Received: 12 March 2020 | Revised: 11 April 2022 | Accepted: 11 May 2022

DOI: 10.1111/1755-0998.13645

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

MHCtools – an R package for MHC high-throughput sequencing data: Genotyping, haplotype and supertype inference, and downstream genetic analyses in non-model organisms

Jacob Roved[1] | Bengt Hansson[2] | Martin Stervander[2,3] | Dennis Hasselquist[2] | Helena Westerdahl[2]

[1]GLOBE Institute, Section for Evolutionary Genomics, University of Copenhagen, Copenhagen K, Denmark

[2]Department of Biology, Molecular Ecology and Evolution Laboratory, Lund University, Lund, Sweden

[3]Department of Biology and Environmental Science, Faculty of Health and Life Sciences, Linnaeus University, Kalmar, Sweden

Correspondence
Jacob Roved, GLOBE Institute, Section for Evolutionary Genomics, University of Copenhagen, 1350 Copenhagen K, Denmark.
Email: jacob.roved@sund.ku.dk

Funding information
H2020 European Research Council, Grant/Award Number: 679799 and 742646; Kungliga Fysiografiska Sällskapet i Lund; Lund University, Department of Biology; Lunds Djurskyddsfond; Vetenskapsrådet, Grant/Award Number: 2015-05149, 2016-00689, 2016-04391, 2020-03976, 2020-04285, 349-2007-8690 and 621-2014-5222; Lund University; Swedish Research Council; Horizon 2020; European Union; European Research Council; Swedish Research Council

Handling Editor: Andrew DeWoody

Abstract

The major histocompatibility complex (MHC) plays a central role in the vertebrate adaptive immune system and has been of long-term interest in evolutionary biology. While several protocols have been developed for MHC genotyping, there is a lack of transparent and standardized tools for downstream analysis of MHC data. Here, we present the R package MHCTOOLS and demonstrate the use of its functions to (i) assist accurate MHC genotyping from high-throughput amplicon-sequencing data, (ii) infer functional MHC supertypes using bootstrapped clustering analysis, (iii) identify segregating MHC haplotypes from family data, and (iv) analyse functional and genetic distances between MHC sequences. We employed MHCTOOLS to analyse MHC class I (MHC-I) amplicon data of 559 great reed warblers (*Acrocephalus arundinaceus*). We identified 390 MHC-I alleles which clustered into 14 functional supertypes. A phylogenetic analysis and analyses of positive selection suggested that the MHC-I alleles belong to several distinct functional groups. We furthermore identified 107 segregating haplotypes among 116 families, and found substantial variation in diversity with 4–21 MHC-I alleles and 3–13 MHC-I supertypes per haplotype. Finally, we show that the great reed warbler haplotypes harboured combinations of MHC-I supertypes with greater functional divergence than observed in simulated populations of possible haplotypes, a result that is in accordance with the divergent allele advantage hypothesis. Our study demonstrates the power of MHCTOOLS to support genotyping and analysis of MHC in non-model species, which we hope will encourage broad implementation among researchers in MHC genetics and evolution.

KEYWORDS
cluster analysis, divergent allele advantage, functional divergence, major histocompatibility complex, MHC haplotypes, MHC supertypes

Dennis Hasselquist and Helena Westerdahl Shared last author.

1 | INTRODUCTION

The major histocompatibility complex (MHC) is a multigene family that plays a vital role in the vertebrate adaptive immune system (Klein & Sato, 2000). Ongoing coevolution with pathogens has caused MHC genes to exhibit both high levels of genetic diversity and preservation of polymorphisms over evolutionary time; consequently, these genes have attracted broad interest in studies of adaptive genetic variation (Ejsmond & Radwan, 2015; Kaufman, 2018; Klein et al., 2007; Piertney & Oliver, 2006). To assist studies on MHC genes, we developed the R package MHCTOOLS, that contains 12 tools for analysis of MHC data (Table 1). The core functions in MHCTOOLS are focused on two fields of data analysis that are prominent in contemporary MHC research: (i) analysing variation in functional properties between sequences, and (ii) analysing how MHC alleles are inherited on segregating haplotypes. In addition, MHCTOOLS offers some useful tools that facilitate the bioinformatics involved in MHC genotyping using amplicon sequencing data.

The function of classical MHC molecules is to present peptides to T cells, which is a decisive step in the initiation of adaptive immune responses. In MHC research, much attention has been focused on the functional divergence of alleles in the parts that encode the peptide binding region of the MHC molecule. Functional divergence between MHC alleles has been analysed using both quantitative and qualitative approaches. The quantitative approach is based on the fundamental principle in the "divergent allele advantage" hypothesis (DAA), that genotypes that combine more divergent MHC alleles enable hosts to mount adaptive immune responses against a more diverse array of pathogens (Wakeland et al., 1990). In silico studies have demonstrated support for the DAA using computational binding predictions of pathogen peptides for combinations of human leucocyte antigen alleles (Lenz, 2011; Pierini & Lenz, 2018). Divergence between MHC alleles has been quantified based on for example, the proportion of varying codons (p-distance) (Lenz et al., 2013), tree-based distances (UniFrac) (Leclaire et al., 2017), and physicochemical properties of amino acids (Grantham or Sandberg distance) (Pierini & Lenz, 2018). The qualitative approach has focused on identifying groups of MHC alleles that share functional properties, commonly referred to as MHC supertypes (Sidney et al., 1996, 2008). MHC supertypes may be regarded as balanced polymorphisms that persist over evolutionary time, despite the potential gain

TABLE 1 Overview of the functions included in MHCTOOLS v. 1.4.2

BootKmeans	A wrapper for the kmeans function of the STATS package in R, allowing for greatly improved confidence in estimated clusters. BootKmeans performs multiple runs of kmeans while estimating optimal k-values based on a set threshold for stepwise reduction in BIC
ClusterMatch	Performs an evaluation of the extent to which different kmeans clustering models identify similar clusters and summarizes bootstrap model stats as means for different estimated values of k. ClusterMatch is designed to take files produced by BootKmeans as input, but other data can be analysed if the descriptions of the required data formats are observed
CreateFas	Creates a FASTA file with all sequences from a DADA2 sequence table
CreateSamplesFas	Creates a set of FASTA files with the sequences present in each sample in a DADA2 sequence table
DistCalc	Calculation of Grantham distances (Grantham, 1974), Sandberg distances (Sandberg et al., 1998), or p-distances (proportion of varying nucleotide or amino acid codons) in pairwise comparisons of sequences. When calculating Sandberg distances, the function additionally outputs five tables with physicochemical z-descriptor values (Sandberg et al., 1998), which can be used for inference of MHC supertypes
HpltFind	Automatically infers major histocompatibility complex (MHC) haplotypes from genotypes of parents and offspring in families. The functions GetHpltTable and GetHpltStats provide evaluation of the output files.
PapaDiv	Calculation of joint 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 (if haplotype analysis is not feasible)
ReplMatch	Automatically compares technical replicates in an amplicon sequencing data set and reports mismatches. The functions **GetReplTable** and **GetReplStats** provide evaluation of the output files

and loss of individual MHC alleles that they comprise (Lighten et al., 2017; Richman, 2000). Several methods have been employed for inference of MHC supertypes. In humans, where the structure of the folded MHC molecules and their peptide binding properties have been described, methods for MHC supertype inference take such detailed information into account (e.g., Hertz & Yanover, 2007; Lund et al., 2004). In studies of organisms where structural information is limited, MHC supertypes are commonly inferred from the physicochemical properties of the amino acid sequences by nonhierarchical clustering analysis combined with a discriminant analysis on principal components (DAPC) (Buczek et al., 2016; Gonzalez-Quevedo et al., 2014; Lighten et al., 2017; Lillie et al., 2015; Sepil et al., 2012; Trujillo et al., 2021; Winternitz et al., 2015). DAPC can be performed with existing R packages such as ADEGENET (Jombart, 2008), but it is associated with several assumptions that MHC data may not meet (in particular absence of multivariate outliers and, with small or unequal sample sizes, multivariate normality and homogeneity of variances) (Tabachnick & Fidell, 2014). Furthermore, a recent evaluation criticized a priori specification of clusters in DAPC and called for a standardized reporting of how a priori clusters are inferred, including how optimal numbers of clusters are detected (Miller et al., 2020). To improve MHC supertype inference and facilitate transparent reporting, we developed the functions BootKmeans and ClusterMatch in MHCTOOLS, which offer bootstrapped nonhierarchical clustering analysis and quantitative evaluation of the results to identify an optimal clustering model. BootKmeans and ClusterMatch work as a standalone method for MHC supertype inference, but outputs may also subsequently be employed in DAPC.

In MHC studies, the heterozygote advantage hypothesis describes the principle that individuals with 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, because the MHC often spans multiple paralogous loci, MHC molecules may be encoded by MHC alleles harboured either on the same or on different multilocus haplotypes. MHC alleles have been found to be nonrandomly associated within haplotypes, and disease associations suggest that selection acts on combined multilocus effects that span entire MHC haplotypes (Buhler et al., 2016; Huchard et al., 2008; Kaufman, 1999; Rioux et al., 2009). If the principles behind the heterozygote advantage and the DAA are applied to multilocus MHC haplotypes, a derived hypothesis emerges: haplotypes that combine a larger number of MHC alleles and/or more divergent MHC alleles will be favoured by natural selection, because they confer an advantage in terms of binding antigens from a broader range of pathogens (Gaigher et al., 2018). Because MHC supertypes represent clusters of alleles that share similar functional properties, this derived hypothesis may also be extended to the level of MHC supertypes.

Detailed knowledge about MHC haplotype structure is mostly limited to humans and a few model organisms, but there is a growing interest in characterizing MHC haplotypes and investigating their effects also in wild non-model species (Gaigher et al., 2016, 2018; Okano et al., 2020; Stervander et al., 2020). For example, Gaigher et al. (2016) inferred MHC haplotypes from the segregation patterns of MHC alleles within families of barn owls (*Tyto alba*), thereby obtaining information about linkage and recombination between alleles, the number of MHC gene copies, and presence of gene copy number variation. In a follow-up study, they used these haplotype data to investigate nonrandom associations of MHC alleles in haplotypes (Gaigher et al., 2018). The studies by Gaigher et al. (2016, 2018) demonstrate the value of family-assisted haplotype inference in MHC studies within evolutionary biology and ecology. To assist such studies, MHCTOOLS includes the function HpltFind which is designed to automatically infer MHC haplotypes from genotypes of parents and offspring. MHCTOOLS additionally includes the functions GetHpltTable and GetHpltStats for posthoc evaluation of the haplotype inference.

