The genomic architecture of the passerine MHC region: High repeat content and contrasting evolutionary histories of single copy and tandemly duplicated MHC genes

Helena Westerdahl1 | Samantha Mellinger1 | Hanna Sigeman1
Verena E. Kutschera3 | Estelle Proux-Wéra3 | Max Lundberg1
Matthias Weissensteiner2 | Allison Churcher4 | Ignas Bunikis5 | Bengt Hansson1
Jochen B. W. Wolf2 | Maria Strandh1

1Molecular Ecology and Evolution Laboratory, Department of Biology, Lund University, Lund, Sweden
2Division of Evolutionary Biology, Faculty of Biology, LMU Munich, Munich, Germany
3Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Solna, Sweden
4National Bioinformatics Infrastructure Sweden, Department of Molecular Biology, Umeå University, Umeå, Sweden
5Uppsala Genome Center, Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, BMC, Uppsala, Sweden

Correspondence
Helena Westerdahl, Molecular Ecology and Evolution Laboratory, Department of Biology, Lund University, Sölvegatan 37, SE-223 62 Lund, Sweden.
Email: helena.westerdahl@biol.lu.se

Funding information
Vetenskapsrådet, Grant/Award Number: 621-2013-4510 and 621-2016-689; H2020 European Research Council, Grant/Award Number: 336536 and 679799; Knut and Alice Wallenberg Foundation; National Bioinformatics Infrastructure Sweden; SciLifeLab; NGI/Uppsala Genome Center; UPPMAX; Science for Life Laboratory, Sweden

Handling Editor: David Coltman

Abstract
The major histocompatibility complex (MHC) is of central importance to the immune system, and an optimal MHC diversity is believed to maximize pathogen elimination. Birds show substantial variation in MHC diversity, ranging from few genes in most bird orders to very many genes in passerines. Our understanding of the evolutionary trajectories of the MHC in passerines is hampered by lack of data on genomic organization. Therefore, we assembled and annotated the MHC genomic region of the great reed warbler (Acrocephalus arundinaceus), using long-read sequencing and optical mapping. The MHC region is large (>5.5 Mb), characterized by structural changes compared to hitherto investigated bird orders and shows higher repeat content than the genome average. These features were supported by analyses in three additional passerines. MHC genes in passerines are found in two different chromosomal arrangements, either as single copy MHC genes located among non-MHC genes, or as tandemly duplicated tightly linked MHC genes. Some single copy MHC genes are old and putative orthologues among species. In contrast tandemly duplicated MHC genes are monophyletic within species and have evolved by simultaneous gene duplication of several MHC genes. Structural differences in the MHC genomic region among bird orders seem substantial compared to mammals and have possibly been fuelled by clade-specific immune system adaptations. Our study provides methodological guidance in characterizing complex genomic regions, constitutes a resource for MHC research in birds, and calls for a revision of the general belief that avian MHC has a conserved gene order and small size compared to mammals.

Keywords
gene order, major histocompatibility complex (MHC), Passeriformes, repeats, single copy genes, tandemly duplicated genes
INTRODUCTION

Birds and mammals have similar gene numbers, but birds have considerably smaller genomes than mammals (Szarski, 1976; Tiersch & Wachtel, 1991). This is largely explained by lower repeat content, shorter introns and large segmental deletions (Botero-Castro et al., 2017; Zhang et al., 2014). The small genome size is a feature birds share with bats, which has prompted the hypothesis that selection associated with metabolic characteristics may act to streamline genomes of flying species (Kapusta et al., 2017; Zhang et al., 2013). In addition to their smaller size, bird genomes are karyotypically rather stable compared to mammalian genomes, hence the gene order and chromosome structure are kept relatively constant over very long evolutionary time scales (Botero-Castro et al., 2017; Ellegren, 2010).

In contrast, birds of the order Passeriformes (passerines), the most species rich bird order, have highly duplicated immune genes, so called major histocompatibility complex (MHC) genes (reviewed in Minias et al., 2019; O’Connor et al., 2019). This expansion of MHC genes in passerines, as evidenced by high throughput amplicon sequencing (O’Connor et al., 2016, 2020), has apparently occurred despite selection for a streamlined genome. The MHC gene expansion seen in passerines resembles that of both mammals and fish (Shiina et al., 2017; Star et al., 2011), though the genomic organization of these duplicated MHC genes in passerines is yet to be characterized.

The classical MHC molecules are encoded by high polymorphic MHC genes (Murphy & Weaver, 2017). Strong selection from pathogens results in both increasing MHC polymorphism (number of different MHC alleles per locus, a population estimate Lenz et al., 2013; Prugnolle et al., 2005), and increasing MHC diversity (number of different MHC alleles/genes per individual O’Connor et al., 2018; Radwan et al., 2012; Wegner et al., 2003). The MHC molecules are crucial for initiating every T cell mediated adaptive immune response. There are two main classes of the so called classical MHC molecules; classical MHC class I (MHC-I) molecules present antigens from intracellular pathogens (for example from viruses, but also self-peptides) and classical MHC class II (MHC-II) molecules present antigens from extracellular pathogens (for example from bacteria) (Murphy & Weaver, 2017; Neefjes et al., 2011). An MHC-I molecule is encoded by a single MHC-I allele and is stabilized by a monomorphic gene called beta-2-microglobulin, whereas an MHC-II molecule is encoded by one MHC-IIa allele and one MHC-IIb allele.

The structure of nonclassical MHC molecules is similar to classical MHC molecules but nonclassical MHC molecules have other immune system related functions and are not subjected to diversifying selection, resulting in low levels of polymorphism (reviewed in Shiina et al., 2009).

The MHC genomic region has gained considerable attention because it holds a large number of genes associated with different functions in the immune system, and alleles/haplotypes which are directly associated with resistance or susceptibility to diseases (reviewed in Shiina et al., 2009). Some of the most convincing associations between specific MHC haplotypes and disease resistance are reported from the avian model system chicken, Gallus gallus, of the order Galliformes (reviewed in Shiina et al., 2007). The core MHC region (MHC-B) of chicken is located on chromosome 16, contains 46 genes and covers 242 kb, and a striking co-evolutionary pattern between the antigen peptide transporter genes (TAP1 and TAP2) and the MHC-I genes has been reported (Kaufman, 2015). Compared to the human MHC region, which covers 4 Mb, the chicken MHC region is compact and the genes are small, a so called minimal essential MHC (Kaufman et al., 1999; Shiina et al., 2007, 2009). This core MHC region has been characterized in bird species from four different bird orders, Galliformes, Anseriformes, Ciconiformes and Pelecaniformes, and 1–7 classical MHC-I genes and 1–2 classical MHC-IIB genes have been detected (Chaves et al., 2009; Chen et al., 2015; Hosomichi et al., 2006; Moon et al., 2005; Tsuji et al., 2017; Wang et al., 2012).

The gene order in the core MHC region is highly conserved among Galliformes species (chicken, turkey Meleagris gallopavo, quail Coturnix japonica, and black grouse Tetrao tetrix (Kaufman et al., 1999; Hosomichi et al., 2006; Chaves et al., 2009; Wang et al., 2012)). Three conserved gene orders in the core MHC region are; (i) the classical MHC-I genes flanking the TAP genes, (ii) the classical MHC-IIB genes flanking the chaperone tapasin (TAPBP), and then in between the classical MHC-I and MHC-IIB genes (iii) the nonclassical MHC-II genes (DMC-I and DMB) and the gene bromodomain containing protein 2 (BRD2) are found. However, in the mallard duck Anas platyrhynchos (Anseriformes), oriental stork Ciconia boyciana (Ciconiiformes) and crested ibis Nipponia nippon (Pelecaniformes), these genes are ordered slightly different (Chen et al., 2015; Fleming-Canepa et al., 2016; Moon et al., 2005; Tsuji et al., 2017): (i) classical MHC-I genes remain close to the TAP genes though on one side only, (ii) TAPBP seems to have been lost from the classical MHC-IIB genes, whereas (iii) the gene order of the nonclassical MHC-II genes and BRD2 remains.

