives in the absence of sufficient information. Following the recommendations of the authors, we set high α-levels of 0.25 for single likelihood ancestral counting and fixed effects likelihood and Bayes factor cutoff at 50 for random effects likelihood (Pond and Frost, 2005a). The authors recommend considering all three methods and following the consensus. We therefore regarded a codon to be selected only if it was identified by at least two of the methods.

Figure 3 Amino acid alignment of *H-2Aa* and *H-2Eb* exon 2. Alignment of loci and codon numbering follows the IMGT (international ImMunoGeneTics information system) (Lefranc *et al*., 2005), based on structural and functional comparisons. Codons present at one locus but absent in the other are in grey. Parts of *H-2Aa* sequence that were not genotyped are dotted. Sequences are aligned against GenBank reference sequences for mouse MHC. The alleles that have been sequenced previously have the GenBank accession numbers attached. Antigen-binding codons (ABS) and codons under selection are shown below the alignments. Selection was analysed for three groups of sequences: *musculus*, *domesticus* and pooled *musculus*, *domesticus* and hybrid-specific sequences (m + d + H). Positively selected codons identified by random effects likelihood (REL), fixed effects likelihood (FEL) and single likelihood ancestral counting (SLAC) are shown as black squares, by OmegaMap as grey squares, negatively selected codons as shaded squares and antigen-binding codons as asterisks.

To compare the strength of positive versus negative selection across different groups of sequences, we estimated whether the proportion of positively versus negatively selected codons was significantly different between groups, and if so, in which direction. We used generalized linear mixed model with binomial distribution



and logit link function. For each gene and all the three groups of sequences, we estimated the proportion of positively versus negatively selected codons. These proportions were used in the generalized linear mixed model as a response variable. We evaluated the significance of the following explanatory variables: species identity (*musculus* vs *domesticus* vs *musculus* + *domesticus* + hybrid), type of selection (positive or negative), gene identity (*H-2Eb* vs *H-2Aa*) and interactions. Significance levels were estimated using the standard deletion procedure based on likelihood ratio tests (that is, Crawley, 2007) assuming χ^2 distribution of deviance change. These data suffer from several sources of non-independence. To control at least partially for this, we included gene identity nested within species identity as a random intercept of the model.

To test for possible differences in selection regimes between the allelic lineages, we used the genetic algorithm (GA)-branch method implemented in HYPHY (Pond and Frost, 2005b). This codon-based method allows for varying dN/dS in phylogeny without *a priori* specification of the distinct lineages.

The OMEGAMAP software is based on a population genetics approximation to the coalescent with recombination. It does a Bayesian co-estimation of the dN/dS ratio (ω) and population recombination rate ($\rho = 4Nr$), allowing each to vary along the sequence (Wilson and McVean, 2006). The analyses used 'objective' priors, with the probable values of the mutation rate and the transition/transversion rate ratio adjusted to follow improper inverse distributions. The ω and ρ were set to fit exponential ($\lambda = 1$) distributions, with seven-codon blocks for ω and ρ. Two Markov chain Monte Carlo runs of 400 000 iterations (5000 iteration burn-in) were compared for convergence. Posterior probability of a positively selected codon was set at 0.95. Total mutation rate per base pair ($\theta tot = 4\,N\mu$) was calculated using an R-script provided by Daniel Wilson (personal communication). As OMEGAMAP assumes population-level sampling, we were not forced to reduce our samples to the subset of unique sequences.

Phylogenetic analyses: We constructed neighbour-joining trees for exon 2 alleles at the two loci with MEGA (Tamura *et al*., 2007), using the same simple model of sequence evolution as for assessing evolutionary divergence (K2P, $\gamma = 0.5$) and tested phylogenies using 100 000 bootstraps. We included unique *musculus*, *domesticus* and hybrid-derived exon 2 sequences discovered in this study together with sequences from six *Mus* taxa retrieved from the GenBank database (14 *H-2Aa* and 43 *H-2Eb* sequences). We used only those taxa sampled for both loci and the sequences with proven origin with regard to species/subspecies. Water vole (*Arvicola terrestris*) sequences were used to root the trees.

Distribution of genetic variation: As a measure of genetic diversity of the two loci, we calculated their expected heterozygosity with GENEPOP 4.0.7 (Raymond and Rousset, 1995). Tests for Hardy–Weinberg equilibrium were performed in GENEPOP; we tested separately for heterozygote deficiency and excess. Genotypic (linkage) disequilibrium between *H-2Aa* and *H-2Eb* was assessed separately for each locality sample by exact tests in GENEPOP.

We estimated standardized F-statistics between six 'pure' populations to assess the pattern of population structure for the two loci. We used RECODEDATA (Meirmans, 2006) to generate two types of data sets from the original genotypes (separately for each locus): first, a data set where each locality sample was recoded to have unique alleles, the second where the *musculus* and the *domesticus* groups were recoded to have unique alleles. With the original and the two recoded data sets, we performed hierarchical analysis of molecular variance using ARLEQUIN 3.11 (Excoffier *et al*., 2005). Results from the original data set gave actual values of F_{ST}, F_{CT} (an index describing the amount of population differentiation between the groups of locality samples, that is, *musculus* and *domesticus*) and F_{SC} (differentiation between the populations within the groups). Results from the first recoded data set gave the maximum value of F_{ST} and F_{SC} possible given the within-population variance present. Similarly, the second recoded data set gave maximal F_{CT}. Standardized values were calculated by dividing the original values by the maximal values (Hedrick, 2005). In the same way we calculated standardized F_{ST} for the locality pairs. These standardized values allow us to control for the different within-population variance (because of the different levels of polymorphism) of the two loci. To assess population differentiation within the subspecies and to compare between them, we used two-sided tests comparing the two groups of locality samples (10 000 permutations) in FSTAT (Goudet, 2001).

