- Supporting Information -

Alternative haplotypes of antigen processing genes in zebrafish diverged early in vertebrate evolution

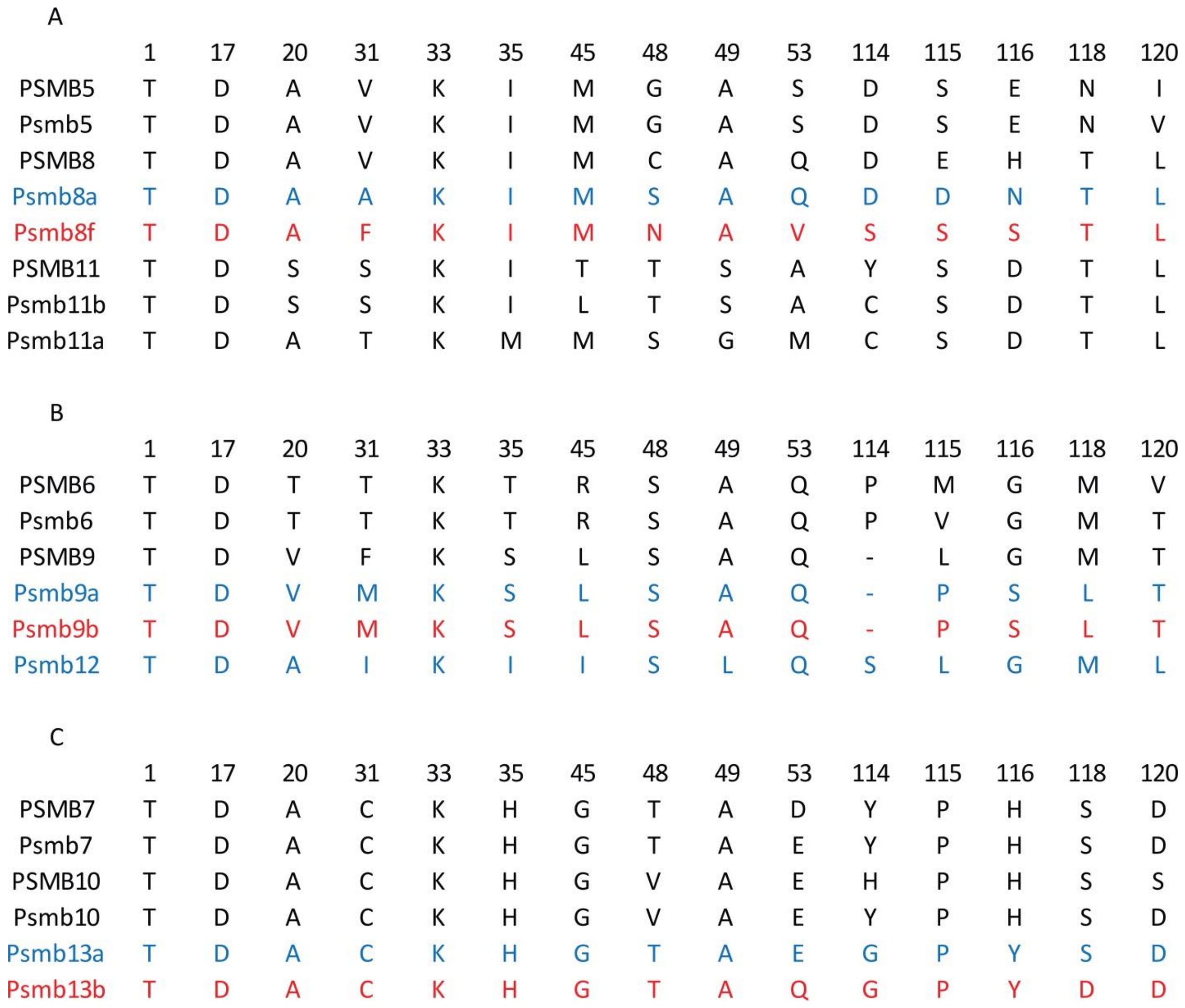
Sean C. McConnell[*], Kyle M. Hernandez, Dustin J. Wcisel, Ross N. Kettleborough, Derek L. Stemple, Jeffrey A. Yoder, Jorge Andrade, and Jill L.O. de Jong[*]

***Corresponding authors:**

Sean C. McConnell
email: scmcconnell@uchicago.edu

Jill L.O. de Jong
email: jdejong@peds.bsd.uchicago.edu

Contents



A

	1	17	20	31	33	35	45	48	49	53	114	115	116	118	120
PSMB5	T	D	A	V	K	I	M	G	A	S	D	S	E	N	I
Psmb5	T	D	A	V	K	I	M	G	A	S	D	S	E	N	V
PSMB8	T	D	A	V	K	I	M	C	A	Q	D	E	H	T	L
Psmb8a	T	D	A	A	K	I	M	S	A	Q	D	D	N	T	L
Psmb8f	T	D	A	F	K	I	M	N	A	V	S	S	S	T	L
PSMB11	T	D	S	S	K	I	T	T	S	A	Y	S	D	T	L
Psmb11b	T	D	S	S	K	I	L	T	S	A	C	S	D	T	L
Psmb11a	T	D	A	T	K	M	M	S	G	M	C	S	D	T	L

B

	1	17	20	31	33	35	45	48	49	53	114	115	116	118	120
PSMB6	T	D	T	T	K	T	R	S	A	Q	P	M	G	M	V
Psmb6	T	D	T	T	K	T	R	S	A	Q	P	V	G	M	T
PSMB9	T	D	V	F	K	S	L	S	A	Q	-	L	G	M	T
Psmb9a	T	D	V	M	K	S	L	S	A	Q	-	P	S	L	T
Psmb9b	T	D	V	M	K	S	L	S	A	Q	-	P	S	L	T
Psmb12	T	D	A	I	K	I	I	S	L	Q	S	L	G	M	L

C

	1	17	20	31	33	35	45	48	49	53	114	115	116	118	120
PSMB7	T	D	A	C	K	H	G	T	A	D	Y	P	H	S	D
Psmb7	T	D	A	C	K	H	G	T	A	E	Y	P	H	S	D
PSMB10	T	D	A	C	K	H	G	V	A	E	H	P	H	S	S
Psmb10	T	D	A	C	K	H	G	V	A	E	Y	P	H	S	D
Psmb13a	T	D	A	C	K	H	G	T	A	E	G	P	Y	S	D
Psmb13b	T	D	A	C	K	H	G	T	A	Q	G	P	Y	D	D

Figure S1. Predicted peptide cleavage site residues selected from the alignment of human and zebrafish proteasome subunits.

Names for the human proteasome subunits are labeled in uppercase; zebrafish subunits are labeled with both uppercase and lowercase according to standard nomenclature and introduced names (Table S5). Sequences from zebrafish core MHC haplotype 19B (Zv9 reference genome) are highlighted in blue; sequences from core MHC haplotype 19D (CG2 clonal zebrafish) are highlighted in red. Sequences are provided in Dataset S1.

A) The predicted cleavage site residues of non-MHC linked proteasome subunits are mostly well-conserved between humans and zebrafish, e.g. PSMB5 shares 14 of 15 residues with psmb5, indicating highly-conserved functions for these genes. Positions 1, 17 and 33 are completely conserved throughout all subunits, representing catalytic residues in the active site. Substitutions at positions 31 and 53 in particular have been proposed to alter peptide cleavage specificities for Psmb8 subunits with divergent lineages. The Psmb8a subunit (chymotrypsin-like catalytic activity) is found in the MHC haplotypes of most species examined including humans. The Psmb8f subunit appears to be missing from mammals, but is found in many other representative vertebrates. Altered activity of the Psmb8f subunit, as a consequence of these substitutions (predicted elastase-like catalytic activity), may thus contribute to functionally-specialized antigen processing pathways in organisms such as zebrafish.