Passeriformes species have up to 68 MHC-I alleles (i.e., at least 34 MHC-I genes) and 95 MHC-IIB alleles (i.e., at least 48 MHC-IIB genes) per individual, as indicated by amplicon sequencing of one target exon (Biedrzycka et al., 2017; O’Connor et al., 2020; Mellinger et al., Manuscript). These MHC-I and -IIB diversity estimates are much higher than those reported in birds from other orders (up to eight MHC-I genes and eight MHC-IIB genes, respectively Minias et al., 2019) and suggest a massive expansion of the MHC genomic region(s) in passerines. This pattern was recently supported using genome data (He et al., 2021).

Previous work has shown that phylogeny explains a significant part of the variance in MHC-I and IIB diversity among passerines (Minias et al., 2019; O’Connor et al., 2016), with especially high MHC-I diversity in warblers in the superfamily Sylvioidae (O’Connor et al., 2016, 2020). Balakrishnan et al. (2010) used several techniques (sequencing of BAC, Southern blot and FISH) to characterize the MHC genomic regions(s) in the zebra finch Taenopygia guttata. They described the gene order of MHC related genes, and found low MHC-I and high MHC-IIB diversities. Ekblom et al. (2011) continued to work with the MHC genomic regions(s) in the zebra finch and mapped MHC related genes to chromosome 16 using intragenic
SNPs. This knowledge of MHC diversity in passerines and the partly characterized MHC genomic region in the zebra finch is a good starting point to investigate the passerine MHC genomic region further. Yet, to date, we are still ignorant about the detailed organisation of the passerine MHC genomic region. Without advancing our knowledge of the gene order in the MHC region the evolutionary trajectories of the massive MHC gene expansions in passerines cannot be understood.

We therefore set out to characterize the MHC genomic region in an oscine passerine with highly duplicated MHC genes, the great reed warbler *Acrocephalus arundinaceus* (superfamily Sylvioidae, infraorder Passerida). We generated a high-quality genome assembly, using a combination of long-read, linked-read, and short-read sequencing, as well as optical mapping, of a single individual for which the genomic and expressed MHC-I and -IIB gene diversities had been thoroughly characterized previously (Westerdahl et al., 1999, 2000), as had the MHC diversity in its family (Westerdahl et al., 2004). With this prior knowledge on the MHC in the genome individual, we had expectations on both the number of MHC gene copies and their linkage, valuable information when setting out to assemble a complex genomic region with a large number of highly similar MHC paralogues and a putatively high repeat content due limited recombination (Roved et al., 2020). We hypothesise that the core MHC region in the great reed warbler will be considerably larger and contain several gene rearrangements compared to hitherto characterized birds. To verify the consistency of such putative gene rearrangements in the core MHC region among passerines, we also analysed, and partly annotated the core MHC region in three additional species where the core MHC region among passerines, we also analysed, and partly characterized compared to hitherto characterized birds.

To verify the consistency of such putative gene rearrangements in the core MHC region among passerines, we also analysed, and partly characterized the core MHC region in three additional species where high-quality long-read genomes were available; the hooded crow *Corvus cornix* (Weissensteiner et al., 2017), jackdaw, *Coloeus monedula* (Weissensteiner et al., 2020) and zebra finch, *Taeniopygia guttata* (Tgut_diploid_1.0, Genbank accession GCA_002008985.2). For downstream analyses we only included primary contigs from these four assemblies. The long-read sequenced genome of the chicken, *Gallus gallus* (GRCG6a, GenBank accession GCA_000002315.5) was used for comparison. The overall quality of each genome assembly was assessed with QUAST v. 4.5.4 (Gurevich et al., 2013). Additionally, the number of conserved single copy bird genes (aves_odb9, N = 4,915) present in the genomes was assessed with BUSCO v. 2.0.1 (Simão et al., 2015; Waterhouse et al., 2017).

### 2.2 MHC full-length predictions

To characterize and assess the total number of full-length MHC-I, MHC-IIA and MHC-IIB genes in the avian genomes a custom approach was applied, the MHC exon annotation (GitHub repository, https://github.com/ekol-hwe/MHC_GRW_genome_2022). Briefly, blastn (nucleotide BLAST) searches were run using BLAST v. 2.6.0+ (Camacho et al., 2009) with the exon sequences of genes of interest as queries. For the great reed warbler, we partitioned one MHC-I and one MHC-IIB cDNA Sanger sequence from the genome individual (GenBank accession numbers AJ005507 and AJ404371, respectively) into exons and used these separately (by gene and class) for querying the genome. These exon sequences were also used for initial predictions of MHC-I and MHC-IIB in the other three passerine genomes. After initial prediction, the coding sequence of one gene of each type was extracted from the respective genomes of the hooded crow, jackdaw and zebra finch to use in final MHC gene prediction within each species. With a similar approach, initial MHC-IIA full-length predictions were made using an MHC-IIA sequence from great tit, *Parus major* (GenBank accession XM_015616663), partitioned into exons as queries and then species-specific sequences were identified and used for final MHCIIA exon predictions. Exon sequences from the chicken TAP1 and TAP2 on chromosome 16 (GRCG6a, GenBank accession CM0000108.5) were used for initial prediction of these genes in passerines. Subsequently, the genomic intervals of blastn hits of individual MHC and TAP exons were used to identify full-length genes in each species. If the majority of the exon intervals of one gene category appeared in expected order, it was considered a putative full-length gene (GitHub repository, https://github.com/ekol-hwe/MHC_GRW_genome_2022; exons 2, 3 and 4 within a gene customized distance in the genome, 10,000 bp for MHC-I, 4,000 bp for MHC-IIB and no set distance for MHCIIA, TAP1 and TAP2). Exon-intron boundaries were manually curated in Geneious v. 6.1.8 (https://www.geneious.com) with guidance from local mapping of the MHC exon sequences used for blastn predictions to each predicted full-length gene.
2.3 | Characterization of the MHC genomic region in passerines

Blast searches with all annotated MHC region genes in chicken and crested ibis were used to identify more MHC region scaffolds/contigs and/or further characterize MHC and TAP scaffolds/contigs. Tblastx searches using BLAST v. 2.6.0+ with an E-value cutoff of 1E-10 were performed. All CDS with functional annotation in the chicken extended MHC-B region (500 kb surrounding the 92 kb B locus (Kaufman et al., 1999), chromosome 16, GRCg6a GenBank accession CM000108.5: 2,344,602–2,844,601 bp) and in the crested ibis MHC region (GenBank accession numbers KP182407-KP182409) were used as queries to local nucleotide blast databases of each passerine genome. The top hit for each chicken and crested ibis MHC region CDS was considered for identification of top MHC-region candidates among the passerine scaffolds/contigs. In addition to MHC and TAP scaffolds/contigs, we included scaffolds/contigs containing tripartite motif (TRIM) genes, which are found in the extended MHC region in chicken (Shiina et al., 2007). Homologous MHC region candidate scaffolds/contigs among species were identified using SatsumaSynteny2 in the software Satsuma v. 2 (Grabherr et al., 2010). The MHC, TRIM and TAP scaffolds/contigs in the great reed warbler genome were used as queries to the hooded crow, jackdaw, zebra finch and chicken genomes. For each great reed warbler MHC, TRIM and TAP scaffold, the hit contig with the highest number of syntenic regions in each of these four species were considered candidates for homologous contigs. The homologous MHC, TRIM and TAP scaffolds/contigs in the four passerines were annotated ab initio (using chicken parameters) with Augustus v. 3.2.3 (Stanke et al., 2006). The predicted CDS for each gene model was translated and searched against swissprot vertebrate proteins (N = 86,131, downloaded 20211118) using blastp in BLAST v. 2.6.0+ (Camacho et al., 2009) with an E-value cutoff of 1E-05. Augustus gene models with a significant hit to a swissprot protein were renamed and kept for downstream analyses.

All great reed warbler scaffolds containing MHC region genes were identified in the genome-wide annotation of the great reed warbler genome assembly (Sigeman et al., 2021). The MHC region gene predictions were manually curated one more time for the great reed warbler, taking the gene predictions from both the genome-wide annotation and from the MHC exon annotation into account. In this process, the genome-wide annotation was updated with the manually curated MHC region genes (Dryad, https://doi.org/10.5061/dryad.fqz612jv6). MHC contig annotations of the other three passerines are also available in Dryad.