Results

Allelic diversity of *H-2Aa* and *H-2Eb* exon 2

We identified 27 *H-2Aa* and 17 *H-2Eb* alleles in 367 individuals from 17 locality samples. All alleles of *H-2Aa* were amplified by both the MusDQA and MusAa primers, that is, no *H-2Aa*-null alleles occurred in our samples. By verifying MusDRB primer results with the second pair of *H-2Eb* primers (JS primers), we found that 19 supposedly MusDRB-homozygous individuals were heterozygous for a total of four MusDRB-null alleles. Three of these null alleles (MumuEb*15/MudoEb*25, Mumuxdo-Eb*2 and Mumuxdo-Eb*3) were in fact amplified by MusDRB in homozygotes, but for the above-mentioned heterozygotes their amplification was at background noise level. For the fourth null allele (Mudo-Eb*20), there was no homozygous individual present within this study. However, we found Mudo-Eb*20 in the homozygous state in the STRB strain derived from the Straas locality (Piálek *et al*., 2008). A STRB individual was amplified with MusDRB primers to get the full Mudo-Eb*20 sequence.

Out of 27 *H-2Aa* and 17 *H*-2Eb sequences found in this study, 17 and 6 sequences, respectively, had not been previously identified in *Mus musculus*. We deposited all the sequences in the GenBank database under accession numbers: HM443553-HM443603. Apart from one *H-2Eb* sequence, we detected no insertions, deletions or stop codons in either *H-2Aa* or *H-2Eb* sequences. A single substitution was found to separate allele MumuEb*8 from allele MudoEb*23, changing the second codon of allele MumuEb*8 to a stop codon (Figure 3).

Transcription analysis

In the 76 individuals with available RNA, SSCP revealed 16 *H-2Aa* alleles out of the 27 present in our DNA samples and 11 alleles out of 17 for *H-2Eb*. For *H-2Aa*, all the SSCP patterns obtained from cDNA were identical to those from genomic DNA, implying no aberrant expression at the level of mRNA. For *H-2Eb*, we detected two cDNA patterns that did not match genomic DNA patterns. First, the profile of the MumuEb*11/MudoEb*27 allele was absent in independent cDNA amplicons, suggesting that its transcript was not present in the sample. Second, there were three alternative transcription levels of the MumuEb*12 allele in heterozygotes. In the first of these cDNA samples, we detected intense expression and the SSCP pattern was identical to the profile obtained from genomic DNA. In the second individual, no transcript of MumuEb*12 was detected and in the third one, the signal was barely detected, suggesting a very low level of transcription. Each time, the signal of the second allele of the heterozygote amplified from cDNA reached the same height as that from genomic DNA, indicating specific decrease in transcription of MumuEb*12.

Sequence evolution: nucleotide substitutions

Within-locus allelic divergence, as estimated by K2P distances, was higher for *H-2Eb* than *H-2Aa* (Table 1). Similarly, the amount of both synonymous and non-synonymous substitutions was higher for *H-2Eb* when considering both unique sequences (dS and dN, distance-based methods, Table 2) and population samples (θsyn and θnon, OMEGAMAP, Table 3).

Table 1 Number of alleles found for *H-2Eb* and *H-2Aa* loci in the *musculus* (H-2Eb mus, H-2Aa mus), *domesticus* (H-2Eb dom, H-2Aa dom) and pooled *musculus*, *domesticus* and hybrid (H-2Eb m+d+H, H-2Aa m+d+H) locality samples in *N* individuals

	N	*No. of alleles*	*He*	*K2P*	*min d*	*max d*
H-2Aa m+d+H	367	27	0.906	0.075 ± 0.013	0.005	0.170
H-2Aa mus	62	14	0.838	0.070 ± 0.014	0.005	0.153
H-2Aa dom	64	11	0.834	0.074 ± 0.013	0.026	0.137
H-2Eb m+d+H	367	17	0.720	0.104 ± 0.016	0.004	0.229
H-2Eb mus	62	10	0.784	0.079 ± 0.013	0.011	0.180
H-2Eb dom	64	8	0.683	0.107 ± 0.017	0.027	0.193

Evolutionary divergence between the alleles was measured as mean Kimura's two-parameter distance (K2P); min d and max d give minimal and maximal pairwise K2P distances; and genetic diversity is given as expected heterozygosity (He).
Mean values and s.e. estimated by bootstrap with 10 000 replicates are shown.

Sequence evolution: recombination

For both loci, and all groups of sequences, recombination (or homologous gene conversion) was indicated by at least one of the three PERMUTE statistics (Table 3). The signal of recombination was stronger for *H-2Aa*, regarding both number of positive tests and level of significance. A higher level of recombination was confirmed by the average recombination rate per codon (ρ) calculated by OMEGAMAP, which was about twice as high for *H-2Aa* than for *H-2Eb*. To find out which mechanism (substitution or recombination) dominated in the evolution of the two loci, we examined the ratio of the OMEGAMAP average mutational rate (θtot) and the average recombinational rate (ρ). By this measure, recombination was of approximately the same importance in the evolution of *H-2Aa* as were substitutions. The evolution of *H-2Eb* was, on the contrary, about two times more affected by substitutions than by recombination (Table 3). No particular recombination hot spots were indicated by OMEGAMAP, either for *H-2Eb* or for *H-2Aa*, with similar ρ values found all along the sequences.

Selection analysis

The dN/dS ratio (Table 2) indicated approximately the same amount of positive selection for both loci, if averaged over the two methods used (2.63 and 2.52 for *H-2Aa* and *H-2Eb*, respectively). *H-2Aa* ABS showed signs of intense positive selection (35 times more non-synonymous substitutions if averaged over the methods) whereas at non-ABS negative selection was detected. Similarly, for *H-2Eb*, the ABS were found to be under positive selection, although much weaker than in *H-2Aa* (dN value just three times higher). Interestingly, positive selection was also detected for non-ABS in *H-2Eb* (dN twice dS).

