JOURNAL OF Evolutionary Biology
European Society for Evolutionary Biology
doi: 10.1111/jeb.12476

Selective pressures on MHC class II genes in the guppy (*Poecilia reticulata*) as inferred by hierarchical analysis of population structure

M. HERDEGEN*, W. BABIK* & J. RADWAN†

*Institute of Environmental Sciences, Jagiellonian University, Kraków, Poland
†Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

Keywords:

balancing selection;
frequency-dependent selection;
local adaptation;
major histocompatibility complex.

Abstract

Genes of the major histocompatibility complex, which are the most polymorphic of all vertebrate genes, are a pre-eminent system for the study of selective pressures that arise from host–pathogen interactions. Balancing selection capable of maintaining high polymorphism should lead to the homogenization of MHC allele frequencies among populations, but there is some evidence to suggest that diversifying selection also operates on the MHC. However, the pattern of population structure observed at MHC loci is likely to depend on the spatial and/or temporal scale examined. Here, we investigated selection acting on MHC genes at different geographic scales using Venezuelan guppy populations inhabiting four regions. We found a significant correlation between MHC and microsatellite allelic richness across populations, which suggests the role of genetic drift in shaping MHC diversity. However, compared to microsatellites, more MHC variation was explained by differences between populations within larger geographic regions and less by the differences between the regions. Furthermore, among proximate populations, variation in MHC allele frequencies was significantly higher compared to microsatellites, indicating that selection acting on MHC may increase population structure at small spatial scales. However, in populations that have significantly diverged at neutral markers, the population-genetic signature of diversifying selection may be eradicated in the long term by that of balancing selection, which acts to preserve rare alleles and thus maintain a common pool of MHC alleles.

Introduction

The major histocompatibility complex (MHC) is present in all jawed vertebrates and encompasses a group of genes involved in the immune response. Classical MHC genes encode proteins that bind to antigens derived from pathogens and present them to the immune system. The MHC-antigen complex is then recognized by T cells, which elicit a highly specific response against the pathogen (Janeway *et al.*, 2004). MHC class I genes typically present antigens derived from intracellular parasites, whereas MHC class II present those from extracellular ones. Together, MHC class I and class II genes are the most polymorphic genes in vertebrates (Garrigan & Hedrick, 2003).

It is believed that the main force driving the evolution of MHC genes is pressure from parasites, which are themselves under strong selection to evade detection by the host immune system. Two types of parasite-driven balancing selection have been proposed as the mechanisms maintaining the extremely high levels of MHC polymorphism (reviewed in Spurgin & Richardson, 2010). The negative frequency-dependent selection hypothesis (also called rare-allele advantage; Clarke & Kirby, 1966; Snell, 1968; Borghans *et al.*, 2004) proposes that in fast-evolving pathogens, mutations that allow pathogens to evade recognition by the most common host MHC genotypes will spread quickly. Thus, hosts that possess rare or novel MHC alleles will have an advantage in detecting pathogens. The second type of balancing selection, heterozygote advantage (i.e.

Correspondence: Magdalena Herdegen, Institute of Environmental Sciences, Jagiellonian University, Gronostajowa 7, 30-387 Krakow, Poland.
Tel.: +48 12 664 51 54; fax: +48 12 664 69 12;
e-mail: magdalena.herdegen@uj.edu.pl

'overdominant selection'; Doherty & Zinkernagel, 1975; Takahata & Nei, 1990; Hughes & Nei, 1992), proposes that MHC heterozygous individuals are able to present more types of antigens to their immune system and, as a consequence, are able to cope with a wider range of pathogens than homozygotes. Both types of selection are expected to favour the establishment, but not the fixation, of new alleles, which may explain both the molecular signatures of positive selection typically found in MHC genes and also the high genetic variation in these sequences (Hughes & Nei, 1988, 1989; Bernatchez & Landry, 2003; Garrigan & Hedrick, 2003).

Apart from molecular evidence, support for the role of parasites as drivers of MHC diversity comes from association of MHC genotypes with infections and comparisons of actual MHC diversity with that expected under neutrality (reviewed in Bernatchez & Landry, 2003; Garrigan & Hedrick, 2003; Sommer, 2005; Piertney & Oliver, 2006; Spurgin & Richardson, 2010). Whereas correlations between MHC heterozygosity and resistance to parasites have been documented in some studies (e.g. Arkush *et al.*, 2002; Froeschke & Sommer, 2005; Oliver *et al.*, 2009), associations between resistance and particular MHC alleles appear to be more common (e.g. Paterson *et al.*, 1998; Meyer-Lucht & Sommer, 2005; Schad *et al.*, 2005; Tollenaere *et al.*, 2008; Fraser & Neff, 2010; Biedrzycka *et al.*, 2011). In addition to selection from parasites, balancing selection on MHC may result from preferences for MHC dissimilar mates (Hedrick, 1992) and from accumulation of recessive deleterious mutations in the MHC region, which favours MHC heterozygosity (van Oosterhout 2009).

Compared to a neutral scenario with no selective forces, balancing selection is thought to result in more even allele frequencies (Garrigan & Hedrick, 2003), but a recent simulation has shown that this assumption is valid only in the case of heterozygote advantage, and not for frequency dependence (Ejsmond *et al.*, 2010). However, it can be expected that balancing selection should result in lower between-population differentiation in MHC genes compared to that found at neutral markers. This is because balancing selection should prevent the loss of rare variants due to drift, as exemplified by the long persistence time of MHC allelic lineages (Figueroa *et al.*, 1988), and also because balancing selection increases the effective rate of gene flow among populations by promoting initially rare immigrant alleles (Schierup *et al.*, 2000; Castric *et al.*, 2008; Nadachowska-Brzyska *et al.*, 2012). A clear pattern of reduced differentiation between populations and species has been observed in plant self-incompatibility genes, where the significance of the rare-allele advantage as the dominant mechanism of balancing selection has been well established (Castric & Vekemans, 2004; Glémin *et al.*, 2005; Castric *et al.*, 2008; Ruggiero *et al.*, 2008). Thus, it can be predicted that under balancing selection, MHC genes should flow between populations and/or species more readily than neutral genes, leading to weaker population structure in MHC. Alternatively, if adaptations to local pathogenic fauna result in diverse MHC allele pools being maintained in populations or species, these genes will display more pronounced population structure than neutral genes. Additionally, temporal changes in parasite fauna and their resulting effects on MHC allele frequencies are likely to be uncoupled across populations, which could be another mechanism leading to interpopulation differentiation.

Empirical data on wild populations do not provide concordant results on this subject. Numerous studies have found a similar degree of population structure in MHC and neutral genes (Boyce *et al.*, 1997; Gutierrez-Espeleta *et al.*, 2001; Hedrick *et al.*, 2001; Biedrzycka & Radwan, 2008). Others have reported that MHC genes are more highly structured among populations than neutral markers, although some of these results are inconsistent across different populations of the same species. In studies of salmonids, for example, Landry & Bernatchez (2001) and Miller *et al.* (2001) observed more clearly delineated population structure at MHC genes compared to neutral loci, but in less than half of the studied populations. Aguilar & Garza (2006) also found more structure in the MHC than in neutral genes, but only in some of the investigated populations of steelhead trout (*Oncorhynchus mykiss*). MHC alleles also provided clearer evidence of large-scale population structure in a migratory bird, the great snipe (*Gallinago media*), than was obtained from microsatellite data (Ekblom *et al.*, 2007); the same pattern was reported in breeding colonies of grey seals (*Halichoerus grypus*; Cammen *et al.*, 2011) and in house sparrows (*Passer domesticus*, MHC I alleles; Loiseau *et al.*, 2009). However, several studies have found weaker population structure at MHC genes compared to neutral markers: a comparison of two Trinidadian guppy (*Poecilia reticulata*) populations revealed that MHC genes provided less evidence of population structure than microsatellite data did (van Oosterhout *et al.*, 2006), and likewise, two populations of giant jumping rats (*Hypogeomys antimena*) were more similar in their MHC alleles than in the mitochondrial control region (Sommer, 2003).