The size of the MHC region in the great reed warbler was estimated as the total length of all scaffolds containing predicted full-length MHC genes (MHC scaffolds) and MHC-related genes (TAP and TRIM scaffolds), as identified above. We decided to exclude one scaffold and one gap from this size estimate; the large scaffold Aaru_104 (2 Mb) with a single MHC-IIB gene only and no additional genes expected to be found in the core MHC region, and also the 200 kb gap inferred by BioNano optical mapping in scaffold Aaru_120.

2.4 | Repeat content in the MHC regions versus the rest of the genome

The total abundance of repeats for each scaffold/contig within species was assessed as follows for the four passerines: RepeatMasker v. 4.0.7 (Smit et al., 2013-2015) was run separately for MHC-I scaffolds/contigs, MHC-IIB scaffolds/contigs, MHC-I+MHC-IIB scaffolds/contigs, and the rest of the genome, using default settings and with published repeat libraries (fAlb15_rm3.0_aves_hc.lib provided by A. Suh) containing Repbase repeats (Bao et al., 2015), repeats from chicken and the zebra finch (Hillier et al., 2004; Warren et al., 2010) and curated repeats from hooded crow and flycatcher (Suh et al., 2018; Vijay et al., 2016). In contrast to Sigeman et al. (2021) we did not include uncurated de novo predicted repeats in the repeat library as we observed a considerable overlap with multicopy genes. This approach underestimates the repeat content but has a higher specificity in its predictions. In the great reed warbler, the combined repeat sizes for both approaches are highly correlated both within the MHC region and in the rest of the genome (r = .98).

2.5 | Repeat content within MHC genes (intragenic) and between tandemly duplicated genes (intergenic)

The intragenic and intergenic repeat content from different types of repeats identified with RepeatMasker were analysed for all the MHC scaffolds/contigs in the four passerines. MHC-I and -IIB genes within 30 kb distance to a neighbouring MHC-I and -IIB gene, respectively, were considered tandemly duplicated MHC genes. Single copy MHC-I and -IIB genes were either found in a context of non-MHC genes or as a single copy gene close to the end of a scaffold/contig. The intergenic repeat content was only analysed between tandemly duplicated MHC-I and -IIB genes whereas the intragenic repeat content was analysed in all MHC-I and -IIB genes in ORF.

2.6 | High throughput amplicon sequencing

MHC-I exon 3 and MHC-IIB exon 2 were amplified using two and four different primer combinations, respectively, in the great reed warbler genome individual, her siblings and parents (Mellinger et al. Manuscript), using standard procedures (Drews & Westerdahl, 2019; Roved et al., 2018). The MHC-I exon 3 and MHC-IIB exon 2 amplicons were sequenced using Illumina MiSeq and the maternal and paternal MHC haplotypes in the genome individual were inferred as described in Mellinger et al. (Manuscript).
2.7 Maximum likelihood trees and statistical analysis

The CDSs of all full-length MHC-I and MHC-IIB genes were extracted from the passerine genomes. After alignment with ClustalW2 in Geneious v. 6.1.8, the nucleotide sequences were manually edited by deleting nucleotides or adding Ns at putative sequencing errors (e.g., indel errors at homopolymer stretches) when needed for in-frame translation. The edited nucleotide sequences were then used for construction of maximum likelihood trees with 500 bootstrap replicates and best-fitting substitution models according to model tests; MHC-I, T92+G (ML, Tamura 3-parameter, 500 bootstrap replicates) and MHC-IIB, GTR+G (ML, general time reversible, 500 bootstrap replicates), in MEGA version X (Kumar et al., 2018; Stecher et al., 2020). The MHC-IIB tree was constructed only from the great reed warbler and zebra finch sequences of the passerines since full length genes were not found in the corvid genomes. Chicken reed warbler and zebra finch sequences of the passerines since full length MHC-IIB genes were not found in the corvid genomes. Chicken BLA (GenBank accession AY357253) and the four MHCIIBA of the crested ibis (cds of DAA, DBA1, DBA2 and DBA3 from Ninii_MHCCore, GenBank accession KP182408) were used as outgroup.

We analysed differences in gene sizes, intragenic repeat content (%), intergenic distances and intergenic repeat content (%) between passerine species (species with less than five data points were excluded from analyses), separately for MHC-I and IIB genes. The statistical analyses were done using nonparametric tests (Mann-Whitney U test and Kruskal Wallis test) in IBM-SPSS version 27.0.

3 RESULTS

3.1 An extended MHC region with high MHC diversity in the great reed warbler

We used PacBio long-read, 10x genomics linked-read and BioNano optical mapping to create a de novo genome assembly of a female great reed warbler (Sigeman et al., 2021). The final assembly (acrAru1) consisted of 3,013 scaffolds with a cumulative length of 1.2 Gb and an N50 of 21.5 Mb (Table S1). Moreover, it had 93% complete single copy bird orthologues as assessed with BUSCO v.3.0.2 (aves_odb9 data set, N = 4,915) (Simão et al., 2015), and 22,524 genes were annotated (Sigeman et al., 2021). Eighteen scaffolds with MHC-genes (MHC-I and/or -IIB genes) and MHC-related genes, that is, genes expected to be found in the MHC-region such as TAP1 and TAP2, were identified in the genome: three MHC-I scaffolds (292, 314 and 476 kb), 11 MHC-IIB scaffolds (50–2,058 kb), two scaffolds with both MHC-I and -IIB genes (755 and 1,288 kb) and two MHC-related scaffolds (122 and 426 kb), covering a total genomic read length of approximately 5.5 Mb (Table S2).

The great reed warbler genome has high MHC diversity, 15 full-length MHC-I and 56 full-length MHC-IIB genes in open reading frame (ORF) were found (Table S2). We also genotyped the MHC-I and MHC-IIB diversity in the genome individual and its parents, using Illumina MiSeq amplicon sequencing, to get an independent estimate of the MHC diversity and to infer maternally and paternally inherited MHC alleles in the genome individual (Figure S1). The amplicon primer pairs were placed in nonvariable gene regions but designed to amplify the most polymorphic gene region across all MHC-I and MHC-IIB paralogues (MHC-I exon 3 and MHC-IIB exon 2 amplicons). The MHC diversity from the MiSeq amplicon data agrees well with the genome data: a total of 22 MHC-I amplicon alleles and 95 MHC-IIB amplicon alleles in ORF were found in the genome individual, which corresponds to 11 MHC-I and 48 MHC-IIB genes under full heterozygosity.

3.2 Gene duplications of trios of MHC genes are likely to explain the high MHC diversity in the great reed warbler

The MHC-I and -IIB genes in the great reed warbler genome are found in two main genomic arrangements: single copy MHC genes amongst non-MHC genes and tandemly duplicated MHC genes with short intergenic distances (Figure 1a). Aaru_Scaffold 18 (1.3 Mb), the largest of the five scaffolds holding MHC-I genes, has several single copy MHC-I genes in ORF. Aaru_Scaffold 508 (292 kb) and Aaru_Scaffold 61 (476 kb) hold tandemly duplicated MHC-I genes, 11 and two, respectively (ten and two MHC-I genes in ORF, Table S2). Phylogenetic reconstructions of both a subset and all MHC-I genes show a monophyletic origin of the tandemly duplicated genes on Aaru_Scaffold 508 and place one single copy MHC-I gene, Aaru-UA*18_03, distant from all other great reed warbler MHC-I genes (Figure 1b, Figure S2a).

