

# Diversifying and Purifying Selection in the Peptide Binding Region of DRB in Mammals

Rebecca F. Furlong · Ziheng Yang

Received: 31 January 2008 / Accepted: 21 February 2008 / Published online: 18 March 2008  
© Springer Science+Business Media, LLC 2008

**Abstract** The class II genes of the major histocompatibility complex encode proteins which play a crucial role in antigen presentation. They are among the most polymorphic proteins known, and this polymorphism is thought to be the result of natural selection. To understand the selective pressure acting on the protein and to examine possible differences in the evolutionary dynamics among species, we apply maximum likelihood models of codon substitution to analyze the *DRB* genes of six mammalian species: human, chimpanzee, macaque, tamarin, dog, and cow. The models account for variable selective pressures across codons in the gene and have the power to detect amino acid residues under either positive or negative selection. Our analysis detected positive selection in the *DRB* genes in each of the six mammals examined. Comparison with structural data reveals that almost all amino acid residues inferred to be under positive selection in humans are in the peptide binding region (PBR) and are in contact with the antigen side chains, although residues outside of but close to the PBR are also detected. Strong purifying selection is also detected in the PBR, at sites which contact the antigen and at sites which may be involved in dimerization or T cell binding. The analysis demonstrates the utility of the random-sites analysis even when structural information is available. The different mammalian species are found to share many positively or

negatively selected sites, suggesting that their functional roles have remained very similar in the different species, despite the different habitats and pathogens of the species.

**Keywords** Major histocompatibility complex · Positive selection · Purifying selection · DRB · Darwinian selection

## Introduction

The major histocompatibility complex (MHC) is a large genomic region found in most vertebrates, which contains many genes with roles in the immune system. One subset of these genes is a family of glycoproteins involved in the recognition and binding of foreign peptides. In humans, they are referred to as human leukocyte antigen (HLA) genes. The family is divided into class I and class II; both collect polypeptides from inside the cell and display them on the cell surface for recognition by T cells. While class I proteins are expressed on all nucleated cell surfaces, class II proteins are expressed only on B lymphocytes, macrophages, and other antigen-presenting cells of the immune system. In mammals, and in many other vertebrates, the two classes are linked in a single gene complex (Kulski et al. 2002; Ohta et al. 2000). In humans, this complex is found on chromosome 6. The crystal structures of the human MHC class I and II proteins are very similar, having two  $\alpha$ -helices flanking a  $\beta$ -sheet to create a groove in which peptides are bound and presented (Bjorkman et al. 1987a; Brown et al. 1988, 1993). This groove contains a number of residues responsible for antigen binding, which are known collectively as the antigen recognition site or peptide-binding region (PBR). The class II protein is a heterodimer, containing an  $\alpha$  and a  $\beta$  chain, both of which contribute one  $\alpha$ -helix and approximately half the  $\beta$ -sheet to

---

R. F. Furlong · Z. Yang (✉)  
Department of Biology, University College London, Darwin  
Building, Gower Street, London WC1E 6BT, UK  
e-mail: z.yang@ucl.ac.uk

*Present Address:*  
R. F. Furlong  
Department of Zoology, University of Oxford, South Parks  
Road, Oxford OX1 3PS, UK

the PBR. Two heterodimers then form a superdimer, a process thought to be important in initiating T cell signaling (Brown et al. 1993). There are three polymorphic class II loci used in this process, known as *HLA-DP*, *HLA-DQ*, and *HLA-DR* (abbreviated *DR* throughout), and each contains one  $\alpha$  gene (abbreviated A) and at least one  $\beta$  gene (abbreviated B, such as *DRB*). Following the publication of an initial crystal structure (Brown et al. 1993), Stern et al. (1994) examined the MHC class II heterodimer protein DR1 (the product of one *DRA1* allele and one *DRB1* allele) and identified 19 amino acid sites from the DRA molecule and 20 sites from the DRB molecule which were in contact with an influenza virus peptide and therefore putatively formed the PBR. Most of these residues line “pockets” within the groove (some residues even line two pockets simultaneously), which accommodate the side chains of the antigen and thus are probably responsible for the different peptide specificities of different alleles.

The MHC class I and II proteins exhibit a remarkable degree of polymorphism (e.g., Bodmer 1972). The role of the proteins in antigen binding and presentation led to an early suggestion that the high level of polymorphism is maintained by heterozygote advantage, due to selective pressures exerted by the host immunological surveillance mechanism (Doherty and Zinkernagel 1975). The reasoning was that mammalian populations are subject to a wide range of diseases, and therefore maintenance of a variety of different alleles would be essential for the population to mount an immune response to any disease. The PBR contains many polymorphic sites (Bjorkman et al. 1987a, b), and overdominant selection (heterozygote advantage) in this region is suggested as the most probable type of balancing selection (Hughes and Nei 1989).

Adaptive changes in protein-coding genes may be detected by comparing the numbers of synonymous ( $d_S$ ) and nonsynonymous ( $d_N$ ) substitutions per site, with the ratio  $\omega = d_N/d_S$  indicating selection on the protein. Neutral, deleterious, and advantageous nonsynonymous mutations are signified by  $\omega = 1$ ,  $\omega < 1$ , and  $\omega > 1$ , respectively. Early studies based on simple pairwise comparisons examined sequences for an excess of nonsynonymous changes, with positive selection detected only if the average  $\omega$  over the whole sequence is  $>1$ . Since most sites in a protein are probably under selective constraints (with  $\omega \ll 1$ ), this method lacks power. If prior information such as the crystal structure of the protein is available to predict a subset of sites which may be under positive selection, one can focus on those sites. This approach has been taken by Hughes and Nei (1988, 1989) and Hughes et al. (1994) to analyze the PBR in MHC class I and II proteins in humans and mice. The crystal structures (Bjorkman et al. 1987a; Brown et al. 1993) were used to identify 57 residues in class I proteins and 44 residues in

class II proteins, which constitute the PBR. The authors then analyzed those PBR codons and detected accelerated nonsynonymous substitution rates, in both MHC class I and MHC class II, while no such signal was detectable when the whole gene was used to calculate  $d_S$  and  $d_N$ . These data have led to the MHC being considered a classic example of positive selection. However, identification of amino acids in the PBR is not a precise exercise. Furthermore, positive selection is not necessarily limited to the PBR, nor is every amino acid in the PBR expected to be under positive selection. Indeed previous analyses of the class I genes identified both non-PBR residues that are under positive selection and PBR residues that are highly conserved and under purifying selection (Suzuki and Gojobori 1999; Yang and Swanson 2002).

