Immunogenetics (2019) 71:395–405
https://doi.org/10.1007/s00251-019-01111-2

ORIGINAL ARTICLE

Distribution of ancient α1 and α2 domain lineages between two classical MHC class I genes and their alleles in grass carp

Zibin Li[1] · Nan Zhang[1] · Lizhen Ma[1] · Zehui Qu[1] · Xiaohui Wei[1] · Zixin Liu[1] · Minghu Tang[2] · Nianzhi Zhang[1] · Yinan Jiang[1] · Chun Xia[1]

Received: 12 December 2018 / Accepted: 5 March 2019 / Published online: 2 April 2019

Abstract

Major histocompatibility complex (MHC) class I molecules play a crucial role in the immune response by binding and presenting pathogen-derived peptides to specific $CD8^+$ T cells. From cDNA of 20 individuals of wild grass carp (*Ctenopharyngodon idellus*), we could amplify one or two alleles each of classical MHC class I genes *Ctid-UAA* and *Ctid-UBA*. In total, 27 and 22 unique alleles of *Ctid-UAA* and *Ctid-UBA* were found. The leader, α1, transmembrane and cytoplasmic regions distinguish between *Ctid-UAA* and *Ctid-UBA*, and their encoded α1 domain sequences belong to the ancient lineages α1-V and α1-II, respectively, which separated several hundred million years ago. However, *Ctid-UAA* and *Ctid-UBA* share allelic lineage variation in their α2 and α3 sequences, in a pattern suggestive of past interlocus recombination events that transferred α2+α3 fragments. The allelic Ctid-UAA and Ctid-UBA variation involves ancient variation between domain lineages α2-I and α2-II, which in the present study was dated back to before the ancestral separation of teleost fish and spotted gar (> 300 million years ago). This is the first report with compelling evidence that recombination events combining different ancient α1 and α2 domain lineages had a major impact on the allelic variation of two different classical MHC class I genes within the same species.

Keywords Fish · MHC class I · Evolution · Lineage · Recombination · Polymorphism

Introduction

Major histocompatibility complex (MHC) class I molecules play a crucial role in the immune response by binding and presenting pathogen-derived peptides to specific $CD8^+$ T cells (Townsend et al. 1985). The result is proliferation of cytotoxic T lymphocytes (CTLs) and, eventually, pathogen clearance from the host. Classical MHC class I complexes consist of an MHC-I heavy chain molecule, β2-microglobulin (β2m) and a peptide ligand of ~ 9 amino acids. MHC-I heavy chain molecules have three extracellular domains, α1, α2, and α3, which each are encoded by different exons. The α1 and α2 domains are the functionally important parts and together constitute a structure with a groove for binding a peptide ligand, and they make important contacts with T cell receptors (TCR) on T cells. The α3 domain, like β2m, is an Ig-like structure. The α3 domain is connected with transmembrane (TM) and cytoplasmic (CY) domains. Classical MHC class I genes have a high degree of genetic polymorphism, particularly among positions associated with peptide-binding sites (Bjorkman et al. 1987), and their levels of sequence diversity are considered the highest among any genes in the vertebrate genome (Vandiedonck and Knight 2009). MHC class I genes have been well studied across species. In humans, classical MHC class I genes are present at three loci (*HLA-A*, *-B*, and *-C*), and thousands of alleles are found for each locus (Robinson et al. 2015). The alleles of each locus are found in varying combinations among different haplotypes of the *Mhc* region on human chromosome 6. *Mhc* haplotype diversity is believed to be associated with susceptibility to numerous infectious diseases (Hill 2001).

In teleost fish, the first MHC gene was isolated from the carp genome by Hashimoto et al. (1990), and different MHC

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00251-019-01111-2) contains supplementary material, which is available to authorized users.

✉ Chun Xia
xiachun@cau.edu.cn

[1] Department of Microbiology and Immunology, College of Veterinary Medicine, China Agricultural University, Beijing, China

[2] Chinese Carp of Yangtze River System and Primitive Breed Fishery, Guangling, Yangzhou, China

genes have subsequently been found in other teleost fish (Ono et al. 1992; Grimholt et al. 1993; Hansen et al. 1999). Our knowledge regarding teleost MHC genes has gradually grown based on whole-genome sequencing and bioinformatic analyses (Howe et al. 2013). In contrast to other vertebrates, teleost fish classical MHC class I and II genes are not linked in the genome (Bingulac-Popovic et al. 1997), and only classical MHC class I genes are found in typical *Mhc* regions (Dijkstra et al. 2013; Grimholt et al. 2015). Previous reports have shown that teleost MHC I genes can be classified into five lineages, Z, U, L, S, and P, with classical genes only being found in the U lineage (Stet et al. 1998; Kruiswijk et al. 2002; Dijkstra et al. 2007; Lukacs et al. 2010; Grimholt et al. 2015). Zebrafish (*Danio rerio*) represents an ideal model organism, and its MHC genes have been well studied, with three MHC class I lineages being described thus far: U, Z, and L. Among U lineage genes, classical MHC class I genes are located in the *Mhc* region found on chromosome 19, and nonclassical MHC class I genes are located on chromosome 22 (Sambrook et al. 2005; Dirscherl et al. 2014). Zebrafish *Mhc* haplotypes can have one, two, or three classical MHC class I genes that are very different from each other and between haplotypes, and which represent ancient α1 and α2 domain lineage variation (McConnell et al. 2014; Grimholt et al. 2015). The variation between zebrafish *Mhc* haplotypes is further amplified by them harbouring highly divergent antigen-processing genes (McConnell et al. 2016).

The grass carp *Ctenopharyngodon idellus*, like zebrafish a member of the Cyprinidae family of teleost fish, has great commercial value and a worldwide distribution as an important aquaculture species. Global production of cultured or farmed grass carp yields approximately 4.6 million tons per year, accounting for 15.6% of the global freshwater aquaculture production in 2011. Completion of the grass carp genome sequence has accelerated studies on molecular immunology and immunogenetics in teleosts (Wang et al. 2015). Our lab previously (Yang et al. 2006; Chen et al. 2010) identified variable classical grass carp MHC class I gene sequences of the U lineage which we now designate together as "*Ctid-UAA*" type. For one grass carp Ctid-UAA molecule, we determined a pMHC-I structure (the complex of heavy chain, β2m, and peptide) using X-ray crystallography (Chen et al. 2017), representing the first elucidation of a pMHC-I structure in fish. These data provided the first direct evidence for a classical manner of peptide presentation by fish MHC class I molecules similar to mammals.

In the present study, by analysing cDNA of 20 wild grass carp individuals, we amplified classical MHC class I sequences of the previously known *Ctid-UAA* type, but also of a newly identified *Ctid-UBA* type. Based on sequence analysis and distribution of alleles among grass carp individuals, we concluded that *Ctid-UAA* and *Ctid-UBA* should represent two different genes. The Ctid-UAA and Ctid-UBA molecules show a unique allelic variation pattern, consisting of various combinations of sequences belonging to ancient α1 and α2 domain lineages.

