Immunogenetics (2005) 57: 16–24
DOI 10.1007/s00251-004-0765-z

ORIGINAL PAPER

S. Consuegra · H. J. Megens · K. Leon ·
R. J. M. Stet · W. C. Jordan

Patterns of variability at the major histocompatibility class II alpha locus in Atlantic salmon contrast with those at the class I locus

Received: 7 October 2004 / Revised: 7 December 2004 / Published online: 22 February 2005

Abstract In order to investigate the mechanisms creating and maintaining variability at the major histocompatibility (MH) class II alpha (*DAA*) locus we examined patterns of polymorphism in two isolated Atlantic salmon populations which share a common post-glacial origin. As expected from their common origin, but contrary to the observation at the MH class I locus, these populations shared the majority of *DAA* alleles: out of 17 sequences observed, 11 were common to both populations. Recombination seems to play a more important role in the origin of new alleles at the class II alpha locus than at the class I locus. A greater than expected proportion of sites inferred to be positively selected (potentially peptide binding residues, PBRs) were found to be involved in recombination events, suggesting a mechanism for increasing MH variability through an interaction between recombination and natural selection. Thus it appears that although selection and recombination are important mechanisms for the evolution of both class II alpha and class I loci in the Atlantic salmon, the pattern of variability differs markedly between these classes of MH loci.

S. Consuegra · W. C. Jordan
Institute of Zoology, Zoological Society of London,
Regent's Park,
London NW1 4RY, UK

S. Consuegra (✉)
Fish Muscle Research Group,
Gatty Marine Laboratory, University of St Andrews,
St. Andrews, Fife KY16 8LB, UK
e-mail: sonia.consuegra@st-andrews.ac.uk
Tel.: +44-1334-463449
Fax: +44-1334-463443

H. J. Megens · K. Leon · R. J. M. Stet
Cell Biology and Immunology Group,
Department of Animal Sciences,
Wageningen University,
Marijkeweg 40,
6709 PG Wageningen, The Netherlands

R. J. M. Stet
Scottish Fish Immunology Research Centre,
University of Aberdeen,
Tillydrone Avenue,
Aberdeen, AB24 2TZ, Scotland, UK

Keywords MHC · Class II alpha · Recombination · Natural selection · Atlantic salmon

Introduction

The genes of the major histocompatibility complex (MHC) are among the best studied loci in vertebrates as they play a critical role in triggering the immune response (Klein 1986) and because they are among the most variable genes known (Parham and Ohta 1996). Numerous studies suggest that these high levels of variability are maintained by natural and sexual selection in a wide range of organisms (reviewed by Apanius et al. 1997; Jordan and Bruford 1998; Bernatchez and Landry 2003). MHC variation is generated and new alleles created by point mutation and by recombination and gene conversion events between existing alleles (Belich et al. 1992; Watkins et al. 1992; Parham and Ohta 1996; Martinsohn et al. 1999). Most of the amino acid variation is concentrated at peptide binding residues (PBRs): those amino acids responsible for binding peptides derived from pathogens. The high frequency of non-synonymous substitutions per site in PBR codons (Hughes and Nei 1988, 1989), long persistence times of alleles (Takahata 1990) and patterns of intra-population allelic variation indicate that MHC genes are under balancing selection (Apanius et al. 1997; Hedrick 1999), probably in relation to pathogen exposure (Edwards and Hedrick 1998; Hedrick 2002; Penn et al. 2002).

Most knowledge of the structure and function of MHC genes come from studies in mammals, particularly in humans (Hughes and Nei 1988, 1989; Hughes et al. 1994; Parham and Ohta 1996) and many studies in other vertebrates assume that the mechanisms that create and maintain MHC variability are similar across taxa. PBRs, for example, are commonly deduced by comparing MHC sequences from a species of interest with human counterparts (e.g. Grimholt et al. 1993; Kim et al. 1999; Hoelzel et al. 1999; Hambuch and Lacey 2002), where the crystal structure has been analysed (Brown et al. 1993). Often, patterns of selection on MHC genes are then studied with respect to PBRs/non-PBRs identified in this manner. However, methods for detecting positive selection that do not require a priori identification of candidate sites have recently been developed (Yang and Bielawski 2000).

Further insight into the molecular evolution of MHC genes can be obtained from the study of other vertebrate species, especially those with a relatively simple MHC structure, such as amphibians (Nonaka et al. 1997), chickens (Kaufman et al. 1999) or salmonids (Shum et al. 2001). Salmonids in general, and Atlantic salmon in particular, have proved to be a good model to analyse the mechanisms of MHC evolution because (1) as in the rest of the teleosts, class I and class II loci are not physically linked, allowing independent evolution of both classes of genes (Grimholt et al. 2002) [as the class I and class II genes do not form a complex they are known as MH genes in teleosts (Stet et al. 2003)], (2) single class I and class II MH loci are expressed (Shum et al. 2001; Stet et al. 2002; Grimholt et al. 2002), making analyses simpler, and (3) most of the present distribution of the species was colonized from refugia after the last glaciation (Hewitt 1999), making possible the study of adaptation of similar lineages colonising different environments and the analysis of molecular evolution on a relatively short temporal scale.

A comparison of the allelic composition at the MH class I locus between two geographically separated Atlantic salmon populations isolated since the last glaciation (15,000 years ago) showed that despite their common origin, the populations possessed almost non-overlapping sets of alleles although they share major ancient allelic lineages (Consuegra et al., submitted). As the substitution rate of single base mutations is relatively slow, even at MHC loci (Klein and O'Huigin 1994), the most plausible mechanism for the rapid divergence of alleles in these populations is recombination between alleles. Strong evidence for recombination was found, particularly in the regions close to and including sites under positive selection and therefore potential PBRs. It is not known, however, whether the class II loci (*Sasa-DAA* and *Sasa-DAB*), that are tightly linked to each other but not linked to class I (Grimholt et al. 2000; Stet et al. 2002), are evolving the same way in the short term. In particular, it is not known whether the role of recombination in generating new alleles is as important in class II as it appears to be in class I for salmonid fish (Shum et al. 2001; Consuegra et al., submitted).