The “random-sites” models (Nielsen and Yang 1998; Yang et al. 2000) use a statistical distribution to describe the variation among sites in the selective pressure indicated by  $\omega$  but do not assume *a priori* which sites may be under positive selection. Maximum likelihood is used to estimate parameters in models of codon substitution, and sites under positive selection are identified using the Bayes empirical Bayes (BEB) approach (Yang et al. 2005). Yang and Swanson (2002) used a random-sites method to examine human MHC class I genes, and the results suggested extreme variability in selective pressure among sites in MHC class I alleles. The authors identified 22 of the 53 sites in the PBR, as well as 3 additional sites outside but close to the PBR, to be under positive selection. While the 25 sites identified to be under positive selection are scattered along the primary sequence, they are all clustered in the PBR in the tertiary structure. Highly conserved sites were also found within the PBR.

In this study we apply random-sites models to MHC class II genes from humans for the first time. We ask whether the pattern found for class I genes also applies to class II. For example, are amino acid sites under positive selection in class II genes clustered in the PBR, as in the class I genes? Furthermore, many class II alleles have been sequenced from nonhuman mammalian species which have not previously been analyzed using this method and in comparison with each other. We investigate whether the same sets of sites are under diversifying or purifying selection across mammalian species. Despite the high allelic diversity, class II genes have a high sequence similarity outside the PBR across mammalian species, indicating that the structure and function of the protein have probably remained unchanged among mammals.

#### Data and Methods

It is not possible to align class II  $\alpha$  and  $\beta$  genes together reliably, particularly over the hypervariable PBR region.

We concentrated on the *DRB* subset of  $\beta$  genes, which are the most polymorphic genes in class II. They consist of a group of up to nine genes and pseudogenes, the number varying between haplotypes (Arvidsson et al. 1995), and when the bona fide human *DRB* genes *DRB1*, *DRB3*, *DRB4*, and *DRB5* are considered together, hundreds of alleles are known. These loci have been studied in a range of mammals, most of which also have several genes and pseudogenes within the locus. Only three alleles were available for the corresponding  $\alpha$  gene in humans (*DRA*), making it unsuitable for a similar analysis.

Human *DRB* gene sequences were obtained from the International Immunogenetics Information System HLA sequence database (<http://www.ebi.ac.uk/imgt/hla>). Mammalian *DRB* sequences were obtained from the Immuno Polymorphism Database (<http://www.ebi.ac.uk/ipd/mhc>). Five nonhuman species were chosen for data availability. They represent a variety of mammals, including nonhuman apes (chimpanzee, *Pan troglodytes*), Old World monkeys (rhesus macaque, *Macaca mulatta*), New World monkeys (cotton-top tamarin, *Saguinus oedipus*), carnivores (dog, *Canis familiaris*), and cetartiodactyls (cow, *Bos taurus*). Pseudogene data were not included, and data for all true genes were collected into a single dataset for each species. We note that use of orthologues and paralogues in the same dataset should not invalidate maximum likelihood analysis under the site models since the phylogenetic tree can reflect the relationships among the sequences due to speciations, gene duplications, and between-allele gene conversions. Within-locus gene conversion will have a similar impact as recombination, which is considered below. Each dataset contained  $\geq 40$  different *DRB* sequences. The human dataset is by far the largest. Summary information for each dataset is reported in Table 1.

Sequence data of each species were first analyzed separately. In view of the fact that some polymorphisms at the MHC II loci might have originated prior to divergence of the species, we also analyzed a combined dataset including sequences from all six species. Due to the nature of the

available data, the sequences for different species are of different lengths, and the maximal segment shared across all species is used in the combined dataset. The translated amino acid sequences were aligned using ClustalW (Thompson et al. 1994), and the nucleotide sequence alignment was constructed accordingly, followed by slight manual adjustments. Many of the sequences were not full-length, but all include the hypervariable second exon of the gene, which contains the PBR. Missing data were treated as “unknown,” to allow information from full-length sequences to be retained. The alignments are available from the authors upon request.

Phylogenetic trees were inferred using neighbor joining (NJ; Saitou and Nei 1987), with distances calculated under the K80 model (Kimura 1980), implemented in the PHYLIP package (Felsenstein 2005). The NJ tree topologies were used but the branch lengths were re-estimated under codon models using the CODEML program in the PAML package (Yang 1997, 2007). Previous studies (e.g., Yang et al. 2000) have suggested that minor differences in the tree topology make little impact on inference of positive selection under site models (see also below). For recombination analysis using exon 2 of human sequences, the tree was inferred using PHYML (Guindon and Gascuel 2003).

Codon-based likelihood analysis was conducted under the random-sites models: M0 (one ratio), M1a (neutral), M2a (selection), M7 (beta), and M8 (beta& $\omega$ ) (Nielsen and Yang 1998; Yang et al. 2000, 2005). Model M0 (one ratio) assumes one  $\omega$  ratio for all codons in the sequence. This is the simplest model and can be used to check that parameter estimates in more complex models are consistent. M1a (neutral) assumes two site classes, with  $\omega_0 < 1$ , estimated from the data, and  $\omega_1 = 1$ , respectively. M2a (selection) adds a third site class to M1a, with  $\omega_2 > 1$  estimated from the data. M1a and M2a are nested and their log likelihood values may be compared using a likelihood ratio test (LRT) to test for positive selection (that is, presence of sites with  $\omega_2 > 1$ ). M7 (beta) is a flexible null model, in which the  $\omega$  ratio for a codon is a random draw from the  $\beta$  distribution,