Materials and methods

Animals

Twenty wild grass carp individuals were captured from the Yangtze River in China. Blood (100 μl to 1 ml) was collected from the fishtail vein using a heparinized syringe with a 23G needle. Blood cells were washed with PBS, resuspended in up to 1 ml of RPMI 1640 medium with 10% DMSO and stored at − 80 °C for follow-up total RNA extraction. Grass carp individuals were named G1-G20 upon capture.

RNA isolation, cDNA synthesis, PCR, cloning, and sequencing

According to a standard protocol, total RNA from the blood was extracted using TRIzol reagent, and RNA concentrations were measured via spectrophotometry. Once the integrity was confirmed by analysis on a 1.5% (*w*/*v*) agarose gel, RNA was reverse transcribed into cDNA using the ExScript RT Reagent Kit. According to the two known grass carp MHC class I genes (U-EF584535 and U-EF584535), two pairs of primers were manually designed that could amplify full-length *Ctid-UAA* and *Ctid-UBA* sequences (Supplementary Table S1). Amplifications were performed in a 50 μl reaction volume, including 100–200 ng of cDNA, 25 pmol of each primer, 4 μl of 2.5 mM dNTP, 10 × PCR buffer with 15 mM $MgCl_2$, and 1 unit of HiFi DNA polymerase. The thermal cycling conditions were set for an initial denaturation at 98 °C for 5 min, followed by 35 cycles at 98 °C for 10 s, 62 °C or 61 °C for 15 s, and 72 °C for 1 min 20 s, with a final extension at 72 °C for 10 min. PCR products were confirmed on 1.5% (*w*/*v*) agarose gels, inserted into the pMD 18-T vector and transformed into *E. coli* DH5α competent cells. Positive clones were selected and sent to BioSune (Shanghai, China) for sequencing. Once the same sequence was verified from two independent PCR reactions in an individual sample or from two different grass carp individuals, this sequence was regarded as a new allele. Finally, all cloned alleles were submitted to GenBank (National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/genbank/), and their accession numbers are listed in Supplementary Table S2.

Phylogenetic tree analysis

The obtained sequence data were sorted using Chromas Ver. 2.0 (http://technelysium.com.au/wp/chromas/). Multiple sequence alignments were conducted using ClustalW Ver. 1.

83 and Jalview Ver. 1.2. Signal peptide and transmembrane helix predictions in proteins was performed using the SignaIP 4.1 online tool (http://www.cbs.dtu.dk/services/SignalP/) and the TMHMM Server Ver. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/), respectively. The phylogenetic tree was constructed by the MEGA-6 program using the neighbour-joining method (Saitou and Nei 1987), and bootstrap values were obtaining using 1000 replicates. Evolutionary genetic distance nucleotides were estimated by the Kimura-2 parameter method (Kimura 1980) using MEGA-6.

Analysis of grass carp genomic sequences deposited in databases

Whole-genome sequencing of grass carp has been accomplished (Wang et al. 2015), and all grass carp data have been released on the official National Center for Gene Research website (http://www.ncgr.ac.cn/grasscarp/). BioEdit Ver. 7.0.5 and Notepad++ Ver. 7.5.1 were used to identify the new grass carp MHC class I gene in the current reference genome. Two known grass carp MHC class I genes (U-EF584535 and U-EF584535) were identified by grass carp DNA library (Chen et al. 2010). Intron-exon organisations of the known grass carp MHC class I genes were predicted by using GENSCAN (http://genes.mit.edu/GENSCAN.html) and comparison with known genes.

Estimation of the *Ctid-UAA* and *Ctid-UBA* allele variability index

The Wu-Kabat variability index was applied to estimate the variance coefficients of the different lineages of the UAA/UBA α1 and α2 domains (Wu and Kabat 1970); variability scores > 6 were defined as highly variable sites (HVSs), as was done previously (Fan et al. 2018; Zhang et al. 2019). The crystal structure of a grass carp MHC class I molecule was uploaded to the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do) under accession number 5Y91. The model was generated with PyMOL Molecular Graphics System Ver. 1.7.4.2 distributed by Schrödinger. These sites were then analysed statistically and located on the crystal structure model and labelled.

Results

Grass carp express two classical MHC class I genes, *Ctid-UAA* and newly identified *Ctid-UBA*

In our previous studies (Yang et al. 2006; Chen et al. 2010), a number of classical MHC class I sequences were found for grass carp. Because we hypothesised that the previously reported fosmid sequences EF584535 and EF584536 (Chen et al. 2010) might represent different MHC class I genes (and not alleles), we derived primers from each of them for amplification of full-length ORF from cDNA. The new primer sets were used to amplify cDNA of 20 wild grass carp individuals, and between two and four different sequences were found per individual. Based on sharing similar leader, α1 and TM/CY coding regions, they were distinguished into two types of sequences, *Ctid-UAA* and *Ctid-UBA*, for which 27 and 22 different sequences were found, respectively. The detection of one or two *Ctid-UAA* and *Ctid-UBA* sequences each in all individuals is in agreement with them representing different genes (Table 1). Based on overall similarity levels, alleles were numbered in *01, *02 and *03 (for *Ctid-UAA*) and *01 and *02 (for *Ctid-UBA*) categories, with the two numbers following those indications (e.g. *0101) used for specifying the individual allele. Representative deduced amino acid sequences are shown in Fig. 1, and all deduced amino acids are shown per domain in Figs. S1a (α1 domain), S1b (α2 domain), S1c (α3 domain), and S1d (leader and TM/CY regions). Phylogenetic trees for the α1, α2, and α3 domains are shown in Fig. 2, and highlight that the α1 sequences are specific for either Ctid-UAA or Ctid-UBA, whereas the lineage variation found among the α2 and α3 sequences is shared between Ctid-UAA and Ctid-UBA. Figures 1 and 2 show that the α2 and α3 sequences are found in consistent lineage pairs, whereas Fig. 1 and Fig. S1d show that this is also true for the leader, α1 and TM/CY regions. For the α1 and α2 domain lineages we use a nomenclature first introduced by Aoyagi et al. (2002), with Ctid-UAA-α1 belonging to domain lineage α1-V (see also Yang et al. 2006), Ctid-UBA-α1 belonging to α1-II, and the α2 domain sequences belonging to either lineage α2-I or α2-II. Sequences belonging to lineages α1-II, α1-V, α2-I, and α2-II were also found in salmonid fishes like rainbow trout and Atlantic salmon (Fig. 2; Aoyagi et al. 2002; Kiryu et al. 2005) with which grass carp shared the last common ancestor > 200 million years ago.