The positions of codons found to be positively and negatively selected by the three standard methods implemented in the HYPHY package are shown in Figure 3. The generalized linear mixed model analysis of positive to negative selected codon ratios did not detect any significant effect of species identity (either as the main effect or included in statistical interactions with other variables). This suggests that the overall level of selection is comparable for *musculus* and *domesticus*

Table 2 Synonymous (dS) and non-synonymous (dN) distances and their ratio (dN/dS) for all the sites, ABS and non-ABS in *H-2Aa* and *H-2Eb* sequences

Locus	*Methods*	*All sites*			*ABS*			*Non-ABS*		
		dS	*dN*	*dN/dS*	*dS*	*dN*	*dN/dS*	*dS*	*dN*	*dN/dS*
H-2Aa	K	0.032 ± 0.013	0.079 ± 0.018	2.47	0.009 ± 0.007	0.338 ± 0.097	37.56	0.035 ± 0.015	0.029 ± 0.008	0.83
	NG+JC	0.029 ± 0.010	0.081 ± 0.017	2.79	0.010 ± 0.007	0.332 ± 0.095	33.20	0.034 ± 0.013	0.030 ± 0.009	0.88
H-2Eb	K	0.046 ± 0.014	0.096 ± 0.018	2.09	0.130 ± 0.079	0.318 ± 0.076	2.45	0.031 ± 0.012	0.045 ± 0.012	1.45
	NG+JC	0.037 ± 0.010	0.109 ± 0.021	2.95	0.091 ± 0.040	0.354 ± 0.103	3.89	0.025 ± 0.009	0.058 ± 0.015	2.32

Abbreviation: ABS, antigen-binding site.
The distances were calculated using the Kumar (K) and Nei–Gojobori method with Jukes–Cantor correction (NG+JC).
The s.e. were estimated by bootstrap with 10 000 replicates.

Table 3 Intralocus recombination in the *musculus* (H-2Eb mus, H-2Aa mus), *domesticus* (H-2Eb dom, H-2Aa dom) and pooled *musculus*, *domesticus* and hybrid (H-2Eb m+d+H, H-2Aa m+d+H) sequences measured as r^2, D and G4, and OmegaMap's recombination rate (σ)

	r^2	$\|D'\|$	*G4*	ρ	θ*syn*	θ*non*	θ*tot*/σ
H-2Aa m+d+H	**	*	*	—	—	—	—
H-2Aa mus	**	**	**	0.119	0.028	0.102	1.098
H-2Aa dom	**	NS	NS	0.143	0.030	0.107	0.958
H-2Eb m+d+H	NS	NS	NS	—	—	—	—
H-2Eb mus	*	NS	NS	0.072	0.031	0.125	2.180
H-2Eb dom	*	NS	NS	0.077	0.036	0.127	2.107

θsyn and θnon give population synonymous and nonsynonymous mutation rates per base pair; θtot/σ ratio is the relative contribution of substitution versus recombination in the evolution of the sequences (calculated by OmegaMap).
Statistical significance: $*P<0.01$, $**P<0.001$, NS, not significant $P>0.01$.

groups and that likewise the proportion of positive to negative selection is comparable (that is, nonsignificant interaction '*gene* × *subspecies*' identity). On the other hand, the interaction '*gene identity* × *type of selection*' (positive vs negative) was significant (Δd.f. = 1, $\chi^2 = 15.693$, $P<0.0001$). We conclude that the proportions of loci under negative and positive selection differ between these two loci. Two separate generalized linear mixed models with the proportion of codons under negative and positive selection as a response variable and gene identity as the only explanatory variable revealed that there is a significantly higher proportion of positively and significantly lower proportion of negatively selected codons in *H-2Aa* compared with *H-2Eb* (Δd.f. = 1, $\chi^2 = 4.141$, $P = 0.04186$ and Δd.f. = 1, $\chi^2 = 11.572$, $P = 0.0006694$, respectively).

OMEGAMAP identified 9 and 20% of the codons to be positively selected for *H-2Aa musculus* and *domesticus* groups, respectively, and 18 and 13% for *H-2Eb* (negative selection is not detected by this method). These percentages should be taken cautiously as they are likely influenced by the choice of block size for ω variation along the sequence (we chose seven-codon blocks, as further shortening resulted in poor Markov chain Monte Carlo convergence). Because of this blockwise treatment, strongly selected codons may obscure signal in their flanking neighbours. This can be seen, for example, at codons 3–9 of *H-2Eb* (Figure 3), where negative selection at codon 5 detected by HYPHY seems to have been obscured in OMEGAMAP by positive selection signal from neighbouring codons. However, despite these differences between methods, OMEGAMAP and HYPHY showed very high congruence in identifying parts of the sequences under selection (Figure 3).

Consistency between ABS based on crystalography and codons determined as positively selected by at least two HYPHY methods was better for *H-2Aa* (56% matched) than for *H-2Eb* (18%).

Using the GA-branch method, we found sequences in the *H-2Eb* genealogy with significantly different dN/dS ratio (Δ-cAIC = 11.28). The sequences under positive selection (dN/dS = 6.98) included 61% of tree branches and 10 extant sequences, whereas the sequences under negative selection (dN/dS = 0.46) included 39% of tree branches and 7 extant sequences. Negatively selected sequences are indicated in Figure 4. For *H-2Aa* sequences, we found one positively selected clade (dN/dS = 4.223).

Phylogenetic analysis

Neighbour-joining trees calculated from the available sequences showed a strong trans-species pattern. For both loci, we found that the alleles cluster without regard to the species or subspecies from which they were sampled (Figure 4).

Distribution of genetic variation

Expected allelic heterozygosity was higher for *H-2Aa* than *H2-Eb* (Table 1). In testing deviations from Hardy–Weinberg equilibrium, only the Arzdorf, Hurka and Mostov localities showed significant heterozygote deficiency. Because deviation was detected for both loci in these samples, it is likely that it was caused by demographic effects rather than null alleles.

Highly significant ($P<0.001$) genotypic (linkage) disequilibrium between *H-2Aa* and *H-2Eb* was detected in all locality samples, consistent with their map distance (0.01 cM).

For both loci we found alleles specific to single localities and subspecies. This differentiation was however more pronounced for *H-2Aa* than for *H-2Eb* (see Supplementary Table 2 for the allelic distribution over the localities). For *H-2Eb*, 24% of sequences (4 out of 17) had identical matches between the subspecies, representing 73% of the total allele copies. For *H-2Aa*, 11% of sequences had matches between the subspecies (3 out of 27), and only 15% of total copies. The overall population differentiation between the 'pure' subspecies was maximal for *H-2Aa* ($F_{STstand} = 0.99$), whereas for *H-2Eb* it was $F_{STstand} = 0.60$. Similarly, the *H-2Aa* allelic distribution was highly subspecies specific ($F_{CTstand} = 0.97$) whereas for *H-2Eb* specificity was low ($F_{CTstand} = 0.16$). Locality differentiation within the *musculus* and *domesticus* groups was $F_{SCstand} = 0.65$ and 0.52 for *H-2Aa* and *H-2Eb*, respectively. Taking both loci together, we found no difference in the differentiation (F_{ST}) between the *musculus* and *domesticus* groups. Standardized pairwise F_{ST} values are shown in Table 4.

