Tradeoffs in expressed major histocompatibility complex diversity seen on a macro evolutionary scale among songbirds

O'Connor, Emily; Westerdahl, Helena

Published in:
Evolution: international journal of organic evolution

DOI:
10.1111/evo.14207

2021

doi:10.1111/evo.14207

Trade-offs in expressed major histocompatibility complex diversity seen on a macroevolutionary scale among songbirds

Emily A. O'Connor[1,2] **and Helena Westerdahl**[1]

[1]*Molecular Ecology and Evolution Lab, Department of Biology, Lund University, Lund, Sweden*

[2]*E-mail: emily.o_connor@biol.lu.se*

Received July 3, 2020

Accepted February 17, 2021

To survive organisms must defend themselves against pathogens. Classical Major Histocompatibility Complex (MHC) genes play a key role in pathogen defense by encoding molecules involved in pathogen recognition. MHC gene diversity influences the variety of pathogens individuals can recognize and respond to and has consequently been a popular genetic marker for disease resistance in ecology and evolution. However, MHC diversity is predominantly estimated using genomic DNA (gDNA) with little knowledge of expressed diversity. This limits our ability to interpret the adaptive significance of variation in MHC diversity, especially in species with very many MHC genes such as songbirds. Here, we address this issue using phylogenetic comparative analyses of the number of MHC class I alleles (MHC-I diversity) in gDNA and complementary DNA (cDNA), that is, expressed alleles, across 13 songbird species. We propose three theoretical relationships that could be expected between genomic and expressed MHC-I diversity on a macroevolutionary scale and test which of these are best supported. In doing so, we show that significantly fewer MHC-I alleles than the number available are expressed, suggesting that optimal MHC-I diversity could be achieved by modulating gene expression. Understanding the relationship between genomic and expressed MHC diversity is essential for interpreting variation in MHC diversity in an evolutionary context.

KEY WORDS: **Birds, Gene expression, Major Histocompatibility Complex, MHC diversity, MHC genes.**

The ability of organisms to survive attack from pathogens is a strong determinant of fitness (Roy and Kirchner 2000; Heidel and Dong 2006; Orgil et al. 2007). To eliminate a pathogen successfully, the host's immune system must recognize the antigens as foreign and mount an appropriate response. Classical Major Histocompatibility Complex genes (henceforth "MHC genes") encode molecules that are central to this process in vertebrates, as they present antigens to Tcells, for identification as either self or nonself (Murphy et al. 2008). If an antigen is identified as nonself, then a highly specific adaptive immune response is initiated. The number of MHC gene copies an individual has plays a pivotal role in determining the array of pathogens that can be recognized (Wegner et al. 2004; Meyer-Lucht and Sommer 2009; Eizaguirre and Lenz 2010; Kloch et al. 2010). New copies of MHC genes arise through duplication events resulting in copies that either diverge functionally, become pseudogenes through the accumulation of deleterious mutations, or are deleted from the genome—the so called "birth-and-death" mode of gene evolution (Nei et al. 1997; Sato et al. 2001; Piontkivska and Nei 2003). MHC genes represent one of the most polymorphic regions of the vertebrate genome and this polymorphism is believed to be primarily driven and maintained by pathogen-mediated selection (Hedrick 2002; Spurgin and Richardson 2010; Radwan et al. 2020).

Measuring MHC gene diversity (henceforth "MHC diversity"), through counting the number of different MHC alleles per individual, is a common approach to estimate the immunogenetic competence of individuals in ecological and evolutionary studies of nonmodel organisms (Wegner et al. 2003; Westerdahl et al. 2005; Meyer-Lucht and Sommer 2009; Oliver et al. 2009; Brouwer et al. 2010; Radwan et al. 2012). However, our

interpretation of the functional relevance of MHC diversity is limited by a lack of data on how many of these genes are expressed especially given that, according to the birth-and-death model, not all copies will necessarily be expressed (Nei et al. 1997). This leads to uncertainty over how to interpret differences in MHC diversity both within and among populations and species, particularly for species with highly duplicated MHC genes such as many songbirds (Sepil et al. 2013; O'Connor et al. 2016; Biedrzycka et al. 2017; Minias et al. 2019). For example, the difference between a species with 10 versus five MHC genes may have little adaptive relevance if only five MHC genes are expressed, given that the number of MHC molecules interacting with pathogens is then likely to be similar in both species.

The number of expressed MHC genes has been proposed to be limited by negative selection during T-cell maturation in the thymus where developing T-cells with a high affinity for self-antigens presented by MHC molecules are removed (Abbas et al. 2014). Thus, the higher the MHC diversity the greater the number of developing T-cells that will be lost during thymic selection, resulting in potential gaps in the T-cell repertoire (Nowak et al. 1992; Woelfing et al. 2009; Migalska et al. 2019). However, it has also been suggested that the level of MHC diversity required to curtail the T-cell repertoire through negative selection is extremely high, making it more probable that MHC diversity is limited by the increased risk of self-reactivity conferred by having very many different MHC molecules (Borghans et al. 2003). Having too few expressed MHC genes is also selected against as this reduces the number of antigens that can be presented. Thus, "optimal" expressed MHC diversity is likely to be determined by a trade-off between the benefits of recognizing many pathogens and the potential costs of immunopathology, mediated either through T-cell repertoire depletion or autoimmunity (Woelfing et al. 2009). Consequently, optimal expressed MHC diversity may vary substantially between species depending on the degree of pathogen pressure they are exposed to, which is determined by factors such as the type of environments they occupy and life-history traits (Eizaguirre and Lenz 2010; O'Connor et al. 2018; Minias et al. 2019; O'Connor et al. 2020).