Since the advent of high-throughput DNA sequencing, MHC genotyping in non-model organisms is often carried out using PCR-based amplicon sequencing (Biedrzycka et al., 2017; Burri et al., 2014; Lighten, et al., 2014; Promerová et al., 2012; Stervander et al., 2020; Zagalska-Neubauer et al., 2010). However, in many species, amplification of specific MHC loci is impeded by sequence similarity across loci, and it is often necessary to coamplify multiple MHC loci (Alcaide et al., 2013; Burri et al., 2014). While this technique is useful for estimating the overall MHC genetic diversity, the resulting data contain no information about linkage or spatial organization of the alleles (Alcaide et al., 2013; Biedrzycka 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 alleles detected in each sample, and associating alleles with specific loci becomes difficult, in particular in species with highly duplicated MHC genes (Lighten et al., 2014). This lack of resolution of the MHC diversity severely challenges contemporary studies of MHC in evolutionary ecology (Gaigher et al., 2016; O'Connor et al., 2019). The use of family data to infer segregating MHC haplotypes may be useful towards overcoming this challenge by supplying information about linkage of alleles (Gaigher et al., 2016, 2018; Okano et al., 2020).

In this study, we employed MHCTOOLS to analyse an MHC class I (MHC-I) amplicon sequencing data set from a wild population of great reed warblers (*Acrocephalus arundinaceus*), a songbird with highly duplicated MHC genes (Roved et al., 2018; Westerdahl et al., 2004). We demonstrated the use of the BootKmeans and ClusterMatch functions for inference of MHC-I supertypes and the function HpltFind for inferring segregating MHC-I haplotypes. We subsequently employed the generated MHC-I supertype and haplotype data to (i) characterize the structure and properties of MHC-I haplotypes, (ii) analyse the functional relationships between MHC-I supertypes, and (iii) analyse how MHC-I supertypes associate on haplotypes. Finally, we investigated whether natural selection has favoured MHC-I haplotypes that harbour high levels of functional divergence. Specifically, we tested the null hypothesis that the observed diversity of MHC-I supertypes in haplotypes is similar to the diversity expected from simulated haplotypes to which MHC-I alleles are randomly assigned. We also tested the null hypotheses that the divergence and degree of overlap between MHC-I supertypes in haplotypes are similar to those expected from simulated haplotypes.

2 | MATERIALS AND METHODS

2.1 | Data set

We used data on 141 adult males, 131 adult females, and 287 offspring from our long-term study population of great reed warblers at Lake Kvismaren in Sweden (Bensch et al., 1998; Hasselquist, 1998; Roved et al., 2018). The adults in our data set were breeding between 1984 and 2004, and the offspring constitute the 1998 and 1999 cohorts, with addition of one family each from 1992 and 1996. Paternity and maternity of all offspring were verified by molecular methods (Hansson et al., 2004; Hasselquist et al., 1995). Fieldwork and DNA sampling were approved by the Malmö/Lund Animal Ethics Committee and the Swedish Bird Ringing Centre.

2.2 | DNA sampling and sequencing

We extracted DNA following Roved et al. (2018) and amplified a 262-bp region of MHC-I exon 3 using the primers HNalla and 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 offspring) were sequenced in a Roche 454 GS FLX (Hoffmann-La Roche), see Roved et al. (2018). The remaining amplicons were sequenced using 300-bp paired-end sequencing in an Illumina MiSeq (Illumina Inc.; see Supporting Information Methods for details). Samples from 23 adult males, 32 adult females, and 150 offspring (including 11 replicates from the 454-sequencing experiment) were included in a first run and a smaller batch with samples from 30 adult males were added in a second run.

2.3 | Filtering of 454 data

The 454-sequencing data were demultiplexed using JMHC (Stuglik et al., 2011). Sequences with <3 reads (across all amplicons) and amplicons with <240 reads were removed. The data were then filtered according to Galan et al. (2010), but additionally optimizing the relative abundance filtering threshold by evaluating the proportion of mismatching sequences between technical replicates (relative abundance = no. reads per sequence / total no. reads per amplicon). In this step, sequences with relative abundance <0.012 were removed. Subsequently, the data were manually screened to remove artificial sequences (i.e., PCR chimeras and single nucleotide substitution errors occurring with lower abundance than parent sequences) and nonfunctional sequence variants (i.e., sequences containing stop codons or indels obstructing the reading frame) (Roved et al., 2018).

2.4 | Filtering of Illumina MiSeq data

The Illumina sequencing outputs were trimmed to remove adapters, primers, and tag sequences using the software CUTADAPT version 1.14 (Martin, 2011). The trimmed sequences were filtered using DADA2 version 1.4.0 (Callahan et al., 2016) in R version 3.4.2 (R Core Team, 2017). Chimeras were removed using removeBimeraDenovo in DADA2.

2.5 | Optimization of filtering settings using MHCtools

DADA2 has a sample inference step that removes PCR errors by employing a machine learning algorithm to cluster low-abundance sequence variants with higher-abundance relatives. The sample inference is sensitive to the accuracy of the input sequencing data, which is adjusted by setting a truncation parameter (truncLen or truncQ) and a threshold of expected error rates (maxEE fw/rv) in DADA2's filterAndTrim function. The truncLen filter truncates reads by a set length, while (alternatively) truncQ truncates reads at the first base that has a Phred quality score (Q score) below a threshold value. maxEE removes reads with an expected error rate greater than a threshold value. Because these parameters affect the sample inference, they also affect the final output from DADA2. Appropriate filterAndTrim settings may be unpredictable for new data sets, and it is therefore beneficial to optimize the settings by evaluating the repeatability of the sample inference. To facilitate such evaluation, we developed the function ReplMatch which compares technical replicates in a sequencing data set and reports mismatching sequences. The supporting functions GetReplTable and GetReplStats provide summaries of the output from ReplMatch.

2.6 | Illumina sequencing output

We optimized the filterAndTrim settings in DADA2 by comparing 25 sets of genotype replicates in filtering runs with different settings. For each run, we employed ReplMatch in MHCTOOLS to calculate the mean proportion of mismatching sequences among replicates. We evaluated ranges of truncQ from 18 to 30 and maxEE (fw and rv) from 0.05 to 2. The optimal truncQ setting (i.e., providing the lowest mean proportion of mismatching sequences) was 20, while maxEE settings produced optima at 0.05 and 0.1 (Figures S1a–b). With MaxEE <0.05, too few forward and reverse reads passed the filtering step to be successfully merged, thus obstructing the sample inference (data not shown). As we conducted manual inspection of our sequences after filtering in DADA2, we proceeded with the less restrictive maxEE = 0.1. Paired reads that did not match exactly in the overlapping region were removed from the data set.

With the settings truncQ = 20 and maxEE (fw and rv) = 0.1, DADA2 inferred 295 and 200 unique sequence variants in our first and second Illumina data sets, respectively. We inspected these manually in BIOEDIT version 7.2.5 (Hall, 1999) to remove nonfunctional variants. To obtain maximum repeatability, we added a final step of filtering the remaining sequence variants by their relative abundance within each amplicon (cf. Biedrzycka et al., 2017), again with optimization of the filtering threshold. We evaluated the mean proportion of mismatching sequences obtained with no relative abundance threshold and with 14 threshold values ranging from 0.005 to 0.03 using ReplMatch. The threshold that produced the lowest mean proportion of mismatching variants was 0.014 (Figure S1c).

After filtering and screening, our first and second Illumina data set, respectively, contained 226 and 162 alleles in 205 and 31 samples (in total 277 alleles, with 111 occurring in both data sets). Number of reads per sample were normally distributed with mean = 20,920 (min = 12,532; max = 32,781) and mean = 22,492 (min = 12,062; max = 38,521), respectively. Two samples with low read numbers (0 and 3150 reads) were removed from the second Illumina data set.

2.7 | Comparing and merging 454 and Illumina outputs

We collated the Illumina and 454-sequencing data sets for downstream analyses. We found 216 alleles that occurred in both the Illumina and the 454-sequencing data sets, 61 alleles that were unique to the Illumina data sets, and 113 unique to the 454-sequencing data set. The larger number of alleles unique to the 454-sequencing data set was expected, because many samples in the Illumina data sets were related to samples in the 454 data set (e.g., offspring of which one or both parents were genotyped in the 454 data set).

Two genotyped samples were excluded from downstream analyses due to labelling errors. All alleles were blasted against the NCBI nr/nt database using BLASTN (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and novel alleles named following MHC standardized nomenclature (Klein et al., 1990).

2.8 | Repeatability assessment

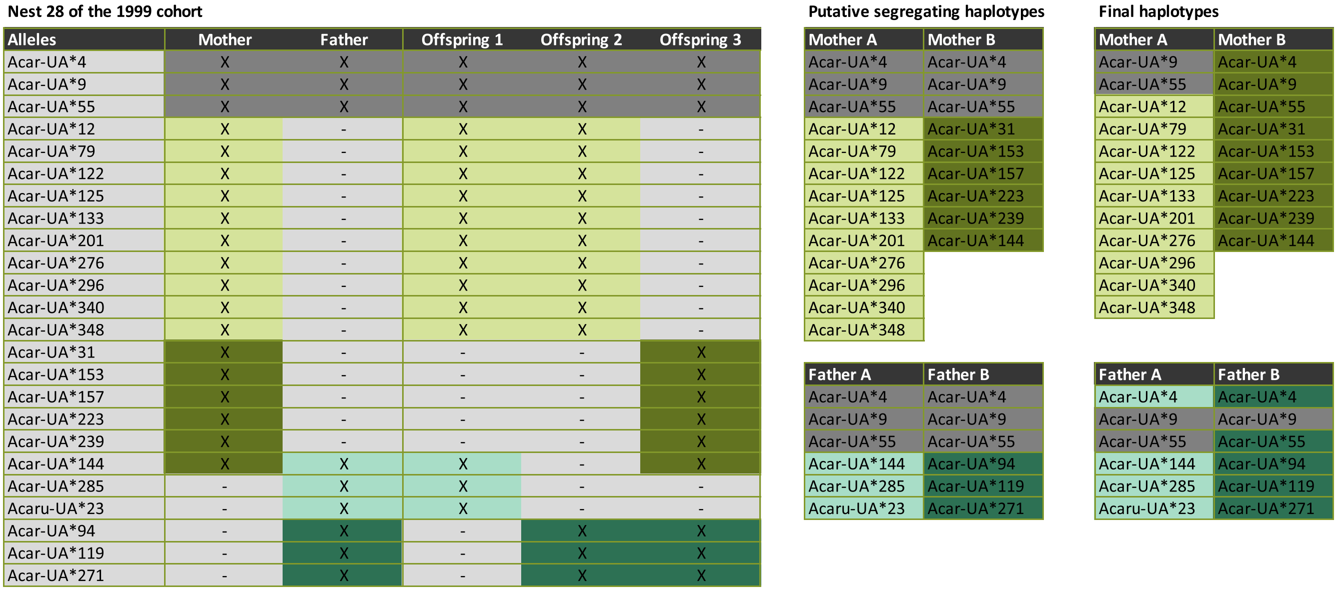
For details on technical replicates, see Technical replicates & filtering in Supporting Information Methods. We calculated the repeatability of our sequencing experiments as 1 minus the mean across all replicate sets of the mean proportion of mismatching sequence variants within each replicate set. Comparisons between replicate sets were performed using ReplMatch and repeatabilities calculated using GetReplStats.

