Immunogenetics
https://doi.org/10.1007/s00251-022-01255-8

ORIGINAL ARTICLE

Intra-specific copy number variation of MHC class II genes in the Siamese fighting fish

Anson Tsz Chun Wong[1] · Derek Kong Lam[1] · Emily Shui Kei Poon[1] · David Tsz Chung Chan[1] · Simon Yung Wa Sin[1]

Received: 12 September 2021 / Accepted: 4 February 2022

Abstract

Duplicates of genes for major histocompatibility complex (MHC) molecules can be subjected to selection independently and vary markedly in their evolutionary rates, sequence polymorphism, and functional roles. Therefore, without a thorough understanding of their copy number variation (CNV) in the genome, the MHC-dependent fitness consequences within a species could be misinterpreted. Studying the intra-specific CNV of this highly polymorphic gene, however, has long been hindered by the difficulties in assigning alleles to loci and the lack of high-quality genomic data. Here, using the high-quality genome of the Siamese fighting fish (*Betta splendens*), a model for mate choice studies, and the whole-genome sequencing (WGS) data of 17 *Betta* species, we achieved locus-specific amplification of their three classical MHC class II genes — *DAB1*, *DAB2*, and *DAB3*. By performing quantitative PCR and depth-of-coverage analysis using the WGS data, we revealed intra-specific CNV at the *DAB3* locus. We identified individuals that had two allelic copies (i.e., heterozygous or homozygous) or one allele (i.e., hemizygous) and individuals without this gene. The CNV was due to the deletion of a 20-kb-long genomic region harboring both the *DAA3* and *DAB3* genes. We further showed that the three DAB genes were under different modes of selection, which also applies to their corresponding DAA genes that share similar pattern of polymorphism. Our study demonstrates a combined approach to study CNV within a species, which is crucial for the understanding of multigene family evolution and the fitness consequences of CNV.

Keywords Copy number variation · Major histocompatibility complex · Gene duplication · Multi-gene family · *Betta splendens* · Whole-genome sequencing

Introduction

Copy number variation (CNV) refers to the differences in the number of repeats of a genomic segment among individuals arising from duplications and deletions (Spielmann et al. 2018; Sudmant et al. 2015). Duplications and deletions of both coding genes and non-coding regulatory elements are important sources for the evolution of genomic variability (Zarrei et al. 2015). CNV can affect phenotypes through mechanisms such as gene dosage, gene interruption, and gene fusion (Gamazon and Stranger 2015), which may drive ecological adaptation (Niimura et al. 2014; Rinker et al. 2019; Sin et al. 2021) and reproductive isolation that may ultimately promote speciation (Malmstrøm et al. 2016; Völker et al. 2010). The mutation rate (i.e., the frequency of *de novo* mutations in an organism over time) of CNV is orders of magnitude higher than that of single-nucleotide polymorphisms, but it may vary throughout the genome depending on local genome architecture (Zhang et al. 2009).

One of the notable examples of genes exhibiting CNV in vertebrates is the major histocompatibility complex (MHC), a multigene family that plays important roles in both pathogen resistance and mate choice (Hoover et al. 2018; Piertney and Oliver 2006; Sin et al. 2015; Yamaguchi and Dijkstra 2019). MHC genes comprise the class I and II genes, which encode cell surface proteins that bind and can present foreign peptides to T-cells and subsequently trigger immune cascades (Roche and Furuta 2015). In addition to their exceptionally high polymorphism, believed to be maintained primarily by pathogen-mediated balancing selection (Radwan et al. 2020; Spurgin and Richardson 2010) and sexual selection (Winternitz et al. 2013), MHC genes are also known to evolve through a birth-and-death process

✉ Simon Yung Wa Sin
sinyw@hku.hk

[1] School of Biological Sciences, The University of Hong Kong, Pok Fu Lam Road, Hong Kong SAR, China

Published online: 01 March 2022

by frequent duplications, inactivation, and deletions (Nei et al. 1997). As a result, extensive variations in both MHC genomic organization and gene copy number are commonly observed between species (Grimholt 2016; O'Connor et al. 2019).

Inter-specific CNV has been documented in some vertebrate groups. For example, MHC class IIb genes have undergone many duplications in some avian superfamilies (e.g., Sylvioidea, Passeroidea, and Muscicapoidea) since the ancestral duplication of a class IIb locus preceding major avian radiation (Burri et al. 2010). Avian species with highly duplicated MHC genes include *Hirundo rustica* (barn swallow) that has 43 copies and *Manacus vitellinus* (Golden-collared manakin) that has 193 copies (He et al. 2021). In teleosts, MHC class II genes are completely lost in the entire order of Gadiformes (Dijkstra and Grimholt 2018; Malmstrøm et al. 2016), *Lophius piscatorius* (angler fish; Dubin et al. 2019), and some seahorse species (Roth et al. 2020) but became highly duplicated in other species such as *Oreochromis niloticus* (Nile tilapia) that has 33 copies (Sato et al. 2012). For class I genes, copy number can exceed 100 in several gadiform species (Dijkstra and Grimholt 2018; Malmstrøm et al. 2016). The large inter-specific CNV may reflect the difference in optimal MHC diversity between species, which was shaped by factors that vary within and between populations of a species (O'Connor and Westerdahl 2021). The underlying evolutionary processes that shape the optimal MHC diversity may thus also drive CNV within a species.

MHC gene copy number within a species was commonly discussed along with MHC diversity to infer immunocompetence (Eizaguirre et al. 2011; Qurkhuli et al. 2019; Radwan et al. 2014; Stervander et al. 2020). A higher number of genes in theory represent a larger MHC repertoire, from which the encoded proteins bind a broader range of antigens. However, the antigen-binding breadth of MHC may also be limited by the intrinsic costs of having a high MHC diversity and superfluous MHC gene expression, which could increase the risk of autoimmunity due to the depletion of circulatory T-cell receptor repertoire (Bentkowski and Radwan 2019; Migalska et al. 2019). The optimal number of MHC genes may thus be different between populations of a species, depending on the difference in pathogen-mediated selective force (Llaurens et al. 2012). Although pathogen resistance due to MHC heterozygote advantage, rare-allele advantage (e.g., Sin et al. 2014), or optimal diversity (e.g., Sammut et al. 2002; Wegner et al. 2003) have been widely investigated, most studies on non-model species only determined the total number of MHC alleles. Only a few studies have revealed different number of loci among individuals within a species (Chain et al. 2014; Cheng et al. 2012; Málaga-Trillo et al. 1998). Without the knowledge on the gene copy number, zygosity of each locus, and the origin of each allele (i.e., which locus the allele belongs to), it is not possible to accurately interpret the effect of selection. Intra-specific CNV and its underlying evolutionary significance thus remain largely unexplored (Bentkowski and Radwan 2019).

One reason that intra-specific CNV is rarely reported is the difficulty of locus-specific characterizations, because MHC genes are highly polymorphic and duplicated in the genome. Previous studies predominantly relied on degenerate primers that co-amplify alleles from multiple loci (Babik 2010; Bernatchez and Landry 2003; Sin et al. 2012a, 2012b), but this approach usually failed to assign alleles to their corresponding locus. Even though locus-specific primers were used, homozygosity was usually assumed if only one allele was amplified (e.g., Liu et al. 2017; Zhai et al. 2017), whereas the possibility of hemizygosity was rarely considered (Lighten et al. 2014a). In addition, publicly available genomic data for non-model organisms has remained suboptimal for MHC studies until recently. Genomes of non-model organisms are mostly *de novo* assembled from short reads, hence highly duplicated MHC paralogs may collapse into a single location (O'Connor et al. 2019). This introduces alignment ambiguity when resequencing reads are mapped against a fragmented pseudo-haploid MHC reference assembly (i.e., a consensus sequence with unresolved haplotypes; Ekblom 2014; Nielsen et al. 2011), hindering the application of mapping results for MHC gene mining, locus-specific primer design, genotyping, and copy number estimation.

With the advance of long-read sequencing technology, highly contiguous genomes reaching chromosome-level become available for non-model species, which will facilitate characterization of MHC genes. In this study, using a chromosome-level genome assembly and WGS data of the Siamese fighting fish (*Betta splendens*) and other *Betta* species, we present an approach to assign MHC alleles to corresponding loci, to identify intra-specific CNV, and to determine the selection pattern in different loci. *Betta splendens* are freshwater labyrinth fishes indigenous to Southeast Asia and have a long history of selective breeding (Zhang et al. 2021). They exhibit remarkable intra-specific phenotypic and behavioral diversity (Monvises et al. 2009; Zhang et al. 2021) and have been widely studied for their mate choice behavior (e.g., Clotfelter et al. 2006; Cram et al. 2019; Dzieweczynski et al. 2014). The knowledge of their MHC variability will facilitate subsequent MHC-related studies such as on mate choice, pathogen resistance, and MHC gene evolution. Here, we (1) identify and genotype multiple MHC class IIa and IIb genes and (2) determine their intra-specific CNV in *B. splendens* using both bioinformatics and laboratory-based approaches. We then (3) investigate the sequence variability, selection patterns, and comparative genomic and phylogenetic relationships of the identified DAB loci to provide a comprehensive picture of MHC gene evolution in *B. splendens* and its close relatives.

Materials and methods

MHC class II gene mining and genomic region reconstruction

The full-length and individual exons of DAB coding sequences (CDS) from multiple teleost species (*Monopterus albus* (KC427217.1), *O. niloticus* (MH220769.1), and *Gasterosteus aculeatus* (DQ016400.1)) were used as queries to blast against the high-quality *B. splendens* genome (GCF_900634795.2), using BLASTN and TBLASTX (Altschul et al. 1990). Genes identified were verified by performing reciprocal BLASTN against the GenBank database and translated amino acid sequence alignment using MEGA X (Kumar et al. 2018). The exon–intron organizations of full-length DAB genes were determined following the GT/AG rule. Three DAB genes were eventually identified and named as *Besp-DAB1*, *Besp-DAB2*, and *Besp-DAB3*, respectively, following the MHC nomenclature rule (Klein et al. 1990).

The organization of all MHC IIa and IIb genes in the *B. splendens* genome was reconstructed according to the genome annotation. Genomic organizations of classical DAB regions in the *B. splendens*, anabantoids, and other teleost species were then compared. With reference to previous studies on teleost MHC II synteny (Dijkstra et al. 2013; Grimholt 2016), syntenic genes up- and down-stream of DAB loci were searched against annotated genomes of *B. splendens*, *Anabas testudineus* (climbing perch; GCF_900324465.2), *Channa argus* (northern snakehead; GCA_004786185.1), and *Mastacembelus armatus* (zigzag eel; GCF_900324485.2). The genes were further validated by performing reciprocal BLASTN and assembled to reconstruct the putative MHC II regions. Using the most updated genome assemblies of *Danio rerio* (zebrafish; GCF_000002035.6), *Gasterosteus aculeatus* (three-spined stickleback; GCA_006229165.1), and *Oreochromis niloticus* (GCF_001858045.2), we also reconstructed their MHC II regions of which previous assemblies had been analyzed by Dijkstra et al. (2013).

Locus-specific DAB primer design

We analyzed the WGS data of *B. splendens* ($n=6$) to design primers at the conserved sites specific to each of the three DAB loci (Table S1). Individual raw reads were retrieved from the Sequence Read Archive (SRA) database and evaluated with the FastQC v0.11.7 (Andrews 2010). Adapter sequences, reads with quality score lower than 20, and length lower than 50 were trimmed using the Trimmomatic v0.38 (Bolger et al. 2014). Trials of read mapping to the *B. splendens* genome were performed using BWA v0.7.17 (Li and Durbin 2009) and NextGenMap v0.5.5 (Sedlazeck et al. 2013) with "sensitive" preset parameters. The performance of the two software on mapping polymorphic MHC genes was assessed. Mapped SAM files were converted to BAM format and sorted using the SAMtools v0.1.09 (Li et al. 2009). Duplicates were marked with MarkDuplicates in Picard Tools (http://broadinstitute.github.io/picard/). Processed BAM files were indexed and loaded into the Integrative Genomics Viewer (IGV) v2.8.0 (Thorvaldsdóttir et al. 2012) to visualize the mapping results.

Exon 2 consensus sequences of each DAB locus were extracted for each individual and aligned with GenBank sequences from synbranchids (JQ236680.1) and pleuronectids (KJ784489), and more distantly related mugilids (AF134941), cichlids (JN967618, AH002424), cyprinids (GU441571, AY103492), and salmonids (FJ597523, AF296385) using ClustalW (Thompson et al. 1994). Locus-specific primers were designed at the exon 1, intron 1, and intron 2 regions (Table S2) using NetPrimer (http://www.premierbiosoft.com/netprimer/index.html).

DNA extraction, PCR, cloning, and sequencing

We extracted gDNA from the muscle tissues of *B. splendens* ($n=17$) sourced from aquatic pet shops using the E.Z.N.A. Tissue DNA Kit (Omega, USA), following the manufacturer's protocol. We quantified the DNA concentration using the Qubit dsDNA HS kit (ThermoFisher Scientific, USA). PCR amplification was performed in a 20-μl reaction mix containing 1 × GoTaq reaction buffer, 1% DMSO, 3 mM $MgCl_2$, 0.2 mM dNTP, 0.2 μM forward primer, 0.2 μM reverse primer, 10–50 ng of gDNA, and 1 unit of GoTaq polymerase (Promega). The PCR cycle began with an incubation at 95 °C for 2 min, followed by 30 cycles of incubation at 95 °C for 30 s, 56–60 °C for 30 s (Table S2), and 72 °C for 1 min, and ended with a final extension at 72 °C for 10 min. The PCR products were electrophoresed on 1.5% agarose gel, and those with the expected band size were sent for Sanger sequencing (BGI, Hong Kong).

Next, PCR products containing more than one sequences were proceeded to cloning and sequencing for allele identification. The PCR for cloning was performed with the described condition in a 30-μl reaction mix. The PCR products were electrophoresed on 1% agarose gel, followed by gel purification using the PureLink Quick Gel Extraction Kit (Invitrogen, USA). The purified PCR products were then ligated into a T-Vector pMD19 (TaKaRa, US) using DNA ligation kit (TaKaRa). Transformation was performed using heat shock and DH5α competent cells with blue-white screening. Colony PCRs were performed on white colonies using the universal M13 primers (forwards: 5′-CACGAC GTTGTAAAACGAC-3′; reverse: 5′-CAGGAAACAGCT

ATGACC-3′) with the same PCR condition as described earlier. At least 8 clones per gene were sequenced for each individual. Allele identity was confirmed by at least two independent PCRs. DAB alleles were named with reference to the IPD-MHC Database requirement (Maccari et al. 2018).