The reasons for these discordant results are not clear. Higher differentiation at MHC loci compared to neutral loci is usually interpreted as evidence for adaptation to local environments and local parasite assemblages (Bernatchez & Landry, 2003; Eizaguirre *et al.*, 2012), but it is not clear how to reconcile this interpretation with long persistence of MHC alleles (Figueroa *et al.*, 1988), which requires some form of balancing rather than directional selection. A recent study demonstrated, via computer simulations, that metapopulation structure can cause high turnover of alleles at loci under overdominant selection in sink populations (McMullan & van Oosterhout, 2012). This temporal structure may

appear as spatial structure if different sink populations are sampled at different times. Here, we propose that whether MHC genes show higher or lower spatial structure than neutral markers may also depend on the spatial or temporal scales involved.

We hypothesize that in neighbouring populations experiencing considerable gene flow which homogenizes allele frequencies at neutral loci, selection on the MHC may lead to increased differentiation in this gene region due to local adaptation, frequency fluctuations caused by frequency-dependent selection or metapopulation dynamics. However, among populations that have diverged at neutral markers due to some barrier to gene flow (e.g. geographic barriers, spatial distance or reproductive barriers), the situation will be reversed because MHC genes may flow more easily than other parts of the genome due to frequency-dependent selection. Indeed, two recent studies have demonstrated that MHC alleles can introgress into recently diverged species more easily than neutral markers can (Abi-Rached *et al.*, 2011; Nadachowska-Brzyska *et al.*, 2012). Furthermore, balancing selection will retain MHC alleles in populations for longer than neutral markers (Takahata & Nei, 1990).

To test our hypothesis, we compared population structure in MHC genes to that found in neutrally evolving microsatellites across different spatial scales in the guppy, *P. reticulata*, a small tropical ovoviviparous fish originally native to northern regions of South America as well as to the islands of Trinidad and Tobago. Signatures of balancing selection acting on MHC sequences in guppies were reported (Lighten *et al.*, 2014), and associations between MHC supertypes and infection with gyrodactylid parasites known to affect guppy survival (van Oosterhout *et al.*, 2007) have been documented (Fraser & Neff, 2010). Guppy populations inhabiting different streams are partially isolated from each other, but non-negligible gene flow between populations also occurs (Crispo *et al.*, 2006). In this study, we take advantage of the hierarchical structure of Venezuelan guppy populations to investigate two levels of population structure: well-defined geographic regions and populations within those regions.

Materials and methods

Samples

A total of 440 guppies (males and females) were collected in continental Venezuela and on Margarita Island in 2011 (Fig. 1). Eighteen localities in four geographic regions were sampled. Of the three sampled regions in continental Venezuela, two were located in adjacent drainages: Cariaco (four localities) and San Juan (ten localities). The populations in the Cariaco drainage have recently been proposed to represent a case of incipient speciation driven by divergent sexual selection (Alexander & Breden, 2004; Poeser *et al.*, 2005; Alexander *et al.*, 2006). These populations are geographically close to the common guppy populations found in the San Juan drainage but are separated from them by a mountain range (Fig. 1), and a recent study (Herdegen *et al.*, 2014) found substantial differentiation between the Cariaco and San Juan populations using both mtDNA and microsatellite data. The third sampled region (hereafter referred to as West) comprised populations from central and western Venezuela (two localities): one near the city of Caracas and one near the city of Guanare. The fourth sampled region was Margarita Island (two localities; Fig. 1, Table S1).

The fish were captured with dip nets, euthanized in a 0.03% solution of MS-222 (tricaine methanosulfonate)

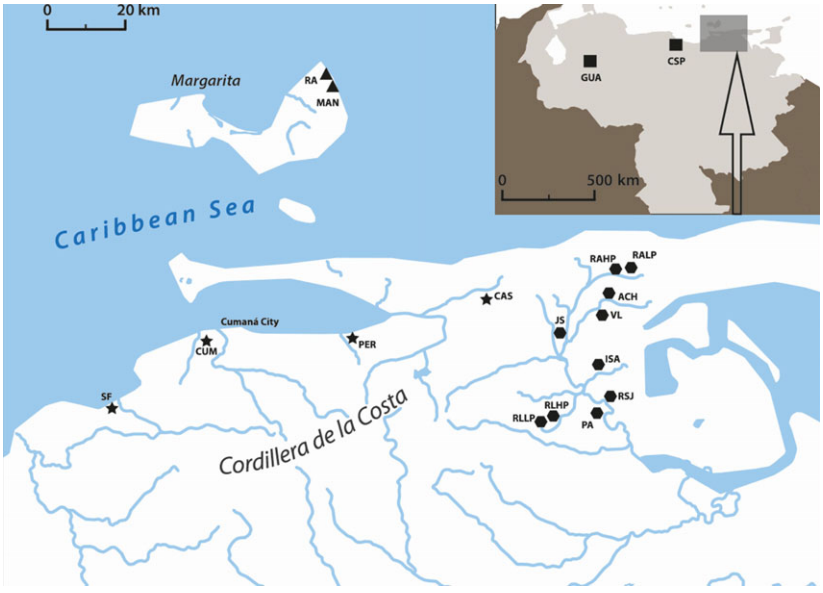

Fig. 1 Map of northeastern Venezuela, including Margarita Island; sampling locations are marked with stars (Cariaco drainage), hexagons (San Juan drainage), rectangles (West) and triangles (Margarita); population names are abbreviated as listed in Table S1.

and stored in 95% alcohol until molecular analyses could be performed.

Microsatellite genotyping

Total DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega, Fitchburg, WI, USA). All individuals were screened for variation at fifteen previously described microsatellite loci: AG1 and AG9 (Olendorf *et al.*, 2004); G75, G183, G255 and G325 (Shen *et al.*, 2007); Pret-52 and Pret-48 (Watanabe *et al.*, 2003); TACA033 (GenBank Acc. No. AY258896); CA061 (GenBank Acc. No. AY258683); TAGA033 (GenBank Acc. No. 258667); Pre15 (GenBank Acc. No. AY830943); and Pre26 (GenBank Acc. No. AY830946). The genotypes at these loci of all individuals from Cariaco and San Juan regions have already been published by Herdegen *et al.* (2014), whereas the amplification and genotyping of the remaining samples was performed here following their methods.

MHC primer design and amplification

The existing primers available to amplify MHC class II sequences in guppies show multiple mismatches when aligned with the homologous sequences of other Poecillids (*Xiphophorus* available in GenBank; *Poecilia lattipina* – M. Herdegen, unpublished data). This indicates that these primers may not amplify all allelic lineages, and we decided to design new primers. First, we extracted RNA from spleens of 20 individuals of *P. reticulata* obtained from our stock population, which had originally been sampled from the Ticaragua river in Trinidad. After conversion to cDNA, we amplified exons 1–4 of MHC II from each sample. The primers for these reactions were designed using sequences of *Poecilia* and *Xiphophorus* available in GenBank (Table S2). We then cloned the PCR products and Sanger-sequenced 86 clones. Based on these sequences, and on sequences available from other Poeciliids, we designed new primers to amplify a 200-bp fragment of the second exon of MHC II (Table S2), which codes for antigen binding groove of the MHC molecule (Janeway *et al.*, 2004). The amplicons were genotyped via sequencing on an Ion Personal Genome Machine (PGM; Life Technologies, Carlsbad, CA, USA).