B) In contrast, predicted cleavage specificity of the Psmb9b subunit encoded by on haplotype D is identical to the Psmb9a subunit encoded by haplotype B. In addition to carrying *psmb8a, psmb9a*, and *psmb13a* genes, the reference genome haplotype B also carries a *psmb12* gene. Unlike the divergent *psmb8f*, *psmb9b*, and *psmb13b* genes, the *psmb12* gene is missing from the divergent haplotype D assembly.

C) Position 53 may play a role in modulating peptide cleavage specificity among divergent zebrafish Psmb13 subunits, as the divergent Psmb13b subunit encoded by haplotype D maintains a unique substitution at position E53Q predicted to reduce trypsin-like cleavage specificity. Position 118 also has another interesting substitution within Psmb13b, S118D, but whether this could help compensate for any loss of charge within the cleavage site from the E53Q substitution remains to be determined.

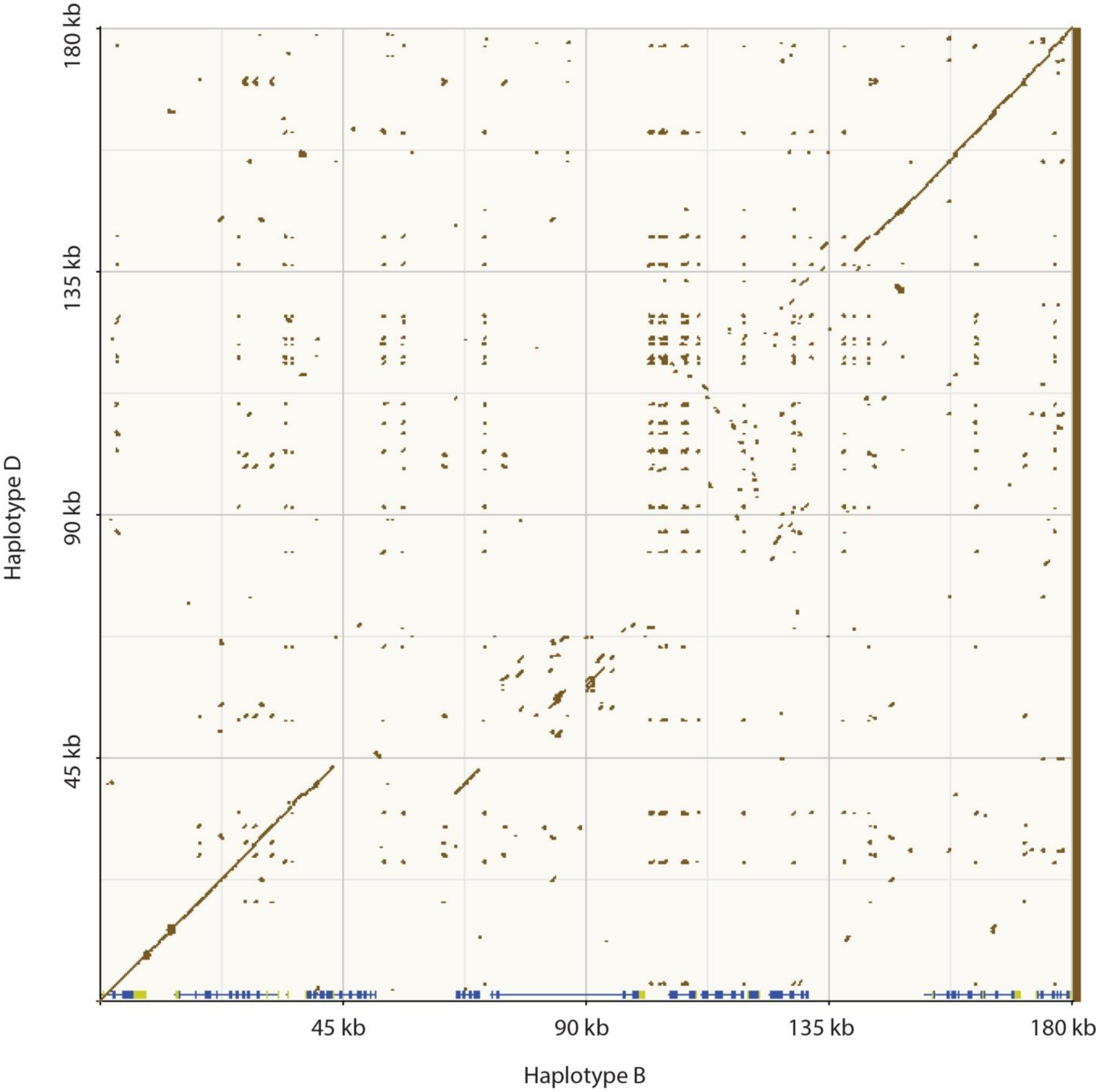


Figure S2. Dotplot comparison of the zebrafish core MHC haplotype 19D assembly with haplotype 19B from the zebrafish reference genome. The dotplot was generated using zPicture (http://zpicture.dcode.org) with sequences from the haplotype 19D assembly (CG2v1.0, this study) compared with sequences from reference haplotype 19B (Zv9 coordinates chr19:7623109:7802660). Exon locations for genes from the reference assembly are highlighted in blue, centered on *mhc1uba* intron 2 positioned at 90 kb.

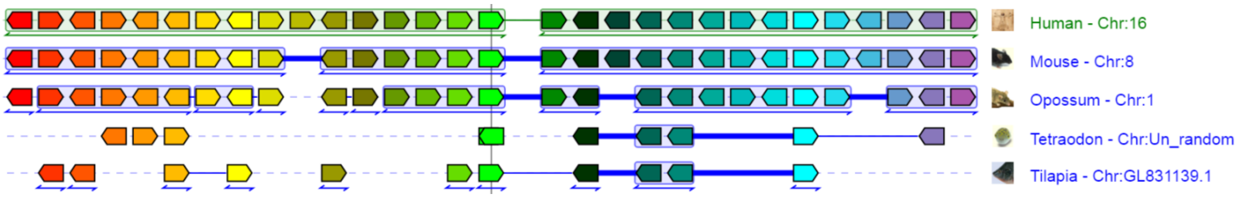


Figure S3. Conserved synteny of the *psmb10* gene outside the core MHC region in jawed vertebrates.

The *psmb10* gene (also called *LMP-10* or *MECL-1*) is the only immunoproteasome gene not linked to the core MHC region. Conserved synteny for the *psmb10* gene (highlighted by the green arrowheads joined by a vertical line in the center) is maintained for some jawed vertebrate species, including sarcopterygii and teleosts. This conserved synteny includes flanking genes *smpd3*, *prmt7*, *slc7a6*, *pla2g15*, *lcat*, *pskh1*, *edc4*, *nutf2*, and *ranbp10*. The region surrounding the *psmb10* gene in some species such as *psmb10* on zebrafish (*Danio rerio*) chromosome 4 appears to have been more highly rearranged (not shown), compared with *psmb10* genomic sequences in other teleost species such as tilapia (*Oreochromis niloticus*), where neighboring syntenic genes for the *psmb10* gene are more readily identified (bottom line). Conserved synteny analysis was performed in 'AlignView' using the Genomicus tool (http://www.genomicus.biologie.ens.fr/), which is based on genomic sequences and annotation as found in Ensembl.

