Diversity and selection of MHC class I genes in the Godlewski's bunting

Wei Huang[1,2,3], Boye Liu[1,4], Tobias L Lenz[2], Yangyang Peng[1], Lu Dong[1*], Yanyun Zhang[1*]

1. MOE Key Laboratory For Biodiversity Science and Ecological Engineering, Beijing Normal University, Beijing, China

2. Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany

3. Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

4. Shaanxi Institute of Zoology, Xi'an, China

*Corresponding author: donglu@bnu.edu.cn

zhangyy@bnu.edu.cn

Abstract

The major histocompatibility complex (MHC) is a multiple-copy immune gene family in vertebrates. Its genes are highly variable and code for antigen-presenting molecules. Characterization of MHC genes in different species and investigating the mechanisms that shape MHC diversity is an important goal in understanding the evolution of biological diversity. Here we developed a next generation sequencing (NGS) protocol to genotype the MHC class I genes of 326 Godlewski's buntings (*Emberiza godlewskii*) sampled in the Western mountain area of Beijing from 2014 to 2016. A total of 184 functional alleles were identified, including both non-classical and classical alleles. Classical alleles could be clustered into nine supertypes. Compared with other passerine birds, the individual diversity of MHC class I genes in Godlewski's buntings is intermediate. Ten amino acid sites in the antigen-binding domain showed signatures of positive selection and eight of them exhibit high amino acid polymorphism. These findings indicate the action of balancing selection and provide a framework for subsequent investigation of selection acting on MHC genes in Godlewski's buntings.

Introduction

The immune system is critical in pathogen resistance as it provides different mechanisms to protect the host from infection. Therefore, the genetic diversity of immune genes is assumed to be strongly linked with the infection of pathogens and is maintained through coevolution between hosts and pathogens (Pilosof et al. 2014). Among the immune genes that have been intensively studied in vertebrates, major histocompatibility complex (MHC) genes play a crucially important role in adaptive immunity (Bernatchez and Landry 2003; Edwards and Hedrick 1998; Piertney and Oliver 2006; Wilson et al. 2010). MHC molecules encoded by MHC genes are capable of binding both self-derived and pathogen-derived peptides and presenting them to T cells to invoke an adaptive immune response. There are two main classes of MHC genes: Class I genes encode molecules that are responsible for binding intracellular peptides while class II genes encode molecules that are responsible for binding extracellular peptides. MHC genes are among the most variable gene families in vertebrates and the mechanisms of driving diversity of MHC genes have received considerable attention (Piertney and Oliver 2006; Radwan et al. 2020; Spurgin and Richardson 2010).

Parasite-mediated selection is thought to be a critical driver for the polymorphism of MHC genes. Three main hypotheses have been proposed to elucidate the association between pathogens and MHC genes (Spurgin and Richardson 2010). 1) *Heterozygote advantage*: individuals which are heterozygous across the MHC can recognize a greater variety of pathogen antigens thus variations at MHC loci are selected and maintained (Takahata and Nei 1990). 2) *Rare-allele advantage*: there is strong selection on pathogens to evade immune protection provided by the most common MHC alleles resulting in rare MHC alleles having a selective advantage (Bodmer 1972). However, the

advantage disappears when the frequency of a protective MHC allele increases and this will maintain MHC variations by negative frequency-dependent selection (Slade and Mccallum 1992; Takahata and Nei 1990). 3) *Fluctuating selection*: a changing environment results in different parasites being abundant in space and time which generates directional selection in different subpopulations at different time points to maintain MHC diversity (Hedrick 2002). Although each of the three hypotheses can be supported by certain studies, the relative importance of these three hypotheses is still not clear (Radwan et al. 2020; Spurgin and Richardson 2010). Besides the paradigm of parasite-mediated selection, sexual selection is viewed as another important force to drive MHC diversity (Milinski 2006). Mate choice based on MHC diversity or MHC compatibility has been reported in several studies (Huchard et al. 2010; Sin et al. 2015) while other studies reported negative results (Paterson and Pemberton 1997). Similar with parasite-mediated selection, it is hard to draw a clear conclusion about MHC-dependent selection from these mixed results and thus more case studies are needed to figure out the detailed selection mechanisms that shape MHC variations.