To be able to assign sequencing reads to their individual guppy of origin, a unique 6-bp tag was added to each of the specific primers. A combination of 12 tagged forward primers and eight tagged reverse primers (we used a mixture of four reverse primers, as the high variation in this genome region precluded the use of only one primer version and we preferred this solution to the highly, 16-folds, degenerated primer, see Table S2) was used to amplify a set of 96 individuals. The 10-μL PCR mixture included 5 μL Multiplex Master Mix (Qiagen, Hilden, Germany), 0.2–0.4 μM primer mixture, containing one forward and four reverse primers, and 20–100 ng of genomic DNA. The reaction conditions were a 15-min denaturation step at 95 °C; 35 cycles of 30 s at 94 °C, 1 min at 53 °C, and 1 min at 72 °C; and a final extension step of 10 min at 72 °C. Following amplification, each set of 96 individuals was pooled approximately equimolarly. Five uniquely barcoded libraries were created from amplicon pools and sequenced using a 318 chip on the PGM at the Functional Genomics Center, Zurich. For 14 individuals, two independent amplicons were obtained and sequenced in separate pools to estimate genotyping error.

To verify whether PGM sequencing captured the majority of MHC II variants present in the guppy genome, we also performed MiSeq (Illumina, San Diego, CA, USA) paired-end sequencing using internal primers that were designed based on the sequences of a large number of alleles obtained earlier from PGM sequencing (see above; Table S2; note that these primers were optimized for Venezuelan guppy populations only). MiSeq sequencing was performed for the same subset of 14 individuals that had been used to estimate genotyping error. As the internal primers were located in conserved portions of the exon and identified from a very large number of MHC II alleles, this sequencing effort could potentially yield additional MHC variants, but concordance between the two genotyping approaches would increase our confidence in the accuracy of genotyping. The primers were barcoded and contained a fragment of Illumina forward or reverse Nextera adaptors at the ends. In the first PCR round, we amplified a 131-bp fragment of the second MHC II exon. The 10-μL PCR mixture included 5 μL Multiplex Master Mix (Qiagen); 0.2–0.4 μM primer mixture, containing three forward and three reverse barcoded primers, and 20–100 ng of genomic DNA. The reaction conditions were a 15-min denaturation step at 95 °C; 33 cycles of 30 s at 94 °C, 1 min at 55 °C, and 1 min at 72 °C; and 10 min of final extension at 72 °C. Amplicons were pooled equimolarly into two pools and purified with a DNA Clean & Concentrator kit (Zymo Research, Irvine, CA, USA). During the second PCR, the remaining parts of the Illumina adaptors were added. The 10-μL PCR mixture included 5 μL Master Mix, 0.2–0.4 μM each of forward and reverse primers and 1 μL of the pooled PCR product diluted 20 times. The reaction conditions were a 15-min denaturation step at 95 °C; 12 cycles of 30 s at 94 °C, 1 min at 50 °C, and 1 min at 72 °C; and 10 min of final extension at 72 °C. The product was again purified as before. The two pools were analysed in two separate sequencing runs.

MHC genotyping

The MHC sequences were extracted from the PGM output and assigned to individuals, then variants were

identified, and individual alignments were prepared with JMHC software (Stuglik *et al.*, 2011). Distinguishing true alleles from artefacts is an essential part of the genotyping process. Artefacts can be generated at various stages. First, during PCR, two kinds of artefacts may be created: point mutations due to *Taq* polymerase errors (Lenz & Becker, 2008; Cummings *et al.*, 2010) and chimeras (Longeri *et al.*, 2002; Lenz & Becker, 2008; Galan *et al.*, 2010), which are the result of recombination between different PCR products. Second, next-generation sequencing involves a relatively high frequency of error, including the creation of substitutions and indels (Huse *et al.*, 2007; reviewed in Shendure & Ji, 2008).

To decide on a method to detect and discard all types of artefacts, we first compared the repeatability of genotyping using two approaches: a procedure based on identifying artefacts that originate from true variants within the same amplicon, as described by Radwan *et al.* (2012), and the degree of change (DOC) approach proposed by Lighten *et al.* (2014), which relies on the assumption that all artefacts will be markedly less common than any true allele within an amplicon. The method of Radwan *et al.* (2012) proved superior (see Results) and was therefore used to genotype the whole dataset. First, all variants differing from the expected amplicon length were discarded, as these were likely to be artefacts which were created by insertions or deletions, the most common type of sequencing errors on the PGM platform. In our case, all variants that departed from the expected length of 200 bp were excluded from the analysis.

In the next step, which was performed at the level of the entire dataset, the threshold frequencies of sequence variants within amplicons were determined. Above these frequencies, all variants could safely be assumed to be true alleles, and below this threshold, most variants could be explained as artefacts (substitutions, chimeras) which arose from true variants within the same amplicon. In the 'grey zone' between the two thresholds, discrimination between true alleles and artefacts was performed on a case-by-case basis by comparing a given variant to those represented by more numerous reads within the same amplicon. To find the appropriate thresholds, the maximum per-amplicon frequency (MPAF) was first calculated for each sequence variant; MPAF is the highest per-amplicon frequency which a variant has reached among all amplicons (Radwan *et al.*, 2012). We then sorted the variants according to their MPAFs. Starting from the variants of 1% MPAF and going up, we checked whether each variant could be explained as an artefact, that is derived by a maximum of two substitutions from a more abundant variant present in the same amplicon or by recombination of two more abundant variants. Using this procedure, all variants with MPAF between 1% and 3% were rejected as artefacts, whereas all variants with MPAF > 12% were accepted as true alleles. The variants with MPAF between 3% and 12% were manually examined in the three amplicons in which they occurred most frequently. In this group, variants that were determined to be artefacts in all three amplicons were discarded, whereas the remaining variants were evaluated on a case-by-case basis using the criteria described above.

With the DOC approach, for each amplicon, X most abundant variants (X depending on the maximum expected number of loci) are sorted by the number of reads and the relative DOC in sequencing depth for the adjacent variants is calculated. The variant with the highest DOC is considered the last true allele (Lighten *et al.*, 2014).

To assess the accuracy and repeatability of the genotyping approaches used here, a subset of 14 individuals was resequenced, as described above. After performing the genotyping, we calculated the correlation coefficient for the number of alleles obtained from each PGM sequencing. We also calculated the percentage of alleles present in both PGM replicates with respect to the total number of alleles detected. Using only the alleles that were recovered in both PGM replicates, we inferred consensus genotypes and calculated the correlation coefficient of the number of alleles, as well as the repeatability of genotypes, between consensus PGM genotypes and those obtained from MiSeq sequencing. We used measures of mean genotype repeatability and mean correlation of allele number to estimate the quality of genotyping. We repeated these calculations for both allele calling methods (see above) and used these calculations to decide which of them to use.