Discussion

This study analyses the genetic diversity and its distribution at two MHC class II genes, *H-2Aa* and *H-2Eb*, in two subspecies of wild house mouse and their hybrids, arguably one of the most intensively studied models of speciation. We studied their transcription patterns, analysed their molecular mechanisms of sequence evolution (for example, substitution distances, recombination vs point mutations, positive selection), described their population structure and compared this across both MHC genes and subspecies. Despite the close linkage between the two genes, we found substantial differences between the loci in almost all analysed features. We did not find any substantial differences between the orthologues in *musculus* and *domesticus* for either locus.

Figure 4 Neighbour-joining trees of (**a**) *H-2Aa* and (**b**) *H-2Eb* exon 2 sequences of mice, including sequences discovered in this study (in bold, labelled as, for example, **Mumu-Aa*1**). Sequences stored in GenBank have the accession numbers attached. The following taxa were included: *Mus musculus castaneus* (Muca), *Mus musculus domesticus* (Mudo), *Mus musculus molossinus* (Mumo), *Mus musculus musculus* (Mumu), *Mus spicilegus* (Musi) and *Mus spretus* (Musp). *Arvicola terrestris* (Arte) sequences were used to root the trees. Bootstrap values >60 are shown. The *H-2Eb* sequences highlighted in grey are under negative selection according to GA-branch analysis of sequences found in this study.

Differences in genetic polymorphism between two MHC class II genes

Extensive allelic diversity in both allele numbers and pairwise nucleotide distances is characteristic for MHC genes (Klein *et al*., 1993). Mouse *H-2Aa* showed higher allelic polymorphism than *H-2Eb* (27 vs 17 alleles). In other rodents the pattern seems to be the opposite (Smulders *et al*., 2003; Sommer, 2003; Bryja *et al*., 2007; Tollenaere *et al*., 2008). However, such contrasts are perhaps not surprising, given the rapid evolutionary dynamics of MHC loci (Nei *et al*., 1997). In all, 17 *H-2A* and 6 *H-2Eb* alleles were previously unidentified in *Mus musculus*, confirming the inadequacy of inbred mouse strains as indicators of variability in wild mice. Although *H2-Aa* had more alleles, *H-2Eb* had greater allelic diversity in terms of nucleotide distance, both mean and maximum number of substitutions and synonymous and nonsynonymous substitutions.

These differences in polymorphism could be due to either a difference in the speed of allelic diversification or in time of their persistence in the gene pool (Klein, 1987; Takahata *et al*., 1992). We propose that the higher diversity of *H-2Eb* alleles is because of longer persistence. Otherwise, a higher mutation rate must be invoked to explain the higher synonymous distances at *H-2Eb*. Fast diversification of the *H-2Aa* alleles is likely because of greater intralocus recombination (or gene conversion), almost twice that in *H-2Eb* and twice the rate of substitution. Recombination creates new allelic variants—new combinations of existing variation—and may be a dominant source of allelic polymorphism (Takahata and Satta, 1998). High frequency of intralocus recombination (dominating substitutions) has already been described in classical MHC genes in various vertebrates (see, for example, Alcaide *et al*., 2007; Promerová *et al*., 2009) including rodents (Richman *et al*., 2003).

Differences in the level of TSP

In the phylogenetic trees constructed from the available sequences, alleles at both loci clustered regardless of their species/subspecies origin. Trans-species origin (TSO) and subsequent maintenance of polymorphism

Table 4 Standardized pairwise F_{ST} for locality sample pairs

H2-Eb		*H2-Aa*	
Loc. pair	$F_{ST\ stand}$	*Loc. pair*	$F_{ST\ stand}$
STU-STR	0.206	CEJ-BUS	0.357
STU-SCH	0.233	SCH-ARZ	0.360
SCH-ARZ	0.268	CEJ-STU	0.648
STU-ARZ	0.393	STR-SCH	0.651
CEJ-STU	0.394	STR-ARZ	0.749
STR-ARZ	0.414	STU-STR	0.966
STR-SCH	0.449	CEJ-STR	0.976
CEJ-BUS	0.450	STU-SCH	0.983
CEJ-STR	0.494	STU-BUS	0.988
CEJ-SCH	0.514	STU-ARZ	0.998
CEJ-ARZ	0.562	CEJ-SCH	1.007
BUS-STR	0.911	BUS-STR	1.009
STU-BUS	0.922	BUS-SCH	1.010
BUS-ARZ	0.934	BUS-ARZ	1.016
BUS-SCH	0.950	CEJ-ARZ	1.018

Between-subspecies pairs are highlighted in grey.

through speciation events is often invoked to explain such trans-species patterns of polymorphism in MHC genes, including rodent homologues of *H-2Aa* (Seddon and Baverstock, 2000; Bryja *et al*. 2006) and *H-2Eb* (Edwards *et al*. 1997; Smulders *et al*., 2003; Musolf *et al*., 2004). However, convergent evolution leading to homoplasy (see, for example, the review Klein *et al*., 2007) and horizontal gene transfer between taxa are other evolutionary processes generating TSP.

In this study, we had the unique opportunity to discriminate between the above TSP-generating processes. To exclude convergent evolution, we consider only identical sequenced alleles shared by *musculus* and *domesticus*, the probability of complete sequence identity (that is, including synonymous sites) by convergence being negligible. For *H-2Aa*, the amount of TSP seems to be twice as low as *H-2Eb*, that is, 11 vs 24% of all the unique alleles at the two loci were shared by the subspecies. However, the amount of TSP is only expected to be reflected by the proportion of shared unique alleles under the simple symmetric-overdominance model of selection, which assumes that all the alleles have the same selective advantages and thus the same expected frequencies. This seems unlikely for our genes and species (and probably for most other natural species/populations). Controlling for allelic frequencies, 73% of all the sampled *H-2Eb* allele copies had allelic states shared across the subspecies, compared with only 15% for *H-2Aa*, indicating an even higher relative TSP for *H-2Eb*. Comparing these numbers is however problematic because of the different allelic polymorphism of the two loci (Hedrick, 1999). Standardized F-statistics, controlling for this bias, showed that *H-2Aa* variation was highly subspecies specific ($F_{CTstand} = 97\%$). On the contrary, there was almost no subspecific differentiation at *H-2Eb* ($F_{CTstand} = 16\%$), confirming much higher levels of TSP in *H-2Eb*.

