Cryptic MHC Polymorphism Revealed but Not Explained by Selection on the Class IIB Peptide-Binding Region

V. Llaurens,†,[1,2] M. McMullan,†,[1,3] and C. van Oosterhout*,[1,3]

[1]Evolutionary Biology Group, Department of Biological Sciences, University of Hull, Hull, United Kingdom

[2]Laboratoire Origine, Structure et Evolution de la Biodiversité (UMR CNRS 7205), Museum National d'Histoire Naturelle, Paris, France

[3]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom

†These authors contributed equally to this work.

***Corresponding author:** E-mail: c.van-oosterhout@uea.ac.uk.

Associate editor: Hideki Innan

Abstract

The immune genes of the major histocompatibility complex (MHC) are characterized by extraordinarily high levels of nucleotide and haplotype diversity. This variation is maintained by pathogen-mediated balancing selection that is operating on the peptide-binding region (PBR). Several recent studies have found, however, that some populations possess large clusters of alleles that are translated into virtually identical proteins. Here, we address the question of how this nucleotide polymorphism is maintained with little or no functional variation for selection to operate on. We investigate circa 750–850 bp of MHC class II *DAB* genes in four wild populations of the guppy *Poecilia reticulata*. By sequencing an extended region, we uncovered 40.9% more sequences (alleles), which would have been missed if we had amplified the exon 2 alone. We found evidence of several gene conversion events that may have homogenized sequence variation. This reduces the visible copy number variation (CNV) and can result in a systematic underestimation of the CNV in studies of the MHC and perhaps other multigene families. We then focus on a single cluster, which comprises 27 (of a total of 66) sequences. These sequences are virtually identical and show no signal of selection. We use microsatellites to reconstruct the populations' demography and employ simulations to examine whether so many similar nucleotide sequences can be maintained in the populations. Simulations show that this variation does not behave neutrally. We propose that selection operates outside the PBR, for example, on linked immune genes or on the "sheltered load" that is thought to be associated to the MHC. Future studies on the MHC would benefit from extending the amplicon size to include polymorphisms outside the exon with the PBR. This may capture otherwise cryptic haplotype variation and CNV, and it may help detect other regions in the MHC that are under selection.

Key words: ABC evolution, parasite selection, multigene family, copy number variation, gene conversion.

Introduction

Balancing selection is an evolutionary force that can maintain many haplotypes over long periods of evolutionary time. Loci under balancing selection are characterized by an excess in nucleotide polymorphism distributed across sequences, typically with intermediate frequencies (Maruyama and Nei 1981). The best-studied loci that experience balancing selection are the self-incompatibility locus in plants (Castric and Vekemans 2004), the mating-type loci in fungi (Uyenoyama 2005), and the immune genes of vertebrates (Piertney and Oliver 2006). Recent whole-genome studies show that balancing selection is a key force in the evolution of many genes in humans and other vertebrates (e.g., Andres et al. 2009).

The major histocompatibility complex (MHC) loci in vertebrates are a gene family involved in the immune response to pathogens. In particular, MHC class II loci encode cell surface proteins that bind nonself peptides from pathogens. These peptides are presented to the T cells, and this interaction elicits the adaptive immune response. The MHC loci exhibit a high diversity, sometimes even in small or bottlenecked populations (Aguilar et al. 2004; van Oosterhout, Joyce, Cummings, Blais, et al. 2006; Mona et al. 2008). Larger vertebrate populations, such as humans, can have hundreds of distinct sequences at a single MHC locus (Garrigan and Hedrick 2003). Natural selection by parasites and pathogens has been shown to maintain the genetic variation at the MHC, and the host immune genes are coevolving in a Red Queen arms race with pathogens (Piertney and Oliver 2006). The principal target of pathogen-mediated selection in the MHC is the so-called peptide-binding region (PBR). This part of the MHC molecule recognizes and binds to antigens. Many studies have shown that the ratio of nonsynonymous to synonymous mutations (d*N*/d*S*) is higher in the PBR as compared with the non-PBR and interpreted it as the signal of pathogen-mediated selection (Hughes and Nei 1988, 1989). Such selection can maintain a balanced polymorphism, and hence, it is called balancing selection.

Other selective pressures have been hypothesized to affect the evolution of multigene families that can have implications for the maintenance of genetic polymorphisms.

For example, a theoretical study by van Oosterhout (2009) showed that a locus under balancing selection can accumulate recessive deleterious mutations, particularly when the gene diversity (heterozygosity) is high and the recombination rates are low. Under these conditions, mutations are not often expressed in homozygote condition, which means that they can accumulate in the MHC region in a Muller's Ratchet-like process. This "sheltered genetic load" (Stone 2004) could shape the diversification rates in loci under balancing selection because new arising sequences would share their genetic load with their ancestor. The genetic load can dramatically alter the evolutionary dynamics of multigene families, which has been theoretically demonstrated also by Uyenoyama (2003) for the *S*-locus in plants. There is growing empirical support for the association of a high mutational load in the MHC multigene family, which could explain the >100 heritable diseases that are attributed to SNPs in the human MHC class I (HLA-A and HLA-B/-C) (Shiina et al. 2006).

The MHC polymorphism is also affected by gene conversion (Spurgin et al. 2011), concerted evolution, and other processes typical to multigene families, such as birth and death evolution (Nei and Rooney 2005). These processes can explain the considerable amount of copy number variation (CNV) that exists among species (Mehta et al. 2009), within species (Bonhomme et al. 2008), and within populations (Eimes et al. 2011). Less well recognized is the bias that can be introduced by gene conversion in estimates of haplotype variation and CNV. This may lead to a systematic underestimation of MHC haplotype diversity because sequence variation between paralogous gene copies can become homogenized. This may be particularly relevant when relatively small amplicons (200–300 bp) are being sequenced, as is the case in many MHC class II studies that focus on the exon containing the PBR only. The MHC multigene family can also include pseudogenes that may not easily be distinguished from functional MHC genes. Theory suggests that such silenced genes may nevertheless contribute to the variability of expressed paralogs through gene conversion (Takuno et al. 2008). Furthermore, not all duplicated loci within a gene family need to be functionally equivalent. For instance, divergence of gene expression among duplicated genes has been observed in different species (Li et al. 2005). It is thus possible that newly duplicated genes could exhibit reduced or null expression patterns. The gene diversity (heterozygosity) at those pseudogenes should, however, be relatively low because this variation ought to behave neutrally and be in a drift-mutation balance.

In recent studies on MHC genes, sequence phylogenies show clusters of distinct MHC sequences, which exhibit very low nucleotide diversity. For instance, in a study based on 454 sequencing of MHC class IIB in flycatchers, Zagalska-Neubauer et al. (2010) demonstrated that the existence of a cluster of 19 MHC pseudogene alleles commonly differed by only one substitution. This indicates either a series of recent duplications or extensive concerted evolution. Within these kinds of clusters, the signal of balancing selection was not detected (see for instance studies from Aguilar et al. 2006 and Bollmer et al. 2010), and these clusters were therefore interpreted as pseudogenes or nonclassical MHC genes.

Irrespective of whether these clusters of sequences consist of pseudogenes or functional genes, the question remains how can so many apparently similar sequences be maintained in the population in the face of drift. In particular, sequences with an identical PBR are predicted to recognize the same epitope and ignoring dosage effects, they act as "functional equivalents." Random genetic drift should reduce the gene diversity within loci with such functional equivalent alleles according to neutral theory. Alternatively, if these sequences are gene paralogous, individuals should show large CNV, with some individuals carrying many gene copies.

Here, we investigate how the polymorphism of sequences of MHC can be maintained in natural populations of the guppy *Poecilia reticulata*. We focus on MHC class IIB genes that have been extensively studied in this species (van Oosterhout, Joyce, and Cummings 2006; van Oosterhout, Joyce, Cummings, Blais, et al. 2006; Fraser and Neff 2009; Fraser et al. 2010). We focus on one observation in particular; the extensive number of similar MHC sequences (2.3 ± 0.6 bp mean pairwise difference) observed at multiple copies in the guppy genome. We find that the nucleotide substitution pattern does not show evidence for positive selection within this cluster. We then analyze genetic variation at 13 microsatellite loci to estimate the effective population sizes and rates of gene flow and design an individual based model that uses the demographic parameters as input to examine whether this putative functionally equivalent MHC variation can be maintained in the finite gene pools.

Materials and Methods

Populations Sampled

Guppies (*P. reticulata*) are small tropical live-bearing fish that are native to streams and rivers of Trinidad, Tobago, and parts of South America. We sampled 80 guppies in 4 populations (20 individuals per population). The Upper Naranjo [UN], Mid Naranjo [MN], and Lower Aripo [LA] are populations located along the Aripo River, which is part of the Caroni Drainage in the Northern Mountain Range of Trinidad. These populations differ in parasite fauna, population size, and demography (van Oosterhout, Joyce, and Cummings 2006), and there are also large differences in neutral microsatellite (Barson et al. 2009) and SNP variation (Willing et al. 2010). The fourth population is from the Pitch Lake [PL], which is located more than 80 km away and separated from the Caroni drainage. Willing et al. (2010) analyzed 34 populations in the main drainages in Trinidad (and 3 in Venezuela) and have shown that although the PL guppy population is distinct, they are most closely related to the Caroni drainage cluster. Guppies were captured using a seine net, given a lethal dose of MS222, and stored separately in molecular grade ethanol.