2.9 | Haplotype inference using MHCtools

We employed the HpltFind function in MHCTOOLS to infer MHC-I haplotypes in our data set. HpltFind is designed to automatically infer haplotypes by analysing the segregation of individual alleles from parents to offspring, as illustrated in Figure 1. The function requires a table specifying the occurrence of alleles in each individual and a table associating individuals with families. It assigns alleles to a putative haplotype if they occur in either parent and in one or more offspring and produces a set of lists that specify the assignment of alleles to putative haplotypes. Alleles with problematic segregation patterns are indicated in the output, including (i) alleles that cannot be resolved on haplotypes because they occur in both parents and (ii) alleles that are missing in samples in which they would be expected to occur (incongruent alleles; e.g., an allele that is observed in offspring but in neither parent, or an allele that occurs only in some offspring, despite all offspring sharing the same putative haplotype containing the allele). The functions GetHpltTable and GetHpltStats provide posthoc summaries of the proportion of incongruent alleles for each family and across all families, respectively.

HpltFind offers a transparent automated method for analysis of allele segregation in a large number of families—a process that may require significant effort to perform manually. However, MHC genotyping from non-model species using degenerate primers may result in coamplification of MHC loci that carry common MHC alleles (i.e., alleles that are found in almost all individuals), which can give rise to increased numbers of unresolved alleles in haplotype analyses. Among the 390 MHC-I alleles that we observed in our data set, the five most common were present in 98%, 97%, 87%, 83%, and 50% of the samples, respectively. Those alleles were often present in both parents in families, and it was therefore difficult to resolve their presence in haplotypes through analysis of segregation patterns in single families. Such a challenge can potentially be solved by extending the analysis of segregation patterns across multiple generations, e.g. tracing how alleles segregated from great grandparents to offspring. Thereby, a haplotype can be observed segregating from several other haplotypes, which increases the likelihood that segregation patterns of common alleles can be resolved. The availability of a detailed pedigree of our great reed warbler study population (Hansson et al., 2005) allowed us to carry out such an analysis across multiple generations.

We initially employed HpltFind to analyse MHC-I allele segregation patterns in 67 families from the 1998 and 1999 cohorts (in total 78 parents and 282 offspring), one family from 1996 (2 parents and 3 offspring), and one from 1991 (2 parents and 5 offspring). Among 26 parents from those 69 families, we were able to trace ancestry up to five generations back (Table S1). For the remaining parents, no pedigree data were available. We investigated allele segregation patterns in 50 ancestral families of the 26 parents, that we traced in our pedigree, using HpltFind. These analyses included 80 additional individuals that were ancestors to the 26 parents. We subsequently compared putative haplotypes vertically within lines of ancestry and laterally between concurrent families using a stepwise protocol, which is described in Supporting Information Methods and Figure S2. This procedure allowed us to reduce the number of putative haplotypes in our data set by solving a number of unresolved allele assignments and observations of incongruent alleles (Table S2). We were unable to analyse the segregation of alleles in three



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

FIGURE 1 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 colours. Dark grey colour 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 (see Figure S2)

ancestral families. The segregation patterns suggested that blood samples from two individuals in these families had been mislabelled, and these samples and families were excluded from downstream analyses.

2.10 | Estimating the recombination rate

In two families, the MHC-I allele segregation patterns indicated that recombination had taken place between parental haplotypes (Figures S3 and S4). The recombinant haplotypes were Acar-HPLT*72 (recombinant from Acar-HPLT*15 and Acar-HPLT*21) and Acar-HPLT*73 (recombinant from Acar-HPLT*14 and Acar-HPLT*27). Based on this observation, we estimated the recombination rate as the number of recombinant haplotypes divided by the total number of gametes (i.e., two times the number of offspring in families for which we successfully inferred haplotypes).

2.11 | MHC-I supertype inference using MHCtools

We conducted a phylogenetic analysis of our MHC-I alleles using PHYML version 3.1 (Guindon et al., 2010; Guindon & Gascuel, 2003), and tested for positive selection using CODEML from the PAML software package (Yang, 1997, 2007) with Bayes Empirical Bayes (BEB) analysis (Yang et al., 2005) to identify codons that showed evidence of positive selection (see Supporting Information Methods for details). We then applied the DistCalc function in MHCTOOLS on an alignment of our MHC-I alleles to extract values of five physicochemical z-descriptors (Sandberg et al., 1998) for the amino acids in 14 positively selected codons. Based on the z-descriptor values, we employed the BootKmeans function in MHCTOOLS to identify MHC-I supertypes using bootstrapped k-means clustering. BootKmeans runs sets of k-means clustering models while evaluating the incremental reduction in Bayesian information criterion (ΔBIC) for increasing values of the number of clusters (k). In our analysis, each set of models evaluated k-values from 1 to 40, and we set BootKmeans to estimate the number of clusters (k_{est}) as the value of k, for which ΔBIC was <1% of the largest ΔBIC observed in each set. This procedure is comparable to visually inspecting an elbow plot of BIC versus k values as illustrated in Figure S5. In each set of models, BootKmeans output the set of clusters inferred by the k-means model that estimated k_{est} clusters and statistics including total within-cluster sums of squares, total between-cluster sums of squares, AIC, and BIC. Based on observations across all k-means models in each set (i.e., representing one scan of k-values from 1 to 40), BootKmeans also recorded BIC_{min}, BIC_{max}, ΔBIC (i.e., BIC_{max} minus BIC for the model estimating k_{est} clusters), ΔBIC divided by BIC_{max}, and ΔBIC divided by k_{est}. We set BootKmeans to run 1000 sets of k-means clustering models on our data set.

Following the bootstrapped k-means clustering, we employed the ClusterMatch function to quantify to which extent k-means models that found equal k_{est}-values inferred similar clusters of MHC-I alleles. For each value of k_{est} among the bootstrapped k-estimation scans, ClusterMatch conducts pairwise comparisons between all models that found k_{est} clusters (i.e., the selected model from each scan). In each pairwise comparison, the number of allele assignments that fall outside the k_{est} most abundant clusters is recorded (illustrated in Figure S6) and the proportion of such assignments out of the total number of allele assignments to clusters is calculated. Finally, means of these values across all pairwise comparisons for each value of k_{est} are calculated. In addition, ClusterMatch summarizes the number of k-estimation scans that estimated k_{est} clusters and calculates means of total within-sums of squares, AIC, BIC, $\Delta BIC/BIC_{max}$, and $\Delta BIC/k_{est}$ for each value of k_{est}. As our final estimated number of clusters, we selected the value of k_{est} that was associated with the lowest mean proportion of allele assignments falling outside the k_{est} most abundant clusters. Among the k-means clustering models that produced the final estimated number of clusters, we selected the models that had the smallest residual BIC, based on the rationale that these combined the MHC-I alleles into the most informative clusters. These were six models, all of which inferred identical clusters of MHC-I alleles. We used the inferred clusters of MHC-I alleles in these selected models as the definition of our MHC-I supertypes, which we named Acar-ST*1 to Acar-ST*14. We inferred the centroid of each MHC-I supertype cluster by calculating the arithmetic mean of the z-descriptors of its constituent alleles.

We repeated the analysis of positive selection on aligned subsets of alleles corresponding to each of the 14 supertypes. We first built phylogenetic trees for each subset using the GTR substitution model with the same settings as for the tree that included all alleles (Supporting Information Methods). These trees were used as input for CODEML along with the alignments of each subset of alleles. The analyses followed the protocol described in Supporting Information Methods with respect to CODEML settings, nested site models, and likelihood ratio tests. For supertypes that showed evidence of positive selection, we inferred positively selected codons by BEB analysis.

2.12 | Functional divergence within and between MHC-I supertypes

To analyse the functional divergence between MHC-I alleles within MHC-I supertypes, we employed the DistCalc function in MHCTOOLS to calculate the means of pairwise amino acid distances between the alleles in each supertype. We quantified functional divergence by three measures of amino acid distance: Grantham distance (Grantham, 1974), p-distance (i.e., the proportion of variable codons), and Sandberg distance (Sandberg et al., 1998). The distances were calculated for the 14 codons that were inferred to have evolved under positive selection. We subsequently evaluated the distributions of the mean Grantham distance, p-distance, and Sandberg distance values within MHC-I supertypes, and compared the three distance measures using Pearson's correlation tests.

To analyse the functional divergence between MHC-I supertypes, we calculated pairwise Sandberg distances between supertype centroids as means of the Euclidian distances between all sets of z1–z5 descriptors in each centroid pair. Furthermore, we calculated the pairwise overlap between MHC-I supertypes as the sum of the mean Sandberg distances between the alleles in each supertype minus the Sandberg distance between the centroids. We used the qgraph function in the R package QGRAPH v. 1.9.2 to create a network visualization based on the pairwise Sandberg distances between supertype centroids.