The number of MHC gene copies varies between different species. In avian species, Chickens are believed to have "minimum essential MHC" with only two class I and class II genes (Kaufman et al. 1999). However, Passerine birds possess the most variable and complex MHC systems in vertebrates (Westerdahl 2007). For instance, a study of MHC class I genes across Passerida has found the number of MHC genes per individual varies from 3-7 in bluethroat to 27-35 in willow warbler (O'Connor et al. 2016). A more recent study of sedge warbler *(Acrocephalus schoenobaenus)* reported the highest diversity of MHC genes in passerine birds with a maximum of 65 alleles per individual (Biedrzycka et al. 2017a). However, despite the high diversity of MHC genes in passerine birds, some alleles are limited expressed (non-classical allele) while others are highly expressed (classical allele). Patterns of classical alleles and non-classical alleles have been identified in three sparrow species which indicates the complexity of Passerine MHC genes (Drews et al. 2017). Thus, carefully characterizing MHC genes in passerine birds is crucially important for understanding the mechanisms shaping MHC diversity because of the high diversity and complexity of MHC genes in passerine birds. Previous studies of Passerine MHC genes suffered from technical problems to genotype passerine MHC genes efficiently and accurately. With the recent application of next generation sequencing (NGS) methods in MHC genotyping (Babik 2010; Babik et al. 2009), passerine MHC genes have attracted considerable attentions and the number of related studies has increased rapidly (Biedrzycka et al. 2017a; Dunn et al. 2013; Jones et al. 2014; Promerova et al. 2012; Sepil et al. 2012; Zagalska-Neubauer et al. 2010).

Godlewski's Bunting (*Emberiza godlewskii*) is a common resident bird in mountains of North China. Almost 90% of the Godlewski's Buntings in Beijing were infected by haemosporidian parasites and most of the infections were caused by three dominant *Plasmodium* lineages with annual stability (Liu et al. 2019), which makes it an interesting study system to explore local adaptation and host-parasite

coevolution. In this study, we characterized MHC class I exon 3 of MHC class I genes in Godlewski's buntings. Exon 3 encodes an important part of the peptide binding domain of MHC class I molecules. MHC class I genes are more polymorphic and associated with parasite resistance and mate choice in other passerine birds (Griggio et al. 2011; Westerdahl et al. 2005). After establishing an NGS-based genotyping method to identify MHC alleles in Godlewski's buntings, our further analysis focused on the following aspects: 1) Characterize the population level and individual level of genetic diversity of MHC class I genes in Godlewski's buntings. 2) Determine the functional variation of MHC class I genes in Godlewski's buntings. 3) Identify signatures of selection on MHC genes in Godlewski's buntings.

Methods

Sampling

We captured 326 Godlewski's buntings by mist net and collected blood samples from brachial vein between 2014 and 2016 in Beijing, China (Sampling Approval by Beijing Normal University: No. CLS-EAW-2013-007). Seven blood samples were stored in RNAfixer (Bioro yee, Beijing, CHN) for RNA extraction while the others were stored in ethanol.

DNA and RNA extraction

Total genomic DNA was extracted from blood samples using the TIANamp genomic DNA kit (TIANGEN). RNA was extracted using the RNAstore Blood RNA Kit (Bioro yee) and reverse transcribed to complementary DNA (cDNA) using the PrimeScript™ RT reagent Kit with gDNA Eraser (Perfect Real Time) (TaKaRa, Beijing, CHN) according to the manufacturer's protocol.

PCR amplification and NGS sequencing

To amplify the partial exon 3 of MHC class I genes, the established primers MHCD-F/R (F:TTMYGGCTGTGACCTCCTG, R: TTGCGCTYCAGCTCTTTC) were used to amplify a fragment between 205 bp and 222bp (Sepil et al. 2012). Each of the two primers was extended with a 6-bp barcode at the 5' end. Unique combinations of barcodes in the primer pairs for each sample enable pooled sequencing and subsequent reassignment to samples after sequencing. PCR was performed in a 20ul reaction with 1ul of each primer, 1ul DNA, 7ul ddH2O, and 10ul PCR Mastermix (TaKaRa). PCR conditions included an initial denaturation at 94 degrees C for 5mins, followed by 35 cycles of denaturation at 94 degrees for 30 seconds, annealing at 63 degrees for 30 seconds and extension at 72 degrees for 30 seconds with a final step at 72 degrees for 7 minutes. Successful PCR products verified by electrophoresis were then pooled for Illumina sequencing. In addition, 22 individuals were amplified twice as technical duplicates with the same genotyping strategies.