Tandemly duplicated MHC-I genes are of particular interest since they explain the high MHC-I diversity in the great reed warbler. We therefore inferred the gene duplication history of the 11 tandemly duplicated MHC-I genes in Aaru_Scaffold 508, based on coding sequence (CDS) similarity, gene size and intergenic distance (Figure 1b, Figure S3a). The gene order, CDS similarity, gene size and intergenic distance suggest a single duplication event of three MHC-I genes, either the trio Aaru-UA*508_4, _5, _6 or the trio Aaru-UA*508_8, _9, _10 (Figure 1, CDS gene similarity, Aaru-UA*508_04&10: 98%, _05&09: 99%, _06&08: 97%, in relation to CDS similarities across all MHC genes in ORF on Aaru_Scaffold 508 [Min, Max, 92%, 99%, Figure S3a], gene size, Aaru-UA*508_04 & 10: both 4.6 kb, _05 & 09: both 4.3 kb, _06 & 08: 10.9 and 10.7 kb, respectively, in relation to gene sizes across all MHC genes in ORF on Aaru_Scaffold 508 [Min, Max, 4.3 kb, 11.2 kb] and intergenic distance, Aaru-UA*508_04 & 05 & 06: 15.0 kb and 32.2 kb, Aaru-UA*508_10 & 09 & 08: 15.2 kb and 22.7 kb in relation to intergenic distances across all MHC-I genes on Aaru_Scaffold 508 [Min, Max, 9.7 kb, 32.2 kb]). The substantial differences in MHC-I gene sizes among the tandemly duplicated MHC-I genes on Aaru_Scaffold 508 is explained by a high repeat content in introns one and two (Figure 1b). These repeats not only strengthen the pattern of a single duplication event of three MHC-I genes but may also have been important in their gene duplication history. Repeats are highly abundant in the MHC region (24.7%), compared...
FIGURE 1  Trios of tandemly organized MHC-I and IIB genes have been duplicated in the great reed warbler. (a) Three MHC scaffolds in the great reed warbler drawn to scale (Aaru_Scaffold 18 (1,288 kb), Aaru_Scaffold 508 (292 kb) and Aaru_Scaffold 120 (510 kb)) with single copy and tandemly duplicated MHC genes. The single copy MHC-I gene Aaru-UA*18_03, that is, neighbouring genes are non-MHC genes (grey boxes), is indicated with a light blue box and an orange arrow and the single copy MHC-IIB gene Acar-DAB*18_04 is indicated with a light green box and an orange arrow. The tandemly duplicated MHC-I genes Aaru-UA*508_08, _09, _10 with reverse orientation (indicated with dark blue arrows) suggesting a single gene duplication event of three MHC-I genes. Tandemly duplicated MHC-IIB genes Aaru-DAB*120_11, _12, _13 have forward, reverse and reverse orientation, pairwise highly similar CDS, gene size and intergenic distance to the three tandemly duplicated MHC-I genes. The intergenic distance is the distance to the nearest tandemly duplicated MHC-IIB gene in forward orientation. *Note, the two intergenic distances between the three tandemly duplicated MHC-IIB genes that should be compared are, the distance between Aaru-DAB*120_19 and 18 against Acar-DAB*120_11 and 12 and the distance between Aaru-DAB*120_18 and 17 against Acar-DAB*120_12 and 13 since these genes have different orientations.

| MHC-I gene | Size (bp) | Repeat (bp) | Intergenic distance (kb) |
|------------|-----------|-------------|-------------------------|
| Aaru-UA*18_3 | 4338     | 28*        | 5789                    |
| Aaru-UA*508_4 | 4593    | 0          | 5736                    |
| Aaru-UA*508_5 | 4636    | 0          | 133                     |
| Aaru-UA*508_9 | 4323    | 113        | 112                     |
| Aaru-UA*508_2* | 2853   | 0          | 4593                    |
| Aaru-UA*508_11 | 6506   | 113        | 4631                    |
| Aaru-UA*508_7 | 4837    | 143        | 4837                    |
| Aaru-UA*508_1 | 11228   | 7975       | 11228                   |
| Aaru-UA*508_8 | 10706   | 7772       | 10706                   |
| Aaru-UA*508_6 | 10901   | 9789       | 10901                   |
| Aaru-DAB*18_4 | 2030    | NA         | 2384                    |
| Aaru-DAB*120_5 | 2452   | 30         | 2452                    |
| Aaru-DAB*120_13 | 2402  | 1.9        | 2402                    |
| Aaru-DAB*120_17 | 2597   | 1.3        | 2597                    |
| Aaru-DAB*120_16 | 2457   | 2.7        | 2457                    |
| Aaru-DAB*120_15* | 2554  | 5.6        | 2554                    |
| Aaru-DAB*120_14 | 2440   | 4.0        | 2440                    |
| Aaru-DAB*120_14 | 2440   | 4.0        | 2440                    |
| Aaru-DAB*120_13 | 2575   | 5.5        | 2575                    |
| Aaru-DAB*120_12 | 2347   | >30        | 2347                    |
| Aaru-DAB*120_11 | 2353   | 3.6        | 2353                    |
| Aaru-DAB*120_10 | 2434   | 19.3       | 2434                    |
| Aaru-DAB*120_9 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_16 | 2503   | 4.6        | 2503                    |
| Aaru-DAB*120_15 | 2554   | 5.6        | 2554                    |
| Aaru-DAB*120_14 | 2602   | >30        | 2602                    |
| Aaru-DAB*120_13 | 2347   | >30        | 2347                    |
| Aaru-DAB*120_12 | 2353   | 3.6        | 2353                    |
| Aaru-DAB*120_11 | 2434   | 19.3       | 2434                    |
| Aaru-DAB*120_10 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_9 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_16 | 2503   | 4.6        | 2503                    |
| Aaru-DAB*120_15 | 2554   | 5.6        | 2554                    |
| Aaru-DAB*120_14 | 2602   | >30        | 2602                    |
| Aaru-DAB*120_13 | 2347   | >30        | 2347                    |
| Aaru-DAB*120_12 | 2353   | 3.6        | 2353                    |
| Aaru-DAB*120_11 | 2434   | 19.3       | 2434                    |
| Aaru-DAB*120_10 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_9 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_16 | 2503   | 4.6        | 2503                    |
| Aaru-DAB*120_15 | 2554   | 5.6        | 2554                    |
| Aaru-DAB*120_14 | 2602   | >30        | 2602                    |
| Aaru-DAB*120_13 | 2347   | >30        | 2347                    |
| Aaru-DAB*120_12 | 2353   | 3.6        | 2353                    |
| Aaru-DAB*120_11 | 2434   | 19.3       | 2434                    |
| Aaru-DAB*120_10 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_9 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_16 | 2503   | 4.6        | 2503                    |
| Aaru-DAB*120_15 | 2554   | 5.6        | 2554                    |
| Aaru-DAB*120_14 | 2602   | >30        | 2602                    |
| Aaru-DAB*120_13 | 2347   | >30        | 2347                    |
| Aaru-DAB*120_12 | 2353   | 3.6        | 2353                    |
| Aaru-DAB*120_11 | 2434   | 19.3       | 2434                    |
| Aaru-DAB*120_10 | 2430    | 6.5        | 2430                    |
| Aaru-DAB*120_9 | 2430    | 6.5        | 2430                    |

Eight great reed warbler scaffolds have tandemly duplicated MHC-IIB genes (Aaru_Scaffold 120, 168, 178, 18, 301, 357, 45 and 554, Table S2) and the highest number of tandemly duplicated genes is found in Aaru_Scaffold 120 with 19 MHC-IIB genes within 510 kb (14 MHC-IIB genes in ORF, Figure 1a). Phylogenetic reconstructions of both a subset and all MHC-IIB genes place the single copy MHC-IIB gene Aaru-DAB*18_04 on Aaru_Scaffold 18 distant from all other MHC-IIB genes (Figure 1c, Figure S2b). In contrast to the tandemly duplicated MHC-I genes in Aaru_Scaffold 508, the MHC-IIB genes on Aaru_Scaffold 120 are frequently mixed with MHC-IIB genes from other scaffolds (Figure S2b).
order, the CDS similarity, the gene size and the intergenic distance suggest a single duplication event of three MHC-IIB genes, either the trio Aaru-DAB*120_11, _12, _13 or the trio Aaru-DAB*120_17, _18, _19 (Figure 1c, Figure S3b, CDS gene similarity, Aaru-DAB*120_11 & 19: 99%, _12 & 18: 99%, _13 & 17: 98% compared to CDS similarities across all MHC-IIB genes in Aaru_Scaffold 120 [Min, Max, 86%, 99%] (Figure S3b), gene size, Aaru-DAB*120_11 & 19: both 2.6 kb, Aaru-DAB*120_12 & 18: both 2.5 kb, Aaru-DAB*120_13 & 17: both 2.3 kb compared to gene sizes across all MHC-IIB genes in Aaru_Scaffold 120 [Min, Max, 2.3 kb, 3.1 kb], intergenic distance, Aaru-DAB*120_11 & 12 & 13: 1.3 and 5.3 kb, Aaru-DAB*120_19 & 18 & 17: 1.3 and 5.2 kb compared to intergenic distances across all MHC-IIB genes in Aaru_Scaffold 120 [Min, Max, 1.3 kb, 6.5 kb]).

