Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools

Jacob Roved[1*], Bengt Hansson[1], Martin Stervander[1,2], Dennis Hasselquist[1†], Helena Westerdahl[1†]

[1] Department of Biology, Molecular Ecology and Evolution Lab, Lund University, Ecology Building, 223 62 Lund, Sweden
[2] Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403-5289, USA
[*] Corresponding author: Tel.no. +46 46 222 17 81, email: jacob.roved@biol.lu.se
[†] Shared last author

Abstract

Major histocompatibility complex (MHC) genes play a central role for pathogen recognition by the adaptive immune system. The MHC genes are often duplicated and tightly linked within a small genomic region. This structural organization suggests that natural selection acts on the combined property of multiple MHC gene copies in segregating haplotypes, rather than on single MHC genes. This may have important implications for analyses of patterns of selection on MHC genes. Here, we present a computer-assisted protocol to infer segregating MHC haplotypes from family data, based on functions in the R package MHCtools. We employed this method to identify 107 unique MHC class I (MHC-I) haplotypes in 116 families of wild great reed warblers (*Acrocephalus arundinaceus*). In our data, the MHC-I genes were tightly linked in haplotypes and inherited as single units, with only two observed recombination events among 334 offspring. We found substantial variation in the number of different MHC-I alleles per haplotype, and the divergence between alleles in MHC-I haplotypes was significantly higher than between randomly assigned alleles in simulated haplotypes. This suggests that selection has favored non-random associations of divergent MHC-I alleles in haplotypes to increase the range of pathogens that can be recognized by the adaptive immune system. Further studies of selection on MHC haplotypes in natural populations is an interesting avenue for future research. Moreover, inference and analysis of MHC haplotypes offers important insights into the structural organization of MHC genes, and may improve the accuracy of the MHC region in *de novo* genome assemblies.

Keywords

Major Histocompatibility Complex, MHC diversity, heterozygote advantage, divergent allele advantage, songbird, great reed warbler

Introduction

The major histocompatibility complex (MHC) is a multigene family that plays a vital role in the vertebrate adaptive immune system (Klein & Sato, 2000). An ongoing arms race with pathogens has caused MHC genes to exhibit extreme levels of genetic diversity, and these genes have gathered a broad interest in studies of adaptive genetic variation in vertebrates (Ejsmond & Radwan, 2015; Kaufman, 2018; Piertney & Oliver, 2006). In humans, the MHC genes are found within a genomic region that spans approximately 4 Mb (Trowsdale & Knight, 2015). The MHC genes are often tightly linked and MHC alleles have been found to be non-randomly associated within haplotypes, suggesting that selection acts on multi-locus MHC haplotypes rather than single MHC genes (Begovich et al., 1992, 2001; Buhler, Nunes, & Sanchez-Mazas, 2016; Hollenbach et al., 2001; Kaufman, 1999; Testi et al., 2015). In MHC studies, the heterozygote advantage hypothesis is often used to describe the principle that individuals with two different maternally and paternally inherited MHC alleles should be able to recognize antigens from a larger range of pathogens, than individuals with two identical MHC alleles (Doherty & Zinkernagel, 1975; Hughes & Nei, 1992). However, from a functional perspective it should not matter whether the MHC molecules expressed in a cell are encoded by alleles that are harbored on the same or on different paternally or maternally inherited multi-locus haplotypes. This reasoning implies that haplotypes with a larger number of MHC gene copies may be favored by natural selection, because they confer an advantage in terms of presenting antigens from a larger range of pathogens, given that the genes have the same function and are expressed to the same degree. Though, note that negative selection of T-cells in the thymus is thought to associate too high individual MHC gene copy numbers with a disadvantage (Nowak, Tarczyhornoch, & Austyn, 1992; Woelfing, Traulsen, Milinski, & Boehm, 2009). Additionally, *in silico* models have shown that the degree of divergence between the amino acid sequences of MHC alleles is positively correlated with the combined number of different antigens that can be bound by the corresponding MHC molecules (Lenz, 2011; Pierini & Lenz, 2018). This suggests that haplotypes that combine highly divergent MHC alleles may be favored by natural selection in a manner similar to MHC heterozygotes (*i.e.*, divergent allele advantage, sensu Wakeland et al. (1990)).

Detailed knowledge about MHC haplotype structure is mostly limited to humans and a few model organisms, but in recent years, there has been a growing interest in characterizing MHC haplotypes and investigating their effects also in wild non-model species (Gaigher et al., 2016, 2018; Huchard, Weill, Cowlishaw, Raymond, & Knapp, 2008; Niskanen et al., 2014; Sin et al., 2014). Gaigher et al. (2016) demonstrated how analysis of the segregation patterns of MHC alleles within families in a natural population of barn owls (*Tyto alba*) could be used to infer MHC haplotypes, and thereby obtain information about linkage and recombination, and confidently assess the number of MHC gene copies and the presence of gene copy number variation (Gaigher et al., 2016). In a follow-up study, they showed how MHC haplotype data can be employed to investigate non-random associations of MHC alleles in haplotypes in a wild animal population (Gaigher et al., 2018). The studies by Gaigher et al. demonstrate the significant value of family-assisted haplotype inference in future studies of MHC genes in evolutionary biology and ecology.

Since the advent of high throughput DNA sequencing, MHC genotyping in non-model organisms has mostly been carried out using PCR-based amplicon sequencing (Biedrzycka, Sebastian, Migalska, Westerdahl, & Radwan, 2017; Burri, Promerova, Goebel, & Fumagalli, 2014; Promerová et al., 2012; Zagalska-Neubauer et al., 2010). However, in many non-model species, amplification of specific MHC loci is impeded by the sequence similarity across different loci, caused by recombination and gene conversion within and between MHC genes, and in such cases it is often necessary to co-amplify multiple MHC loci (Alcaide, Liu, & Edwards, 2013; Burri et al., 2014; Zagalska-Neubauer et al., 2010). While this technique is useful for estimating the overall genetic diversity harbored in the MHC, the resulting data contain no information about linkage or spatial organization of the amplified alleles (Alcaide et al., 2013; Biedrzycka, Sebastian, et al., 2017; Burri et al., 2014; Gaigher et al., 2016). Furthermore, the number of loci has to be estimated indirectly from the number of different alleles detected in each sample, and associating alleles with specific loci becomes extremely difficult, in particular in species with highly duplicated MHC genes (Gaigher et al., 2016; Lighten, van Oosterhout, Paterson, Mcmullan, & Bentzen, 2014; O'Connor, Westerdahl, Burri, & Edwards, 2019). This lack of resolution severely challenges studies of linkage and recombination as well as inference of selection, and it is an Achilles heel of many contemporary studies of MHC genes in evolutionary ecology (Gaigher et al., 2016; O'Connor et al., 2019). The use of family data to infer segregating MHC haplotypes offers a powerful method to overcome these challenges (Gaigher et al., 2016, 2018). However, for such studies to capture a significant part of the MHC haplotype variation in wild populations, segregation patterns should be analyzed in a large number of families - especially when studying species with high levels of MHC diversity, such as songbirds (clade Passeri of Passeriformes) (Minias, Pikus, Whittingham, & Dunn, 2018; O'Connor, Strandh, Hasselquist, Nilsson, & Westerdahl, 2016). Fortunately, as the cost of high-throughput DNA sequencing continues to decrease, genotyping the number of samples necessary to infer MHC haplotypes in a large number of families is becoming feasible to many research groups.

To assist such studies, we here present the R package MHCtools that contains a set of functions for automated MHC haplotype inference from family data (Roved, 2019). Besides MHC haplotype inference, MHCtools contains functions that facilitate the bioinformatical steps involved in filtering large amplicon sequencing data sets and downstream data analysis (Table 1). In the present paper, we demonstrate the use of MHCtools to assist filtering of MHC class I (MHC-I) amplicon sequencing data and to carry out automated MHC-I haplotype inference using data from a wild population of great reed warblers (*Acrocephalus arundinaceus*), a songbird with highly duplicated MHC genes and extensive gene copy number variation between individuals (Roved, Hansson, Tarka, Hasselquist, & Westerdahl, 2018; Westerdahl, Wittzell, & von Schantz, 1999; Westerdahl, Wittzell, von Schantz, & Bensch, 2004). We used the resulting haplotype data set to estimate the degree of MHC-I gene copy number variation in this species, and the degree of recombination in the chromosomal region containing the MHC-I genes. Finally, we investigated whether tight linkage among the MHC-I genes has favored evolution of haplotypes that combine highly divergent MHC-I alleles.