MHC allele genotyping and classification

After amplicon sequencing, Trimmomatic v 0.36 (Bolger et al. 2014) was used to filter the sequencing data from genomic DNA and RNA samples by removing: 1) adapter sequences; 2) the reads that more than 10% sites were missing and 3) the reads more than 50% sites were lower than Q20. All qualified reads were uploaded to the Amplisas server (Biedrzycka et al. 2017b; Sebastian et al. 2016) for MHC allele genotyping. Amplisas is an online server for analysing amplicon sequencing result that integrates demultiplexing, variant clustering and allele filtering by user-specified parameters. Reads were first assigned to individuals by the unique barcode combinations and then clustered by default parameters. Amplisas includes a set of rules that aid in the detection and removal of PCR artefacts, which are prone to occur during multi-locus amplification and are a common problem in any MHC genotyping approach (Lenz and Becker 2008). To further separate sequencing artefacts from putatively true alleles, the following criteria were applied: 1) remove all the variants for which maximum per amplicon frequency depth (MPAF) was lower than 1%; and 2) exclude variants that appeared only in a single individual in the dataset to minimize the effect of PCR artefacts. Given that MHC genes often exhibit an excess of rare alleles, the latter criterion might bias the retained allele repertoire towards higher-frequency alleles. However, we here chose this approach to achieve higher confidence in the allele calls and also because of the large sample size. For other studies, corresponding filtering settings need to be chosen carefully with regard to the sample size and the specific analyses that are intended.

Functional allele identification and phylogenetic analysis

Past studies on MHC have revealed three broad classes of allelic sequences that can be detected by amplicon sequencing: 1) *non-functional alleles*, which are characterized by premature stop-codons or other features that prevent their translation and which are usually derived from MHC pseudogenes that are often remnant of the birth-and-death model by which the MHC gene family is thought to evolve (Nei et al. 1997). 2) *non-classical alleles*, which are characterized by low polymorphism and no or low expression as well as a lack of signatures of selection, presumed to derive from genes coding for non-classical MHC molecules that serve basal functions in antigen-loading, and 3) *classical alleles*, which are characterized by high sequence polymorphism, signatures of selection and moderate to high expression. We therefore first examined whether alleles bear complete open reading frame by sequence alignment and those who do not bear a complete open reading frame were considered as non-functional alleles. Previous studies in passerines have found classical as well as substantial numbers of non-classical MHC alleles (Biedrzycka et al. 2017b; Drews et al. 2017). Compared with classical alleles, non-classical ones have lower expression levels and different evolutionary histories. We therefore carried out an additional filter step based on phylogenetic

analysis and expression level to investigate the possibility of non-classical alleles in our study system. This analysis identified a highly supported monophyletic clade that only included non-expressed alleles and was thus designated as non-classical clade and excluded from further analysis.

To infer the phylogeny of MHC alleles, nucleotide substitution models were first tested with jModelTest (Version 2.1.3) and the best-fit model (TN93+I+G) was determined based on Bayesian Information Criterion (BIC). The Bayesian tree was then reconstructed using BEAST (Version1.8.0) and the Markov chain Monte Carlo (MCMC) was set with the length of the chain as 1×10^7. Finally, the maximum credibility tree was estimated by TreeAnnotator (Version 1.8.0) and the selected tree was adjusted and analyzed with FigTree (Version 1.4.3). Two phylogenetic trees were constructed for all 478 alleles and alleles that were 212bp-long and 215-bp long respectively.