MHC Class IIB Screening

Extraction and Amplification

Genomic DNA was extracted from the caudal fin using the HotSHOT protocol (Truett et al. 2000). MHC class IIB was first amplified by polymerase chain reaction (PCR) using a degenerate forward primer (*DABdegfb*:5′-GTG TCT TTA RCT CSH CTG ARC-3′), situated near the 5′ end of exon 2 of the MHC class IIB loci and a reverse primer (*DABR6b*:5′-TTA GGG TAG AAA TCA TAA ACT CTG CA-3′), situated near the 5′ end of exon 3. These primers amplify approximately 750–850 bp of genomic DNA, including 222 bp of exon 2, from codon 22, through intron 2, up to and including the first 68 bp of exon 3. Our primers thus amplified circa 82% of exon 2, the full intron 2, and 32% of exon 3. (These estimates are based on the MHC class IIB structure of the three-spined stickleback [*Gasterosteus aculeatus*] transcript number ENSGACT00000000450 of the sequence Q9GJP3_GASAC.) The forward primer (*DABdegfb*) is known to amplify all previous guppy MHC class IIB sequences (van Oosterhout, Joyce, and Cummings 2006; van Oosterhout, Joyce, Cummings, Blais, et al. 2006). In addition, we screened four guppies with the same set of primers employed by Fraser and Neff (2009) (see below).

The PCR mix of 25 μl contained 2.5 pmol of the specific reverse primer and 12.5 pmol of the degenerate forward primer, 2.5 mM $MgCl_2$, and 0.2 mM of each dNTP and 0.5 U *Taq* polymerase (Bioline Ltd., London). The touchdown PCR consisted of an initial step of 95 °C for 3 min followed by one cycle of 94 °C for 1 min, 61 °C for 1:30 min, 72 °C for 1:30 min; then one cycle of 94 °C for 1 min, 59 °C for 1:30 min, 72 °C for 1:30 min; and then 28 cycles of 94 °C for 1 min, 58 °C for 1:30 min, 72 °C for 1:30 min; with a final step of 72 °C for 30 min. Products were visualized on an agarose gel.

Cloning and Sequencing

PCR products were cloned using DH5α (Invitrogen) competent bacterial cells, using pGEM-T Easy Vector (Promega Ltd.) according to the manufacturer's instructions. For each individual sample, between 12 and 18 (mean 16.5) colonies were picked from plates (for justification of cloning and sequencing effort, see supplementary box 1, Supplementary Material online). The colonies were dipped directly into the second PCR of 35 μl, containing 7 pmol of each M13 primer, 2.5 mM $MgCl_2$, 0.2 mM of each dNTP, and 0.7 U *Taq* polymerase. In addition, we performed PCR, cloning and sequencing of 12 additional colonies from each of four randomly picked guppies from the set of 20 MN fish analyzed and amplified those with different sets of PCR primers used by Fraser and Neff (2009). This additional screening was performed to confirm that we amplified the same set of MHC class IIB sequences that was studied previously (e.g., Fraser and Neff 2009) and that we did not underestimate the true number of MHC sequences or paralogous gene copies. The PCR consisted of an initial step of 95 °C for 5 min followed by 31 cycles of 94 °C for 1 min, 54 °C for 1:30 min, and 72 °C for 1:30 min, with a final step of 72 °C for 30 min. The products were resolved on an agarose gel. ExoSAP cleanup and sequencing were performed by SymBio Corporation, USA on an ABI3730xl.

Sequences were checked and aligned using CodonCode Aligner version 2.0 and Mega 4.1 (Tamura et al. 2007). Exon 2 sequences were aligned to other guppy MHC class IIB sequences (Sato et al. 1996; van Oosterhout, Joyce, and Cummings et al. 2006) and to cichlid class IIB sequences (Figueroa et al. 2000) (for details, see supplementary fig. 2, Supplementary Material online). Errors can be introduced to sequences by heteroduplex mismatch repair or sporadic substitutions caused by *Taq* polymerase misincorporations (Kanagawa 2003). MHC sequences were confirmed when they were observed in at least two independent PCRs (Lukas and Vigilant 2005; Cummings et al. 2010). When a polymorphic site was observed in only a single sequence or in a single individual (i.e., in a single PCR–cloning–sequencing run), it was discarded from the analysis because it was considered as PCR error. Our analysis method is thus very conservative with regard to PCR errors. This stringency provides confidence about the polymorphism detected with this method. Two individuals had MHC sequences that were not confirmed by PCRs of any of the other individuals. Cloning and sequencing were repeated for one of these individuals (30 clones were sequenced), and all four sequences were independently confirmed. However, the second individual from the LA population was not cloned and sequenced again and removed from the study. The LA population thus had $N = 19$ samples analyzed, whereas the UN, MN, and PL all had $N = 20$ each.

We found a large number of sequences that were diverged by up to 3 bp from one another, and this group of sequences was labeled "cluster 1" (see below). We checked whether the predominance of cluster 1 sequences could be explained by cloning or PCR bias or by PCR errors. Three lines of evidence suggest this is not the case. Firstly, the proportion of clones that contain a copy of cluster 1 per individual (mean [SE] $= 0.526 \pm 0.144$) is not significantly higher than the proportion of cluster 1 in all sequences per individual (mean [SE] $= 0.512 \pm 0.27$) (paired *T*-test, $T = 0.32$, $P = 0.754$). Secondly, we observed a significant positive correlation between the proportion of cluster 1 sequences per individual and the proportion of clones containing a distinct cluster 1 sequence in the individual (correlation: $r = 0.616$, $P = 0.001$). (We excluded in this analysis the individuals in which cluster 1 was fixed or absent, as this would inflate the correlation.) The positive correlation suggests that the number of clones that contain cluster 1 sequences is proportional to the number of distinct sequence copies of cluster 1 in an individual. Thirdly, we amplified DNA by PCR and sequenced four individuals twice, and those had two sequences from cluster 1 both of which were found in the two independent PCRs of the same sample. (Note that all sequences were already confirmed by observing them in at least two independent PCRs in different samples.) These analyses suggest there is no PCR error or PCR/cloning bias favoring cluster 1 sequences.

Sequence Polymorphism Analyses

Statistical analyses were performed using the software R 2.13.0 (R Development Core Team 2011). Phylogenetic analysis of MHC sequences was performed with the software Mega 4.0.2 (Tamura et al. 2007). The phylogeny was built on the whole sequence using maximum likelihood method, assuming a Tamura–Nei substitution model. Bootstraps were computed using 1,000 permutations. The PBR of 66 of 223 bp of the exon 2 was defined based on van Oosterhout, Joyce, and Cummings (2006), and the d*N*/d*S* ratio was computed with DnaSP 5.10.01 (Librado and Rozas 2009) for the PBR and non-PBR codons separately. Departures from neutral expectation (null hypothesis is $dN = dS$) were tested using a *Z*-test based on codons using either the alternative hypothesis is $dN > dS$ (positive selection) or $dN < dS$ (purifying selection). Results from these tests should be taken with caution since the comparisons were performed among sequences belonging to different MHC loci, whereas these test are designed to analyze single-copy genes (Innan 2003). The nucleotide divergence between paralogous gene copies is generally higher than that of orthologous comparisons. Hence, by performing the analyses across genes in a multigene family, such elevated divergence can reduce the d*N*/d*S* ratio in the PBR due to substitution saturation effects (Hughes and Friedman 2004; van Oosterhout, Joyce, and Cummings 2006). Phylogenetic networks were built with the software SplitsTree4 (www.splitstree.org), using the neighbor-net method (Huson and Bryant 2006).

A cluster of sequences exhibiting an identical intron 2 sequence (cluster 1) was detected and studied independently. We examined whether the sequence variation within this cluster showed evidence for positive selection ($dN/dS > 1$) at the PBR and negative selection ($dN/dS < 1$) at the non-PBR. Because this was a subsample of only 27 sequences (compared with $66 - 27 = 39$ noncluster 1 sequences), we performed bootstrap analysis to check whether the sample size reduced the statistical power of the test. This subsampling process was repeated 1,000 times to compute the mean and standard deviation (SD) of d*N*/d*S*, and this bootstrap analysis was performed in Minitab 12.1 (see supplementary box 2, Supplementary Material online).

Gene Conversion

Putative recombinant sequences were identified using the recombination detection package (RDP3) (Martin et al. 2010), which includes RDP (Heath et al. 2006), GENECONV (Padidam et al. 1999), Bootscan (Martin et al. 2005), Maxchi (Smith 1992), Chimaera (Posada and Crandall 2001), SiSscan (Gibbs et al. 2000), and 3Seq (Boni et al. 2007). The detection of gene conversion events is based on the comparison of the different polymorphic sites and their location in the sequences of the data set. In the different detection methods implemented in the software RDP3, the significance tests are based on the comparison of the probability that each sequence might be explained by point mutations or by a recombination event between two sequences in the data set. Therefore, when the gene conversion is found significant, it means that it is unlikely that the sequence could be explained by point mutation, regardless whether they originate from natural mutation or PCR error. Furthermore, chimeric sequences may be generated during PCR by splicing two separate alleles. However, our protocol for acceptation of MHC alleles requires that they are observed identically in two separate individuals or reactions. Confirming a chimeric sequence would require the splicing of two of the same alleles found in separate individuals. Therefore, given that we accept only identical alleles from separate PCRs, we are extremely unlikely to accept chimeric sequences.