Statistical analyses: microsatellites

Each locus was checked for deviations from Hardy–Weinberg equilibrium using GENEPOP 4.1.2 (Rousset, 2008). The presence of null alleles was tested and their frequencies estimated in FreeNA (Chapuis & Estoup, 2007) using the algorithm of Dempster *et al.* (1977). Each pair of loci was checked for linkage disequilibrium using an exact test implemented in GENEPOP. Because we conducted multiple tests on our data, we controlled for the false discovery rate (FDR; Benjamini & Hochberg, 1995) in both analyses. Allelic richness, the per-population number of alleles corrected for population size, was estimated for all populations using FSTAT v. 2.9.3 (Goudet, 2001).

Pairwise F_{ST} values between each pair of populations were calculated in FreeNA and were adjusted for the presence of null alleles using the excluding null alleles correction. A neighbour-joining tree was constructed from the matrix of pairwise F_{ST} values in POPTREE (Takezaki *et al.*, 2010) to visualize relationships between populations, and the robustness of the tree was tested with 1000 bootstrap replicates. An analysis of molecular vari-

ance was performed in ARLEQUIN v. 3.5 (Excoffier & Lischer, 2010) to assess genetic differentiation among our four defined regions and among populations within regions. The significance of AMOVA components was tested with 10 000 permutations. Additionally, the AMOVA was repeated for binary-encoded microsatellite data to facilitate direct comparison with binary-encoded MHC data.

To infer the most probable number of genetically differentiated clusters, we analysed the data in STRUCTURE 2.3.3 (Pritchard *et al.*, 2000), using the admixture model with uncorrelated allele frequencies and allowing for the presence of null alleles. The alpha value was set to 7 in order to achieve good mixing. The burn-in length was set to 100 000 and the post-burn-in MCMC was run for one million steps. We ran the analysis for K values ranging from 1 to 18 and performed ten runs for each K value. Structure Harvester (Earl & vonHoldt, 2012) was then used to calculate ΔK, a measure that estimates the number of clusters supported by the data (Evanno *et al.*, 2005).

Statistical analyses: MHC

Because we could not assign MHC alleles to loci, in all the analyses below they were treated as dominant markers with each allele encoded as a dominant biallelic locus. Allelic richness was calculated for all populations. Differences in the number of alleles per individual were analysed among populations using a hierarchical analysis of variance with population as the grouping variable nested in region, the random factor. The analysis was performed in STATISTICA v.10 (Statsoft, Tulsa, OK, USA). Pairwise F_{ST} values between all pairs of populations were calculated in ARLEQUIN, and their statistical significance was tested with 10 000 permutations. An analysis of molecular variance (AMOVA) with region as the higher hierarchical level was also performed in ARLEQUIN. The significance of AMOVA components was tested with 10 000 permutations.

Comparison between the markers

AMOVA analyses described above quantified proportions of variance in microsatellites and in MHC alleles explained by the levels of population and region, but did not allow a direct statistical comparison between the two types of markers. Therefore, we used the following approach based on the hierarchical genome scan method (Excoffier *et al.*, 2009) to compare between-population (F_{ST}) and between-region (F_{CT}) differentiation in MHC and microsatellite data. Unlike microsatellites, MHC loci are not independent and alleles cannot be ascribed to loci, so we could not apply the classic genome scan method to compare MHC and microsatellite diversity. Instead, the expected F_{ST} values for given values of heterozygosity were first simulated in ARLEQUIN for binary-encoded microsatellite loci using 200 000 permutations. Mean expected F_{ST} values were then calculated over heterozygosity intervals of 0.01. Based on this distribution, we calculated the difference between the observed and expected F_{ST}, divided by the standard deviation for a given heterozygosity interval; this was performed for all pairwise F_{ST} values between all pairs of populations for each MHC and binary-encoded microsatellite locus. The degree of deviation between observed and expected values that was found in MHC data was compared with that found in microsatellites using a t-test, but with degrees of freedom adjusted to reflect the fact that the data came from 15 microsatellite loci and all the MHC loci are linked (therefore, we conservatively considered the MHC to be a single locus). The test was performed in STATISTICA v. 10 (Statsoft). This same procedure was also repeated for F_{CT}.

To infer the most probable number of genetically differentiated clusters present in MHC data, we used STRUCTURE 2.3.3, with the same parameters as the microsatellite analysis (see above), and inferred the optimal number of clusters based on the results of a Structure Harvester analysis.

Results

Microsatellites

The number of alleles per locus ranged from 16 to 89 (mean 42.3), with genotyping error estimated at 4.4%. Of 265 tests for deviation from Hardy–Weinberg equilibrium, 95 (36%) were significant after FDR correction (Table S3). This was probably due to the presence of null alleles, which were detected in all populations at low-to-moderate frequencies ranging from 0.00 to 0.36 (Table S4). Allelic richness values are provided in Table S5. Significant linkage disequilibrium was detected in 101 (6%) of 1680 tests following FDR correction. There was no clear pattern of occurrence of these disequilibria across populations or loci, an observation that suggests they were generated by drift/admixture rather than by physical linkage between loci. We therefore retained all loci for further analyses.

Genetic differentiation between pairs of populations (F_{ST}) ranged from 0.02 (between two populations within the San Juan region) to 0.44 (between populations GUAN and PER from San Juan and Cariaco regions, respectively), with global $F_{ST} = 0.22$ (all F_{ST} values were significant, Table S6). The tree based on the pairwise F_{ST} matrix generally reflects the division of sampled populations into four regions (Fig. 2). However, in the STRUCTURE analysis, both Evanno's delta K and the logarithm of the probability of the data supported $K = 6$ as the optimal number of genetically differentiated clusters. Nevertheless, the populations from each region generally grouped together and were separated from those of other regions (Fig. 3b), with

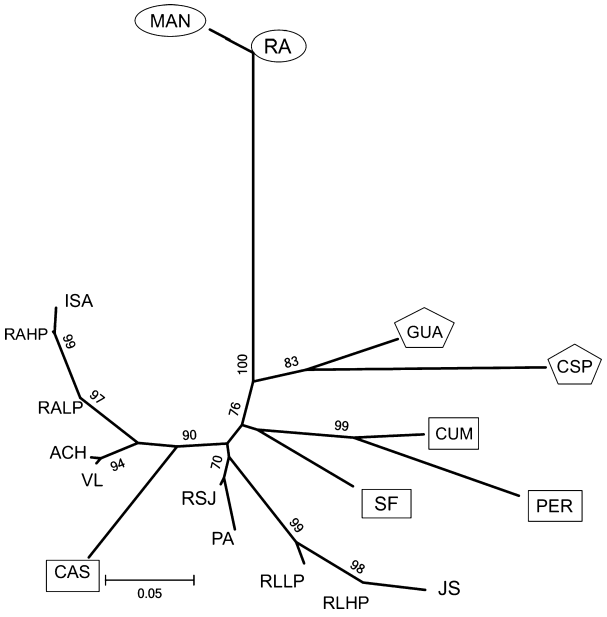

Fig. 2 Neighbour-joining tree constructed from microsatellite data based on the matrix of pairwise F_{ST} values.

the exception of CAS, a Cariaco drainage population that is located near the border with San Juan and grouped with the populations from this region. We observed further subdivision within San Juan, with populations from this region clustering into three groups.

The AMOVA revealed that both region and population explained comparable amounts of variance (14.09% and 15.24%, respectively, 11.75% and 12.88% after excluding the three loci (G183, G325 and Pret-52) exhibiting departures from Hardy–Weinberg expectations; for binary-encoded data, the values were 12.1% and 17.57%, respectively; $P < 0.0001$ in all cases, Table 1).