2.13 | MHC-I supertypes on haplotypes

Having defined the MHC-I supertypes, we proceeded to investigate (i) how supertypes were distributed on segregating haplotypes (i.e., which supertypes, and how many MHC-I alleles representing each, that were observed on each haplotype) and (ii) the distribution of haplotypes that each supertype was represented on. For each MHC-I supertype, we calculated the variance of the number of alleles representing it on each haplotype. We used the leveneTest function in the R package CAR v. 3.0.11 to test for homogeneity of these variances. Furthermore, we used Pearson's correlations to test the association between the number of MHC-I supertypes and the number of different MHC-I alleles observed on each haplotype.

2.14 | Functional divergence in haplotypes

We analysed the functional divergence between MHC-I alleles within haplotypes by calculating the means of the pairwise centroid distances and the pairwise overlaps between the positively selected MHC-I supertypes represented in each haplotype. Haplotypes carrying less than two positively selected MHC-I supertypes were excluded from this analysis.

2.15 | Data simulations

We generated 10,000 in silico simulations of our haplotype data set by randomly assigning alleles from positively selected MHC-I supertypes to haplotypes, while maintaining the number of different alleles (i.e., from positively selected MHC-I supertypes) for each haplotype. We inferred which positively selected MHC-I supertypes were represented in each simulated haplotype and, for haplotypes carrying at least two positively selected MHC-I supertypes, calculated means of the pairwise centroid distances and the pairwise overlaps between these. For each simulation, we then compared the number of positively selected MHC-I supertypes observed in the real haplotypes to the values observed in the simulated haplotypes using paired *t* tests. A two-sided *p*-value was calculated as two times the proportion of simulations where $t \leq 0$. Similarly, we also compared the mean centroid distances and mean overlap between the positively selected MHC-I supertypes observed in the real haplotypes to the values observed in the simulated haplotypes. One-sided *p*-values were calculated as the proportion of simulations where $t \leq 0$ for the comparisons of mean centroid distances, or $t \geq 0$ for the comparisons of mean overlaps between supertypes. The data simulations and *t* tests were carried out in R version 4.1.0 (R Core Team, 2021).

3 | RESULTS

3.1 | MHC-I genotyping assisted by MHCtools

We genotyped 262 bp of the MHC-I exon 3 in 559 great reed warblers using amplicon sequencing on the Illumina MiSeq and Roche 454 platforms; see Materials and Methods. The data from our Illumina MiSeq sequencing runs were filtered using DADA2 in combination with a per amplicon frequency threshold and with optimization of filtering settings using ReplMatch and GetReplStats from MHCTOOLS. Our first Illumina data set achieved a repeatability of 0.998, across 52 samples in 25 replicate sets, and a smaller batch of samples in a second Illumina run showed perfect agreement between two replicates. In comparison, the repeatability among 50 sets of technical replicates in our 454 data set was 0.94 after filtering (Roved et al., 2018). The repeatability among 11 replicated samples between the Illumina and Roche 454 data sets was 0.96. When combining the Illumina and Roche 454 data sets, we identified 390 alleles, of which 324 were unique at the amino acid sequence level. The number of different MHC-I alleles per individual ranged from six to 26 (mean = 14.3, SD = 3.42; Figure S7).

3.2 | Haplotype inference

We employed HpltFind from MHCTOOLS to infer MHC-I haplotypes based on allele segregation patterns in 116 great reed warbler families. We initially identified 225 putative MHC-I haplotypes with a mean proportion of unresolved allele assignments of 0.446 (Table S2). By comparing the putative haplotypes vertically within lines of ancestry and laterally between concurrent families in a stepwise protocol (Figure S2), we solved several unresolved allele assignments and observations of incongruent alleles (i.e., alleles missing in or erroneously assigned to haplotypes; Table S2). This reduced the final number of MHC-I haplotypes to 107, with a mean proportion of unresolved allele assignments of 0.255 (Supporting Information haplotype tables). The mean proportion of unresolved alleles was 0.199 in haplotypes that could be observed in multiple families, and 0.327 in haplotypes observed only in single families.

In the haplotype inference process, we discovered and removed 15 sequencing errors from individual samples in our data set, corresponding to a proportion of 0.0019 of the total number of allele assignments. We inferred 430 putative null alleles (corresponding to a proportion of 0.051 of 8007 allele assignments), which we subsequently added to individual samples (Table S2). These null alleles were known alleles in the data set that produced false negatives in some samples (most probably by amplifying inconsistently during PCR, which increases the risk of allelic dropout, especially with the 454 sequencing technology). The proportion of null alleles is comparable to the expected error rate given the repeatability of 0.96 between the 454 and Illumina sequencing platforms.

We found considerable variation in the number of MHC-I gene copies among haplotypes, with between four and 21 different alleles in single haplotypes (mean = 9.2, SD = 2.80; Figure S8). We found two recombinant haplotypes among the 334 offspring in the 116 families. From this observation, we estimated a recombination rate of 0.0030 within the MHC-I in great reed warblers, corresponding to a distance of 0.3 centimorgan (cM).

3.3 | Phylogenetic analysis

In the phylogenetic tree produced with the GTR model (Figure 2), we identified five tentative groups based on the tree topology and SH-aLRT support values (Supporting Information Methods). These groups are indicated in the phylogenetic tree (by letters A–E) and specified in Table S4.

3.4 | Selection analysis

We tested two sets of nested models in CODEML (M2 vs. M1 and M8 vs. M7) using all 390 MHC-I alleles in our data set. The M2 and M8 models, which allowed for positive selection, fit the data significantly better than the M1 and M7 models ($p < .0001$ in both model comparisons; Table S5a). The M8 model had the largest likelihood value, and in this model 14 out of 87 amino acid codons were estimated to be under positive selection (mean dN/dS = 3.44; Figure S9; Table S5a). The positively selected sites predicted by BEB analysis and associated dN/dS- and p-values are shown in Table S6a.

3.5 | MHC-I supertype inference

Using DistCalc in MHCTOOLS, we extracted five physicochemical z-descriptors for the 14 codons of our MHC-I alleles that were inferred to be under positive selection. We then employed BootKmeans to run 1000 sets of k-means clustering models on the z-descriptor values. BootKmeans estimated between 13 and 23 different clusters, and we employed ClusterMatch to evaluate the agreement between inferred clusters among the bootstrapped models. The proportion of allele assignments to low-ranking clusters ranged from 0.032 for models that inferred 14 clusters to 0.083 for models that inferred 18 or 19 clusters (Table S7). Among the models that inferred 14 clusters, six models shared the smallest residual BIC. These models inferred identical clusters, and we used these as the final definition of our MHC-I supertypes. The categorization of MHC-I alleles into MHC-I supertypes is specified in Table S8a–n and the number of alleles associated with each supertype summarized in Table S9 and Figure S10.

3.6 | Functional divergence within and between MHC-I supertypes

The mean Grantham distance within each MHC-I supertype ranged from 0.31 to 32.44 with an average of 19.95, the mean

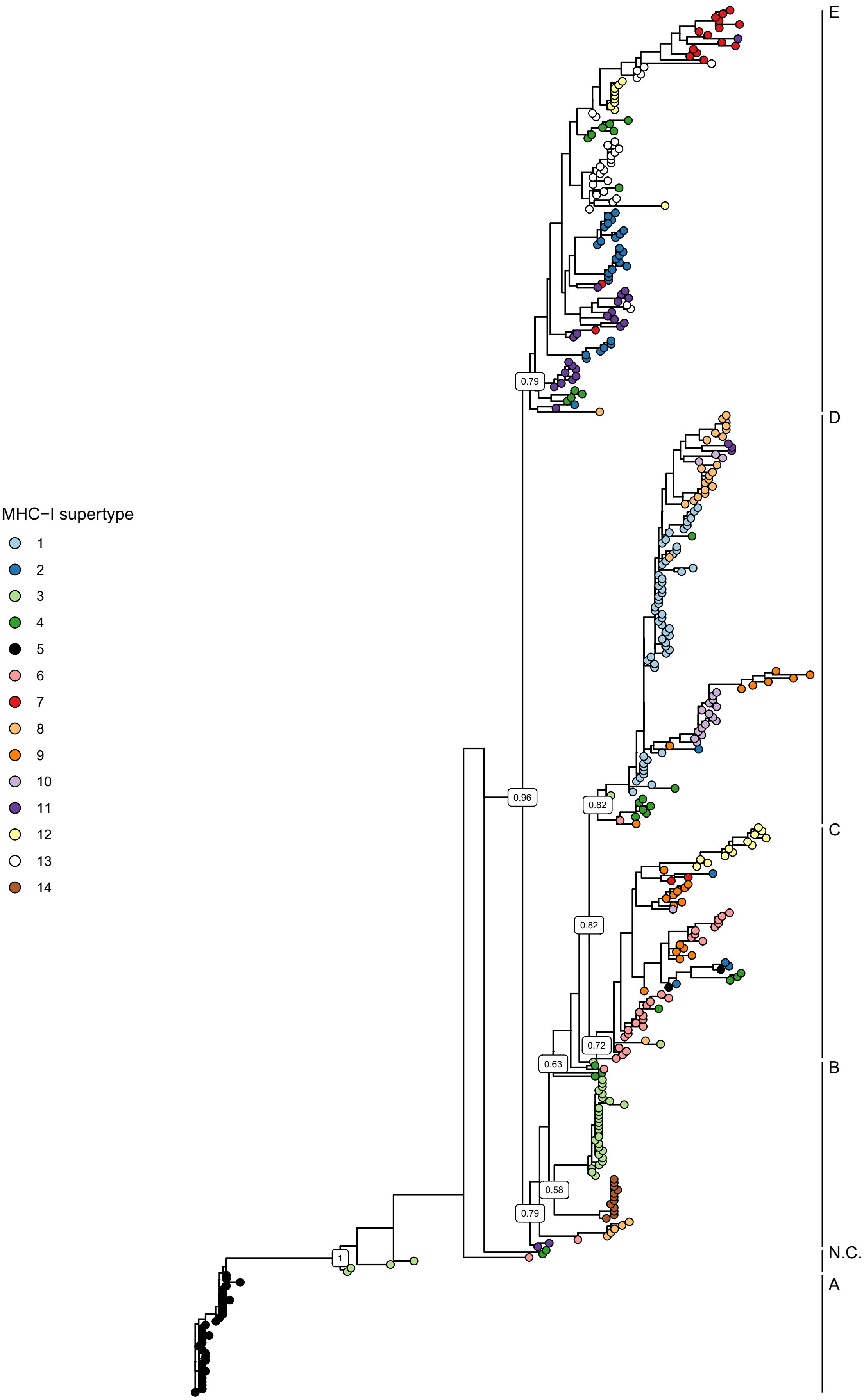


FIGURE 2 Unrooted GTR tree of the 390 MHC-I exon 3 alleles in our data set with SH-aLRT support values shown for selected nodes. Association of alleles with MHC-I supertypes is indicated with coloured circles. The side bars indicate the position of groups A–E and the group of nonclustering alleles (N.C.) in the tree