**Table 1** Summary statistics of DRB datasets analyzed in this paper

Species	No. sequences	No. codons	Tree length	$\hat{\kappa}$ (M0)	$\hat{\omega}$ (M0)
Human ( <i>Homo sapiens</i> )	247	237	5.04	1.77	0.61
Chimpanzee ( <i>Pan troglodytes</i> )	49	91	3.24	1.47	0.65
Macaque ( <i>Macaca mulatta</i> )	92	86	8.29	1.34	0.53
Tamarin ( <i>Saguinus oedipus</i> )	40	90	4.11	1.44	0.56
Cow ( <i>Bos taurus</i> )	102	229	6.18	1.17	0.85
Dog ( <i>Canis familiaris</i> )	41	232	1.58	1.36	2.05
All	571	94	28.95	1.43	0.60

*Note:* Tree length is defined as the sum of branch lengths of the tree, measured by the number of nucleotide substitutions per codon. Tree length and estimates of parameters  $\kappa$  and  $\omega$  are obtained under model M0 (one ratio)

with  $0 < \omega < 1$ . M8 (beta& $\omega$ ) adds an extra site class to M7, with  $\omega_s > 1$  estimated from the data. Comparison between M7 and M8 constitutes another LRT of positive selection. The CODEML program is also used to calculate the posterior probability that each site falls into the different site classes using the BEB approach (Yang et al. 2005). We used a stringent cutoff of  $p > 0.99$  to identify sites under positive selection.

The program PLATO (Grassly and Holmes 1997) was used to analyze exon 2 of the human alleles to detect recombination. This exon contains the PBR. PLATO uses a sliding window to detect “anomalous” regions in the alignment that do not fit the evolutionary model for the complete dataset. Such regions may be the result of recombination, but sites with conflicting rates or branch lengths may also be detected if rate heterogeneity is not allowed for in the model. Two models were used in the analysis: one assuming a single rate of evolution at all sites and another assuming the gamma model of variable rates among sites (Yang 1994). In the latter model, the shape parameter  $\alpha$  was estimated using PHYML. Further recombination analysis was carried out using MaxChi2 (Smith 1992), Geneconv (Sawyer 1989), and RDP (Martin and Rybicki 2000), implemented in the RDP2 program (Martin et al. 2005). The default parameters were used in each case.

## Results

### Positive Selection in the DRB Genes of Every Species

Maximum likelihood estimates of parameters under codon models of variable  $\omega$ s across sites in the human dataset are listed in Table 2. Model M0 (one ratio), which assumes a single  $\omega$  for all codons in the sequence, does not find evidence for positive selection ( $\hat{\omega} = 0.61$ ). Evolution of DRB, like almost every other protein, is on average dominated by purifying selection removing nonsynonymous mutations. Models M2a (selection) and M8 (beta& $\omega$ ), which allow selection at a subset of sites, fit the data significantly better

than models M1a (neutral) and M7 (beta), which do not allow for such sites. Parameter estimates under M2a (selection) suggest 6% of sites to be under positive selection with  $\hat{\omega}_2 = 4.7$ . Parameter estimates under M8 (beta& $\omega$ ) suggest 10% of sites to be under positive selection with a slightly lower  $\hat{\omega}_s = 3.9$ . The BEB posterior probabilities and posterior means of  $\omega$  for sites under model M8 (beta& $\omega$ ) are shown in Fig. 1. Ten sites were inferred to have  $\omega > 1$  with high posterior probabilities ( $p > 0.99$ ) under M8: 11L, 13F, 37S, 47Y, 57D, 67L, 71R, 74A, 86G, and 96E. Site numbers and amino acids refer to the reference sequence DRB1\*1010101 (PDB file 1DLH, chain B). These sites are mapped onto the crystal structure in Fig. 2.

To test whether the tree topology affected the inference of sites under positive selection, a tree was constructed using the alignment of exon 2 only. Use of this tree to analyze the full-length alignment identified the same sites as before, with very similar parameter estimates (results not shown). Thus our results are robust to minor changes to the tree topology. Similar robustness of the site-based analysis to the assumed tree topology was noted in previous analyses (e.g., Yang and Swanson 2002).

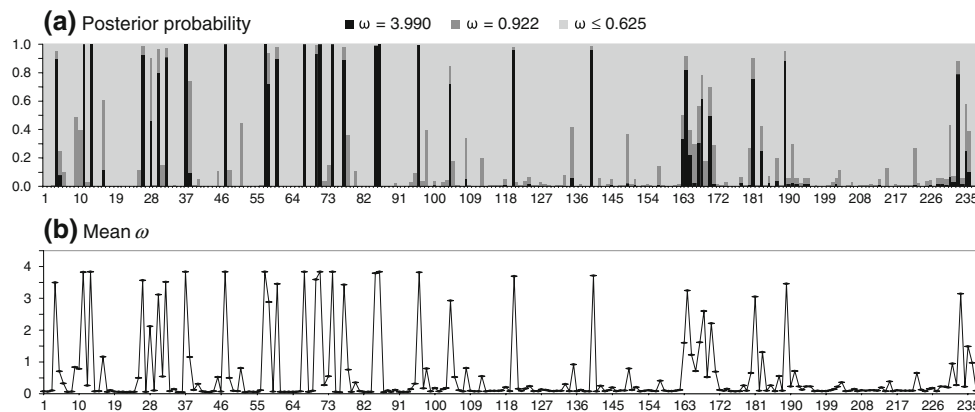
The data for the five other mammalian species were analyzed using the same method. In every dataset, models M2a (selection) and M8 (beta& $\omega$ ), which allow for sites under positive selection, fitted the data significantly better than models M1a (neutral) and M7 (beta), which do not allow for such sites. Positive selection sites identified under model M8 (beta& $\omega$ ) from data of different species are summarized in Fig. 3. The residues are numbered according to the human reference sequence. We also calculated the correlation coefficients between species in the posterior means of  $\omega$  (Table 3). Of the 10 sites found to be under positive selection in the human dataset, nine are also detected in at least one other species. However, the number of sites detected in each species is variable, possibly reflecting the size and information content of the datasets.