Positive selection on MHC alleles

To examine positively selected sites (PSSs) in classical MHC genes, we calculated the ratio of nonsynonymous (d_N) to synonymous sites (d_S) by CodeML, implemented in PAML (Yang 2007). A d_N/d_S ratio >1 indicates positive selection whereas ratios <1 indicates negative (purifying) selection. Here, we used "sites" models implemented in CodeML that allow d_N/d_S to vary across amino acid sites to examine which sites are under strong positive selection. Two hypotheses of codon evolution were tested here, using two pairs of models in CodeML. Models M1a and M7 represent nearly neutral codon substitution models, while M2a and M8 are positive selection codon substitution models that allow sites to have d_N/d_S >1. A likelihood ratio test was used to identify the model that better supports the data (M1a vs. M2a and M7 vs. M8) as well as an empirical Bayes method to identify PSSs that have a d_N/d_S ratio significantly above 1. Finally, we calculated the nucleotide diversity of PSS, non-PSS, and all sites together.

Supertype Identification

The amino acid residues coding for the peptide-binding region (PBR) of the MHC molecule are expected to be most variable across MHC variants and usually bear signatures of strong positive selection, as they are largely responsible for the specific binding and presenting of pathogen-derived peptides. However, MHC research in non-model species suffers from the difficulty to accurately identify PBR sites, so PSSs are commonly used as a proxy for characterizing the specific binding properties of MHC alleles instead. In humans, the PBR and PSS sets do indeed overlap substantially (Furlong and Yang 2008; Reche and Reinherz 2003). In order to identify supertypes, groups of MHC variants that share similar binding properties, we first characterized the physicochemical properties of PSS amino acid by five descriptor values hydrophobicity (z1), steric bulk (z2), polarity (z3), and electronic effects (z4 and z5) (Doytchinova and Flower 2005; Sandberg et al. 1998) and then translated into a matrix. This matrix was then used for K-mean clustering algorithm and discriminant analysis of principal components (DAPC) using "adegenet" package in R (Jombart 2008). The

optimal number of supertypes was determined by Bayesian Information Criterion score. Once the number of clusters was chosen, DAPC was applied to visualize the relationship between supertypes by drawing a scatterplot using the first two PCs (Jombart et al. 2010).

Results

MHC allele diversity and classification

Following the Amplisas genotyping pipeline, a total of 478 unique sequence variants of MHC I Exon 3 were called from 326 buntings, 282 of that were identified in at least two buntings. There was some length variability, with mainly three lengths of the variants found in this study, 212bp, 214bp and 215bp. After alignment with MHC class I alleles of great tits *(Parus major)*, we found only 212bp-long and 215bp-long variants to bear a complete open reading frame. Therefore, the other alleles were considered as non-functional alleles and removed from further analysis. In summary, a total of 184 alleles satisfied the initial filter requirements for functional alleles. An average of 86.7% of the alleles was shared between the technical replicates (n = 22) included in sequencing and genotyping. The maximum number of alleles per individual is 21 with a mean of 10.6 ± 2.6 alleles per individual, suggesting copy number variation of the MHC class I genes, with up to at least 11 loci in Godlewski's bunting.

There were several monophyletic clades identified in the Bayesian tree for the 184 functional alleles with a length of 212bp and 215bp. The 12 alleles with expression evidence from RNA of a subset of birds were distributed throughout the tree, except for one distinct clade. This monophyletic clade (Clade 1) is composed of only 215bp-long alleles and exhibits no alleles with evidence of expression (Fig 1). In addition, we also found that clade 1 clustered with the 214bp-long alleles that were identified as non-functional in a phylogenetic tree using all 478 primarily identified sequence variants (Supplementary information). Therefore, alleles in clade 1 were considered as non-classical alleles, leaving a total of 160 classical alleles identified in the investigated buntings. The number of classical alleles varied among individuals from one to eighteen with an average number of 8.4 ± 2.3 classical alleles per individual. The maximum number of classical alleles per individual suggested that there are up to nine classical MHC class I loci. From these classical alleles, nine supertypes (Fig 2) were identified in the population with a mean of 5.3 ± 1.2 supertypes per individual (table 1).

Positive selection analysis

Between the nearly neutral models (M1 / M7) and positive selection models (M2 / M8), the positive selection models fitted the data better in PAML. Ten positively selected sites (Fig 3) were identified by Bayes empirical analysis in both models (M2 and M8). All of these ten sites were segregating and

eight of them were highly variable, bearing at least four different amino acids. Compared with all sites (π = 0.147) and only non-positively selected sites (π = 0.106), positively selected sites have an outstanding high nucleotide diversity (π = 0.481).