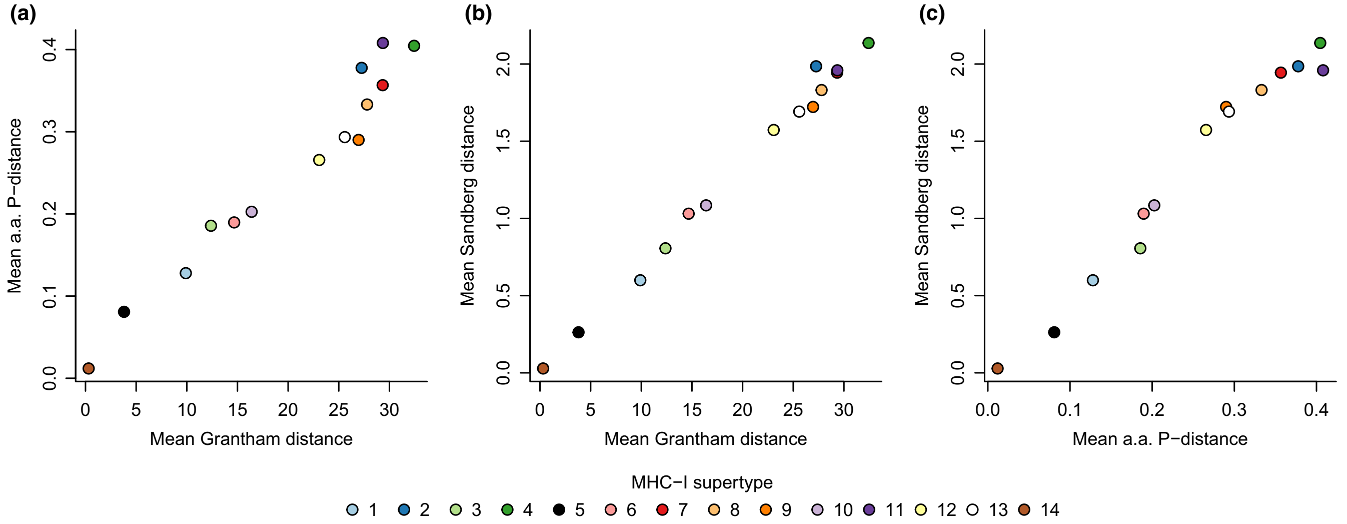


FIGURE 3 (a–c) Scatterplots illustrating the distributions of and associations between the three measures of functional divergence within MHC-I supertypes: mean Grantham distance, mean amino acid p-distance, and mean Sandberg distance. Notes: In plot b, the data point representing Acar-ST*7 is almost completely overlapped by the point representing Acar-ST*11. Values were calculated using the amino acids in 14 codons that showed evidence of positive selection

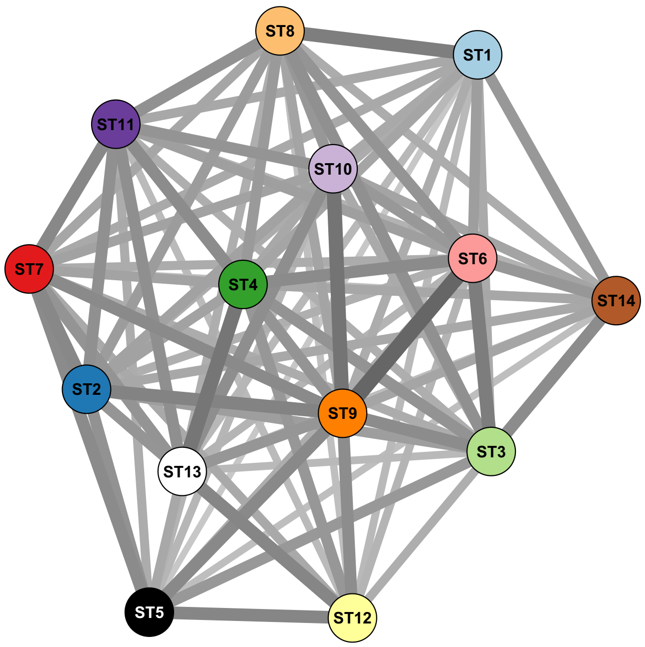


FIGURE 4 Network visualization of the functional divergence between the MHC-I supertypes (ST1–ST14) based on the pairwise Sandberg distance between supertype centroids. The saturation and width of the connecting lines indicate the proximity of each pair of supertype centroids

amino acid (a.a.) p-distance ranged from 0.012 to 0.408 with an average of 0.252, and the mean Sandberg distance ranged from 0.028 to 2.136 with an average of 1.332. The values of these three measures of functional divergence were strongly correlated (Pearson's correlations: r = 0.980–0.997, $p < .0001$; Figure 3). Values for each supertype are specified in Table S9. The pairwise Sandberg distances between MHC-I supertype centroids ranged from 1.73 to 4.75 with an average of 2.89 and the pairwise overlap between MHC-I supertypes ranged from −3.89 to 1.89 with an average of −0.23 (Table S10). The functional divergence between MHC-I supertypes is illustrated with a network visualization in Figure 4.

3.7 | MHC-I supertypes on haplotypes

The MHC-I haplotypes in our data set carried between three and 13 MHC-I supertypes (mean = 7.0, SD = 1.84; Table S11a, Figure 5a). Three supertypes (Acar-ST*3, Acar-ST*5, and Acar-ST*14) were present on almost all haplotypes, Acar-ST*1 was present on 77 haplotypes, while the remaining supertypes each were present on approximately half or less of the haplotypes (mean = 35.6, SD = 12.0; Table S11b, Figure 5b). Accordingly, we observed great diversity in haplotype composition beyond the three most common supertypes (Figure 6). The number of supertypes and number of alleles on haplotypes were strongly correlated (Pearson's correlation: r = 0.86 [0.84 when excluding the outlier Acar-HPLT*84], $p < .0001$; Figure S11). Furthermore, we found considerable variation in the number of alleles from each supertype that were represented on haplotypes (Figure 6), with variances of the number of alleles representing each supertype on haplotypes differing significantly between supertypes (Levene's test: $p < .0001$; Table 2).

3.8 | MHC-I supertypes and evolutionary patterns

To compare the peptide binding properties represented by the 14 MHC-I supertypes with the evolutionary relationships between alleles, we plotted supertype associations in the phylogenetic tree

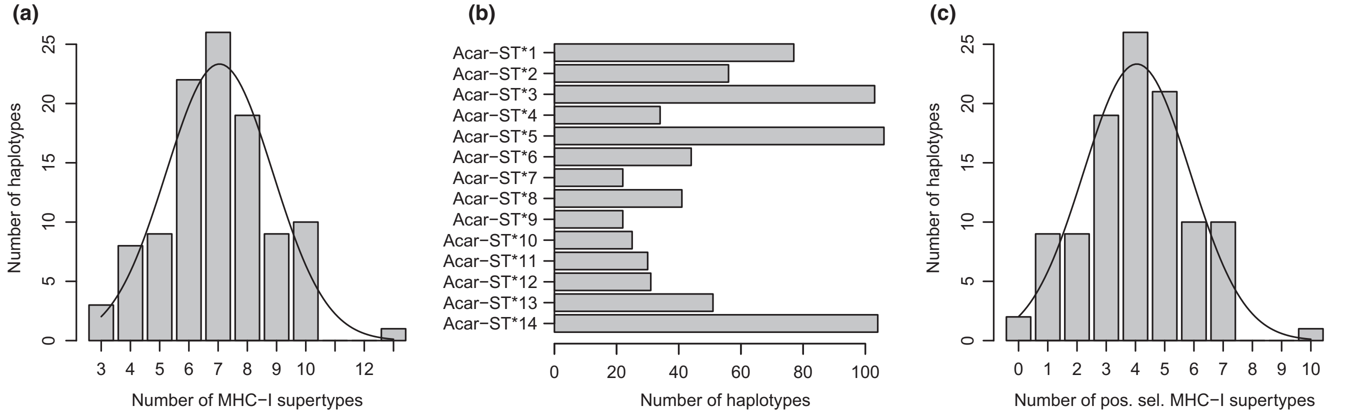


FIGURE 5 (a) Distribution of the number of MHC-I supertypes in haplotypes. The line shows a normal distribution with the observed mean (7.0) and standard deviation (1.84). (b) Bar plot showing the number of different haplotypes that each MHC-I supertype was observed in. (c) Distribution of the number of positively selected MHC-I supertypes in haplotypes. The line shows a normal distribution with the observed mean (4.0) and standard deviation (1.83)

(Figure 2). Most of the alleles representing the three most common supertypes (Acar-ST*3, Acar-ST*5, and Acar-ST*14) reside in three monophyletic clades, that each harbour limited divergence (Acar-ST*5 in group A; Acar-ST*3 and Acar-ST*14 in two monophyletic clades within group B). In contrast, the variation in branch lengths indicate considerable divergence within all other groups (groups C, D and E) in the tree. Yet, interestingly, we found limited divergence between the alleles from Acar-ST*1, even though they reside with alleles from more divergent supertypes in group D (Figure 2).

The divergent patterns evident in the tree spurred us to test for positive selection within each MHC-I supertype. When running CODEML on subsets of alleles representing each supertype, we found that the M2 and M8 models fit the data significantly better than the M1 and M7 models, respectively, for Acar-ST*1, Acar-ST*2, Acar-ST*4, and Acar-ST*[6–13], indicating that alleles associated with these supertypes evolved under positive selection (Table S5b–c, e, g–n). The positively selected sites predicted by BEB analysis for these supertypes and associated dN/dS and *p*-values are shown in Figure S9 and Table S6b–l. In contrast, the models M2 and M8 did not fit the data significantly better than M1 and M7 for Acar-ST*3, Acar-ST*5, and Acar-ST*14, indicating that alleles associated with these supertypes have not evolved under positive selection (Table S5d, f, o).