Although the processes that shape optimal MHC diversity act on the level of the individual and are influenced by factors that vary within and between populations, they scale-up to generate species-level differences in MHC diversity (Woelfing et al. 2009; Eizaguirre et al. 2011; Bentkowski and Radwan 2019). Differences in optimal MHC diversity may help to explain the large between-species variation and lower within-species variation in MHC gene copy number in genomic DNA seen across many songbird species (O'Connor et al. 2018). Comparative studies of MHC diversity have focused on genomic MHC diversity, largely overlooking differences that may exist between species in the number of expressed genes (Westerdahl 2007; O'Connor et al. 2016; Minias et al. 2019) but see Drews et al. (2017) and Drews and Westerdahl (2019). The implicit assumption being that the number of MHC genes in open reading frame in the genome accurately reflects the number that are expressed for all species. However, this may not be the case, as optimality could be resolved at the level of the genome, gene expression, or both. Therefore, the question of whether all MHC genes present in the genome are expressed in any given species has more than one plausible answer from a theoretical perspective.

First, if optimality has been fully resolved at the level of the genome, that is, the number of MHC genes in the genome is optimal, then all MHC genes should be expressed. In this scenario, there would be a one-to-one relationship between the number of MHC genes in genomic DNA (gDNA) and the number of MHC genes in complementary DNA (cDNA), that is, expressed genes ("All" in Fig. 1A). The birth-and-death mode of MHC gene evolution makes this scenario unlikely, especially in species with highly duplicated MHC genes (Nei et al. 1997). However, the assumption that all MHC genes are expressed is implicit in the common practice of counting MHC alleles in gDNA to estimate functional MHC diversity. A more plausible scenario is that gene duplication events have resulted in some species having more MHC genes than the optimal number. In which case, optimal MHC diversity may be achieved by adjusting the number of expressed MHC genes. In species with more than the optimal number of MHC genes, individuals who express only a proportion of their MHC genes may be selected for, leading to species that do not express all the MHC genes available in their genomes. Surplus MHC gene copies could be silenced, for example, through the accumulation of deleterious mutations in the protein coding DNA or the promoter regions (Agrawal and Kishore 2000), leading to a lower than one-to-one relationship between the number of MHC genes in gDNA and cDNA. As the same environments that select for conserving high functional genomic MHC diversity are also likely to favor high expressed MHC diversity, there would be a generally linear relationship between the number of genomic and expressed MHC genes across species with the increase in the number of expressed genes being proportional to the increase in the number of MHC genes in the genome ("Proportion" in Fig. 1A). Alternatively, species with particularly high genomic MHC diversity may express disproportionately fewer MHC genes. This could occur either if the costs of immunopathology increase in a nonlinear fashion with MHC gene copy number or if there are species that have extremely highly duplicated MHC genes in their genomes, such that many copies become nonfunctional. This would result in an increasingly lower proportion of MHC genes being expressed in species with very high MHC diversity ("Disproportion" in Fig. 1A). Both the "Proportion" and "Disproportion" scenarios are plausible under the birth-and-death model of gene evolution.

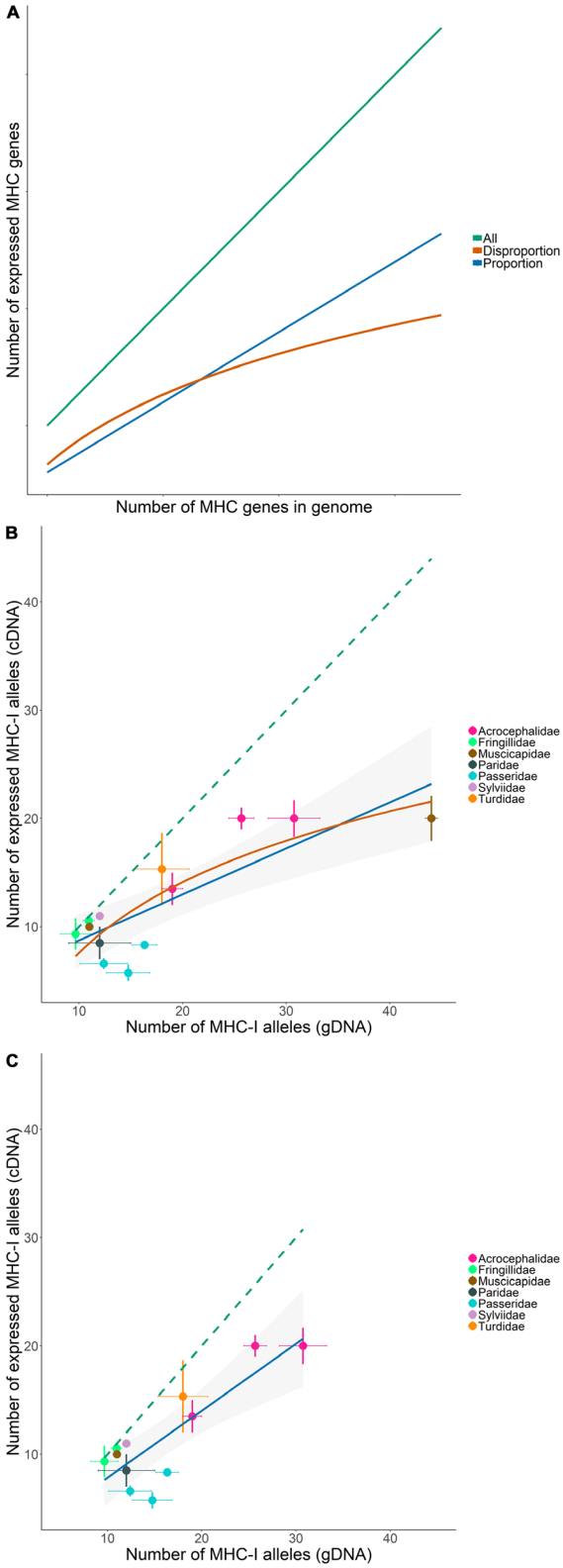

Figure 1. Theoretical and empirical relationships between genomic and expressed MHC diversity. (A) Theoretical scenarios for the relationship between the number of MHC genes in the genome and the number of expressed MHC genes across species. The green line (All) depicts a scenario in which every MHC gene is expressed. The blue line (Proportion) depicts a scenario in which proportionally fewer genes than the total number present in the genome are expressed. The orange line (Disproportion) depicts a scenario in which the number of expressed MHC genes is disproportionally lower in species with particularly high genomic MHC diversity. (B) Empirical relationship between the number of MHC-I alleles detected in genomic DNA (gDNA) and the number that are expressed, that is, detected in complementary DNA (cDNA), across 13 songbird species. Solid blue line shows the linear relationship and shading shows the 95% confidence intervals. Solid orange line shows the quadratic relationship. Dashed green line shows the theoretical one-to-one relationship (All). Each data point is colored according to the taxonomic family of each species. (C) Figure 1B excluding data from *Erithacus rubecula* and consequently ommiting the depiction of a quadratic realtionship.