Genotyping from whole-genome and transcriptome sequencing data

Individual WGS data of *B. splendens* and other 16 *Betta* species ($n = 72$) available in the SRA database were processed using the same method described above (Table S1). Mapping was performed using the NextGenMap v0.5.5 only, which performs better than BWA in aligning the highly polymorphic MHC regions in our analysis. Transcriptomic sequencing (RNA-seq) reads of *B. splendens* individuals ($n = 14$) were mapped against the genome supplied with the RefSeq annotation in GTF format using the STAR v2.7.3 (Dobin et al. 2012), and duplicates were not marked (Table S1). We then visualized the mapping results in the IGV and extracted the exon 2 sequences of *DAA1–3* and *DAB1–3* from all individuals under the following criteria: (1) sequencing depth-of-coverage of the CDS is at least 10 × (Song et al. 2016); (2) sequences extracted from overlapping sequencing reads are contiguous without gaps (Fig. S1); and (3) identified alleles contain conserved teleost-specific key amino residues (Dijkstra et al. 2013).

Copy number variation analysis

For some individuals, no WGS reads were mapped to the *DAB3* locus, or no *DAB3* alleles could be isolated from cloning and sequencing, which indicated possible CNV at this locus. We therefore performed quantitative PCR (qPCR) and depth-of-coverage (DoC) analysis of WGS data to determine the copy number of *DAB3* gene in each individual.

Quantitative PCR

We designed a pair of primers that specifically amplify both the exon 3 regions of *DAB2* and *DAB3* genes (Table S2). Both β-actin (*ACTB*) and ribosomal protein 17 (*RPL17*) genes were used as the reference genes, which should have two allelic copies in each individual. The primers for the reference genes were designed on exon 4 (Table S2). The amplification efficiencies of these three primer pairs were 95–105%. qPCR was performed in triplicates per sample ($n = 17$) in a 15-μl reaction mix containing 2 × iTaq Universal SYBR Green Supermix (Bio-Rad, US), 0.4 μM forward primer, 0.4 μM reverse primer, and 20 ng gDNA using the CFX96 Torch Real-Time PCR Detection System (Bio-Rad, US). qPCR began with an incubation of 95 °C for 2 min, followed by 40 cycles of incubation at 95 °C for 10 s and 60 °C for 30 s. To confirm the specificity of the reactions, melt curve analysis was performed from 60 to 95 °C with 0.5 °C increments per step. The ΔΔCq method was used to calculate the relative copy number (RCN) with the equation $\mathrm{RCN} = 2 \times 2^{-\Delta\Delta C_q}$, where the calibrator has two copies per diploid genome (Weaver et al. 2010).

DoC analysis

From *B. splendens* WGS data ($n = 29$), number of reads uniquely mapped onto the exonic regions of the *DAB1*, *DAB3*, *ACTB*, and *RPL17* genes were calculated using SAMtools v0.1.09 (Li et al. 2009) and normalized by their corresponding sequence lengths. Both *ACTB* and *RPL17* were used as the reference genes. We calculated the numbers of *DAB1*, *DAB3*, and *ACTB* allelic copies relative to *RPL17*. The RCN determined by qPCR and DoC were plotted using R v4.0.2 (R Core Team 2020).

To identify the breakpoint corresponding to the loss of the genomic region containing *DAB3*, we analyzed the DoC of the region spanning the *DAB2* and *DAB3* genes. Normalized DoC in *DAB3*-containing and *DAB3*-missing individuals was compared to determine the position and size of the genomic deletion.

Selection analysis

Translated amino acid of identified alleles was aligned using ClustalW (Thompson et al. 1994). Putative binding sites (PBSs) were assigned according to previous characterization studies (Brown et al. 1993; Li et al. 2014). Nucleotide diversity (π), rates of synonymous (d_S), and non-synonymous (d_N) substitutions for PBS and non-PBS were calculated using DnaSP v6.12.03 (Rozas et al. 2017) according to the modified Nei-Gojobori method (1986) with Jukes-Cantor correction. Standard errors were obtained with 1000 bootstrap replicates. The d_N/d_S ratio (ω) and *Z*-test were computed to infer signs of selection on PBS and non-PBS at statistical significance level of $P < 0.05$, with $\omega > 1$ indicating positive selection while $\omega < 1$ indicating purifying selection.

We further tested for positive selection at specific codons separately in the β1-domain for the three DAB loci using the HyPhy package (Pond et al. 2004) implemented in the Datamonkey web server (www.datamonkey.org; Delport et al. 2010). Because recombination may cause false positive results in selection tests employing likelihood ratio tests (LRT; Anisimova et al. 2003), we used Genetic Algorithm Recombination Detection (GARD; Kosakovsky Pond et al. 2006) in the HyPhy package to identify significant recombination breakpoints and split sequences into partitions accordingly for subsequent selection analyses. We then used mixed effects model of evolution (MEME), fixed effects likelihood (FEL), and fast unconstrained Bayesian AppRoximation (FUBAR) methods implemented in HyPhy to infer signals of positive selection. All methods were used with default settings.

Phylogenetic analysis

Phylogenetic relationships among DAB alleles of *B. splendens* and 16 *Betta* species were reconstructed using the Bayesian inference approach. DAB sequences from *Anabas testudineus* (XM_026369498.2, XM_026352900.2, XM_033326567.1), *Channa argus* (EXN66_Car014117, EXN66_Car014106, ENX66_Car014124), *Monopterus albus* (KC427217.1, KC247222.1), *Paralichthys olivaceus* (HQ635018.1, HQ635058.1), *Trachinotus ovatus* (KX181520.1, KX181525.1), *Oreochromis niloticus* (MH220769.1, MH220773.1, AB677259.1), *Poecilia reticulata* (retrieved from Llaurens et al. (2012)), *Gasterosteus aculeatus* (DQ016400.1, DQ016409.1), and *Salmo salar* (FJ597530.1, FJ597533.1) were also included in the analysis. The human *HLA-DRB1* (KU947990.1) was used as an outgroup. We ran jModeltest v2.1.10 (Darriba et al. 2012) to identify the best-fit nucleotide substitution model, which was the Juke-Cantor model with a gamma distribution, based on the corrected Akaike information criterion (AIC_c). Two independent runs of four Markov chain Monte Carlo (MCMC) chains were then run in MrBayes v3.2.7a (Ronquist et al. 2012) for 6 million generations with a sampling frequency of every 100 generations and the first 25% of the tree samples being discarded as “burn-in”. The consensus tree was plotted using FigTree v1.4.4 (Rambaut 2018). In addition, to visualize recombination events causing the evolutionary relationships among *B. splendens* DAB alleles deviated from a bifurcating tree, the phylogenetic network was reconstructed using the Neighbor-Net analysis implemented in SplitsTree4 (Huson and Bryant 2005).

Results

Sequence variation of the MHC II genes

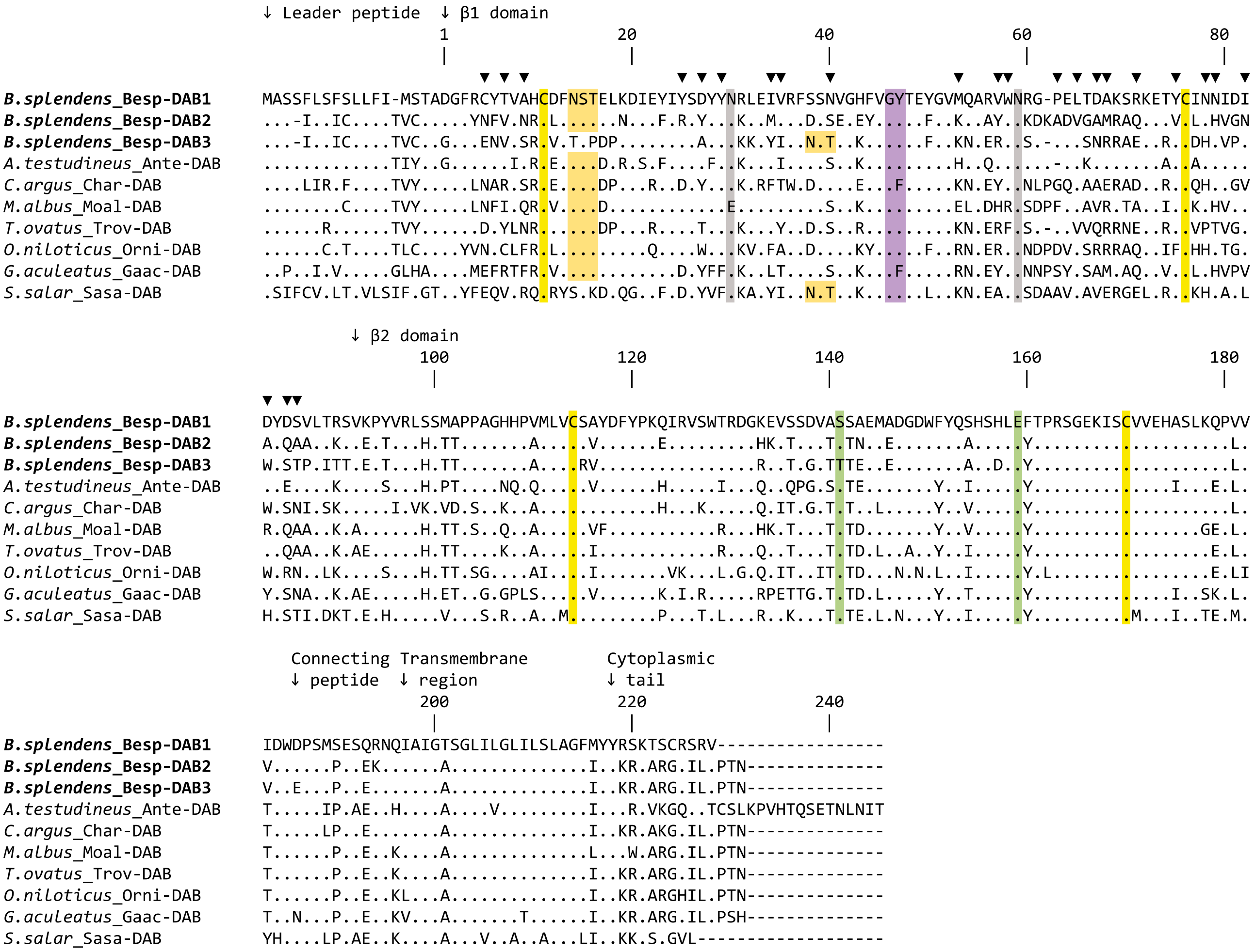
BLAST search against the *B. splendens* genome altogether identified three classical DAB (all in chromosome 2: Fig. 1) and four non-classical DBB loci (all in chromosome 16: Fig. S2). Our study focuses on the classical DAB genes, which are highly polymorphic and contain key residues for peptide ligand binding (Dijkstra et al. 2013; Wu et al. 2021). The full-length CDS of *DAB1*, *DAB2*, and *DAB3* genes were found to encode for 244, 248, and 247 amino acids, respectively, consistent with the three predicting DAB sequences from the genome annotation (GCF_900634795.2: Fig. 1). Along with the homologous sequences of other teleost species, the majority of variable sites in *B. splendens* DAB genes resided in the β1 domain, while the leader peptide, latter part of the β2 domain, and the transmembrane region showed a high amino acid identity (Fig. 1). Among the proteins encoded by the three DAB genes, DAB2 and DAB3 displayed a considerably higher similarity with each other than with DAB1 in the leader peptide and β2 domain, and DAB1 also had a shorter cytoplasmic tail than DAB2 and DAB3 (Fig. 1).

Sequence comparisons of DAB amino acid sequences among *Betta* and other teleost species revealed an overall conservation of several key amino acid residues known to be important for MHC II functions (Fig. 1, Fig. S3; Brown et al. 1993; Dijkstra et al. 2013). These included the four cysteine residues predicted to form disulfide bridges, and the N14, S15, and T16 in DAB1-2 (or N38, S39, T or S40 in *DAB3*) for *N*-linked glycosylation. The S141 and E159 in the β2 domain, known to bind CD4 molecules in mammals (Dijkstra et al. 2013), were almost conserved in *Betta* DAB sequences, except that the serine residue was replaced with threonine in DAB3. On the other hand, the H78 and N79 in the β1 domain that are known to form hydrogen bonds with the backbones of peptide ligands in mammalian species (Fremont et al. 1998) were not entirely conserved among *Betta* species (Fig. S3). The H78 was entirely and partially replaced with asparagine at DAB1 and DAB2-3 respectively, whereas the N79 in seven of the DAB2 sequences was replaced with histidine. The variation of histidine and asparagine at this position was previously found in classical MHC IIb genes of teleosts (e.g., *Danio rerio*; Dijkstra et al. 2013) and cartilaginous fishes (Wu et al. 2021).

Locus-specific genotyping of *B. splendens* individuals from cloning and sequencing ($n = 18$) and high-throughput sequencing data ($n = 27$ for WGS, $n = 12$ for RNA-seq, and $n = 2$ for both) altogether identified 4, 14, and 12 exon 2 alleles for *DAB1*, *DAB2*, and *DAB3*, respectively, with some alleles shared among different individuals (Figs. 2 and 3). Both genotyping methods yielded either one or two alleles per locus in *B. splendens*. Allelic sequences could not be extracted for *DAB2* in 15 WGS individuals (Fig. 3) due to the presence of mapping gaps. This is attributed to poor mapping at *DAB2*, which had a high sequence diversity, so that reads belonging to *DAB2* exon 2 might not map to the reference (Table S4).

Furthermore, cross-species mapping of WGS data ($n = 43$) from other 16 *Betta* species identified a total of 20, 34, and 18 alleles for *DAB1*, *2*, and *3*, respectively. Most of the genotyped WGS individuals contained either one or two alleles per locus, except *B. brownorum*, *B. burdigala*, and *B. pulchra*, which had three allelic sequences identified as *DAB2* (Table S3).