Comparison with reported genomic sequences of grass carp MHC class I

Contrary to our original hypothesis, based on sequence comparisons and phylogenetic tree analysis (Figs. 2 and S1), we now believe that the MHC genes situated in the fosmid sequences EF584535 and EF584536 (Chen et al. 2010; U-EF584535 and U-EF584536 in the figures) may represent *Ctid-UAA* alleles. Their intron-exon organisation is shown in Fig. 3. We also screened the grass carp reference genome sequence for genes most similar to *Ctid-UAA* and *Ctid-UBA*. In an apparent *Mhc* region with also tapasin, TAP2 and PSMB genes (not shown), we found an intact *Ctid-UAA* gene, a *Ctid-UAA-like* gene which may be a pseudogene because it lacks the expected leader and TM/CY regions, and a *Ctid-UCA* gene which does not show close similarities with either *Ctid-*

Table 1 Alleles expressed in cDNA in each grass carp

Grass carp	UAA alleles		UBA alleles	
G1	*UAA**0101 (10)	*UAA**0201 (13)	***UBA**0101 (13)**	*UBA**0201 (10)
G2	***UAA**0102 (7)**	***UAA**0202 (11)**	***UBA**0102 (14)**	***UBA**0103 (4)**
G3	***UAA**0103 (10)**	*UAA**0301 (6)	***UBA**0103 (6)**	***UBA**0202 (10)**
G4	*UAA**0104 (4)		***UBA**0202 (11)**	***UBA**0203 (10)**
G5	***UAA**0105 (19)**	*UAA**0203 (8)	***UBA**0204 (10)**	
G6	***UAA**0102 (5)**	***UAA**0106 (6)**	*UBA**0104 (15)	***UBA**0204 (17)**
G7	*UAA**0107 (3)		***UBA**0203 (9)**	
G8	***UAA**0103 (22)**	*UAA**0204(10)	***UBA**0205 (3)**	
G9	***UAA**0105 (16)**	*UAA**0108 (10)	***UBA**0105 (6)**	***UBA**0205 (9)**
G10	***UAA**0105 (15)**	*UAA**0206 (10)	***UBA**0101 (8)**	*UBA**0106 (11)
G11	***UAA**0205 (6)**	***UAA**0303(5)**	***UBA**0105(10)**	***UBA**0206 (13)**
G12	***UAA**0106 (12)**		*UBA**0107 (21)	***UBA**0206 (10)**
G13	***UAA**0202(10)**	*UAA**0207 (11)	*UBA**0207 (7)	
G14	*UAA**0109 (9)	***UAA**0303 (11)**	*UBA**0108(9)	***UBA**0208 (4)**
G15	*UAA**0208 (10)		***UBA**0208 (3)**	***UBA**0209 (5)**
G16	*UAA**0209 (22)	***UAA**0304(13)**	*UBA**0109 (10)	
G17	*UAA**0110 (19)	*UAA**0305 (4)	*UBA**0110 (4)	***UBA**0209 (8)**
G18	***UAA**0304(13)**	*UAA**0306 (4)	***UBA**0205 (6)**	*UBA**0210 (11)
G19	*UAA**0111 (13)		***UBA**0102 (8)**	*UBA**0211 (15)
G20	*UAA**0210 (13)	***UAA**0302(10)**	*UBA**0212 (10)	

The number of matching clones is indicated in parentheses. Shared alleles are indicated in bold

UAA or *Ctid-UBA*. The intron-exon organisation of these genes is shown in Fig. 3, and their encoded sequences are compared in Figs. 2 and S1. Fig. S1a (and Fig. 2) and Fig. S1d show that the predicted U-EF584535 molecule has a *UAA*-type leader plus α1 domain region, but a *UBA*-type TM/CY region, which complicates the gene assignment in the present study. Immediately, apparent from the intron-exon organisation is the large intron in all the depicted intact genes between the α1 and α2 domain exons, which is much larger than found in human, as was first reported for a classical MHC class I gene in common carp (van Erp et al. 1996). We were not able to find *Ctid-UBA* sequences in the available genomic sequence information. We refrain from showing a more detailed analysis of the reference genome *Mhc* region, because the observed high frequency of apparent pseudogenes (for genes other than MHC) may represent the biological situation but might also derive from sequencing errors. In the future, we hope to address this by sequencing several *Mhc* haplotype regions by ourselves.

Residues typical for classical MHC class I

Amino acids Y7, Y59, Y84, T142, K145, W146, Y155, and Y167 are conserved among mammalian classical MHC class I molecules and are involved in binding the termini of MHC class I-bound peptides in the peptide-binding groove (Hashimoto et al. 1990). Previous studies have found almost complete conservation of these amino acids in classical U lineage sequences in teleosts, except for R instead of Y at position 84. The Ctid-UAA and Ctid-UBA sequences possess these conserved amino acids (Fig. 1). In addition, we found an N-linked glycosylation (NQS) motif at N83 (numbering based on the Fig. 1 alignment; Figs. 1 and S1a; also see Text S3a in Grimholt et al. 2015), and four conserved cysteine residues (C98, C160, C196, and C254) (Fig. 1). Significantly, 9 amino acids, D17, Q31, Q93, D119, K191, T199, E229, Q237, and N278, which participate in interacting with β2m, are highly conserved among Ctid-UAA and Ctid-UBA (Figs. 1 and S1). In addition to interacting with β2m, the α3 domain is indispensable for binding CD8 on cytotoxic T cells during cell-mediated immunity. In mammals, six residues in the α2 (Q114, D121, and E127) and α3 (T225, Q226, and L230) domains have been shown to be key residues involved in interacting with CD8 molecules. Residues D119 and L227 are completely conserved among Ctid-UAA and Ctid-UBA sequences (Figs. 1 and S1), and they can be viewed as positions potentially interacting with CD8 molecules. D119 also participates in MHC class I molecule binding to β2m.

Sequence motifs characteristic for the ancient α1 and α2 domain lineages

The α1-II, α1-V, α2-I, and α2-II lineages are hundreds of millions of years old, and we tried to understand their

Fig. 1 Alignment of representative deduced amino acid sequences. Dots represent identity to the reference Ctid-UAA*0101 allele, and dashes represent indels. The α1 and α2 domain lineages are shown on the right hand side of each sequence. The letters a–f above the alignment indicate positions that in the pCtid-UAA*0101 structure (PDB 5Y91) form part of the binding pockets A–F. Eight conserved residues in mammalian MHC class I sequences are indicated with a circle (●). Four conserved cysteine residues are indicated with a triangle (▲). Conserved residues interacting with β2m are indicated with a square (■). Boxes indicate the N-linked glycosylation target site. Positions for which the variation per gene per lineage surpasses a threshold value for being deemed polymorphic (see Fig. 5) are coloured light blue (participating in pockets A–F) and grey (others), respectively. Lineage-specific residues are shown in different colours: α1-V (green), α1-II (yellow), α2-I (pink), and α2-II (cyan)

characteristic differences by studying sequence comparisons (Fig. S1 and Grimholt et al. 2015) and by mapping the detected characteristic motifs on the elucidated grass carp pMHC-I structure (which is based on a Ctid-UAA*0101 allele that has α1-V and α2-II sequences; Chen et al. 2017). Residues that we deem characteristic are highlighted in colour in Fig. 1, and their structural positions are investigated in Fig. 4.