Tests were conducted using a critical value $\alpha = 0.05$ with the Bootscan permutations set to 1,000. *P* values were Bonferroni corrected for multiple comparisons of sequences. Sequences were linear, and phylogenetic evidence of recombination was required. The window size was set to 20 bp with a step size of 2 (and 4 in the case of Bootscan).

Estimation of Population Effective Sizes

Neutral genetic variation was used to estimate the effective sizes and the migration rates between the UN and MN populations. Extracted DNA was amplified by PCR at 13 microsatellite loci for 50 individuals per population, including two interrupted repeats: *Pr39*, *Pr92* (Becher et al. 2002); ten perfect dinucleotide repeats: *Pret-32*, *Pret-46*, *Pret-69*, *Pret-77* (Watanabe et al. 2003), *G72*, *G82*, *G211*, *G289*, *G350* (Shen et al. 2006), *Hull 9-1*; and a perfect tetranucleotide repeat, *Hull 70-2* (van Oosterhout, Joyce, Cummings, Blais, et al. 2006). Forward primers were labeled with Cy5, Cy5.5 (Eurofins MWG Operon, Germany) and WellRED D2 (Sigma-Aldrich) dyes. Microsatellites were amplified using Qiagen Multiplex PCR Kit according to the manufacturer's instructions with 30 PCR cycles and annealing temperatures of 53 °C (*Pr39*, *Pret-77*, *Hull 70-2*, and *Pret-46*), 56 °C (Pr92, Pret-69, G72 Hull 9-1, and G350), and 58 °C (*Pret-32*, G82, *G289*, and *G211*). Loci of similar annealing temperatures were multiplexed together in a 10 µl PCR (there were no more than five loci in a single reaction). PCR products were resolved on a Beckman Coulter CEQ 8000 sequencer using CEQ size standard kit (400 bp). Individuals for which the genotype of one or more loci was not unambiguously retrieved were excluded from the final analysis.

The software Migrate 2.4 (http://evolution.genetics.washington.edu/lamarc.html) was used to get estimates of population size and migration rate in UN, MN, and LA populations (located, respectively, at high, middle, and low altitude within the same river) (Beerli and Felsenstein 2001). Migrate uses a maximum likelihood (ML) coalescent approach to estimate theta (Θ), which is equal to four times the effective population size, N_e, multiplied by the mutation rate (per generation), μ ($\Theta = 4N_e\mu$), and *M*, the migration rate parameter which is the migration rate, *m*, divided by the mutation rate (m/μ). The migration parameter (*M*) was converted into number of migrants per generation (*Nm*) by

multiplying M by Θ and then dividing by four (nuclear inheritance scalar). Effective population size and migration rate were calculated based on a microsatellite mutation rate of $\mu = 2 \times 10^{-4}$ (Ellegren 2000).

Migrate was run four times, using F_{ST} estimates to start the first run. Subsequent runs were started from estimates (Θ and M) of previous runs. The Brownian motion model was used as an approximation of the stepwise mutation model. The MCMC search criteria used 200 short chains of 10,000 steps and 10 long chains of 400,000 steps with heating scheme of four temperatures (1.0, 1.2, 1.5, and 3.0). The burn-in was set to 100,000. Runs were repeated until Θ and M estimates were consistent between runs, either reaching asymptote or having overlapping 95% confidence intervals (CIs) between runs. ML estimates (with 5–95% CI) were used to compare effective population size (N_e) and migration rate (Nm) between populations.

Likelihood ratio tests of reduced migration models were tested against the full migration model in order to establish the likelihood of a barrier to migration between populations. Tests were conducted in a replicate fourth run, and models were assessed by comparison of log likelihood of the test model in comparison to the full model and using Akaike's Information Criterion in the Migrate output. Reduced migration models included: 1) no upstream migration, 2) barrier to upstream migration between the MN and the UN populations, 3) barrier to upstream migration between the LA and the MN populations, and 4) migration only between adjacent populations (direct UN to LA migration blocked).

Simulations of MHC Diversity

Translating cluster 1 DNA sequences into amino acids shows that 22 of 27 protein sequences are identical for their PBR. These sequences should be able to recognize only the same pathogens (i.e., they are functional equivalents). Hence, they should behave neutrally with respect to one another. Similarly, if they are pseudogenes, balancing selection cannot act on this variation, and the variation in cluster 1 should be prone to drift. We therefore tested whether balancing selection could maintain the total number of haplotypes observed and also whether the observed number of functionally equivalent haplotypes from cluster 1 could be maintained in the finite gene pools without (pathogen-mediated) overdominant selection. An individual based model was thus constructed to simulate the genetic diversity of MHC class II loci in natural guppy populations.

The population size and migration rate estimates were obtained from the software Migrate (see above). The UN and MN effective population size were $N_e = 1085$ and $N_e = 1251$, respectively. Migration occurred every generation, with on average, 1.9 moving downstream from the UN to the MN and 0.55 migrants going upstream from the MN into the UN and the MN receiving 0.61 migrants per generation from the LA population. The LA is a source population that is connected to the Caroni Drainage metapopulation with extremely large N_e (Barson et al. 2009), and hence, the MHC sequence frequency was assumed to remain stable over time in our simulations.

Previous studies indicate that the MHC class II shows CNV in guppies (van Oosterhout, Joyce, and Cummings 2006; van Oosterhout, Joyce, Cummings, Blais, et al. 2006), and hence, we simulated a variable number of distinct MHC sequences per individual. The simulated mean number of distinct sequences per individual was drawn from a uniform distribution with range from 1 to 6. We have not found evidence for linkage disequilibrium between our MHC sequences, and hence, we simulated free recombination between sequences occurring at different MHC loci. Furthermore, at the start of the simulations, populations possessed the same number of distinct MHC exon 2 sequences observed in our UN, MN, and LA populations combined ($N = 39$). Of those 39 sequences, 11 belonged to cluster 1 (see Results), and those are assumed to be selectively functionally equivalent and thus behave neutrally with respect to each other. Symmetric overdominance selection was operating on the sequences within a locus with selection coefficient s ($s = 0$, 0.02, 0.05, 0.1, 0.15, 0.2, and 0.5). When a locus was homozygous or when an individual was heterozygous for two distinct cluster 1 sequences, the fitness conferred by this locus was $1 - s$. Fitness effects were multiplicative across loci. The simulations were run for 20,000 generations, and the data were collected at 10 generation intervals from generation 15,000 to 20,000. Populations had reached a migration–selection–drift equilibrium after 15,000 generations. The output of the model was the number of distinct sequences in the population (A_p), and crucially, the number of neutral sequences within populations (A_{p0}). The mean and 5–95% CIs for A_p and A_{p0} were computed over the final 5,000 generations and averaged across five replicated runs.

Results

Phylogeny of MHC Class II Sequences

A total of 66 distinct sequences of MHC class IIB were detected in the four wild guppy populations. Figure 1 shows the phylogeny of these sequences. A group of 27 sequences, highlighted by a rectangle, exhibited very low diversification (2.3 ± 0.6 bp mean pairwise difference within the whole 1,068 bp sequence; 1.2 ± 0.4 bp within exon 2). These 27 sequences also shared the same intron 2 sequence and were labeled cluster 1. Individuals exhibited between 0 and 5 copies of cluster 1, which shows that this sequence is present at multiple MHC loci and that these sequences are paralogs.

Diversity of MHC Sequences

Based on 79 guppies, 39 distinct MHC class IIB sequences of exon 2 of the 66 distinct sequences of MHC were detected. The number of distinct exon 2 sequences per individuals varied from 1 to 6. Since the guppy is a diploid species, this suggests that some individuals have at least three distinct

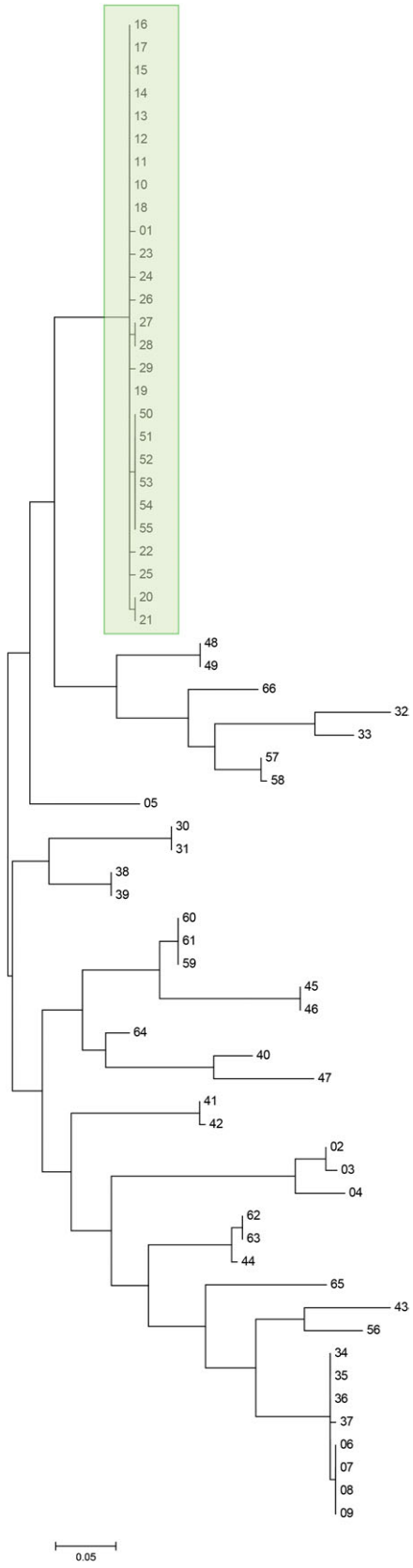

Fig. 1. Phylogeny of MHC class IIB sequences based on the whole sequence (including exon 2, intron 2, and exon 3). The rectangle highlights cluster 1 sequences. Note that a similar, albeit smaller, cluster of sequences is present in the bottom half of the tree.