It is worth noting that our genotyping method may not apply to cases in which the sequence variants within a locus



```
                            ↓ Leader peptide  ↓ β1 domain
                                              1                 20                  40                  60                  80
                                              |                  |                   |                   |                   |
                                                  ▼ ▼ ▼               ▼ ▼ ▼    ▼▼    ▼            ▼   ▼▼    ▼ ▼ ▼▼  ▼   ▼  ▼▼  ▼
B.splendens_Besp-DAB1       MASSFLSFSLLFI-MSTADGFRCYTVAHCDFNSTELKDIEYIYSDYYNRLEIVRFSSNVGHFVGYTEYGVMQARVWNRG-PELTDAKSRKETYCINNIDI
B.splendens_Besp-DAB2       ...-I..IC....TVC.....YNFV.NR.L......N...F.R.Y...K..M...D.SE.EY.....F..K..AY..KDKADVGAMRAQ...V.L.HVGN
B.splendens_Besp-DAB3       ...-I..IC....TVC..G...ENV.SR.V.T.PDP........A...KK.YI..N.T..K......F..KN.ER..S.-...SNRRAE..R..DH.VP.
A.testudineus_Ante-DAB      .............TIY..G......I.R.E....D.R.S.F....F..K...I....S..K.........H..Q......-..K.......A..A.....
C.argus_Char-DAB            ....LIR.F....TVY.....LNAR.SR.E....DP...R..D.Y...K.RFTW.D....E...F.....KN.EY..NLPGQ.AAERAD..R..QH..GV
M.albus_Moal-DAB            ........C....TVY.....LNFI.QR.V....D............E.........S..K.........EL.DHR.SDPF..AVR.TA..I..K.HV..
T.ovatus_Trov-DAB           ......R......TVY......D.YLNR......DP...R....T...K...Y..D.S..R.........KN.ERF.S.-..VVQRRNE..R..VPTVG.
O.niloticus_Orni-DAB        ......C.T....TLC....YVN.CLFR.L.........Q....W...KV.FA..D....KY.....F..RN.ER..NDPDV.SRRRAQ..IF.HH.TG.
G.aculeatus_Gaac-DAB        ..P..I.V.....GLHA....MEFRTFR.V............D.YFF.K..LT....S..K...F.....RN.EY..NNPSY.SAM.AQ..V..L.HVPV
S.salar_Sasa-DAB            .SIFCV.LT.VLSIF.GT..YFEQV.RQ.RYS.KD.QG..F.D.YVF.KA.YI..N.T..K......L..KN.EA..SDAAV.AVERGEL.R..KH.A.L

                                     ↓ β2 domain
                                           100                 120                 140                 160                 180
                                             |                   |                   |                   |                   |
                            ▼ ▼▼
B.splendens_Besp-DAB1       DYDSVLTRSVKPYVRLSSMAPPAGHHPVMLVCSAYDFYPKQIRVSWTRDGKEVSSDVASSAEMADGDWFYQSHSHLEFTPRSGEKISCVVEHASLKQPVV
B.splendens_Besp-DAB2       A.QAA..K..E.T...H.TT.......A.....V......E.........HK.T...T.TN..E.......A.....Y....................L.
B.splendens_Besp-DAB3       W.STP.ITT.E.T...H.TT.......A....RV................R..T.G.TTTE..E.......A..D..Y....................L.
A.testudineus_Ante-DAB      ..E....K....S...H.PT....NQ.Q.....V......H.....I...Q..QPG.S.TE.......Y..I.....Y..............I...E.L.
C.argus_Char-DAB            W.SNI.SK.....I.VK.VD.S..K..A............H...K.....Q.IT.G.T.T..L.....Y..V.....Y....................L.
M.albus_Moal-DAB            R.QAA..K.A......H.TT.S..Q..A.....VF...........R...HK.T...T.TD.......Y..V.....Y.................GE.L.
T.ovatus_Trov-DAB           ..QAA..K.AE.....H.TT....K..A.....I............R...Q..T...T.TD.L..A..Y..I.....Y..................E.L.
O.niloticus_Orni-DAB        W.RN..LK....S...H.TT.SG....AI....I.......VK...L.G.Q.IT..IT.TD...N.N.L..I.....Y.L................E.LI
G.aculeatus_Gaac-DAB        Y.SNA..K.AE.....H.ET..G.GPLS.....V......K.I.R.....RPETTG.T.TD.L.....Y..T.....Y..............I..SK.L.
S.salar_Sasa-DAB            H.STI.DKT.E.H.....V...S.R..A..M.........P...T.L...R..K...T.TE.L.N...Y..I.....Y..........M...I..TE.M.

                              Connecting Transmembrane        Cytoplasmic
                              ↓ peptide  ↓ region             ↓ tail
                                           200                 220                 240
                                             |                   |                   |
B.splendens_Besp-DAB1       IDWDPSMSESQRNQIAIGTSGLILGLILSLAGFMYYRSKTSCRSRV-----------------
B.splendens_Besp-DAB2       V......P..EK......A..............I..KR.ARG.IL.PTN--------------
B.splendens_Besp-DAB3       V..E...P..E.......A..............I..KR.ARG.IL.PTN--------------
A.testudineus_Ante-DAB      T.....IP.AE..H....A....V.........I...R.VKGQ..TCSLKPVHTQSETNLNIT
C.argus_Char-DAB            T.....LP..E.......A..............I..KR.AKG.IL.PTN--------------
M.albus_Moal-DAB            T......P..E..K....A..............L...W.ARG.IL.PTN--------------
T.ovatus_Trov-DAB           T......P..E..K....A..............I..KR.ARG.IL.PTN--------------
O.niloticus_Orni-DAB        T......P..E..KL...A..............I..KR.ARGHIL.PTN--------------
G.aculeatus_Gaac-DAB        T..N...P..E..KV...A.......T......I..KR.ARG.IL.PSH--------------
S.salar_Sasa-DAB            YH....LP.AE..K....A...V..A..A...LI..KK.S.GVL-------------------
```

Fig. 1 Amino acid sequence alignment of the classical *DAB* genes for *Betta splendens* (highlighted in bold), anabantoids and other teleosts. GenBank accession numbers are XP_028994021 (*Besp-DAB1*), XP_028998859 (*Besp-DAB2*), XP_028986550 (*Besp-DAB3*), XP_033182458.1 (*Anabas testudineus*), EXN66_Car014117 (*Channa argus*), AGZ05746 (*Monopterus albus*), APD68833 (*Trachinotus ovatus*), AYN72181 (*Oreochromis niloticus*), and CAA27925 (*Salmo salar*). The sequence for *Gasterosteus aculeatus* was retrieved from Llaurens et al. (2012). The complete amino acid sequence of Besp-DAB1is shown. Numbers above the sequence represent the codon position starting from the β1 domain. Letters and dots represent residues that are distinct from or identical to Besp-DAB1respectively, whereas hyphens represent gaps. Inverted triangles denote known or putative peptide-binding sites (PBS) predicted from previous MHC studies (Dijkstra et al. 2013; Li et al. 2014). The CD4-binding residues are highlighted in green. The conserved cysteine residues known to form a disulfide bridge and N-glycosylation motifs are highlighted in yellow and orange respectively. The jawed-vertebrate-specific asparagine residues are highlighted in grey. The teleost-specific glycine and tyrosine residues are highlighted in purple

are more divergent than between loci, such as the MHC class I genes in some fish species (e.g., Shiina et al. 2005, Miller et al. 2006) or when there was recombination between loci (Dijkstra et al. 2006; Nonaka and Nonaka 2010).

Differences in sequence diversity among the three Besp-DAB genes

All DAB genes are putatively functional given the absence of frameshifts and in-frame stop codons (Fig. 1), and no pseudogenes were observed. They were all identified in the RNA-seq data (Table S4), but the expression level of *DAB1* was much lower than that of *DAB2* and *DAB3*. While most alleles from *DAB2* and *DAB3* differed by more than three amino acids, the four identified alleles at *DAB1* either only differed by 1–2 non-synonymous substitutions or a single synonymous substitution (i.e., *Besp-DAB1*010,101* and *Besp-DAB1*010,102*) that gave rise to identical amino acid sequences (Fig. 2). Besides, the PBS encoded by *DAB2* and *DAB3* had a high nucleotide diversity (π = 0.391 and 0.297), while that of *DAB1* was as low as 0.009 (Table 1). We further genotyped WGS individuals for the three DAA genes that encode α1 domains and found that they shared similar patterns of sequence diversity with their corresponding DAB genes (Fig. S4; Table S5). *DAA1* has a considerably lower sequence diversity, whereas *DAA2* and *DAA3* sequences are

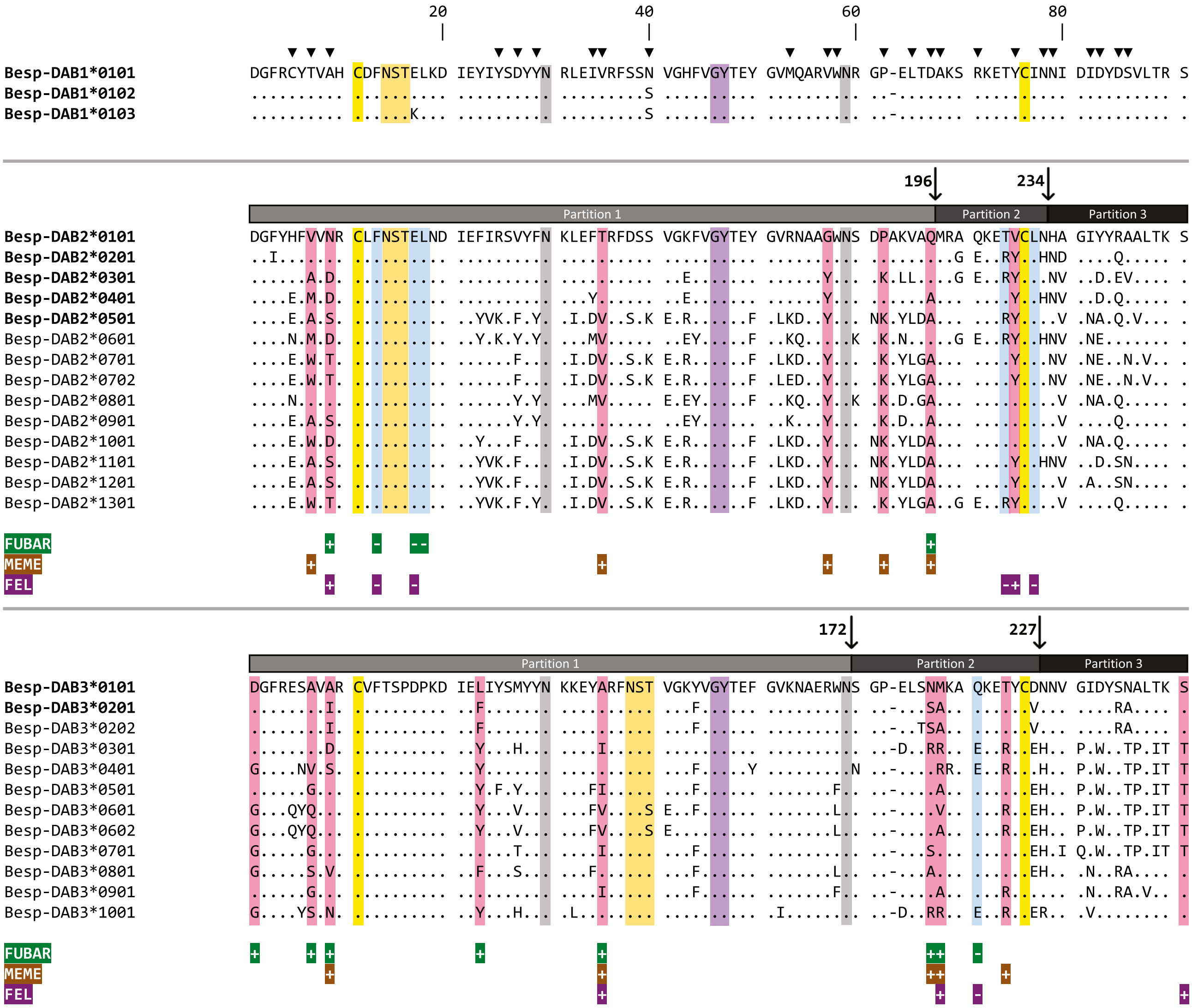

20
40
60
80
Besp-DAB1*0101 DGFRCYTVAH CDFNSTELKD IEYIYSDYYN RLEIVRFSSN VGHFVGYTEY GVMQARVWNR GP-ELTDAKS RKETYCINNI DIDYDSVLTR S
Besp-DAB1*0102 ....S ....-....... ....
Besp-DAB1*0103 ....K... ....S ....-....... ....
196
234
Partition 1
Partition 2
Partition 3
Besp-DAB2*0101 DGFYHFVVNR CLFNSTELND IEFIRSVYFN KLEFTRFDSS VGKFVGYTEY GVRNAAGWNS DPAKVAQMRA QKETVCLNHA GIYYRAALTK S
Besp-DAB2*0201 ..I....... ....G E..RY..HND ....Q..... .
Besp-DAB2*0301 ......A.D. ....E....... ......Y... .K.LL....G E..RY..NV ..D.EV.... .
Besp-DAB2*0401 ....E.M.D. ....Y...... ..E....... ......Y... .....A... ....Y..HNV ..D.Q..... .
Besp-DAB2*0501 ....E.A.S. ....YVK.F.Y. .I.DV..S.K E.R......F .LKD..Y... NK.YLDA... ...RY....V .NA.Q.V... .
Besp-DAB2*0601 ....N.M.D. ....Y.K.Y.Y. ...MV..... ..EY.....F ..KQ.....K .K.N.....G E..RY..HNV .NE....... .
Besp-DAB2*0701 ....E.W.T. ....F... .I.DV..S.K E.R......F .LKD..Y... .K.YLGA... ....Y...NV .NE..N.V.. .
Besp-DAB2*0702 ....E.W.T. ....F... .I.DV..S.K E.R......F .LED..Y... .K.YLGA... ....Y...NV .NE..N.V.. .
Besp-DAB2*0801 ....N..... ....Y.Y. ...MV..... E.EY.....F ..KQ..Y..K .K.D.GA... ....V .NA.Q..... .
Besp-DAB2*0901 ....E.A.S. ....Y.Y. ....EY...... ...K...Y... .K.D..A... ....V ....Q..... .
Besp-DAB2*1001 ....E.W.D. ....Y...F... .I.DV..S.K E.R......F .LKD..Y... NK.YLDA... ....V .NA.Q..... .
Besp-DAB2*1101 ....E.A.S. ....YVK.F... .I.DV..S.K E.R......F .LKD..Y... NK.YLDA... ....Y..HNV ..D.SN.... .
Besp-DAB2*1201 ....E.A.S. ....YVK.F... .I.DV..S.K E.R......F .LKD..Y... NK.YLDA... ....Y....V .A..SN.... .
Besp-DAB2*1301 ....E.W.T. ....YVK.F.Y. .I.DV..S.K E.R......F .LKD..Y... .K.YLGA..G E..RY....V ....Q..... .
FUBAR
MEME
FEL
172
227
Partition 1
Partition 2
Partition 3
Besp-DAB3*0101 DGFRESAVAR CVFTSPDPKD IELIYSMYYN KKEYARFNST VGKYVGYTEF GVKNAERWNS GP-ELSNMKA QKETYCDNNV GIDYSNALTK S
Besp-DAB3*0201 ....I.. ....F....... ....F...... ....-...SA.. .....V.... ....RA.... .
Besp-DAB3*0202 ....I.. ....F....... ....F...... ....-...TSA.. .....V.... ....RA.... .
Besp-DAB3*0301 ....D.. ....Y...H... ....I..... ...F...... ....-D..RR.. E..R...EH.. P.W..TP.IT T
Besp-DAB3*0401 G....NV.S. ....Y....... ....F.....Y ....N ..-....RR. E..R...H.. P.W..TP.IT T
Besp-DAB3*0501 ......G... ....Y.F.Y... ...FI..... ...F...... ....F.. ..-....A.. ......EH.. P.W..TP.IT T
Besp-DAB3*0601 G...QYQ... ....Y...V... ...FV....S E..F...... ....L.. ..-....V.. ...R...EH.. P.W..TP.IT T
Besp-DAB3*0602 G...QYQ... ....Y...V... ...FV....S E......... ....L.. ..-....A.. ...R...EH.. P.W..TP.IT T
Besp-DAB3*0701 G.....G... ....T... ....I..... ...F...... ....-....S.. ......EH.I Q.W..TP.IT T
Besp-DAB3*0801 G.....S.V. ....F...S... ...F...... ...F...... ....L.. ..-....A.. ......EH.. .N..RA.... .
Besp-DAB3*0901 ......G... ....I..... ...F...... ....F.. ..-....A.. ...R...... .N..RA.V.. .
Besp-DAB3*1001 G....YS.N. ....Y...H... .L........ .... .I........ ..-D..RR.. E..R..ER.. .V........ .
FUBAR
MEME
FEL

Fig. 2 Amino acid sequence alignment, recombination breakpoints, and selection analysis results of the exon 2 alleles identified from cloning-sequencing (highlighted in bold) and whole genome or transcriptome data for the three DAB genes of *Betta splendens*. The complete amino acid sequences of Besp-DAB1*0101, Besp-DAB2*0101, and Besp-DAB3*0101 are shown. Numbers above the alignment represent the codon position starting from the β1 domain. Letters and dots in the sequences represent residues that are distinct from or identical to the first sequence of each locus,respectively, whereas hyphens represent gaps. Inverted triangles denote putative peptide-binding sites (PBSs) predicted from previous MHC studies. The conserved cysteine residues known to form a disulfide bridge and *N*-glycosylation motifs are highlighted in yellow and orange respectively. The highly conserved residues among jawed vertebrates and ray-finned fishes are highlighted in grey and purple respectively. Grey bars and numerated arrows above DAB2 and DAB3 sequences refer to their corresponding partitions and recombination breakpoints identified from GARD tests. Codon sites under positive or purifying selection are highlighted in red and indicated as “+” or in blue and indicated as “–”, respectively. Green, brown and purple colors representing results from FUBAR, MEME and FEL, respectively

polymorphic across the α1 domains. Furthermore, *DAB1* sequences appeared to be highly conserved throughout the *Betta* genus (Fig. S3). *B. mahachaiensis*, *B. siamorientalis*, *B. smaragdina*, and *B. imbella*, which are most closely related to *B. splendens* (Panijpan et al. 2014; Rüber et al. 2004), shared at least one identical DAB1 sequence at the amino acid level with *B. splendens*. DAB1 in other *Betta* species were different from Besp-DAB1 in only 2–4 PBSs (Fig. S3). In contrast, the β1 domains of DAB2 and DAB3 were variable among *Betta* species.