Methods

Data set

We used empirical data on 141 adult males, 131 adult females, and 287 chicks from our long term study population of great reed warblers breeding at lake Kvismaren (59°10'N, 15°25'E) in southern Central Sweden (Hansson et al., 2018; Hasselquist, Montras-Janer, Tarka, & Hansson, 2017; Roved et al., 2018). The adult individuals in our data set have been observed, examined, and ringed in the period 1984–2004, and the chicks constitute the 1998 and 1999 cohorts from the same population, with addition of one nest from each of the years 1992 and 1996 (cf. Bensch et al., 1998; Hasselquist, 1998; Tarka et al., 2014). All territorial males and breeding females in our study population were mist-netted and marked

Table 1 *Overview of the functions included in MHCtools v. 1.2.1.*

CalcPdist	Calculates nucleotide or amino acid p-distances (i.e. the proportion of variable positions) from pairwise sequence comparisons and mean p-distances for each sample in a DADA2 sequence table.
CreateFas	Creates a FASTA file with all the sequences in a DADA2 sequence table.
CreateSamplesFas	Creates a set of FASTA files with the sequences present in each sample in a DADA2 sequence table.
HpltFind	Automatically infers major histocompatibility complex (MHC) haplotypes from the genotypes of parents and offspring in families. The functions **GetHpltTable()** and **GetHpltStats()** are designed to evaluate the output files.
PapaDiv	Calculates the joint major histocompatibility complex (MHC) diversity in parent pairs, taking into account alleles that are shared between the parents. The joint diversity in parent pairs is useful for heritability analyses in non-model species, where one wants to estimate the heritability of MHC diversity. The number of unique alleles in offspring may not be directly derived from the parental genotypes if some alleles are shared between the parents.
ReplMatch	Automatically compares technical replicates in an amplicon sequencing data set and reports the proportion of mismatches. The functions **GetReplTable()** and **GetReplStats()** are designed to evaluate the output files.

with aluminum rings and unique combinations of color rings (Bensch et al., 1998). We located > 95% of all nests before fledging of the offspring and registered all breeding attempts (Bensch et al., 1998). 8–10 days after hatching we ringed the nestlings with an aluminum ring and a year-specific color ring. The paternity and maternity of all chicks were verified by molecular methods (Hansson, Hasselquist, & Bensch, 2004; Hasselquist, Bensch, & von Schantz, 1995). Approximately 3% of the total number of offspring resulted from extra-pair mating (Hansson et al., 2004; Hasselquist et al., 1995; Hasselquist, Bensch, & von Schantz, 1996), and these were assigned to their genetic parents in all analyses.

Fieldwork and DNA sampling were approved by the Malmö/Lund Animal Ethics Committee and the Swedish Bird Ringing Centre.

DNA sampling and sequencing

We collected 20–80 µl blood from each individual by puncturing the brachial or tarsus vein. The blood samples were suspended in 500 µl SET buffer (0.15 M NaCl, 0.05 M TRIS, 0.001 M EDTA), and kept frozen at –20°C until DNA extraction. Isolation of genomic DNA was carried out by standard phenol/chloroform-isoamylalcohol extraction (Sambrook, Fritsch, & Maniatis, 1989). The purified DNA was stored in 1×TE buffer and frozen at –50 or –80°C.

We amplified a 262 bp region of MHC-I exon 3 from 559 great reed warblers using previously designed primers: HNalla-HN46 (O'Connor et al., 2016; Westerdahl et al., 2004). Approximately half of the amplicons (samples from 88 adult males, 100 adult females, and 145 chicks) were sequenced in a Roche 454 GS FLX (F. Hoffmann-La Roche AG, Basel, Switzerland) following the manufacturer's instructions at the Department of Biology, Lund University. See Roved et al. (2018) for details on experimental setup, tags, and PCR amplification in the 454-sequencing experiment. The remaining amplicons were sequenced in two independent runs using 300 bp paired-end sequencing in an Illumina MiSeq (Illumina Inc., San Diego, CA, USA) following the manufacturer's instructions at the Department of Biology, Lund University. Samples from 23 adult males, 32 adult females, and 150 chicks (including 11 replicates from the 454-sequencing experiment) were included in the first run and a smaller batch with samples from 30 adult males were added in a second run. In both runs, great reed warbler samples were multiplexed with other samples in libraries comprising 384 samples. Details on tags, PCR amplification, and library preparation for the Illumina MiSeq sequencing experiment are provided in the SI (Supplementary Methods).

Filtering and screening of sequencing data

The 454 sequencing data were demultiplexed using the software jMHC (Stuglik, Radwan, & Babik, 2011) and filtered according to the filtering protocol described in Galan, Guivier, Caraux, Charbonnel, & Cosson (2010). Subsequently, the data were manually screened to remove low-quality and artificial sequences, and non-functional alleles. See (Roved et al., 2018) for further details on the results and bioinformatics protocol of the 454 sequencing experiment.

The Illumina MiSeq sequencing outputs were trimmed to remove adapters, primers, and tag sequences using the software Cutadapt version 1.14 (Martin, 2011). The trimmed sequences were then filtered using the R package DADA2 version 1.4.0 (Callahan et al., 2016) in R version 3.4.2 (R Core Team, 2017). We inspected the error rate plots for deviations from the error rates expected given the observed Q-values, and employed the integrated function removoBimeraDenovo in DADA2 to remove chimeras. The machine learning algorithm employed by DADA2 to estimate the error rates associated with a data set requires the setting of a prior of maximum expected error rates, and the final output from DADA2 depends on this prior setting (Roved et al., unpublished data). Different combinations of primers and/or study species may produce different amounts of sequencing errors; hence, prior settings may be unpredictable for new data sets. Therefore, each study should perform a careful evaluation of the accuracy obtained with a number of different prior settings. Such an evaluation can be performed by comparing technical replicates in the data set.

After preliminary filtering of the first Illumina data set we identified 52 samples that had identical genotypes to one or more other samples (predominantly full siblings) in the data set, and grouped these samples into 25 replicate sets. We then optimized the filterAndTrim settings in DADA2 by comparing the replicate sets in 29 filtering runs with different settings. For each of these runs, we employed the functions ReplMatch and GetReplStats in the R package MHCtools version 1.2.1 to obtain repeatability estimates for the optimization of the filterAndTrim settings in DADA2. We identified the optimal filterAndTrim settings by evaluating the repeatability obtained with each setting. The settings providing the highest repeatability were: maxN = 0, maxEE fw = 0.1, maxEE rv = 0.1, and truncQ = 20. DADA2 inferred 295 sequences in the first Illumina data set with these settings. We then used the function CreateFas in MHCtools to create a fasta file with all the filtered sequences, and inspected these manually in BioEdit version 7.2.5 (Hall, 1999) to remove non-functional variants (sequences containing stop codons or indels that induced shifts in the reading frame). The remaining sequences were filtered by their relative abundance within each amplicon (per amplicon frequency). We again optimized the filtering threshold using repeatability estimates obtained by comparing the replicate sets using ReplMatch and GetReplStats in MHCtools. The per amplicon frequency threshold that gave the highest repeatability was 0.014. After filtering and screening, our first Illumina data set contained 226 unique DNA alleles in 205 samples. 197 alleles were found in both the first Illumina and the 454 sequencing runs, while 29 alleles were only found in the first