To control for horizontal gene transfer between the two taxa (introgression) at their contact zone, we compared only 'pure' *musculus* and *domesticus* localities, geographically distant from any localities showing introgression of allozymes (Macholán *et al*., 2008, Macholán M and Piálek J, unpublished data). However, introgression clines can be of very different widths for different loci (Barton and Hewitt, 1985). If the high level of *H-2Eb* TSP was caused by clines of introgression rather than TSO, we would not expect between-subspecies localities to be more similar ($F_{CTstand} = 16\%$) than within-subspecies ones ($F_{SCstand} = 52\%$). Such a pattern is expected only because of TSO, unless strong introgression has homogenized adjacent *musculus* (Buskovice) and *domesticus* (Straas) localities. This is obviously not the case as Buskovice–Straas pairwise F_{ST} is the seventh highest out of nine between-subspecies pairs (Table 4). We conclude that both loci show TSO, but to very different degrees.

Differences in gene transcription and expression

All analysed *H-2Aa* alleles were successfully transcribed. Altered gene transcription was indicated for some *H-2Eb* alleles, consistent with a high incidence of non-expressed E protein revealed by serology in wild mice (Figueroa *et al*., 1989). Transcription of two *H-2Eb* alleles was either decreased (allele MumuEb*12) or absent (MumuEb*11/MudoEb*27). We also found a stop codon in the coding sequence of MumuEb*8. As only 16 out of 27 *H-2Aa* alleles and 11 out of 17 *H-2Eb* alleles have been analysed here, it is possible that further investigation will reveal further transcription and expression variation.

The aberrant transcription of the MumuEb*12 allele was most probably because of a splicing site mutation identified previously in *M. m. domesticus* (haplotypes q and 302) and *M. m. castaneus* (haplotype w17) (Vu *et al*., 1988; Begovich *et al*., 1990; Tacchini-Cottier *et al*., 1995), because the exon 2 sequences of these alleles from sister taxa are identical to our MumuEb*12 sequence. Similarly, the stop codon at position two of exon 2 has already been described in *M. m. domesticus* (haplotype w29 and w37) and *M. spicilegus* (allele Musieb4) (Golubiæ *et al*., 1987; Tacchini-Cottier *et al*., 1995; Edwards *et al*., 1997).

The TSP of the identical *domesticus* and *castaneus* alleles with splicing site mutations would be explained by TSO of this mutation (Tacchini-Cottier *et al*., 1995) and includes a *musculus* allele found within this study. TSO is also indicated for the stop-codon mutation, as *domesticus* w37 haplotype (Mudo U13656) and *musculus* allele MumuEb*8 from our sample cluster together in the phylogeny (Figure 4), although the bootstrap support for this node is weak.

Although TSO seems the most likely explanation for these TSPs, a single ancestral origin of these alleles appears paradoxical because it implies that the ancestral allele had selective advantage strong enough to maintain its descendants through speciation events. It is known that the I-E protein tends to bind viral or bacterial superantigens (Tomonari *et al*., 1993), leading to a deleterious immune reaction. If these superantigens are present in the environment, it can be advantageous not to express I-E protein (Tacchini-Cottier *et al*., 1995). Temporal or spatial differences in the presence of such superantigens would then allow balancing selection to maintain both functional and non-functional variants. This could explain the maintenance of descendants of a stop codon containing (non-functional) variant through a speciation event. However, it does not explain the observed similarity of sequences containing the stop codon across taxa: if these sequences are descendants of a non-functional allele that arose before the *musculus-castaneus* split, then we would

expect them to have accumulated many mutations, both synonymous and non-synonymous, during the intervening history. In short, if the stop codon clade of alleles originates deep in time and is non-functional, then that clade should be diverse rather than highly conserved. One resolution to this paradox is that, although not transcribed, the sequence of the stop codon allele is highly conserved for non-coding reasons. Certainly, this paradoxical evidence suggests that the loss of function in some *H-2Eb* alleles does not automatically mean that *H-2Eb* is in the classic process of pseudogenization commonly assumed in MHC studies (Nei *et al*., 1997).

Differences in the strength and direction of selection detected on sequences

Both loci showed excess of non-synonymous to synonymous substitutions, a signal of positive selection (Hill and Hastie, 1987). Strong positive selection at ABS and negative selection at non-ABS characterize the *H-2Aa* gene, a pattern typical for classic functional MHC genes (for example, see review Garrigan and Hedrick, 2003). A positive selection signal was also detected at ABS of *H-2Eb* (even if much lower than that at *H-2Aa*), but the non-ABS also showed a considerable signal of positive selection. The most probable explanation is that some *H-2Eb* ABS have not been identified by the crystallography studies we refer to. For some MHC alleles and antigen combinations, alternative codons take part in binding (compare Scott *et al*., 1998 and Fremont *et al*., 1998) and the *H-2Eb* crystallography studies cover only a single allele and two antigens (Fremont *et al*., 1996).

Weaker positive selection acting at *H-2Eb* vs *H-2Aa* was also indicated by significantly fewer positively selected codons. However, in the *H-2Eb* genealogy it was possible to distinguish between a positively selected subset of alleles ($dN/dS = 6.98$) and a negatively selected subset ($dN/dS = 0.46$). The analyses that mix such subsets within groups risk obscuring both the positive signal of the first and the negative signal of the second.

A higher proportion of negatively vs positively selected codons was found for *H-2Eb* vs *H-2Aa*. This proportion might be underestimated for the above reason (but this would make the actual difference even greater). Nevertheless, only one of these codons is a putative ABS (binding to the conservative peptide backbone of an antigen (Stern *et al*., 1994), and hence this probably does not mean that *H2-Eb* has more conservative function in the binding of antigens. Instead, we suggest that it reflects higher structural constraints in the respective parts of the MHC class II β subunits than in α subunits (reviewed in Bondinas *et al*., 2007).