Comparative genomic organization of MHC II regions

All DAA and DAB genes were located on chromosome 2, and all DBA and DBB genes were located on chromosome

Individual	Besp-DAB1*010101	Besp-DAB1*010102	Besp-DAB1*0102	Besp-DAB1*0103	Besp-DAB2*0101	Besp-DAB2*0201	Besp-DAB2*0301	Besp-DAB2*0401	Besp-DAB2*0501	Besp-DAB2*0601	Besp-DAB2*0701	Besp-DAB2*0702	Besp-DAB2*0801	Besp-DAB2*0901	Besp-DAB2*1001	Besp-DAB2*1101	Besp-DAB2*1201	Besp-DAB2*1301	Besp-DAB3*0101	Besp-DAB3*0201	Besp-DAB3*0202	Besp-DAB3*0301	Besp-DAB3*0401	Besp-DAB3*0501	Besp-DAB3*0601	Besp-DAB3*0602	Besp-DAB3*0701	Besp-DAB3*0801	Besp-DAB3*0901	Besp-DAB3*1001	Total number of verified DAB alleles
CA			x		x		x												■	■	■	■	■	■	■	■	■	■	■	■	3
CB		x					x												■	■	■	■	■	■	■	■	■	■	■	■	2
CC	x		x		x		x												■	■	■	■	■	■	■	■	■	■	■	■	4
CD			x		x		x												■	■	■	■	■	■	■	■	■	■	■	■	3
CE		x	x			x				x									x												5
CF			x				x												■	■	■	■	■	■	■	■	■	■	■	■	2
CG	x				x														■	■	■	■	■	■	■	■	■	■	■	■	2
CH	x				x														■	■	■	■	■	■	■	■	■	■	■	■	2
CI		x	x				x												■	■	■	■	■	■	■	■	■	■	■	■	3
CJ	x		x						x	x									x												5
CK		x	x							x	x								x	x											6
CL		x					x			x									x												4
CM	x		x		x	x													■	■	■	■	■	■	■	■	■	■	■	■	4
CN		x					x			x									x												4
CO			x		x														■	■	■	■	■	■	■	■	■	■	■	■	2
CP	x		x		x	x													■	■	■	■	■	■	■	■	■	■	■	■	4
CQ		x					x	x											■	■	■	■	■	■	■	■	■	■	■	■	3
CR	x				x			x											■	■	■	■	■	■	■	■	■	■	■	■	3
GA	x			x											x								x	x							5
GB			x										x						░	░	░	░	░	░	░	░	░	░	░	░	2
GC	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░										x			3
GD	x															x											x				3
GE			x													x											x				3
GF			x		░	░	░	░	░	░	░	░	░	░	░	░	░	░	■	■	■	■	■	■	■	■	■	■	■	■	1
GG			x	x	░	░	░	░	░	░	░	░	░	░	░	░	░	░											x		3
GH	x				░	░	░	░	░	░	░	░	░	░	░	░	░	░										x			2
GI	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░										x			3
GJ	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░	■	■	■	■	■	■	■	■	■	■	■	■	2
GK	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░				x									3
GL	x			x													x								x			x			5
GM			x														x								x			x			4
GN			x											x					■	■	■	■	■	■	■	■	■	■	■	■	2
GO			x		░	░	░	░	░	░	░	░	░	░	░	░	░	░											x		2
GP			x														x								x						3
GQ	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░										x			3
GR	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░							x			x			4
GS			x		░	░	░	░	░	░	░	░	░	░	░	░	░	░										x			2
GT			x		░	░	░	░	░	░	░	░	░	░	░	░	░	░	■	■	■	■	■	■	■	■	■	■	■	■	1
GU	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░										x			3
GV	x			x													x									x	x				5
GW	x			x														x										x		x	5
GX			x														x								x						3
GY	x			x								x									x										4
GZ	x			x	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	2
GAA			x						x										x												3
GAB			x		░	░	░	░	░	░	░	░	░	░	░	░	░	░	■	■	■	■	■	■	■	■	■	■	■	■	1
GAC	x								x										x												3
TA				x	░	░	░	░	░	░	░	░	░	░	░	░	░	░								x					2
TB	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░								x					1
TC			x	x	░	░	░	░	░	░	░	░	░	░	░	░	░	░								x					3
TD	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	0
TE	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	0
TF		x			░	░	░	░	░	░	░	░	░	░	░	░	░	░	x												2
TG	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	x												1
TH	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	0
TI	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░								x					1
TJ	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	0
TK	x		x		░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	2
TL	x		x		░	░	░	░	░	░	░	░	░	░	░	░	░	░	x								x				4

◂**Fig. 3** Summary of DABexon 2 alleles found in *Betta splendens* individuals. For the column "Individual", the first letter denotes the data source for obtaining the alleles: "C" represents cloning and sequencing data; "G" represents whole-genome sequencing data (see Table S1); "T" represents transcriptome sequencing data (Table S1). The second and third letters (if any) denote the individual ID. Boxes corresponding to the locus-specific alleles present in the individuals are denoted by "×" and highlighted in red (*DAB1*), blue (*DAB2*), and green (*DAB3*), respectively. Black boxes indicate the absence of *DAB3* in those individuals based on the cloning-sequencing and DoC results. Samples that did not meet genotyping criteria (see Materials and methods for details), and hence, no alleles were retrieved are indicated by grey boxes, but this does not implicate the absence of locus in those individuals

16 and 8 (Fig. S5, Table S6). Two DAB-containing regions were identified at chromosome 2, separated by around 8 Mb in the *B. splendens* genome (Fig. 4; Fig. S5). The first region (Fig. 4a) spanning 1.337 Mb contained the *DAB1* gene, whereas the second region (Fig. 4b) spanning 1.625 Mb contained the *DAB2* and *DAB3* genes. All DAB genes were in tight linkage with their corresponding α-chain-encoding DAA genes. A further 12 and 16 common syntenic genes were annotated in the first and second regions, respectively. Along with the known teleost MHC II regions (Yamaguchi and Dijkstra 2019), the two *B. splendens* genomic regions were neither linked to their class I-related genes nor did they share syntenic genes residing in the human MHC II region (Horton et al. 2004). Comparison of the two regions with their conspecifics confirmed the lack of defined MHC synteny among species in the order Anabantiformes. This is particularly obvious in region 2 (Fig. 4b), in which the genes were more contracted in the *B. splendens* genome (1625 kb), but the genes were scattered throughout different chromosomes with altered transcriptional orientation in other species, such as *A. testudineus* (8,058 Kb) and *O. niloticus* (30,345 Kb). Additionally, MHC II gene copy number varied across species within the homologous regions. Among the first region, MHC II genes were not found in *M. armatus* and *C. argus*, whereas for the *B. splendens*, *D. rerio*, *G. aculeatus*, and *O. niloticus*, the copy numbers ranged from two to five (Fig. 4a). Similarly, among the second homologous region, the MHC II gene copy number was zero in *D. rerio* and varied from two in *A. testudineus* to up to eight in *O. niloticus* (Fig. 4b).

Intra-specific copy number variation

Cloning and sequencing analysis identified *DAB1* and *DAB2* alleles in all 18 samples, but only 6 of them had their *DAB3* alleles amplified even with multiple pairs of primers (Fig. 3; Table S2). We thus compared the RCN of the *DAB3* gene in individuals with WGS data ($n = 29$) or cloning and sequencing data ($n = 17$). The qPCR result indicated all individuals contained 2 allelic copies of the *ACTB* gene (Fig. 5a). The total number of alleles from both of the *DAB2* and *DAB3* genes was 2 in *DAB3*-missing individuals ($n = 12$) but ranged from 2.5 to 3.8 in *DAB3*-containing individuals ($n = 5$: Fig. 5a). Among the five individuals that had the *DAB3* gene, the one with RCN of 3.8 (i.e., individual CK: Fig. 3) was known to be both *DAB2*- and *DAB3*-heterozygous. *DAB3*-containing individuals that had a total number of alleles of 3 were likely to be *DAB3* hemizygous, containing only one copy of the *DAB3* gene in their diploid genome.

The DoC analysis revealed a consistent CNV pattern with the qPCR results. Individuals consistently contained 2 alleles of *DAB1* and *ACTB* (Fig. 5b), signifying that they are single-copied genes. On the other hand, the number of *DAB3* allelic copies ranged from 0 to 2 (Fig. 5b). Heterozygous individuals (H_e, $n = 6$) contained 2 *DAB3* allelic copies; non-heterozygous individuals (non-H_e, $n = 17$) contain either 1 copy (i.e., hemizygous) or 2 copies (i.e., homozygous); and *DAB3*-missing individuals had zero copy ($n = 5$: Fig. 5b).

Loss of the genomic region containing the *DAA3* and *DAB3* genes

The DoC analysis indicated the deletion of a genomic region of ~ 20 kb in length, containing both the *DAA3* and *DAB3* genes (Fig. 6). *DAB3*-containing individuals had a high DoC in that region, which is comparable to the genome-wide level. However, *DAB3*-missing individuals had the DoC close to zero in the same region, except for the highly repetitive intergenic region (~ 8 kb) between *DAA3* and *DAB3*. BLAST search using this intergenic region as query showed that there were many highly similar sequences in other regions of the *B. splendens* genome; therefore, the high DoC of this intergenic region in *DAB3*-missing individuals was likely due to mapping of reads originated from other regions.

Recombination and selection patterns

GARD analysis indicated the presence of two recombination breakpoints in the β1 domain encoded by *DAB2* and *DAB3* (Fig. 2). The breakpoints are close to the end of exon 2, at positions 196 and 234 in *DAB2* and at positions 172 and 227 in *DAB3*. Recombination events were also inferred from the Neighbor-Net analysis, which shows the reticulated networks in *DAB2* and *DAB3* alleles (Fig. 7).

Signature of positive selection was detected at both *DAB2* and *DAB3* but not at *DAB1* (Table 1). The ratio of non-synonymous to synonymous substitutions (ω) was significantly higher at the PBS of *DAB2* ($p < 0.05$, $Z = 1.77$) and *DAB3* ($p < 0.05$, $Z = 1.74$), whereas their non-PBS did not show significant positive selection (Table 1). No selection

Table 1 Nucleotide diversity (π), rates of non-synonymous (d_N) and synonymous (d_S) substitutions (± standard error), and ratio of d_N to d_S (ω) for peptide-binding sites (PBS), non-PBS, and combined (PBS + non-PBS) at the three *DAB* loci in Siamese fighting fish *Betta splendens*. Regions with signs of positive selection ($\omega > 1$) were highlighted in bold

Region	Number of codons	π	d_N	d_S	ω
*DAB1 β*1					
PBS	24	0.009 ± 0.010	0.012 ± 0.012	0	0
Non-PBS	66	0.008 ± 0.005	0.003 ± 0.003	0.025 ± 0.018	0.120
Combined	90	0.009 ± 0.004	0.006 ± 0.004	0.019 ± 0.014	0.316
*DAB2 β*1					
PBS	24	0.391 ± 0.055	0.453 ± 0.084	0.176 ± 0.093	**2.574***
Non-PBS	67	0.106 ± 0.017	0.079 ± 0.018	0.205 ± 0.057	0.385
Combined	91	0.170 ± 0.019	0.162 ± 0.025	0.198 ± 0.048	0.818
*DAB3 β*1					
Putative PBS	24	0.297 ± 0.048	0.330 ± 0.082	0.199 ± 0.091	**1.658***
Non-PBS	66	0.065 ± 0.012	0.067 ± 0.017	0.059 ± 0.025	**1.136**
Combined	90	0.119 ± 0.015	0.127 ± 0.023	0.092 ± 0.026	**1.380**

*Statistical significance at $P < 0.05$ from *Z*-test

was observed in the entire β1 domain of *DAB1*. Further tests by FEL, MEME, and FUBAR identified site-specific positive selection on 7 codons of *DAB2* and 9 codons of *DAB3* (Fig. 2), the majority of which is known or predicted to encode PBS from previous MHC studies (Dijkstra et al. 2013). On the other hand, results from MEME and FUBAR highlighted several non-PBS codons undergoing purifying selection at *DAB2* (positions 13, 17–18, 74, and 77) and *DAB3* (position 70). Similarly, signatures of positive selection were also found in *DAA2* and *DAA3* but not in *DAA1* (Fig. S4).

Phylogenetic analyses

The phylogenetic tree shows the distinct and highly supported monophyletic clades for each of the three DAB loci (Fig. 8). *B. splendens* alleles within each clade did not form a monophyletic group but were instead interspersed with those of closely related species such as *B. siamorientalis* and *B. smaragdina* (Fig. 8). The groupings of *B. splendens* alleles for each clade were consistent with those from Neighbor-Net results (Fig. 7). Specifically, three sub-groups of the *DAB2* clade and the major sub-group of the *DAB3* clade (containing *Besp-DAB3*0301–0701*) in the phylogenetic tree were grouped similarly in the phylogenetic network. The observed reticulated phylogenetic network among *Besp-DAB3*0101–0202* and *Besp-DAB3*0801–1001* (Fig. 7) was also consistent with the multifurcating pattern in the *DAB3* clade of the phylogenetic tree (Fig. 8).