The α1-II lineage is best characterised by three rather unique residues that may form part of the B pocket (Matsumura et al. 1992; Hashimoto et al. 1990), namely D23, S33 and K42. Figure 4d shows that in the Ctid-UAA pMHC-I structure the residues A23, M33, and A42 do not all participate directly in the B pocket, but from their orientations, it can be assumed that residues with larger sidechains (like D23 and K42) may participate.

The α1-V lineage is more difficult to characterise by conserved and unique motifs. Grimholt et al. (2015) highlighted the D15 and Q87 as rather characteristic for α1-V. Indeed, D15 is found in most α1-V sequences, including in more than half of the Ctid-UAA and Ctid-UBA sequences (Fig. S1a; Grimholt et al. 2015), whereas sequences of other lineages tend to have a hydrophobic residue at this position (Grimholt et al. 2015). A substantial set of the investigated Ctid-UAA and Ctid-UBA sequences has G15 (Fig. S1a), and so also shares the property of not being hydrophobic. Like Grimholt et al. (2015), we also found Q87 conserved in the α1-V sequences, but point out that this residue is also conserved among α1-II sequences (Fig. 1). Nevertheless, we chose to follow the lead by Grimholt et al. (2015) to mark this residue as characteristic for α1-V, because in the other seven distinguished α1 domain lineages this residue is absent or rare (Grimholt et al. 2015). The function of the D/G15 and Q87 residues is not obvious from their structural position (Fig. 4a, c), but their presence/absence probably does not cause differences in peptide-binding properties.

A consistent difference between α2-I and α2-II sequences is the W100 residue in α2-I compared to other residues (mostly leucine) at this position in α2-II, and a single residue insertion in the α2-I sequences a few residues C-terminal of this tryptophan (Fig. 1 and S1b; Grimholt et al. 2015). The position 100 residue resides at the end of the S1 strand of the α2 domain, and although we are not confident about the

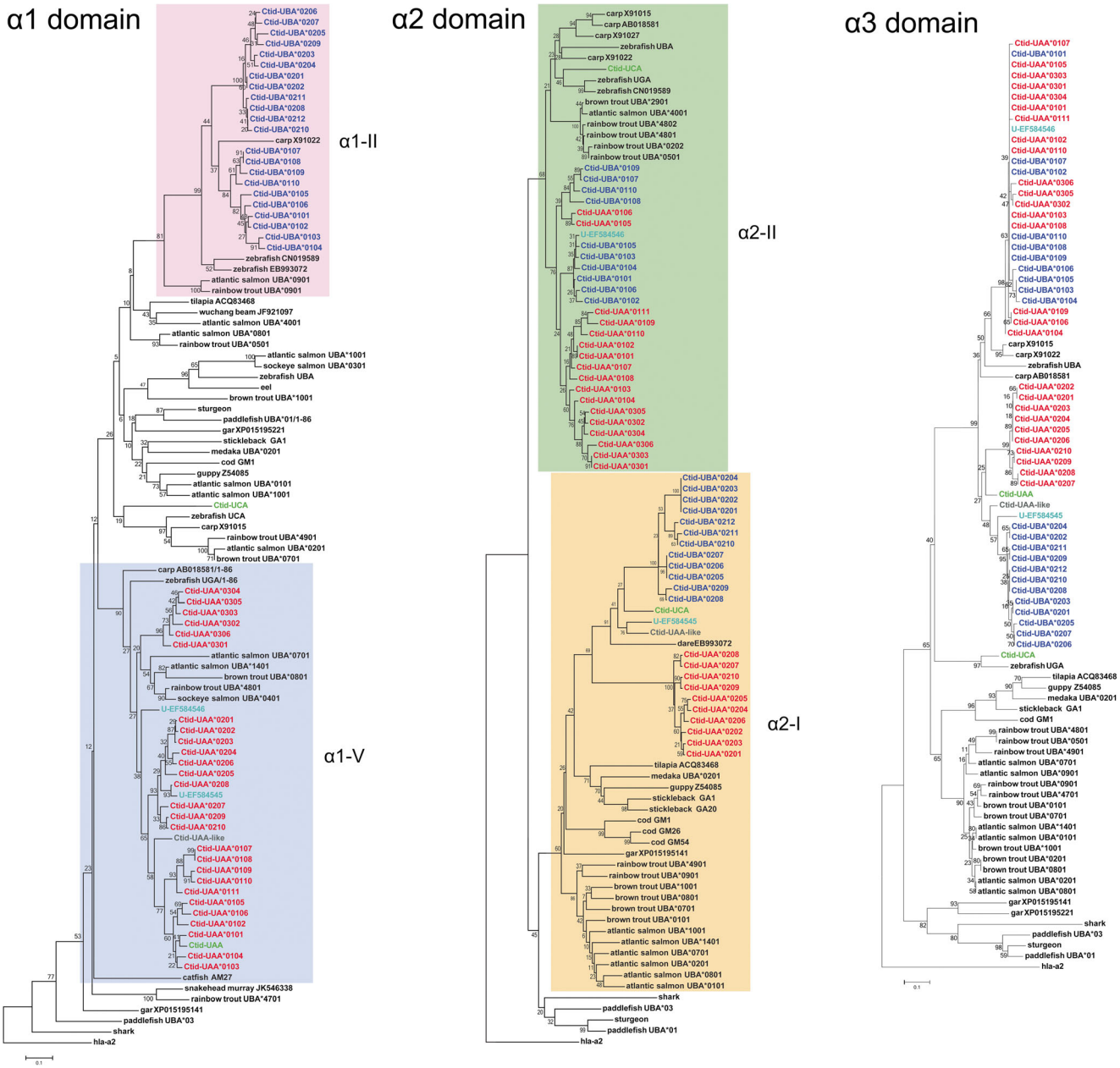

Fig. 2 Molecular phylogenetic analysis of α1, α2, and α3 domains of all cloned Ctid-UAA and Ctid-UBA as determined by the neighbour-joining method. Phylogenetic tree based on handmade alignment (Fig. S1a–c) of selected α1, α2, and α3 domains amino acid sequences. The different lineage divisions of α1 and α2 domains into α1-II, α1-V, α2-I, and α2-II were originally introduced by Aoyagi et al. 2002 (see also Kiryu et al. 2005). Subcategories of Ctid-UAA and Ctid-UBA sequences (e.g. *01 versus *02) are indicated by different colouring of their names. Sequence GenBank references are listed in Fig. S1

precise location of the single amino acid insertion, it maps somewhere in the α2 domain S1S2 loop and should make this loop longer (Fig. 4a, which highlights the matching positions in the α2-II domain of Ctid-UAA*0101). In addition, most α2-II sequences have residues K125, E159, and N160, which all can interact with each other (Fig. 4a, c); in contrast, α2-I sequences typically have a hydrophobic residue at position 125 and varying residues at positions 159 and 160 (Fig. S1b; Grimholt et al. 2015). We have no interesting hypothesis for explaining these differences between α2-I and α2-II.