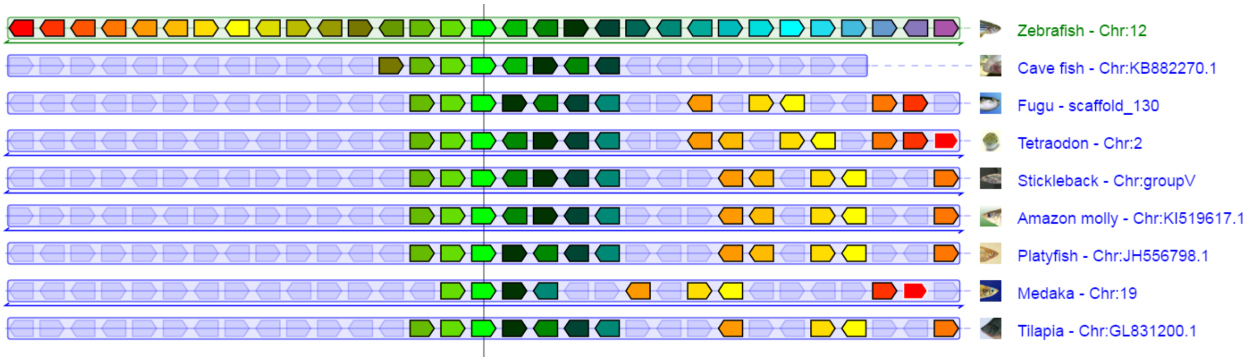


Figure S4. Conserved synteny of the *tap2t* gene outside the core MHC region in teleosts.

Unlike other transporter associated with antigen processing (TAP) genes, the *tap2t* gene was identified only in teleosts. Teleost species such as zebrafish (*Danio rerio*), fugu (*Takifugu rubripes*), stickleback (*Gasterosteus aculeatus*) and medaka (*Oryzias latipes*) maintain conserved synteny surrounding the *tap2t* gene. This conserved synteny surrounding *tap2t* (highlighted by the green arrowheads joined by a vertical line in the center) includes the *kcdt2*, *slc16a5*, *chad*, *acsf2*, *armc7*, *ndel1a*, *srcap*, *bc2*, *ttll6*, and *hoxb6b* genes. Conserved synteny analysis was performed using the Genomicus tool (http://www.genomicus.biologie.ens.fr/), which is based on genomic sequences and annotation as found in Ensembl.

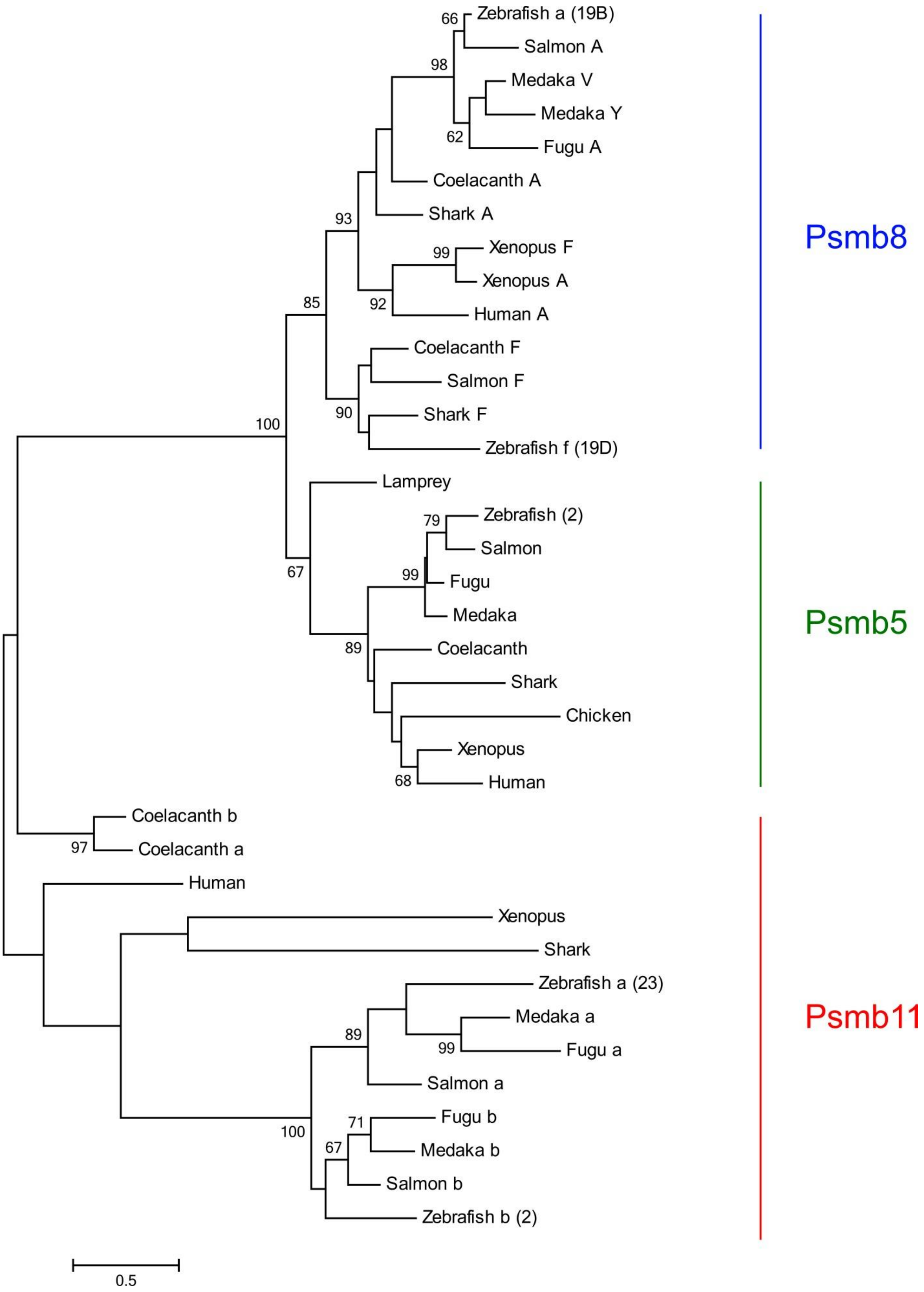

66
Zebrafish a (19B)
Salmon A
98
Medaka V
Medaka Y
62
Fugu A
Coelacanth A
Shark A
93
99
Xenopus F
Xenopus A
85
92
Human A
Coelacanth F
Salmon F
90
Shark F
100
Zebrafish f (19D)
Lamprey
79
Zebrafish (2)
Salmon
67
99
Fugu
Medaka
89
Coelacanth
Shark
Chicken
Xenopus
68
Human
Coelacanth b
97
Coelacanth a
Human
Xenopus
Shark
Zebrafish a (23)
Medaka a
89
99
Fugu a
Salmon a
71
Fugu b
100
67
Medaka b
Salmon b
Zebrafish b (2)
Psmb8
Psmb5
Psmb11
0.5

Figure S5. Phylogenetic tree for Psmb5, Psmb8*,* and Psmb11 predicted amino acid sequences from representative vertebrates.

Three major lineage branches are highlighted. Subunits from all three major lineage branches, including constitutive β5 subunit Psmb5, inducible β5i subunit Psmb8, and thymoproteasome β5t subunit Psmb11, are highly conserved across jawed vertebrates. Jawless vertebrates (e.g. lamprey) only have the constitutive Psmb5 subunit. Only jawed vertebrate species with MHC-based immunity have been shown to have additional (non-constitutive) proteasome subunit genes, reflecting the function of these genes.