Discussion

MHC is known to exhibit between- and within-species CNV, but the latter was rarely examined in detail. As the first MHC study on labyrinth fishes, we characterized this multigene family from a locus-specific approach and revealed intra-specific CNV of MHC class II genes in *B. splendens*, due to the deletion of a ~ 20-kb-long genomic region in some haplotypes. In addition, we found that each MHC locus was under different modes of selection. This information would have been masked by the conventional approach that co-amplifies alleles from multiple genes.

Intra-specific CNV of MHC in *B. splendens*

Due to the lack of locus-specific information in numerous non-model organisms, a common approach of estimating the number of MHC loci was using degenerate primers for PCR and dividing the number of unique alleles per individual by two (Minias et al. 2019, 2021; Wegner et al. 2003). However, this method may underestimate the true copy number and overlook CNV, because it lacks zygosity information on each locus and ignores the possible existence of hemizygosity, in which the MHC gene is located at only one of the homologous chromosomes (Dijkstra et al. 2006; Dirscherl and Yoder 2015; Lighten et al. 2014a). Many variant callers have been developed to estimate genome-wide CNV from high-throughput sequencing data (Zhao et al. 2013), but their performance on analyzing MHC genes is hampered by the poor mapping of highly polymorphic sequences. As a result, the fitness consequences of CNV and hemizygosity were seldom investigated. We demonstrate here that by combining both bioinformatic analysis and wet lab experiment, there is intra-specific CNV of *DAB3* among *B. splendens* individuals. This gene was tightly linked to *DAB2* (separated by ~ 30 kb), which was found in all studied individuals. Specifically, the CNV was due to a segmental deletion of the genomic region containing both the *DAA3* and *DAB3* genes (Fig. 6) in some haplotypes. Furthermore, genotyping results suggest possible

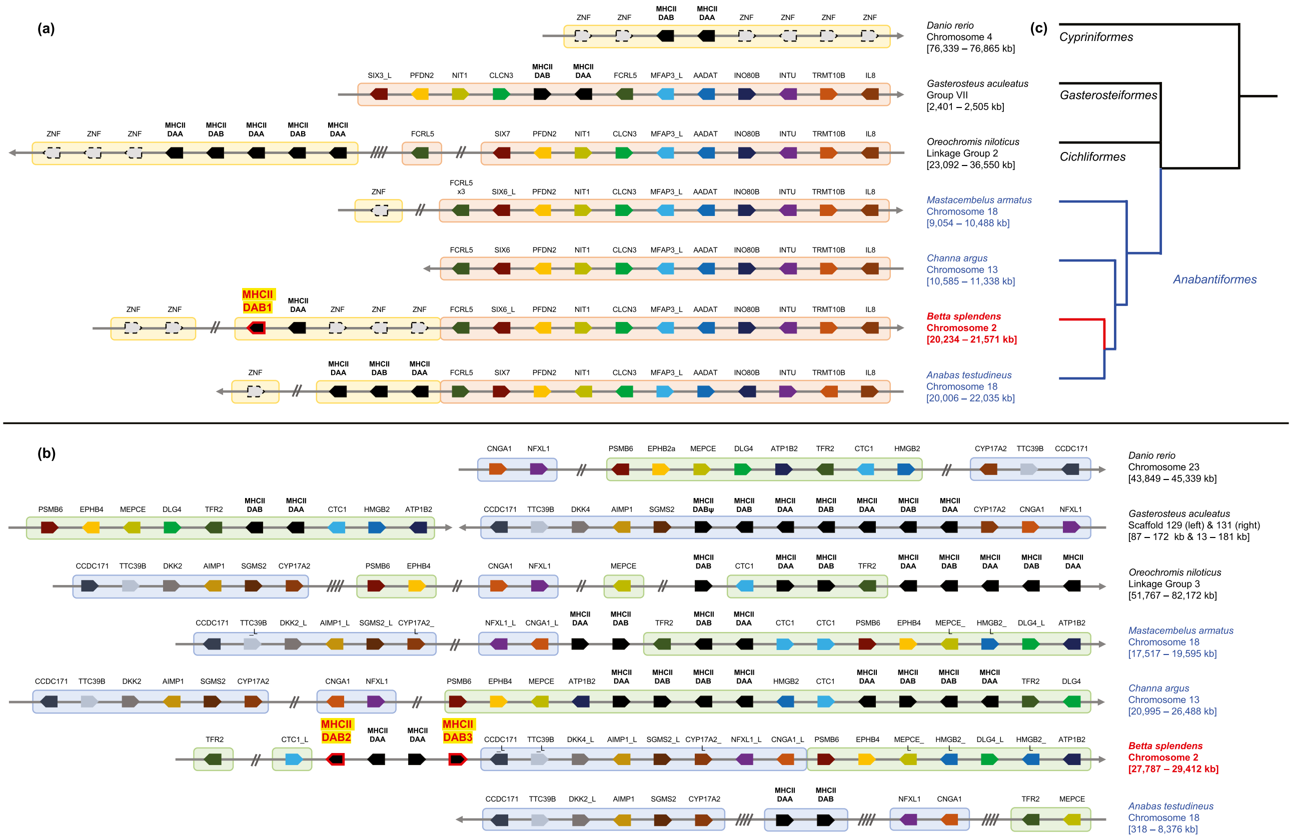
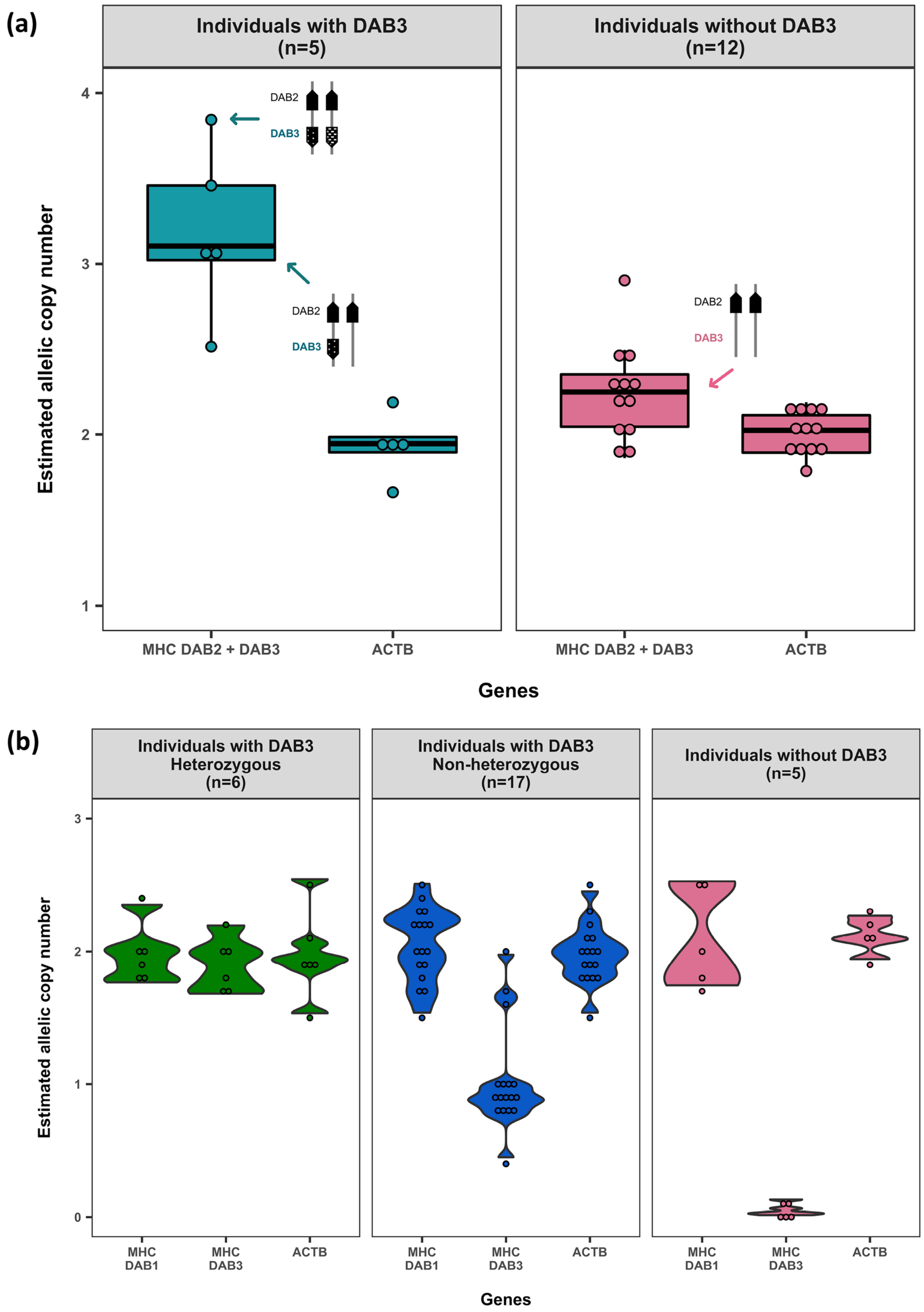

(a)
ZNF ZNF MHCII DAB MHCII DAA ZNF ZNF ZNF ZNF
Danio rerio
Chromosome 4
[76,339 – 76,865 kb]
SIX3_L PFDN2 NIT1 CLCN3 MHCII DAB MHCII DAA FCRL5 MFAP3_L AADAT INO80B INTU TRMT10B IL8
Gasterosteus aculeatus
Group VII
[2,401 – 2,505 kb]
ZNF ZNF ZNF MHCII DAA MHCII DAB MHCII DAA MHCII DAB MHCII DAA FCRL5 SIX7 PFDN2 NIT1 CLCN3 MFAP3_L AADAT INO80B INTU TRMT10B IL8
Oreochromis niloticus
Linkage Group 2
[23,092 – 36,550 kb]
ZNF FCRL5 x3 SIX6_L PFDN2 NIT1 CLCN3 MFAP3_L AADAT INO80B INTU TRMT10B IL8
Mastacembelus armatus
Chromosome 18
[9,054 – 10,488 kb]
FCRL5 SIX6 PFDN2 NIT1 CLCN3 MFAP3_L AADAT INO80B INTU TRMT10B IL8
Channa argus
Chromosome 13
[10,585 – 11,338 kb]
ZNF ZNF MHCII DAB1 MHCII DAA ZNF ZNF ZNF FCRL5 SIX6_L PFDN2 NIT1 CLCN3 MFAP3_L AADAT INO80B INTU TRMT10B IL8
Betta splendens
Chromosome 2
[20,234 – 21,571 kb]
ZNF MHCII DAA MHCII DAB MHCII DAA FCRL5 SIX7 PFDN2 NIT1 CLCN3 MFAP3_L AADAT INO80B INTU TRMT10B IL8
Anabas testudineus
Chromosome 18
[20,006 – 22,035 kb]
(c)
Cypriniformes
Gasterosteiformes
Cichliformes
Anabantiformes
(b)
CNGA1 NFXL1 PSMB6 EPHB2a MEPCE DLG4 ATP1B2 TFR2 CTC1 HMGB2 CYP17A2 TTC39B CCDC171
Danio rerio
Chromosome 23
[43,849 – 45,339 kb]
PSMB6 EPHB4 MEPCE DLG4 TFR2 MHCII DAB MHCII DAA CTC1 HMGB2 ATP1B2 CCDC171 TTC39B DKK4 AIMP1 SGMS2 MHCII DABψ MHCII DAB MHCII DAA MHCII DAB MHCII DAA MHCII DAB MHCII DAA CYP17A2 CNGA1 NFXL1
Gasterosteus aculeatus
Scaffold 129 (left) & 131 (right)
[87 – 172 kb & 13 – 181 kb]
CCDC171 TTC39B DKK2 AIMP1 SGMS2 CYP17A2 PSMB6 EPHB4 CNGA1 NFXL1 MEPCE MHCII DAB CTC1 MHCII DAA MHCII DAB TFR2 MHCII DAA MHCII DAB MHCII DAA MHCII DAB MHCII DAA
Oreochromis niloticus
Linkage Group 3
[51,767 – 82,172 kb]
CCDC171 TTC39B_L DKK2_L AIMP1_L SGMS2_L CYP17A2_L NFXL1_L CNGA1_L MHCII DAA MHCII DAB TFR2 MHCII DAB MHCII DAA CTC1 CTC1 PSMB6 EPHB4 MEPCE_L HMGB2_L DLG4_L ATP1B2
Mastacembelus armatus
Chromosome 18
[17,517 – 19,595 kb]
CCDC171 TTC39B DKK2 AIMP1 SGMS2 CYP17A2 CNGA1 NFXL1 PSMB6 EPHB4 MEPCE ATP1B2 MHCII DAA MHCII DAB MHCII DAB MHCII DAA HMGB2 CTC1 MHCII DAA MHCII DAB MHCII DAB MHCII DAA TFR2 DLG4
Channa argus
Chromosome 13
[20,995 – 26,488 kb]
TFR2 CTC1_L MHCII DAB2 MHCII DAA MHCII DAA MHCII DAB3 CCDC171_L TTC39B_L DKK4_L AIMP1_L SGMS2_L CYP17A2_L NFXL1_L CNGA1_L PSMB6 EPHB4 MEPCE_L HMGB2_L DLG4_L HMGB2_L ATP1B2
Betta splendens
Chromosome 2
[27,787 – 29,412 kb]
CCDC171 TTC39B DKK2_L AIMP1 SGMS2 CYP17A2 MHCII DAA MHCII DAB NFXL1 CNGA1 TFR2 MEPCE
Anabas testudineus
Chromosome 18
[318 – 8,376 kb]

Fig. 4 Schematic diagram showing the MHC IIgenomicregions containing **(a)** *DAB1* and **(b)** *DAB2-3* in *Betta splendens* (highlighted in red) and the corresponding regions in anabantoids — *Anabas testudineus*, *Channa argus*, and *Mastacembelus armatus* (highlighted in blue) and other species (*Oreochromis niloticus*, *Gasterosteus aculeatus*, and *Danio rerio*). The range of each region is indicated in kilobase (kb). Syntenic genes in close physical linkage predicted to be translocated as a unit are grouped inside colored blocks. Black blocks refer to the classical MHC II genes. Other colored blocks in each region refer to the teleost-specific syntenic genes flanking DAB loci, with the gene aliases depicted above the blocks. The symbol "ψ" denotes pseudogenes. Different members of the same gene family are assigned to the same color. The black blocks corresponding to the three DAB loci of *B. splendens* characterized in this study are enclosed by red grids with their names highlighted in yellow. Pointed gene blocks indicate the transcriptional orientation. Only syntenic genes shared among species are included in this figure and several species-specific genes were detailed in the supplementary information of Dijkstra et al (2013). Dotted light grey blocks refer to the zinc finger protein family (*ZNF*) genes, which are extensively distributed across the genome and hence may not be an ideal indication of region synteny (Dijkstra et al. 2013). Predicted genes of low sequence identity or denoted as "low-quality proteins" in the annotation are appended "_L" (-like) at the end of the aliases. Some syntenies drawn in reverse order are indicated by the grey arrows for better illustration. All regions are not drawn to scale, with double slash and quadruple slash inserted for regions where two syntenic genes are > 500 kb but < 1,000 kb, and > 1,000 kb apart from each other respectively. **(c)** A phylogram displaying the phylogenetic relationship among the selected species

linkage of *DAB2* and *DAB3* (Fig. 3). *Besp-DAB2*0101*, *0201*, *0301*, *0401* and *0901* may be associated with a haplotype lacking *DAB3*. The *DAB2*0101–0401* and *0901* alleles were clustered together as clade 2b in the phylogeny (Figs. 7 and 8). This indicates a single-deletion event of the *DAB3* region in the ancestral haplotype of these alleles (Fig. 8).