Among Acar-ST*1, Acar-ST*2, Acar-ST*4, and Acar-ST*[6–13], we found considerable variation in the number and position of the positively selected sites (Figure S9), which tended to cluster within amino acid codons 1–5, 18–21, and 57–62, suggesting that these form regions where the MHC-I protein interacts with antigens.

3.9 | MHC-I diversity and functional divergence within haplotypes

MHC-I haplotypes carried between zero and 10 MHC-I supertypes that showed evidence of positive selection (mean = 4.0, SD = 1.83; Table S12, Figure 5c). Two haplotypes (Acar-HPLT*74 and Acar-HPLT*82) carried only MHC-I supertypes that did not show evidence of positive selection, and nine haplotypes (Acar-HPLT*05, Acar-HPLT*42, Acar-HPLT*49, Acar-HPLT*57, Acar-HPLT*61, Acar-HPLT*66, Acar-HPLT*67, Acar-HPLT*68, and Acar-HPLT*76) carried only one positively selected MHC-I supertype. We quantified the functional divergence harboured within each haplotype as means of the pairwise centroid distances and pairwise overlaps between positively selected MHC-I supertypes in each haplotype. The mean centroid distances ranged from 1.73 to 4.12 (mean = 2.89, SD = 0.30), and mean overlaps ranged from −1.44 to 1.13 (mean = 0.095, SD = 0.51) (Table S12; Figures S12a–b).

3.10 | Nonrandom association of MHC-I supertypes in haplotypes

To investigate whether natural selection has favoured haplotypes that harbour high levels of diversity or functional divergence of positively selected MHC-I supertypes, we compared the observed haplotypes to in silico predictions of possible haplotypes from 10,000 simulated data sets. In these tests, the number of positively selected MHC-I supertypes did not differ significantly between real and simulated haplotypes ($p = .33$). However, the mean centroid distances between the positively selected MHC-I supertypes in real haplotypes were significantly larger than in the simulated haplotypes ($p < .0059$; Figure 7a). Similarly, the mean overlap between the positively selected MHC-I supertypes in real haplotypes was significantly smaller than in the simulated haplotypes ($p < .0037$; Figure 7b).

4 | DISCUSSION

We present the R package MHCTOOLS, which contains tools that facilitate population-wide screening of MHC diversity, functional divergence, and segregating haplotypes. We demonstrate the use of functions from MHCTOOLS on an empirical data set of great reed warblers to (i) optimize settings in bioinformatical filtering of

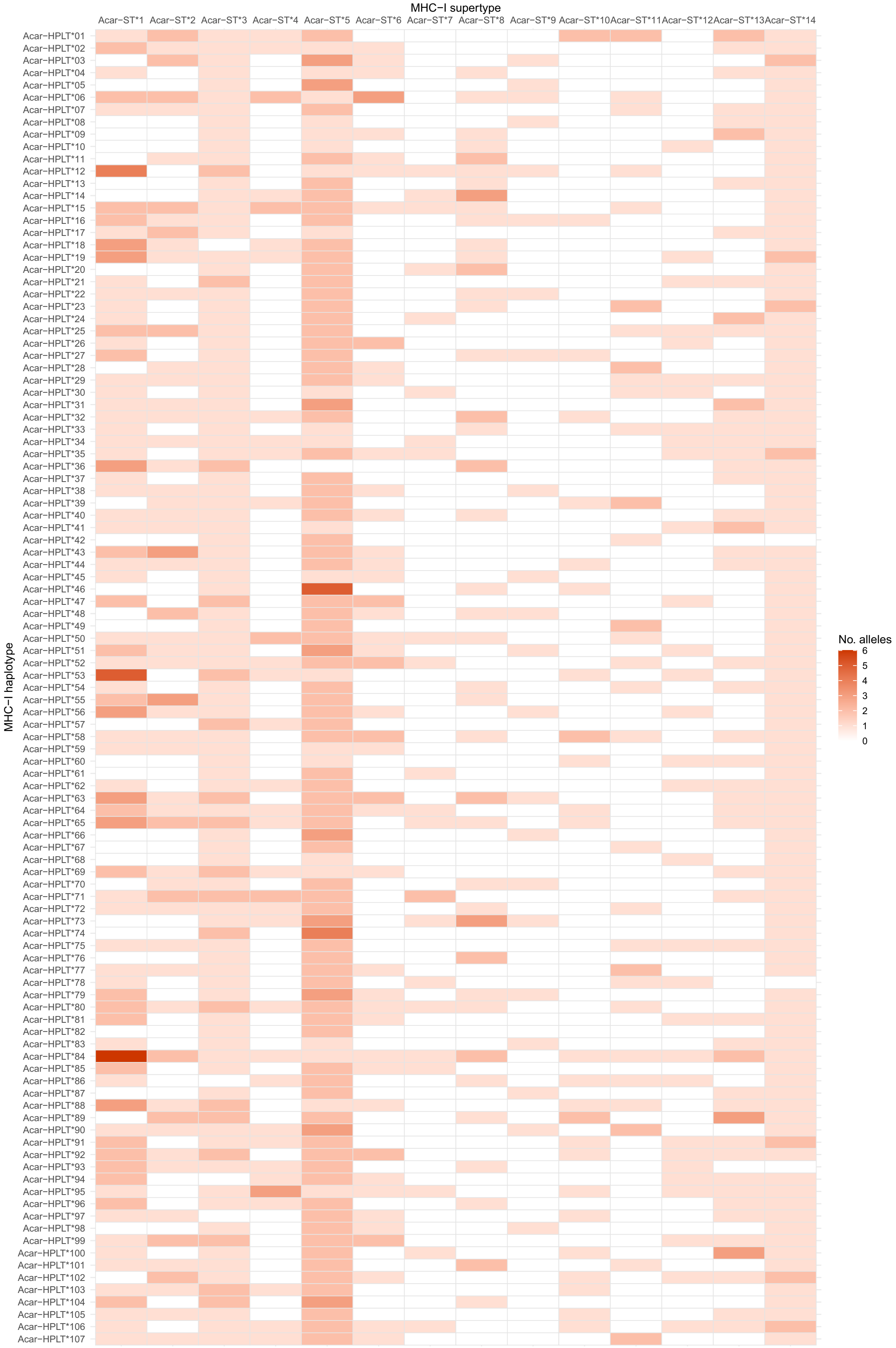


FIGURE 6 Heatmap showing the number of alleles associated with each MHC-I supertype that was observed in each haplotype

TABLE 2 Variances of the number of alleles from each MHC-I supertype observed per haplotype

	Variance of no. alleles per haplotype
Acar-ST*1	1.20
Acar-ST*2	0.58
Acar-ST*3	0.19
Acar-ST*4	0.37
Acar-ST*5	0.44
Acar-ST*6	0.44
Acar-ST*7	0.19
Acar-ST*8	0.52
Acar-ST*9	0.16
Acar-ST*10	0.25
Acar-ST*11	0.38
Acar-ST*12	0.21
Acar-ST*13	0.47
Acar-ST*14	0.09

high-throughput amplicon sequencing data, (ii) characterize segregating MHC-I haplotypes based on family data, (iii) infer MHC-I supertypes based on physicochemical z-descriptors, and (iv) calculate distances for functional divergence estimation.

We used the function ReplMatch from MHCTOOLS for fast and efficient evaluation of replicates in our Illumina sequencing data set, which enabled us to optimize settings for bioinformatical filtering. While our filtering protocol was based on DADA2, our optimization of the settings (i.e., by repeated filtering and evaluation of the data using ReplMatch) was instrumental in achieving a genotyping repeatability of 0.998. Related to MHC genotyping, MHCTOOLS additionally includes the functions CreateFas and CreateSamplesFas, which generate fasta files for manual screening of sequencing data—an important step in any MHC genotyping experiment.

As tightly linked MHC loci often cosegregate, investigating how different alleles associate with different loci is relevant for understanding the organization of and the evolutionary relationship between MHC genes. Understanding how MHC alleles segregate in haplotypes is an obvious step along that road. We therefore developed the function HpltFind in MHCTOOLS to carry out automated analysis of allele segregation patterns in family data, which greatly facilitates haplotype studies on large data sets. In addition to inferred segregating haplotypes, HpltFind reports unresolved and incongruent allele occurrences, which allows both for (i) standardized evaluation of the overall performance of haplotype inference processes and (ii) further investigations of problematic segregation patterns. We provide an example of a stepwise protocol for resolving problematic allele segregation patterns, which can be modified to fit most data sets (Figure S2; Supporting Information Methods). We used HpltFind to characterize MHC-I haplotypes in our empirical data set and identified 107 different MHC-I haplotypes in 116 great reed warbler families. This corresponds to the diversity observed in a similar study on barn owls, where 111 different MHC-I haplotypes were observed among 140 families (Gaigher et al., 2018). Our analyses of allele segregation patterns indicated that MHC-I loci are tightly linked in the great reed warbler, with a recombination rate of 0.0030, corresponding to a genetic distance of 0.3 cM. A recent study confirmed this strong linkage by showing that MHC-I genes in the great reed warbler mainly are organized as tandemly duplicated genes within a small genomic region (Westerdahl et al., 2022).

Functional divergence between MHC alleles is of great biological interest due to its association with resistance of vertebrate hosts to pathogens (Pierini & Lenz, 2018; Wakeland et al., 1990). MHCTOOLS includes the functions BootKmeans and DistCalc that facilitate analyses of functional divergence in the qualitative and quantitative sense. Employing these, we identified 14 MHC-I supertypes, each representing a subset of alleles that share similar functional properties. Our bootstrapped clustering approach allowed us to identify the number of clusters that was associated with the greatest accuracy in the cluster assignment of alleles, and to select the models with the most informative clusters (i.e., the ones with the greatest ΔBIC). Our method thus lends a transparency and a degree of confidence to inference of MHC supertypes that have not been possible to achieve with previous methods, and which have been called for in a recent evaluation of studies that inferred genetic clusters (Miller et al., 2020). We recommend that future studies of MHC supertypes employ our method, potentially in combination with subsequent DAPC, if further qualitative analyses of the physicochemical differences between inferred MHC supertypes are required. We subsequently employed the DistCalc function to quantify and compare levels of functional divergence harboured within MHC-I.