Comparative datasets of MHC diversity in both gDNA and cDNA are required to test which of the scenarios, "All," "Proportion," or "Disproportion," are supported by empirical data. However, few studies have published data on gene expression in songbirds. We therefore generated such data and set out to test which of the three aforementioned theoretical scenarios best fits the patterns of genomic versus expressed MHC diversity we see across the songbird clade Passerida. We compared the number of MHC class I (MHC-I) alleles in gDNA and cDNA across 13 species with different levels of MHC-I diversity and used phylogenetically informed analyses to test which of the relationships, "All," "Proportion," or "Disproportion," is evident. By comparing across songbirds within a phylogenetic framework, our study provides novel macroevolutionary evidence of how gene expression could be instrumental in achieving optimal MHC diversity across species with highly duplicated MHC genes.

Materials and Methods

We MHC-I genotyped gDNA and cDNA samples from seven Passerida species and used this data, in combination with similar datasets from a further six Passerida species, to compare the number of MHC-I alleles in gDNA and cDNA across species with different levels of MHC-I diversity. The standard approach for studies that estimate MHC diversity in nonmodel species is currently amplicon high-throughput sequencing (HTS) in which sets of degenerate primers are used to amplify a section of an allele across all MHC genes of a particular class within an individual simultaneously (O'Connor et al. 2019). These multiplexed MHC alleles are then sequenced and filtered to eliminate artifacts, which results in an MHC-I genotype for each individual sample. As alleles cannot be assigned to specific loci, but rather to an MHC multilocus, the total number of different alleles is used in place of the number of genes. The exact relationship

between these two variables is determined by levels of heterozygosity/homozygosity at each gene but should be highly correlated. Although many studies have used the approach of counting HTS alleles from gDNA to estimate MHC diversity, very few have simultaneously investigated the number of expressed alleles (but see Biedrzycka et al. 2017; Drews et al. 2017; Drews and Westerdahl 2019).

SAMPLE COLLECTION

Blood samples (20–40 μL) were collected from the brachial veins of individuals from each of the following seven species before they were released back into the wild: Eurasian reed warbler (*Acrocephalus scirpaceus*, n = 3), European greenfinch (*Carduelis chloris*, n = 3), European robin (*Erithacus rubecula*, n = 3), Blue tit (*Cyanistes caeruleus*, n = 2), Common redstart (*Phoenicurus phoenicurus*, n = 1), Garden warbler (*Sylvia borin*, n = 1), and Eurasian blackbird (*Turdus merula*, n = 3). Blood samples were collected from a total of 16 individuals. For details of samples, see Supporting information Table S1. These species represent a broad span of the phylogenetic range of Passerida (Johansson et al. 2008). Each blood sample was split in two and stored either at −20°C in SET buffer (150 mM NaCl, 50 mM TRIS, 1 mM EDTA, pH 8.0) for DNA extraction or at −40°C in 100 μL K2EDTA and 500 μL TRIzol LS (Life Technologies, Carlsbad, CA, USA) for RNA extraction.

DNA AND RNA EXTRACTION

Genomic DNA was extracted using a standard ammonium acetate protocol (Sambrook et al. 1989). RNA was extracted with a combination of the TRIzol LS protocol (Life Technologies) and the RNeasy Mini kit (QIAGEN, Hilden, Germany). Homogenization and phase separation were performed according to the TRIzol LS protocol, resulting in an aqueous phase. One volume of 70% EtOH was added to the aqueous phase. Next, the RNeasy protocol was performed, which included a column-based DNase treatment (Chiari and Galtier 2011). The RNA (mRNA) was reverse transcribed to complementary DNA (cDNA) using the RETROscript kit (Life Technologies) according to the manufacturer's protocol.

MHC-I GENOTYPING

Genotyping was performed on replicated samples for each DNA and RNA sample. Fragments of exon 3 of MHC-I alleles were amplified using three different primer pairs (i.e., 12 different exon 3 amplicons were sequenced per individual: six for each gDNA and cDNA sample). These primer pairs amplify partially overlapping sets of MHC-I alleles (Fig. S1). See Supporting information Appendix S1 for primer details. Amplicons were high-throughput DNA sequenced using bidirectional pyrosequencing on the 454 GS FLX system by 454/Roche at Lund University Sequencing Facility (Faculty of Science). Full details of all laboratory and sequencing procedures can be found in O'Connor et al. (2018). Sequencing was performed across three separate 454 runs (for details of the distribution of samples across runs and sequencing depth see Supporting information Table S2 and Appendix S1). The raw 454 data were demultiplexed, clustered, and filtered using the program AmpliSAS (Sebastian et al. 2016). Full details of the filtering and genotyping procedures along with details of primer performance and the approach to dealing with missing data can be found in Supporting information Appendix S1.

The final set of verified alleles for each individual for each of the three primer pairs was combined to merge any identical alleles. Alleles were considered identical and merged if they had 100% sequence identity for the full length of their sequence overlap (Supporting information Fig. S1). We used the de novo assemble option within Geneious Prime version 2019 with customized sensitivity settings to identify and merge identical alleles and create a FASTA file of the final genotype for each individual. Merged alleles were approximately 239 bp in length if fragments were amplified by all three primer pairs with identical sequences in their overlapping sections (Supporting information Fig. S1). Manual curation of remaining sequences was conducted to remove any likely pseudogenes, which was assessed based on the reading frame, the absence of highly conserved amino acids, and/or the presence of stop codons (a total of seven out of 265 alleles were manually removed: four in *Turdus merula*, two in *Phoenicurus phoenicurus*, and one in *Cyanistes caeruleus*). The sequence length of 239 bp represents approximately 25% of the full sequence for MHC-I molecules (Westerdahl et al. 1999). Therefore, it is possible that evidence of pseudogenes exists outside of this region. However, based upon the evidence available in this study, all MHC-I sequences in the final dataset were assumed to encode functional MHC-I molecules.