Discussion

MHC variation in godlewski's bunting

Passerine birds have been reported to have complex MHC systems with a large number of MHC alleles (and thus gene copies) as well as extensive existence of pseudogenes (Westerdahl 2007). Prior to the usage of next generation sequencing (NGS) in MHC genotyping, the identification of passerine MHC genes was hampered significantly by high cost and incomplete allele identification because of limited sequencing depth. With the advent of NGS technology, an increasing number of studies have started applying this methodology to genotype passerine MHC genes, yielding improved results with lower cost and a higher number of alleles identified (Biedrzycka et al. 2017a; Dunn et al. 2013; Jones et al. 2014; Promerova et al. 2012; Sepil et al. 2012; Zagalska-Neubauer et al. 2010). It is still a matter of discussion whether the increase in number of alleles per individual, which is often observed upon switching from older methods to NGS-based protocols, indicates that some alleles were previously missed, or whether NGS-based methods are more prone to overestimating allele number. However, NGS technology is undoubtedly facilitating MHC genotyping in many species that could otherwise not be studied in this context, and this is particularly true for passerine birds. In this study, we established an NGS-based protocol to successfully genotype the MHC class I genes in Godlewski's bunting and identified 160 functional alleles in a large population of the buntings. The overall repeatability of the allele calling in our dataset was 86.7%. Similar levels of repeatability have also been reported in other NGS-based genotyping studies and the inconsistency of results may either come from "allelic dropout", i.e. missed alleles due to the stochastic nature of PCR amplification (Biedrzycka et al. 2017b), or from mis-incorporation of PCR artefacts, which are prone to occur specifically with high PCR cycle numbers, such as used here. Future genotyping efforts should thus aim to minimize the number of PCR cycles further and also consider employing other measures to avoid artefact formation (Lenz and Becker 2008).

The average number of functional alleles per individual (8.4 ± 2.3) in Godlewski's buntings falls into the range of individual MHC variability in other passerine birds, such as great tit (23.8 ±3.9) (Sepil et al. 2012), rufous-collared sparrow (*Zonotrichia capensis*) (4.4 ±1.8) (Jones et al. 2014), common yellowthroat *(Geothlypis trichas)* (8.4 ±0.4) (Dunn et al. 2013), house sparrow *(Passer domesticus)* (4.7 ±1.5) (Karlsson and Westerdahl 2013) and scarlet rosefinch *(Carpodacus erythrinus)* (8.0 ±1.6) (Promerova et al. 2012) which were also genotyped by NGS-based methods. Compared with other passerine birds, Godlewski's bunting thus has an intermediate level of individual diversity at MHC class I genes.

Classical and non-classical MHC alleles

Among the initially identified MHC class I sequences, we found a substantial number of non-functional sequence variants. These variants were readily identified by lacking an open reading frame, and are likely representing MHC pseudogenes, i.e. remnants of the birth-and-death process of MHC gene evolution (Nei et al. 1997). Such non-functional sequence variants have also been detected in previous studies of other passerines (Jones et al. 2014; Sepil et al. 2012). More interestingly, we also found a clear pattern of subdivision among the putatively functional MHC class I alleles in Godlewski's bunting. A group of alleles clustered in a monophyletic clade with shallow branches. Analysis of MHC gene expression corroborated that those putatively non-classical alleles exhibit limited or no expression in blood. Indeed, a similar pattern has been reported in a previous study, revealing the existence of both classical and non-classical MHC I alleles in house sparrows. Compared to classical MHC alleles, non-classical alleles appear to have a different evolutionary history and exhibit limited expression (Drews et al. 2017). Thus, our results reflected the existence of classical and non-classical MHC alleles in Godlewski's buntings.

Due to the more limited availability of RNA samples, the expression of MHC I alleles is rarely studied comprehensively in other passerine birds. However, without using expression data and careful phylogenetic analysis, these patterns might be missed. This could in turn lead to inaccurate identification and classification of MHC alleles and thus overestimate the functional MHC diversity and bias downstream analysis. Here, our results emphasize the importance of using integrative methods, which combined phylogenetic reconstruction and gene expression analysis to identify classical MHC alleles.