Class I and class II MHC genes are known to evolve differently not only in primates (Boyson et al. 1996) but also in other vertebrates such as cyprinids (Kruiswijk 2002) and sparrows (Bonneaud et al. 2004), although in different ways. In primates, class II loci show trans-species sharing of allelic lineages that is not seen at class I loci (Boyson et al. 1996; Seddon and Ellegren 2002). In contrast, in cyprinids and salmonids class I alleles represent highly divergent and ancient allelic lineages while class II alleles are more recent (Shum et al. 2001; Stet et al. 2002; Kruiswijk 2002).

Here we analyse variation at the class II alpha locus (*Sasa-DAA*) in two recently diverged Atlantic salmon populations, using maximum likelihood methods for detecting molecular adaptation and recombination to examine if (1) both populations differ in allelic composition at the class II alpha locus (as for class I), (2) there is evidence of positive selection on any sites that may indicate that they are putative PBRs, and (3) recombination at the PBRs is involved in creating new alleles.

Materials and methods

Samples

Juvenile Atlantic salmon were sampled from four west coast Irish rivers with natural populations of Atlantic salmon: Owenmore, Owenduff, Burrishole and Carrowinskey. Samples of white muscle or adipose fins were stored in 95% ethanol while anterior kidney tissue was stored in RNAlater buffer (Qiagen) for subsequent extraction of RNA.

DNA isolation, cDNA synthesis, amplification and sequencing

Genomic DNA was isolated from muscle samples using the Geneclean DNA Purification kit (Qiagen), resuspended in 100 μl of elution buffer and stored at 4°C until use in PCR amplifications.

Total RNA was extracted from anterior kidney tissue of 17 individuals using the Purescript RNA Isolation kit from GENTRA (Gentra Systems, Minneapolis, Minn., USA) and 11 μl of purified RNA digested with DNAse I was used to synthesise first strand cDNA using the First-Strand cDNA Synthesis kit (Amersham Pharmacia Biotech UK). First strand cDNAs were used as templates for PCR amplification of the Atlantic salmon β-actin locus to check for possible genomic DNA contamination (the presence of an intron between the primers results in products of different size in genomic and cDNA) with the primers (Act_fwd 5′-ATGGAAGATGAAATCGCCGC-3′ and Act_rev 5′-TGC CAGATCTTCTCCATGTCG-3′). Samples that gave a band of the correct size (~200 bp), with no evidence of genomic DNA contamination (a product of ~450 bp), were then used for amplification of the MHC class II alpha locus.

A region of ~214 bp of the cDNA was amplified with a 50:50 mix of the following primers:DAAexon2_fwd: 5′-GGGTTTCTTTTCTCAGTTCTGC-3′, and DAAexon2_rev: 5′-CTTCTCTCTCTTACCTATTTTCTCTTCTG-3′. This region spans most of exon 2 (α_1 domain) of the class II *Sasa-DAA* locus (Grimholt et al. 2002). The final amplification volume was 25 μl, distributed as follows: 16.6 μl sterilized distilled water, 2.5 μl 10× amplification buffer, 3.5 μl 2 mM dNTPs, 1.5 μl 50 mM $MgCl_2$, 0.075 μl of each primer (100 pM), 0.5 μl DMSO and 1.25 U of *Taq* polymerase (Invitrogen). PCR conditions were 95°C for 5 min, then 5 cycles of 94°C for 30 s, 60°C for 30 s and 72°C for 1 min, then another 5 cycles of 94°C for 30 s, 55°C for 30 s and 72°C for 1 min and 25 cycles of 94°C for 30 s, 50°C for 30 s and 72°C for 1 min followed by an extension of 72°C for 10 min.

The PCR products were run in a 1% agarose gel, bands of the expected size were excised and DNA was purified using the Qiaquick Gel Purification Kit (Qiagen) and resuspended in 30 µl of elution buffer. The purified products were sequenced on both strands using the same forward and reverse exon 2 primers with the ABI Prism BigDye Terminator Cycle Sequencing Kit diluted with Better Buffer (Microzone) following the manufacturer's protocol and sequences were resolved on an ABI Prism 377 automated sequencer. The sequences were compared with previously described Atlantic salmon sequences (Stet et al. 2002; Megens et al., in preparation). In the case of previously undescribed sequences the purified products were cloned into the pCR2.1 plasmid vector (TA-cloning kit, Invitrogen) and transformed in INF'strain of *Escherichia coli*. Plasmid DNA from at least 5 colonies per individual was isolated using the Quiaprep Spin Miniprep Kit (Qiagen) and sequenced as described above.

Sequence analysis

Only sequences represented by at least two clones from independent PCRs were considered in subsequent analyses. Sequences from class II alleles were aligned with Sequencher (Genecodes) software and BioEdit v5.0.9 (using the Clustal W program included in the package). MEGA 2.1 (Kumar et al. 2001) was used to calculate the gamma distance from the amino acid sequences and to build a neigbour-joining phylogenetic tree with 1,000 bootstrap iterations to assess support for nodes in the phylogeny. *Onchorynchus mykiss* class II alpha (*Onmy-DAA*) sequences (GenBank accession nos. AJ251431–33) were used as an outgroup to root the tree (Grimholt et al. 2000). As in previous studies (Stet et al. 2002; Grimholt et al. 2002), alleles were defined on the basis of deduced amino acid sequences, not on nucleotide sequence.