There were 258 verified MHC-I alleles in this dataset (GenBank Accession codes: MF477947 - MF477998; MF478236 - MF478248; MF478321; MF478323 - MF478336; MF478339 - MF478420; MF478607 - MF478612; MF478622 - MF478624; MF478626; MF478678 - MF478680; MF478682 - MF478684; MF478689; MF478692 - MF478693; MF478695; MF478697 - MF478731 and MT655253- MT655301).

To rule out the possibility that differences in the number of MHC-I alleles between individuals reflected differences in sequencing depth we examined the relationship between the number of sequencing reads and the number of MHC-I alleles in gDNA and cDNA. We found no evidence that sequencing depth determined the number of MHC-I alleles in the genotypes for individuals in either gDNA (Supporting information Fig. S2, Spearman's rank correlations $r = 0.01$, $P = 0.99$, Supporting information Table S2) or cDNA (Supporting information Fig. S2, $r = 0.15$, $P = 0.58$, Table S2). It is

clear from inspecting the data that there are three individuals with notably higher reads for gDNA than the other individuals (Supporting information Fig. S2). However, even after the removal of the data from these individuals there remained no significant relationship between read number and the number of MHC-I alleles detected in gDNA ($r = 0.26$, $P = 0.38$).

ADDITIONAL DATA

To increase the sample size, we added estimates of the number of MHC-I alleles in gDNA and cDNA for the following six Passerida species: Great reed warbler (*Acrocephalus arundinaceus*, n = 2), Sedge warbler (*Acrocephalus schoenobaenus*, n = 4), House sparrow (*Passer domesticus*, n = 5), Spanish sparrow (*Passer hispaniolensis*, n = 3), Tree sparrow (*Passer montanus*, n = 4), and Eurasian siskin (*Spinus spinus*, n = 18). Total n = 36 individuals. The data for these species came from three published studies: *Passer* species from Drews et al. (2017), *S. spinus* from Drews and Westerdahl (2019), *A. schoenobaenus* from Biedrzycka et al. (2017), and one currently unpublished *A. arundinaceus* dataset from the research group of H. Westerdahl. All four studies used generally comparable HTS and filtering methods to those described above. Drews et al. (2017) used 454, the same technology as the current study, whereas the other three studies used illumina MiSeq. A study comparing the performance or 454 and illumina MiSeq for MHC-I genotyping in *P. domesticus* reported highly similar MHC-I genotypes, suggesting that differences between these approaches should not have had a significant impact on the comparability of this data (Razali et al. 2017). In addition, differences in filtering strategy have been shown to have only a minor impact on the final MHC-I genotypes, given sufficient coverage, in a study on *A. schoenobaenus*, which is a species with very high MHC-I diversity in gDNA (Biedrzycka et al. 2017). Nevertheless, different approaches to primer design, sequencing and genotyping can affect estimates of MHC diversity. Therefore, we only used data from studies that we had full working knowledge of (all were supervised by H. Westerdahl) to ensure the additional data were comparable within the standardized framework of our current study. Furthermore, each of these studies used the same HTS and filtering methods for both cDNA and gDNA samples, ensuring that the genotypic and expressed MHC-I genotypes were comparable within studies.

DATA ANALYSES

We investigated the relationship between the number of MHC-I allele in gDNA and cDNA across species using Bayesian Phylogenetic Mixed Models (BPMM) implemented in the R package "MCMCglmm" (Hadfield 2010). The response variable was the number of MHC-I alleles in cDNA per individual and the fixed effects were the number of MHC-I alleles in gDNA, as well as a quadratic function of the number of gDNA alleles ($gDNA^2$). The fixed effect of the number of MHC-I alleles in gDNA allowed us to test for a linear relationship between the number of gDNA and cDNA MHC-I alleles and to test whether this was significantly lower than one-to-one. A one-to-one relationship would be consistent with the "All" scenario, whereas a lower than one-to-one relationship would be consistent with the "Proportion" scenario. The quadratic term was used to test for a nonlinear relationship between the number of gDNA and cDNA MHC-I alleles, consistent with the "Disproportion" scenario. Species was included as a random effect in the model to account for the nonindependence of multiple individuals from the same species. A Poisson error distribution was used to model the number of alleles.

To account for the nonindependence of data due to species ancestry, we included a phylogenetic relationship matrix as a random effect in all models. Using the subset tool on the Bird Tree website (http://birdtree.org/), we downloaded a sample of 1500 trees from the posterior distribution of the Hackett all-species backbone tree for the species in our dataset (Jetz et al. 2012). We ran our models on all 1500 trees to accommodate uncertainty in phylogenetic relationships, discarding the first 500 trees as a burn-in. For each tree we ran 10,000 iterations with a burn-in of 9999 and saved the final iteration resulting in a posterior sample of 1000 estimates. Parameter estimates were summarized using the posterior mode (PM), 95% credible interval (CIs) and pMCMC values (the number of iterations greater or less than zero divided by the total number of iterations). Terms were considered statistically significant when 95% CIs did not span 0 and pMCMC values were less than 0.05 (Hadfield 2010). We specified inverse-Wishart priors (V = 1, nu = 0.002) for all random effects that led to all models converging. Model convergence was tested by repeating each analysis three times and examining the correspondence between chains in R using the "coda" package version 0.16-1 (Plummer et al. 2016) by: (1) visually inspecting the traces of the MCMC posterior estimates and their overlap; (2) calculating the autocorrelation and effective sample size of the posterior distribution of each chain; and (3) using Gelman and Rubin's convergence diagnostic test, which compares within- and between-chain variance using a potential scale reduction factor (Gelman and Rubin 1992). Values above 1.1 indicate chains with poor convergence properties. The potential scale reduction factor was less than 1.1 for all the parameter estimates presented.

As the number of individuals sampled from each species varied, we tested whether this influenced our results by also running models on species-level data with the number of individuals sampled per species included as a covariate. We also tested how the results of our analyses were shaped by the addition of data from the other studies by running a model on a dataset that only included the seven species genotyped specifically for this study.