High CDS similarity between MHC genes is frequent among the tandemly duplicated MHC genes and indicates either recent duplication events of single gene copies and/or gene conversion (Figure S3). However, larger duplication events, such as homologous unequal crossing-over of trios of MHC-I or MHC-IIB genes, are particularly important for explaining the evolution of the high numbers of tandemly duplicated MHC genes in the great reed warbler (Figure 1).

### 3.3 Single copy MHC-I and IIB genes in the core MHC region are putative orthologues among passerines

The great reed warbler scaffold Aaru_Scaffold 18 shares large-scale homology with chicken chromosome 16, that is, the chromosome holding the core MHC region (Table S4), but the gene content and gene order partly differ. To investigate whether this genomic reorganization is common among passerines, we characterized MHC-genes and MHC-related genes in three additional passerine species with long-read sequenced high quality assemblies: the hooded crow, jackdaw and zebra finch (Tables S1 and S2). Aaru_Scaffold 18 shares large-scale homology with large contigs in the zebra finch (Tgut_Contig 1045, 634 kb), jackdaw (Cmon_Contig 195, 481 kb) and hooded crow (Ccor_Contig 220, 438 kb, Table S4). These four passerine scaffold/contigs with shared large-scale homology to the core MHC region contain one MHC-I gene in ORF in the zebra finch (Tgut-UA*1045_01), jackdaw (Cmon-UA*195_01) and hooded crow (Ccor-UA 220_01) and three in the great reed warbler (Aaru-UA*18_01-03; gene 18_01 and 18_03 in ORF, Table S2). They also contain one MHC-IIB gene in ORF in the zebra finch (Tgut-DAB*1045_01) and jackdaw (Cmon-DAB*195_01) and four MHC-IIB genes in the great reed warbler (Aaru-DAB*18_1-4; genes 18_01, 18_03 and 18_04 in ORF), but no MHC-IIB gene in the hooded crow.

The MHC-I gene Aaru-UA*18_03 in the great reed warbler and Tgut-UA*1045_01 in the zebra finch are found in the same gene order, next to the genes Flotillin-1 (FLOT1), Tubulin beta (TUBB) and mediator of DNA-damage checkpoint 1 (MDC1). The genes FLOT1, TUBB and MDC1 were not successfully annotated in the jackdaw and hooded crow but was recently annotated in the New Caledonian crow Corvus moneduloides. The MHC-I genes Acar-UA*18_03 and Tagu-UA*1045_01 in the great reed warbler and zebra finch, respectively, are putatively orthologous, based on both gene order and phylogenetic reconstruction of full-length MHC-I CDS in ORF (Figure 2a and Figure S2a). Likewise, are the MHC-I genes Cmon-UA*195_01 and Ccor-UA*220_01 in the jackdaw and hooded crow, respectively, putative orthologues, based on both large-scale homology of contigs and phylogenetic reconstruction of full-length MHC-I CDS in ORF (Figure 2a and Table S4). Based on gene order, large-scale homology of scaffold/contigs and phylogenetic reconstruction of MHC-IIB CDS in ORF one MHC-IIB gene in the core MHC region is a putative orthologue in the great reed warbler, zebra finch and jackdaw (Aaru-DAB*18_04, Tgut-DAB*1045_01 and Cmon-DAB*195_01, note Cmon-DAB*261_01 is the putative second allele at this class IIB locus, Figure 2b and Table S4).

All four passerines have tandemly duplicated MHC-IIB genes, whereas only the great reed warbler has tandemly duplicated MHC-I genes (except for jackdaw Cmon_Contig 327 with two MHC-I genes 30 kb apart, Table S2). As previously reported, the zebra finch has one MHC-I gene (or possibly two Balakrishnan et al., 2010; Ekbloom et al., 2011). Both the single copy MHC-IIB genes and the tandemly duplicated MHC-IIB genes are monophyletic within species in the four passerines, that is, unique to each species without any signs of trans-species polymorphisms based on full CDS, except for the putatively orthologous MHC-IIB gene in the core MHC region mentioned above (Aaru-DAB*18_04, Tgut-DAB*1045_01 and Cmon-DAB*195_01 [Figure 2b]). Two different MHC-IIB gene families have previously been reported among birds, DAB1 and DAB2, and in line with previous findings (Burri et al., 2010), all the MHC-IIB genes reported here, both single copy and tandemly duplicated, in the four passerines belong to the gene family DAB1 (Figure S4).

To conclude, the high MHC diversity in passerines is found in two different gene arrangements in their genomes, either as single copy MHC genes among non-MHC genes or as tandemly duplicated tightly linked MHC genes. The latter are of young evolutionary age and appear species specific in our data set whereas some of the single copy MHC genes are found in a conserved gene order and are older than the four studied passerine species.

### 3.4 Long terminal repeats (LTRs) are frequent across the whole MHC region

The total repeat content in the MHC-scaffolds/contigs is considerably higher than the genome averages in all the passerines (Figure 3a). Long terminal repeats (LTRs) are clearly overrepresented in the MHC-scaffolds/contigs compared with other repeats and this overrepresentation of LTRs is not seen in the remaining part of the genomes (Figure 3a, Figure S5). The rather large difference between the four species in total repeat content in the MHC-scaffolds/contigs can be explained by actual differences in repeat content between species but can also be due to differences in representation of MHC scaffolds/contigs among species (Table S2).
Repeats within the MHC-I and IIB genes (intragenic repeats) are frequent

The highest MHC diversities among the four passerines are found in the great reed warbler and the jackdaw (15 and eight MHC-I genes, respectively; Figure 2c). Twenty of these 23 MHC-I genes have intragenic repeats, that is, repeats within introns, and five of these 20 MHC-I genes also have LTRs. The MHC-I genes with repeats other than LTRs are small in size ($N = 15$, Mean = 4,445 bp, SD = 843 bp) compared to genes with LTRs ($N = 5$, Mean = 10,044 bp, SD = 1,526 bp), and the five largest MHC-I genes (7,534–11,022 bp) all have intragenic LTRs (Figure 4a).

The highest MHC diversities in the passerines are found among the MHC-IIB genes (Figure 2d). Repeats are frequent within the MHC-IIB genes in the zebra finch, jackdaw and hooded crow and are found in 100, 97 and 67% of the genes, respectively, although only in 9% of MHC-IIB genes in the great reed warbler. The intragenic MHC-IIB repeat content (proportion of repeats per gene) differs significantly between species (Figure 3b, Indep samples, Kruskal-Wallis, $N = 127$, KW(3) = 95.80, $p < .001$). The MHC-IIB gene size appears to be positively correlated with the intragenic repeat content in zebra finch (linear regression, $R = .63$, $p < .001$, Figure 4b), but not in the great reed warbler and jackdaw.
3.6 | Repeats between tandemly duplicated MHC-IIB genes (intergenic repeats) are frequent

All scaffolds/contigs with tandemly duplicated MHC-IIB genes have repeats in the intergenic regions, that is, between the MHC-IIB genes, and the repeat content differs significantly between species (Figure 3c; Indep samples, Kruskal-Wallis, $N = 142$, $KW_{(2)} = 54.61$, $p < .001$). This difference is mainly explained by the intergenic LTR content (Indep samples, Kruskal-Wallis, $N = 142$, $KW_{(2)} = 52.55$, $p < .001$) and less so by other repeats (Indep samples, Kruskal-Wallis, $N = 142$, $KW_{(2)} = 5.58$, $p = .061$).