Selection on MHC genes in Godlewski's buntings

Ten amino acid residues in the exon 3 of MHC class I genes of Godlewski's buntings showed signatures of positive selection, eight of them exhibiting a particularly high number of variants (more than four different amino acids). In addition, nucleotide diversity was found to be comparatively higher in positively selected sites. These patterns indicate that MHC class I genes in godlewski's buntings evolve under strong balancing selection.

However, these patterns can only suggest the existence of historical selection. Further evidence of contemporary selection on MHC genes is essential to clarify selection on MHC genes in Godlewski's buntings and thus elucidates immunogenetic adaptation of this species to its pathogen community. Heterozygote advantage, negative frequency-dependent selection and fluctuation selection are believed to be three main selection mechanisms that drive the maintenance of MHC polymorphism (Radwan et al. 2020; Spurgin and Richardson 2010). To determine and differentiate mechanisms of

contemporary selection, we can compare MHC diversity with expected diversity under neutrality by either examining proportions of genotypes within populations or comparing population structure of MHC genotypes across different populations. Alternatively, the most powerful and popular way is to investigate MHC-parasite associations (Spurgin and Richardson 2010). Avian malaria is widespread among passerine birds (Bensch et al. 2009; Clark et al. 2014) and has been found to be associated with MHC genes in some passerine species, indicating the potential role of malaria parasites in shaping the diversity of MHC genes (Dunn et al. 2013; Jones et al. 2015; Loiseau et al. 2011; Sepil et al. 2013; Westerdahl et al. 2005). For example, in sedge warblers, prevalence of avian malaria was found to be linked with allele frequency changes of some MHC class I supertypes and individuals with more MHC supertypes were discovered to have a higher resistance to avian malaria (Biedrzycka et al. 2018). The Godlewski's buntings in our study area have been found to bear a high prevalence of haemosporidian parasites which contributed mainly by three dominant specialist lineages (Liu et al. 2019). This raises the possibility that the diversity of MHC class I genes in Godlewski's buntings could be shaped by its interaction with malaria parasites. Future work on our study system could thus focus on testing associations between malaria parasites and MHC alleles or MHC supertypes to investigate the mechanism of contemporary selection on MHC genes in this species.

Reference

Babik W (2010) Methods for MHC genotyping in non-model vertebrates. Molecular Ecology Resources 10*:*237-251

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. Molecular Ecology Resources 9*:*713-719

Bensch S, Hellgren O, Perez-Tris J (2009) MalAvi: a public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages. Molecular Ecology Resources 9*:*1353-1358

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? Journal of Evolutionary Biology 16*:*363-377

Biedrzycka A, Bielanski W, Cmiel A, Solarz W, Zajac T, Migalska M, Sebastian A, Westerdahl H, Radwan J (2018) Blood parasites shape extreme major histocompatibility complex diversity in a migratory passerine. Molecular Ecology 27*:*2594-2603

Biedrzycka A, O'Connor E, Sebastian A, Migalska M, Radwan J, Zajac T, Bielanski W, Solarz W, Cmiel A, Westerdahl H (2017a) Extreme MHC class I diversity in the sedge warbler (Acrocephalus schoenobaenus); selection patterns and allelic divergence suggest that different genes have different functions. Bmc Evolutionary Biology 17

Biedrzycka A, Sebastian A, Migalska M, Westerdahl H, Radwan J (2017b) 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

Bodmer WF (1972) Evolutionary Significance of the HL-A System. Nature 237*:*139-145

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30*:*2114-20

Clark NJ, Clegg SM, Lima MR (2014) A review of global diversity in avian haemosporidians (Plasmodium and Haemoproteus: Haemosporida): new insights from molecular data. International Journal for Parasitology 44*:*329-338

Doytchinova IA, Flower DR (2005) In silico identification of supertypes for class II MHCs. J Immunol 174*:*7085-95