Full details of all model specifications and results can be found in Supportinng information Tables S3 to S6.

Table 1. Mean number of MHC-I alleles detected in genomic DNA (gDNA) and complementary DNA (cDNA) and the proportion of expressed MHC-I alleles.

Species	Family	Individuals sampled	Mean MHC-I alleles (gDNA)	Mean MHC-I alleles (cDNA)	Proportion of expressed MHC-I alleles
Acrocephalus arundinaceus	Acrocephalidae	2	19.0	13.5	0.71
Acrocephalus schoenobaenus	Acrocephalidae	4	30.8	20.0	0.65
Acrocephalus scirpaceus	Acrocephalidae	3	25.7	20.0	0.78
Sylvia borin	Sylviidae	1	12.0	11.0	0.92
Phoenicurus phoenicurus	Muscicapidae	1	11.0	10.0	0.91
Erithacus rubecula	Muscicapidae	3	44.0	20.0	0.45
Turdus merula	Turdidae	3	18.0	15.3	0.84
Cyanistes caeruleus	Paridae	2	12.0	8.5	0.72
Passer domesticus	Passeridae	5	12.4	6.6	0.58
Passer hispaniolensis	Passeridae	3	16.3	8.3	0.51
Passer montanus	Passeridae	4	14.8	5.8	0.40
Carduelis chloris	Fringillidae	3	9.7	9.3	0.97
Spinus spinus	Fringillidae	18	10.9	10.6	0.97

Species names are colored according to their taxonomic family in line with Figure 1.

Results and Discussion

The mean number of MHC-I alleles in gDNA per individual across species ranged from 9.7 to 44.0 (Table 1) and the proportion of these that were expressed varied considerably (40 to 97%). In the cases where there were several species per family, that is, Acrocephalidae and Passeridae, the proportion of expressed alleles appeared to be fairly conserved within families (Table 1).

Overall, we found strong support for a linear positive relationship between the number of MHC-I alleles detected in gDNA and cDNA, which was significantly lower than one-to-one as demonstrated by the CIs spanning a range well below one (PM = 0.0330, CI = 0.0179 to 0.0482, pMCMC < 0.001, Fig. 1B, Supporting information Table S3). This shows that fewer MHC-I alleles are expressed than the total number available, which is consistent with the "Proportion" scenario (Fig. 1A). There was also a significant quadratic relationship between the number of MHC-I alleles in gDNA and cDNA (PM = −0.0979, −0.2036 to −0.0075, pMCMC = 0.02, Fig. 1B, Supporting information Table S3), indicating a nonlinear decrease in the number of expressed alleles as the number of gDNA alleles increases. Furthermore, no species expressed more than an average of 20 MHC-I alleles, despite there being three species with over 20 MHC-I alleles in gDNA: *A. scirpaceus* had 25.7, *A. schoenobaenus* had 30.8 and *E. rubecula* had 44.0 MHC-I alleles per individual in gDNA. This supports the "Disproportion" scenario in which the number of expressed MHC genes is disproportionally lower in species with particularly high MHC diversity (Fig. 1A). However, it is clear from inspecting the data that the significance of the quadratic term was likely to have been driven by *E. rubecula*, which expressed fewer than half of their gDNA alleles (Table 1). Indeed, when *E. rubecula* was removed from the analysis the significant quadratic relationship was no longer evident (PM = −0.0901, CI = −0.2302 to 0.0596, pMCMC = 0.10, Fig. 1C, Supporting information Table S4). The results from *E. rubecula* could reflect a general feature of species with very high MHC-I diversity, but as there were no other species with comparable MHC-I diversity in the current study we cannot be certain. However, in other taxa with very many MHC genes, such as neoteolost fish, species have also been shown to limit the number of expressed MHC genes. For example, Nile tilapia (*Oreochromis niloticus*) express fewer than half of their 17 MHC-I genes (Murray et al. 2000). Future studies involving more songbird species with high MHC-I diversity are required to determine whether there is a limit to the number of expressed MHC-I genes across songbirds in line with the "Disproportion" scenario.

A similar pattern of a lower than one-to-one relationship between the number of MHC-I alleles in gDNA and cDNA, with evidence of a quadratic relationship, was also found when we analyzed the dataset containing only the seven species genotyped for the current study (Supporting information Table S5). This suggests that the addition of data from other studies served to strengthen patterns that were evident even in this smaller dataset. We also found these patterns were robust to differences in the number of individuals sampled from each species (Supporting information Table S6). Visual inspection of the raw data shows that the number of MHC-I alleles, both in gDNA and cDNA, was generally similar among individuals of the same species

(Supporting information Table S1). Although this is based on a small number of individuals and it is known that numbers of MHC-I alleles in gDNA can vary more within species when more individuals are sampled (e.g., Biedrzycka et al. 2017). Few studies on songbirds have investigated the proportion of expressed alleles in more than a handful of individuals. To our knowledge, Drews and Westerdahl (2019) reports data on the most individuals in terms of the number of expressed MHC-I genes (n = 18) and these results indicate that within Siskins (*S. spinus*) the number of expressed MHC-I alleles is fairly conserved. Furthermore, in our data the proportion of expressed MHC-I alleles appear to be more similar within than between species (Tables 1 and Supporting information Table S1). Taken together, these findings suggest that the proportion of expressed MHC-I alleles may be stable within species and is therefore unlikely to vary greatly over time.

The finding that songbirds express fewer MHC-I alleles than the number they possess is in line with the concept that gene expression enables genomic MHC diversity to be adjusted to an optimal level, as has been previously suggested by Drews et al. (2017) for three songbird species belonging to the family Passer. Expressing fewer alleles than the number present in the genome is also consistent with the expectations of the birth-and-death model of multigene family evolution, in which some duplicated gene copies become nonfunctional over time (Nei et al. 1997; Edwards et al. 2000; Hess et al. 2000). These pseudogenes are silenced in the genome and would not be expected to be expressed. Although we only included MHC-I alleles that appeared functional in the current study, that is, alleles in open reading frame, there could be deleterious mutations outside the sequenced exon resulting in pseudogenization. The occurrence of pseudogenized alleles is increasingly likely in species with extremely highly duplicated MHC genes, which could partly explain why we saw evidence of disproportionally fewer expressed MHC-I alleles in species with very many MHC-I alleles in gDNA. However, we lack sufficient species with very high MHC-I diversity in this dataset to statistically disentangle the "Disproportion" and "Proportion" relationships.