MHC II genotyping

Personal Genome Machine sequencing yielded 2.3 mln reads; 0.6 mln reads which could not be assigned to amplicons and 0.3 mln reads departing from the expected length of 200 bp were discarded, as was one amplicon with coverage < 500 reads. The remaining 1.4 mln reads were distributed among 440 individuals (453 amplicons: 440 samples plus 13 replicates; the 14th repeated amplicon was discarded due to low coverage). The average per-amplicon coverage was 3 070 (SD ± 1275) reads, from which true alleles constituted on average 62%.

Comparisons performed on the 14 individuals for which two independent amplicons were analysed (see Materials and methods) allowed us to estimate the repeatability of sequencing methods and to choose the most accurate allele calling procedure. For the replicates analysed in two separate PGM runs, the correlation coefficient for the number of alleles was 0.788 ($P < 0.05$), whereas this value was 0.685 ($P < 0.05$) for a comparison of the consensus PGM genotypes with those obtained from MiSeq. In examining the alleles obtained in the two PGM runs, the mean repeatability of alleles was 87.07%; for MiSeq vs. consensus PGM runs, it was 82.77%. Comparison of the genotyping results obtained from PGM with those from MiSeq revealed that the external primers used for PGM genotyping failed to capture only one variant which was amplified with internal primers and detected by the MiSeq sequencing.

When we applied the alternative DOC method (Lighten *et al.*, 2014; see Methods), the correlation coefficients of the two PGM runs dropped to 0.572 ($P < 0.05$). The average repeatability of identified alleles was likewise lower, at 75.87% for the two-PGM-run comparison. An examination of individual cases in the subset of 14 genotypes allowed us to identify the reasons for the discrepancies between the two allele calling methods. Most of the inconsistencies were due to

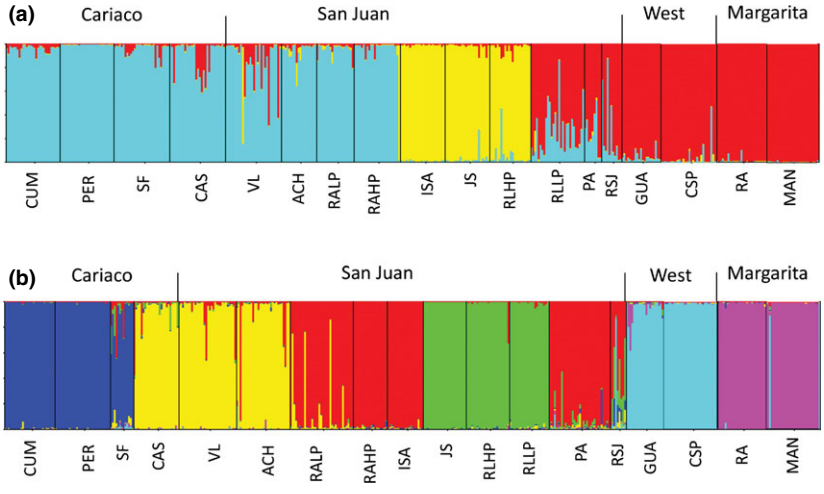

Fig. 3 STRUCTURE results based on (a) MHC ($K = 3$) and (b) microsatellite ($K = 6$) data, using the optimal number of genetically differentiated clusters.

Table 1 Analysis of molecular variance; variation explained is divided between three levels of population structure ($P < 0.0001$ in all cases except for microsatellites 'within populations' where $P < 0.001$).

	Percentage of variation explained	
Source of variation	Microsatellites	MHC
Among regions	14.09	9.28
Among populations within regions	15.24	19.04
Within populations	70.67	71.68

underestimation of the number of alleles by the DOC method when compared to the method of Radwan *et al.* (2012). In those cases, we observed more than one high DOC value: one between very rare and medium-frequency variants, probably representing the border between artefacts and true alleles, and the other between medium-frequency and high-frequency variants. This latter apparently did not separate true alleles from artefacts, but appeared to result from lower efficiency of amplification of the medium-frequency variants. As equal amplification of all variant copies present in the individual is one of the most important assumptions on which the DOC method relies (Lighten *et al.*, 2014), this approach was not appropriate for analysis of our data. All the above analyses demonstrated the higher reliability of the method of Radwan *et al.* (2012), which takes into account unequal amplification of variants; as a result, we decided to apply this method to genotype our whole set of MHC data.

MHC diversity

The total number of MHC II alleles identified in the 440 analysed individuals was 380 (GenBank Acc. No. KM041310–KM041689), with the per-amplicon number of alleles ranging from 1 to 8 (average 4.25). Among the 34 alleles with the highest overall frequency (appearing in at least 10% of samples), 1 was present in all four regions, six in three regions, and ten in two regions. Most remaining alleles were present only in a single population (287) or in a single individual (180; Table S7). However, as they occurred at low frequencies, they contributed little to the overall pattern of MHC diversification. Regions differed significantly in the average number of alleles per individual ($F_{3,419} = 5.06$, $P = 0.014$; Fig. 4); this was true even when per-population allelic richness (Table S5) was used as a covariate to test whether the number of alleles per individual could be explained by differences in the number of alleles segregating in populations ($F_{2,429} = 9.26$, $P < 0.001$). Genetic differentiation among populations (F_{ST}) ranged from 0.03 between two populations from the San Juan region to 0.54 between populations JS and MANG from the San Juan and Margarita regions, respectively (Table S8). All F_{ST} values were significant.

In the STRUCTURE analysis, both Evanno's delta K and the logarithm of the probability of the data supported $K = 3$ as the optimal number of genetically differentiated clusters (Fig. 3a). The clusters showed very little correspondence to the geographic regions: one cluster consisted of populations from both San Juan and Cariaco, another cluster was formed by three other San Juan populations, and the third cluster contained Margarita, West, and still another group of populations from San Juan. In contrast to results obtained from microsatellite data, AMOVA indicated that population explained almost twice as much (19.04%, $P < 0.0001$) of the total MHC variation as region did (9.28%, $P < 0.0014$). This result held when we grouped the populations into the three clusters suggested by STRUCTURE (within clusters 19.07% and between clusters 9.69%). Thus, regardless of whether the highest hierarchical level was based on a geographic or a genetic criterion, it explained a much lower fraction of MHC variation than the lower, among-population level did.

Comparison between MHC and microsatellites

Allelic richness at MHC and at microsatellite loci were highly correlated ($r = 0.70$, $P = 0.001$). The deviation of observed from expected F_{ST} values in microsatellites was significantly lower than in the MHC (t-test for independent samples: $t = 5.02$, $P < 0.001$, d.f. = 14). However, there was no difference in the deviations of observed from expected F_{CT} values between microsatellite and MHC data.