Detecting positive selection

The ratio of non-synonymous/synonymous substitutions (ω=d_N/d_S) is the most common measure used for detecting positive selection acting on protein coding genes. A ratio of ω>1 is interpreted as evidence that non-synonymous mutations result in fitness advantages and are fixed at a higher rate than synonymous mutations (are positively selected). Neutral amino acid changes will result in ω=1 while amino acid sites under purifying selection will produce ratios ω<1. As a high number of the sites in the protein will be invariant due to structural constrains, classical analysis comparing rates of non-synonymous and synonymous substitutions in the complete coding sequence can be inadequate to detect positive selection. To overcome this problem, we used maximum-likelihood models of codon substitution to address the question of whether the rate of non-synonymous substitution (d_N) is greater than the rate of synonymous substitution (d_S) over the entire set of sequences, taking into account the phylogenetic structure of the sequences. To detect positive selection we used different codon-based models that allow for variable selection among sites as recommended by Yang et al. (2000) and implemented in the program CODEML of the PAML 3.14 package (Yang 1997). Five different models that allow for different intensity of selection among sites (and deduced from the data) were tested. We compared the scenario where non-synonymous mutations are either neutral or deleterious (models M1 and M7, respectively) with models that allow for positive selection including an additional category for advantageous substitutions (models M2, M3 and M8). Three of the models assume a discrete distribution of the ω statistic (*dN*/*dS*) among sites: M1 (neutral) assumes two categories of sites conserved (ω=0) and neutral (ω=1); M2 (selection) includes an additional category of sites with ω estimated from the data; M3 (discrete) assumes a discrete distribution of K different ω ratios. Two additional models assume a continuous distribution for heterogeneous ω ratios among sites: M7 (beta) that assumes a beta distribution and does not allow for positively selected sites and M8 (beta and ω) that accounts for positively selected sites (ω>1). Nested models can be compared in pairs using the likelihood ratio test (LRT): twice the log-likelihood difference is compared with a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between both models. The null model has a fixed ω=1 while the alternative models have an estimate of ω as a free parameter. In this way, the more general models M2 and M3 can be tested against M1 and M8 against M7. Maximum likelihood trees to provide the phylogenetic information were constructed using DNAML from PHYLIP (Felsenstein 1989). A Bayesian approach implemented in CODEML was used to identify residues under positive selection in the α1 domain and sites with a posterior probability >95% were considered as positively selected under the model that best fitted the data.

We performed an analysis of sequence variability using a variability metric (V) (Reche and Reinherz 2003) that is formally similar to the Shannon entropy index (Shannon 1949) and that allows identification of variable amino acid residues. For a multiple protein sequence alignment the modified Shannon entropy (V) for every site follows the equation:

$$V = -\sum_{i=1}^{M} P_i \log_2 P_i$$

Where P_i is the fraction of residues of amino acid type i, and M is the number of amino acid types (20). Values of V range from 0 (only one residue in present at that position) to 4.322 (all 20 residues equally represented in that position). Amino acid sites with V>1.0 are considerered variable, whereas those with V<1 are considered conserved. Variable amino acid residues estimated in this way may be functionally relevant in immune recognition through involvement in peptide contact (i.e. potential PBRs) (Stewart et al. 1997).

Recombination analysis

The likelihood-based method of Grassly and Holmes (1997) for detection of phylogenetically anomalous regions or "spatial phylogenetic variation" (SPV) was implemented using the program Plato (Partial Likelihoods Assessed Through Optimisation) (Grassly and Holmes 1997). SPV may arise either as a result of selection or of conversion/recombination. In this method a likelihood for each site is calculated on the basis of the overall (or global) maximum-likelihood phylogeny, and a 'sliding window' technique is used to identify the region with the lowest likelihood score value for each window size. The window size is varied from a minimum of 5 base pairs up to half the sequence length. The standardised normal deviate (*Z*) is used to test the statistical significance of each of the lowest likelihood regions against a null normal distribution generated by simulation (100 replicates). A value of $Z \geq 3$ is taken to be equivalent to $P \leq 0.05$ adjusted for multiple tests (Grassly and Holmes 1997). The maximum likelihood phylogeny used to test in PLATO was constructed with DNAml from the PHYLIP 3.6 package (Felsenstein 1989) based on the most appropriate model of nucleotide substitution determined by MODEL TEST v1.06 (Posada and Crandall 1998).

DnaSP software (Rozas and Rozas 1999) was used to estimate the minimum number of recombination events (Rm) in the history of the sample using the four-gamete test (Hudson and Kaplan 1985). The regions potentially involved in recombination identified by DnaSP were compared with regions of SPV detected with PLATO.

The program Ldhat (McVean et al. 2002) was used to assess the importance of the recombination relative to point mutation in the patterns of genetic variation of all the exon 2 sequences. This program uses a composite-likelihood method based on the coalescent theory to estimate population recombination rates (4Ner), under an infinite-sites model of sequence evolution, allowing for recurrent mutation. The method also implements a likelihood permutation test for testing the null hypothesis of no recombination (4Ner=0) based on the loss of the interchangeable character of the sites when recombination occurs (McVean et al. 2002). Recombination analyses were also conducted after excluding sites affected by positive selection in order to avoid the confounding effect of selection in the estimates of recombination rates.

Results

Sequence variability

A fragment of 214 bp of the *DAA* gene corresponding to the α1 domain of the protein was sequenced and aligned with class II *DAA* alleles described in previous studies (Stet et al. 2002; Megens et al., in preparation). No more than two sequences were amplified from each individual as expected from a single expressed locus. Seventeen alleles were identified (Fig. 1), 64% of them common to both populations (11 out of 17). Three of the alleles were unique to the Irish population (*Sasa-DAA*0304*, *Sasa-DAA*0305* and *Sasa-DAA*1202*) and three were unique to the Norwegian sample (*Sasa-DAA*0801*, *Sasa-DAA*0901* and *Sasa-DAA* 1101*). A total number of 14 different alleles were obtained from Irish Atlantic salmon while 14 different alleles were previously described for Norwegian Atlantic salmon. Three unique substitutions were present in the sequences: one in the Norwegian *Sasa-DAA*1101* allele (S instead of W at position 19) and two in the Irish *Sasa-DAA*0305* allele (R instead of Q/G at position 47, and I instead of K/T/V at position 61; Fig. 1).