MHC loci. The number of MHC genes could be even higher if some sequences were not amplified using our primers (i.e., null sequences) or when sequences at different MHC loci are identical at the amplicon (i.e., possess no distinguishable SNPs). Indeed, by including the intron 2 and exon 3 sequence variation, we uncovered $66 - 39 = 27$ sequences (40.9%) that would have been missed if we had amplified the exon 2 variation only. (Note that these 27 are not the same as the 27 sequences of cluster 1.) The mean number of distinct exon 2 sequences per individual in the UN and MN population was 2.75 ± 1.16 and 2.25 ± 0.78, respectively.

Cluster 1 sequences were detected in only three of the four populations (UN, MN, and LA), and they are present at an extraordinarily high frequency in the UN and MN (61.8% and 46.6% of sequences detected in UN and MN, respectively, belonged to cluster 1, compared with only 14.3% in the LA). Figure 2 shows the presence of the different exon 2 sequences of cluster 1 in UN and MN. In the UN population, we found nine sequences, whereas MN population exhibited seven sequences, and five sequences were found in both populations.

Nucleotide Polymorphism in the Whole Sample

Exon 2

No premature stop codons or frameshift mutations were detected in exon 2 sequences, and there was no evidence to suggest any of these sequences represent a pseudogene. Exon 2 exhibited a large mean nucleotide diversity ($\pi = 0.17 \pm 0.01$) across all sequences. The nucleotide diversity was higher in the PBR ($\pi = 0.30 \pm 0.025$) than in the non-PBR ($\pi = 0.11 \pm 0.01$) (paired Wilcoxon test: $V = 2.10^6$, $P < 0.001$). There was a significant difference in the number of segregating sites between the PBR and non-PBR ($\chi^2 = 21.97$, degree of freedom [df] $= 1$, $P < 0.001$), suggesting diversifying selection targeting the PBR. Given that they are synonymous substitutions, they should be neutral and not be experiencing diversifying selection. The relatively elevated dS in the PBR may be explained by gene conversion (see below). The high value of the Tajima's D ($D = 2.34$) and the relatively large proportion of nonsynonymous substitutions ($dN/dS = 2.26$), confirmed the signal of balancing selection in the PBR (Z-test for positive selection, $P = 0.02$). On the contrary, in the non-PBR, both Tajima's D ($D = 0.38$) and the dN to dS ratio ($dN/dS = 0.37$) were much smaller, indicating a significant signal of purifying selection (Z-test for purifying selection: $P = 0.001$).

Intron 2

Intron 2 was highly variable in terms of size (ranging from 473 to 778 bp). Nevertheless, intron 2 exhibited more than four times lower nucleotide diversity ($\pi = 0.04 \pm 0.005$) than the exon 2 ($\pi = 0.17 \pm 0.01$). The number of segregating sites was also significantly lower in intron 2 than in exon 2 ($\chi^2 = 210.22$, df $= 1$, $P < 0.001$), consistent with diversifying selection acting on large parts of exon 2 and neutral evolution in intron 2.

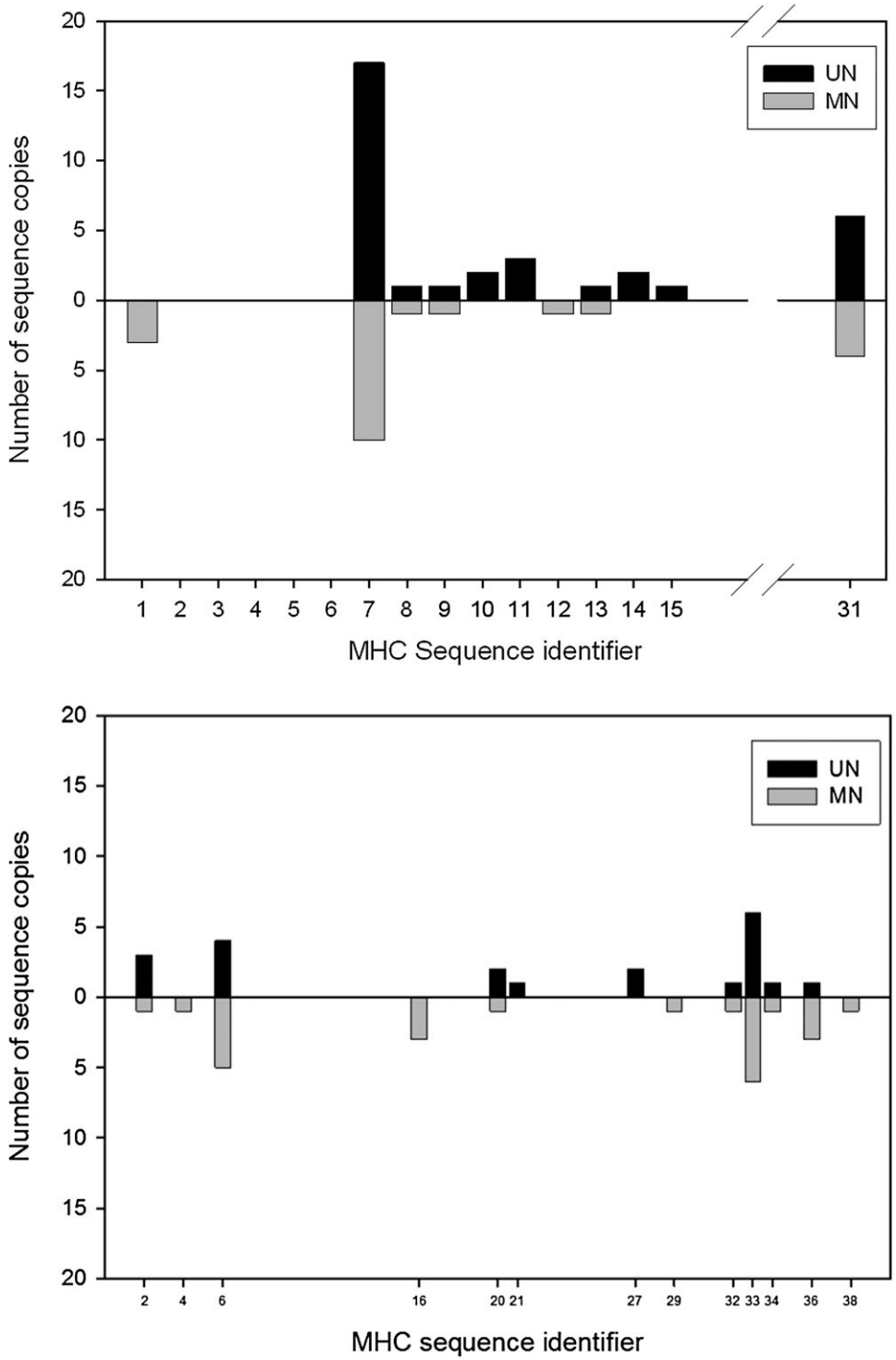

Fig. 2. Distribution of copy number of distinct sequences belonging to cluster 1 (top graph) or within the other (noncluster 1) sequences (bottom graph) in the UN (black bars) and MN (light bars) populations. Each distinct exon 2 sequence is identified by a number (ranging from 1 to 38).

Exon 3

Exon 3 also exhibited a low nucleotide diversity ($\pi = 0.04 \pm 0.01$), had a Tajima's D close to zero ($D = 0.23$), and a d*N*/d*S* which did not significantly depart from unity (d*N*/d*S* = 0.71, $P = 0.63$). The number of segregating sites was also significantly lower in exon 3 than in exon 2 ($\chi^2 = 30.34$, df = 1, $P < 0.001$). Exon 3 contained no stop codons or indels in any of the sequences. The exon 3 thus appears to evolve slowly with its genetic polymorphisms being restricted by purifying selection.