3.7 | Tandemly duplicated MHC-IIB genes share both scaffold/contig homology and gene order among passerine species

Two great reed warbler scaffolds with tandemly duplicated MHC-IIB genes (Aaru_Scaffold 554 and 357) share large-scale homology with the zebra finch (Tgut_Contig 910 and 310) and jackdaw (Cmon_Contig 277F and 380F) contigs (Table S4). The tandemly duplicated MHC-IIB genes on these scaffolds/contigs are flanked by BRD2 and nonclassical MHC-II genes (DMA and DMB) on one side in the great reed warbler and zebra finch (Figure 5a), a gene order shared Galliformes, Ciconiiformes and Pelecaniformes species (Chen et al., 2015; Shiina et al., 2007; Tsuji et al., 2017). However, these tandemly duplicated MHC-IIB genes in the great reed warbler and zebra finch are uniquely flanked by a single MHC-IIA gene on the opposite side (Figure 5a). Finding a single MHC-IIA gene in each of the four passerines was unexpected since tandem MHC-IIA-MHC-IIB dyads have been reported in both Ciconiiformes and Pelicaniformes (Figure 5a) (Chen et al., 2015; Tsuji et al., 2017). Moreover, tandem MHC-IIA-MHC-IIB dyads are the common organization in mammals (Shiina et al., 2009, 2017). MHC-IIA is annotated as a full-length gene in the great reed warbler (Acar-DAA 554) and zebra finch (Tagu-DAA 910 and Tagu-DAA 375 are likely to be alleles of the same gene, Figure S6), but only found as short blast hits in the jackdaw and hooded crow. Finally, the chaperone gene TAPBP (tapasin) has been lost from this MHC-IIB region in passerines compared to birds of Galliformes species; note a similar loss in Ciconiiformes and Pelecaniformes species (Figure 5a) (Chen et al., 2015; Tsuji et al., 2017).

3.8 | Several major rearrangements in the core MHC region in passerines compared to chicken

The great reed warbler scaffold with TAP1 and TAP2 genes (Aaru_Scaffold 200) share large-scale homology with zebra finch, jackdaw and hooded crow contigs (Figure 5b, Table S4). Galliformes species have classical MHC-I gene(s) on either side of the TAP genes, but in the four passerines these MHC-I genes flanking the TAP1 and TAP2 genes have been lost. A similar loss of functional classical MHC-I gene(s) next to the TAP genes is seen in mallard duck (Anseriformes) and oriental stork (Ciconiiformes), and in these species MHC-I genes are only found on one side of the TAP genes (Figure 5b) (Chen et al., 2015; Fleming-Canepa et al., 2016; Moon et al., 2005; Tsuji et al., 2017). There are remnants of one MHC-I gene next to the TAP genes in the four passerines, although this gene seems pseudogenized (Figure S7).
In summary, there have been several major rearrangements in the core MHC region in passerines, compared to other bird orders, and these rearrangements seem to be conserved among the four passerine species studied. In passerines large-scale homology analyses reveal a core MHC region (Aaru_Scaffold 18 [1,288 kb], Tgut_Contig 1045 [634 kb], Cmon_Contig 195 [481 kb] and Ccor_Contig 220 [438 kb]) which hold the genes FLOT, TUBB and MDC1, a gene order shared between passerines and chicken. However, the genomic region holding the genes TAP1, TAP2 and tenascin X (TNXB) has left the core MHC region in passerines (Figure 5b, passerine scaffold/contig; Aaru_Scaffold 200 [425 kb], Tgut_Contig 57 [641 kb], Cmon_Contig 214F [320 kb] and Ccor_Contig 247F [261 kb]). Likewise, the genomic region holding the genes MHC-IIA, DMA, DMB and BRD2 has left the core MHC region in passerines (Figure 5a, passerine contigs; Aaru_Scaffold 357 [178 kb] and 554 [175 kb], Tgut_Contig 910 [733 kb] and Cmon_Contig 380F [55 kb]). Also, the tandemly duplicated MHC-I and IIB genes are found outside the core MHC region in passerines (Figure 1a). Finally, of the three important gene orders in the avian MHC core region outlined in the Introduction, only one remains in passerines, the gene order of the nonclassical MHC-II genes (DMA and DMB) and BRD2, whereas the only remaining MHC-I gene close to the TAP genes has been pseudogenized and TAPBP has been lost from the classical MHC-IIB genes. Unfortunately, we were not able to scaffold the three large MHC regions mentioned here. However, when the scaffold/contig information from the zebra finch and great reed warbler in the current study is combined with earlier linkage analyses in the zebra finch (Ekblom et al., 2011), the following structural organization is supported: the core MHC region (Aaru_Scaffold 18/Tgut_Contig 1045) is placed next to MHC-IIA, DM (A and B) and BRD2 (Aaru_Scaffold 357/Tgut_Contig 910), hence a continuous approximately 2 Mb genomic region.
**DISCUSSION**

There have been several major rearrangements in the core MHC region in the four studied passerines, the great reed warbler, zebra finch, jackdaw, and hooded crow, compared to the thoroughly characterized core MHC region in Galliformes species (Kaufman et al., 1999; Shiina et al., 2007). Genes expected to be found in the core MHC region seem to have been lost from this region in passerines, for example, classical and nonclassical MHC-II genes, TAP1 and TAP2 genes, and are found in separate scaffolds/contigs. Moreover, the gene TAPBP and the classical MHC-I gene(s) on one side of the TAP genes have been completely lost from all the MHC-scaffolds/contigs in the four studied passerines. These latter gene losses are also seen among Anseriformes, Ciconiiformes, and Pelecaniformes.

**FIGURE 5** The structural differences in the MHC genomic region among bird orders is substantial and a multitude of MHC class I and II gene organizations are seen in birds. (a) The gene order and organization of the MHC-IIA (yellow) and MHC-IIB (green) genes differ among bird orders (within Galliformes there is some gene copy number variation for MHC-IIB, Chen et al., 2015; Kaufman et al., 1999; Shiina et al., 2007; Tsuji et al., 2017), and only the passerines, with data from the great reed warbler and zebra finch, have many tandemly duplicated MHC-IIB genes, also note their single MHC-IIA gene copy. The gene order and organization of BRD2, DMA, DMB1, and DMB2 (grey colour) in this region are shared among the bird orders, *DB is better represented in the genome than the DMA in the zebra finch whereas *DNA is better represented than DMB in the great reed warbler, and in jackdaw only *DNA is indicated. (b) The gene order and organization of the MHC-I (blue) genes differ slightly among the bird orders (Chen et al., 2015; Kaufman et al., 1999; Moon et al., 2005; Shiina et al., 2007; Tsuji et al., 2017), and the passerines, with data from the great reed warbler, zebra finch, jackdaw, and hooded crow, seem to have lost their functional MHC-I genes from this region. The gene order and organization of DMB1, DMB2, TAP1, and TAP2 (grey colour) in this region are shared among birds of the orders Galliformes, Ciconiiformes, and Pelecaniformes; note that the DMB genes align with Figure 5a, whereas the gene TNXB (pink) has rearranged in Passeriformes, with data from the great reed warbler, zebra finch, jackdaw, and hooded crow. *Several of the MHC-I genes are not expressed in Anseriformes. Pseudogenes are indicated with ψ. Simplified comparative genomic maps, not drawn to scale.
species, suggesting they occurred rather early in the bird phylogeny (Chen et al., 2015; Fleming-Canepa et al., 2016; Moon et al., 2005; Tsuji et al., 2017). The second MHC-I gene close to the TAP genes, which is found in all birds studied so far, has been pseudogenized in the four passerines studied. Loss of the MHC-I genes close to the TAP genes agrees with findings by Balakrishnan et al. (2010) in zebra finch. A shared feature of the core MHC region in the studied passerines is therefore MHC gene loss and MHC gene pseudogenization.

However, there is a large expansion in genetic diversity of the MHC-IIIB genes in the four passerines compared to previously characterized Galliformes, Anseriformes, Ciconiiformes and Pelicaniformes species. Interestingly, the expansion in genetic diversity of the MHC-I genes is less consistent; the great reed warbler has high MHC-I diversity while the zebra finch only has a single MHC-I gene. All the passerine MHC-IIIB genes are monophyletic within species in our data set which represent three rather distant bird families, except for one single copy MHC-IIIB gene found in a shared gene order in the core MHC region. This specific MHC-IIIB gene predates the common ancestor of the studied passerines and is thus a putative orthologue among the studied species. Additionally, all the MHC-IIIB genes in passerines in the current study belong to the DAB1 gene family, in line with previous studies (Burri et al., 2010). As for MHC-IIB, all the passerine MHC-I genes are monophyletic within species except for one single copy MHC-I gene found in the core MHC region. This specific MHC-I gene may predate the age of the Passerida clade and is found in a shared gene order in the zebra finch and great reed warbler, that is, it is a putative orthologue of the zebra finch and great reed warbler. Similarly, the single copy MHC-I gene found in the core MHC region of hooded crow and jackdaw is a putative orthologue among corvids, that is, species from the Corvida clade, based on large scale homology and phylogenetic reconstruction.