To demonstrate the power of combining analyses of MHC haplotypes and MHC functional divergence—which are facilitated by the functions available in MHCTOOLS—we dedicate the next paragraphs to discussing the results of the analyses on our empirical data set.

4.1 | MHC-I diversity on haplotypes in the great reed warbler

The MHC exhibits extraordinary evolutionary dynamics with rapid expansions and contractions of MHC gene copy number, and substantial variation in MHC haplotype structure (Kelley et al., 2005; Minias et al., 2018; Nei & Rooney, 2005). Previous studies have reported considerable variation in the number of different MHC alleles between individuals within species, suggesting that MHC gene copy number variation (CNV) may be a common trait, at least among birds (O'Connor et al., 2019). Our analyses of MHC-I haplotypes confirmed previous indications of substantial MHC-I CNV in the great reed warbler (O'Connor et al., 2016; Roved et al., 2018), with a minimum of four and a maximum of 21 different MHC-I alleles per haplotype. The

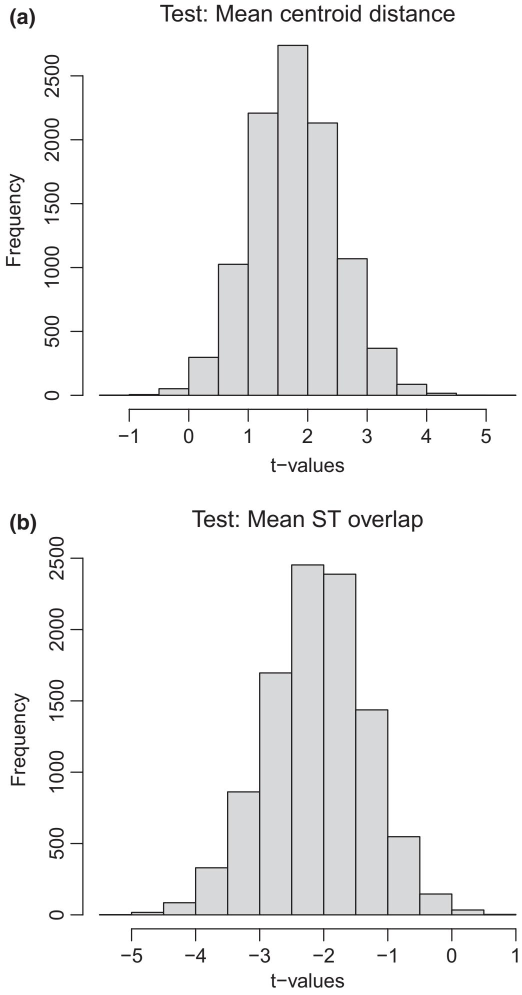


FIGURE 7 The distribution of t-values from pairwise *t*-tests of (a) the mean centroid distances and (b) the mean overlap between positively selected MHC-I supertypes in real haplotypes versus in 10,000 simulated populations of possible haplotypes

number of MHC-I supertypes observed on haplotypes varied between three and 13 and was positively correlated with the number of different MHC-I alleles per haplotype (Figure S11). Interestingly, despite the strong correlation, we found significant variation in the number of alleles that represented each MHC-I supertype on haplotypes, suggesting that gene duplication may be more frequent in loci that harbour certain supertypes (e.g., Acar-ST*1 and Acar-ST*5), while also differing between haplotypes (Figure 6). The observed variation in the number of MHC-I supertypes between haplotypes may be a consequence of previous diversification and neo-functionalization of duplicated loci driven by selection from pathogens, as indicated by a recent simulation study (Bentkowski & Radwan, 2019).

4.2 | Signatures of selection differ between MHC-I supertypes

A phylogenetic reconstruction of our MHC-I alleles placed most alleles in three monophyletic clades that harboured considerable divergence (clades C, D, E; Figure 2), as expected for MHC alleles that coevolve with pathogens (Edwards, 2009). These clades harboured alleles from 11 MHC-I supertypes that showed evidence of positive selection. The alleles that represent the remaining three MHC-I supertypes (Acar-ST*3, Acar-ST*5, and Acar-ST*14) mainly grouped in three other monophyletic clades, that each harboured limited divergence (Figure 2). As expected from this pattern, we found no evidence for positive selection among the alleles associated with these three supertypes. This suggests that the topology in the phylogenetic tree may reflect MHC-I genes with slightly different biological functions. Acar-ST*3, Acar-ST*5, and Acar-ST*14 are represented on most haplotypes in our data set (Figure 5b), and we propose that the genes harbouring these may be nonclassical MHC genes, that serve other functions than pathogen recognition. Among the MHC-I supertypes that showed evidence of positive selection, our analysis revealed considerable variation in the number and position of positively selected codons, consistent with these supertypes having evolved as balanced polymorphisms with divergent peptide-binding properties (Figure S9), that is, features expected for classical MHC-I genes involved in antigen presentation and pathogen recognition. While most MHC-I supertypes were predominantly associated with single clades in the phylogeny (the exceptions being Acar-ST*4 and Acar-ST*12), we observed some degree of admixture of all MHC-I supertypes between clades. This intriguing pattern can potentially be a consequence of gene conversion or a signature of convergent evolution, where selection favoured similar binding properties at different MHC-I loci. Future studies on the expression of individual MHC-I alleles in the great reed warbler would be essential to verify these suggestions.

To estimate the breadth of each MHC-I supertype, we calculated the functional divergence between the alleles associated with each supertype (Figure 3). Among the putative nonclassical MHC supertypes, the alleles in Acar-ST*5 and Acar-ST*14 were considerably less divergent than alleles in all other supertypes, while Acar-ST*3 had the fourth lowest level of functional divergence (Figure 3). Among the positively selected MHC-I supertypes, Acar-ST*1, Acar-ST*6, and Acar-ST*10 harboured relatively low levels of functional divergence. Notably, we also observed a large variance in the number of alleles from Acar-ST*1 represented on haplotypes (Table 2), and we propose that this observation indicates recent duplication of loci belonging to a putative classical MHC-I gene harbouring Acar-ST*1. Under this scenario, the limited divergence within Acar-ST*1 may be explained by recently duplicated loci not having had time to diverge. The remaining eight MHC-I supertypes (Acar-ST*2, Acar-ST*4, Acar-ST*[7–9], and Acar-ST*[11–13]) all harboured considerably higher levels of functional divergence, consistent with the divergent branch lengths in the phylogenetic tree (Figures 2 and 3).

4.3 | Functional divergence within MHC-I haplotypes

Standing genetic variation in MHC haplotypes serves an important evolutionary function by enabling rapid adaptive shifts in response to dynamics of pathogen communities (cf. Alves et al., 2019), and accordingly we found great variation in the composition of MHC-I haplotypes in our data set (Figure 6). That haplotypes harboured on average four positively selected MHC-I supertypes (Figure 5c) is in agreement with the principle that increased MHC diversity enables the adaptive immune system to recognize more pathogens. Altogether, 96 out of 107 haplotypes harboured at least two positively selected MHC-I supertypes. However, we were surprised to find two haplotypes that harboured no positively selected MHC-I supertypes and nine that harboured only one. A potential explanation for our observation of these low-diversity MHC-I haplotypes is that they contain additional alleles from classical MHC-I loci that were not detected by our primers. In a previous study, the primers that we employed amplified ~80% of the total population of MHC-I alleles detected when genotyping was carried out using one additional set of primers (O'Connor et al., 2016). It is, however, also possible that deviations from optimal MHC diversity are offset by selective advantages associated with particular MHC alleles (Sepil et al., 2013; Westerdahl et al., 2005), or that haplotypes with lower than optimal diversity are maintained in combinations with haplotypes that harbour high diversity. In such combinations, the presence of low-diversity haplotypes would cause diploid levels of MHC-I diversity not to extend too far beyond a hypothetical optimum (cf. Nowak et al., 1992; Woelfing et al., 2009). Hence, the existence of low-diversity haplotypes may allow populations to maintain a standing genetic variation that includes haplotypes with higher-than-optimal MHC diversity.

Finally, we also investigated the functional divergence of MHC-I supertypes within haplotypes. The haplotypes in our great reed warbler population harboured positively selected MHC-I supertypes that were more divergent and overlapped to a lesser degree than expected from in silico simulations, where MHC-I alleles were randomly assigned to haplotypes (Figures 7a, b). This indicates that natural selection has favoured nonrandom combinations of MHC-I supertypes that increase the functional divergence harboured in haplotypes. Such nonrandom association of MHC-I supertypes in haplotypes may be evolutionarily advantageous by increasing the range of pathogens that can be recognized by the adaptive immune system, consistent with the principle of the DAA (cf. Wakeland et al., 1990). To our knowledge, our study is the first to investigate MHC diversity in haplotypes by MHC supertype analysis. However, association of highly divergent MHC alleles in haplotypes has previously been shown in chacma baboons (*Papio ursinus*), where it was suggested that selection favours haplotypes that combine MHC-DRB alleles with dissimilar physicochemical properties (Huchard et al., 2008). Similarly, Gaigher et al. (2018) found that differences in amino acid sequences between two known MHC-IIB loci in barn owls had reached fixation. In contrast, Gaigher et al. (2018) found no evidence for a shift towards highly divergent allele combinations in MHC class I haplotypes in barn owls.

5 | CONCLUSION

We have presented the R package MHCTOOLS and demonstrated the power of its functions to (i) support accurate MHC genotyping in non-model species, (ii) quantify functional divergence between MHC alleles, and (iii) carry out population-wide screening of MHC supertypes and segregating haplotypes. We believe that MHCTOOLS will be valuable to future MHC studies in non-model species, and that it offers methodological improvements to the field of MHC research, that may help to advance our understanding of MHC genetics and evolution.

AUTHOR CONTRIBUTIONS

Jacob Roved designed the data analysis protocols and conceived, designed, and created the R package MHCTOOLS; Jacob Roved, Bengt Hansson, Dennis Hasselquist, and Helena Westerdahl jointly conceived the study of MHC-I haplotypes in great reed warblers; Jacob Roved conceived the method for inference of MHC-I supertypes and all downstream data analyses; Martin Stervander constructed amplicon sequencing libraries for Illumina sequencing; Jacob Roved carried out bioinformatics, analysed the data, and wrote the manuscript, with input from all authors.