Gene Conversion

The haplotype network (fig. 3) of the 66 whole MHC class IIB sequences exhibits a large number of loops (parallel branches that join allele phylogenies). These loops are typical for the MHC of many other species, and they indicate

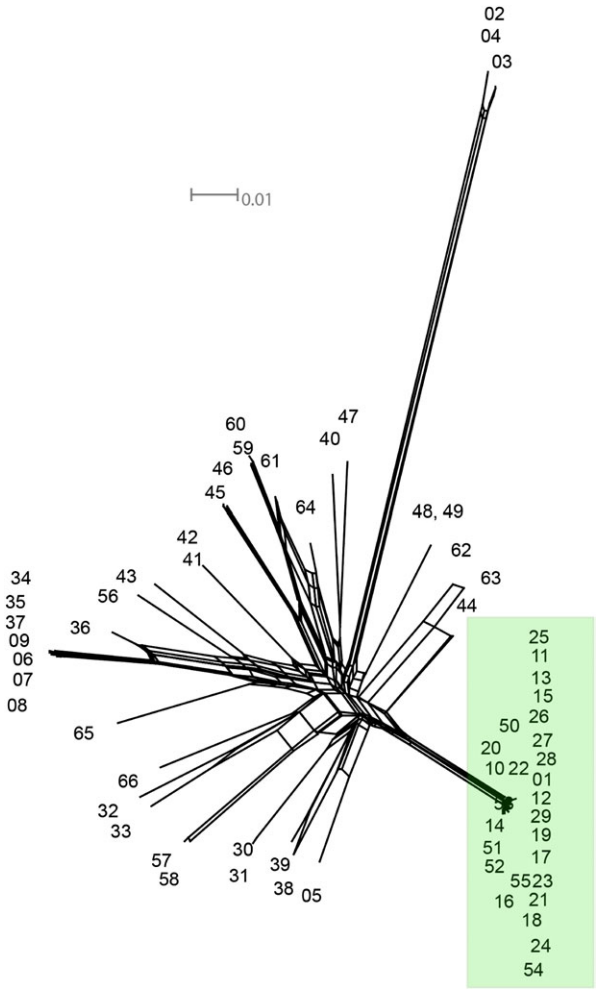

FIG. 3. Network of MHC class IIB sequences. The rectangle highlights cluster 1 sequences.

distinctly dissimilar haplotypes that share areas with a high level of sequence similarity. This is indicative of exchange of fragments among diverged sequences, which could be due to gene conversion or recombination. The RDP3 indeed detected significant recombination events in 12 of 66 sequences with various length of fragments exchanged (mean = 551 ± 330 SD) (table 1). Three of these events were located within the exon 2, four within intron 2, and five were overlapping this exon and intron (fig. 4).

Polymorphism within Cluster 1

The 27 sequences from cluster 1 exhibit 2.3 ± 0.6 bp mean pairwise difference. Only 21 sites were polymorphic: 10 of those sites were located in the exon 2 (4 within PBR and 6 in the non-PBR), whereas the 11 others sites were located within intron 2.

Exon 2 of Cluster 1

The 27 sequences that shared an almost identical intron 2 sequence in cluster 1 also exhibited a low nucleotide diversity ($\pi = 0.005 \pm 0.002$) in their exon 2. This diversity is more than 40 times lower than among the 39 sequences not belonging to cluster 1 ($\pi = 0.21 \pm 0.015$). Among the sequences of cluster 1, the nucleotide diversity was not significantly higher in the PBR ($\pi = 0.007 \pm 0.003$) than in the non-PBR ($\pi = 0.005 \pm 0.002$). Furthermore, there was no significant difference in number of segregating sites between PBR and non-PBR ($\chi^2 = 2.75$, df = 1, $P = 0.097$).

The cluster 1 sequences had a low Tajima's D ($D = -1.54$) and dN to dS ratio ($dN/dS = 0.06$, Z-test for positive selection: $P = 0.33$) in the PBR as well as in the non-PBR ($D = -1.56$ and $dN/dS = 0.11$, Z-test for positive selection: $P = 0.41$), indicating no signal of balancing selection. Among the 27 distinct sequences belonging to cluster 1, 22 sequences shared an identical amino acid motif for the putative PBR, which makes them functionally equivalent with regard to pathogen-mediated selection. These 22 functionally equivalent sequences should thus behave neutrally with regards to one another if pathogen selection was the only selective force operating on this genetic variation.

Intron 2 of Cluster 1

All intron 2 sequences of cluster 1 had the same length, and this intron exhibited a very low nucleotide diversity ($\pi = 0.002 \pm 0.001$), which is more than 20 times lower than among the 39 sequences not belonging to cluster 1

Table 1. Recombinant and Parental (major and minor) Sequences Identified Using the RDP3 Package (Martin et al. 2010).

Sequences Involved			Length of Fragment	Detection Methods						
Recombinant	Major	Minor	Exchanged (bp)	RDP	GENECONV	Bootscan	Maxchi	Chimaera	SiSscan	3Seq
1	48	63	810	NS	NS	NS	2.80×10^{-5}	2.09×10^{-5}	6.59×10^{-5}	2.58×10^{-6}
3	58	−6	812	4.90×10^{-4}	1.81×10^{-4}	NS	3.05×10^{-8}	3.26×10^{-4}	7.06×10^{-14}	NS
33	57	43	52	4.94×10^{-2}	NS	NS	NS	2.63×10^{-2}	NS	2.23×10^{-3}
35	44	−48	618	NS	NS	NS	NS	NS	1.30×10^{-2}	1.21×10^{-2}
36	42	35	219	NS	4.92×10^{-2}	2.50×10^{-5}	2.05×10^{-6}	2.05×10^{-6}	NS	9.34×10^{-16}
38	31	5	842	NS	NS	5.49×10^{-3}	2.91×10^{-3}	NS	NS	NS
40	3	66	809	1.43×10^{-3}	9.13×10^{-4}	1.78×10^{-9}	5.25×10^{-9}	3.04×10^{-4}	7.86×10^{-14}	NS
48	5	−62	240	NS	NS	NS	2.42×10^{-3}	NS	1.88×10^{-3}	8.38×10^{-3}
56	49	37	362	NS	NS	NS	7.48×10^{-3}	9.88×10^{-4}	1.48×10^{-4}	6.50×10^{-4}
57	66	37	809	NS	NS	NS	2.15×10^{-2}	8.59×10^{-3}	1.69×10^{-3}	NS
62	36	33	100	NS	NS	NS	1.59×10^{-3}	NS	1.60×10^{-2}	1.65×10^{-4}
63	−48	44	934	NS	NS	NS	NS	NS	0.039785	6.46×10^{-6}

NOTE.—The start and end points of the recombination events are indicated. Putative recombination events were confirmed when two or more software packages indicated significant evidence for gene conversion (with $\alpha = 0.05$). Sequences preceded by a dash indicated that the actual sequence was unknown but the sequence preceded by a dash is the closest likely candidate. NS, not significant.

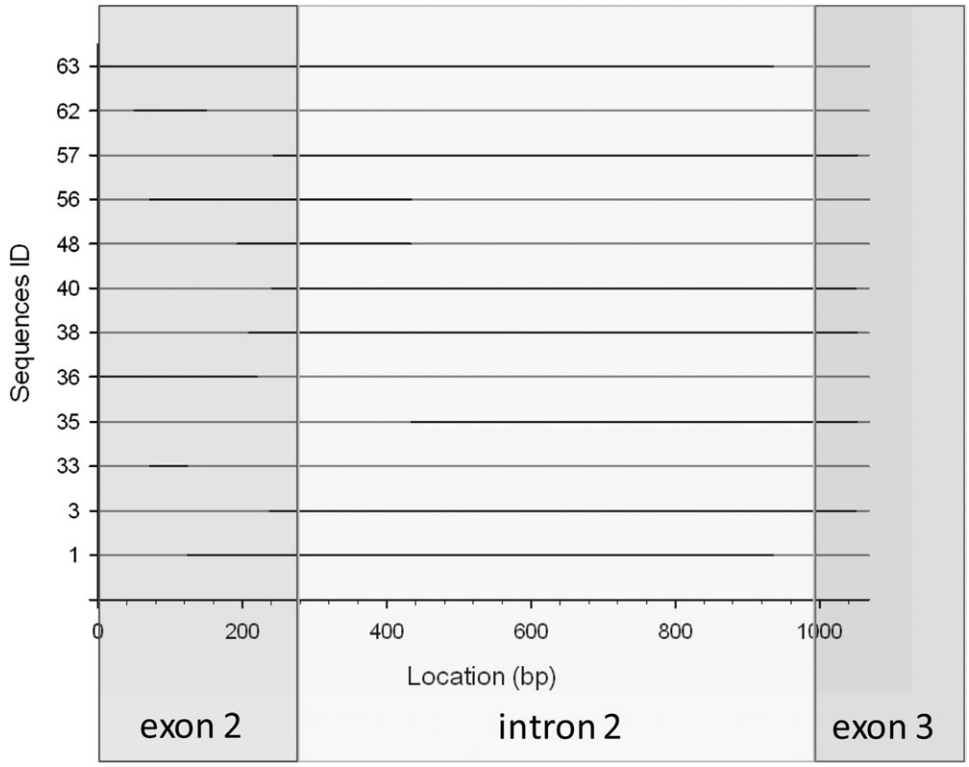

Fig. 4. Gene conversion events detected using RDP3 package. Sequences are represented as a gray line, and the parts of sequences that have been exchanged through gene conversion events are figured in black.