Single residue position allelic polymorphism and HVSs analysis of Ctid-UAA and Ctid-UBA

To analyse the single residue position allelic polymorphism characteristics of Ctid-UAA and Ctid-UBA, the HVSs of the α1 and α2 domains were analysed based on the Wu-Kabat method. This was done separately per gene and per domain lineage (Fig. 5), to allow an estimation of the polymorphism of each molecule independent of the variation caused by "lineage swapping". The positions that were deemed polymorphic beyond a threshold

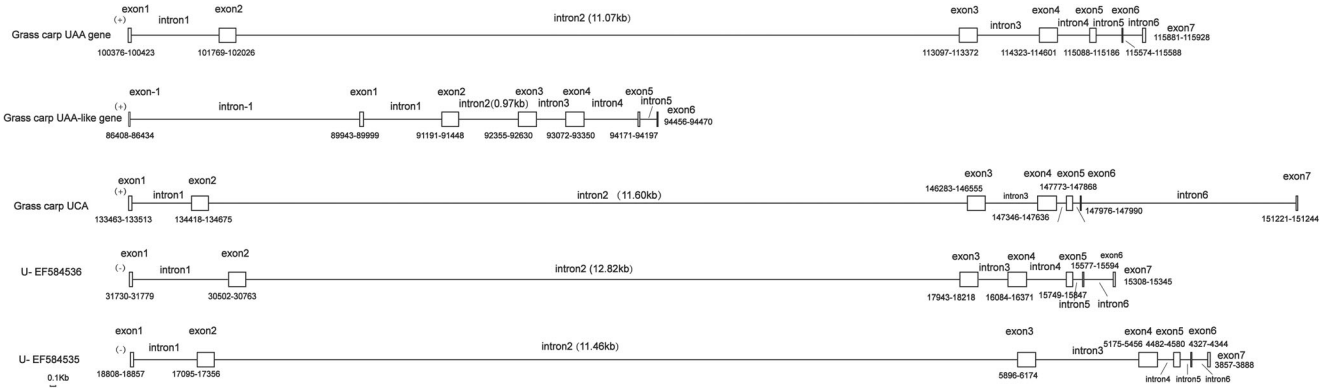

Fig. 3 Intron-exon organisation of grass carp MHC class I genes. In the reference grass carp genome, in scaffold CI01000319, grass carp MHC class I genes *UAA* (scaffold positions 100,376–115,928 kb), *UAA*-like (86408–94,470 kb), and *UCA* (133463-151,244 kb) were identified. The MHC class I genes *U-EF584535* and *U-EF584535* were identified in genomic fosmid sequences in our previous report (Chen et al. 2010). Boxes indicate exons separated by introns. Intron2 sequences are positioned between the α1 and α2 domain exons (exon2 and exon3), and, compared to mammals, are unusually long in several of the depicted genes (see van Erp et al. 1996). The genes are oriented in the forward sequence direction (+), while genes are oriented in the reverse sequence direction (−). The intron and exon sizes of all the genes are drawn to scale

level (variability scores > 6) were visually labelled on the Ctid-UAA pMHC-I structure (Fig. 5). An interesting observation is that the Ctid-UAA and *C*tid-UBA α1 domains, which are gene specific and for which allelic lineage variation was not detected in the present study, appear to show more functional within lineage polymorphism in the peptide-binding groove than found for the α2 domains which do show lineage variation between alleles. If comparing the combined Ctid-UAA and Ctid-UBA α2-I sequences, or their combined α2-II sequences, higher degrees of variation in the expected peptide-binding grooves are found. The latter suggests that although there may not be much allelic variation other than "lineage swapping" in the α2 sequences that contributes to differences in peptide binding, within lineage but between genes α2 point mutation variation is used for increasing the pool of different peptides that can be bound.

Discussion

In the present study, we found that grass carp have at least two expressed classical MHC class I genes which share the property of having various combinations of α1 and α2 domain sequences belonging to ancient lineages. Based on shared leader, α1 and TM/CY regions among the cDNA sequences, we categorised the sequences as *Ctid-UAA* and *Ctid-UBA* alleles, and assume that they belong to two different gene loci.

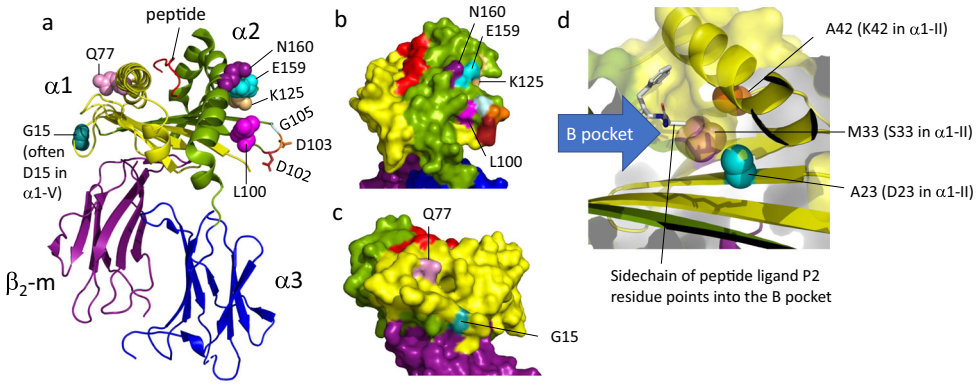

Fig. 4 Structural positions of residues characteristic for the detected α1 and α2 domain lineages. Positions of interests are highlighted on a pMHC-I structure of Ctid-UAA*0101 (PDB accession 5Y91), which has sequences belonging to lineages α1-V and α2-II. Pictures were made using Pymol (Schrödinger) software. **a** The pCtid-UAA structure is shown in cartoon format with different colours for the α1, α2, α3, and β2m domains, and the peptide ligand. Residues at positions at which for α1-V, α2-I, or α2-II characteristic residues are found, are highlighted by showing the respective pCtid-UAA residues in spheres format and individual colours. We are not confident about the position of the single residue insertion in lineage a2-I, except that it should be in the S1S2 loop, for which we highlighted the DDG stretch with different colours for each residue (residue numbers follow Fig. 1 alignment). The figures (**b**) and (**c**) use the same colouring as in (**a**), but are shown in surface format, and from different orientations. **d** The residues D23, S33, and K42 characteristic for α1-II may participate in B pocket formation. A pCtid-UAA structure is shown in cartoon format plus transparent surface format, and its peptide ligand is shown in sticks format. The pCtid-UAA residues A23, M33, and A42 are highlighted by individual colours and by showing their sidechains and Cα atoms in spheres format. We speculate that in case of D23, S33, and K42, all three sidechains may form part of the B pocket