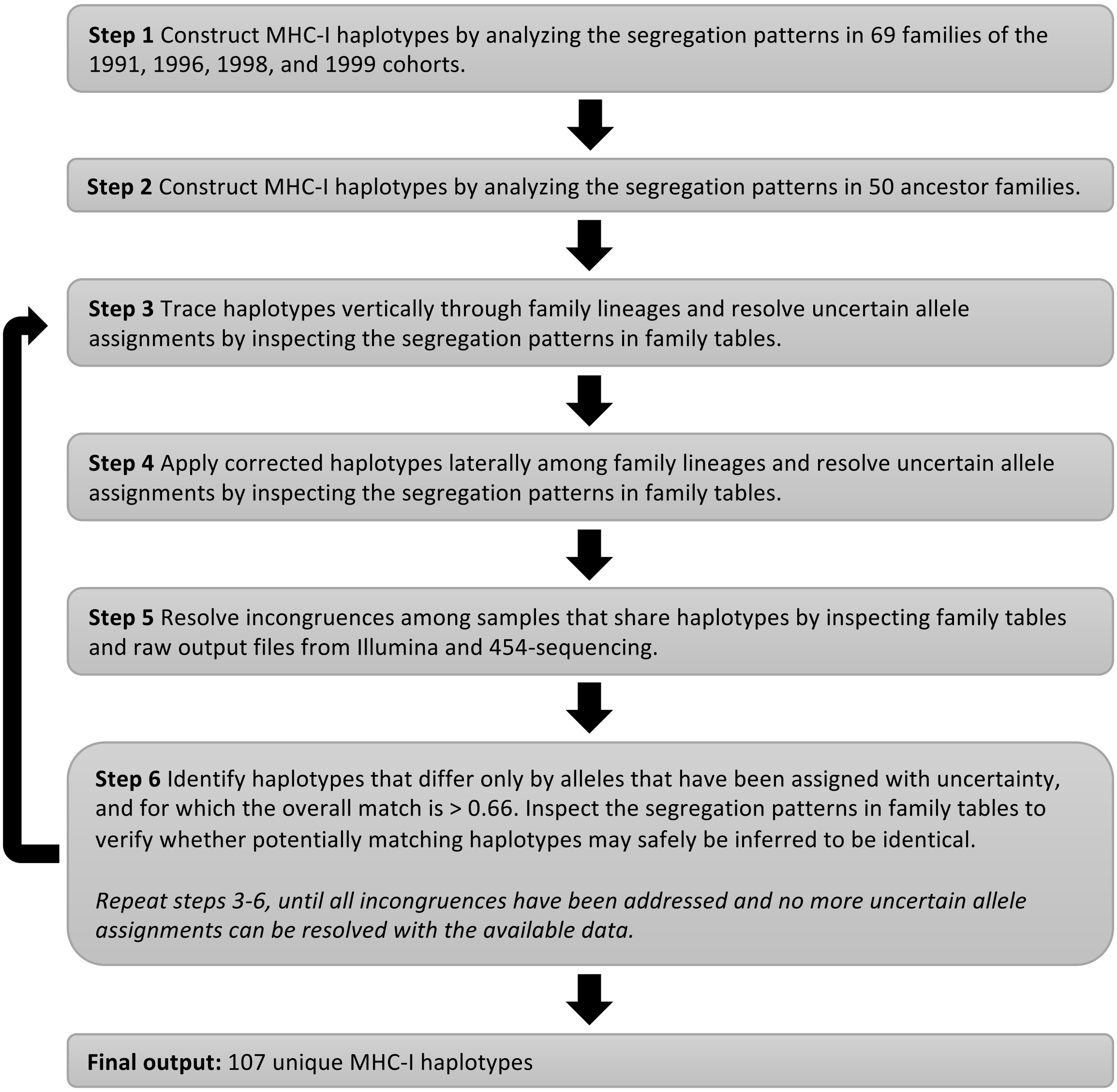
Step 1 Construct MHC-I haplotypes by analyzing the segregation patterns in 69 families of the 1991, 1996, 1998, and 1999 cohorts.
Step 2 Construct MHC-I haplotypes by analyzing the segregation patterns in 50 ancestor families.
Step 3 Trace haplotypes vertically through family lineages and resolve uncertain allele assignments by inspecting the segregation patterns in family tables.
Step 4 Apply corrected haplotypes laterally among family lineages and resolve uncertain allele assignments by inspecting the segregation patterns in family tables.
Step 5 Resolve incongruences among samples that share haplotypes by inspecting family tables and raw output files from Illumina and 454-sequencing.
Step 6 Identify haplotypes that differ only by alleles that have been assigned with uncertainty, and for which the overall match is > 0.66. Inspect the segregation patterns in family tables to verify whether potentially matching haplotypes may safely be inferred to be identical.
Repeat steps 3-6, until all incongruences have been addressed and no more uncertain allele assignments can be resolved with the available data.
Final output: 107 unique MHC-I haplotypes

Fig. 1 *Flow chart of the haplotype inference protocol.*

Illumina run and 132 alleles were only found in the 454 sequencing run. The number of reads per sample was normally distributed with a mean of 20,920, a minimum of 12,532, and a maximum of 32,781.

The data from the second Illumina run was filtered and inspected using the same procedure and settings as the first Illumina data set, resulting in an initial inference of 200 sequences which, after filtering by per amplicon frequency and removal of two samples with low read numbers (0 and 3,150 reads, respectively) resulted in 162 unique DNA alleles in 31 samples. 32 of these alleles were not observed in the previous Illumina or 454 data sets. The number of reads per sample in the second Illumina data set was normally distributed with a mean of 22,492, a minimum of 12,062, and a maximum of 38,521.

After filtering and screening, we merged the Illumina MiSeq and 454 sequencing data sets for downstream analyses. Two genotyped samples were excluded from further analyses due to labelling errors. All DNA sequences were blasted against the NCBI database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and novel alleles were named following the MHC standardized nomenclature (Klein et al., 1990).

Haplotype inference

We employed the HpltFind function in MHCtools to infer MHC-I haplotypes in our data set. We initially used the genotyping data from 372 samples from 67 families that constituted the 1998 and 1999 cohorts from our study population (*i.e.*, all families from those years, in total 78 parents and 282 chicks), one family from 1991 (2 parents and 5 chicks), and one from 1996 (2 parents and 3 chicks). Hereafter, these families will be referred to as 'the 69 families'.

Among the 390 MHC-I alleles that we observed in our data set, the five most common ones were present in 98%, 97%, 87%, 83%, and 50% of the samples, respectively. Such common alleles are often present in bothparents in families, and it can be difficult to ascertain their presence in one or both haplotypes in each parent through analysis of segregation patterns. To obtain the highest possible resolution, we therefore investigated segregation patterns across multiple generations, using the detailed pedigree of our study population (Hansson, Åkesson, Slate, & Pemberton, 2005; Tarka et al., 2014). Among 26 parents from the 69 families, we were able to trace ancestry up to five generations back, while the remaining parents were immigrants, for which no pedigree data were available (SI, Pedigree table). We investigated allele segregation patterns in 50 ancestral families of the 26 parents from the 69 families, that we were able to trace in our pedigree (hereafter referred to as 'the ancestral families') (SI, Pedigree table). This data included the genotypes of in total 80 adult individuals, of which 18 were also observed as parents in the 1998 and 1999 cohorts.

We inferred haplotypes based on allele segregation patterns by the following steps (illustrated with a flow chart in Fig. 1). The haplotype inference process on our data set is summarized in Table 2.

1. We used the HpltFind function in MHCtools to construct putative haplotypes for each individual in the 69 families. HpltFind assigns alleles to a parental haplotype if they occur in either parent and in one or more offspring. In cases when both parents and/or all or most offspring share an allele, it may not be possible to assign the allele to a parental haplotype with certainty. HpltFind assigns such alleles to all of their potential haplotypes and flags them as unresolved. We have illustrated this analysis for one family in Fig. 2.
2. We then constructed haplotypes for each individual in the ancestral families, using the same procedure as in step 1.
3. We traced the inheritance of all haplotypes through generations as far as data were available within each family lineage (SI, Pedigree table). When a haplotype mismatched between generations by alleles that had been flagged as unresolved, we inspected the family tables to verify whether we could resolve the allele assignments.
4. We then applied each corrected ancestral haplotype throughout all families in which it occurred, and resolved lateral mismatches between family lineages by the same procedure as we applied to the vertical mismatches observed within family lineages in step 3.
5. If incongruences were found between individuals that inherited or passed down a haplotype, and these incongruences could not be resolved by inspection of the family tables as described in steps 3–4, the most likely cause of the incongruence was identified. In rare cases, some alleles (i) had failed to amplify during PCR or had been erroneously deleted from samples in the filtering process (usually due to low PCR amplification success, *i.e.* null alleles), (ii) turned out to be sequencing errors that survived the filtering process (only alleles that had low read numbers and could be derived from more abundant alleles in the same sample by single nucleotide substitutions), or (iii) had not initially been assigned to a haplotype in the inference process (usually