($\pi = 0.053 \pm 0.005$). All pairs of sequences from cluster 1 exhibit at most one nucleotide difference only. These are not due to PCR or sequencing errors because all sequences were independently confirmed by multiple independent PCRs (mean number of PCRs per cluster 1 sequence: 26.2; range [4–244]).

Exon 3 of Cluster 1

All the 27 cluster 1 sequences were identical for their exon 3, suggesting strong purifying selection, a recent origin of duplication, and/or a high rate of gene conversion. The absence of variation prevented the computation of summary statistics. Strong purifying selection on exon 3 could result in a selective sweep eroding variation in the neighboring intron 2, but this explanation seems unlikely given that we do detect considerable polymorphisms in the adjacent exon 2.

Prediction of the MHC Variation in UN and MN

The high occurrence of sequences belonging to cluster 1 in UN ($N = 9$) and MN ($N = 7$) is puzzling because first, they do not exhibit a signature of diversifying selection and second, 7 of 11 exon 2 sequences showed an identical PBR motif and are thus likely to recognize the same pathogens. Based on this nucleotide variation, they should evolve neutrally with respect to each other. We performed computer simulations to ascertain 1) the selection coefficients required to maintain the MHC polymorphism in the populations and 2) test whether this large number of cluster 1 sequences can be maintained in migration–drift equilibrium. The simulations use the estimated effective population size and migration rates from the migrate analysis. Figure 5 shows that for selection coefficients ($0.1 \leq s \leq 0.2$), the 5–95% CI of the number of distinct MHC sequences simulated in UN and MN population (A_p) overlap the observed values. The observed MHC sequences diversity is thus consistent with the expectations for loci under overdominance selection in the demographic scenario inferred from the microsatellites. However, figure 5 also shows that the simulated number of functionally equivalent MHC sequences (A_{p0}) is significantly smaller than the observed values. The observed haplotype diversity of cluster 1 in the UN and MN population is thus unlikely to be maintained through neutral evolution alone. This suggests that other selective forces are operating on the sequence variation outside the studied amplicon.

Discussion

In this study on the MHC class II variation of four populations of guppies, we amplified MHC class II *DAB* loci using different primer combinations. From a total of 66 sequences, there was a group of 27 sequences (cluster 1) with high sequence similarity (2.3 ± 0.6 bp mean pairwise difference). This compares with an overall mean pairwise sequence difference of 55.4 ± 4.0 bp. Guppies can possess up to five distinct copies of a cluster 1 sequence, which shows that sequence similarity does not necessarily reflect locus affiliation (cf. MHC of passerine birds, Sato et al. 2011) and that

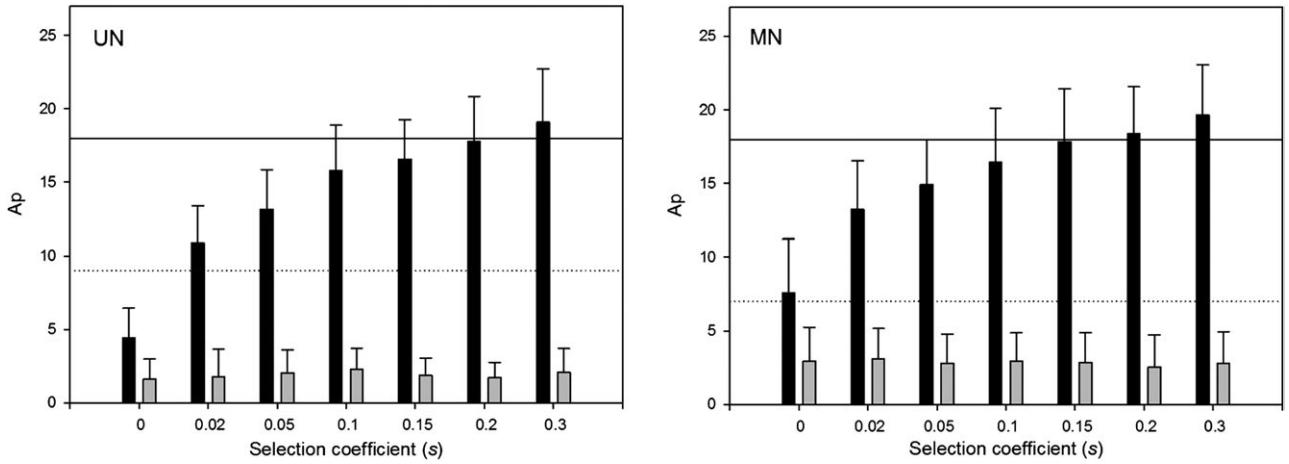

FIG. 5. Observed (lines) and simulated values (columns) of the number of MHC sequences per population (A_p) for UN (left graph) and MN (right graph) with symmetric overdominance across a range of selection coefficients (s = 0, 0.02, 0.05, 0.1, 0.15, 0.2 and 0.3). The black bars represent the mean number (±5–95% CI) of all sequences (expected A_p) and the gray bars represent the number of cluster 1 sequences that are "functionally equivalent" and neutral with respect to one another (expected A_{p0}). The solid lines represent the observed total number of distinct MHC sequences (observed A_p) and dotted lines are the observed number of cluster 1 sequences (observed A_{p0}). The simulated populations had an effective population size $N_{e(UN)}$ = 1,085 and $N_{e(MN)}$ = 1,251, migration rates $Nm_{(UN-MN)}$ = 1.9, $Nm_{(MN-UN)}$ = 0.55, $m_{(LA-MN)}$ = 0.61. (For details, see text.)

cluster 1 sequences have been duplicated in their genome. These observations are consistent with recent gene duplication and/or a high rate of gene conversion. We found evidence of gene conversion using the software RDP3 (Martin et al. 2010). Gene conversion explained the large number of loops observed in the MHC sequences network (see also Spurgin et al. 2011 for a detailed analysis on gene conversion in the MHC in recently founded bird populations). The most puzzling observation was the large number of almost identical MHC class IIB sequences that was maintained in three small guppy populations (UN, MN, and LA). When translated, 22 were identical for their amino acids at the PBR, and we found no evidence of selection acting on sequence variation within this cluster 1. Previously, we showed that the same sequences were also found when amplifying cDNA, which suggests that these genes are expressed and not pseudogenes (van Oosterhout, Joyce, and Cummings 2006). Even if these sequences represent pseudogenes, they should behave selectively neutral with respect to one another. An individual based model showed that this large number of apparently functionally equivalent sequences (or pseudogenes) could not be maintained in drift–migration equilibrium at one to three loci without balancing selection. We also explored whether we may have underestimated the CNV by reamplifying a subset of guppies with a different primer set, but this did not uncover new sequences. Altogether, these results suggest that besides pathogen-mediated selection, other evolutionary forces are acting on this MHC variation or that gene conversion leads us to significantly underestimate the true CNV.

Cluster 1 Sequence Variation

Across all sequences, we observed a significant departure from neutral expectations in the PBR of exon 2. The high nucleotide diversity and the relative excess in amino acid replacement substitutions observed in the PBR is a signature for diversifying or balancing selection (Wright and Gaut 2005), and this has been observed in many other species (Spurgin and Richardson 2010). However, more intriguing, we found 27 marginally differentiated sequences of cluster 1 that did not exhibit a signal of diversifying selection in the PBR.

These cluster 1 alleles were detected in the three populations of the Caroni drainage (UN, MN, and LA) but not in the Pitch Lake (PL) population. The PL is an isolated population in which the fish are exposed to extreme environmental conditions, such as high temperatures and high levels of hydrocarbons. The parasite composition differs from that of the three other populations (Schelkle et al. 2011), which could alter the parasite-mediated selection pressures and result in differences in the composition of MHC class II alleles.

Among the cluster 1 sequences, the PBR appeared to evolve according to the neutral expectation, based on Tajima's *D*, d*N*/d*S* ratio, and number of substitutions in the PBR relative to the non-PBR. Our analysis on the number of clones containing cluster 1 in each individual also indicates that the observed diversity was not due to cloning or PCR bias and that the high copy number of cluster 1 was thus not an artifact of our screening method. Possibly, the expression level may differ among the *DAB* copies of the cluster 1 alleles, making them functionally distinct. There is a growing number of studies reporting clusters of highly similar MHC sequences (e.g., Fraser and Neff 2009; Zagalska-Neubauer et al. 2010), and it is likely that the occurrence of such clusters has been grossly underestimated due to the validation protocols which advocate the binning of sequences that are distinct from each other by one or few SNPs. This observation is intriguing, particularly

because the evolutionary significance of this variation remains unknown.

Despite a signal of neutral evolution, the high frequency of cluster 1 sequences in two populations (61.8 and 46.6% in UN and MN, respectively), the high numbers of copies in some individuals (up to five per individual), and the existing CNV in the population (the mean number of cluster 1 sequences per individual was 2.6, 1.45, 0.33, and 0 in UN, MN, LA, and PL populations, respectively) suggest that the molecular evolution of cluster 1 is not strictly neutral. Our computer simulations indeed indicated that this high degree of diversity cannot be explained by neutral evolution alone. Furthermore, it is unlikely that cluster 1 sequence variation exists in the guppy genome in many paralogous gene copies, given that we cannot amplify more than six distinct exon 2 sequences per individual despite using different sets of primer combinations and a more in depth cloning and sequencing of a subset of individuals. However, gene conversion may have homogenized sequence variation across gene paralogs, which in turn could have masked CNV. Hence, we cannot completely rule out the possibility that some individuals may possess more than three *DAB* genes in their genome. The number of duplicated genes in our model was, however, based on the observed CNV in guppies. Increasing the number of duplicated genes inflates the number of alleles per individual above what we have observed in our populations. We parameterized our model based on the available empirical data, and given these model assumptions, we cannot explain the existence of so many functional equivalent cluster 1 alleles in populations.