Discussion

Despite the role of MHC genes as a model of positive selection at the molecular level, the details of the

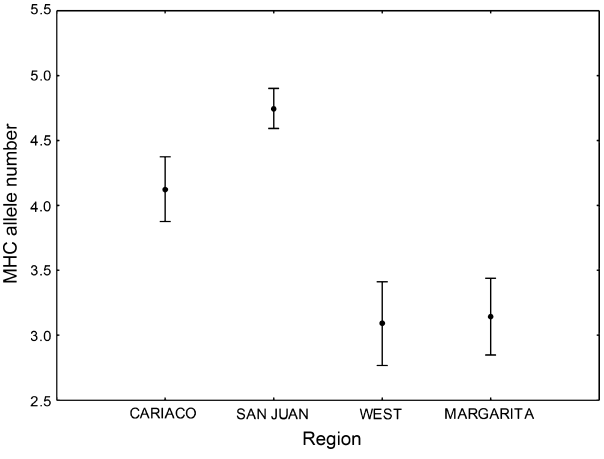

Fig. 4 Average number of MHC alleles per individual in each region; error bars represent 95% confidence interval.

selective mechanisms acting on them are not well understood (Spurgin & Richardson, 2010). Whereas high polymorphism strongly suggests the influence of balancing selection, strong between-population differentiation at MHC genes that has been observed in some systems (Landry & Bernatchez, 2001; Miller *et al.*, 2001; Aguilar & Garza, 2006; Cammen *et al.*, 2011) suggests that diversifying selection is also acting on the MHC. These two forces may or may not be compatible and may interact with each other to produce patterns of diversity and population structure that are difficult to interpret. For example, diversifying selection should result in highly structured populations, but this outcome conflicts with that of the balancing selection proposed by the heterozygote advantage hypothesis, which should homogenize allele frequencies within and between populations. However, balancing selection coupled with metapopulation dynamics may in fact increase temporal variation in MHC allele frequencies in bottlenecked sink populations (McMullan & van Oosterhout, 2012). This may result in increased spatial structure if temporal fluctuations are asynchronous across populations. To further complicate the picture, frequency-dependent balancing selection may skew allele frequencies compared to neutrality (Ejsmond *et al.*, 2010), and given that frequency-dependent selection is highly unlikely to be synchronized between populations, it is likely that this would amplify differences between populations. In the long run, however, frequency-dependent selection is expected to maintain MHC polymorphism (Borghans *et al.*, 2004). Therefore, for populations which have diverged substantially at neutral markers, MHC alleles should display relatively weaker population structure as they are likely to be maintained in the regional pool for a long time, while fluctuating in frequency within populations. Our results from an analysis of molecular variance and STRUCTURE are largely consistent with the above scenario, highlighting the role of balancing selection in shaping MHC allele frequencies among guppy populations.

Whereas for microsatellites, the fraction of total variation explained by within-region, among-population differences was comparable to that explained by differences among regions, for MHC alleles, the amount of variance explained by region was only half of that explained by interpopulation differences. The difference between the two types of markers was statistically supported by the analysis of observed vs. expected F_{ST} values which showed that indeed at the interpopulation level, population structure was more pronounced for MHC genes than for microsatellites, a finding which indicates the influence of diversifying and/or frequency-dependent selection. Recent simulations by McMullan & van Oosterhout (2012) have shown that MHC genes, just because they are duplicated, have effective rates of gene flow that are three times higher than neutral markers. In this context, our result on the within-region between-population level of divergence is conservative, in that we observed higher divergence despite the higher expected effective MHC gene flow due.

Our findings are concordant with the results of a study of Canadian sockeye salmon (Miller *et al.*, 2001) and another of Atlantic salmon (Landry & Bernatchez, 2001). Both found evidence for stronger population structure at MHC loci than at microsatellites on a small, but not on a larger, geographic scale. Interestingly, our results contrast with those of two studies which examined the same species we did. van Oosterhout *et al.* (2006) found less differentiation at MHC genes than in microsatellites between guppy populations in Trinidad. However, this study was based on only one pair of populations located in the same river but in different habitats. Also, using guppy populations in Trinidad, Fraser *et al.* (2010) performed a similar analysis to ours, and they observed lower F_{ST} in MHC than in microsatellites within drainages (no assessment was carried out between drainages). However, the discrepancy between this study and ours could be explained by differences in the sampling scheme. Among our 18 sampling localities, there were only two pairs within the same river. In contrast, Fraser *et al.* (2010) collected specimens from two locations in each of the rivers they sampled, which could have resulted in reduced between-population divergence.

On the regional scale, no significant differences were observed between microsatellites and MHC in their deviations from the expected F_{CT} values, although this deviation was slightly lower for MHC data, a result consistent with the lower proportion of MHC genetic variance that was explained by region in AMOVA. Similarly, STRUCTURE analysis of MHC data showed little differentiation between regions: two of three major supported clusters contained populations from different regions, a result that contrasts with microsatellite data, which mostly grouped individuals by region (Fig. 3). Thus, both AMOVA and STRUCTURE support the idea that MHC genes are more differentiated among populations than among regions. Together, our findings suggest that at a larger spatial scale, signatures of balancing selection prevail. This result also indicates no regional adaptations to different parasite communities, such as those observed among riverine and lake populations of sticklebacks, which are characterized by differentiated MHC allele pools (Eizaguirre *et al.*, 2012). This may be due to the ubiquitous presence of gyrodactylid flukes among guppy populations (Lyles, 1990; Martin and Johnsen 2007). Infection with these parasites appears to exert significant selective pressure on guppies (van Oosterhout *et al.*, 2007) and was shown to be associated with MHC genotypes (Fraser & Neff, 2010). The effects of balancing selection in our system were however not strong enough to result in a lower F_{CT} for MHC than

for microsatellites. This observation could result from the interaction of multiple selective forces, for example balancing and diversifying selection, which operate at the same time across populations and counteract the effects of each other (Spurgin & Richardson, 2010). Additionally, demographic factors may obscure the effects of adaptive processes. Indeed, the high correlation we found between allelic richness in MHC and microsatellite loci indicates that much of the observed structure can be attributed to the common history of these genes in the populations. In particular, genetic drift is expected to considerably influence population dynamics, as its effects are most pronounced in populations with relatively small effective sizes, as it is the case for guppies (Barson *et al*., 2009).

Interestingly, we observed significant differences in the number of MHC alleles per individual across the four regions sampled (see Fig. 4). The most pronounced difference was observed in the Cariaco and San Juan drainages vs. the Margarita and West regions, the latter two regions characterized by a lower number of alleles per individual. One possible explanation for this could be that performance of our primers differed between regions. However, as our MHC primers were designed using cDNA from Trinidadian guppies, which are more closely related to the Margarita and West populations than to Cariaco guppies (Alexander *et al*., 2006), and as our primers were shown to capture all but one variant obtained using a set of internal primers (see Results), this is unlikely. Indeed, we detected more (up to eight) MHC alleles per individual than has been previously reported (up to six, van Oosterhout *et al*., 2006; up to four, Fraser & Neff, 2010), indicating high quality of our primers.

Furthermore, the among-region differences in the number of alleles segregating in a population were not explained by the overall level of MHC polymorphism as measured by allelic richness. Hence, the observed differences in allelic number across the regions sampled here are likely to result from baseline differences in the number of MHC loci. Alternatively, the differences may result from different alleles sharing the same exon 2 sequence, as documented by Llaurens *et al.* (2012). In any case, populations differ in the number of variants in exon 2, the region coding for a part of MHC molecule responsible for foreign antigen recognition. Such differences may reflect differing pressures from parasites (Wegner *et al*., 2003; Eizaguirre *et al*., 2011), a possibility that awaits future investigation.

To conclude, our results indicate that selection operating on MHC genes in the guppy has led to increased divergence between populations within regions. This finding could be explained by two potential mechanisms: diversifying selection resulting from adaptation to local pathogen communities or balancing selection leading to changes in the frequencies of MHC genes among partially isolated populations. The latter, but not the former, mechanism should also lead to the long-term maintenance of MHC alleles in a population and counteract between-region differentiation at MHC loci in populations that have accumulated differences at neutral markers. In showing that on the broader, regional scale, MHC population structure tends to be less defined than that found in microsatellites, our results are more consistent with a scenario involving balancing selection. However, the signal of balancing selection was not strong, an observation that could be attributed to the confounding effects of genetic drift and/or local adaptation which may be operating across regions and counteracting its effects.