The presence of nonclassical MHC-I genes in our dataset could also influence the number of expressed alleles detected. Nonclassical MHC genes exhibit low polymorphism, limited expression, and have a less well-defined role in adaptive immunity than classical MHC genes (Rodgers and Cook 2005). However, the occurrence of nonclassical MHC-I genes across the songbird radiation has not been well-characterized. Putatively nonclassical MHC-I genes have been reported in the songbird families Passeridae and Fringallidae, although these nonclassical genes are expressed in blood and are not orthologous to nonclassical MHC genes in Galliformes (MHC-Y genes) (Drews et al. 2017; Drews and Westerdahl 2019). Therefore, we do not expect putatively nonclassical MHC genes to have any large effect on the expression profiles in our dataset. The limited knowledge of nonclassical MHC-I genes in songbirds reflects the need for a more detailed characterization of the MHC gene region across this taxonomic group. Similarly, it is not known whether MHC-I gene expression differs between tissues and as we sampled only blood the number of expressed MHC-I genes may be underestimated (Pang et al. 2013; Chen et al. 2015; Shiina and Blancher 2019).

On balance, we believe that our dataset provides robust evidence that fewer MHC-I alleles are expressed across songbirds than the total number available in gDNA ("Proportion") and suggestive evidence that the number of expressed alleles may be disproportionately lower in species with very many MHC-I genes ("Disproportion"). Therefore, studies that measure MHC-I diversity solely from gDNA may overestimate functional MHC-I diversity. Exposure to more pathogens selects for higher MHC diversity based on evidence from genomic DNA across species in songbirds (O'Connor et al. 2018) and between populations in three-spined sticklebacks, *Gasterosteus aculeatus* (Eizaguirre et al. 2011). Similar conclusions can be drawn from higher MHC polymorphism in human populations that are exposed to more pathogens (Prugnolle et al. 2005). To what extent the differences in expressed MHC-I diversity demonstrated between species in our study are explained by difference in pathogen pressure requires further research.

There are still many gaps in our knowledge of MHC gene expression in nonmodel species. A better understanding of the relationship between genomic and expressed MHC diversity is essential if we are to be able to interpret the adaptive significance of variation in MHC diversity across species. Here, we provide an important step toward this goal by showing that not all MHC-I alleles are expressed across birds on a macroevolutionary scale, especially in species with high MHC-I diversity. This provides evidence that optimal MHC-I diversity could be achieved by modulating gene expression in species with highly duplicated MHC genes.

AUTHOR CONTRIBUTIONS

EO and HW authors were responsible for designing the study, data interpretation, and writing the manuscript. EO conducted the lab work and statistical analyses.

ACKNOWLEDGMENTS

We are grateful to Anna Drews, Aleksandra Biedrzycka, and Samantha Mellinger for access to data. We wish to thank Lars Råberg for feedback on the manuscript. This project was funded by grants awarded to H. W. by the Swedish Research Council (621−2011−3674 and 2015−05149) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 679799.

CONFLICT OF INTEREST

The authors have no conflict of interests to declare.

DATA ARCHIVING

All raw data are available in Supporting information Table S1. All MHC-I sequences from the individuals genotyped for this study are publicly available on GenBank (Accession codes: MF477947-MF477998; MF478236-MF478248; MF478321; MF478323-MF478336; MF478339-MF478420; MF478607-MF478612; MF478622-MF478624; MF478626; MF478678-MF478680; MF478682-MF478684; MF478689; MF478692-MF478693; MF478695; MF478697-MF478731, and MT655253-MT655301).

LITERATURE CITED

Abbas, A. K., A. H. Lichtman, and S. Pillai. 2014. Immunological tolerance and autoimmunity. In Basic immunology: functions and disorders of the immune system. 4th ed., 171–188.Elsevier Saunders, Philadelphia.

Agrawal, S., and M. C. Kishore. 2000. MHC Class I gene expression and regulation. J. Hematother. Stem Cell Res. 9:795–812.

Bentkowski, P., and J. Radwan. 2019. Evolution of major histocompatibility complex gene copy number. PLoS Comput. Biol. 15:1–15.

Biedrzycka, A., E. O'Connor, A. Sebastian, M. Migalska, J. Radwan, T. Zając, W. Bielański, W. Solarz, A. Ćmiel, and H. Westerdahl. 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 Evol. Biol. 17:1–12.

Biedrzycka, A., A. Sebastian, M. Migalska, H. Westerdahl, and J. Radwan. 2017. Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC Class I in a passerine bird. Mol. Ecol. Resour. 17:642–655.

Borghans, J. A. M., A. J. Noest, and R. J. D.e. Boer. 2003. Thymic selection does not limit the individual MHC diversity. Eur. J. Immunol. 33:3353–3358.

Brouwer, L., I. Barr, M. Van De Pol, T. Burke, J. Komdeur, and D. S. Richardson. 2010. MHC-Dependent survival in a wild population: Evidence for hidden genetic benefits gained through extra-pair fertilizations. Mol. Ecol. 19:3444–3455.

Chen, L. C., H. Lan, L. Sun, Y. L. Deng, K. Y. Tang, and Q. H. Wan. 2015. Genomic organization of the crested Ibis MHC provides new insight into ancestral avian MHC Structure. Sci. Rep. 5:1–11.

Chiari, Y., and N. Galtier. 2011. RNA extraction from sauropsids blood: Evaluation and improvement of methods. Amphibia Reptilia 32:136–139.

Drews, A., M. Strandh, L. Råberg, and H. Westerdahl. 2017. Expression and phylogenetic analyses reveal paralogous lineages of putatively classical and non-classical MHC-I genes in three sparrow species (*Passer*). BMC Evol. Biol. 17:1–12.