The combined dataset for all six mammalian species was analyzed similarly. The phylogenetic tree, constructed using the PHYML program (Guindon and Gascuel 2003), is shown

**Table 2** Log-likelihood values and parameter estimates under random-sites models for human DRB alleles

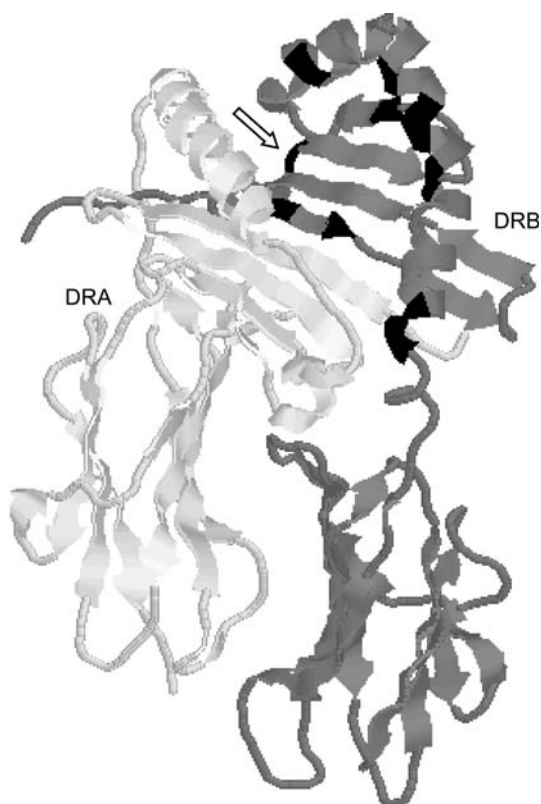
Model	$\ell$	Parameter estimate(s)	Positively selected sites
M0 (one ratio)	-6096.65	$\hat{\omega} = 0.613$	None
M1a (neutral)	-5687.96	$\hat{p}_0 = 0.781, (\hat{p}_1 = 0.219), \hat{\omega} = 0.045, \omega_1 = 1$	Not allowed
M2a (selection)	-5559.35	$\hat{p}_0 = 0.720, \hat{p}_1 = 0.219 (\hat{p}_2 = 0.060), \hat{\omega}_0 = 0.045, \omega_1 = 1, \hat{\omega}_2 = 4.709$	11L, 13F, 37S, 47Y, 57D, 67L, 71R, 74A, 86G
M7 (beta)	-5964.51	$\hat{p} = 0.104, \hat{q} = 0.334$	Not allowed
M8 (beta& $\omega$ )	-5563.48	$\hat{p}_0 = 0.900, \hat{p} = 0.171, \hat{q} = 0.626, (\hat{p}_1 = 0.100), \hat{\omega}_s = 3.990$	11L, 13F, 37S, 47Y, 57D, 67L, 71R, 74A, 86G, 96E, 120S, 140A

*Note:* Positive selection sites are identified at the cutoff  $p > 95\%$ , with those with  $p > 99\%$  shown in boldface. Estimates of  $\kappa$  range from 1.5 to 1.8 among models



**Fig. 1** (a) Posterior probabilities of site classes for codons along the *DRB* gene under model M8 (beta& $\omega$ ) for the human dataset. The 11  $\omega$  ratios are 0.0000, 0.0000, 0.0006, 0.0045, 0.0196, 0.0624, 0.1605, 0.3454, 0.6247, 0.9217, and  $\omega_s = 3.9897$ . The first 10 categories are from the beta distribution, each with proportion 0.0899, and the last category has proportion 0.1010 (see Table 2). The first nine categories are collapsed into one, represented as  $\omega < 0.625$ . Sites for which the

posterior probability for the site class of positive selection (with  $\omega_s$ ) exceeds 0.95 are listed in Table 2 as positively selected sites. (b) Approximate posterior means of  $\omega$ , calculated as a weighted average of  $\omega$  over the 11 site classes, weighted by the posterior probabilities. Sites with low mean  $\omega$ 's are inferred to be under purifying selection. Sites are numbered according to the reference sequence DRB1\*1010101



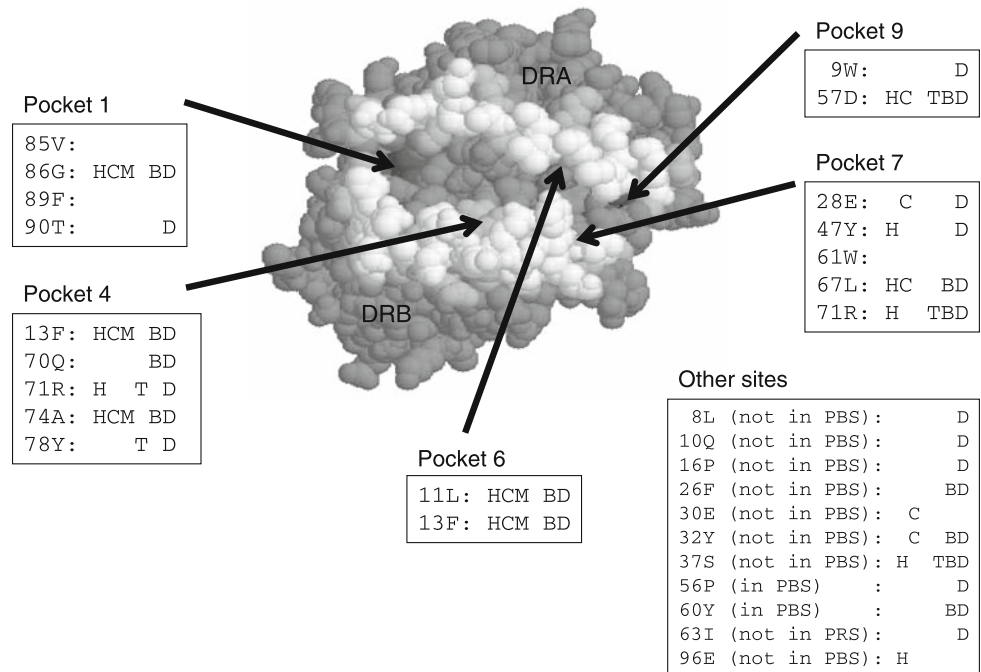
**Fig. 2** Structure of the MHC allele DRA, DRB1\*1010101 (PDB file 1DLH, chains A and B). DRA is not analyzed in this paper and is colored pale gray. DRB is shaded dark gray, and amino acid residues identified to be under positive selection ( $p > 0.99$ ) under model M8 (beta& $\omega$ ) are shown in black. All sites detected with a high probability to be under positive selection are located in the PBR. Residues 37S and 96E, indicated with an arrow, were detected with a high posterior probability but are not thought to be in contact with the antigen