The *psmb8* gene has two distinct forms in zebrafish associated with alternative haplotypes; *psmb8a* is present in the zebrafish reference genome chromosome 19 core MHC haplotype B, while *psmb8f* is expressed from core MHC haplotype D. These *psmb8* forms are ancient, being conserved for approximately 500 million years in sharks, some teleosts, and coelacanths. Interestingly, the *psmb8f* form appears to have been lost from most tetrapods and other teleosts (1). However, *psmb8f*-like forms appear to have been re-derived in some teleost lineages, as in medaka via a V31Y substitution. Furthermore, the *psmb8f* form appears to have been eroded in Xenopus, by sequence exchange events on either side of the bulky cleavage site residue A31F, making this *psmb8f* gene appear more *psmb8a*-like (2). In zebrafish, the Psmb8f subunit has the bulky A31F, and also a Q53V substitution predicted to alter peptide cleavage specificity. These observations indicate that the alternative forms of Psmb8 subunits conserve ancient, distinctive functions.

In another branch, the β5t subunit Psmb11 maintains two duplicate genes that appear to be specific to teleosts, diverging on separate chromosomes. These genes are predicted to have distinct functions, with *psmb11b* more conserved than the *psmb11a* copy that is found adjacent to the *psmb5* gene (3). Coelacanths also maintain two *psmb11* copies, but these appear on the same chromosome and are much more closely related, thus appearing to be the result of a more recent tandem duplication event.

The Maximum Likelihood method based on the JTT matrix-based model was used to construct the phylogenetic tree within the MEGA6 program. To model evolutionary rate differences among sites, a discrete Gamma distribution was used (5 categories, +*G* parameter = 0.5036), which allowed some sites to be evolutionarily invariable (0.0000% sites). Tree is drawn to scale, and branch lengths represent the number of substitutions per site. Bootstrap values greater than or equal to 60% are provided next to the branches, as calculated using 500 replicate trees. Positions with less than 95% site coverage were eliminated, providing a total of 234 positions for the 37 sequences in the dataset. For zebrafish genes, chromosome number is provided in parentheses, while for chromosome 19 core MHC genes a specific haplotype suffix is also provided.

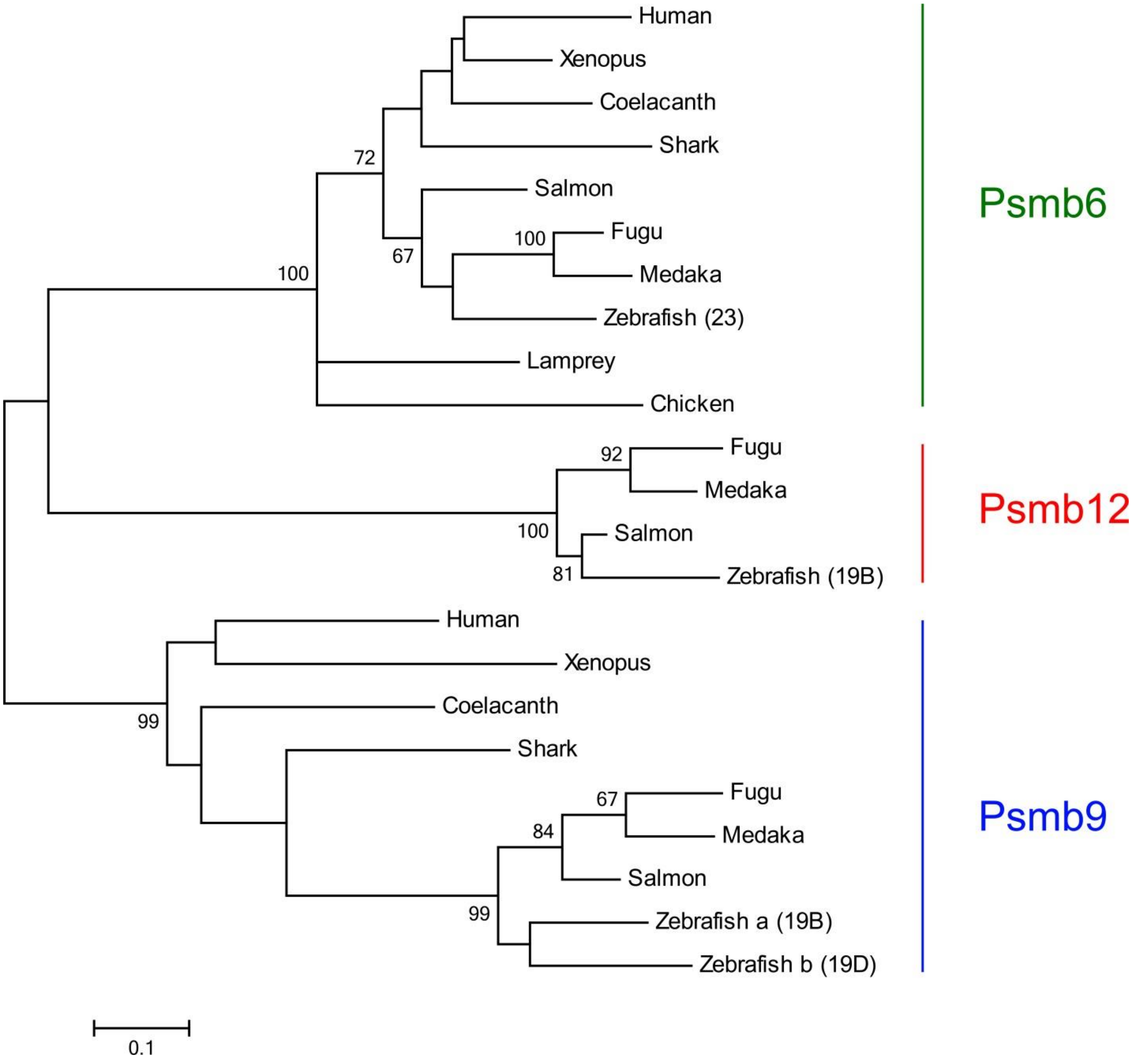

Human
Xenopus
Coelacanth
Shark
72
Salmon
Psmb6
100
Fugu
67
Medaka
100
Zebrafish (23)
Lamprey
Chicken
92
Fugu
Medaka
Psmb12
100
Salmon
81
Zebrafish (19B)
Human
Xenopus
Coelacanth
99
Shark
67
Fugu
84
Medaka
Psmb9
Salmon
99
Zebrafish a (19B)
Zebrafish b (19D)
0.1

Figure S6. Phylogenetic tree for Psmb6, Psmb9, and Psmb12 predicted amino acid sequences from representative vertebrates.

Three major lineage branches are highlighted. Both β1 subunit Psmb6, and β1i subunit Psmb9, are highly conserved across jawed vertebrates. Jawless vertebrates (e.g. lamprey) only have constitutive Psmb6. The β1t subunit Psmb12 appears to be teleost-specific as it is not found in other vertebrate lineages. The *psmb12* gene undergoes presence/absence variation in zebrafish, e.g. while present in the zebrafish reference genome chromosome 19 core MHC haplotype B, it is missing from core MHC haplotype D.

The Maximum Likelihood method based on the JTT matrix-based model was used to construct the phylogenetic tree within the MEGA6 program. To model evolutionary rate differences among sites, a discrete Gamma distribution was used (5 categories, +*G* parameter = 0.9091), which allowed some sites to be evolutionarily invariable (5.2118% sites). Tree is drawn to scale, and branch lengths represent the number of substitutions per site. Bootstrap values greater than or equal to 60% are provided next to the branches, as calculated using 500 replicate trees. Positions with less than 95% site coverage were eliminated, providing a total of 214 positions for the 23 sequences in the dataset. For zebrafish genes, chromosome number is provided in parentheses, while for chromosome 19 core MHC genes a specific haplotype suffix is also provided.