ACKNOWLEDGEMENTS

We wish to thank A. Drews and T. Johansson for library constructions and high-throughput 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 (2016-04391, 2020-03976 to DH; 621-2014-5222, 2016-00689 to BH; 2015-05149, 2020-04285 to HW), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Starting 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).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

MHCTOOLS version 1.4.2 (including user manual and documentation) has been made available at CRAN: https://cran.r-project.org/package=MHCtools. Our data set is available at the Zenodo repository: https://doi.org/10.5281/zenodo.3716048 (Roved et al., 2020). DNA sequences are available at GenBank: https://ncbi.nlm.nih.gov (accession numbers: MH468831–MH469159; MT193762–MT193822).

ORCID

Jacob Roved https://orcid.org/0000-0001-6977-714X
Bengt Hansson https://orcid.org/0000-0001-6694-8169
Martin Stervander https://orcid.org/0000-0002-6139-7828
Dennis Hasselquist https://orcid.org/0000-0002-0056-6616
Helena Westerdahl https://orcid.org/0000-0001-7167-9805

REFERENCES

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*, *1*, e86. https://doi.org/10.7717/peerj.86

Alves, J. M., Carneiro, M., Cheng, J. Y., de Matos, A. L., Rahman, M. M., Loog, L., Campos, P. F., Wales, N., Eriksson, A., Manica, A., Strive, T., Graham, S. C., Afonso, S., Bell, D. J., Belmont, L., Day, J. P., Fuller, S. J., Marchandeau, S., Palmer, W. J., … Jiggins, F. M. (2019). Parallel adaptation of rabbit populations to myxoma virus. *Science*, *363*(6433), 1319–1326. https://doi.org/10.1126/science.aau7285

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

Bentkowski, P., & Radwan, J. (2019). Evolution of major histocompatibility complex gene copy number. *PLoS Computational Biology*, *15*(5), 1–15. https://doi.org/10.1371/journal.pcbi.1007015

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

Buczek, M., Okarma, H., Demiaszkiewicz, A. W., & Radwan, J. (2016). MHC, parasites and antler development in red deer: No support for the Hamilton & Zuk hypothesis. *Journal of Evolutionary Biology*, *29*(3), 617–632. https://doi.org/10.1111/jeb.12811

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

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), 581–583. https://doi.org/10.1038/nmeth.3869

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

Edwards, S. V. (2009). Natural selection and phylogenetic analysis. *Proceedings of the National Academy of Sciences*, *106*(22), 8799–8800. https://doi.org/10.1073/pnas.0904103106

Ejsmond, M. J., & Radwan, J. (2015). Red queen processes drive positive selection on major histocompatibility complex (MHC) genes. *PLoS Computational Biology*, *11*(11), e1004627. 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*, *120*, 396–406. 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*, 296. https://doi.org/10.1186/1471-2164-11-296

Gonzalez-Quevedo, C., Davies, R. G., & Richardson, D. S. (2014). Predictors of malaria infection in a wild bird population: landscape-level analyses reveal climatic and anthropogenic factors. *Journal of Animal Ecology.*, *83*, 1091–1102. https://doi.org/10.1111/1365-2656.12214

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, *185*(4154), 862–864.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by Maximum Likelihood. *Systematic Biology*, *52*(5), 696–704. https://doi.org/10.1080/10635150390235520

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*. (Vol. *41*, pp. 95–98). Oxford University Press.

Hansson, B., 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, 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

Hasselquist, D. (1998). Polygyny in great reed warblers: A long-term study of factors contributing to male fitness. *Ecology*, *79*(7), 2376–2390.

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

Hertz, T., & Yanover, C. (2007). Identifying HLA supertypes by learning distance functions. *Bioinformatics*, *23*(2), e148–e155. https://doi.org/10.1093/Bioinformatics/btl324

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

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

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. *Annual Review of Immunology*, *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/10.1007/s00251-004-0717-7

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

Klein, J., & 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

Klein, J., Sato, A., & Nikolaidis, N. (2007). MHC, TSP, and the origin of species: From immunogenetics to evolutionary genetics. *Annual Review of Genetics*, *41*(1), 281–304. https://doi.org/10.1146/annurev.genet.41.110306.130137

Leclaire, S., Strandh, M., Mardon, J., Westerdahl, H., & Bonadonna, F. (2017). Odour-based discrimination of similarity at the major histocompatibility complex in birds. *Proceedings of the Royal Society B*, *284*(1846), 20162466. https://doi.org/10.1098/rspb.2016.2466

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

Lenz, T. L., Mueller, B., Trillmich, F., & Wolf, J. B. W. (2013). Divergent allele advantage at MHC-DRB through direct and maternal genotypic effects and its consequences for allele pool composition and mating. *Proceedings of the Royal Society B*, *280*(1762), 20130714. https://doi.org/10.1098/rspb.2013.0714

Lighten, J., Papadopulos, A. S. T., Mohammed, R. S., Ward, B. J., Paterson, I. G., Baillie, L., Bradbury, I. R., Hendry, A. P., Bentzen, P., & van Oosterhout, C. (2017). Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. *Nature Communications*, *8*(1297), 1–10. https://doi.org/10.1038/s41467-017-01183-2

Lighten, J., Van Oosterhout, C., & Bentzen, P. (2014). Critical review of NGS analyses for de novo genotyping multigene families. *Molecular Ecology.*, *23*, 3957–3972. https://doi.org/10.1111/mec.12843

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), 753–767. https://doi.org/10.1111/1755-0998.12225

Lillie, M., Grueber, C. E., Sutton, J. T., Howitt, R., Bishop, P. J., Gleeson, D., & Belov, K. (2015). Selection on MHC class II supertypes in the New Zealand endemic Hochstetter's frog phylogenetics and phylogeography. *BMC Evolutionary Biology*, *15*(1), 1–11. https://doi.org/10.1186/s12862-015-0342-0

Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Røder, G., Justesen, S., Buus, S., & Brunak, S. (2004). Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, *55*(12), 797–810. https://doi.org/10.1007/s00251-004-0647-4

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

Miller, J. M., Cullingham, C. I., & Peery, R. M. (2020). The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity*, *125*(5), 269–280. https://doi.org/10.1038/s41437-020-0348-2

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

Nei, M., & 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

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, E. 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

Okano, M., Miyamae, J., Suzuki, S., Nishiya, K., Katakura, F., Kulski, J. K., Moritomo, T., & Shiina, T. (2020). Identification of novel alleles and structural haplotypes of major histocompatibility complex Class I and DRB genes in domestic cat (*Felis catus*) by a newly developed NGS-based genotyping method. *Frontiers in Genetics*, *11*, 1–15. https://doi.org/10.3389/fgene.2020.00750

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*. R Foundation for Statistical Computing. https://www.r-project.org/

Richman, A. (2000). Evolution of balanced genetic polymorphism. *Molecular Ecology*, *9*(12), 1953–1963. https://doi.org/10.1046/j.1365-294X.2000.01125.x

Rioux, J. D., Goyette, P., Vyse, T. J., Hammarstroem, L., Fernando, M. M. A., Green, T., De Jager, P. L., Foisy, S., Wang, J., de Bakker, P. I. W., Leslie, S., McVean, G., Padyukov, L., Alfredsson, L., Annese, V., Hafler, D. A., Pan-Hammarstroem, Q., Mattell, R., Sawcer, S. J. … Hauser, S. L. (2009). Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(44), 18680–18685. https://doi.org/10.1073/pnas.0909307106

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

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., & Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry*, *41*(14), 2481–2491. https://doi.org/10.1021/jm9700575

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), 20130134. https://doi.org/10.1098/rspb.2013.0134

Sepil, I., Moghadam, H. K., Huchard, E., & Sheldon, B. C. (2012). Characterization and 454 pyrosequencing of major histocompatibility complex class I genes in the great tit reveal complexity in a passerine system. *BMC Evolutionary Biology*, *12*, 68. https://doi.org/10.1186/1471-2148-12-68

Sidney, J., Grey, H. M., Kubo, R. T., & Sette, A. (1996). Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. *Immunology Today*, *17*(6), 261–266.

Sidney, J., Peters, B., Frahm, N., Brander, C., & Sette, A. (2008). HLA class I supertypes: A revised and updated classification. *BMC Immunology*, *9*, 1–15. https://doi.org/10.1186/1471-2172-9-1

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. *Molecular Ecology*, *29*, 3578–3592. https://doi.org/10.1111/mec.15471

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

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th end. ed.). Pearson Education Limited.

Trujillo, A. L., Hoffman, E. A., Becker, C. G., & Savage, A. E. (2021). Spatiotemporal adaptive evolution of an MHC immune gene in a frog-fungus disease system. *Heredity*, *126*(4), 640–655. https://doi.org/10.1038/s41437-020-00402-9

Wakeland, E. K., Boehme, S., She, J. X., Lu, C. C., McIndoe, R. A., Cheng, I., Ying, Y., & 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., Mellinger, S., Sigeman, H., Kutschera, V. E., Proux-Wéra, E., Lundberg, M., Weissensteiner, M., Churcher, A., Bunikis, I., Hansson, B., Wolf, J. B. W., & Strandh, M. (2022). The genomic architecture of the passerine MHC region: high repeat content and contrasting evolutionary histories of single copy and tandemly duplicated MHC genes. *Molecular Ecology Resources*, 1–17. https://doi.org/10.1111/1755-0998.13614

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., & 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

Winternitz, J. C., Promerova, M., Polakova, R., Vinkler, M., Schnitzer, J., Munclinger, P., Babik, W., Radwan, J., Bryja, J., & Albrecht, T. (2015). Effects of heterozygosity and MHC diversity on patterns of extra-pair paternity in the socially monogamous scarlet rosefinch. *Behavioral Ecology and Sociobiology*, *69*(3), 459–469. https://doi.org/10.1007/s00265-014-1858-9

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

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, *13*(5), 555–556. https://doi.org/10.1093/bioinformatics/13.5.555

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., Wong, W. S. W., & Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, *22*(4), 1107–1118. https://doi.org/10.1093/molbev/msi097

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*, 1–15. https://doi.org/10.1186/1471-2148-10-395

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Roved, J., Hansson, B., Stervander, M., Hasselquist, D., & Westerdahl, H. (2022). MHCtools – an R package for MHC high-throughput sequencing data: Genotyping, haplotype and supertype inference, and downstream genetic analyses in non-model organisms. *Molecular Ecology Resources*, *22*, 2775–2792. https://doi.org/10.1111/1755-0998.13645