Most sequences consisted of highly conserved motifs with a low number of variable residues concentrated in particular sites (Fig. 1). The estimated average difference in nucleotides among sequences was k=10.132 for the Norwegian and k=9.69 for the Irish sequences. In total, 22 nucleotide sites out of 214 (10.3%) were variable and 15 out of 71 amino

```
              ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| .
                      10         20         30         40         50         60         70
Sasa-DAA*0101 GCSDSDGLDM YGLDGEEMWY ADFNKQEGVV ALPPFADPFT FPGFYEQAVG NQGVCKGNLA KCIKAYKNPE E
Sasa-DAA*0201 .......... .......... .....G.... .......... ......G... ......A... VN........ .
Sasa-DAA*0301 .......V.. .......... .....G...M P......... Y..A..G... ...I..A... T......... .
Sasa-DAA*0302 .......... .......... .....G...M P......... ...A..G... ...I..A... T......... .
Sasa-DAA*0303 ........N. .......... .....G...M P......... Y..A..G... ...I..A... T......... .
Sasa-DAA*0304 .......V.. .......... .....G...M P......... Y..A...... ......A... T......... .
Sasa-DAA*0305 ........N. .......... .....G...M P......... Y..A..R... ......A... I......... .
Sasa-DAA*0401 .......V.. .......... .....G.... .......... ......G... ......A... VN........ .
Sasa-DAA*0501 .......V.. .......... .....G.... .......... ...H..G... ......A... TS........ .
Sasa-DAA*0601 .......V.. .......... .....G...M P......... Y..A...... ......A... VN........ .
Sasa-DAA*0602 .......VE. .......... .....G...M P......... Y..A...... ...I..A... VN........ .
Sasa-DAA*0603 .......VE. .......... .....G...M P......... ...A...... ......A... VN........ .
Sasa-DAA*0701 .......... .......... .......... .......... YH.A...... ...I...... .......... .
Sasa-DAA*0901 .......V.. .......... .....G.... .......... .H.A..G... ......A... VN........ .
Sasa-DAA*1001 ........N. .......... .....G.... .......... .......... ......A... TS........ .
Sasa-DAA*1101 .......V.. ........S. .....G.... .......... .......... ......A... TS........ .
Sasa-DAA*1201 .......... .......... .......... .......... ...H..G... .......... .......... .
```

Fig. 1 Alignment of MH class II *Sasa-DAA* exon 2 amino acid sequences for Atlantic salmon from Irish and Norwegian populations. *Dots* indicate identity. Nucleotide sequences were deposited in GenBank under accession numbers AY780908-AY780917

acid residues were polymorphic (21.1%). Only one nucleotide change was synonymous and every sequence therefore corresponded to a different allele.

Sequences from both populations were pooled in the rest of the analyses in order to determine the number of shared sequences and assess similarity between unique and common sequences.

Phylogenetic analysis

We performed phylogenetic analyses by building a neighbour-joining tree based on gamma distance (Fig. 2). Branch lengths were generally short, but three possible allelic lineages may be defined although they were supported by low bootstrap values. The newly identified Irish sequences cluster in one of the lineages (I) consisting of two sub-lineages (*Sasa-DAA*0301*, *Sasa-DAA*0302*, *Sasa-DAA*0303*, *Sasa-DAA*0304*, *Sasa-DAA*0305* and *Sasa-DAA*0601*, *Sasa-DAA*1201*, *Sasa-DAA*1202*) differing from the Norwegian sequences only by a few (2–5) base changes. The sequences absent in the Irish population (*Sasa-DAA*0801*, *Sasa-DAA*0901*, *Sasa-DAA*1101*), on the other hand, are distributed across the other three clusters.

Patterns of positive selection

Maximum likelihood models that allow positive selection fitted the data significantly better than those that assume only neutral or conserved mutations (Table 1). LRT tests suggested that model M2, which allows for positive selection, fitted the data better than model M1 (which considers only conserved and neutral sites) (P<0.001). Estimates using the M2 model suggest that 20% of sites in the α1 domain were under strong positive selection (ω=32.7) and the rest of the sites were under purifying selection (ω=0). All variable sites were identified as positively selected by this model. Model M3, which assumes three site classes, fitted the data significantly better than M1 (P<0.001) and M2 (P<0.01). The results of model M3 suggested that 4.2% of the sites in the α1 domain were under strong positive selection (ω=46.2).

The LRT test comparing the two models that assume a beta distribution of ω over sites (M7 and M8) indicated that M8 (which allows for selection) fitted the data better than M7 (which does not allow selection) (P<0.001). Estimates from M8 indicate that 16% of the sites are under strong positive selection in the sequences (ω=44.8).

Results from the different models indicate that there is variable selective pressure across sites of the MHC class II alpha sequences and the presence of a number of posi-