in Fig. 4. The cow and dog alleles form separate monophyletic clades but sequences from other species do not form strictly monophyletic clades, even though the tree is highly structured, supporting the notion that MHC II polymorphism in mammals may be transspecies (Lundberg and McDevitt 1992; Musolf et al. 2004; Takahata and Nei 1990). Positive selection was detected in both the M1a-M2a and the M7-M8 comparisons (results not shown), demonstrating the presence of sites under positive selection. Both model M2a (selection) and model M8 (beta& $\omega$ ) identified nine sites under positive selection at the  $p > 0.99$  cutoff: 11L, 13F, 32Y, 37S, 57D, 67L, 71R, 74A, and 86G. This list is very similar to the list based on the human data alone (Table 2).

#### Amino Acid Residues in the DRB Under Purifying Selection

Despite the extreme polymorphism found within the *DRB* genes, the overall structure of the protein remains highly conserved. For example, *DRB* sequences from different mammalian sequences can also be easily aligned together. Purifying selection must therefore have had an essential role in the evolution of the protein. Even the hypervariable PBR includes highly conserved amino acids. Thus we examined whether the same set of amino acids in the PBR are under purifying selection across the mammalian species. Of the 20 residues that have contact with the antigen in the human DRB protein, four sites have posterior mean  $\omega < 0.1$  under model M8 (beta& $\omega$ ) and are inferred to be undergoing strong purifying selection in humans. These are 61W, 81H, 82N, and 89F (Fig. 1b and Table 4). At sites 82N and 89K, the mean  $\omega$  is  $< 0.2$  across all the species considered (Table 4),

**Fig. 3** Top view of the DRA/DRB molecule indicating antigen-binding pockets. The  $\alpha$  helices are shaded white. The DRB residues that line the pockets, and the species in which positive selection was detected at that site, are listed. H, human; C, chimpanzee; M, macaque; T, tamarin; B, cow (bovine); D, dog



**Table 3** Pearson correlation coefficients of posterior mean  $\omega$  between species, estimated under model M8 (beta& $\omega$ ) for exon 2

	Chimp	Macaque	Tamarin	Cow	Dog
Human	0.845	0.661	0.700	0.675	0.807
Chimp		0.679	0.780	0.679	0.734
Macaque			0.615	0.631	0.771
Tamarin				0.646	0.669
Cow					0.716

whereas the mean  $\omega$  at residue 81H varies from 0.62 in the cow to 0.09 in the human and the mean  $\omega$  at residue 61W varies from 2.53 in the tamarin to 0.06 in the human.

Sites outside the PBR may be expected to be under various levels of selective constraint. Therefore we considered only the 10 sites with the lowest  $\omega$  values under M8 for each species (data not shown). There is a strong overlap between species, consistent with the fairly strong correlations in Table 3. Eight sites are detected with a very low mean  $\omega$  in at least four of the mammal species: 22E, 35E, 36E, 46E, 52E, 54G, 76D and 87E.

**Detection of Recombination**

It has been suggested that both inter- and intralocus recombination may play a role in *HLA* gene family evolution (She et al. 1991). Several recombination-detection methods were thus applied to the human dataset. PLATO (Grassly and Holmes 1997) detected six anomalous regions when the NJ tree was used, as reported in Table 5. Since selection acts most strongly on first and second codon positions, analysis is

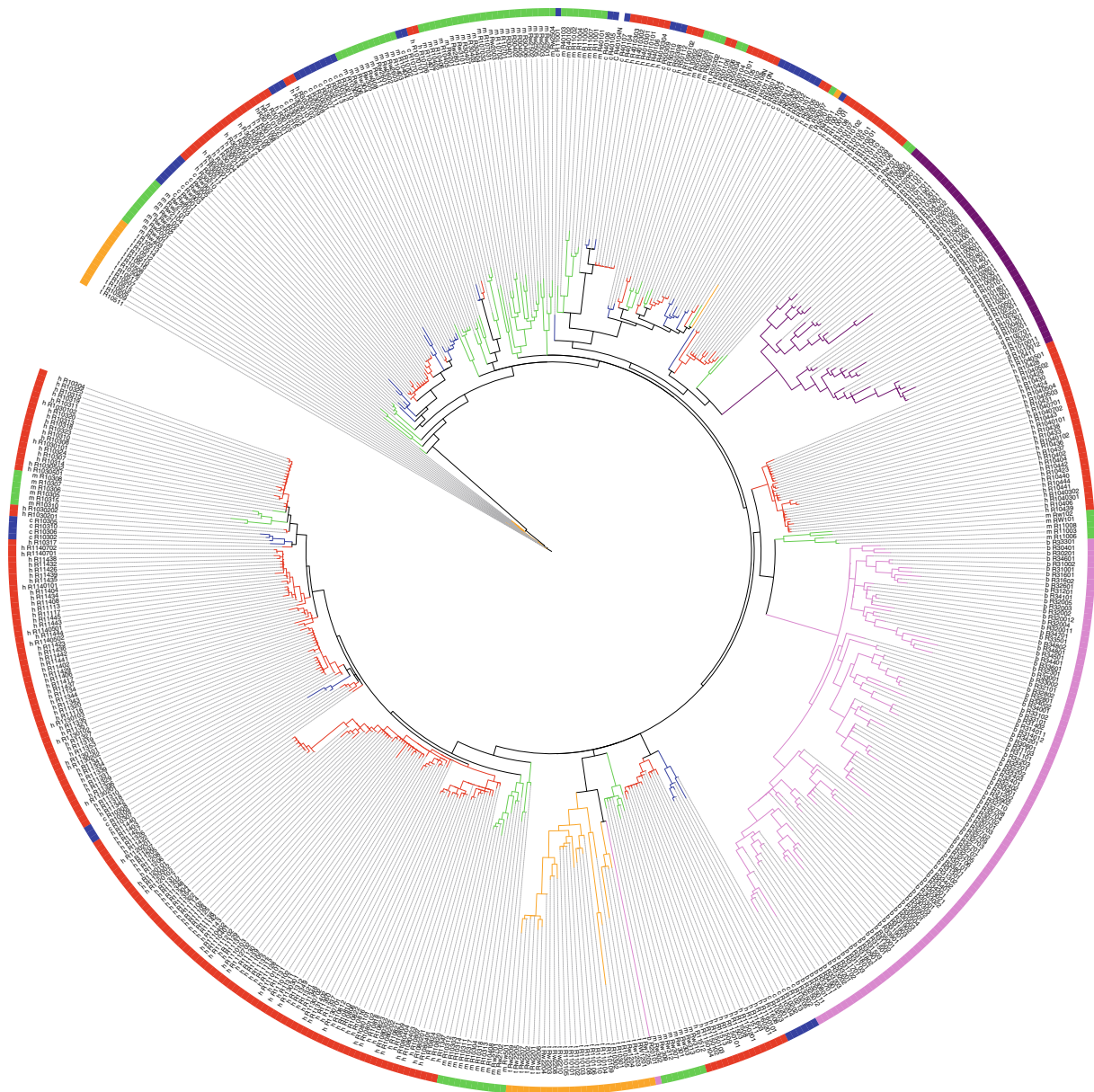
often carried out using only third positions to minimize the effect of selection upon the method. However, we used all three codon positions of exon 2, because the exon is rather short, and using only third positions would seriously compromise the phylogenetic information in the data. The ML estimates of parameters were  $\hat{\kappa} = 1.5$  for the transition/transversion rate ratio and  $\hat{\alpha} = 0.53$  for the gamma shape parameter. The same anomalous regions were identified when the ML tree was used.