Drews A, Strandh M, Raberg L, Westerdahl H (2017) Expression and phylogenetic analyses reveal paralogous lineages of putatively classical and non-classical MHC-I genes in three sparrow species (Passer). BMC Evol Biol 17*:*152

Dunn PO, Bollmer JL, Freeman-Gallant CR, Whittingham LA (2013) Mhc Variation Is Related to a Sexually Selected Ornament, Survival, and Parasite Resistance in Common Yellowthroats. Evolution 67*:*679-687

343 Edwards SV, Hedrick PW (1998) Evolution and ecology of MHC molecules: from genomics to sexual
344 selection. Trends in Ecology & Evolution 13*:*305-311
345 Furlong RF, Yang Z (2008) Diversifying and purifying selection in the peptide binding region of
346 DRB in mammals. Journal of Molecular Evolution 66*:*384-394
347 Griggio M, Biard C, Penn DJ, Hoi H (2011) Female house sparrows "count on" male genes:
348 experimental evidence for MHC-dependent mate preference in birds. Bmc Evolutionary
349 Biology 11
350 Hedrick PW (2002) Pathogen resistance and genetic variation at MHC loci. Evolution 56*:*1902-1908
351 Huchard E, Knapp LA, Wang JL, Raymond M, Cowlishaw G (2010) MHC, mate choice and
352 heterozygote advantage in a wild social primate. Molecular Ecology 19*:*2545-2561
353 Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers.
354 Bioinformatics 24*:*1403-5
355 Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new
356 method for the analysis of genetically structured populations. Bmc Genetics 11
357 Jones MR, Cheviron ZA, Carling MD (2014) Variation in positively selected major histocompatibility
358 complex class I loci in rufous-collared sparrows (Zonotrichia capensis). Immunogenetics
359 66*:*693-704
360 Jones MR, Cheviron ZA, Carling MD (2015) Spatially variable coevolution between a
361 haemosporidian parasite and the MHC of a widely distributed passerine. Ecol Evol 5*:*1045-60
362 Karlsson M, Westerdahl H (2013) Characteristics of MHC Class I Genes in House Sparrows Passer
363 domesticus as Revealed by Long cDNA Transcripts and Amplicon Sequencing. Journal of
364 Molecular Evolution 77*:*8-21
365 Kaufman J, Milne S, Gobel TWF, Walker BA, Jacob JP, Auffray C, Zoorob R, Beck S (1999) The
366 chicken B locus is a minimal essential major histocompatibility complex. Nature 401*:*923-925
367 Lenz TL, Becker S (2008) Simple approach to reduce PCR artefact formation leads to reliable
368 genotyping of MHC and other highly polymorphic loci - Implications for evolutionary
369 analysis. Gene 427*:*117-123
370 Liu BY, Deng ZQ, Huang W, Dong L, Zhang YY (2019) High prevalence and narrow host range of
371 haemosporidian parasites in Godlewski's bunting (Emberiza godlewskii) in northern China.
372 Parasitology International 69*:*121-125
373 Loiseau C, Zoorob R, Robert A, Chastel O, Julliard R, Sorci G (2011) Plasmodium relictum infection
374 and MHC diversity in the house sparrow (Passer domesticus). Proc Biol Sci 278*:*1264-72
375 Milinski M (2006) The major histocompatibility complex, sexual selection, and mate choice. Annual
376 Review of Ecology Evolution and Systematics 37*:*159-186
377 Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of
378 the vertebrate immune system. Proceedings of the National Academy of Sciences of the
379 United States of America 94*:*7799-7806

O'Connor EA, Strandh M, Hasselquist D, Nilsson JA, Westerdahl H (2016) The evolution of highly variable immunity genes across a passerine bird radiation. Molecular Ecology 25:977-989

Paterson S, Pemberton JM (1997) No evidence for major histocompatibility complex-dependent mating patterns in a free-living ruminant population. Proceedings of the Royal Society B-Biological Sciences 264:1813-1819

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

Pilosof S, Fortuna MA, Cosson JF, Galan M, Kittipong C, Ribas A, Segal E, Krasnov BR, Morand S, Bascompte J (2014) Host-parasite network structure is associated with community-level immunogenetic diversity. Nature Communications 5