Acknowledgments

This work was funded by Polish National Science Center (project number 2011/03/N/NZ8/00017) to MH. We thank Sebastian Książkiewicz for producing the map and three anonymous reviewers for their helpful comments.

Conflict of interest

The authors declare that they have no competing interests.

References

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L. *et al.* 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334**: 89–94.

Aguilar, A. & Garza, J.C. 2006. A comparison of variability and population structure for major histocompatibility complex and microsatellite loci in California coastal steelhead (*Oncorhynchus mykiss* Walbaum). *Mol. Ecol.* **15**: 923–937.

Alexander, H.J. & Breden, F. 2004. Sexual isolation and extreme morphological divergence in the Cumana guppy: a possible case of incipient speciation. *J. Evol. Biol.* **17**: 1238–1254.

Alexander, H.J., Taylor, J.S., Wu, S.S.-T. & Breden, F. 2006. Parallel evolution and vicariance in the guppy (*Poecilia reticulata*) over multiple spatial and temporal scales. *Evolution* **60**: 2352–2369.

Arkush, K.D., Giese, A.R., Mendonca, H.L., McBride, A.M., Marty, G.D. & Hedrick, P.W. 2002. Resistance to three pathogens in the endangered winter-run chinook salmon (*Oncorhynchus tshawytscha*): effects of inbreeding and major histocompatibility complex genotypes. *Can. J. Fish. Aquat. Sci.* **59**: 966–975.

Barson, N.J., 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.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.* **57**: 289–300.

Bernatchez, L. & Landry, C. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J. Evol. Biol.* **16**: 363–377.

Biedrzycka, A. & Radwan, J. 2008. Population fragmentation and major histocompatibility complex variation in the spotted suslik, *Spermophilus suslicus*. *Mol. Ecol.* **17**: 4801–4811.

Biedrzycka, A., Kloch, A., Buczek, M. & Radwan, J. 2011. Major histocompatibility complex DRB genes and blood parasite loads in fragmented populations of the spotted suslik *Spermophilus suslicus*. *Mamm. Biol.* **76**: 672–677.

Borghans, J.A.M., Beltman, J.B. & De Boer, R.J. 2004. MHC polymorphism under host-pathogen coevolution. *Immunogenetics* **55**: 732–739.

Boyce, W.M., Hedrick, P.W., MuggliCockett, N.E., Kalinowski, S., Penedo, M.C.T. & Ramey, R.R. 1997. Genetic variation of major histocompatibility complex and microsatellite loci: a comparison in bighorn sheep. *Genetics* **145**: 421–433.

Cammen, K., Hoffmann, J.I., Knapp, L.A., Harwood, J. & Amos, W. 2011. Geographic variation of the major histocompatibility complex in Eastern Atlantic grey seals (*Halichoerus grypus*). *Mol. Biol.* **20**: 740–752.

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

Castric, V., Bechsgaard, J., Schierup, M.H. & Vekemans, X. 2008. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet.* **4**: 8.

Chapuis, M.P. & Estoup, A. 2007. Microsatellite null alleles and estimation of population differentiation. *Mol. Biol. Evol.* **24**: 621–631.

Clarke, B. & Kirby, D.R.S. 1966. Maintenance of histocompatibility polymorphisms. *Nature* **211**: 999–1000.

Crispo, E., Bentzen, P., Reznick, D. N., Kinnison, M. T. & Hendry, A. P. 2006. The relative influence of natural selection and geography on gene flow in guppies. *Mol. Ecol.* **15**: 49–62.

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

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**: 1–38.

Doherty, P.C. & Zinkernagel, R.M. 1975. Enhanced immunological surveillance in mice heterosytous at H-2 gene complex. *Nature* **256**: 50–52.

Earl, D.A. & vonHoldt, B.M. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**: 359–361.

Eizaguirre, C., Lenz, T.L., Sommerfeld, R.D., Harrod, C., Kalbe, M. & Milinski, M. 2011. Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. *Evol. Ecol.* **25**: 605–622.

Eizaguirre, C., Lenz, T.L., Kalbe, M. & Milinski, M. 2012. Divergent selection on locally adapted major histocompatibility complex immune genes experimentally proven in the field. *Ecol. Lett.* **15**: 723–731.

Ejsmond, M.J., Babik, W. & Radwan, J. 2010. MHC allele frequency distributions under parasite-driven selection: a simulation model. *BMC Evol. Biol.* **10**: 332.

Ekblom, R., Saether, S.A., Jacobsson, P., Fiske, P., Sahlman, T., Grahn, M. *et al.* 2007. Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Mol. Ecol.* **16**: 1439–1451.

Evanno, G., Regnaut, S. & Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**: 2611–2620.

Excoffier, L. & Lischer, H.E.L. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**: 564–567.

Excoffier, L., Hofer, T. & Foll, M. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285–298.

Figueroa, F., Gunther, E. & Klein, J. 1988. MHC polymorphism predating speciation. *Nature* **335**: 265–267.

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

Fraser, A., Ramnarine, I. & Neff, B. 2010. Selection at the MHC class IIB locus across guppy (*Poecilia reticulata*) populations. *Heredity* **104**: 155–167.

Froeschke, G. & Sommer, S. 2005. MHC class II DRB variability and parasite load in the striped mouse (*Rhabdomys pumilio*) in the southern Kalahari. *Mol. Biol. Evol.* **22**: 1254–1259.

Galan, M., Guivier, E., Caraux, G., Charbonnel, N. & Cosson, J.F. 2010. A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics* **11**: 296.

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

Glémin, S., Gaude, T., Guillemin, M.-L., Lourmas, M., Olivieri, I. & Mignot, A. 2005. Balancing selection in the wild: testing population genetics theory of self-incompatibility in the rare species *Brassica insularis*. *Genetics* **17**: 279–289.

Goudet, J. 2001. FSTAT, version 2.9.3, a program to estimate and test gene diversities and fixation indices. http://www2.unil.ch/popgen/softwares/fstat.htm.

Gutierrez-Espeleta, G.A., Hedrick, P.W., Kalinowski, S.T., Garrigan, D. & Boyce, W.M. 2001. Is the decline of desert bighorn sheep from infectious disease the result of low MHC variation? *Heredity* **86**: 439–450.

Hedrick, P.W. 1992. Female choice and variation in the major histocompatibility complex. *Genetics* **132**: 575–581.

Hedrick, P.W., Parker, K.M. & Lee, R.N. 2001. Using microsatellite and MHC variation to identify species, ESUs, and MUs in the endangered Sonoran topminnow. *Mol. Biol.* **10**: 1399–1412.

Herdegen, M., Alexander, H. J., Babik, W., Mavárez, J., Breden, F. & Radwan, J. 2014. Population structure of guppies in north-eastern Venezuela, the area of putative incipient speciation. *BMC Evol. Biol.* **14**: 28.

Hughes, A.L. & Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**: 167–170.

Hughes, A.L. & Nei, M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc. Natl. Acad. Sci. USA* **86**: 958–962.

Hughes, A.L. & Nei, M. 1992. Maintenance of MHC polymorphism. *Nature* **355**: 402–403.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L. & Mark Welch, D. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**: R143.

Janeway, C.A., Travers, P., Walport, D. & Shlomchik, M.J. 2004. *Immunobiology: The Immune System in Health and Disease*. Garland Publishing, New York, NY.