Along with other recently duplicated genes, MHC has been known to contribute to a significant portion of CNV in populations (Chain et al. 2014). When a gene is newly duplicated in a genome, it may persist in an evolutionary stage of CNV polymorphism in populations for a few million years until being fixed or lost (Chain et al. 2014; Lynch and Conery 2000). Because new paralogs are assumed to be functionally redundant at the time of origin, they must have acquired novel functions or retained all or parts of the ancestral gene function, or otherwise be pseudogenized by degenerative mutations (Lynch and Conery 2000; Zhang 2003). In this context, *DAB3* may have been playing a role complement to or distinct from the other two DAB genes in pathogen resistance (Eizaguirre et al. 2009). Phylogenetic analyses show that *DAB3* alleles form a distinct monophyletic clade, indicating a possible functional divergence from *DAB2* alleles. This locus-specific information obtained is crucial to MHC-based hypothesis testing. For example, Dearborn et al. (2016) used a locus-specific approach to study the MHC IIb genes in the Leach's storm petrel (*Oceanodroma*


(a)
Individuals with DAB3
(n=5)
Individuals without DAB3
(n=12)
Estimated allelic copy number
4
3
2
1
DAB2
DAB3
DAB2
DAB3
DAB2
DAB3
MHC DAB2 + DAB3
ACTB
MHC DAB2 + DAB3
ACTB
Genes
(b)
Individuals with DAB3
Heterozygous
(n=6)
Individuals with DAB3
Non-heterozygous
(n=17)
Individuals without DAB3
(n=5)
Estimated allelic copy number
3
2
1
0
MHC DAB1
MHC DAB3
ACTB
MHC DAB1
MHC DAB3
ACTB
MHC DAB1
MHC DAB3
ACTB
Genes

◀**Fig. 5** **(a)** Combined boxplot anddotplot of the total allelic copy numbers of *DAB2* and *DAB3* exon 3, and the *ACTB* gene based on the qPCR results, normalized by *RPL17*. Three schematic diagrams denote the *DAB2* and *DAB3* loci of chromosome 2 in respect to *DAB3*-heterozygous or -homozygous, *DAB3*-hemizygous, and *DAB3*-missing individuals. **(b)** Combined violin plot and dotplot of the *DAB1*, *DAB3*, and *ACTB* allelic copy numbers estimated using whole genome sequencing data for *DAB3*-heterozygous, *DAB3*-non-heterozygous (homozygous or hemizygous), and *DAB3*-missing individuals

leucorhoa) and showed that if genes became functionally divergent after gene duplication, the offspring will always inherit diverse multilocus genotypes, and MHC diversity will be maintained even under random mating. However, if alleles of duplicated genes did not diverge from each other, MHC-disassortative mating is required to increase the MHC diversity in offspring.

Notably, some fishes contain only a single *DAB3* copy in their diploid genomes. Hemizygosity was rarely reported from previous MHC research. Lighten et al. (2014b) estimated a range of 1–5 class IIb loci from seven populations of guppies (*Poecilia reticulata*) and indirectly predicted the presence of hemizygous loci from allelic copy number estimation. These hemizygous guppies were presumably hybrids originated from mating between parents from two locally adapted populations and may possess MHC haplotypes with different number of loci from two diverging gene pools (Eizaguirre et al. 2009). In theory, individuals with super-optimal MHC diversity confer enhanced antigen-binding breadth, but also come with a detrimental reduction of T cell repertoire and hence a lower fitness (Nowak et al. 1992; Woelfing et al. 2009). Alternatively, hemizygous individuals could benefit from recognizing the target antigen(s) while avoiding superfluous expression of identical MHC alleles (i.e., in homozygous individuals), given that doubling the number of alleles could lead to more than twofold elevation in the expression level (Loehlin and Carroll 2016). The optimal number of MHC genes in an individual likely depends on the selective pressure from pathogens. An approach to test the fitness consequence of CNV is to compare the pathogen load between individuals with different MHC copy numbers, i.e., *DAB3*-heterozygous, *DAB3*-homozygous, *DAB3*-hemizygous, and *DAB3*-missing individuals in our *B. splendens* case. On the

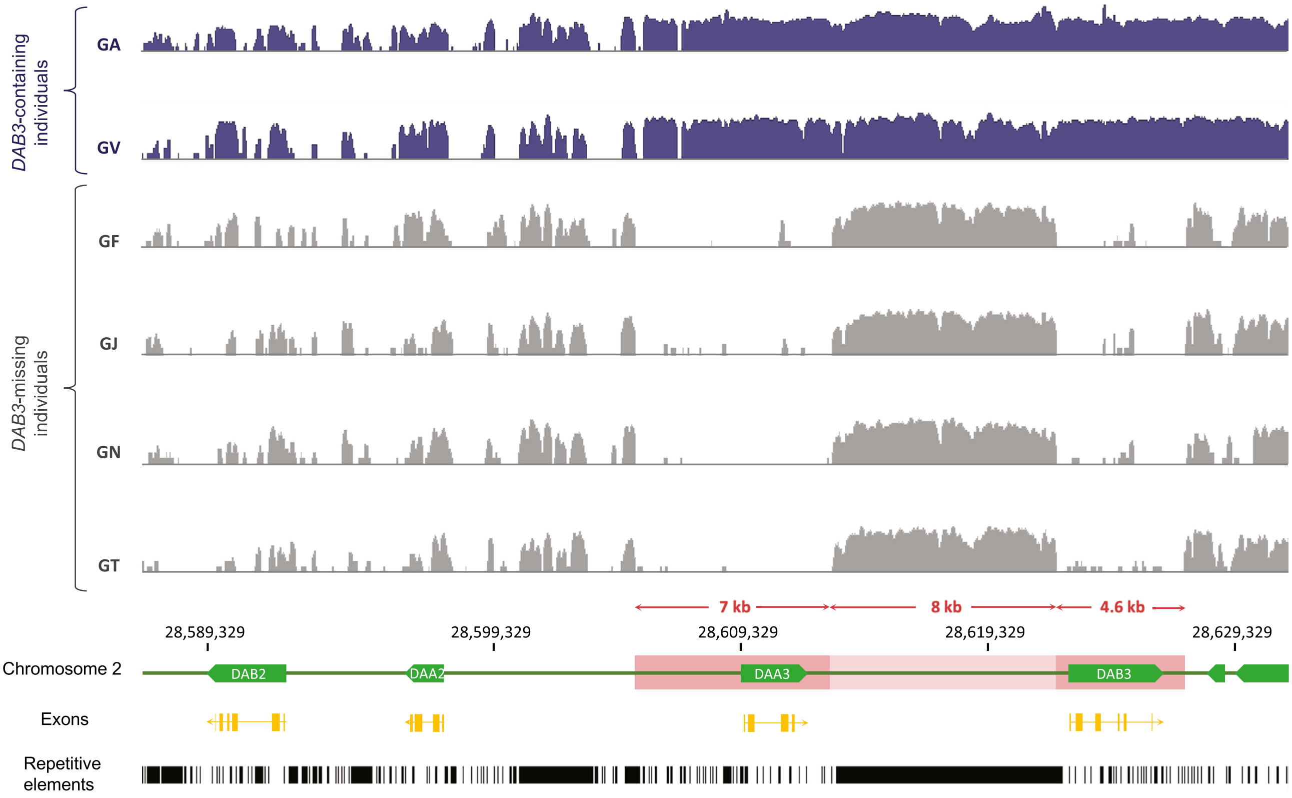


Fig. 6 Sequencing depth-of-coverage (DoC) across the genomic region spanning from *DAB2* to *DAB3*. DoC of *DAB3*-containing individuals (GA and GV) and *DAB3*-missing individuals (GF, GJ, GN and GT) are shown in purple and grey, respectively. Green and yellow bars denote annotation of genes and exons, respectively, with arrows indicating transcriptional orientation. The red bar highlights the potential genomic region (~ 20 kb) that was lost in haplotypes without the *DAB3* gene.Thelight red bar (~ 8 kb) denotes the highly repetitive intergenic region between the *DAA3* and *DAB3* genes, which likely led to wrong mapping. Black bars denote repeat-masked regions in the genome

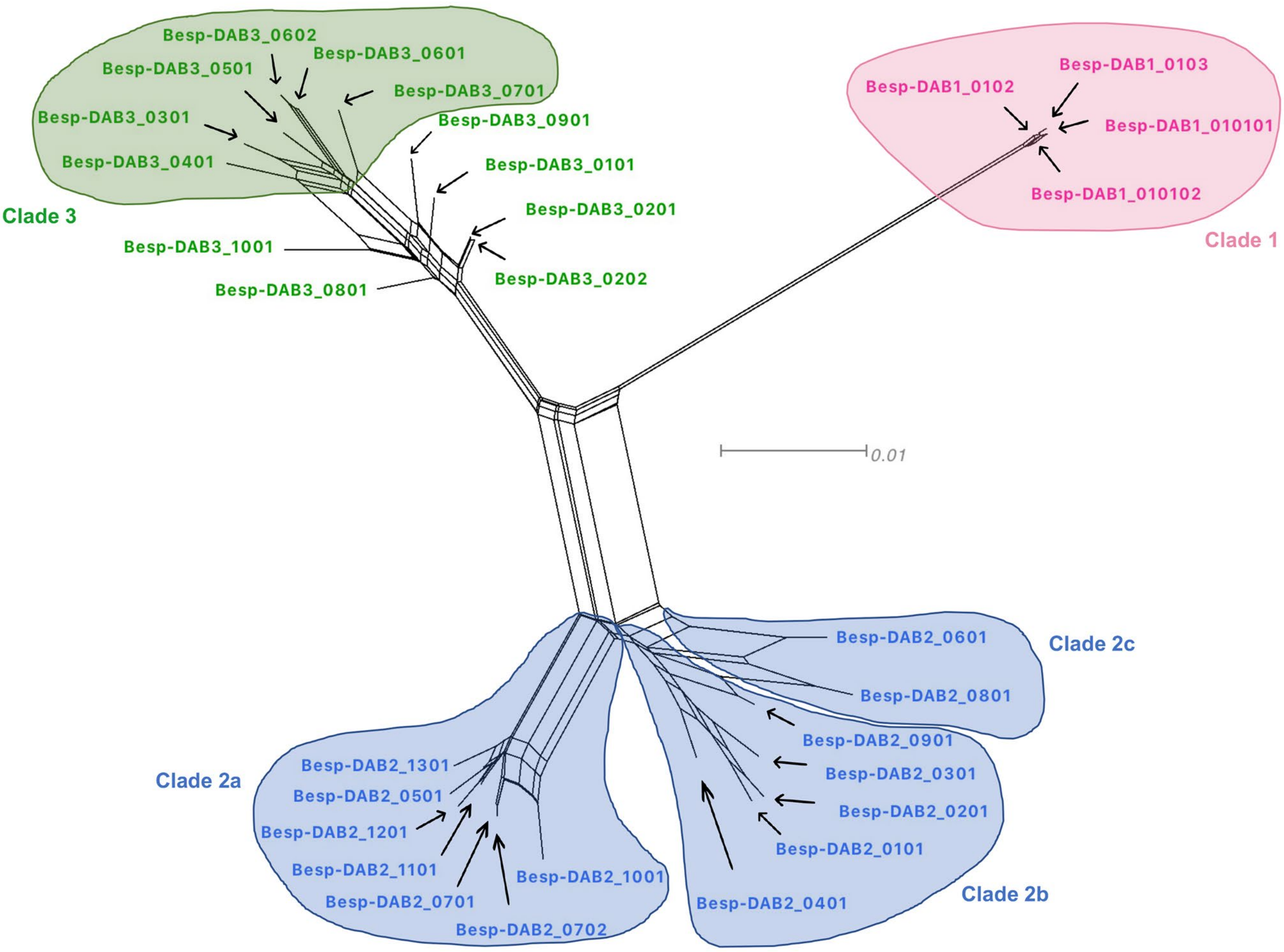


Fig. 7 A neighbor-net phylogenetic network of DAB exon 2alleles from *Betta splendens*. Labels of *B. splendens* allelesbelonging to *DAB1*, *DAB2*, and *DAB3* are colored in red, blueand green respectively. Within each clade, well-defined clusters shown in thephylogenetic tree (Fig. 8) are enclosed and shaded. The scale bar at the centershows the branch length in substitutions per site

other hand, the number of genomic copies could be disproportional to that of expressed copies (O'Connor and Westerdahl 2021); therefore, comparing the expression level of *DAB3* genes between individuals of different genotypes upon pathogen exposure will disentangle the relationship between genomic and expressed MHC diversity. Both findings will shed light on the evolution of MHC CNV within and between species.

CNV among anabantoids

Our comparative genomic analysis of anabantoids revealed frequent translocation and inversion of the MHC II genes in different chromosomes throughout the anabantoid diversification. This coincides with the long-standing idea that teleost MHC is diverse but not complex (Shand and Dixon 2001). In bony fishes, the most notable differences of their MHC genes from mammalian ones are their non-linkage of MHC class I and II regions (Sato et al. 2000) and the wide dispersion of genes associated with MHC-associated pathways (Reusch et al. 2004). It had therefore been suggested to designate the regions in teleost fishes as major histocompatibility (MH) instead of MHC (Stet et al. 2003). Since the divergence of ancestral teleost from *Lepisosteus oculatus* (spotted gar) and cartilaginous fishes, which, along with tetrapods, have their class I and II loci closely clustered in a single chromosome (Braasch et al. 2016), the primordial MHC class II has been translocated out of a typical MHC region into different linkage groups in the early stages of ray-finned fish evolution (Yamaguchi and Dijkstra 2019). The absence of linkage thus provides rooms for frequent inter- and intra-locus recombination (Stet et al. 2003), giving rise to a variety of syntenic organization during teleost diversification.