Copy Number Variation

We are alerted to the possibility that many studies could be underestimating CNV, given that we uncovered 40.9% more haplotypes by also sequencing intron 1 and exon 3. This haplotype variation would have been missed if we had only amplified the exon 2. We believe that the true MHC haplotype variation and CNV may be systematically underestimated in the literature because many studies sequence only the exon containing the PBR. Given that gene conversion tends to homogenize nucleotide variation among member genes, we anticipate that this has resulted in a significant underestimation of the actual CNV of the MHC. Possibly, given that gene conversion is also implicated in the evolution of R-resistance genes (Ribas et al. 2011), the self-incompatibility locus (*S*-locus) in plants (Charlesworth et al. 2003), and mating types genes in oomycetes (Cvitanich et al. 2006), CNV could be underestimated in other multigene families as well.

Effect of Gene Conversion on the Polymorphism at MHC Loci

Takuno et al. (2008) suggest that duplicated genes could be maintained even after loss of function in the case of genes under diversifying selection: the duplicated copy could promote the polymorphism through gene conversion. We find that the cluster 1 sequences exhibit low diversity that does not appear to be promoted by pathogen-mediated selection. However, cluster 1 sequences do not appear to be disproportionally involved in gene conversion (1 of 12 sequences exhibiting a recombination event belonged to cluster 1). Nevertheless, gene conversion events among cluster 1 sequences may not have been detected because of insufficient nucleotide polymorphisms weakens the statistical power of these analyses (van Oosterhout C, unpublished data). Indeed, gene conversion tends to decrease the nucleotide diversity among sequences but increase the number of haplotypes (Nei and Rooney 2005). It is therefore possible that in guppy MHC class II, gene conversion may have reduced the nucleotide variation in cluster 1, thereby reducing the statistical power to detect it whilst increasing the number of distinct sequences in this cluster. At present, we are unable to test this hypothesis.

Selection Promoting Cluster 1 Gene Duplication

The retention of duplicated (paralogous) gene copies has been explained by various models, including the neofunctionalization (Ohno 1970) and subfunctionalization (Lynch and Force 2000) models (the latter is also known as the duplication–degeneration–complementation [DDC] model). Duplicated gene copies can be maintained when they perform different or new functions or when paralogous genes become specialized in their expression patterns (Lynch and Force 2000; Li et al. 2005). The DDC model postulates that a pair of duplicate genes degenerates complementary parts of their *cis*-regulatory motifs, which means that the ancestral gene product can only be fully produced when both copies are expressed. These models can explain the existence of two or perhaps a small number of paralogous gene copies. However, they are unlikely to explain the retention of such a large number of sequences (n = 27 in 79 individuals analyzed) because this would require that all copies have degenerated different parts of their motifs.

Selection Favoring Cluster 1 Sequences

The PBR sequence of cluster 1 may recognize an antigen of a common pathogen. Parasite-challenge experiments demonstrated that a particular MHC sequence in guppies was consistently associated to a low parasite load of a common group of fish pathogens (*Gyrodactylus* spp.) (Fraser and Neff 2010; Fraser et al. 2010). It may therefore offer an important selective advantage to hosts carrying one or more copies of this sequence (cf. gene dosage effect as has been reported in human, see for instance Engelmann et al. 2010). This could promote gene duplication (cf. accordion model of MHC evolution, see Klein et al. 1993) and/or gene convergence, resulting in many paralogous gene copies. If cluster 1 sequences are critical for pathogen defense, they could constitute a stronger selective advantage than any other MHC sequence (cf. asymmetric overdominance selection). Selection could also act on other pleiotropic fitness effects (not necessarily immune related) that may be associated to the cluster 1 alleles, potentially resulting in an

unconditional fitness advantage of individual carrying one or more cluster 1 sequences.

In the supplementary information (see supplementary fig. 1, Supplementary Material online), we therefore analyzed the effects of asymmetric overdominance, attributing cluster 1 sequences an unconditional advantage over other sequences when in homozygote state (this asymmetric overdominance model is described, e.g., by Hartl and Clark 2007). Although this increased the equilibrium frequency and number of cluster 1 sequence copies, the effect is insufficient to explain the large number of copies. We therefore rejected the hypothesis that asymmetric overdominance alone (due to a higher fitness conferred by cluster 1 alleles) can explain the large number of cluster 1 sequences in the UN and MN guppy populations.

Selection Outside the Studied Amplicon

Balancing selection could be operating also outside the studied amplicon. The class II MHC is a heterodimeric protein composed of an α and β chain (Hughes and Yeager 1998). The PBR is coded by two different exons present in the class IIA and IIB gene. Balancing selection could be operating on the PBR of class IIA. If both classes are linked, the large number of highly similar cluster 1 sequences (of class IIB) could be maintained through genetic hitchhiking. However, in the human MHC, the class IIA has consistently fewer alleles than the class IIB genes (Robinson et al. 2011). If teleost class II MHC is comparable to that of humans, this explanation becomes unlikely. Nevertheless, in human MHC class II, there are specific αβ pairings that are dependent on the polymorphisms exhibited by these alleles (Bondinas et al. 2007). It may therefore be possible that the almost monomorphic cluster 1 alleles pair with specific alleles present in the A genes. Epistasis in combination with balancing selection on the class IIA genes could theoretically maintain the diversity in this cluster 1.

It is increasingly recognized that immune genes other than the classical MHC class I and II could be subject to selection (Jensen et al. 2008). A recent theoretical study has shown that if selection operates outside the PBR—for example, on the "sheltered load" that is thought to be associated to the MHC—it can also maintain a high level of polymorphism (cf. ABC evolution, van Oosterhout 2009). Analogous to the *S*-locus evolution of selfing plants (Llaurens et al. 2009), deleterious mutation can accumulate around loci evolving under balancing selection. For example, the human MHC (HLA) is associated to a high mutational load as is evident from the more than 100 genetic disorders associated to this gene complex (Shiina et al. 2006). The sequences of cluster 1 may carry a lower mutational load or have a sheltered load that is less expressed when they unify with other MHC sequences (because of having few recessive deleterious SNPs in common with the other sequences). Van Oosterhout (2009) showed that such sequences can be maintained in the population at a considerably elevated frequency. The accumulation of deleterious mutations in the flanking regions of the MHC is dependent on the degree of linkage in the genomic region considered. We did test the rate of gene conversion and microrecombination, and they were low, albeit not zero. However, van Oosterhout (2009) showed that epistatic selection can reduce the effective recombination rate, reinforce linkage disequilibria, and thus maintain a sheltered load. We therefore favor the hypothesis that the cluster 1 sequences are associated with nonneutral (functional or recessive deleterious) polymorphisms outside the studied amplicon on which selection is acting.

Further experiments are warranted to examine the fitness effects of cluster 1 sequences, particularly with regards to pathogen resistance and the mutational load. An important advantage of guppies as a model system is that vast numbers of experimental crosses can be made to test deviations from Mendelian segregation of sequences in order to quantify the mutational load associated to the MHC. In addition, a comparative analysis of much longer MHC amplicons using second generation sequencing techniques could be used to examine sequence (dis)similarity between different sequences to test whether cluster 1 is indeed strongly differentiated from other sequences in the surrounding MHC region as predicted by ABC evolution.

Supplementary Material

Supplementary boxes 1 and 2 and figures 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank the Marie Curie Fellowship PIEF-GA-2009-254065 to V.L. and a Natural Environment Research Council (NE/E529241/1) studentship to M.M. for funding, Christine Dreyer and all members of her research team in Max Planck Institute in Tübingen for stimulating scientific discussions on guppy MHC, and David Richardson and Lewis Spurgin for helpful comments on the manuscript.

References

Aguilar A, Edwards SV, Smith TB, Wayne RK. 2006. Patterns of variation in MHC class II beta loci of the little greenbul (*Andropadus virens*) with comments on MHC evolution in birds. *J Hered.* 97:133–142.

Aguilar A, Roemer G, Debenham S, Binns M, Garcelon D, Wayne RK. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proc Natl Acad Sci U S A.* 101:3490–3494.

Andres AM, Hubisz MJ, Indap A, et al. (12 co-authors). 2009. Targets of balancing selection in the human genome. *Mol Biol Evol.* 26:2755–2764.

Barson NJ, Cable J, van Oosterhout C. 2009. Population genetic analysis of microsatellite variation of guppies (*Poecilia reticulata*) in Trinidad and Tobago: evidence for a dynamic source-sink metapopulation structure, founder events and population bottlenecks. *J Evol Biol.* 22:485–497.

Becher SA, Russell ST, Magurran AE. 2002. Isolation and characterization of polymorphic microsatellites in the Trinidadian guppy (*Poecilia reticulata*). *Mol Ecol Notes.* 2:456–458.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A.* 98:4563–4568.