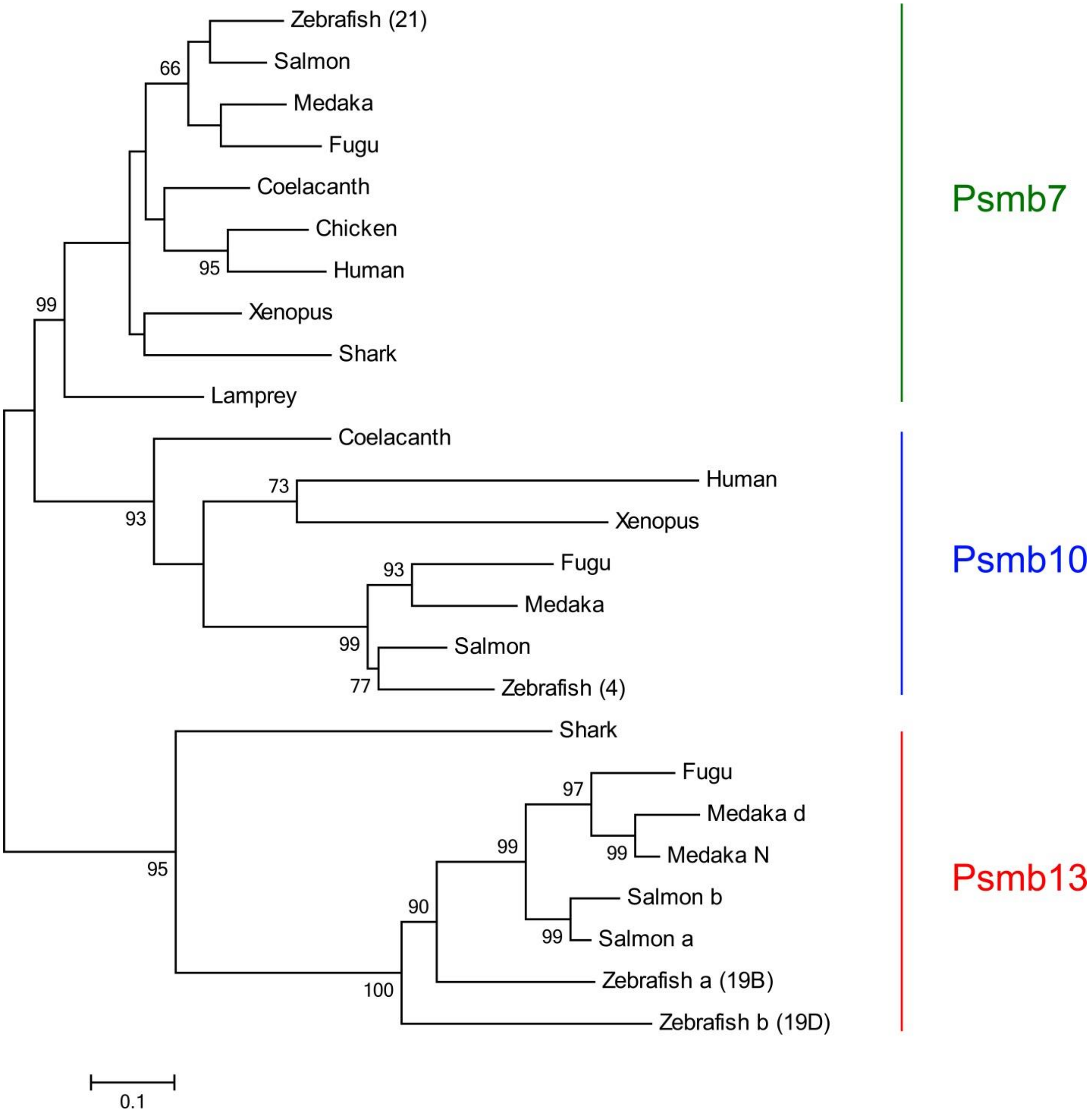

Zebrafish (21)
Salmon
66
Medaka
Fugu
Coelacanth
Chicken
95
Human
99
Xenopus
Shark
Lamprey
Coelacanth
73
Human
93
Xenopus
93
Fugu
Medaka
99
Salmon
77
Zebrafish (4)
Shark
Fugu
97
Medaka d
99
99
Medaka N
95
Salmon b
90
99
Salmon a
Zebrafish a (19B)
100
Zebrafish b (19D)
0.1
Psmb7
Psmb10
Psmb13

Figure S7. Phylogenetic tree for Psmb7, Psmb10, and Psmb13 predicted amino acid sequences from representative vertebrates.

Three major lineage branches are highlighted. The β2 subunit Psmb7, as well as the β2i subunit Psmb10, are both highly conserved across most jawed vertebrates. Jawless vertebrates such as lamprey only have the constitutive Psmb7. The β2t subunit Psmb13 was not found in some other vertebrate lineages and appears to be specific to teleost and sharks. The *psmb13* gene has two forms in zebrafish associated with alternative haplotypes; *psmb13a* is present in the zebrafish reference genome chromosome 19 core MHC haplotype B, while *psmb13b* is expressed from core MHC haplotype D. The divergent Psmb13b zebrafish proteasome subunit has an E53Q substitution predicted to alter peptide cleavage specificity.

The Maximum Likelihood method based on the JTT matrix-based model was used to construct the phylogenetic tree within the MEGA6 program. To model evolutionary rate differences among sites, a discrete Gamma distribution was used (5 categories, +*G* parameter = 0.9279), which allowed some sites to be evolutionarily invariable (12.3292% sites). Tree is drawn to scale, and branch lengths represent the number of substitutions per site. Bootstrap values greater than or equal to 60% are provided next to the branches, as calculated using 500 replicate trees. Positions with less than 95% site coverage were eliminated, providing a total of 274 positions for the 25 sequences in the dataset. For zebrafish genes, chromosome number is provided in parentheses, while for chromosome 19 core MHC genes a specific haplotype suffix is also provided.