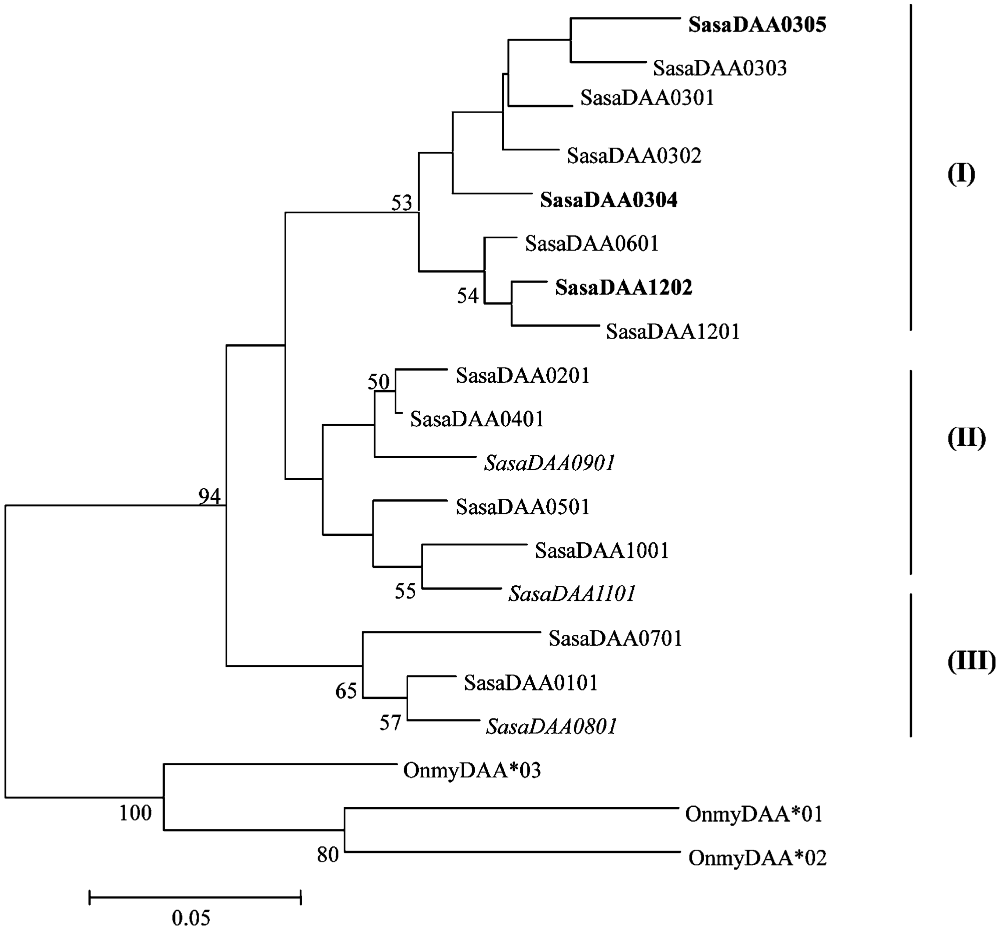

Fig. 2 Phylogenetic tree of α1 domain amino acid sequences of the MHC class II gene of Atlantic salmon for Irish and Norwegian populations based on gamma distance with a=1.53. The reliability of the cluster analysis was tested by 1,000 bootstrap iterations and the results are shown in the nodes. Sequences in *bold* are unique to Irish populations and sequences in italics are unique to Norwegian populations

Table 1 Log-likelihood values (L) and parameter estimates (ω=non-synonymous/synonymous rates ratio estimates; pi=fraction of the codons having as discrete ω or following a ω distribution, p & q=statistical parameters of the beta distribution) under random-sites models for the Irish and Norwegian class II MH alleles. Positively selected sites (* represents 95%, ** represents 99%) were identified by a Bayesian method implemented in CODEML. Amino acid positions refer to those in *Sasa-DAA*0101*

Model	L	Estimates of Parameters	Positively selected sites
M1	−566.74	p_0=0.751 (ω_0=0) p_1=0.249 (ω_1=1)	Not allowed
M2	−530.01	p_0=0.793 p_1=0.000 p_2=0.207 (ω_2=32.7)	8L**, 9D**, 19W**, 26Q**, 30V**, 31A**, 41F**, 42P**, 44F**, 47Q**, 54V**, 57G**, 61K**, 62C**
M3	−525.05	p_0=0.006 (ω_0=0.341) p_1=0.831 (ω_1=0.342) p_2=0.042 (ω_2=46.2)	8L**, 9D**, 26Q*, 30V**, 41F**, 42P**, 44F**, 47Q**, 54V**, 61K**, 62C**
M7	−567.08	p=0.007 ω=0.202 q=0.025	Not allowed
M8	−525.98	p_0=0.836 p_1=0.163 (ω=44.8) p=49.697 q=99.000	8L**, 9D**, 26Q*, 41F**, 42P**, 44F**, 47Q**, 54V**, 61K**, 62C**

tively selected sites. The patterns of distribution of positively selected sites were consistent among models (M2, M3 and M8), although a larger number of sites was inferred under model M2 than the alternative models. The results from the M8 model, which were the most conservative, are presented in Fig. 3.

Following the criteria of Reche and Reinherz (2003) for HLA, four of the sites identified as positively selected were also identified as highly polymorphic (V>1) by the entropy analysis, and on this basis could be considered as potential PBRs. All other positively selected sites were classified as polymorphic (V>0.5) (Fig. 3).