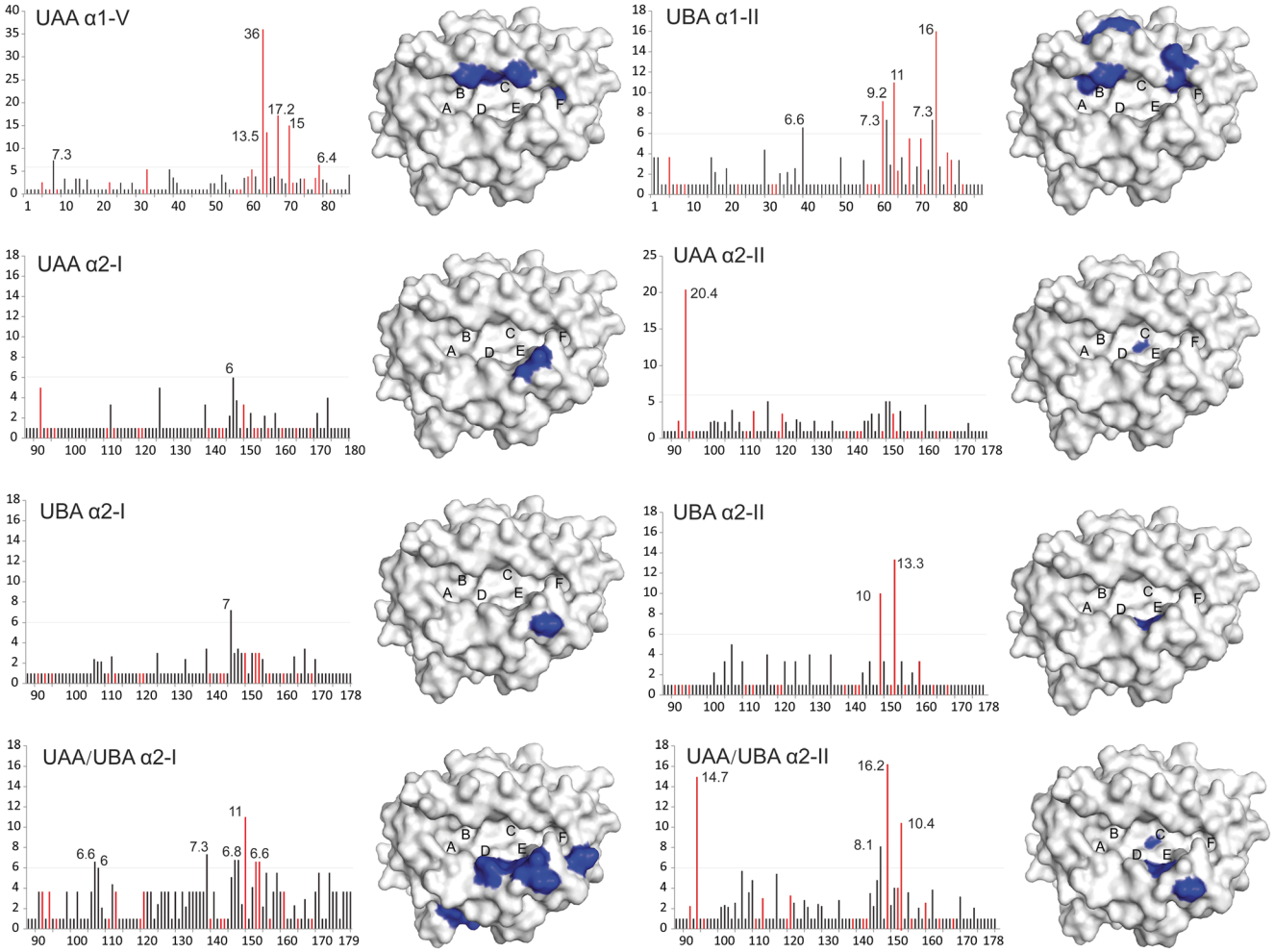

Fig. 5 Wu-Kabat plot of amino acid variability among the sequences of the different lineages of α1 and α2 domains of Ctid-UAA and Ctid-UBA and highly variable sites located in the peptide-binding groove (PBG) of the crystal/Ctid-UAA structure. In the above eight comparisons, only sequences of one domain lineage of either Ctid-UAA or Ctid-UBA are compared with each other. In the bottom two comparisons, all sequences of one α2 domain lineage, regardless whether belonging to Ctid-UAA or Ctid-UBA, are compared with each other. In the Wu-Kabat plots, the positions of the A–F pocket residues are indicated by red bars, and numbers added to bars indicate the variability scores of HVSs. For each comparison, a top-view of the binding groove of the crystal/Ctid-UAA structure is shown in surface format, with most residues coloured white but with the HVSs coloured blue. Approximate locations of the A–F pockets are indicated with the corresponding capital letter; the upper part of the molecule views consist importantly of the α1 domain helices, and the lower part importantly of the α2 domain helices

However, clarification of the zebrafish classical MHC class I situation took more than a few studies (e.g. Takeuchi et al. 1995; Michalova et al. 2000), and only quite recently, it was realised that the zebrafish situation can only be understood at the haplotype level, and that the widely divergent haplotypes with differing numbers of classical MHC class I genes that may undergo recombination prohibit a narrow definition of gene locus identity and "allelic" variation (McConnell et al. 2014, 2016). For a definite understanding of the grass carp classical MHC class I situation, in the future, we hope to sequence several grass carp *Mhc* haplotype regions.

Regardless of the gene assignment of the grass carp cDNA sequences, it is obvious that they are derived from different genes that experienced interlocus recombination events exchanging peptide-binding domain sequences. Among the grass carp cDNA sequences, the leader, α1 and TM/CY regions show lineage consistency, and so do the α2 and α3 domains. We interpret this as most likely caused by transfers of α2+α3 regions between *Ctid-UAA* and *Ctid-UBA* loci at some point in evolution. Recombination probably is stimulated by an unusually large intron between the α1 and α2 domain exons (Fig. 3), which was also found in the case of, for example, common carp (van Erp et al. 1996) and rainbow trout (Shiina et al. 2005). In salmonid fishes, this intron is extensively used for interallelic recombination creating different combinations of ancient α1 and α2 domain sequences (Shum et al. 2001; Aoyagi et al. 2002; Xia et al. 2002; Kiryu et al. 2005; Grimholt et al. 2015). Also in zebrafish there is evidence for recombination involving this intron, although that has not been studied intensively (Michalova et al.

2000; McConnell et al. 2014), and in our previous study in grass carp, we found that the leader plus α1 domain sequence of a *Ctid-UAA* sequence could be replaced by a sequence belonging to a seemingly rare α1 lineage which we will not discuss in the present paper (Yang et al. 2006). Our present study may not have been exhaustive yet in finding all possible *Ctid-UAA* and *Ctid-UBA* allelic variation, but it is unique in its observation of a "lineage-exchange" signature shared by two different genes and constituting a major part of their allelic variations. Although concerning different domain regions and types of recombination, and not resulting in the large allelic variation shared between two different genes as found for *Ctid-UAA* and *Ctid-UBA*, interlocus recombination events were also concluded for two neighbouring classical MHC class I genes in the fish medaka (Nonaka and Nonaka 2010), and for two neighbouring nonclassical MHC class I genes in rainbow trout (Dijkstra et al. 2006). Probably, *Ctid-UAA* and *Ctid-UBA* are also close neighbours in the grass carp genome, but that remains to be determined.