Nest 28 of the 1999 cohort

Alleles	Mother	Father	Offspring 1	Offspring 2	Offspring 3
Acar-UA*4	X	X	X	X	X
Acar-UA*9	X	X	X	X	X
Acar-UA*55	X	X	X	X	X
Acar-UA*12	X	-	X	X	-
Acar-UA*79	X	-	X	X	-
Acar-UA*122	X	-	X	X	-
Acar-UA*125	X	-	X	X	-
Acar-UA*133	X	-	X	X	-
Acar-UA*201	X	-	X	X	-
Acar-UA*276	X	-	X	X	-
Acar-UA*296	X	-	X	X	-
Acar-UA*340	X	-	X	X	-
Acar-UA*348	X	-	X	X	-
Acar-UA*31	X	-	-	-	X
Acar-UA*153	X	-	-	-	X
Acar-UA*157	X	-	-	-	X
Acar-UA*223	X	-	-	-	X
Acar-UA*239	X	-	-	-	X
Acar-UA*144	X	X	X	-	X
Acar-UA*285	-	X	X	-	-
Acaru-UA*23	-	X	X	-	-
Acar-UA*94	-	X	-	X	X
Acar-UA*119	-	X	-	X	X
Acar-UA*271	-	X	-	X	X

Putative segregating haplotypes

Mother A	Mother B
Acar-UA*4	Acar-UA*4
Acar-UA*9	Acar-UA*9
Acar-UA*55	Acar-UA*55
Acar-UA*12	Acar-UA*31
Acar-UA*79	Acar-UA*153
Acar-UA*122	Acar-UA*157
Acar-UA*125	Acar-UA*223
Acar-UA*133	Acar-UA*239
Acar-UA*201	Acar-UA*144
Acar-UA*276	
Acar-UA*296	
Acar-UA*340	
Acar-UA*348	

Father A	Father B
Acar-UA*4	Acar-UA*4
Acar-UA*9	Acar-UA*9
Acar-UA*55	Acar-UA*55
Acar-UA*144	Acar-UA*94
Acar-UA*285	Acar-UA*119
Acaru-UA*23	Acar-UA*271

Final haplotypes

Mother A	Mother B
Acar-UA*9	Acar-UA*4
Acar-UA*55	Acar-UA*9
Acar-UA*12	Acar-UA*55
Acar-UA*79	Acar-UA*31
Acar-UA*122	Acar-UA*153
Acar-UA*125	Acar-UA*157
Acar-UA*133	Acar-UA*223
Acar-UA*201	Acar-UA*239
Acar-UA*276	Acar-UA*144
Acar-UA*296	
Acar-UA*340	
Acar-UA*348	

Father A	Father B
Acar-UA*4	Acar-UA*4
Acar-UA*9	Acar-UA*9
Acar-UA*55	Acar-UA*55
Acar-UA*144	Acar-UA*94
Acar-UA*285	Acar-UA*119
Acaru-UA*23	Acar-UA*271

Fig. 2 *Family table from nest number 28 of the 1999 cohort showing MHC-I allele segregation patterns with inferred putative segregating MHC-I haplotypes (Mother A, Mother B, Father A, Father B) marked by different colors. Dark gray color indicates that a segregation pattern could not be determined for an allele, because it was present in both parents and in all offspring (uncertain allele). In the final haplotypes, a number of uncertain alleles were resolved by applying steps 3-6 in the haplotype inference protocol (Fig. 1).*

Table 2 *Summary of the haplotype inference process.*

Steps 1–2	
Initial number of putative haplotypes	225
Initial mean proportion of unresolved alleles in putative haplotypes	0.446
Steps 3–4	
Number of putative haplotypes identical by descend to other haplotypes	165
Number of unresolved alleles verified in putative haplotypes	358
Number of alleles added to putative haplotypes (failed initial assignment)	155
Number of alleles removed from putative haplotypes (invalid initial assignment)	121
Step 5	
Number of sequencing errors called as alleles, removed from individual samples	15
- proportion of the total number of allele assignments to genotypes	*0.0019*
Number of null alleles added to individual samples	430
- proportion of the total number of allele assignments to genotypes	*0.051*
Step 6	
Number of putative haplotypes estimated to be homologous to other haplotypes	41
Number of unresolved alleles verified in putative haplotypes	128
Number of alleles added to putative haplotypes (failed initial assignment)	7
Number of alleles removed from putative haplotypes (invalid initial assignment)	33
Final number of putative haplotypes	**107**
Final mean proportion of unresolved alleles in putative haplotypes	**0.255**

because it was a null allele in one or more samples in the family). Each incongruence was investigated by inspecting the family tables and the raw sequencing output files for presence or absence of the mismatching allele prior to filtering. In cases of multiple solutions to an incongruence, the solution involving the fewest assumptions was applied. By comparing haplotypes both within and between family lineages as described in steps 3–5, we were able to resolve a large proportion of the uncertain allele assignments (Table 2).

6. Finally, we identified potentially identical haplotypes throughout the data set, using a lower threshold of 0.66 for the proportion of matching sequences between any two haplotypes. We investigated all potential matches by manual inspection. Whenever a set of haplotypes only mismatched by alleles that had been assigned as uncertain (and had not been resolved in steps 3–5), we inspected the family tables to verify whether we could safely assume identity between the potentially matching haplotypes. This enabled us to resolve some additional uncertain allele assignments (Table 2).

We repeated steps 3–6, further improving the exactness of our haplotype inference by reapplying the corrected haplotypes both within and between families and family lineages. This process was repeated until all incongruences had been addressed and no more uncertain allele assignments could be resolved with the available data.

We were unable to resolve the segregation of alleles in three of the ancestral families, as the genotypes did not overlap. The segregation patterns suggested that blood samples from two individuals in these families have been exchanged or mislabeled, so these samples and nests were excluded from further analyses.

Estimating the recombination rate

We calculated the recombination rate as the number of recombinant haplotypes divided by the total number of gametes (*i.e.*, the number of offspring in families for which we successfully inferred haplotypes multiplied by two).

Genetic divergence between alleles in MHC-I haplotypes

We used the CalcPdist function in MHCtools to quantify the proportion of nucleotide and amino acid differences (p-distances) between all pairs of MHC-I alleles, and calculated the means of the pairwise p-distances between all alleles in each haplotype.

To test whether alleles were more divergent within haplotypes than expected by chance, we generated 1,000 *in silico* simulations of our haplotype data set. The simulated data sets were generated by randomly assigning existing alleles to haplotypes while maintaining the number of different alleles for each haplotype. We calculated means of the pairwise nucleotide and amino acid p-distances between all alleles in each simulated haplotype and, for each simulation, calculated a mean across all haplotypes. Finally, for both nucleotide and amino acid p-distances, we compared the mean of the p-distances observed across all real haplotypes against the distribution of the mean p-distances derived from the simulated haplotype sets. The p-distance calculations, data simulations, and t-tests were carried out in R version 3.6.1 (R Core Team, 2019).

Results

We developed a stepwise protocol for inference of MHC haplotypes in non-model species, based on the function HpltFind in the R package MHCtools (Roved, 2019), which carries out automated analysis of allele segregation patterns in family data (Fig. 1). We demonstrated the utility of this and a number of auxiliary functions from MHCtools by carrying out MHC-I genotyping and haplotype inference on an empirical data set of 559 great reed warblers from our long-term study population in Lake Kvismaren in southern Central Sweden.

We genotyped the individuals in our data set by amplifying and sequencing the MHC-I exon 3 using high-throughput Roche 454 and Illumina MiSeq platforms. The data set that was sequenced using the Roche 454 platform was filtered according to the principles in Galan et al. (2010), and we achieved a repeatability of 0.94 for these samples (estimated across 50 replicate sets) (Roved et al., 2018). The data set that was sequenced using Illumina MiSeq was filtered with DADA2 (Callahan et al., 2016), and we achieved a near perfect repeatability of 0.998

(estimated across 25 replicate sets). When comparing 11 replicated MHC-I genotypes between the Roche 454 and Illumina MiSeq runs, we achieved a repeatability of 0.96 between the sequencing platforms. The final, merged data set contained 390 unique MHC-I exon 3 alleles in 559 samples, of which 324 alleles were unique at the amino acid sequence level.