It is known that PLATO may detect fast-evolving sites under positive selection as anomalous regions, even when no recombination has taken place (Grassly and Holmes 1997). Thus we used three additional methods: MaxChi2 (Smith 1992), Geneconv (Sawyer 1989), and RDP (Martin and Rybicki 2000), all implemented in the RDP2 program (Martin et al. 2005). MaxChi2 and Geneconv are so-called “substitution” methods and do not rely on a phylogeny. RDP is a phylogenetic method based on pairwise comparison and calculates the probability of recombination using a binomial model. All three methods have previously been tested on real and simulated datasets and found to be reliable (Posada and Crandall 2001). No recombination was detected by any of these methods.

**Discussion**

**Diversifying Selection in Mammalian DRB Alleles**

ML analysis of *DRB* alleles in humans revealed considerable variation in selective pressure among sites. Ten sites



**Fig. 4** The maximum likelihood tree for all 571 MHC alleles from the six mammalian species, reconstructed using the PHYML program (Guindon and Gascuel 2003) under the default HKY85 nucleotide-substitution model (Hasegawa et al. 1985). The tree is generated using the iTOL tool (<http://www.itol.embl.de/>). Sequences from the same species are represented using the same color: red for human, blue for chimpanzee, green for macaque, orange for tamarin, pink for cow, and purple for dog. The dog alleles form a monophyletic group.

were inferred to be under positive selection with high posterior probabilities ( $>0.99$ ) under model M8 (beta& $\omega$ ). Model M2a (selection) is more stringent but produced similar sites and  $\omega$  values to M8. These values are comparable to those obtained for human MHC class I genes (Yang and Swanson 2002).

Twenty-four codons from the *DRB* gene were originally suggested to form the PBR: 9W, 11L, 13F, 28E,

One cow sequence (b\_R20101) is from locus 2, while all other cow sequences are from locus 3, and they form a monophyletic clade. Sequences from other species do not form monophyletic groups. The branch lengths, measured by the expected number of nucleotide substitutions per codon, are re-estimated using the CODEML program (Yang 1997, 2007) under model M0 (one ratio). The parameter estimates under this model are  $\hat{\kappa} = 1.43$  and  $\hat{\omega} = 0.60$

30C, 32Y, 37S, 38V, 47Y, 56P, 60Y, 61W, 65K, 68L, 70Q, 71R, 74A, 78Y, 81H, 82N, 85V, 86G, 88S, and 89F (Brown et al. 1993). Hughes et al. (1994) used these 24 putative PBR codons to calculate  $d_S$  and  $d_N$  within the PBR, obtaining a ratio of  $\omega = 4.8$ . The  $\omega$  values we found for the subset of positively selected sites are remarkably similar to that estimate, given the small size of their dataset (32 alleles).

**Table 4** Posterior mean  $\omega$  for sites within the PBR, calculated using BEB under M8 (beta& $\omega$ )

Site	Human	Chimp	Macaque	Tamarin	Cow	Dog
9W	0.84	1.64	0.80	0.50	0.53	10.18
11L	3.83	3.38	4.56	2.64	5.00	10.25
13F	3.83	3.38	4.56	2.61	5.00	10.25
28E	2.12	3.36	0.94	2.31	0.81	10.25
47Y	3.83	0.75	0.93	0.96	0.92	10.25
56P	0.11	0.26	3.54	1.36	4.66	10.25
57D	3.83	3.38	3.80	2.66	5.00	10.25
60Y	3.46	1.13	0.94	0.82	5.00	10.25
61W*	0.06	1.30	0.58	2.53	0.79	0.10
67L	3.83	3.38	1.13	2.63	5.00	10.25
70Q	3.60	2.89	0.86	2.59	5.00	10.25
71R	3.83	3.14	0.94	2.65	4.95	10.25
74A	3.83	3.38	4.56	1.68	5.00	10.25
78Y	0.76	3.12	0.89	2.66	0.88	10.25
81H*	0.09	0.16	0.44	0.13	0.62	0.09
82N*	0.06	0.07	0.06	0.15	0.06	0.09
85V	3.79	3.31	0.97	2.26	0.88	0.09
86G	3.83	3.38	4.56	2.47	5.00	10.25
89F*	0.06	0.09	0.05	0.18	0.17	0.10
90T	0.13	0.81	0.56	2.21	0.60	10.25