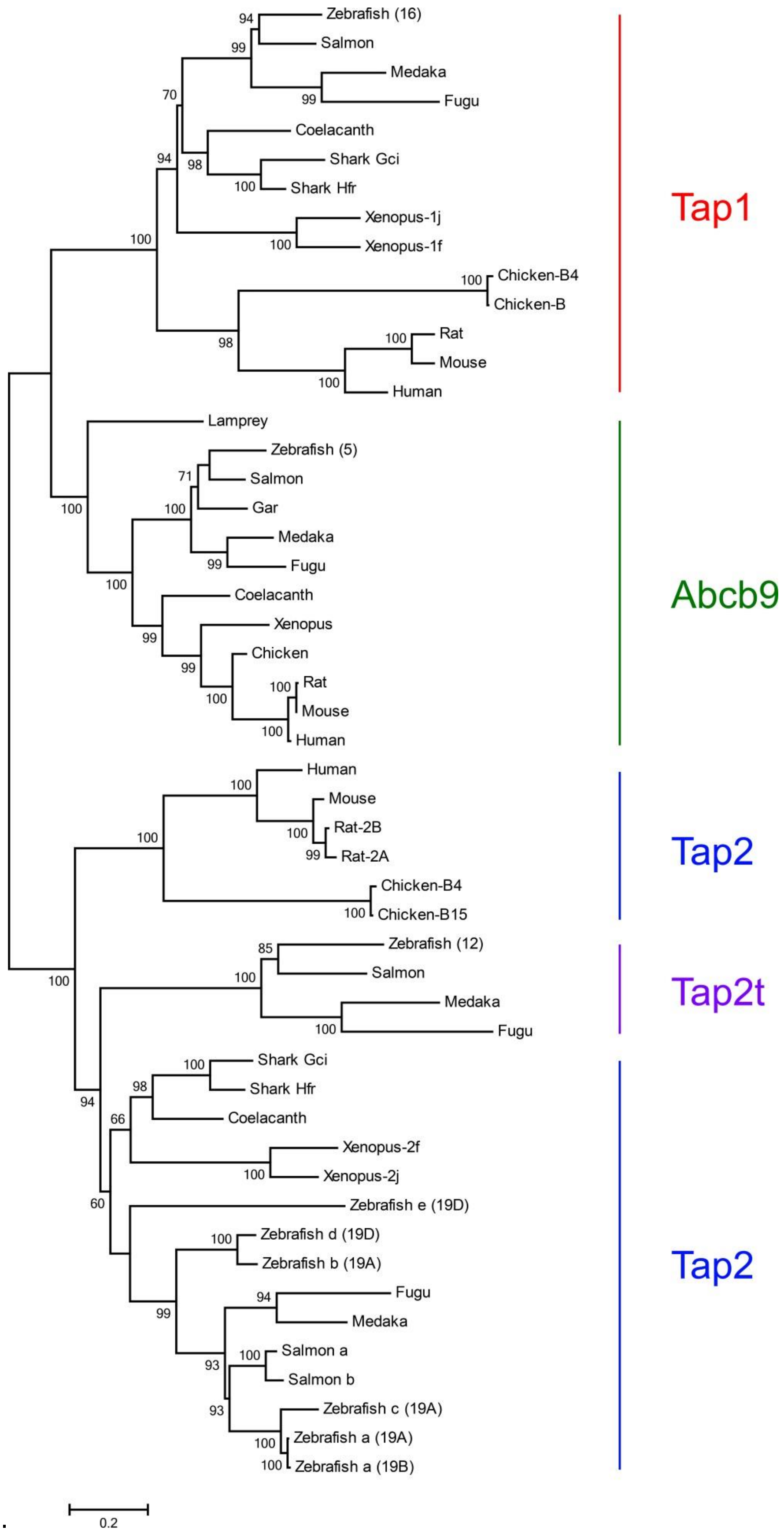

94
Zebrafish (16)
99
Salmon
Medaka
70
99
Fugu
Coelacanth
94
98
Shark Gci
100
Shark Hfr
Xenopus-1j
100
100
Xenopus-1f
100
Chicken-B4
Chicken-B
100
Rat
98
Mouse
100
Human
Lamprey
Zebrafish (5)
71
Salmon
100
100
Gar
Medaka
99
Fugu
100
Coelacanth
Xenopus
99
Chicken
99
100
Rat
100
Mouse
100
Human
Human
100
Mouse
100
100
Rat-2B
99
Rat-2A
Chicken-B4
100
Chicken-B15
85
Zebrafish (12)
100
Salmon
100
Medaka
100
Fugu
100
Shark Gci
98
Shark Hfr
94
66
Coelacanth
Xenopus-2f
100
Xenopus-2j
60
Zebrafish e (19D)
100
Zebrafish d (19D)
Zebrafish b (19A)
94
Fugu
99
Medaka
100
Salmon a
93
Salmon b
93
Zebrafish c (19A)
100
Zebrafish a (19A)
100
Zebrafish a (19B)
Tap1
Abcb9
Tap2
Tap2t
Tap2
0.2

Figure S8. Phylogenetic tree for Abcb9, Tap1*,* Tap2 and Tap2t predicted amino acid sequences across representative vertebrates.

Three major lineage branches are highlighted. The presumed ancestral Abcb9 subunit, as well as the derived heterodimeric Tap1 and Tap2 subunits, are each highly conserved across most jawed vertebrates. Jawless vertebrates such as lamprey only have the Abcb9 subunit. The Tap2t subunit was not found in other vertebrate lineages and appears to be specific to teleosts; Tap2t subunits from teleosts cluster as a distinctive branch within the larger Tap2 family. The *tap2* gene has five forms in zebrafish associated with alternative haplotypes. The *tap2a* gene is present in the zebrafish reference genome chromosome 19 core MHC haplotype B. The *tap2a*, *tap2b*, and *tap2c* are all found in core MHC haplotype A. In addition, *tap2d* and tap2e are expressed from core MHC haplotype D. The divergent zebrafish Tap2 subunits maintain several substitutions predicted to alter peptide cleavage specificity.

The Maximum Likelihood method based on the JTT matrix-based model was used to construct the phylogenetic tree within the MEGA6 program. To model evolutionary rate differences among sites, a discrete Gamma distribution was used (5 categories, +*G* parameter = 1.9675), which allowed some sites to be evolutionarily invariable (2.7601% sites). Tree is drawn to scale, and branch lengths represent the number of substitutions per site. Bootstrap values greater than or equal to 60% are shown are provided next to the branches, as calculated using 500 replicate trees. Positions with less than 95% site coverage were eliminated, providing a total of 560 positions for the 51 sequences in the dataset. For zebrafish genes, chromosome number is provided in parentheses, while for some chromosome 19 core MHC genes a specific haplotype letter suffix is also given. Three major lineage branches are highlighted.

RNA-Seq contig	Genomic scaffold(s)	Hap D gene	%ID	Hap B gene (best match)	Chr.
17561	13206	*daxx*	99	*daxx*	19
3129	13206	*tapbp*	98	*tapbp*	19
218	13206, 51738	*mhc1uga*	49	*mhc1uba*	19
10154	15837	*tap2d*	65	*tap2a*	19
11205	15837	*psmb9b*	86	*psmb9a*	19
2067	15837, 2546	*psmb13b*	71	*psmb13a*	19
818	2546	*psmb8f*	64	*psmb8a*	19
14104, 23407	2546	*tap2e*	50	*tap2a*	19
2621	2546	*brd2a*	100	*brd2a*	19
12208	2546	*hsd17b8*	99	*hsd17b8*	19

Table S1. Comparison of genes from the haplotype 19D assembly with genes from haplotype 19B from the zebrafish reference genome. For each gene found within the CG2 haplotype D assembly, RNA-Seq contigs and genomic scaffold identifiers are provided. Blue font highlights genes with high levels of pairwise percent identity (%ID) relative to the reference genome sequences, and red font highlights genes having divergent sequences relative to the reference genome (below 90%). Pairwise percent identity (%ID) was calculated with BLAST using predicted amino acid sequences for genes from haplotype 19D, compared with sequences from the most closely matched genes from reference haplotype 19B (Zv9).