We successfully identified 107 unique MHC-I haplotypes based on allele segregation patterns in 116 great reed warbler families (SI, Haplotype Tables). Segregation patterns may be hard to resolve for alleles that are observed at high frequencies in the population, and our data set contained five MHC-I alleles with allele frequencies > 0.5. However, by analyzing allele segregation across multiple generations, we were able to reduce the mean proportion of unresolved alleles to 0.199 for haplotypes that were observed in multiple families, while it was 0.327 for haplotypes that were only observed in single families. For all samples, the mean proportion of unresolved alleles was 0.255 (Table 2). Furthermore, our haplotype inference process revealed a high congruence between the genotypes of related individuals in our data set. In the haplotype inference process, we discovered and removed 15 sequencing errors from individual samples in our data set (type I errors), corresponding to a proportion of 0.0019 of the total number of allele assignments (Table 2). We discovered 430 null alleles (type II errors), corresponding to a proportion of 0.051 of the total number of allele assignments, which were subsequently added to individual samples (Table 2). The larger number of type II compared to type I errors was expected, since the amplification of individual loci may be more or less successful in different samples.

We found considerable variation in the number of MHC-I gene copies between haplotypes, with as little as four and as many as 21 different alleles in single haplotypes. The mean number of different MHC-I alleles per haplotype was 9.2 with a standard deviation (S.D.) of 2.80 (Fig. 3). Furthermore, we found two recombinant haplotypes among the 334 offspring of the 116 families. From this observation, we estimated a recombination rate of 0.0030 within the genomic region spanned by the MHC-I in great reed warblers, corresponding to a distance of 0.3 centimorgan (cM).

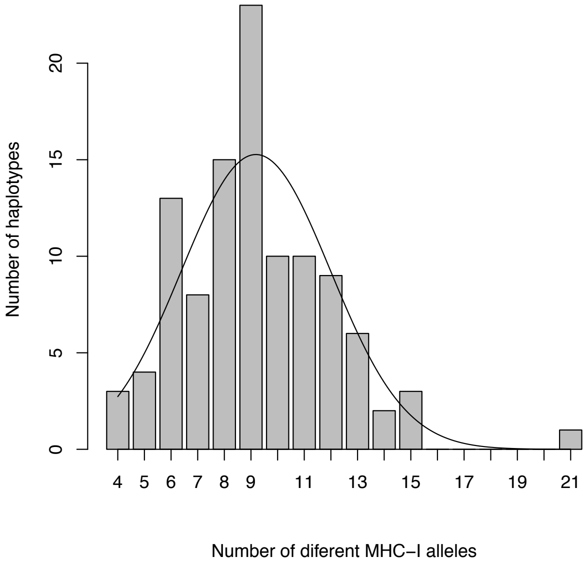

Fig. 3 *Distribution of the number of different MHC-I alleles on haplotypes. The line shows a normal distribution with the observed mean (9.2) and standard deviation (2.80*).

The mean nucleotide divergence in the MHC-I haplotypes in our data set ranged from 0.083 to 0.139 with a mean of 0.118 (S.D. = 0.0097). The mean amino acid divergence in the MHC-I haplotypes in our data set ranged from 0.145 to 0.217 with a mean of 0.186 (S.D. = 0.0130). To investigate whether natural selection may have favored combinations of MHC-I alleles that are highly divergent from each other, we tested the mean nucleotide and amino acid p-distances observed in real haplotypes against similar measures obtained from 1,000 simulated data sets, in which alleles were randomly assigned to haplotypes. The mean p-distances across the real haplotypes were significantly larger than across the simulated haplotypes, both on the nucleotide

Table 3 *Observed mean nucleotide and amino acid p-distances in real haplotypes and across 1,000 simulated haplotype data sets.*

	Nucleotide p-distance	**Amino acid p-distance**
Mean p-distance in real haplotypes	0.118	0.186
Mean p-distance across 1,000 simulated haplotype sets (S.D.)	0.101 (± 0.0014)	0.163 (± 0.0020)

and amino acid levels ($p < 0.001$ in both tests; Table 3; Fig. 4).

Discussion

In the present study, we demonstrated the use of the newly developed R package MHCtools (Roved, 2019) for analysis of allele segregation patterns in species with highly duplicated gene families. We employed MHCtools in a stepwise protocol to characterize segregating MHC-I haplotypes in 116 families of a wild songbird, the great reed warbler. We identified 107 unique MHC-I haplotypes, and overall, observed a high degree of congruence in our haplotype inference process, which confirmed the accuracy of our genotyping methods. This provides a solid validation of our protocols, in particular given that haplotype inference is difficult in species with highly duplicated MHC genes (such as the great reed warbler) (Gaigher et al., 2018). The number of different MHC-I haplotypes that we observed in our population corresponds to the number observed in a previous study on barn owls, where 111 MHC-I haplotypes were observed among 140 families (Gaigher et al., 2018). Such standing genetic variation in MHC diversity could serve an important evolutionary function by enabling rapid adaptive shifts in response to the dynamics of faster evolving pathogens (Alves et al., 2019; O'Connor et al., 2019). Our analyses of the allele segregation patterns confirmed strong linkage of the MHC-I loci in the great reed warbler with a recombination rate of 0.0030, corresponding to a genetic distance of 0.3 cM.

As tightly linked MHC loci often co-segregate, investigations of the structure of MHC genes, haplotypes, and recombination between MHC loci may improve our understanding of correlations between MHC variation, fitness, and disease in wild populations (O'Connor et al., 2019). The MHC exhibits extraordinary evolutionary dynamics with rapid expansions and contractions of MHC gene copy number, and substantial variation in MHC sequence and haplotype structure (Kelley, Walter, & Trowsdale, 2005; Minias et al., 2018; Masatoshi Nei & Rooney, 2005; O'Connor et al., 2016; Ohta, 1991; Spurgin et al., 2011). Thus, previous studies have reported considerable variation in the number of different MHC alleles between individuals within species, suggesting that MHC gene copy number variation may be a common trait, at least among birds (Biedrzycka, O'Connor, et al., 2017; Gaigher et al., 2016; Roved et al., 2018; Stervander, Dierickx, Thorley, Brooke, & Westerdahl, 2020; Whittingham, Dunn, Freeman-Gallant, Taff, & Johnson, 2018). Our analyses of MHC-I haplotypes confirmed previous indications of substantial MHC-I gene copy number variation in the great reed warbler (O'Connor et al., 2016; Roved et al., 2018), with a minimum of 4 and a maximum of 21 different MHC-I alleles per haplotype (Fig. 3). Such variation in the number of different alleles on MHC haplotypes may seem puzzling, as one

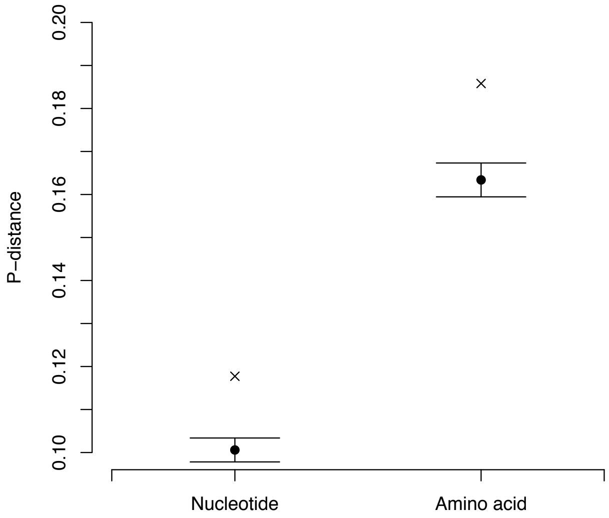

Fig. 4 *Observed mean nucleotide and amino acid p-distances on real haplotypes (crosses) versus mean nucleotide and amino acid p-distances ± 1 S.D. across 1,000 simulated haplotype data sets (dots w. error bars).*

would expect selection to purge haplotypes that deviate too much from the optimal level of MHC diversity in the population. One explanation could be that variation in the number of alleles per haplotype is offset by variation either in the functional divergence between alleles or in the degree of specificity in the peptide binding properties of the alleles (cf. Chappell et al., 2015). It is also possible that deviations from optimal MHC diversity are offset by selective advantages associated with particular alleles (e.g. Bateson et al., 2016; Bonneaud, Perez-Tris, Federici, Chastel, & Sorci, 2006; Sepil, Lachish, Hinks, & Sheldon, 2013; Westerdahl et al., 2005), or that haplotypes with lower than optimal diversity are maintained in combinations with other haplotypes that harbor high diversity - and vice versa. Gene expression could also play a role, since not all MHC-I genes may be expressed to the same degree (Chappell et al., 2015; Drews, Strandh, Råberg, & Westerdahl, 2017). Furthermore, we have previously found evidence for a sexual conflict over MHC-I diversity in the great reed warbler, as having a higher than average number of different MHC-I alleles was advantageous for males, while the opposite was true for females (Roved et al., 2018). It is plausible that the selective advantage of having high MHC-I diversity in males could help maintain haplotypes with high numbers of alleles in great reed warblers, while, on the other hand, the disadvantage of such haplotypes in females could help maintain haplotypes with only few alleles.