*Note:* Residues which contact the antigen are in boldface. Residues under purifying selection in humans ( $\omega < 0.1$ ) are indicated with an asterisk

**Table 5** Anomalous regions detected by PLATO in the human DRB dataset

Base (codon) positions	$z$ Value
(a) Model of one rate for all sites	
254–258 (84–86)	41.65
170–174 (56–58)	38.32
109–113 (36–37)	32.36
199–221 (66–73)	27.97
230–234 (76–78)	12.28
139–143 (46–47)	6.97
(b) Model of $\Gamma$ rates for sites	
170–174 (56–58)	8.13
254–258 (84–86)	7.66
109–113 (36–37)	6.40
199–221 (66–73)	5.26

*Note:* The critical value for the  $z$  value is 3.565

A later analysis of DR1 crystal structure by Stern et al. (1994) listed 20 codons that made contact with the antigen—9W, 11L, 13F, 28E, 47Y, 56P, 57D, 60Y, 61W, 67L, 70Q, 71R, 74A, 78Y, 81H, 82N, 85V, 86G, 89F, and 90T—with the residues underlined forming pockets for the antigen side chains. This list is in general agreement with recent

summaries of important residues in the *HLA-DRB* by Bondinas et al. (2007). Here we use the new 2007 list of Bondinas et al. for comparison with our analysis under the random-sites models, since contact with the antigen is essential for the role of the PBR. The 10 sites inferred to be under positive selection under M8 (beta& $\omega$ ) with  $p > 0.99$  are 11L, 13F, 37S, 47Y, 57D, 67L, 71R, 74A, 86G, and 96E. The first nine of them are “pocket” sites; site 37S was not identified as a pocket site by Stern, but is now listed as being in pocket 9 by Bondinas et al. (2007). The only exception is 96E, indicated with an arrow in Fig. 2, which is outside the PBR and not identified as being in contact with the antigen or as having any other important role. It is likely that this site relates to an interaction or structural variant that has not yet been investigated, as the 2007 list is based on eight different allele/antigen combinations. Therefore we conclude that positive selection is acting within antigen side-chain pockets of the human DRB protein. This is in agreement with the theory that the high level of polymorphism found in this gene is driven by positive selection.

The majority of the sites inferred to be under positive selection in other primates are also under selection in humans. One exception is residue 78Y, which is detected in the cotton-top tamarin but not in any other primate examined. In humans, this site has a posterior probability of 0.00 of being under positive selection and a mean  $\omega$  of 0.70 under model M8 and therefore is not inferred to be under positive selection. However, this site corresponds to an antigen-pocket residue and selection at this site appears to be unique to the tamarin among the primates. Outside the PBR, three sites are inferred to be under positive selection in chimpanzee and not in any other primate: 28E, 30C, and 32Y. In humans, sites 30C and 32Y have posterior probabilities of 0.66 and 0.98 and mean  $\omega$  values of 3.01 and 4.06, respectively, under M8. Therefore these sites are very likely to be under positive selection in humans as well. In contrast, site 28E has a posterior probability of 0.06 and a mean  $\omega$  of 1.09 in humans. Site 28E is an antigen-pocket residue.

Sites inferred to be under positive selection in the bovine dataset were also similar to those detected in humans. The majority of sites detected at a high posterior probability ( $>0.99$ ) under model M8 are within the antigen side chain pockets and are also selected in humans and other primates. The exceptions are 26F, which is a pocket site (Bondinas et al. 2007) and 60Y, which is in contact with the antigen although it is not a pocket site; neither of these sites were detected as being under positive selection in any primate at the  $p > 0.99$  cutoff. However, both sites have high mean  $\omega > 3.8$  and posterior probability ( $p > 0.90$ ) in humans and are very likely to be under positive selection in humans as well.

The canine dataset was unusual in that evidence for positive selection was found under model M0 (one ratio); that is, the  $\omega$  ratio averaged over all sites and sequences is 2.05. A similar number of sites is detected to be undergoing positive selection as in other mammals examined here, but the  $\omega$  estimate for the positive selection sites is much higher, at 12.02 under M8, compared with 3.99 (human) and 5.03 (bovine). Some sites are inferred to be undergoing positive selection in the canine dataset but not in other datasets: 8L, 9W, 10Q, 16H, 56P, 63S, and 90T (the sites and residues refer to the reference human sequence). Sites 9W and 90T are pocket sites and have mean  $\omega$  values of 0.80 and 0.12, respectively, in humans. Site 56P is a PBR nonpocket site, with the posterior mean  $\omega$  to be 0.10 in humans. These sites thus appear to be evolving adaptively in response to selective pressures unique to canines. Sites 8L, 10Q, 16H, and 63S are not known to have contact with the antigen (although their positions in the human crystal structure suggest this is possible) and show no evidence of positive selection in humans. Without a crystal structure of the canine protein it is not possible to speculate what adaptive roles these residues may have.

Several sites are found to be under diversifying selection in at least five of the six mammals: 11L, 13F, 57D, 74A, and 86G. These sites probably play very important roles in accommodating different side-chain structures in all mammalian species.

The similarity among species in the selective pressure on amino acid residues in the *DRB* alleles, as indicated by the shared sites detected to be under positive selection and by the high correlation of the estimated  $\omega$  ratios (Table 3), may be due to two reasons. First, the existence of trans-species polymorphism means that the phylogenetic trees for different species may overlap in their evolutionary histories. Second, the selective pressures on the *DRB* alleles may indeed be similar in different mammals. While both factors contribute to the pattern in the data, similar selective pressures among species appear to be the major reason, because the tree for sequences from all species is highly structured, so that most of the evolution occurred among alleles within species; for instance, sequences from the cow form a monophyletic clade, as do those from the dog.