Selection, function and linkage disequilibrium

We have shown substantial differences in evolutionary patterns between the two discussed genes, despite their very close linkage. Close linkage implies shared history and hence presumably these differences were caused by differences in selection, which, in turn, implies different function of the two genes. For paralogous proteins arising from gene duplication, one possible fate is functional diversification (for example, see review Hughes, 2005). Immunological studies have reported functional differences between the I-E and I-A proteins (Figueroa *et al*., 1990; Cosgrove *et al*., 1992; Tomonari *et al*., 1993). More recently, it has been proposed that different selection mechanisms shape variation at the rodent orthologues of *H-2Aa* and *H-2Eb* based on population genetic and parasite genotype association studies (Bryja *et al*., 2007; Tollenaere *et al*., 2008).

The main differences found within this study are recapped as follows: (1) *H-2Eb* alleles are maintained in the gene pool for longer, manifested by higher TSO and sequence divergence, whereas *H-2Aa* shows faster allelic turnover due to higher (substitution + recombination) rates; (2) antigen binding sites of all *H-2Aa* alleles evolve under strong positive selection, whereas the pattern for *H-2Eb* is not so clear, probably because some *H-2Eb* alleles evolve under positive and others under negative selection; (3) *H-2Aa* shows higher genetic structure than *H-2Eb* between subspecies locality samples, where gene flow is expected; and (4) some *H-2Eb* alleles are not expressed.

Speculating on possible selective pressures, given the first three points above, it would appear that at least some *H-2Eb* alleles are coevolving with multiple widespread and slow-evolving antigens. Slow or cyclical evolution of these antigens would explain trans-species maintenance of *H-2Eb* alleles. Selection for optimal binding of these antigens would cause negative selection signal for some of the allelic lineages. Similarly, widespread antigen occurrence would explain the relatively uniform spatial distribution of *H-2Eb* variation. The properties of these hypothetical antigens resemble those of pathogen-associated molecular motifs common to a group of pathogen taxa and representing targets of innate immunity, for example, the lipopolysaccharide motif shared across many bacteria. For *H-2Eb* homologues, such a link with innate immunity has already been suggested (Kriener *et al*., 2000). On the other hand, fast *H-2Aa* evolution evokes a Red Queen arms race with fast-evolving antigens, in which diversification is what matters. Because our measures of population structure control for allelic diversity, we are able to compare genetic structure between the loci and conclude that *H-2Aa* is more structured. This higher structuring could reflect differences in pathogen communities between the localities, possibly because of non-synchronized fluctuations in their densities.

MHC haplotypes appear to be haploblocks of high linkage disequilibrium interspersed by recombinational hot spots (Jeffreys *et al*., 2001). The strong linkage disequilibrium that we detect between the two markers is consistent with the map distance between them, and rules out a previously undetected recombination hot spot as an explanation for the apparent difference in evolutionary history of the genes. Although close linkage between the two loci should couple their evolutionary histories, it does not prevent quicker vs slower evolution of *H-2Aa* vs *H-2Eb*, via a higher (recombination + substitution) rate in *H-2Aa*. Similarly, even if balancing selection is homogenizing *H-2Eb* allelic frequencies between locality samples (Muirhead, 2001), rapid mutational diversification of linked *H-2Aa* alleles can recover population structure at *H-2Aa*.

Conclusion

Most studies of MHC in wild-living vertebrates focus on a single MHC gene, usually the orthologue of human

DRB (that is, mouse *H-2Eb*). This choice has been influenced by the finding that *H-2Eb* is the most polymorphic MHC class II gene in humans (Janeway *et al*., 1999) and shows high levels of polymorphism in other vertebrates including rodents (see, for example, Babik *et al*. 2005 and references therein). As such, it is implicitly assumed to reflect the diversity of the whole MHC class II haplotype (Babik *et al*., 2005; Radwan *et al*., 2007).

If we had included only *H-2Eb* in our study, we would have further reinforced this general picture by stating high *H-2Eb* polymorphism. It is therefore interesting to note that *H-2Aa is* more polymorphic than *H-2Eb* in wild house mice, evolving under a different selective regime. It follows that studies analyzing causes and consequences of population structure at MHC, or dealing with MHC trans-species evolution, could reach very different conclusions depending on which MHC gene is chosen, despite even tight linkage between them. This should be kept in mind especially when study conclusions lead to management decisions, for example, regarding evolutionary significant units in conservation management. Care should also be taken in studies of MHC and mate choice, or pathogen/MHC genotype associations. An obvious example is correlation with heterozygosity: we show that individuals can differ considerably in heterozygosity even at closely linked MHC loci. We recommend careful choice of MHC gene in study design, or better, consideration of multiple MHC genes.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank all colleagues who participated in the field work and the local people who kindly allowed us to trap on their properties. We are grateful to Jakub Kreisinger for his help with statistics. We thank three anonymous reviewers for their useful comments. This study was funded by the Grant Agency of the Academy of Sciences of the Czech Republic (project no. IAA600930608) and Czech Science Foundation (206/08/0640).

References

Alcaide M, Edwards SV, Negro JJ (2007). Characterization, polymorphism, and evolution of MHC class II B genes in birds of prey. *J Mol Evol* **65**: 541–554.

Arden B, Klein J (1982). Biochemical comparison of major histocompatibility complex molecules from different subspecies of *Mus musculus*: evidence for trans-specific evolution of alleles. *Proc Natl Acad Sci USA* **79**: 2342–2346.

Babik W, Durka W, Radwan J (2005). Sequence diversity of the MHC DRB gene in the Eurasian beaver (*Castor fiber*). *Mol Ecol* **14**: 4249–4257.

Babik W, Pabijan M, Radwan J (2008). Contrasting patterns of variation in MHC loci in the Alpine newt. *Mol Ecol* **17**: 2339–2355.

Barton NH, Hewitt GM (1985). Analysis of hybrid zones. *Annu Rev Ecol Syst* **16**: 113–148.

Begovich AB, Vu TH, Jones PP (1990). Characterization of the molecular defects in the mouse E beta f and E beta q genes. Implications for the origin of MHC polymorphism. *J Immunol* **144**: 1957–1964.

Bondinas GP, Moustakas AK, Papadopoulos GK (2007). The spectrum of HLA-DQ and HLA-DR alleles, 2006: a listing correlating sequence and structure with function. *Immunogenetics* **59**: 539–553.