Drews, A., and H. Westerdahl. 2019. Not all birds have a single dominantly expressed MHC-I gene: Transcription suggests that siskins have many highly expressed MHC-I genes. Sci. Rep. 9:1–11.

Edwards, S. V., J. Gasper, D. Garrigan, D. Martindale, and B. F. Koop. 2000. A 39-Kb sequence around a blackbird MHC class II gene: Ghost of selection past and songbird genome architecture. Mol. Biol. Evol. 17:1384–1395.

Eizaguirre, C., and T. L. Lenz. 2010. Major histocompatability complex polymorphism: Dynamics and consequences of parasite-mediated local adaptation in fishes. J. Fish Biol. 77:2023–2047.

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

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–511.

Hadfield, J. D. 2010. MCMC methods for multi-respoinse generalized linear mixed models: The MCMCglmm R package. J. Stat. Softw. 33:1–22.

Hedrick, P. W. 2002. Pathogen resistance and genetic variation at MHC loci. Evolution 56:1902–1908.

Heidel, A. J., and X. Dong. 2006. Fitness benefits of systemic acquired resistance during *Hyaloperonospora parasitica* infection in *Arabidopsis thaliana*. Genetics 173:1621–1628.

Hess, C. M., J. Gasper, H. E. Hoekstra, C. E. Hill, and S. V. Edwards. 2000. MHC Class II pseudogene and genomic signature of a 32-Kb cosmid in the house finch (*Carpodacus Mexicanus*). Genome Res. 10:613–623.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444–448.

Johansson, U. S., J. Fjeldså, and R. C. K. Bowie. 2008. Phylogenetic relationships within passerida (Aves: *Passeriformes*): A review and a new molecular phylogeny based on three nuclear intron markers. Mol. Phylogenet. Evol. 48:858–876.

Kloch, A., W. Babik, A. Bajer, E. Sinski, and J. Radwan. 2010. Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole myodes glareolus. Mol. Ecol. 19:255–265. https://doi.org/10.1111/j.1365-294X.2009.04476.x

Meyer-Lucht, Y., and S. Sommer. 2009. Number of MHC alleles is related to parasite loads in natural populations of yellow necked mice, *Apodemus flavicollis*. Evol. Ecol. Res. 11:1085–1097.

Migalska, M., A. Sebastian, and J. Radwan. 2019. Major histocompatibility complex class I diversity limits the repertoire of T cell receptors. PNAS 116:5021–5026. https://doi.org/10.1073/pnas.1807864116

Minias, P., E. Pikus, L. A. Whittingham, and P. O. Dunn. 2019. Evolution of copy number at the MHC varies across the avian tree of life. Genome Biol. Evol. 11:17–28. https://doi.org/10.1093/gbe/evy253.

Murphy, K., C. A. Jr. Janeway, P. Travers, M. Walport, and M. Ehrenstein. 2008. Janeway's immunobiology. 7th ed. Garland Science, New York.

Murray, B. W., P. Nilsson, H. S. Z. Zaleska-Rutczynska, and J. Klein. 2000. Linkage relationships and haplotype variation of the major histocompatibility complex class I A genes in the cichlid fish *Oreochromis niloticus*. Mar. Biotechnol. 2:437–448.

Nei, M., X. Gu, and T. Sitnikova. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. 94:7799–7806. https://doi.org/10.1073/pnas.94.15.7799.

Nowak, M. A., K. Tarczy-Hornoch, and J. M. Austyn. 1992. The optimal number of major histocompatibility complex molecules in an individual. Proc. Natl. Acad. Sci. 89:10896–10899. https://doi.org/10.1073/pnas.89.22.10896.

O'Connor, E. A., C. K. Cornwallis, D. Hasselquist, J. Å. Nilsson, and H. Westerdahl. 2018. The evolution of immunity in relation to colonization and migration. Nat. Ecol. Evol. 2:841–849. https://doi.org/10.1038/s41559-018-0509-3.

O'Connor, E. A., D. Hasselquist, J.-Å. Nilsson, H. Westerdahl, and C. K. Cornwallis. 2020. Wetter climates select for higher immune gene diversity in resident, but not migratory, songbirds. Proc. Royal Soc. B 287. https://doi.org/10.1098/rspb.2019.2675

O'Connor, E. A., M. Strandh, D. Hasselquist, J. Å. Nilsson, and H. Westerdahl. 2016. The evolution of highly variable immunity genes across a *Passerine* bird radiation. Mol. Ecol. 25:977–989. https://doi.org/10.1111/mec.13530.

O'Connor, E. A., H. Westerdahl, R. Burri, and S. V. Edwards. 2019. Avian MHC Evolution in the era of genomics: Phase 1.0. Cells 8:1152. https://doi.org/10.3390/cells8101152.

Oliver, M. K., S. Telfer, and S. B. Piertney. 2009. Major Histocompatibility Complex (MHC) heterozygote superiority to natural multi-parasite

infections in the water vole (*Arvicola Terrestris*). Proc. Royal Soc. B 276:1119–1128. https://doi.org/10.1098/rspb.2008.1525.

Orgil, U., H. Araki, S. Tangchaiburana, R. Berkey, and S. Xiao. 2007. Intraspecific genetic variations, fitness cost and benefit of RPW8, a disease resistance locus in *Arabidopsis thaliana*. Genetics 176:2317–2333. https://doi.org/10.1534/genetics.107.070565.

Pang, J.c., F.y. Gao, M.x. Lu, X. Ye, H.p. Zhu, and X.l. Ke. 2013. Major histocompatibility complex class IIA and IIB Genes of nile tilapia *Oreochromis niloticus*: Genomic structure, molecular polymorphism and expression patterns. Fish Shellfish Immunol. 34:486–496. https://doi.org/10.1016/j.fsi.2012.11.048.

Piontkivska, H., and M. Nei. 2003. Birth-and-death evolution in primate MHC class I genes: Divergence time estimates. Mol. Biol. Evol. 20:601–609. https://doi.org/10.1093/molbev/msg064.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2016. CODA: Convergence diagnosis and output analysis for MCMC. R News 6:7–11.