Landry, C. & Bernatchez, L. 2001. Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (*Salmo salar*). *Mol. Ecol.* **10**: 2525–2539.

Lenz, T.L. & Becker, S. 2008. Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci – implications for evolutionary analysis. *Gene* **427**: 117–123.

Lighten, J., Van Oosterhout, C., Paterson, I.G., McMullan, M. & Bentzen, P. 2014. Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol. Ecol. Resour.* **14**: 753–767.

Llaurens, V., McMullan, M. & van Oosterhout, C. 2012. Cryptic MHC polymorphism revealed but not explained by selection on the class IIB peptide-binding region. *Mol. Biol. Evol.* **29**: 1631–1644.

Loiseau, C., Richard, M., Garnier, S., Chastel, O., Julliard, R., Zoorob, R. *et al.* 2009. Diversifying selection on MHC class I in the house sparrow (*Passer domesticus*). *Mol. Ecol.* **18**: 1331–1340.

Longeri, M., Zanotti, M. & Damiani, G. 2002. Recombinant DRB sequences produced by mismatch repair of heteroduplexes during cloning in *Escherichia coli*. *Eur. J. Immunogenet.* **29**: 517–523.

Lyles, A. M. 1990. Genetic variation and susceptibility to parasites: *Poecilia reticulata* infected with *Gyrodactylus turnbulli*. PhD dissertation, Princeton University, Princeton, NJ.

Martin, C.H. & Johnsen, S. 2007. A field test of the Hamilton-Zuk hypothesis in the guppy Poecilia reticulata. *Behav. Ecol. Sociobiol.* **61**: 1897–1909.

McMullan, M. & van Oosterhout, C. 2012. Inference of selection based on temporal genetic differentiation in the study of highly polymorphic multigene families. *PLoS ONE* **7**: 8.

Meyer-Lucht, Y. & Sommer, S. 2005. MHC diversity and the association to nematode parasitism in the yellow-necked mouse (*Apodemus flavicollis*). *Mol. Ecol.* **14**: 2233–2243.

Miller, K.M., Kaukinen, K.H., Beacham, T.D. & Withler, R.E. 2001. Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. *Genetica* **111**: 237–257.

Nadachowska-Brzyska, K., Zielinski, P., Radwan, J. & Babik, W. 2012. Interspecific hybridization increases MHC class II diversity in two sister species of newts. *Mol. Ecol.* **21**: 887–906.

Olendorf, R., Reudi, B. & Hughes, K.A. 2004. Primers for 12 polymorphic microsatellite DNA loci from the guppy (*Poecilia reticulata*). *Mol. Ecol. Notes* **4**: 668–671.

Oliver, M.K., Telfer, S. & Piertney, S.B. 2009. Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (*Arvicola terrestris*). *Proc. R. Soc. Lond. B Biol. Sci.* **276**: 1119–1128.

van Oosterhout, C., Joyce, D.A., Cummings, S.M., Blais, J., Barson, N.J., Ramnarine, I.W. *et al.* 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.

van Oosterhout, C., Mohammed, R.S., Hansen, H., Archard, G.A., McMullan, M., Weese, D.J. *et al.* 2007. Selection by parasites in spate conditions in wild Trinidadian guppies (*Poecilia reticulata*). *Int. J. Parasitol.* **37**: 805–812.

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

Paterson, S., Wilson, K. & Pemberton, J.M. 1998. Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population (*Ovis aries* L.). *Proc. Natl. Acad. Sci. USA* **95**: 3714–3719.

Piertney, S.B. & Oliver, M.K. 2006. The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**: 7–21.

Poeser, F.N., Kempkes, M. & Isbrücker, I.J.H. 2005. Description of *Poecilia* (Acanthophacelus) *wingei* n. sp. from the París Peninsula, Venezuela, including notes on Acanthophacelus Eigenmann, 1907 and other subgenera of Poecilia Bloch and Schneider, 1801 (Teleostei, Cyprinodontiformes, Poeciliidae). *Contrib. Zool.* **74**: 97–115.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Radwan, J., Zagalska-Neubauer, M., Cichon, M., Sendecka, J., Kulma, K., Gustafsson, L. *et al.* 2012. MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Mol. Ecol.* **21**: 2469–2479.

Rousset, F. 2008. GENEPOP '007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol. Ecol. Resour.* **8**: 103–106.

Ruggiero, M., Jacquemin, B., Castric, V. & Vekemans, X. 2008. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in *Arabidopsis halleri*. *Genet. Res.* **90**: 37–46.

Schad, J., Ganzhorn, J.U. & Sommer, S. 2005. Parasite burden and constitution of major histocompatibility complex in the malagasy mouse lemur, *Microcebus murinus*. *Evolution* **59**: 439–450.

Schierup, M.H., Vekemans, X. & Charlesworth, D. 2000. The effect of hitchhiking on genes linked to a balanced polymorphism in a subdivided population. *Genet. Res.* **76**: 63–73.

Shen, X., Guanpin, Y. & Meijie, L. 2007. Development of 51 genomic microsatellite DNA markers of guppy (*Poecilia reticulata*) and their application in closely related species. *Mol. Ecol. Notes* **7**: 302–306.

Shendure, J. & Ji, H. 2008. Next-generation sequencing. *Nat. Biotechnol.* **26**: 1135–1145.

Snell, G. D. 1968. The H-2 locus of the mouse: observations and speculations concerning its comparative genetics and its polymorphism. *Folia Biol. (Prague)* **14**: 335–358.

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

Sommer, S. 2005. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.* **12**: 16.

Spurgin, L.G. & Richardson, D.S. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. Lond. B Biol. Sci.* **277**: 979–988.

Stuglik, M.T., Radwan, J. & Babik, W. 2011. jMHC: software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Mol. Ecol. Resour.* **11**: 739–742.

Takahata, N. & Nei, M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**: 967–978.

Takezaki, N., Nei, M. & Tamura, K. 2010. POPTREE2: software for constructing population trees from allele frequency data and computing other population statistics with Windows Interface. *Mol. Biol. Evol.* **27**: 747–752.

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

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.

Wegner, K.M., Reusch, T.B.H. & Kalbe, M. 2003. Multiple parasites are driving major histocompatibility complex polymorphism in the wild. *J. Evol. Biol.* **16**: 224–232.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Tables S1–S8.

Table S1 Sampling sites names and abbreviations, sampling regions, sample sizes and coordinates of the sampling locations.

Table S2 List of primers used for amplification of MHC II; 1–2 are primers amplifying exons 1–4 of the MHC II, 3–7 were used in PGMruns, 8–13 were used in the Illumina run.

Table S3 Hardy–Weinberg equilibrium tests for each locus in each population (GENEPOP); missing values are due to monomorphism of the locus in this particular population; *P*-values significant after FDR are in red.

Table S4 Null alleles frequencies for each locus in each population.

Table S5 Allelic richness for each population at MHC and microsatellite loci.

Table S6 Pairwise F_{ST} values for microsatellites for all pairs of populations; ENA correction applied; significance level = 0.05; all values significant.

Table S7 Proportion of individuals in each region bearing a given MHC variant; reported only variants present in at least 10% of all individuals.

Table S8 Pairwise F_{ST} values for all population pairs, binary encoded MHC data; significance level = 0.05; all values significant.

Data deposited at Dryad: doi:10.5061/dryad.h7m7t

Received 18 March 2014; revised 8 August 2014; accepted 11 August 2014