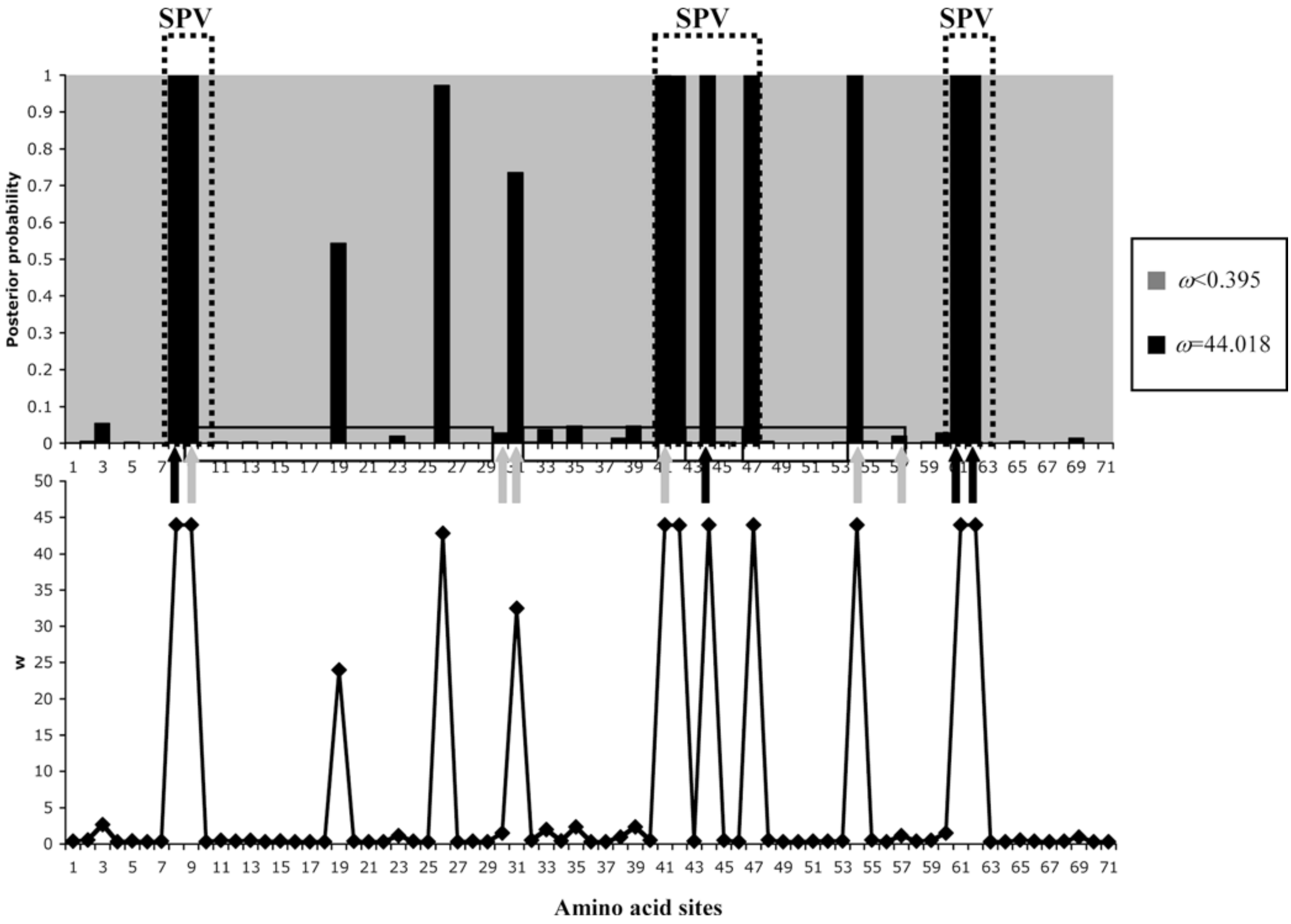

Fig. 3 Posterior probabilities of site classes for sites along the MH class II alpha 1 domain under the random-sites model M8 (beta & ω). Ten equal-probability categories were used to approximate the beta distribution (Yang et al. 2000), so that the model has 11 categories. Posterior means of ω, calculated as the average of ω over the 11 site classes, weighted by the posterior probabilities. (▭) Sites involved in recombination, (▪ ▪ ▪ ▪ ▪) areas of SPV, polymorphic sites determined by their entropy values ➡(V>0.5), ➡(V>1.0)

Analysis of recombination

A test of alternative models of sequence evolution (Modeltest v3.06, Posada and Crandall 1998) indicated that the sequences followed a F81 model of nucleotide substitution with a gamma distribution shape parameter a=1.536 and those parameters where used to run the analysis of spatial phylogenetic variation in PLATO that was carried out for the whole set of sequences described in both Norwegian and Irish. The result of the test indicated that three regions of SPV (23–28, 123–141 and 182–187; Fig. 3).

The results from the maximum likelihood recombination test suggested significant evidence of recombination (P<0.001) in the class II alpha sequences. Estimated recombination rate in the population (ρ=4 Ner=96) was higher than the mutation rate (Watterson estimate for population rate of mutation, θ=6.5). Moreover, the results of the recombination analysis from DnaSP carried out including all sequences pooled revealed that at least six recombination events have occurred and involved regions between 3 and 27 nucleotides long (Fig. 3).

Most of the sites identified as positively selected fall into a region of spatial variable region coinciding in many cases with sites involved in recombination (Fig. 3). However, significant evidence of recombination was observed after repeating the maximum likelihood recombination test without sites identified as positively selected (P<0.001). After removing sites identified as positively selected, the estimated recombination rate in the population (ρ=23) remained higher than the mutation rate (θ=2.7). Only two of the six recombination events initially detected by DNAsp remained after excluding selected sites (regions from positions 23–89 and 94–123).

Discussion

Population variability

Irish and Norwegian Atlantic salmon populations had substantially overlapping sets of alleles, differing in only a few class II alpha alleles (14 alleles described in each population, with only 3 of them private to each). Only a single locus was expressed in all individuals, as reported previously in Norwegian farmed Atlantic salmon (Stet et al. 2002). The 19 alleles described here, including three new sequences found only in Irish populations to date, were polymorphic and differed from one another by ~9 substitutions on average, similar to values found previously in class II *DAA* genes in farmed Atlantic salmon (Stet et al. 2002). The fragment of the exon 2 sequenced corresponds to the region where PBRs are potentially located and hence is likely to constitute the most variable region of the gene (Stern et al. 1994).

The phylogeny of the exon 2 sequences revealed three possible allelic lineages of the *Sasa-DAA* locus in Atlantic salmon, although these lineages were supported by low bootstrap values. Branch lengths in each lineage were short, with low genetic distances suggesting a recent origin of the alleles at this locus (Shum et al. 2001; Stet et al. 2002). The class II alleles described in this study in both Norwegian and Irish populations have low divergence and all three new Irish alleles are located in one of the pre-existing clusters of shared alleles.

MH class I versus class II patterns of variability

Contrary to the observation in class I, where Irish and Norwegian populations had almost non-overlapping sets of alleles (Consuegra et al., submitted), they share the majority of class II *DAA* alleles (11 of 17 alleles). Although class I and class II molecules have similar tertiary structures, they differ in the class of peptides that they bind and the proteolytic routes for processing them (Kaufman et al. 1994; Castellino et al. 1997; Gromme and Neefjes 2002). They also differ in the class of T cells that they react with (Housset and Malissen 2003). Different modes of binding the peptide by class I and class II molecules, relatively strict in class I and more permissive in class II, may be responsible for functional differences between the two classes of MHC molecule, e.g differences in positive and negative selection of T cells (Huseby et al. 2003). The different nature of peptide binding may be one reason for the different evolutionary rates between class I and class II genes, as has been proposed to explain the higher turnover of the class I genes with respect to the class II genes in primates (Go et al. 2003). Under this model, stricter peptide binding by class I molecules would promote a higher rate of change in order to adapt to new infections (Go et al. 2003). However, the nature of class I peptide binding specificities has not been determined in fish (Grimholt et al. 2002) and our present results for Atlantic salmon contrast with the situation in the Lake Tana barbs species flock where closely related species have been shown to share class I alleles but differ in their class II alleles. The class II alleles are completely partitioned among the 10 different species studied without any sharing between species (Kruiswijk 2002).