Prugnolle, F., A. Manica, M. Charpentier, J. F. Guégan, V. Guernier, and F. Balloux. 2005. Pathogen-driven selection and worldwide HLA Class I diversity. Curr. Biol. 15:1022–1027. https://doi.org/10.1016/j.cub.2005.04.050.

Radwan, J., W. Babik, J. Kaufman, T. L. Lenz, and J. Winternitz. 2020. Advances in the evolutionary understanding of MHC polymorphism. Trends Genet. 36:298–311. https://doi.org/10.1016/j.tig.2020.01.008.

Radwan, J., M. Zagalska-Neubauer, M. Cichoń, J. Sendecka, K. Kulma, L. Gustafsson, and W. Babik. 2012. MHC diversity, malaria and lifetime reproductive success in collared flycatchers. Mol. Ecol. 21:2469–2479. https://doi.org/10.1111/j.1365-294X.2012.05547.x.

Razali, H., E. O'Connor, A. Drews, T. Burke, and H. Westerdahl. 2017. A quantitative and qualitative comparison of Illumina MiSeq and 454 amplicon sequencing for genotyping the highly polymorphic Major Histocompatibility Complex (MHC) in a non-model species. BMC Res. Notes 10:1–10. https://doi.org/10.1186/s13104-017-2654-1.

Rodgers, J. R., and R. G. Cook. 2005. MHC Class IB Molecules bridge innate and acquired immunity. Nat. Rev. Immunol. 5:459–471. https://doi.org/10.1038/nri1635.

Roy, B. A., and J. W. Kirchner. 2000. Evolutionary dynamics of pathogen resistance and tolerance. Evolution 54:51–63. https://doi.org/10.1111/j.0014-3820.2000.tb00007.x.

Sambrook, J., E. F. Fritsch, and T. M. T. Maniatis. 1989. Molecular cloning: A laboratory manual. 2nd ed. Cold Spring Habour Laboratory Press, Cold Spring Habour.

Sato, A., W. E. Mayer, H. Tichy, P. R. Grant, R. B. Grant, and J. Klein. 2001. Evolution of Mhc Class II B genes in Darwin's finches and their closest relatives: Birth of a new gene. Immunogenetics 53:792–801. https://doi.org/10.1007/s00251-001-0393-9.

Sebastian, A., M. Herdegen, M. Migalska, and J. Radwan. 2016. Amplisas: A web server for multilocus genotyping using next-generation amplicon sequencing data. Mol. Ecol. Resour. 16:498–510. https://doi.org/10.1111/1755-0998.12453.

Sepil, I., S. Lachish, A. E. Hinks, and B. C. Sheldon. 2013. MHC Supertypes confer both qualitative and quantitative resistance to avian malaria infections in a wild bird population. Proc. Royal Soc. B 280. https://doi.org/10.1098/rspb.2013.0134.

Shiina, T., and A. Blancher. 2019. The cynomolgus macaque MHC polymorphism in experimental medicine. Cells 8:1–31. https://doi.org/10.3390/cells8090978.

Spurgin, L. G., and D. S. Richardson. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc. Royal Soc. B 277:979–988. https://doi.org/10.1098/rspb.2009.2084.

Wegner, K. A. S., M. Kalbe, H. Schaschl, and T. B. H. Reusch. 2004. Parasites and individual major histocompatibility complex diversity—An optimal choice? Microbes Infect. 6:1110–1116. https://doi.org/10.1016/j.micinf.2004.05.025.

Wegner, K.A. S., T. B. H. Reusch, and M. Kalbe. 2003. Multiple parasites are driving major histocompatibility complex polymorphism in the wild. J. Evol. Biol. 16:224–232. https://doi.org/10.1046/j.1420-9101.2003.00519.x.

Westerdahl, H., Waldenström, J., Hansson, B., Hasselquist, D., von Schantz T., and Bensch, S. 2005. Associations between Malaria and MHC genes in a migratory songbird. Proc. Royal Soc. B 272:1511–1518. https://doi.org/10.1098/rspb.2005.3113

———. 2007. Passerine MHC: Genetic variation and disease resistance in the wild. J. Ornithol. 148:469–477. https://doi.org/10.1007/s10336-007-0230-5

Westerdahl, H., T. V. Schantz, and H. Wittzell. 1999. Polymorphism and transcription of MHC Class I Genes in a passerine bird, the great reed warbler. Immunogenetics 49:158–170. https://doi.org/10.1007/s002510050477

Woelfing, B., A. Traulsen, M. Milinski, and T. Boehm. 2009. Does intra-individual major histocompatibility complex diversity keep a golden mean? Proc. Royal Soc. B 364:117–128. https://doi.org/10.1098/rstb.2008.0174.

Associate Editor: L. Spurgin
Handling Editor: A. G. McAdam

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Overlap in the MHC-I exon 3 sequences amplified by each of the three primer pairs (PP1, PP2 and PP3).

Figure S2 Relationship between the total number of 454 sequencing reads per individual and the number of MHC-I alleles in the (a) gDNA and (b) cDNA MHC-I genotype for that individual.

Figure S3 Relationship between the difference in sequencing depth between gDNA and cDNA and the proportion of expressed MHC-I alleles.

Table S1: Details of individual samples and the number of MHC-I alleles in gDNA and cDNA

Table S2a: Distribution of samples across the three 454 sequencing runs (Jan 2013, Dec 2013 & May 2015). Technical replicates are donated '_a' or '_b'.

Table S3: The relationship between the number of MHC-I alleles in gDNA and cDNA analyzed using a BPMM implemented in the R package 'MCMCglmm'.

Table S4: The relationship between the number of MHC-I alleles in gDNA and cDNA, excluding data from Erithacus rubecula, analyzed using a BPMM implemented in the R package 'MCMCglmm'.

Table S5: The relationship between the number of MHC-I alleles in gDNA and cDNA, including only the genotypes generated for this study (n = 7 species), analyzed using a BPMM implemented in the R package 'MCMCglmm'.

Table S6: The relationship between the number of MHC-I alleles in gDNA and cDNA, accounting for the number of individuals sampled from each species, analyzed using a BPMM implemented in the R package 'MCMCglmm'.