Bollmer JL, Dunn PO, Whittingham LA, Wimpee C. 2010. Extensive MHC class II B gene duplication in a passerine, the common yellowthroat (*Geothlypis trichas*). *J Hered.* 101:448–460.

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

Bonhomme M, Doxiadis GGM, Heijmans CMC, Vervoort V, Otting N, Bontrop RE, Crouau-Roy B. 2008. Genomic plasticity of the immune-related Mhc class I B region in macaque species. *BMC Genomics* 9(No. 514).

Boni MF, Posada D, Feldman MW. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* 176:1035–1047.

Castric V, Vekemans X. 2004. Plant self incompatibility in natural populations: a critical assessment of recent theoretical and empirical advance. *Mol Ecol.* 13:2873–2889.

Charlesworth D, Mable BK, Schierup MH, Bartolome C, Awadalla P. 2003. Diversity and linkage of genes in the self-incompatibility gene family in *Arabidopsis lyrata*. *Genetics* 164:1519–1535.

Cummings SM, McMullan M, Joyce DA, van Oosterhout C. 2010. Solutions for PCR, cloning and sequencing errors in population genetic analysis. *Conserv Genet.* 11:1095–1097.

Cvitanich C, Salcido M, Judelson HS. 2006. Concerted evolution of a tandemly arrayed family of mating-specific genes in *Phytophthora* analyzed through inter- and intraspecific comparisons. *Mol Genet Genomics.* 275:169–184.

Eimes JA, Bollmer JL, Whittingham LA, Johnson JA, van Oosterhout C, Dunn PO. 2011. Rapid loss of MHC class II variation in a bottlenecked population is explained by drift and copy number variation. *J Evol Biol.* 24:1847–1856.

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* 16:551–558.

Engelmann R, Eggert M, Neeck G, Mueller-Hilke B. 2010. The impact of HLA-DRB alleles on the subclass titres of antibodies against citrullinated peptides. *Rheumatology* 49:1862–1866.

Figueroa F, Mayer WE, Sultmann H, O'HUigin C, Tichy H, Satta Y, Takezaki N, Takahata N, Klein J. 2000. Mhc class IIB gene evolution in East African cichlid fishes. *Immunogenetics* 51:556–575.

Fraser BA, Neff BD. 2009. MHC class IIB additive and non-additive effects on fitness measures in the guppy *Poecilia reticulata*. *J Fish Biol.* 75:2299–2312.

Fraser BA, Neff BD. 2010. Parasite mediated homogenizing selection at the MHC in guppies. *Genetica* 138:273–278.

Fraser BA, Ramnarine IW, Neff BD. 2010. Selection at the MHC class IIB locus across guppy (*Poecilia reticulata*) populations. *Heredity* 104:155–167.

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

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16:573–582.

Hartl DL, Clark AG. 2007. Principles of population genetics. 4th ed. Sunderland (MA): Sinauer Associates, Inc.

Heath L, van der Walt E, Varsani A, Martin DP. 2006. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J Virol.* 80:11827–11832.

Hughes AL, Friedman R. 2004. Recent mammalian gene duplications: robust search for functionally divergent gene pairs. *J Mol Evol.* 59:114–120.

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 U S A.* 86:958–962.

Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet.* 32:415–435.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23(2):254–267.

Innan H. 2003. The coalescent and infinite-site model of a small multigene family. *Genetics* 163:803–810.

Jensen LF, Hansen MM, Mensberg KL, Loeschcke V. 2008. Spatially and temporally fluctuating selection at non-MHC immune genes: evidence from TAP polymorphism in populations of brown trout (*Salmo trutta, L.*). *Heredity* 100:79–91.

Kanagawa T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng.* 96:317–323.

Klein J, Ono H, Klein D, O'Huigin C. 1993. The accordion model of MHC evolution. *Prog Immunol.* 8:137–143.

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet.* 21:602–607.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Llaurens V, Gonthier L, Billiard S. 2009. The sheltered genetic load linked to the S-locus: new insights from theoretical and empirical approaches in sporophytic self-incompatibility. *Genetics* 183: 1105–1118.

Lukas D, Vigilant L. 2005. Reply: facts, faeces and setting standards for the study of MHC genes using noninvasive samples. *Mol Ecol.* 14:1601–1602.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26:2462–2463.

Martin DP, Posada D, Crandall KA, Williamson C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses.* 21:98–102.

Maruyama T, Nei M. 1981. Genetic variability maintained by mutation and over-dominant selection in finite populations. *Genetics* 98:441–459.

Mehta RB, Nonaka MI, Nonaka M. 2009. Comparative genomic analysis of the major histocompatibility complex class I region in the teleost genus Oryzias. *Immunogenetics* 61:385–399.

Mona S, Crestanello B, Bankhead-Dronnet S, Pecchioli E, Ingrosso S, D'Amelio S, Rossi L, Meneguz PG, Bertorelle G. 2008. Disentangling the effects of recombination, selection, and demography on the genetic variation at a major histocompatibility complex class II gene in the alpine chamois. *Mol Ecol.* 17:4053–4067.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet.* 39:121–152.

Ohno S. 1970. Evolution by gene duplication. Heidelberg (Germany): Springer-Verlag.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* 265: 218–225.

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

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A.* 98:13757–13762.

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/

Ribas AF, Cenci A, Combes MC, Etienne H, Lashermes P. 2011. Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. *BMC Genomics* 12(No. 240).

Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SGE. 2011. The IMGT/HLA database. *Nucleic Acids Res.* 39:D1171–D1176.

Sato A, Figueroa F, O'Huigin C, Reznick DN, Klein J. 1996. Identification of major histocompatibility complex genes in the guppy, *Poecilia reticulata*. *Immunogenetics* 43:38–49.

Sato A, Tichy H, Grant PR, Grant BR, Sato T, O'Huigin C. 2011. Spectrum of MHC class II variability in Darwin finches and their close relatives. *Mol Biol Evol.* 28:1943–1956.

Schelkle B, Paladini G, Shinn AP, King S, Johnson M, van Oosterhout C, Mohammed RS, Cable J. 2011. *Ieredactylus rivuli* gen. et sp. nov. (Monogenea, Gyrodactylidae) from *Rivulus hartii* (Cyprinodontiformes, Rivulidae) in Trinidad. *Acta Parasitol.* 56:360–370.

Shen XY, Yang GP, Liao MJ. 2006. Development of 51 genomic microsatellite DNA markers of guppy (*Poecilia reticulata*) and their application in closely related species. *Mol Ecol Notes.* 7: 302–306.

Shiina T, Ota M, Shimizu S, et al. (26 co-authors). 2006. Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* 173:1555–1570.

Smith JM. 1992. Analysing the mosaic structure of genes. *J Mol Evol.* 34:126–129.

Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc R Soc Lond B Biol Sci.* 277:979–988.

Spurgin LG, van Oosterhout C, Illera JC, Bridgett S, Gharbi K, Emerson BC, Richardson DS. 2011. Gene conversion rapidly generates major histocompatibility complex diversity in recently founded bird populations. *Mol Ecol.* 20:5213–5225.

Stone JL. 2004. Sheltered load associated with S-alleles in *Solanum carolinense*. *Heredity* 92:335–342.

Takuno S, Nishio T, Sattay Y, Innany H. 2008. Preservation of a pseudogene by gene conversion and diversifying selection. *Genetics* 180:517–531.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.

Truett GE, Heeger P, Mynatt RL, Truett AA, Walker JA, Warman ML. 2000. Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). *Biotechniques* 29:52–54.

Uyenoyama MK. 2003. Genealogy-dependent variation in viability among self-incompatibility genotypes. *Theor Popul Biol.* 63: 281–293.

Uyenoyama MK. 2005. Evolution under tight linkage to mating type. *New Phytol.* 165:63–70.

van Oosterhout C. 2009. A new theory of MHC evolution: beyond selection on the immune genes. *Proc R Soc Lond B Biol Sci.* 276:657–665.

van Oosterhout C, Joyce DA, Cummings SM. 2006. Evolution of MHC class IIB in the genome of wild and ornamental guppies, *Poecilia reticulata*. *Heredity* 97:111–118.

van Oosterhout C, Joyce DA, Cummings SM, Blais J, Barson NJ, Ramnarine IW, Mohammed RS, Persad N, Cable J. 2006. Balancing selection, random genetic drift, and genetic variation at the major histocompatibility complex in two wild populations of guppies (*Poecilia reticulata*). *Evolution* 60:2562–2574.

Watanabe T, Yoshida M, Nakajima M, Taniguchi N. 2003. Isolation and characterization of 43 microsatellite DNA markers for guppy (*Poecilia reticulata*). *Mol Ecol Notes.* 3:487–490.

Willing EM, Bentzen P, van Oosterhout C, Hoffmann M, Cable J, Breden F, Weigel D, Dreyer C. 2010. Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Mol Ecol.* 19:968–984.

Wright SI, Gaut BS. 2005. Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol.* 22:506–519.

Zagalska-Neubauer M, Babik W, Stuglik M, Gustafsson L, Cichon M, Radwan J. 2010. 454 sequencing reveals extreme complexity of the class II major histocompatibility complex in the collared flycatcher. *BMC Evol Biol.* 10(No. 395).