All four passerine long-read genome assemblies are of reasonably high quality, yet high MHC-I and -IIB gene diversities and occurrences of tandemly duplicated MHC-I and IIB genes are most pronounced in the great reed warbler genome. The great reed warbler belongs to the superfamily Sylvioida where high MHC diversities were expected based on previous studies (O'Connor et al., 2016, 2018), and we envision the higher MHC-IIIB diversity seen in the great reed warbler compared to the MHC-IIIB diversity in the zebra finch and jackdaw to be due to the MHC expansion in Sylvioida rather than differences in quality of the genome assemblies. However, in the hooded crow, the MHC-IIIB diversity is even lower than in the critically endangered, strongly bottlenecked Hawaiian crow Corvus hawaiiensis (Sutton et al., 2018), and our present estimate of MHC-I and MHC-IIIB diversities in the hooded crow is therefore likely to be an underestimation. Assembly metrics, such as genome coverage and N50, are used to estimate the overall genome quality, although as read coverage fluctuates locally it is the read coverage in the region of interest that matters (Clifton et al., 2020), as well as the complexity of the genomic region. Our target is the core MHC region, which holds tandemly duplicated MHC-I and MHC-IIB paralogues and has a high repeat content, thereby it is a region with high complexity and therefore very long reads are needed for a reliable assembly. Recent improvements in high fidelity long-read sequencing will aid the assembly of the core MHC region in future studies.

The MHC-I genes in passerines are considerably larger, more similar between loci, and have larger intergenic distances than the MHC-IIIB genes. These factors make it more challenging to assemble tandemly duplicated MHC-I genes compared to tandemly duplicated MHC-IIIB genes. Contigs that are 50 kb are readily assembled from long-read sequencing, and 50 kb contigs can hold up to eight tandemly duplicated MHC-IIIB genes, but not more than two MHC-I genes. The shorter average read lengths and lower sequencing coverage in the jackdaw and hooded crow compared to the great reed warbler may explain why tandemly duplicated MHC-I genes were not found or were very rare in corvids in the present study.

Tandemly duplicated MHC genes are often highly similar within species which makes these genomic regions demanding to first assemble correctly and second to infer haplotypes. Moreover, an additional complicating factor is the high repeat content in the MHC region, with LTRs being particularly frequent. In the great reed warbler we were able to infer the haploid MHC genomic region, not only using genome assembly tools, but also using segregation analyses based on high-throughput amplicon sequencing data from the genome individual and its parents (Mellinger et al., Manuscript). The number of MHC-IIIB genes in the haploid representation of the great reed warbler genome had been slightly overestimated by the genome assembly, and the scaffolds Aaru-DAB*357 and Aaru-DAB*554 are likely to be the two haplotypes of a single MHC-II gene region. Scaffold Aaru-DAB*357 is inherited from the father (indicated by six paternal amplicon alleles), and Aaru-DAB*554 from the mother (indicated by eight maternal amplicon alleles). The full-length genes on scaffolds Aaru-DAB*357 and Aaru-DAB*554 are frequently found in the same clusters in phylogenetic reconstructions, further suggesting they are alleles of the same genes rather than novel gene copies. The zebra finch contigs Tgut-DAB*310 and Tgut-DAB*910, with four and 15 MHC-IIIB genes in ORF, respectively, are likewise very likely to represent the two haplotypes of a single MHC-II gene region, as are the two jackdaw contigs Cmon-DAB*330 and Cmon-DAB*334, with 11 and seven MHC-IIIB genes in ORF.

Ectopic recombination, that is, unequal crossing-over at meiosis (Ohta, 2013), is a plausible mechanism for the MHC gene family expansions seen in tandemly duplicated MHC genes in passerines, since it generates duplications of neighbouring genes (Cordaux & Batzer, 2009; Hastings et al., 2009). According to the birth-and-death evolution model some of these duplicated genes then become pseudogenized while others are expressed and subject to positive selection (Nei et al., 1997). Our great reed warbler data aligns well with ectopic recombination and the birth-and-death model of evolution; a large proportion of the MHC genes in the great reed warbler are tandemly duplicated (49 out of 56 MHC-IIIB genes), and approximately 15% of the MHC genes have become pseudogenized. Note that we only included high quality gene hits in our analyses (MHC exon annotation). Moreover, past MHC gene duplication events of trios of MHC-I and MHC-IIIB genes within scaffolds can be seen in the great reed warbler genome. Simultaneous duplication of several
MHC genes should be particularly important for explaining the massive expansions of the tandemly duplicated MHC genes in passerines. However, when one chromosome homologue gains MHC gene copies the other chromosome homologue loses its copies. The mechanism of unequal crossing-over may therefore not only explain the massive expansion of the MHC-I and IIB in the great reed warbler but possibly also the loss of the tandemly duplicated MHC-I genes in the zebra finch. Tandemly duplicated MHC-I genes exist in the jackdaw, infraorder Corvida, and in the great reed warbler, infraorder Passerida, but not in the zebra finch, infraorder Passerida, yet it remains to be investigated when in evolutionary time the corresponding genes to the zebra finch MHC-I genes were lost. High MHC-I diversities dominate among passerines, and zebra finch is most probably the exception to the rule (He et al., 2021; O’Connor et al., 2016, 2018).

The repeat content in the MHC region in the great reed warbler is high (24.7%) compared to the overall repeat content in the rest of the great reed warbler genome (9.5%). These repeats may have contributed both to the expansion of the MHC genes and the gene rearrangements in the core MHC region compared to earlier bird orders. Repeats cannot be efficiently removed from low-recombining chromosomal regions, and as expected the great reed warbler ancestral W-chromosome, a sex-linked region that does not recombine, is extremely enriched for repeats compared to the rest of the genome (Sigeman et al., 2021). How then can the repeat content be so high in the rather large MHC region, approximately 5.5 Mb in the great reed warbler? Previous studies on MHC-I haplotypes in the great reed warbler and house sparrow P. domesticus pedigrees suggest recombination is rare among MHC genes (Karlsson & Westerdahl, 2013; Razali et al., 2017; Roved et al., 2020; Westerdahl et al., 2004), a finding which agrees with our current findings of a high repeat content. It should be noted that recombination in the MHC regions is low on a microevolutionary scale (within populations), whereas gene rearrangements have occurred repeatedly on a macroevolutionary scale (among species). Bursts of expansions of MHC genes during the passerine radiation may very well have been initiated by repeat insertions, a stochastic process. These repeats then become duplicated when, for example, trios of MHC genes were duplicated, as shown in the present study, whereafter the repeats remain among the tightly linked, rarely recombining, tandemly duplicated MHC genes. The expansion and reduction in MHC gene copy numbers is likely to be a stochastic process upon which selection can act. Massive expansion in MHC-I diversity has not only been seen in passerines, but also among fish in the Gadiformes lineage; the latter has been suggested to be an evolutionary response to the loss of the MHC-IIB genes (Malmström et al., 2016; Star et al., 2011). In line with this reasoning the loss of MHC-I in the zebra finch may have resulted in an expansion of the MHC-IIB diversity, although this is yet to be investigated in a phylogenetic context. Previous studies have reported negative consequences of too high MHC diversity, and according to theory there is a trade-off in MHC diversity (Nowak et al., 1992; Woelfling et al., 2009). Although the optimal expressed MHC diversity will be context dependent (O’Connor & Westerdahl, 2021), for example related to adaptations in the immune system and selection from pathogens, it will certainly differ between species, but also between populations within a species (O’Connor et al., 2018, 2020).