Besides investigating MHC diversity in terms of the number of different alleles, we also took advantage of our haplotype data to investigate whether selection may have favored haplotypes that combine highly divergent MHC-I alleles in the great reed warbler. We showed that the mean proportions of nucleotide and amino acid differences between alleles were significantly higher in real haplotypes than in simulated haplotypes, to which alleles were assigned at random (Fig. 4). Because of the strong linkage among the MHC-I loci in the great reed warbler, non-random association of MHC-I alleles in haplotypes may be adaptive because it ensures that optimal MHC diversity is transmitted from parents to offspring. Our results suggest that selection favors combinations of highly divergent alleles in MHC-I haplotypes, because it is likely to increase the range of pathogens that can be recognized by the adaptive immune system, following the principle of the divergent allele advantage hypothesis (Wakeland et al., 1990). However, optimal individual MHC diversity in offspring may also be achieved by parents commonly choosing to mate with MHC-compatible partners (Aeschlimann, Haberli, Reusch, Boehm, & Milinski, 2003; Penn & Potts, 1999; Strandh et al., 2012). Even so, selection may favor non-random association of MHC alleles in haplotypes, because it may be advantageous for offspring irrespective of the degree of compatibility between maternally and paternally transmitted alleles. This may in particular be true under conditions that preclude random mating, e.g. in inbred or bottlenecked populations. Non-random association of highly divergent MHC alleles in haplotypes has previously been shown in wild chacma baboons (*Papio ursinus*), where it was suggested that selection favors haplotypes that combine MHC-DRB alleles with dissimilar physicochemical properties across multiple loci (Huchard et al., 2008). In contrast, Gaigher et al. (2018) found no evidence for a shift towards highly divergent allele combinations in MHC class I or II haplotypes in barn owls.

To our knowledge, this is the first study attempting to characterize MHC haplotypes in a species with highly duplicated MHC genes. However, while our results provide clear evidence for non-random association of MHC-I alleles in haplotypes in the great reed warbler, further investigations are required to establish whether alleles at different loci share the same evolutionary history. In humans, the MHC-I contains three highly polymorphic classical loci involved in antigen presentation (HLA-A, -B, and -C), and these have evolved divergent antigen binding properties (Buhler et al., 2016; Nei, Gu, & Sitnikova, 1997; Paul et al., 2013; Pierini & Lenz, 2018; Rao, De Boer, van Baarle, Maiers, & Kesmir, 2013). In birds, extraordinary divergence has been shown for MHC class IIB, where alleles separate into two lineages that predate the radiation of extant species (Goebel et al., 2017), but no such pattern has been described for MHC-I. To achieve a better understanding of the organization of and the evolutionary relationship between MHC-I genes in birds and other non-model species, it is necessary to investigate how different alleles associate with different loci. Understanding how MHC alleles segregate in haplotypes is an obvious step along that road, and may be complemented by analyses of the phylogenetic relationship between alleles, as exemplified in e.g. Goebel et al. (2017) and Burri et al. (2008), and by studies that attempt to map the physical structure of the MHC region (O'Connor et al., 2019). The latter may be possible to achieve by genome sequencing, in particular using long-read or linked read technologies, but current genome assembly methods have yet to overcome obstacles presented by the large number of duplicated sequences in the MHC region (Näpflin et al., 2019; O'Connor et al., 2019). Until such methods are developed and can be applied across a range of both species and individuals, linkage maps derived from haplotype inference may offer valuable insights into the structural organization of the MHC region. In addition, recent advances in studies on humans have demonstrated that incorporating prior knowledge about population-wide haplotype diversity may improve the accuracy of MHC genotyping in *de novo* genome assemblies (Dilthey, Cox, Iqbal, Nelson, & McVean, 2015). This underlines both the current and future value of methods and tools, such as those presented in this paper, that facilitate population-wide screening of MHC haplotypes in non-model species, where no such data are available *a priori*.

Acknowledgements

We wish to thank A. Drews and T. Johansson for library constructions and high-troughput sequencing. Furthermore, we are grateful to S. Bensch, B. Nielsen, M. Åkesson, and staff at Kvismare Bird Observatory for considerable help with the great reed warbler fieldwork. Funding and support were provided by the Swedish Research Council (621-2013-4357, 2016-04391 to DH; 621-2014-5222, 2016-00689 to BH; 2015-05149 to HW), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Start Up grant 679799 to HW; Advanced grant 742646 to DH), a Linnaeus research excellence grant (349-2007-8690) from the Swedish Research Council and Lund University to the Centre for Animal Movement Research (CAnMove), Lunds Djurskyddsfond, Kungl. Fysiografiska Sällskapet i Lund (JR), and Kvismare Bird Observatory (report number 193).

Conflicts of interest

The authors have no conflicting interests to declare.

References

Aeschlimann, P. B., Haberli, M. A., Reusch, T. B. H., Boehm, T., & Milinski, M. (2003). Female sticklebacks Gasterosteus aculeatus use self-reference to optimize MHC allele number during mate selection. *Behavioral Ecology and Sociobiology*, *54*(2), 119–126. https://doi.org/10.1007/s00265-003-0611-6

Alcaide, M., Liu, M., & Edwards, S. V. (2013). Major histocompatibility complex class I evolution in songbirds: universal primers, rapid evolution and base compositional shifts in exon 3. *PeerJ*. https://doi.org/10.7717/peerj.86

Alves, J. M., Carneiro, M., Cheng, J. Y., de Matos, A. L., Rahman, M. M., Loog, L., … Jiggins, F. M. (2019). Parallel adaptation of rabbit populations to myxoma virus. *Science*, *363*(6433), 1319–1326. https://doi.org/10.1126/science.aau7285

Bateson, Z. W., Hammerly, S. C., Johnson, J. A., Morrow, M. E., Whittingham, L. A., & Dunn, P. O. (2016). Specific alleles at immune genes, rather than genome-wide heterozygosity, are related to immunity and survival in the critically endangered Attwater's prairie-chicken. *Molecular Ecology*. https://doi.org/10.1111/mec.13793

Begovich, A. B., McClure, G. R., Suraj, V. C., Helmuth, R. C., Fildes, N., Bugawan, T. L., … Klitz, W. (1992). Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. *The Journal of Immunology*, *148*, 249–258.

Begovich, A. B., Moonsamy, P. V., Mack, S. J., Barcellos, L. F., Steiner, L. L., Grams, S., … Klitz, W. (2001). Genetic variability and linkage disequilibrium within the HLA-DP region: Analysis of 15 different populations. *Tissue Antigens*, *57*(5), 424–439. https://doi.org/10.1034/j.1399-0039.2001.057005424.x

Bensch, S., Hasselquist, D., Nielsen, B., & Hansson, B. (1998). Higher fitness for philopatric than for immigrant males in a semi-isolated population of great reed warblers. *Evolution*, *52*(3), 877–883. https://doi.org/10.2307/2411282

Biedrzycka, A., O'Connor, E., Sebastian, A., Migalska, M., Radwan, J., Zając, T., … Westerdahl, H. (2017). Extreme MHC class I diversity in the sedge warbler (Acrocephalus schoenobaenus); Selection patterns and allelic divergence suggest that different genes have different functions. *BMC Evolutionary Biology*, *17*(1), 1–12. https://doi.org/10.1186/s12862-017-0997-9

Biedrzycka, A., Sebastian, A., Migalska, M., Westerdahl, H., & Radwan, J. (2017). Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. *Molecular Ecology Resources*, *17*(4), 624–655. https://doi.org/10.1111/1755-0998.12612

Bonneaud, C., Perez-Tris, J., Federici, P., Chastel, O., & Sorci, G. (2006). Major histocompatibility alleles associated with local resistance to malaria in a passerine. *Evolution*, *60*(2), 383–389. https://doi.org/10.1554/05-409.1

Buhler, S., Nunes, J. M., & Sanchez-Mazas, A. (2016). HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*, *68*, 401–416. https://doi.org/10.1007/s00251-016-0918-x