Although trans-species allelic variation in classical MHC class I is also known for mammals (Figueroa et al. 1988), in teleost fish, this variation is substantially older and includes α1 and α2 domain lineages that do not homogenise with each other and which are hundreds of millions of years old (Shum et al. 2001; Kiryu et al. 2005; Grimholt et al. 2015). Their conservation in evolution implies that these domain lineages confer properties that are under balancing selection (selection for variation within a population) and that are not easily replicated by point mutations in sequences belonging to other lineages. But it is still not known what those properties are. Some lineages like α1-II appear to be characterised by unique peptide-binding features, but for other lineages like α1-V such features are not obvious (Fig. 4; Grimholt et al. 2015).

The first realisation that in teleost fish unusual deep lineages of classical MHC class I variation can be found was by Hansen et al. (1996). After Grimholt et al. (1993) had found classical MHC class I in Atlantic salmon with an α2 sequence belonging to domain lineage α2-I, Hansen et al. (1996) identified sequences in rainbow trout (another salmonid fish) with α2 domains belonging to that same α2-I lineage, but also sequences with a "carp-type" α2 domain that belong to what we now call the α2-II lineage. The ancestors of carp and salmon separated > 200 million years ago, implying that these lineages are ancient (see also Fig. 2). The α2-I versus α2-II designation derives from the study by Aoyagi et al. (2002), who were the first to provide evidence, in rainbow trout, that these ancient lineages could be used in allelic fashion. Whereas definitions for most α1 domain lineages and for the α2-II lineage follow readily from clear clustering borders observed upon phylogenetic tree analysis, the α2-I lineage shows wide variation and there have been several attempts to change the lineage definition to smaller subsets without clear borders in phylogenetic trees (e.g. Kiryu et al. 2005; Yang et al. 2006). In the present study, we like to reintroduce the simple α2-I versus α2-II nomenclature, based on our phylogenetic clustering analysis (Fig. 2), and distinct sequence differences (Figs. 1 and S1b). Based on database sequence analysis, we could find both α2-I and α2-II sequences already at the level of spotted gar (Figs. 2 and S1b), a primitive bony fish with a last common ancestor with teleost fishes > 300 million years ago. This is the first time that the age of the teleost fish classical MHC class I α2 lineages can be traced back so far in evolution, which is even a bit further back than has been done for the α1 lineages so far (Grimholt et al. 2015). Within bony fish evolution, the α2-II lineage appears to represent the more conservative sequences, as also exemplified by the lengths of the arms in Fig. 2 phylogenetic tree. Motif differences between α2-I and α2-II mostly concern residues which appear to have been newly acquired or lost in α2-I compared to classical MHC class I in more primitive ray-finned fish, and consist of the acquisition of a tryptophan at or near the end of β-strand S1 and an extra residue in the S1S2 loop, and the loss of an ancestral pair of interacting residues K125 and E159 (Figs. 1, 4, S1b). These residues are all positioned at the outer part of the α2 domain structure (Fig. 4b), and are not expected to directly affect peptide ligand binding or interaction with the α3 or β2m domains. We do not have a reasonable speculation for explaining this motif at the functional level.

The distribution of Ctid-UAA and Ctid-UBA allelic variation by point mutations that are expected to affect the peptide-binding groove suggests that the presence of allelic lineages lessens the need for variation by point mutations (Fig. 5).

In summary, our study makes an important contribution to the elucidation of classical MHC class I variation in teleost fish. It concludes an important role for "ancient domain lineage swapping" for creating the allelic variation in two different genes. Future studies should try to understand why these lineages are maintained for such long evolutionary periods.

Author's contributions Conceived and designed the experiments: N Zhang, and C Xia. To assist the sampling: L Ma, Z Qu, X Wei, and M Tang. Gene clone and analysis: Z Li, Y Jiang, N Zhang, and Z Liu. Provided the resources: M Tang and C Xia. Wrote the paper: Z Li. Reviewed the paper: N Zhang and C Xia. All authors approved the final version of the paper.

Funding information This work was supported by the 973 Project of the China Ministry of Science and Technology (Grant 2013CB835302), and the National Natural Science Foundation of China (Grant 31572653).

References

Aoyagi K, Dijkstra JM, Xia C, Denda I, Ototake M, Hashimoto K, Nakanishi T (2002) Classical MHC class I genes composed of highly divergent sequence lineages share a single locus in rainbow trout (Oncorhynchus mykiss). J Immunol 168:260–273

Bingulac-Popovic J, Figueroa F, Sato A, Talbot WS, Johnson SL, Gates M, Postlethwait JH, Klein J (1997) Mapping of mhc class I and class II regions to different linkage groups in the zebrafish, Danio rerio. Immunogenetics 46:129–134

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature 329:512–518

Chen W, Jia Z, Zhang T, Zhang N, Lin C, Gao F, Wang L, Li X, Jiang Y, Li X, Gao GF, Xia C (2010) MHC class I presentation and regulation by IFN in bony fish determined by molecular analysis of the class I locus in grass carp. J Immunol 185:2209–2221

Chen Z, Zhang N, Qi J, Chen R, Dijkstra JM, Li X, Wang Z, Wang J, Wu Y, Xia C (2017) The structure of the MHC class I molecule of bony fishes provides insights into the conserved nature of the antigen-presenting system. J Immunol 199:3668–3678

Dijkstra JM, Kiryu I, Yoshiura Y, Kumánovics A, Kohara M, Hayashi N, Ototake M (2006) Polymorphism of two very similar MHC class Ib loci in rainbow trout (Oncorhynchus mykiss). Immunogenetics 58: 152–167

Dijkstra JM, Katagiri T, Hosomichi K, Yanagiya K, Inoko H, Ototake M, Aoki T, Hashimoto K, Shiina T (2007) A third broad lineage of major histocompatibility complex (MHC) class I in teleost fish; MHC class II linkage and processed genes. Immunogenetics 59: 305–321

Dijkstra JM, Grimholt U, Leong J, Koop BF, Hashimoto K (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

Dirscherl H, McConnell SC, Yoder JA, de Jong JL (2014) The MHC class I genes of zebrafish. Dev Comp Immunol 46:11–23

Fan S, Wang Y, Wang S, Wang X, Wu Y, Li Z, Zhang N, Xia C (2018) Polymorphism and peptide-binding specificities of porcine major histocompatibility complex (MHC) class I molecules. Mol Immunol 93:236–245

Figueroa F, Günther E, Klein J (1988 Sep 15) MHC polymorphism predating speciation. Nature. 335(6187):265–267

Grimholt U, Hordvik I, Fosse VM, Olsaker I, Endresen C, Lie O (1993) Molecular cloning of major histocompatibility complex class I cDNAs from Atlantic salmon (Salmo salar). Immunogenetics 37: 469–473

Grimholt U, Tsukamoto K, Azuma T, Leong J, Koop BF, Dijkstra JM (2015) A comprehensive analysis of teleost MHC class I sequences. BMC Evol Biol 15:32