Promerova 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 12:285-292

Radwan J, Babik W, Kaufman J, Lenz TL, Winternitz J (2020) Advances in the Evolutionary Understanding of MHC Polymorphism. Trends in Genetics 36:298-311

Reche PA, Reinherz EL (2003) Sequence variability analysis of human class I and class II MHC molecules: Functional and structural correlates of amino acid polymorphisms. Journal of Molecular Biology 331:623-641

Sandberg M, Eriksson L, Jonsson J, Sjostrom 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 41:2481-2491

Sebastian A, Herdegen M, Migalska M, Radwan J (2016) amplisas: a web server for multilocus genotyping using next-generation amplicon sequencing data. Molecular Ecology Resources 16:498-510

Sepil I, Lachish S, Hinks AE, Sheldon BC (2013) Mhc supertypes confer both qualitative and quantitative resistance to avian malaria infections in a wild bird population. Proceedings of the Royal Society B-Biological Sciences 280

Sepil I, Moghadam HK, Huchard E, Sheldon BC (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

Sin YW, Annavi G, Newman C, Buesching C, Burke T, Macdonald DW, Dugdale HL (2015) MHC class II-assortative mate choice in European badgers (Meles meles). Mol Ecol 24:3138-50

Slade RW, Mccallum HI (1992) Overdominant Vs Frequency-Dependent Selection at Mhc Loci. Genetics 132:861-862

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proceedings of the Royal Society B-Biological Sciences 277:979-988

Takahata N, Nei M (1990) Allelic Genealogy under Overdominant and Frequency-Dependent Selection and Polymorphism of Major Histocompatibility Complex Loci. Genetics 124*:*967-978

Westerdahl H (2007) Passerine MHC: genetic variation and disease resistance in the wild. Journal of Ornithology 148*:*S469-S477

Westerdahl H, Waldenstrom J, Hansson B, Hasselquist D, von Schantz T, Bensch S (2005) Associations between malaria and MHC genes in a migratory songbird. Proc Biol Sci 272*:*1511-8

Wilson AJ, Reale D, Clements MN, Morrissey MM, Postma E, Walling CA, Kruuk LEB, Nussey DH (2010) An ecologist's guide to the animal model. Journal of Animal Ecology 79*:*13-26

Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24*:*1586-1591

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

Table 1. MHC diversity of MHC Class I exon 3 in Godlewsiki's Bunting

	Range	Mean	SD	Minimum number of loci
Number of alleles per individual	1-21	10.61	2.59	11
Number of classical alleles per individual	1-18	8.42	2.33	9
Number of supertypes per individual	1-9	5.28	1.18	NA

Figure Legends

Figure 1. Bayesian phylogenetic tree depicting the major subset of MHC alleles in Godlewski's bunting. Red dots indicate the expressed alleles. The clade 1 in orange with high support without any expressed alleles was identified as non-classical clade.

Figure 2. DAPC scatterplot of the nine MHC class I supertypes. Seven Principal components (PCs) were retained during analyses to describe the relationship between the clusters. The scatterplot shows only the first two PCs of the DAPC of MHC supertypes. The bottom right graph illustrates the variation explained by the PCs. Alleles are represented as dots and supertypes as ellipses.

Figure 3. Positive selection on amino acid sites. A sequence logo showing the relative frequencies of amino acids of the fragment of the MHC class I exon 3. The plot is based on sequences of all putative expressed alleles. Positions under positive selection are indicated with asterisks.

Declarations

Ethics approval and consent to participate

Sampling Approval issued by College of Life Sciences, Beijing Normal University: No. CLS-EAW-2013-007

Data Accessibility Statement

All obtained sequences in this study will be submitted to GenBank after acceptance.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Science Foundation of China (No. 31772444 to LD) and the Monitoring Fund for the Haemosporidian Parasites by the National Forestry and Grassland Administration of China.

Authors' contributions

WH, YZ, and LD conceived the study, WH, BL, YP, and LD collected samples and carried out lab work, WH, TLL and LD analysed data with the help from BL and YP. WH and LD wrote the manuscript with inputs from all authors.

Acknowledgements

We thank Jiaxin Cao for technique support of RNA extraction and Beijing Baihuashan National Natural Reserve for assistance in field work.