Patterns of positive selection and recombination

Our results suggested that the observed diversity in the class II alpha sequences may be to a large extent generated by positive selection on PBRs. Maximum likelihood models allowing selection fitted the data significantly better than models that considered only neutral or conserved sites. The selected sites identified by Bayesian analysis coincided largely with sites that aligned with PBRs in human sequences (Stet et al. 2002) and which represent most of the variability among the sequences. Ten sites were identified with a strong signal of positive selection by model M8 (Table 1; Fig. 3) from which five correspond to potential PBRs (41F, 44F, 54V, 61K, 62C) and one to a conserved residue (42P) according to the comparison of Stet et al. (2002) with *HLA-DRA* sequences. Models M2 and M3 included all the sites identified by M8; M3 added one more site and model M2 identified three additional sites, one of them potentially a PBR site (57G) according to Stet et al. 2002. The fact that sites identified as PBRs by crystallography in humans have been independently identified as positively selected in Atlantic salmon strongly suggests that they are implicated in peptide binding in this species.

The observed sequence diversity is probably not only the result of point mutations but is also generated by recombination events. The higher estimated recombination rate in relation to the mutation rate in the population suggests that recombination has been an important force in creating the present allelic diversity. Even after removing positively selected sites, results from the maximum likelihood tests showed evidence for recombination in the sequences.

Three of the four regions of SPV identified by PLATO overlapped with sites involved in recombination (as identified by DnaSP). Moreover, the regions identified as areas of SPV included in all cases sites identified as positively selected. Although the method implemented by PLATO cannot discriminate between recombination and selection as the cause for SPV (Grassly and Holmes 1997), the fact that after removing all positively selected sites evidence for recombination remained, indicate that at least part of the SPV regions detected correspond to recombination events.

Positively selected sites coincided in half the cases with the regions involved in recombination deduced by DnaSP. The fact that there is a non-random association between sites showing recombination and selection (G-test=11.02, P=0.001) suggests that alleles with recombinant PBRs have been positively selected, thus increasing variability, as suggested by Otha (1996). More positively selected sites (10) were detected than recombination events (6), although the latter cover most of the sequence (66%).

However, the coincidence of sites identified as both involved in recombination and under diversifying selection could also be due to the effect of recombination on the maximum-likelihood methods. Recombination can introduce stochasticity in the phylogenies (Schierup and Hein 2000) making it difficult to infer the number of ancestral allelic lineages. Current phylogeny-based models of codon substitution that include heterogeneous selective pressures across sites do not take the effects of recombination into account, although Bayesian methods for inferring positively selected sites seem to be little affected by recombination (Anisimova et al. 2003). Although the number of sequences analysed is not particularly large (n=17), and they are not very divergent (tree length S=1.0), the strength of selection estimated for positively selected sites, their coincidence with human PBRs and their independent identification as polymorphic sites (V>1) suggest that our results are robust to the effects of recombination. In theory, it is possible that some of the sites identified by the M2 model and not by the M8 model are the result of the confounding effects of the recombination in the phylogeny. However, we have taken that into account by using the M8 model results (apparently less affected by recombination; Anisimova et al. 2003) to compare the results of recombination and selection tests.

In common with the situation for class I sequences, recombination involving the putative PBR sites seems to play an important role in the origin of variability in class II sequences. The existence of a large intron between α1 and α2 domains of the salmonid MH class I gene, where potential PBRs are located, may provide an additional mechanism for increasing variability through allele shuffling (Shum et al. 2001; Grimholt et al. 2002) compared to class II, where potential PBRs are only encoded in exon 2. In fact, the number of recombination events at the class II alpha locus was lower than that observed for the same populations for the class I locus, even when considering both class I domains independently. This may be the result of a more recent origin of class II alleles compared to those at the class I locus; class I alleles may simply have maintained the signature of old recombination events and accumulated more variation (Takahata and Satta 1998). Also, point mutation rates at the class II locus were lower than recombination rates, contrary to that the situation found at the class I locus.

In summary, recombination seems to play an important role in the origin of new alleles at the class II alpha locus of Atlantic salmon, both in Irish and Norwegian natural populations. A relatively large proportion of positively selected sites (potential PBRs) were found to be involved in recombination events. Although similar processes (selection and recombination) appear to be shaping variability at the class I and class II alpha loci in Atlantic salmon, the resulting distribution of variability across populations is very different for the two classes of MH loci.

Acknowledgements We are grateful to Helmut Schaschl for technical guidance and Elvira De Eyto and Phil McGinnity for providing the samples on which this study is based. Funding was provided by a contract from the European Commission (SALIM PACT: QLRT-2000-01185).

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

Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. Crit Rev Immunol 17:179–224

Belich MP, Madrigal JA, Hildebrand WH, Zemmour J, Williams RC, Luz R, Petzl-Erler ML, Parham P (1992) Unusual HLA-B alleles in two tribes of Brazilian Indians. Nature 357:326–329

Bernatchez L, Landry C (2003) MHC studies in non-model vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16:363–377

Bonneaud C, Sorci G, Morin V, Westerdahl H, Zoorob R, Wittzell H (2004) Diversity of MHC class I and IIB genes in house sparrows (*Passer domesticus*). Immunogenetics 55:855–865

Boyson JE, Shufflebotham C, Cadavid LF, Urvater JA, Knapp LA, Hughes AL, Watkins DI (1996) The MHC class I genes of the rhesus monkey. Different evolutionary histories of MHC class I and class II genes in primates. J Immunol 156:4656–4665

Brown JH, Jardetzky T, Saper MA, Samaraoui B, Bjorkman PJ, Wiley DC (1993) Three dimensional structure of the human class II histocompatibility molecules. Nature 364:33–39

Castellino F, Zhong G, Germain RN (1997) Antigen presentation by MHC class II molecules: invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture. Hum Immunol 54:159–169