Burri, R., Promerova, M., Goebel, J., & Fumagalli, L. (2014). PCR-based isolation of multigene families : lessons from the avian MHC class IIB. *Molecular Ecology Resources*, *14*, 778–788. https://doi.org/10.1111/1755-0998.12234

Burri, Reto, Hirzel, H. N., Salamin, N., Roulin, A., & Fumagalli, L. (2008). Evolutionary patterns of MHC class II B in owls and their implications for the understanding of avian MHC evolution. *Molecular Biology and Evolution*, *25*(6), 1180–1191. https://doi.org/10.1093/molbev/msn065

Callahan, B. J., Mcmurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2 : High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7). https://doi.org/10.1038/nmeth.3869

Chappell, P., Meziane, E. K., Harrison, M., Magiera, L., Hermann, C., Mears, L., … Kaufman, J. (2015). Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife*, *4*. https://doi.org/10.7554/eLife.05345

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., & McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, *47*(6), 682–688. https://doi.org/10.1038/ng.3257

Doherty, P. C., & Zinkernagel, R. M. (1975). Enhanced immunological surveillance in mice heterozygous at H-2 gene complex. *Nature*, *256*(5512), 50–52. https://doi.org/10.1038/256050a0

Drews, A., Strandh, M., Råberg, L., & Westerdahl, H. (2017). Expression and phylogenetic analyses reveal paralogous lineages of putatively classical and non-classical MHC-I genes in three sparrow species (Passer). *BMC Evolutionary Biology*, *17*(1). https://doi.org/10.1186/s12862-017-0970-7

Ejsmond, M. J., & Radwan, J. (2015). Red queen processes drive positive selection on major histocompatibility complex (MHC) genes. *PLoS Computational Biology*, *11*(11). https://doi.org/10.1371/journal.pcbi.1004627

Gaigher, A., Burri, R., Gharib, W. H., Taberlet, P., Roulin, A., & Fumagalli, L. (2016). Family-assisted inference of the genetic architecture of major histocompatibility complex variation. *Molecular Ecology Resources*, *16*(6), 1353–1364. https://doi.org/10.1111/1755-0998.12537

Gaigher, A., Roulin, A., Gharib, W. H., Taberlet, P., Burri, R., & Fumagalli, L. (2018). Lack of evidence for selection favouring MHC haplotypes that combine high functional diversity. *Heredity*. https://doi.org/10.1038/s41437-017-0047-9

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*. https://doi.org/10.1186/1471-2164-11-296

Goebel, J., Promerová, M., Bonadonna, F., McCoy, K. D., Serbielle, C., Strandh, M., … Fumagalli, L. (2017). 100 million years of multigene family evolution: Origin and evolution of the avian MHC class IIB. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3839-7

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic Acids Symposium Series No. 41* (pp. 95–98).

Hansson, B, Hasselquist, D., & Bensch, S. (2004). Do female great reed warblers seek extra-pair fertilizations to avoid inbreeding? *Proceedings of the Royal Society B*, *271*, S290–S292. https://doi.org/10.1098/rsbl.2004.0164

Hansson, Bengt, Åkesson, M., Slate, J., & Pemberton, J. M. (2005). Linkage mapping reveals sex-dimorphic map distances in a passerine bird. *Proceedings of the Royal Society B*, *272*(1578), 2289–2298. https://doi.org/10.1098/rspb.2005.3228

Hansson, Bengt, Sigeman, H., Stervander, M., Tarka, M., Ponnikas, S., Strandh, M., … Hasselquist, D. (2018). Contrasting results from GWAS and QTL mapping on wing length in great reed warblers. *Molecular Ecology Resources*, *18*(4), 867–876. https://doi.org/10.1111/1755-0998.12785

Hasselquist, D, Bensch, S., & von Schantz, T. (1995). Low frequency of extrapair paternity in the polygynous great reed warbler, Acrocephalus arundinaceus. *Behavioral Ecology*, *6*(1), 27–38. https://doi.org/10.1093/beheco/6.1.27

Hasselquist, D, Bensch, S., & von Schantz, T. (1996). Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler. *Nature*, *381*(6579), 229–232. https://doi.org/10.1038/381229a0

Hasselquist, Dennis. (1998). Polygyny in great reed warblers: A long-term study of factors contributing to male fitness. *Ecology*, *79*(7), 2376–2390. https://doi.org/10.1890/0012-9658(1998)079[2376:pigrwa]2.0.co;2

Hasselquist, Dennis, Montras-Janer, T., Tarka, M., & Hansson, B. (2017). Individual consistency of long-distance migration in a songbird: significant repeatability of autumn route, stopovers and wintering sites but not in timing of migration. *Journal of Avian Biology*, *48*(1), 91–102. https://doi.org/10.1111/jav.01292

Hollenbach, J. A., Thomson, G., Cao, K., Fernandez-Vina, M., Erlich, H. A., Bugawan, T. L., … Klitz, W. (2001). HLA diversity, differentiation, and haplotype evolution in mesoamerican natives. *Human Immunology*, *62*(4), 378–390. https://doi.org/10.1016/S0198-8859(01)00212-9

Huchard, E., Weill, M., Cowlishaw, G., Raymond, M., & Knapp, L. A. (2008). Polymorphism, haplotype composition, and selection in the Mhc-DRB of wild baboons. *Immunogenetics*, *60*(10), 585–598. https://doi.org/10.1007/s00251-008-0319-x

Hughes, A. L., & Nei, M. (1992). Maintenance of MHC polymorphism. *Nature*, *355*(6359), 402–403. https://doi.org/10.1038/355402b0

Kaufman, J. (1999). Co-evolving genes in MHC haplotypes: The “rule” for nonmammalian vertebrates? *Immunogenetics*, *50*(3–4), 228–236. https://doi.org/10.1007/s002510050597

Kaufman, J. (2018). Unfinished Business: Evolution of the MHC and the Adaptive Immune System of Jawed Vertebrates. *Annu. Rev. Immunol*, *36*, 383–409. https://doi.org/10.1146/annurev-immunol

Kelley, J., Walter, L., & Trowsdale, J. (2005). Comparative genomics of major histocompatibility complexes. *Immunogenetics*, *56*, 683–695. https://doi.org/DOI 10.1007/s00251-004-0717-7

Klein, J, Bontrop, R. E., Dawkins, R. L., Erlich, H. A., Gyllensten, U. B., Heise, E. R., … Watkins, D. I. (1990). Nomenclature for the major histocompatibility complexes of different species - a proposal. *Immunogenetics*, *31*(4), 217–219.

Klein, Jan, & Sato, A. (2000). The HLA system - First of two parts. *The New England Journal of Medicine*, *343*(10), 702–709. https://doi.org/10.1056/NEJM200009073431006

Lenz, T. L. (2011). Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution*, *65*(8), 2380–2390. https://doi.org/10.1111/j.1558-5646.2011.01288.x

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). *Molecular Ecology Resources*, *14*(4). https://doi.org/10.1111/1755-0998.12225

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Minias, P., Pikus, E., Whittingham, L. A., & Dunn, P. O. (2018). Evolution of copy number at the MHC varies across the avian tree of life. *Genome Biology and Evolution*, *11*(1), 17–28. https://doi.org/10.1093/gbe/evy253

Näpflin, K., O'Connor, E. A., Becks, L., Bensch, S., Ellis, V. A., Hafer-Hahmann, N., … Edwards, S. V. (2019). Genomics of hosts-pathogen interactions: challenges and opportunities across ecological and spatiotemporal scales. *PeerJ*, *7*(e8013), 1–37. https://doi.org/10.7717/peerj.8013

Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, *94*(15), 7799–7806. https://doi.org/10.1073/pnas.94.15.7799

Nei, Masatoshi, & Rooney, A. P. (2005). Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics*, *39*(1), 121–152. https://doi.org/10.1146/annurev.genet.39.073003.112240

Niskanen, A. K., Kennedy, L. J., Ruokonen, M., Kojola, I., Lohi, H., Isomursu, M., … Aspi, J. (2014). Balancing selection and heterozygote advantage in major histocompatibility complex loci of the bottlenecked Finnish wolf population. *Molecular Ecology*, *23*(4), 875–889. https://doi.org/10.1111/mec.12647

Nowak, M. A., Tarczyhornoch, K., & Austyn, J. M. (1992). The optimal number of major histocompatibility complex-molecules in an individual. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10896–10899. https://doi.org/10.1073/pnas.89.22.10896