Hansen JD, Strassburger P, Du Pasquier L (1996) Conservation of an alpha 2 domain within the teleostean world, MHC class I from the rainbow trout Oncorhynchus mykiss. Dev Comp Immunol 20:417–425

Hansen JD, Strassburger P, Thorgaard GH, Young WP, Du Pasquier L (1999) Expression, linkage, and polymorphism of MHC-related genes in rainbow trout, Oncorhynchus mykiss. J Immunol 163: 774–786

Hashimoto K, Nakanishi T, Kurosawa Y (1990) Isolation of carp genes encoding major histocompatibility complex antigens. Proc Natl Acad Sci U S A 87:6863–6867

Hill AV (2001) The genomics and genetics of human infectious disease susceptibility. Annu Rev Genomics Hum Genet 2:373–400

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch GJ, White S, Chow W, Kilian B, Quintais LT, Guerra-Assuncao JA, Zhou Y, Gu Y, Yen J, Vogel JH, Eyre T, Redmond S, Banerjee R, Chi J, Fu B, Langley E, Maguire SF, Laird GK, Lloyd D, Kenyon E, Donaldson S, Sehra H, Almeida-King J, Loveland J, Trevanion S, Jones M, Quail M, Willey D, Hunt A, Burton J, Sims S, McLay K, Plumb B, Davis J, Clee C, Oliver K, Clark R, Riddle C, Elliot D, Threadgold G, Harden G, Ware D, Begum S, Mortimore B, Kerry G, Heath P, Phillimore B, Tracey A, Corby N, Dunn M, Johnson C, Wood J, Clark S, Pelan S, Griffiths G, Smith M, Glithero R, Howden P, Barker N, Lloyd C, Stevens C, Harley J, Holt K, Panagiotidis G, Lovell J, Beasley H, Henderson C, Gordon D, Auger K, Wright D, Collins J, Raisen C, Dyer L, Leung K, Robertson L, Ambridge K, Leongamornlert D, McGuire S, Gilderthorp R, Griffiths C, Manthravadi D, Nichol S, Barker G et al (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kiryu I, Dijkstra JM, Sarder RI, Fujiwara A, Yoshiura Y, Ototake M (2005) New MHC class Ia domain lineages in rainbow trout (Oncorhynchus mykiss) which are shared with other fish species. Fish Shellfish Immunol 18:243–254

Kruiswijk CP, Hermsen TT, Westphal AH, Savelkoul HF, Stet RJ (2002) A novel functional class I lineage in zebrafish (Danio rerio), carp (Cyprinus carpio), and large barbus (Barbus intermedius) showing an unusual conservation of the peptide binding domains. J Immunol 169:1936–1947

Lukacs MF, Harstad H, Bakke HG, Beetz-Sargent M, McKinnel L, Lubieniecki KP, Koop BF, Grimholt U (2010) Comprehensive analysis of MHC class I genes from the U-, S-, and Z-lineages in Atlantic salmon. BMC Genomics 11:154

Matsumura M, Fremont DH, Peterson PA, Wilson IA (1992 Aug 14) Emerging principles for the recognition of peptide antigens by MHC class I molecules. Science. 257(5072):927–934

McConnell SC, Restaino AC, de Jong JL (2014) Multiple divergent haplotypes express completely distinct sets of class I MHC genes in zebrafish. Immunogenetics 66:199–213

McConnell SC, Hernandez KM, Wcisel DJ, Kettleborough RN, Stemple DL, Yoder JA, Andrade J, de Jong JL (2016) Alternative haplotypes of antigen processing genes in zebrafish diverged early in vertebrate evolution. Proc Natl Acad Sci U S A 113:E5014–E5023

Michalova V, Murray BW, Sultmann H, Klein J (2000) A contig map of the Mhc class I genomic region in the zebrafish reveals ancient synteny. J Immunol 164:5296–5305

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

Ono H, Klein D, Vincek V, Figueroa F, O'HUigin C, Tichy H, Klein J (1992) Major histocompatibility complex class II genes of zebrafish. Proc Natl Acad Sci U S A 89:11886–11890

Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43:D423–D431

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sambrook JG, Figueroa F, Beck S (2005) A genome-wide survey of Major Histocompatibility Complex (MHC) genes and their paralogues in zebrafish. BMC Genomics 6:152

Shiina T, Dijkstra JM, Shimizu S, Watanabe A, Yanagiya K, Kiryu I, Fujiwara A, Nishida-Umehara C, Kaba Y, Hirono I, Yoshiura Y, Aoki T, Inoko H, Kulski JK, Ototake M (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

Shum BP, Guethlein L, Flodin LR, Adkison MA, Hedrick RP, Nehring RB, Stet RJ, Secombes C, Parham P (2001) Modes of salmonid MHC class I and II evolution differ from the primate paradigm. J Immunol 166:3297–3308

Stet RJ, Kruiswijk CP, Saeij JP, Wiegertjes GF (1998) Major histocompatibility genes in cyprinid fishes: theory and practice. Immunol Rev 166:301–316

Takeuchi H, Figueroa F, O'hUigin C, Klein J (1995) Cloning and characterization of class I Mhc genes of the zebrafish, *Brachydanio rerio*. Immunogenetics 42(2):77–84

Townsend AR, Gotch FM, Davey J (1985) Cytotoxic T cells recognize fragments of the influenza nucleoprotein. Cell 42:457–467

van Erp SH, Dixon B, Figueroa F, Egberts E, Stet RJ (1996) Identification and characterization of a new major histocompatibility complex class I gene in carp (Cyprinus carpio L.). Immunogenetics. 44(1): 49–61

Vandiedonck C, Knight JC (2009) The human major histocompatibility complex as a paradigm in genomics research. Brief Funct Genomic Proteomic 8:379–394

Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, Lu H, Huang R, Xia X, Feng Q, Liang X, Liu K, Zhang L, Lu T, Huang T, Fan D, Weng Q, Zhu C, Lu Y, Li W, Wen Z, Zhou C, Tian Q, Kang X, Shi M, Zhang W, Jang S, Du F, He S, Liao L, Li Y, Gui B, He H, Ning Z, Yang C, He L, Luo L, Yang R, Luo Q, Liu X, Li S, Huang W, Xiao L, Lin H, Han B, Zhu Z (2015) The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation. Nat Genet 47:625–631

Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 132:211–250

Xia C, Kiryu I, Dijkstra JM, Azuma T, Nakanishi T, Ototake M (2002 Apr) Differences in MHC class I genes between strains of rainbow trout (Oncorhynchus mykiss). Fish Shellfish Immunol 12(4):287–301

Yang TY, Hao HF, Jia ZH, Chen WH, Xia C (2006) Characterisation of grass carp (Ctenopharyngodon idellus) MHC class I domain lineages. Fish Shellfish Immunol 21:583–591

Zhang L, Lin D, Yu S, Bai J, Jiang W, Su W, Huang Y, Yang S, Wu J (2019) Polymorphism of duck MHC class molecules. Immunogenetics 71:49–59