In most of our assembled scaffolds/contigs the tandemly duplicated MHC genes are placed separately in a gene context without any non-MHC genes. However, one set of tandemly duplicated MHC-IIB genes in the great reed warbler and zebra finch are placed in a known gene order, between a single classical MHC-IIA gene and the nonclassical MHC-II genes (DM). A single classical MHC-IIA gene, or fragments of a single MHC-IIA gene, was found in each of the passerines. It is therefore plausible that a large proportion of the MHC-IIB proteins encoded by the highly duplicated MHC-IIB genes, pair up with the proteins encoded by the single MHC-IIA gene to produce MHC-II molecules, a heterodimer with one alpha (IIA) and one beta (IIB) chain. It is possible that passerines compensate for the lack of MHC-IIA diversity by having a larger and more diverse complement of MHC-IIB genes, which would be particularly advantageous if the proteins encoded by the single MHC-IIA gene combine with all different MHC-IIB proteins to form a broad variety of MHC-II heterodimers. This organization of MHC-II genes is very different from the organisation in birds of the orders Ciconiiformes and Pelecaniformes in and mammals, where the MHC-II genes are found in tandem MHC-IIA-MHC-IIB dyads (Chen et al., 2015; Shiina et al., 2007; Tsuji et al., 2017). The MHC-II molecules in mammals are then formed per locus with one alpha and one beta copy, for example in humans there are three classical MHC-II loci (DP, DR and DQ) with at least one alpha- (DPA, DRA and DQA) and one beta- (DPB, DRB and DQB) gene per locus.

There is a multitude of organizations of the MHC region in birds from different orders, for example, in Galliformes the classical MHC-IIB genes are flanking tapasin, in Ciconiiformes and Pelecaniformes the classical MHC-IIB genes are found in dyads with MHC-IIA genes (and tapasin has been lost), and in Passeriformes the tandemly duplicated classical MHC-IIB genes are found next to a single MHC-IIA gene (and tapasin has been lost). Moreover, classical MHC-I genes flanking the TAP-genes in Galliformes, are found on one side only in Ciconiiformes and Pelecaniformes, and rather distant from the TAP-genes in Passeriformes. Indeed, there are comparatively large differences in the gene order of classical MHC-I and MHC-IIB genes among birds compared to mammals. In mammals, the classical MHC-II genes are found in dyads in the class II region, as are the TAP genes, while tapasin is found in the extended class II region, and the classical MHC-I genes are placed as single genes in the class I region far away from the TAP-genes (Shiina et al., 2009, 2017). Furthermore, the genomic MHC region differs considerably in size among different bird species and orders; the MHC genomic region in the great reed warbler (5.5 Mb) is actually more similar in size to human (4.0 Mb) than to chicken (0.24 Mb) (Shiina et al., 2007, 2009).

The quality and read lengths of long-read sequencing continuously improve which means better assemblies and more confident
ACKNOWLEDGEMENTS
This study was funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grants 679799 to HW and 336536 to JW), the Swedish Research Council (621-2016-689 to BH and 621-2013-4510 to JW) and the Knut and Alice Wallenberg Foundation (to JW). We thank NBIS for long-term bioinformatic support. ACR, EPW and VKA are financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. We would like to acknowledge support of NGI/Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and work performed at NGI/Uppsala Genome Center has been funded by RFI/the Swedish Research Council and Science for Life Laboratory, Sweden. We also would like to acknowledge the computational infrastructure provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Centre for Advanced Computational Science (UPPMAX) under the projects b2016230 and uppstore2017099. Finally, thanks to Inger Ekström for improving and arranging the figures.

CONFLICT OF INTEREST
The authors declare no competing interests.

AUTHOR CONTRIBUTIONS
Conceptualization, Helena Westerdahl and Maria Strandh; Methodology, Maria Strandh and Samantha Mellinger; Bioinformatic assembly, Hanna Sigeman and Iagnas Bunikis; Bioinformatic annotation and analyses, Samantha Mellinger, Verena E. Kutschera, Estelle Proux-Wéra, Allison Churcher and Max Lundberg; Formal analysis, Helena Westerdahl; Writing - original draft, Helena Westerdahl and Maria Strandh; Revision of manuscript, all coauthors.

DATA AVAILABILITY STATEMENT
Genome data are found at GenBank: Great reed warbler genome assembly acrAr1 is accessible under BioProject ID PRJNA765537 (Sigeman et al., 2021), hooded crow, Corvus cornix (Genbank accession GCA_002023255.1, Weissensteiner et al., 2017), jackdaw, Coloeus monedula (an earlier version of the current jackdaw reference assembly with Genbank accession GCA_013407035.1, Weissensteiner et al., 2020), zebra finch, Taeniopygia guttata (Tgtu_diploid_1.0, Genbank accession GCA_002008985.2) and Gallus gallus (GRG6a, GenBank accession GCA_000002315.5). Scripts are deposited on GitHub (https://github.com/ekol-hwe/MHC_GRW_genome_2022). The generated data files, the GFF files and the jackdaw MHC contigs are found in Dryad (https://doi.org/10.5061/dryad.fqz612jv6) along with manually curated fasta-files of full-length MHC-I and MHC-IIb genes in open reading frame.

ORCID
Helena Westerdahl https://orcid.org/0000-0001-7167-9805
Verena E. Kutschera https://orcid.org/0000-0002-8930-534X
Bengt Hansson https://orcid.org/0000-0001-6694-8169

REFERENCES
Balakrishnan, C. N., Ekblom, R., Völker, M., Westerdahl, H., Godínez, R., Kotkiewicz, H., Burt, D. W., Graves, T., Griffin, D. K., Warren, W. C., & Edwards, S. V. (2010). Gene duplication and fragmentation in the zebra finch major histocompatibility complex. BMC Biology, 8. https://doi.org/10.1186/1741-7007-8-29
Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6, 11.
Biedrzycka, A., O’Connor, E., Sebastian, A., Migalska, M., Radwan, J., Zając, T., Bielański, W., Solarz, W., Cmiel, A., & Westerdahl, H. (2017). Extreme MHC class I diversity in the sedge warbler (Acrocephalus schoenobaenus); Selection patterns and allelic divergence suggest that different genes have different functions. BMC Evolutionary Biology, 17, 1–2. https://doi.org/10.1186/s12862-017-0977-9
Botero-Castro, F., Fuguet, E., Tilak, M. K., Nabholz, B., & Galtier, N. (2017). Avian genomes revisited: Hidden genes uncovered and the rates versus traits paradox in birds. Molecular Biology and Evolution, 34, 3123–3131. https://doi.org/10.1093/molbev/msx236
Burri, R., Salamin, N., Studer, R. A., Roulin, A., & Fumagalli, L. (2010). Adaptive divergence of ancient gene duplicates in the avian MHC class II B. Molecular Biology and Evolution, 27, 2360–2374.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics, 10, 421. https://doi.org/10.1186/1471-2105-10-421
Chaves, L. D., Krueth, S. B., & Reed, K. M. (2009). Defining the Turkey MHC: Sequence and genes of the B locus. The Journal of Immunology, 183, 6530–6537. https://doi.org/10.4049/jimmunol.0901310
Chen, L. C., Lan, H., Sun, L., Deng, Y. L., Tang, K. Y., & Wan, Q. H. (2015). Genomic organization of the crested ibis MHC provides new insight into ancestral avian MHC structure. Scientific Reports, 5, 1–11. https://doi.org/10.1038/srep07963
Clifton, B. D., Jimenez, J., Kimura, A., Chahine, Z., Librado, B., Sanchez-Gracia, A., Abbassi, M., Carranza, F., Chan, C., Marchetti, M., & Zhang, W. (2020). Understanding the early evolutionary stages of a tandem drosophila melanogaster-specific gene family: A structural and functional population study. Molecular Biology and Evolution, 37, 2584–2600.
Wang, J. (2013). Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science*, 339(6118), 456–460. [https://doi.org/10.1126/science.1230835](https://doi.org/10.1126/science.1230835)

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., & Ödeen, A. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Proceedings of the Royal Society B: Biological Sciences*, 346, 1311–1321.

**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:** 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*, 22, 2379–2395. [https://doi.org/10.1111/1755-0998.13614](https://doi.org/10.1111/1755-0998.13614)