Furthermore, CNV plausibly exists among *Betta* species. Other *Betta* species show varied number of unique alleles and uneven sequencing coverage at the three *B. splendens* DAB loci based on the mapping results. For example, *B.*

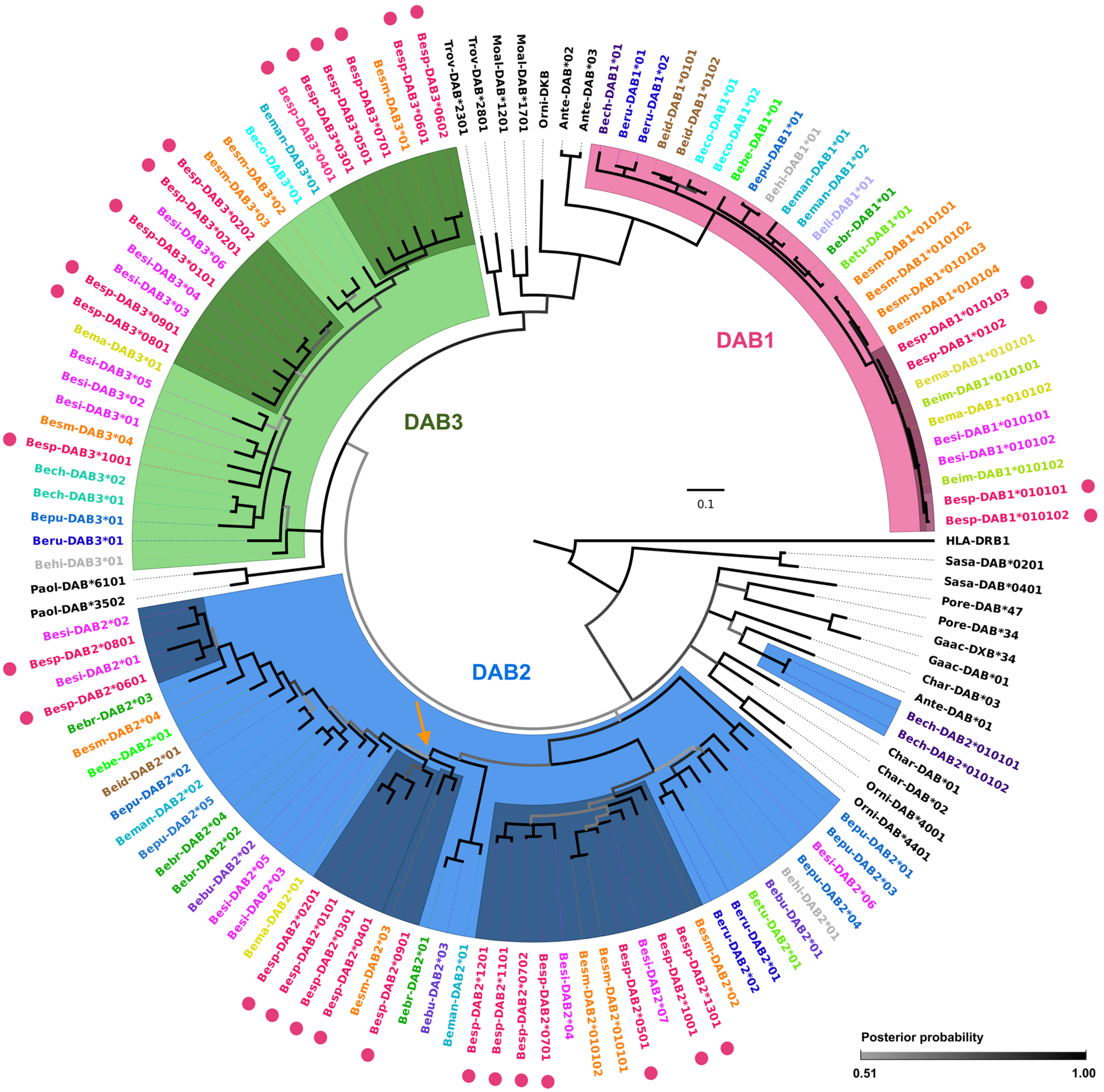


Fig. 8 A Bayesian phylogenetic tree *DAB* exon 2 alleles from *Betta* species, anabantoids, and other teleost species. A human *HLA-DRB* allele is used as the outgroup. Alleles of each *Betta* species are indicated by same colored labels, whereas labels of other species are black. Pink dots represent allele sequences of *B. splendens*.Branches belonging to the *DAB1*, *DAB2*, and *DAB3* clusters are highlighted in red, blue, and green, respectively. Within each cluster, the darkened color represents major lineages containing alleles of *B. splendens*. The scale bar at the center shows the branch length in substitutions per site.The branch color intensities correspond to the posterior probability, which are predominantly higher than 0.90 in most of the branches. The orange arrow indicates a potential single-deletion event of the *DAB3* region in the ancestral haplotype of *Besp-DAB2*0101*, *0201*, *0301*, *0401*,and *0901*

burdigala had a 7- to 14-fold higher coverage at *DAB2* compared with *DAB1* and *DAB3*, whereas reads of MHC from *B. rubra* were predominantly mapped to *DAB1*. The observed uneven coverage is likely due to the presence of additional MHC gene copies. It is worth noting that our findings may only reflect CNV among *B. splendens* individuals that have been subjected to long history of domestication (Zhang et al. 2021). Further studies on wild-type individuals and other *Betta* species are needed to unravel the inter-specific CNV pattern in this genus.

Positive selection at *DAB2* and *DAB3* but not *DAB1*

Signature of positive selection at the PBS of *DAB2* and *DAB3* and the trans-species polymorphism shown by the phylogenetic analysis suggests that balancing selection maintains the polymorphism of these two genes over long periods of time across speciation events (Klein et al. 1998). In contrast, *DAB1* is almost monomorphic, which is uncommon for classical MHC genes (Yamaguchi and Dijkstra 2019). *DAB1* in *B. splendens* contains key structural residues and shares high sequence similarity with *DAB2* and *DAB3* that are both under the same DA lineage in teleosts (Dijkstra et al. 2013). However, the identified exon 2 sequences of *DAB1* are almost identical, and there was no signature of positive selection. This pattern was observed in closely related *B. smaragdina*, *B. mahachiensis*, *B. siamorientalis*, and *B. imbellis*, which had their *DAB1* sequences identical to the *Besp-DAB1*0102* allele (Fig. S3). In the phylogenetic tree, the *DAB1* clade had a much shorter branch length than *DAB2* and *DAB3* clades (Fig. 8), indicating a low substitution rate and small sequence difference. A potential homolog of *Besp-DAB1* was found in *G. aculeatus* (stickleback) and *A. testudineus* (climbing perch) but not in other two anabantoids (Fig. 4a). In teleostei, while some species have a higher sequence diversity in either the DAB or DAA genes (Gómez et al. 2010), other species have similar diversity at both genes (Reusch et al. 2004). The corresponding DAA and DAB genes in *B. splendens* share similar patterns of polymorphism, e.g., both *DAA1* and *DAB1* are almost monomorphic without signs of positive selection, and both *DAA2* and *DAB2* are the most polymorphic one among DAA and DAB genes, respectively (Fig. S4).

MHC genes without extensive polymorphism or signs of positive selection were occasionally reported in other studies (e.g., Aguilar et al. 2006, Bollmer et al. 2010, Jeon et al. 2019, Llaurens et al. 2012 and Zagalska-Neubauer et al. 2010). This is likely to be masked in many studies that co-amplified multi-locus alleles and performed strict clustering and elimination of so called "redundant" alleles (e.g., Gerdol et al. 2019). Since individuals with reduced MHC diversity could be more vulnerable to pathogen infection (Radwan et al. 2010), the low polymorphism of *DAB1* found in multiple species is pointing to a functional difference of *DAB1* to *DAB2* and *DAB3*. *DAB1* appeared to have a lower expression level compared to the more polymorphic *DAB2* and *DAB3* (Table S4), which indicates a differentiated functional role of this gene from the others. One possibility is that *DAB1* may act as a promiscuous binder that recognizes a broad range of antigens from common pathogens, opposite to fastidious binders that are specialized for one or a few peptide motifs (Kaufman 2018). Alternatively, *DAB1* could be maintained through genetic hitchhiking given their tight linkage with the α-chain coding *DAA* gene (Radwan et al. 2020). Llaurens et al. (2012), who found similar monomorphic pattern in a cluster of DAB sequences from *P. reticulata*, proposed that positive selection may act on the PBR of DAA while the neighboring DAB "hitchhikes" and increases in frequency along with the selective sweep. But this mechanism might not apply to *B. splendens* as we found no evidence of positive selection in *DAA1* (Fig. S4).

Conclusion

We investigated three classical MHC IIb genes in *B. splendens* using a locus-specific approach and uncovered copy number variation of MHC genes in this species. Specifically, different individuals may carry different copy numbers of the *DAB3* gene — some individuals with two copies (heterozygous or homozygous), some have one (hemizygous), and some do not possess this gene. We demonstrated that there was a potentially single deletion event of the ~ 20-kb-long genomic region that had led to the lack of *DAA3* and *DAB3* in descendant haplotypes. We also investigated the evolutionary history of the three DAB genes in anabantoids and 16 *Betta* species, revealing considerable differences in the mode of selection and sequence polymorphism specific to each locus. Their corresponding DAA genes in *B. splendens* share similar pattern of polymorphism and selection. These results highlight the importance of investigating locus-specific characteristics and intra-specific CNV, which can shape pathogen resistance and subsequently lead to an effect on the fitness consequences of individuals and population, ultimately affecting the evolutionary trajectory of this multigene family.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00251-022-01255-8.

Acknowledgements The computations were performed using research computing facilities offered by the Information Technology Services, the University of Hong Kong.

Author contribution S.Y.W.S. and A.T.C.W. designed research; A.T.C.W., D.K.L., E.S.K.P., D.T.C.C. and S.Y.W.S. performed research; A.T.C.W. analyzed data; A.T.C.W. wrote the paper and all authors contributed to revised versions.

Funding This research was supported by the start-up fund (the University of Hong Kong) granted to SYWS.

Data availability Information of raw read sequences (including their NCBI Sequence Read Archive accession numbers) used in this study were detailed in the supplementary file 1. The DNA sequences of *Betta splendens* exon 2 alleles were assigned to GenBank accession numbers: MW863583 – MW863612.

Code availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

Aguilar A, Edwards SV, Smith TB et al (2006) Patterns of variation in MHC Class II β loci of the Little Greenbul (*Andropadus virens*) with comments on MHC evolution in birds. J. Hered 97:133-142. https://doi.org/10.1093/jhered/esj013

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164*:*1229–1236. https://doi.org/10.1093/genetics/164.3.1229

Babik W (2010) Methods for MHC genotyping in non-model vertebrates. Mol Ecol Resour 10:237–251. https://doi.org/10.1111/j.1755-0998.2009.02788.x

Bentkowski P, Radwan J (2019) Evolution of major histocompatibility complex gene copy number. PLoS Comput Biol 15:e1007015. https://doi.org/10.1371/journal.pcbi.1007015

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16:363–377. https://doi.org/10.1046/j.1420-9101.2003.00531.x

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bollmer JL, Dunn PO, Whittingham LA et al (2010) Extensive MHC Class II B gene duplication in a passerine, the common yellowthroat (*Geothlypis trichas*). J Hered 101:448-460. https://doi.org/10.1093/jhered/esq018

Braasch I, Gehrke AR, Smith JJ et al (2016) The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet 48:427–437. https://doi.org/10.1038/ng.3526

Brown JH, Jardetzky TS, Gorga JC et al (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364:33–39. https://doi.org/10.1038/364033a0

Burri R, Salamin N, Studer RA et al (2010) Adaptive divergence of ancient gene duplicates in the avian MHC class II β. Mol Biol Evol 27:2360–2374. https://doi.org/10.1093/molbev/msq120

Chain FJJ, Feulner PGD, Panchal M et al (2014) Extensive copy-number variation of young genes across stickleback populations. PLoS Genet 10:e1004830. https://doi.org/10.1371/journal.pgen.1004830

Cheng Y, Stuart A, Morris K et al (2012) Antigen-presenting genes and genomic copy number variations in the Tasmanian devil MHC. BMC Genomics 13:87. https://doi.org/10.1186/1471-2164-13-87

Clotfelter ED, Curren LJ, Murphy CE (2006) Mate choice and spawning success in the fighting fish *Betta splendens*: the importance of body size, display behavior and nest size. Ethology 112:1170-1178. https://doi.org/10.1111/j.1439-0310.2006.01281.x

Cram RA, Lawrence JM, Dzieweczynski TL (2019) Mating under the influence: male Siamese fighting fish prefer EE2-exposed females. Ecotoxicology 28:201–211. https://doi.org/10.1007/s10646-018-02012-y

Darriba D, Taboada GL, Doallo R et al (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772–772. https://doi.org/10.1038/nmeth.2109

Dearborn DC, Gager AB, McArthur AG et al (2016) Gene duplication and divergence produce divergent MHC genotypes without disassortative mating. Mol Ecol 25:4355–4367. https://doi.org/10.1111/mec.13747

Delport W, Poon AFY, Frost SDW et al (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455–2457. https://doi.org/10.1093/bioinformatics/btq429

Dijkstra JM, Grimholt U (2018) Major histocompatibility complex (MHC) fragment numbers alone - in Atlantic cod and in general - do not represent functional variability. F1000Res 7:963. https://doi.org/10.12688/f1000research.15386.2

Dijkstra JM, Grimholt U, Leong J et al (2013) Comprehensive analysis of MHC class II genes in teleost fish genomes reveals dispensability of the peptide-loading DM system in a large part of vertebrates. BMC Evol Biol 13:260. https://doi.org/10.1186/1471-2148-13-260

Dijkstra JM, Kiryu I, Yoshiura Y et al (2006) Polymorphism of two very similar MHC class Ib loci in rainbow trout (*Oncorhynchus mykiss*). Immunogenetics 58:152–167. https://doi.org/10.1007/s00251-006-0086-5

Dirscherl H, Yoder JA (2015) A nonclassical MHC class I U lineage locus in zebrafish with a null haplotypic variant. Immunogenetics 67:501–513. https://doi.org/10.1007/s00251-015-0862-1

Dobin A, Davis CA, Schlesinger F et al (2012) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635

Dubin A, Jørgensen TE, Moum T et al (2019) Complete loss of the MHC II pathway in an anglerfish. Lophius Piscatorius Biol Lett 15:20190594. https://doi.org/10.1098/rsbl.2019.0594

Dzieweczynski TL, Russell AM, Forrette LM et al (2014) Male behavioral type affects female preference in Siamese fighting fish. Behav Ecol 25:136–141. https://doi.org/10.1093/beheco/art095

Eizaguirre C, Lenz TL, Sommerfeld RD et al (2011) Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three-spined stickleback ecotypes. Evol Ecol 25:605–622. https://doi.org/10.1007/s10682-010-9424-z

Eizaguirre C, Lenz TL, Traulsen A et al (2009) Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. Ecol Lett 12:5–12. https://doi.org/10.1111/j.1461-0248.2008.01247.x

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042. https://doi.org/10.1111/eva.12178

Fremont DH, Monnaie D, Nelson CA et al (1998) Crystal structure of I-A^{k} in complex with a dominant epitope of lysozyme. Immunity 8:305–317. https://doi.org/10.1016/S1074-7613(00)80536-1

Gamazon ER, Stranger BE (2015) The impact of human copy number variation on gene expression. Brief Funct Genom 14:352–357. https://doi.org/10.1093/bfgp/elv017

Gerdol M, Lucente D, Buonocore F et al (2019) Molecular and structural characterization of MHC class II β genes reveals high diversity in the cold-adapted icefish *Chionodraco hamatus*. Sci Rep 9:5523. https://doi.org/10.1038/s41598-019-42003-5