Braunstein NS, Germain RN (1986). The mouse E beta 2 gene: a class II MHC beta gene with limited intraspecies polymorphism and an unusual pattern of transcription. *EMBOJ* **5**: 2469–2476.

Bryja J, Galan M, Charbonnel N, Cosson JF (2005). Analysis of major histocompatibility complex class II gene in water voles using capillary electrophoresis-single stranded conformation polymorphism. *Mol Ecol Notes* **5**: 173–176.

Bryja J, Galan M, Charbonnel N, Cosson JF (2006). Duplication, balancing selection and trans-species evolution explain the high levels of polymorphism of the DQA MHC class II gene in voles (Arvicolinae). *Immunogenetics* **58**: 191–202.

Bryja J, Charbonnel N, Berthier K, Galan M, Cosson JF (2007). Density-related changes in selection pattern for major histocompatibility complex genes in fluctuating populations of voles. *Mol Ecol* **16**: 5084–5097.

Cosgrove D, Bodmer H, Bogue M, Benoist C, Mathis D (1992). Evaluation of the functional equivalence of major histocompatible complex class II A and E complexes. *J Exp Med* **176**: 629–634.

Crawley MJ (2007). *R Book*. John Wiley and Sons: Chichester.

Cucchi T, Vigne JD, Auffray JC (2005). First occurrence of the house mouse (*Mus musculus domesticus* Schwartz & Schwartz, 1943) in the western Mediterranean: a zooarchaeological revision of sub-fossil house mouse occurrences. *Biol J Linn Soc* **84**: 429–445.

Edwards SV, Chesnut K, Satta Y, Wakeland EK (1997). Ancestral polymorphism of MHC class II genes in mice: implications for balancing selection and the mammalian molecular clock. *Genetics* **146**: 655–668.

Excoffier L, Laval G, Schneider S (2005). Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online* **1**: 47–50.

Figueroa F, Gutknecht J, Tichy H, Klein J (1990). Class II MHC genes in rodent evolution. *Immunol Rev* **113**: 27–46.

Figueroa F, Tichy H, Singleton G, Franguedakis-Tsolis S, Klein J (1989). High frequency of H-2E0 alleles among wild mice. *Immunogenetics* **30**: 222–225.

Fremont DH, Hendrickson WA, Marrack P, Kappler J (1996). Structures of an MHC class II molecule with covalently bound single peptides. *Science* **272**: 1001–1004.

Fremont DH, Monnaie D, Nelson CA, Hendrickson WA, Unanue ER (1998). Crystal structure of I-Ak in complex with a dominant epitope of lysozyme. *Immunity* **8**: 305–317.

Garrigan D, Hedrick PW (2003). Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution* **57**: 1707–1722.

Golubiæ M, Budimir O, Schoepfer R, Kasahara M, Mayer WE, Figueroa F *et al.* (1987). Nucleotide sequence analysis of class II genes borne by mouse t chromosomes. *Genet Res* **50**: 137–146.

Götze D, Nadeau J, Wakeland EK, Berry RJ, Bonhomme F, Egorov IK *et al.* (1980). Histocompatibility-2 system in wild mice. X. Frequencies of H-2 and Ia antigens in wild mice from Europe and Africa. *J Immunol* **124**: 2675–2681.

Goudet J (2001). *FSTAT, a Program to Estimate and Test Gene Diversities and Fixation Indices*, version 2.9.3, Available from http://www.unil.ch/izea/softwares/fstat.html.

Guenet JL, Bonhomme F (2003). Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet* **19**: 24–31.

Hedrick PW (1999). Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**: 313–318.

Hedrick PW (2005). A standardized genetic differentiation measure. *Evolution* **59**: 1633–1638.

Hill RE, Hastie ND (1987). Accelerated evolution in the reactive center regions of serine protease inhibitors. *Nature* **326**: 96–99.

Huang SW, Yu HT (2003). Genetic variation of microsatellite loci in the major histocompatibility complex (MHC) region in the southeast Asian house mouse (*Mus musculus castaneus*). *Genetica* **119**: 201–218.

Hughes AL (2005). Gene duplication and the origin of novel proteins. *Proc Natl Acad Sci USA* **102**: 8791–8792.

Hughes AL, Nei M (1990). Evolutionary relationships of class II major-histocompatibility-complex genes in mammals. *Mol Biol Evol* **7**: 491–514.

Janeway C, Travers P, Walport D, Capra J (1999). Immunobiology: The Immune System in Health and Disease. 4th edn. Current Biology Publications: London.

Jeffreys AJ, Kauppi L, Neumann R (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**: 217–222.

Kimura M (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111–120.

Klein J (1986). *Natural History of the Major Histocompatability Complex*. John Wiley and Sons: New York.

Klein J (1987). Origin of major histocompatibility complex polymorphism—The trans-species hypothesis. *Hum Immunol* **19**: 155–162.

Klein J, Benoist C, David CS, Demant P, Lindahl KF, Flaherty L *et al.* (1990a). Revised nomenclature of mouse H-2 genes. *Immunogenetics* **32**: 147–149.

Klein J, Bontrop RE, Dawkins RL, Erlich HA, Gyllensten UB, Heise ER *et al.* (1990b). Nomenclature for the major histocompatibility complexes of different species: a proposal. *Immunogenetics* **31**: 217–219.

Klein J, O'Huigin C, Kasahara M, Vincek V, Klein D, Figueroa F (1991). Frozen haplotypes in MHC evolution. In: Klein J, Klein D (eds). *Molecular Evolution of the Major Histocompatibility Complex*. Springer Verlag: Berlin, pp 261–286.

Klein J, Satta Y, O'hUigin C, Takahata N (1993). The molecular descent of the major histocompatibility complex. *Annu Rev Immunol* **11**: 269–295.

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

Knapp LA (2007). Selection on MHC: a matter of form over function. *Heredity* **99**: 241–242.

Kriener K, O'hUigin C, Tichy H, Klein J (2000). Convergent evolution of major histocompatibility complex molecules in humans and New World monkeys. *Immunogenetics* **51**: 169–178.

Kumanovics A (2007). Genomic organization of the mouse major histocompatibility complex. In: Fox J, Barthold S, Davisson M, Newcomer C., Quimby F, Smith A (eds). *The Mouse in Biomedical Research*, 2nd edn. Academic Press: New York. Vol. 4, pp 119–135.