RNA-Seq contig	Genomic scaffold(s)	CG2 gene	%ID	Zv9 (best match)	Chr.
4054	75501	*psmb5*	99	*psmb5*	2
12947	6167, 48866, 28756	*psmb6*	99	*psmb6*	23
6794	70674, 78660, C10987108	*psmb7*	100	*psmb7*	21
10534	18415, 26616	*psmb10*	100	*psmb10*	4
NA	75501	*psmb11a*	99	*psmb11a*	23
NA	26712	*psmb11b*	100	*psmb11b*	2
4392	86824	*abcb9*	99	*abcb9*	5
1051	16168, 7509, 70724	*tap1*	100	*tap1*	16
9121	10598, 26828	*tap2t*	99	*tap2t*	12
18099	17582	tapbpl	99	tapbpl	16
777	17862	*b2m*	100	*b2m*	4
NA	11566	*b2ml*	99	*b2ml*	8
15805	68974	cd8a	99	CD8a	21
32479	71708	cd8b	100	CD8b	7
6200	1788	*mhc2daa*	99	*mhc2daa*	8
2953	1788	*mhc2dab*	91	*mhc2dab*	8
14906	44652	*mhc2dbb*	98	*mhc2dbb*	18

Table S2. MHC-pathway related genes found outside the core MHC locus of the Zv9 reference genome compared with sequences within the CG2 clonal zebrafish. For each gene found within the CG2 genomic assembly, RNA-Seq contigs and genomic scaffold identifiers are provided. Blue font highlights genes with high levels of pairwise percent identity (%ID) relative to the reference genome sequences, and red font highlights genes having divergent sequences relative to the reference genome (below 90%). Pairwise percent identity (%ID) was calculated with BLAST using predicted amino acid sequences for genes from CG2 clonal zebrafish, compared with sequences from the most closely matched genes from zebrafish reference genome (Zv9). NA indicates that no expressed transcripts were found, as expected for the thymus-specific *psmb11* genes as no thymic tissue was included in the RNA preparation.

	215	216	**217**	218	219	220	260	261	**262**	263	264	265	**266**	267	268	269
Tap2a (19B)	M	C	**A**	I	N	S	M	G	**R**	A	V	A	**L**	N	V	N
Tap2a (19A)	M	C	**A**	I	N	S	M	G	**R**	A	V	A	**L**	N	V	N
Tap2c (19A)	T	C	**V**	I	Y	S	M	G	**D**	A	V	A	**L**	N	V	N
Tap2b (19A)	M	C	**T**	L	S	R	M	S	**Q**	S	V	A	**M**	N	V	N
Tap2d (19D)	M	L	**T**	I	S	R	M	S	**Q**	S	V	A	**M**	N	V	N
Tap2e (19D)	M	C	**T**	L	S	R	M	S	**R**	S	L	A	**A**	N	V	N
Tap2t (12)	M	F	**S**	L	N	R	M	G	**R**	A	V	A	**M**	N	V	N
Salmon-2t	M	C	**S**	L	S	R	M	G	**R**	S	V	A	**L**	N	S	N
Salmon-2a	M	C	**A**	I	N	S	M	G	**R**	A	V	A	**L**	N	V	N
Salmon-2b	M	C	**A**	I	N	S	M	A	**R**	A	L	A	**L**	N	V	N
Shark Hfr	M	F	**T**	M	S	R	M	S	**R**	S	I	G	**L**	N	V	N
Shark Gci	M	F	**T**	M	Y	R	M	S	**R**	S	I	A	**A**	N	V	N
Chick-B4	T	F	**I**	R	F	R	A	S	**N**	V	L	A	**L**	N	I	N
Chick-B15	T	F	**I**	G	F	R	A	S	**N**	V	L	T	**L**	N	I	N
Xenopus-2f	M	F	**S**	H	S	S	V	S	**R**	A	I	A	**A**	N	V	N
Xenopus-2j	M	F	**S**	L	A	R	V	S	**R**	S	I	A	**A**	N	V	N
Mouse	L	F	**T**	M	S	R	M	S	**R**	W	L	P	**F**	N	A	N
Rat-2B	L	F	**T**	M	S	R	M	S	**R**	W	L	P	**F**	N	A	N
Rat-2A	L	F	**A**	E	S	R	M	S	**Q**	W	L	S	**L**	N	A	N
Human	T	Y	**T**	M	S	R	M	S	**N**	W	L	P	**L**	N	A	N

Table S3. Selected residues of the specificity loop within Tap2 molecules from species with polymorphic alleles. An alignment of deduced amino acid sequences highlights three cis-modification (cim) residues at positions **217**, **262**, and **266** (above in bold) that have been shown to have functional roles in determining Tap2 peptide transport specificity (4). Zebrafish chromosome and core MHC haplotype identifiers are provided in parentheses (when applicable). Accession numbers and species identifiers are provided in Dataset S1.

Human

β5	β6	β7	
5	6	7	Constitutive
8	6	7	Intermediate I
8	9	7	Intermediate II
8	9	10	Immuno-
11	9	10	Thymo-

Zebrafish

β5	β6	β7	
5	6	7	Constitutive
8a	6	7	Intermediate 1
8f	6	7	Intermediate 2
8a	9a	7	Intermediate 3
8f	9a	7	Intermediate 4
8a	9b	7	Intermediate 5
8f	9b	7	Intermediate 6
8a	9a	10	Immuno- 1
8f	9a	10	Immuno- 2
8a	9b	10	Immuno- 3
8f	9b	10	Immuno- 4
8a	9a	13a	Immuno- 5
8f	9a	13a	Immuno- 6
8a	9b	13a	Immuno- 7
8f	9b	13a	Immuno- 8
8a	9a	13b	Immuno- 9
8f	9a	13b	Immuno- 10
8a	9b	13b	Immuno- 11
8f	9b	13b	Immuno- 12
11a	9a	10	Thymo- 1
11b	9a	10	Thymo- 2
11a	9b	10	Thymo- 3
11b	9b	10	Thymo- 4
11a	9a	13a	Thymo- 5
11b	9a	13a	Thymo- 6
11a	9b	13a	Thymo- 7
11b	9b	13a	Thymo- 8
11a	9a	13b	Thymo- 9
11b	9a	13b	Thymo- 10
11a	9b	13b	Thymo- 11
11b	9b	13b	Thymo- 12

Table S4. Proteasome subunit combinations in zebrafish. Distinct proteasome subunit compositions in teleosts may number as high as 30, even when constrained according to cooperative subunit assembly rules (5), which specify for example Psmb8 incorporation before Psmb9. This suggests many additional subunit combinations in zebrafish compared with humans, where only five combinations are known.

	Hom.	Het.	% Het.
CG1	396319	7254	1.80
CG2	533084	8618	1.59
WT1	37161	6332	14.56
WT2	57103	6381	10.05

Table S5. Comparison of SNP variant calls identified from whole exome sequencing. CG1 and CG2 double haploid clonal zebrafish samples were subjected to whole exome sequencing and analyzed for genetic variation. Pooled samples of each clonal line showed a substantial number of SNPs that varied from the reference genome (approximately 400,000-500,000), but the vast majority (>98%) were called as homozygous, approaching fixation of homozygous SNPs within each clonal line. Comparison of the clonal zebrafish heterozygous SNP percentages with the percentage of heterozygous SNPs identified from two individual wild-type samples (WT1 and WT2) showed that heterozygous SNPs are enriched by nearly an order of magnitude in the wild-type samples. This was observed even though each wild-type was sequenced and processed as an individual and not pooled, which may lead to fewer confident calls for each individual wild-type fish compared with four pooled clonal samples, especially for heterozygous SNPs due to fewer reads for each possible genotype. Of note, this SNP analysis likely over-estimates the genetic variation of the double haploid clonal lines, due to possible ambiguous mapping of paralogs or repeats, along with potential errors in sequencing from either genome. ‘Hom.’ represents number of homozygous SNPs, ‘Het.’ is the number of heterozygous SNPs, and ‘%Het.’ is the percentage of heterozygous SNPs when divided by the total, where Hom. + Het. = 100%.