O'Connor, E. A., Strandh, M., Hasselquist, D., Nilsson, J., & Westerdahl, H. (2016). The evolution of highly variable immunity genes across a passerine bird radiation. *Molecular Ecology*, *25*(4), 977–989. https://doi.org/10.1111/mec.13530

O'Connor, Emily A, Westerdahl, H., Burri, R., & Edwards, S. V. (2019). Avian MHC evolution in the era of genomics: Phase 1.0. *Cells*, *8*(1152), 1–21. https://doi.org/10.3390/cells8101152

Ohta, T. (1991). Multigene families and the evolution of complexity. *Journal of Molecular Evolution*, *33*(1), 34–41. https://doi.org/10.1007/BF02100193

Paul, S., Weiskopf, D., Angelo, M. A., Sidney, J., Peters, B., & Sette, A. (2013). HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *The Journal of Immunology*, *191*(12), 5831–5839. https://doi.org/10.4049/jimmunol.1302101

Penn, D., & Potts, W. (1999). The evolution of mating preferences and major histocompatibility complex genes. *American Naturalist*, *153*(2), 145–164. https://doi.org/10.1086/303166

Pierini, F., & Lenz, T. L. (2018). Divergent allele advantage at human MHC genes: Signatures of past and ongoing selection. *Molecular Biology and Evolution*, *35*(9), 2145–2158. https://doi.org/10.1093/molbev/msy116

Piertney, S. B., & Oliver, M. K. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity*, *96*(1), 7–21. https://doi.org/10.1038/sj.hdy.6800724

Promerová, M., Babik, W., Bryja, J., Albrecht, T., Stuglik, M., & Radwan, J. (2012). Evaluation of two approaches to genotyping major histocompatibility complex class I in a passerine-CE-SSCP and 454 pyrosequencing. *Molecular Ecology Resources*, *12*(2), 285–292. https://doi.org/10.1111/j.1755-0998.2011.03082.x

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rao, X., De Boer, R. J., van Baarle, D., Maiers, M., & Kesmir, C. (2013). Complementarity of binding motifs is a general property of HLA-A and HLA-B molecules and does not seem to effect HLA haplotype composition. *Frontiers in Immunology*, *4*(NOV), 1–6. https://doi.org/10.3389/fimmu.2013.00374

Roved, J. (2019). MHCtools: Analysis of MHC data in non-model species. CRAN. Retrieved from https://cran.r-project.org/package=MHCtools

Roved, J., Hansson, B., Stervander, M., Hasselquist, D., & Westerdahl, H. (2020). Data from: Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. Zenodo. https://doi.org/10.5281/zenodo.3716048

Roved, J., Hansson, B., Tarka, M., Hasselquist, D., & Westerdahl, H. (2018). Evidence for sexual conflict over MHC diversity in a wild songbird. *Proceedings of the Royal Society B*, *285*, 20180841. https://doi.org/10.1098/rspb.2018.0841

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). *Molecular cloning: a laboratory manual*. *Molecular cloning: a laboratory manual.* (2nd ed.). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sepil, I., Lachish, S., Hinks, A. E., & Sheldon, B. C. (2013). Mhc supertypes confer both qualitative and quantitative resistance to avian malaria infections in a wild bird population. *Proceedings of the Royal Society B*, *280*(1759). https://doi.org/10.1098/rspb.2013.0134

Sin, Y. W., Annavi, G., Dugdale, H. L., Newman, C., Burke, T., & MacDonald, D. W. (2014). Pathogen burden, co-infection and major histocompatibility complex variability in the European badger (Meles meles). *Molecular Ecology*, *23*(20). https://doi.org/10.1111/mec.12917

Spurgin, L. G., Van Oosterhout, C., Illera, J. C., Bridgett, S., Gharbi, K., Emerson, B. C., & Richardson, D. S. (2011). Gene conversion rapidly generates major histocompatibility complex diversity in recently founded bird populations. *Molecular Ecology*, *20*(24), 5213–5225. https://doi.org/10.1111/j.1365-294X.2011.05367.x

Stervander, M., Dierickx, E. G., Thorley, J., Brooke, M. d. L., & Westerdahl, H. (2020). High MHC gene copy number maintains diversity despite homozygosity in a Critically Endangered single-island endemic bird, but no evidence of MHC-based mate choice. *BioRxiv*. https://doi.org/10.1101/2020.02.03.932590

Strandh, M., Westerdahl, H., Pontarp, M., Canback, B., Dubois, M.-P., Miquel, C., … Bonadonna, F. (2012). Major histocompatibility complex class II compatibility, but not class I, predicts mate choice in a bird with highly developed olfaction. *Proceedings of the Royal Society B*, *279*(1746), 4457–4463. https://doi.org/10.1098/rspb.2012.1562

Stuglik, M. T., Radwan, J., & Babik, W. (2011). jMHC: Software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular Ecology Resources*, *11*(4), 739–742. https://doi.org/10.1111/j.1755-0998.2011.02997.x

Tarka, M., Akesson, M., Hasselquist, D., & Hansson, B. (2014). Intralocus sexual conflict over wing length in a wild migratory bird. *American Naturalist*, *183*(1), 62–73. https://doi.org/10.1086/674072

Testi, M., Battarra, M., Lucarelli, G., Isgro, A., Morrone, A., Akinyanju, O., … Sanchez-Mazas, A. (2015). HLA-A-B-C-DRB1-DQB1 phased haplotypes in 124 Nigerian families indicate extreme HLA diversity and low linkage disequilibrium in Central-West Africa. *Tissue Antigens*, *86*(4), 285–292. https://doi.org/10.1111/tan.12642

Trowsdale, J., & Knight, J. (2015). Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet*, (67), 301–323. https://doi.org/10.1146/annurev-genom-091212-153455.Major

Wakeland, E. K., Boehme, S., She, J. X., Lu, C. C., McIndoe, R. A., Cheng, I., … Potts, W. K. (1990). Ancestral polymorphisms of MHC class II genes: Divergent allele advantage. *Immunologic Research*, *9*(2), 115–122. https://doi.org/10.1007/bf02918202

Westerdahl, H., Waldenstrom, J., Hansson, B., Hasselquist, D., von Schantz, T., & Bensch, S. (2005). Associations between malaria and MHC genes in a migratory songbird. *Proceedings of the Royal Society B*, *272*(1571), 1511–1518. https://doi.org/10.1098/rspb.2005.3113

Westerdahl, H., Wittzell, H., & von Schantz, T. (1999). Polymorphism and transcription of Mhc class I genes in a passerine bird, the great reed warbler. *Immunogenetics*, *49*(3), 158–170. https://doi.org/10.1007/s002510050477

Westerdahl, H., Wittzell, H., von Schantz, T., & Bensch, S. (2004). MHC class I typing in a songbird with numerous loci and high polymorphism using motif-specific PCR and DGGE. *Heredity*, *92*(6), 534–542. https://doi.org/10.1038/sj.hdy.6800450

Whittingham, L. A., Dunn, P. O., Freeman-Gallant, C. R., Taff, C. C., & Johnson, J. A. (2018). Major histocompatibility complex variation and blood parasites in resident and migratory populations of the common yellowthroat. *Journal of Evolutionary Biology*, *31*(10), 1544–1557. https://doi.org/10.1111/jeb.13349

Woelfing, B., Traulsen, A., Milinski, M., & Boehm, T. (2009). Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philosophical Transactions of the Royal Society B-Biological Sciences*, *364*(1513), 117–128. https://doi.org/10.1098/rstb.2008.0174

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 Evolutionary Biology*, *10*. https://doi.org/10.1186/1471-2148-10-395

Data Accessibility

MHCtools version 1.2.1 (including user manual and documentation) is available at CRAN: https://cran.r-project.org/package= MHCtools. Our data set is available at the Zenodo repository: http://dx.doi.org/10.5281/ zenodo.3716048 (Roved, Hansson, Stervander, Hasselquist, & Westerdahl, 2020). DNA sequences are available at GenBank: https://ncbi.nlm.nih.gov (accession numbers: MH468831–MH469159; MT193762–MT193822).

Author contributions

JR designed the data analysis protocols and conceived, designed, and created the R package MHCtools; JR, BH, DH, and HW jointly conceived the study of MHC-I haplotypes in great reed warblers; MS constructed the amplicon sequencing libraries for Illumina sequencing; JR carried out bioinformatics, analyzed the data, and wrote the paper, with input from all authors.