#### Purifying Selection Within the PBR of Mammalian *DRB* Alleles

Of the 20 residues determined by Stern et al. (1994) to have contact with the antigen, two sites have very low  $\omega$  values and are inferred to be undergoing purifying selection across all mammalian species examined: 82N and 89F. Despite being within the PBR, these sites may be important to the basic structure of the PBR and are thus highly

conserved. For example, site 82N is known to form a hydrogen bond with the antigen, facilitating binding within the PBR. Thus purifying selection appears to have maintained the role of this site across the mammals. However, some differences among species are noticeable. Like site 82N, site 61W also forms a hydrogen bond with the antigen, and has a mean  $\omega < 0.1$  in humans as well as a low  $\omega$  value in canines. However, in other animals this site has posterior mean  $\omega$ 's close to 1, and in the tamarin it has  $\omega = 2.53$ . It is likely that this hydrogen bond is not maintained across all mammalian species, and that some *DRB* proteins may use other residues to form an equivalent bond.

#### The Role of Recombination in Human *DRB* Alleles

The codon-based analysis assumes a single tree topology for the whole alignment and does not account for recombination. Simulation studies (Anisimova et al. 2003; Shriner et al. 2003) suggest that detection of positive selection using site models may be affected if the sequences have undergone excessive recombination events. It has been suggested that recombination plays a role in generating polymorphism in some MHC genes, including *DRB* (Gaur and Nepom 1996). We tested for recombination using four different methods, using the human exon 2 data: PLATO (Grassly and Holmes 1997), RDP (Martin and Rybicki 2000), MaxChi2 (Smith 1992), and Geneconv (Sawyer 1989). PLATO and RDP use a phylogenetic tree to detect anomalous regions of the alignment, while MaxChi2 and Geneconv search for unusual substitution patterns. When the model does not allow for among-site rate heterogeneity, PLATO detected short anomalous stretches of nucleotide sites, which flank codons detected to be under positive selection in this paper (sites 37S, 47Y, 57D, 67L, 71R, and 86G). Under the gamma model of variable rates among sites, PLATO detected fewer regions with weaker support. Simulations suggest that with highly variable rates among sites, PLATO may identify fast-evolving sites under positive selection as anomalous regions even if the model accounts for rate variation and even if no recombination has taken place (Posada 2002). The small sizes of the anomalous regions suggest that recombination is an unlikely explanation for the sites detected by PLATO. Furthermore, none of the other methods (RDP2, MaxChi2, and Geneconv) detected recombination. Therefore we conclude that recombination has not made significant contributions to variations in the *DRB* alleles analyzed here. Sites identified to be under positive selection in this paper are in contact with the antigen and are most likely genuine positive-selection sites.

## Summary

In sum, our likelihood analysis revealed positive selective pressure in every mammalian species examined. The amino acid sites detected to be under positive selection are either within or close to the PBR. These results are in agreement with the previous analysis of a much smaller dataset by Hughes et al. (1994), but extend their conclusions by identifying exactly which PBR sites are under positive selection. Indeed, we detected positive-selection sites outside the PBR as well as negative-selection sites within the PBR, results that are impossible to obtain using the approach of Hughes and Nei (1988), which uses structural information to designate amino acid residues potentially under positive selection. Our results highlight the utility of random-sites models even when structural information is available. We discover striking similarities among the mammalian species in the amino acid residues detected to be under positive selection, suggesting that DRB performs more or less the same function of antigen binding and presentation in the different species, involving the same amino acid residues.

**Acknowledgments** We thank Nicolas Galtier and two anonymous referees for many critical comments. Andrew Rambaut, Joe Bielawski, and Gabriela Aguilera provided many helpful comments. This work is supported by a Biotechnology and Biological Sciences Research Council grant to Z.Y.

## References

- Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229–1236
- Arvidsson AK, Svensson AC, Widmark E, Andersson G, Rask L et al (1995) Characterization of three separated exons in the HLA class II DR region of the human major histocompatibility complex. *Hum Immunol* 42:254–264
- Bjorkman PJ, Saper SA, Samraoui B, Bennet WS, Strominger JL et al (1987a) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 329:512–518
- Bjorkman PJ, Saper SA, Samraoui B et al (1987b) Structure of the class I histocompatibility antigen, HLA-A2. *Nature* 329:506–512
- Bodmer W (1972) Evolutionary significance of the HLA system. *Nature* 237:139–145
- 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
- Brown JH, Jardetzky T, Saper MA et al (1988) A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. *Nature* 332:845–850
- Brown JH, Jardetzky TS, Gorga JC et al (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364:33–39
- Doherty PC, Zinkernagel RM (1975) A biological role for the major histocompatibility antigens. *Lancet* 1:1406–1409
- Felsenstein J (2005) *Phylib: phylogenetic inference program*, version 3.6. University of Washington, Seattle
- Gaur LK, Nepom GT (1996) Ancestral major histocompatibility complex DRB genes beget conserved patterns of localized polymorphisms. *Proc Natl Acad Sci USA* 93:5380–5383
- Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol* 14:239–247
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704
- Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174
- Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170
- Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci USA* 86:958–962
- Hughes AL, Hughes MK, Howell CY, Nei M (1994) Natural selection at the class II major histocompatibility complex loci of mammals. *Philos Trans R Soc Lond B Biol Sci* 346:359–366
- Kimura M (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120
- Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H (2002) Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunol Rev* 190:95–122
- Lundberg AS, McDevitt HO (1992) Evolution of major histocompatibility complex class II allelic diversity: direct descent in mice and humans. *Proc Natl Acad Sci USA* 89:6545–6549
- Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16:562–563
- Martin DP, Williamson C, Posada D (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21:260–262
- 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). *J Immunogenet* 56:420–426
- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936
- Ohta Y, Okamura K, McKinney EC, Bartl S, Hashimoto K et al (2000) Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. *Proc Natl Acad Sci USA* 97:4712–4717
- Posada D (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol Biol Evol* 19:708–717
- Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA* 98:13757–13762
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Sawyer SA (1989) Statistical tests for detecting gene conversion. *Mol Biol Evol* 6:526–538
- She JX, Boehme SA, Wang TW, Bonhomme F, Wakeland EK (1991) Amplification of major histocompatibility complex class II gene diversity by intraexonic recombination. *Proc Natl Acad Sci USA* 88:453–457
- Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res* 81:115–121
- Smith JM (1992) Analyzing the mosaic structure of genes. *J Mol Evol* 34:126–129

- Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG et al (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 368:215–221
- Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* 16:1315–1328
- Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124:967–978
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
- Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306–314
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591
- Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* 19:49–57
- Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118