Lefranc MP, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G (2005). IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* **29**: 917–938.

Macholán M, Munclinger P, Šugerková M, Dufková P, Bímová B, Božíková E *et al.* (2007). Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. *Evolution* **61**: 746–771.

Macholán M, Baird SJ, Munclinger P, Dufková P, Bímová B, Piálek J (2008). Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? *BMC Evol Biol* **8**: 271.

Meirmans PG (2006). Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* **60**: 2399–2402.

Muirhead CA (2001). Consequences of population structure on genes under balancing selection. *Evolution* **55**: 1532–1541.

Musolf K, Meyer-Lucht Y, Sommer S (2004). Evolution of MHC-DRB class II polymorphism in the genus *Apodemus* and a comparison of DRB sequences within the family Muridae (Mammalia: Rodentia). *Immunogenetics* **56**: 420–426.

Nadeau JH, Britton-Davidian J, Bonhomme F, Thaler L (1988). H-2 polymorphisms are more uniformly distributed than allozyme polymorphisms in natural populations of house mice. *Genetics* **118**: 131–140.

Nei M, Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418–426.

Nei M, Gu X, Sitnikova T (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci USA* **94**: 7799–7806.

Nizetiæ D, Figueroa F, Dembiæ Z, Nevo E, Klein J (1987). Major histocompatibility complex gene organization in the mole rat *Spalax ehrenbergi*: evidence for transfer of function between class II genes. *Proc Natl Acad Sci USA* **84**: 5828–5832.

Piálek J, Vyskočilová M, Bímová B, Havelková D, Piálková J, Dufková P *et al.* (2008). Development of unique house mouse resources suitable for evolutionary studies of speciation. *J Hered* **99**: 34–44.

Piertney SB, Oliver MK (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**: 7–21.

Pond SLK, Frost SDW, Muse SV (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.

Pond SLK, Frost SDW (2005a). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22**: 1208–1222.

Pond SLK, Frost SDW (2005b). A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol* **22**: 478–485.

Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* **23**: 1891–1901.

Promerová M, Albrecht T, Bryja J (2009). Extremely high MHC class I variation in a population of a long-distance migrant, the Scarlet Rosefinch (*Carpodacus erythrinus*). *Immunogenetics* **61**: 451–461.

Radwan J, Kawałko A, Wójcik JM, Babik W (2007). MHC-DRB3 variation in a free-living population of the European bison, *Bison bonasus*. *Mol Ecol* **16**: 531–540.

Raufaste N, Orth A, Belkhir K, Senet D, Smadja C, Baird SJE *et al.* (2005). Inferences of selection and migration in the Danish house mouse hybrid zone. *Biol J Linn Soc* **84**: 593–616.

Raymond M, Rousset F (1995). GENEPOP version 1.2: population genetics software for exact tests and ecumenicism. *J Hered* **86**: 248–249.

Richman AD, Herrera LG, Nash D, Schierup MH (2003). Relative roles of mutation and recombination in generating allelic polymorphism at an MHC class II locus in *Peromyscus maniculatus*. *Genet Res* **82**: 89–99.

Saha BK, Shields JJ, Miller RD, Hansen TH, Shreffler DC (1993). A highly polymorphic microsatellite in the class II Eb gene allows tracing of major histocompatibility complex evolution in mouse. *Proc Natl Acad Sci USA* **90**: 5312–5316.

Schad J, Sommer S, Ganzhorn JU (2004). MHC variability of a small lemur in the littoral forest fragments of Southeastern Madagascar. *Conserv Genet* **5**: 299–309.

Scott CA, Peterson PA, Teyton L, Wilson IA (1998). Crystal structures of two I-Ad-peptide complexes reveal that high affinity can be achieved without large anchor residues. *Immunity* **8**: 319–329.

Seddon JM, Baverstock PR (2000). Evolutionary lineages of RT1.Ba in the Australian *Rattus*. *Mol Biol Evol* **17**: 768–772.

Smulders MJM, Snoek LB, Booy G, Vosman B (2003). Complete loss of MHC genetic diversity in the Common Hamster (*Cricetus cricetus*) population in the Netherlands. Consequences for conservation strategies. *Conserv Genet* **4**: 441–451.

Sommer S (2003). Effects of habitat fragmentation and changes of dispersal behaviour after a recent population decline on the genetic variability of noncoding and coding DNA of a monogamous Malagasy rodent. *Mol Ecol* **12**: 2845–2851.

Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL *et al.* (1994). Crystal structure of the human

class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* **368**: 215–221.

Tacchini-Cottier F, Mayer WE, Begovich AB, Jones PP (1995). Inactivation of E alpha and E beta expression in inbred and wild mice by multiple distinct mutations, some of which predate speciation within *Mus* species. *Int Immunol* **7**: 1459–1471.

Tollenaere C, Bryja J, Galan M, Cadet P, Deter J, Chaval Y *et al.* (2008). Multiple parasites mediate balancing selection at two MHC class II genes in the fossorial water vole: insights from multivariate analyses and population genetics. *J Evol Biol* **21**: 1307–1320.

Takahashi K, Rooney AP, Nei M (2000). Origins and divergence times of mammalian class II MHC gene clusters. *J Hered* **91**: 198–204.

Takahata N, Satta Y, Klein J (1992). Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* **130**: 925–938.

Takahata N, Satta Y (1998). Selection convergence, and intragenic recombination in HLA diversity. *Genetica* **102-3**: 157–169.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4 0. *Mol Biol Evol* **24**: 1596–1599.

Tomonari K, Fairchild S, Rosenwasser OA (1993). Influence of viral superantigens on V beta- and V alpha-specific positive and negative selection. *Immunol Rev* **131**: 131–168.

Tucker PK, Sage RD, Warner J, Wilson AC, Eicher EM (1992). Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. *Evolution* **46**: 1146–1163.

Van Valen L (1973). A new evolutionary law. *Evol Theory* **1**: 1–30.

Vu TH, Tacchini-Cottier FM, Day CE, Begovich AB, Jones PP (1988). Molecular basis for the defective expression of the mouse Ew17 beta gene. *J Immunol* **141**: 3654–3661.

Wilson DJ, McVean G (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**: 1411–1425.

Zhang Z, Yu J (2006). Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. *Genomics Proteomics Bioinformatics* **4**: 173–181.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)