Gómez D, Conejeros P, Marshall SH et al (2010) MHC evolution in three salmonid species: a comparison between class II alpha and beta genes. Immunogenetics 62:531–542. https://doi.org/10.1007/s00251-010-0456-x

Grimholt U (2016) MHC and Evolution in Teleosts Biology 5:6. https://doi.org/10.3390/biology5010006

He K, Minias P, Dunn PO (2021) Long-read genome assemblies reveal extraordinary variation in the number and structure of MHC loci in birds. Genome Biol. Evol. 13:evaa270. https://doi.org/10.1093/gbe/evaa270

Hoover B, Alcaide M, Jennings S et al (2018) Ecology can inform genetics: disassortative mating contributes to MHC polymorphism in Leach's storm-petrels (*Oceanodroma leucorhoa*). Mol Ecol 27:3371–3385. https://doi.org/10.1111/mec.14801

Horton R, Wilming L, Rand V et al (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889–899. https://doi.org/10.1038/nrg1489

Huson DH, Bryant D (2005) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267. https://doi.org/10.1093/molbev/msj030

Jeon H-B, Won H, Suk HY (2019) Polymorphism of MHC class IIB in an acheilognathid species, *Rhodeus sinesis* shaped by historical selection and recombination. BMC Genet. 20:74. https://doi.org/10.1186/s12863-019-0775-3

Kaufman J (2018) Generalists and specialists: a new view of how MHC class I molecules fight infectious pathogens. Trends Immunol 39:367–379. https://doi.org/10.1016/j.it.2018.01.001

Klein J, Bontrop RE, Dawkins RL et al (1990) Nomenclature for the major histocompatibility complexes of different species: a proposal. Immunogenetics 31:217–219. https://doi.org/10.1007/BF00204890

Klein J, Sato A, Nagl S et al (1998) Molecular trans-species polymorphism. Annu Rev Ecol Syst 29:1–21. https://doi.org/10.1146/annurev.ecolsys.29.1.1

Kosakovsky Pond SL, Posada D, Gravenor MB et al (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098. https://doi.org/10.1093/bioinformatics/btl474

Kumar S, Stecher G, Li M et al (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li W, Sun W, Hu J et al (2014) Molecular characterization, polymorphism and expression analysis of swamp eel major histocompatibility complex class II B gene, after infection by *Aeromonas Hydrophilia*. J Anim Plant Sci 24:481–491

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

Lighten J, van Oosterhout C, Paterson IG et al (2014b) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). Mol Ecol Resour 14:753–767. https://doi.org/10.1111/1755-0998.12225

Liu H-Y, Xue F, Gong J et al (2017) Limited polymorphism of the functional MHC class II B gene in the black-spotted frog (*Pelophylax nigromaculatus*) identified by locus-specific genotyping. Ecol Evol 7:9860–9868. https://doi.org/10.1002/ece3.3408

Llaurens V, McMullan M, van Oosterhout C (2012) Cryptic MHC polymorphism revealed but not explained by selection on the class IIB peptide-binding region. Mol Biol Evol 29:1631–1644. https://doi.org/10.1093/molbev/mss012

Loehlin DW, Carroll SB (2016) Expression of tandem gene duplicates is often greater than twofold. Proc Natl Acad Sci 113:5988–5992. https://doi.org/10.1073/pnas.1605886113

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155. https://doi.org/10.1126/science.290.5494.1151

Maccari G, Robinson J, Bontrop RE et al (2018) IPD-MHC: nomenclature requirements for the non-human major histocompatibility complex in the next-generation sequencing era. Immunogenetics 70:619–623. https://doi.org/10.1007/s00251-018-1072-4

Málaga-Trillo E, Zaleska-Rutczynska Z, McAndrew B et al (1998) Linkage relationships and haplotype polymorphism among cichlid Mhc class II B loci. Genetics 149:1527–1537. https://doi.org/10.1093/genetics/149.3.1527

Malmstrøm M, Matschiner M, Tørresen OK et al (2016) Evolution of the immune system influences speciation rates in teleost fishes. Nat Genet 48:1204–1210. https://doi.org/10.1038/ng.3645

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

Miller KM, Li S, Ming TJ et al (2006) The salmonid MHC class I: more ancient loci uncovered. Immunogenetics 58:571-589. https://doi.org/10.1007/s00251-006-0125-2

Minias P, Pikus E, Whittingham LA et al (2019) Evolution of copy number at the MHC varies across the avian tree of life. Genome Biol Evol 11:17–28. https://doi.org/10.1093/gbe/evy253

Minias P, Włodarczyk R, Remisiewicz M et al (2021) Distinct evolutionary trajectories of MHC class I and class II genes in Old World finches and buntings. Heredity. https://doi.org/10.1038/s41437-021-00427-8

Monvises A, Nuangsaeng B, Sriwattanarothai N et al (2009) The Siamese fighting fish: well-known generally but little-known scientifically. ScienceAsia 35:8–16. https://doi.org/10.2306/scienceasia1513-1874.2009.35.008

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

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

Nielsen R, Paul JS, Albrechtsen A et al (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451. https://doi.org/10.1038/nrg2986

Niimura Y, Matsui A, Touhara K (2014) Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res 24:1485–1496. https://doi.org/10.1101/gr.169532.113

Nonaka MI, Nonaka M (2010) Evolutionary analysis of two classical MHC class I loci of the medaka fish, *Oryzias latipes*: haplotype-specific genomic diversity, locus-specific polymorphisms, and interlocus homogenization. Immunogenetics 62:319–332. https://doi.org/10.1007/s00251-010-0426-3

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

O'Connor EA, Westerdahl H (2021) Trade-offs in expressed major histocompatibility complex diversity seen on a macroevolutionary scale among songbirds. Evolution 75:1061–1069. https://doi.org/10.1111/evo.14207

O'Connor EA, Westerdahl H, Burri R et al (2019) Avian MHC evolution in the era of genomics: phase 1.0. Cells 8:1152. https://doi.org/10.3390/cells8101152

Panijpan B, Kowasupat C, Laosinchai P et al (2014) Southeast Asian mouth-brooding *Betta* fighting fish (Teleostei: Perciformes) species and their phylogenetic relationships based on mitochondrial COI and nuclear ITS1 DNA sequences and analyses. Meta Gene 2:862–879. https://doi.org/10.1016/j.mgene.2014.10.007

Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. Heredity 96:7–21. https://doi.org/10.1038/sj.hdy.6800724

Pond SLK, Frost SDW, Muse SV (2004) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679. https://doi.org/10.1093/bioinformatics/bti079

Qurkhuli T, Schwensow N, Brändel SD et al (2019) Can extreme MHC class I diversity be a feature of a wide geographic range? The example of Seba's short-tailed bat (*Carollia perspicillata*). Immunogenetics 71:575–587. https://doi.org/10.1007/s00251-019-01128-7

R Core Team (2020) R: a language and environment for statistical computing.

Radwan J, Babik W, Kaufman J et al (2020) Advances in the evolutionary understanding of MHC polymorphism. Trends Genet 36:298–311. https://doi.org/10.1016/j.tig.2020.01.008

Radwan J, Biedrzycka A, Babik W (2010) Does reduced MHC diversity decrease viability of vertebrate populations? Biol Conserv 143:537–544. https://doi.org/10.1016/j.biocon.2009.07.026

Radwan J, Kuduk K, Levy E et al (2014) Parasite load and MHC diversity in undisturbed and agriculturally modified habitats of the ornate dragon lizard. Mol Ecol 23:5966–5978. https://doi.org/10.1111/mec.12984

Rambaut A (2018) FigTree v1.4.4.

Reusch TBH, Schaschl H, Wegner KM (2004) Recent duplication and inter-locus gene conversion in major histocompatibility class II genes in a teleost, the three-spined stickleback. Immunogenetics 56:427–437. https://doi.org/10.1007/s00251-004-0704-z

Rinker DC, Specian NK, Zhao S et al (2019) Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. Proc Natl Acad Sci 116:13446–13451. https://doi.org/10.1073/pnas.1901093116

Roche PA, Furuta K (2015) The ins and outs of MHC class II-mediated antigen processing and presentation. Nat Rev Immunol 15:203–216. https://doi.org/10.1038/nri3818

Ronquist F, Teslenko M, Van Der Mark P et al (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542. https://doi.org/10.1093/sysbio/sys029

Roth O, Solbakken MH, Tørresen OK et al (2020) Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. Proc Natl Acad Sci 117:9431–9439. https://doi.org/10.1073/pnas.1916251117

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC et al (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34:3299–3302. https://doi.org/10.1093/molbev/msx248

Rüber L, Britz R, Tan HH et al (2004) Evolution of mouthbrooding and life-history correlates in the fighting fish genus *Betta*. Evolution 58:799–813. https://doi.org/10.1111/j.0014-3820.2004.tb00413.x

Sammut B, Marcuz A, Pasquier LD (2002) Correction, Vol. 32(6) 2002, pp 1593–1604 The fate of duplicated major histocompatibility complex class Ia genes in a dodecaploid amphibian. Xenopus Ruwenzoriensis Eur J Immunol 32:2698–2709. https://doi.org/10.1002/1521-4141(200209)32:9%3c2698::AID-IMMU2698%3e3.0.CO;2-U

Sato A, Dongak R, Hao L et al (2012) Organization of Mhc class II A and B genes in the tilapiine fish *Oreochromis*. Immunogenetics 64:679–690. https://doi.org/10.1007/s00251-012-0618-0

Sato A, Figueroa F, Murray BW et al (2000) Nonlinkage of major histocompatibility complex class I and class II loci in bony fishes. Immunogenetics 51:108–116. https://doi.org/10.1007/s002510050019

Sedlazeck FJ, Rescheneder P, von Haeseler A (2013) NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics 29:2790–2791. https://doi.org/10.1093/bioinformatics/btt468

Shiina T, Dijkstra JM, Shimizu S et al (2005) Interchromosomal duplication of major histocompatibility complex class I regions in rainbow trout (*Oncorhynchus mykiss*), a species with a presumably recent tetraploid ancestry. Immunogenetics 56:878-893. https://doi.org/10.1007/s00251-004-0755-1

Shand R, Dixon B (2001) Teleost major histocompatibility genes: diverse but not complex. Mod Asp Immunobiol 2:66–72

Sin SYW, Cloutier A, Nevitt G et al (2021) Olfactory receptor subgenome and expression in a highly olfactory procellariiform seabird. Genetics. https://doi.org/10.1093/genetics/iyab210

Sin YW, Annavi G, Dugdale HL et al (2014) Pathogen burden, co-infection and major histocompatibility complex variability in the European badger (*Meles meles*). Mol Ecol 23:5072–5088. https://doi.org/10.1111/mec.12917

Sin YW, Annavi G, Newman C et al (2015) MHC class II-assortative mate choice in European badgers (*Meles meles*). Mol Ecol 24:3138–3150. https://doi.org/10.1111/mec.13217

Sin YW, Dugdale HL, Newman C et al (2012a) Evolution of MHC class I genes in the European badger (*Meles meles*). Ecol Evol 2:1644–1662. https://doi.org/10.1002/ece3.285

Sin YW, Dugdale HL, Newman C et al (2012b) MHC class II genes in the European badger (*Meles meles*): characterization, patterns of variation, and transcription analysis. Immunogenetics 64:313–327. https://doi.org/10.1007/s00251-011-0578-9

Song K, Li L, Zhang G (2016) Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. Sci. Rep. 6:35736. https://doi.org/10.1038/srep35736

Spielmann M, Lupiáñez DG, Mundlos S (2018) Structural variation in the 3D genome. Nat Rev Genet 19:453–467. https://doi.org/10.1038/s41576-018-0007-0

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Royal Soc B 277:979–988. https://doi.org/10.1098/rspb.2009.2084

Stervander M, Dierickx EG, Thorley J et al (2020) High MHC gene copy number maintains diversity despite homozygosity in a critically endangered single-island endemic bird, but no evidence of MHC-based mate choice. Mol Ecol 19:3578–3592. https://doi.org/10.1111/mec.15471

Stet RJ, Kruiswijk CP, Dixon B (2003) Major histocompatibility lineages and immune gene function in teleost fishes: the road not taken. Crit Rev Immunol 23:441–471. https://doi.org/10.1615/critrevimmunol.v23.i56.50

Sudmant PH, Rausch T, Gardner EJ et al (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81. https://doi.org/10.1038/nature15394

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680. https://doi.org/10.1093/nar/22.22.4673

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. https://doi.org/10.1093/bib/bbs017

Völker M, Backström N, Skinner BM et al (2010) Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. Genome Res 20:503–511. https://doi.org/10.1101/gr.103663.109

Weaver S, Dube S, Mir A et al (2010) Taking qPCR to a higher level: analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. Methods 50:271–276. https://doi.org/10.1016/j.ymeth.2010.01.003

Wegner KM, Kalbe M, Kurtz J et al (2003) Parasite selection for immunogenetic optimality. Science 301:1343–1343. https://doi.org/10.1126/science.1088293

Winternitz JC, Minchey SG, Garamszegi LZ et al (2013) Sexual selection explains more functional variation in the mammalian major histocompatibility complex than parasitism. Proc Royal Soc B 280:20131605. https://doi.org/10.1098/rspb.2013.1605

Woelfing B, Traulsen A, Milinski M et al (2009) Does intra-individual major histocompatibility complex diversity keep a golden mean? Philos. Trans. R. Soc. Lond., B. Biol Sci 364:117–128. https://doi.org/10.1098/rstb.2008.0174

Wu Y, Zhang N, Hashimoto K et al (2021) Structural comparison between MHC classes I and II; in Evolution, a Class-II-Like Molecule Probably Came First. Front Immunol 12:621153. https://doi.org/10.3389/fimmu.2021.621153

Yamaguchi T, Dijkstra JM (2019) Major histocompatibility complex (MHC) genes and disease resistance in fish. Cells 8:378. https://doi.org/10.3390/cells8040378

Zarrei M, MacDonald JR, Merico D et al (2015) A copy number variation map of the human genome. Nat Rev Genet 16:172–183. https://doi.org/10.1038/nrg3871

Zhai T, Yang H-Q, Zhang R-C et al (2017) Effects of population bottleneck and balancing selection on the Chinese alligator are revealed by locus-specific characterization of MHC genes. Sci Rep 7:5549. https://doi.org/10.1038/s41598-017-05640-2

Zhang F, Gu W, Hurles ME et al (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genom Hum Genet 10:451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang W, Wang H, Brandt DYC et al (2021) The genetic architecture of phenotypic diversity in the betta fish (*Betta splendens*). bioRxiv https://doi.org/10.1101/2021.05.10.443352

Zhao M, Wang Q, Wang Q et al (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinform 14:S1. https://doi.org/10.1186/1471-2105-14-S11-S1

Zagalska-Neubauer M, Babik W, Stuglik M et al (2010) 454 sequencing reveals extreme complexity of the class II major histocompatibility complex in the collared flycatcher. BMC Evol Biol 10:395. https://doi.org/10.1186/1471-2148-10-395

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.