Edwards SV, Hedrick PW (1998) Evolution and ecology of MHC molecules: from genomics to sexual selection. Trends Ecol Evol 13:305–311

Felsenstein J (1989) PHYLIP-Phylogeny Inference Package (Version 32). Cladistics 5:164–166

Go YY, Satta Y, Kawamoto G, Rakotoarisoa A, Randrianjafy N, Koyama H, Hirai H (2003) Frequent segmental sequence exchanges and rapid gene duplication characterize the MHC class I genes in lemurs. Immunogenetics 55:450–461

Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using sequence data. Mol Biol Evol 14:239–247

Grimholt U, Hordvik I, Fosse VM, Olsaker I, Endresen C, Lie O (1993) Molecular cloning of major histocompatibilty complex class I cDNA from Atlantic salmon (*Salmo salar*). Immunogenetics 37:469–473

Grimholt U, Getahun A, Hermsen T, Stet RJ (2000) The major histocompatibility class II alpha chain in salmonid fishes. Dev Comp Immunol 24:751–763

Grimholt U, Drablös F, Jörgensen SM, Höyheim B, Stet RJM (2002) The major histocompatibility class I locus in Atlantic salmon (*Salmo salar* L.): polymorphism, linkage analysis and protein modelling. Immunogenetics 54:570–581

Gromme M, Neefjes J (2002) Antigen degradation or presentation by MHC class I molecules via classical and non-classical pathways. Mol Immunol 39:181–202

Hambuch TM, Lacey EA (2002) Enhanced selection for mhc diversity in social tuco-tucos. Evolution 56:841–845

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

Hedrick PW (2002) Pathogen resistance and genetic variation at MHC loci. Evolution 56:1902–1908

Hewitt GM (1999) Post-glacial re-colonization of European biota. Biol J Linn Soc 68:87–112

Hoelzel AR, Stephens JC, O'Brien SJ (1999) Molecular genetic diversity and evolution at the MHC *DQB* locus in four species of pinnipeds. Mol Biol Evol 16:611–618

Housset D, Malissen B (2003) What do TCR-pMHC crystal structures teach us about MHC restriction and alloreactivity? Trends Immunol 24:429–437

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

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 86l:958–962

Hughes AL, Hughes MK, Howell CY, Nei M (1994) Natural selection at the class II major histocompatibility complex loci of mammals. Phil Trans R Soc Lond B 345:359–367

Huseby ES, Crawford F, White J, Kappler J, Marrack P (2003) Negative selection imparts peptide specificity to the mature T cell repertoire. Proc Natl Acad Sci USA 20:11565–11570

Jordan WC, Bruford MW (1998) New perspectives on mate choice and the MHC. Heredity 81:127–133

Kaufman J, Salomonsen J, Flajnik M (1994) Evolutionary conservation of MHC class I and class II molecules —different yet the same. Semin Immunol 6:411–424

Kaufman J, Milne S, Gobel TWF, Walker BA, Jacob JP, Auffray C, Zoorob R, Beck S (1999) The chicken *B* locus is a minimal essential major histocompatibility complex. Nature 401:923–925

Kim TJ, Parker KM, Hedrick PW (1999) Major histocompatibility complex differentiation in Sacramento River Chinook salmon. Genetics 151:1115–1122

Klein J (1986) Natural history of the major histocompatibility complex. Wiley, New York

Klein J, O'Huigin C (1994) MHC polymorphism and parasites. Phil Trans R Soc Lond B 346:351–358

Kruiswijk CP (2002) Evolution of major histocompatibilitygenes in cyprinid fish. Molecular analyses and phylogenies. Dissertation, Wageningen University. ISBN 90-5808-735-2

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244

Martinsohn JT, Sousa AB, Guethlein LA, Howard JC (1999) The gene conversion hypothesis of MHC evolution: a review. Immunogenetics 50:168–200

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241

Nonaka M, Namikawa C, Kato Y, Sasaki M, Salter-Cid L, Flajnik MF (1997) Major histocompatibility complex gene mapping in the amphibian *Xenopus* implies a primordial organisation. Proc Natl Acad Sci USA 94:5789–5791

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science 272:67–74

Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc Natl Acad Sci USA 99:11260–11264

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818

Reche PA, Reinherz EL (2003) Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. J Mol Biol 331:623–641

Rozas J, Rozas R (1999) DnaSP v3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–891

Seddon JM, Ellegren H (2002) MHC class II genes in European wolves: a comparison with dogs. Immunogenetics 54:490–500

Shannon CE (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Shum BP, Guethlein L, Flodin LR, Adkison MD, Hedrick RP, Nehring RB, Stet RJM, Secombes C, Parham P (2001) Modes of Salmonid MHC class I and II evolution differ from the primate paradigm. J Immunol 166:3297–3308

Stern L, Brown J, Jardetzky T, Gorga J, Urban R, Strominger J (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 368:215–221

Stet RJM, de Vries B, Mudde K, Hermsen T, van Heerwaarden J, Shum BP, Grimholt U (2002) Unique haplotypes of co-segregating major histocompatibility. Immunogenetics 54:320–331

Stet RJM, Kruiswijk CP, Dixon B (2003) Major histocompatibility lineages and immune gene function in fish: the road not taken. Crit Rev Immunol 23:441–471

Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomick M, Weigert S, Litwin S (1997) A Shannon entropy analysis of immunoglobin and T cell receptor. Mol Immunol 34:1067–1082

Takahata N (1990) A simple genealogical structure of strongly balanced allelic lines and transspecies evolution of diversity. Proc Natl Acad Sci USA 87:2419–2423

Takahata N, Satta Y (1998) Selection, convergence, and intragenic recombination in HLA diversity. Genetica 102:320–331

Watkins DI, McAdam SN, Liu X, Strang CR, Milford EL, Levine CG, Garbert TL, Dogon AL, Lord CI, Ghim SH, Troup GM, Hughes AL, Letvin NL (1992) New recombinant *HLA-B* alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. Nature 357:329–333

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503

Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449