Gene names	Subunit names	Additional Names	Previous names	ZFIN ID	Ensembl Gene	Chr. Haplo.
psmb5	beta5	LMPX, constitutive proteasome subunit β5	-	ZDB-GENE-990415-215	ENSDARG00000075445	2
psmb6	beta1	LMPY, delta, constitutive proteasome subunit β6	-	ZDB-GENE-990415-216	ENSDARG00000002240	23
psmb7	beta2	LMPZ, constitutive proteasome subunit β7	-	ZDB-GENE-001208-4	ENSDARG00000037962	21
psmb8a	beta5ia	LMP7A, immune-proteasome subunit 5ia	*psmb8*	ZDB-GENE-990415-141	ENSDARG00000001303	19 A
psmb8f	beta5if	LMP7F, immune-proteasome subunit β5if	-	ZDB-GENE-050417-319		19 D
psmb9a	beta1ia	LMP2A, immune-proteasome subunit β1ia	-	ZDB-GENE-990415-140	ENSDARG00000000656	19 A,B
psmb9b	beta1ib	LMP2B, immune-proteasome subunit β1ib	-	ZDB-GENE-001208-3		19 D
psmb10	beta2i	LMP10, MECL1, immune-proteasome subunit β2i	*PSMB10*	ZDB-GENE-040718-278	ENSDARG00000043781	4
psmb11a	beta5ta	Thymoproteasome subunit β5ta	-	ZDB-GENE-040724-32	ENSDARG00000078253	23
psmb11b	beta5tb	Thymoproteasome subunit β5tb	-		ENSDARG00000068086	2
psmb12	beta1t	LMP2/delta-like, teleost proteasome subunit β1t	*psmb9l, psmb11*	ZDB-GENE-001208-1	ENSDARG00000031885	19 A,B
psmb13a	beta2ta	PSMB7/10-like, teleost proteasome subunit β2ta	*psmb10, psmb12*	ZDB-GENE-001208-2	ENSDARG00000001656	19 A,B
psmb13b	beta2tb	PSMBb7/10-like, teleost proteasome subunit β2tb	-			19 D

Table S6. Zebrafish proteasome subunit nomenclature. Gene names approved by the zebrafish nomenclature committee are listed in the first column of this table, and names advanced by this study highlighted in bold (accommodating the novel genes). The 'Subunit names' and 'Additional names' columns both provide alternate identifiers. 'Previous names' refers to other potentially conflicting zebrafish nomenclature. The ZFIN and Ensembl database identifiers are listed for each gene when available. The final column provides chromosome (Chr.) and specific haplotype identifiers (Haplo.) for the zebrafish core MHC locus (when applicable).

Gene names	Additional Names	Previous names	ZFIN ID	Ensembl Gene	Chr. Haplo.
tap1	transporter associated with antigen processing (*TAP*) 1	*abcb2*	ZDB-GENE-050517-43	ENSDARG00000079766	16
tap2a	*tap2* subunit type a	*abcb3l1*	ZDB-GENE-030616-245	ENSDARG00000036787	19 A,B
tap2b	*tap2* subunit type b	*abcb3l2*	ZDB-GENE-030616-225		19 A
tap2c	*tap2* subunit type c	*abcb3*	ZDB-GENE-990415-260		19 A
tap2d	*tap2* subunit type d	-			19 D
tap2e	*tap2* subunit type e	-			19 D
tap2t	*tap2* subunit type t, teleost-specific	*tap2-like*	ZDB-GENE-130531-48	ENSDARG00000033446	12
tapbp	tap binding protein, tapasin, tpsn	-			19 A,D
tapbp.1	tap binding protein tandem duplicate 1, tapasin gene 1	*TAPBP*	ZDB-GENE-010110-2	ENSDARG00000045011	19 B
tapbp.2	tap binding protein tandem duplicate 2, tapasin gene 2	*tapbp*		ENSDARG00000079402	19 B

Table S7. Zebrafish TAP and *tapbp* gene nomenclature. Gene names approved by the zebrafish nomenclature committee are listed in the first column of this table, and names proposed during this study are highlighted in bold (accommodating novel genes). The 'Additional names' column provides alternative identifiers. 'Previous names' refers to other potentially conflicting zebrafish nomenclature. The ZFIN and Ensembl database identifiers are listed for each gene when available. The final column provides chromosome (Chr.) and specific haplotype identifiers (Haplo.) for the zebrafish core MHC locus (when applicable).

Psmb13_Salmon_(*Salmo_salar*)	ABQ59647
Psmb13_Fugu_(*Takifugu_rubripes*)	XP_003978905
Psmb13_Medaka_(*Oryzias_latipes*)	NP_001171882
Psmb13_Damselfish_(*Stegastes_partitus*)	XP_008301222
Psmb13_A_Zebrafish_(19B)	NP_571752
Psmb13_B_Zebrafish_(19D)	GDQH01002062.1

Table S8. Accession numbers for selected teleost Psmb13 subunits. For zebrafish (*Danio rerio*) sequence names, chromosome number and core MHC haplotype identifier are appended in parentheses.

HLA-A Specif.	C	FS T		**M**		L	Q	QH	V	S	**H**	I**S**	NH	A				
HLA-B Specif.	H	DH	PT	EG K**T**	H**F**	P	D	**I**T	C**F** SY	**R**I T	NK S		G**Y**		CG	NK	P	DH
HLA Common	**Y**	**Y**	**AS**	**M**	**Y**	**RQ** EG	**NE**	NK	M	A**G**	Q	**AT**	D	E**V**	**ND** S	**IT**	**AL**	Y
Position	7	9	24	45	59	62	63	66	67	69	70	73	74	76	77	80	81	84
Uaa	**Y**	**Y**	V	**T**	**Y**	**R**	**E**	**I**	**F**	**G**	A	V	F	N	**N**	V	I	R
Uba	A	**Y**	**A**	F	**Y**	**Q**	**Q**	**I**	L	**G**	Y	V	F	N	**N**	V	V	R
Uca	A	**Y**	**A**	F	**Y**	**Q**	**Q**	**I**	L	**R**	**H**	K	F	N	**N**	V	**A**	R
Uda	**Y**	**F**	**A**	S	**Y**	**R**	**Q**	L	A	**G**	Y	V	**Y**	**V**	**D**	**T**	**L**	R
Uea	**Y**	**Y**	V	**T**	**Y**	**R**	**E**	**I**	**F**	**G**	A	V	F	N	**N**	V	I	R
Ufa	V	**Y**	**A**	V	**F**	**R**	**N**	**I**	R	N	M	L	F	N	**N**	**I**	**A**	R
Uga	**Y**	**Y**	M	A	**Y**	I	**N**	R	L	**A**	T	**A**	F	N	**N**	V	**A**	R
Uha	A	**Y**	**A**	F	**Y**	**Q**	**E**	**I**	L	**G**	**H**	**S**	F	N	**N**	V	**A**	R
Uia	**Y**	I	**S**	**M**	**Y**	**R**	**E**	R	E	**G**	I	V	F	N	**N**	**I**	**A**	R
Uja	**Y**	**Y**	I	**M**	**F**	**R**	**N**	**I**	A	**G**	T	N	F	N	**N**	V	**A**	R
Uka	A	**Y**	**A**	F	**Y**	**Q**	**Q**	**I**	L	**R**	**H**	**T**	F	N	**N**	V	**A**	R

Table S9. Polymorphic peptide binding residues conserved between zebrafish and human MHCI molecules. MHC Class I (MHCI) sequence polymorphisms predicted to be in contact with peptides in the binding cleft are shown for human HLA-A and HLA-B molecules. Residues are listed as specific to either HLA-A or HLA-B alleles, or common to both, as described (6). Columns provide MHCI residue numbering according to IMGT nomenclature. Amino acid polymorphism at these peptide binding sites is highly pervasive for the 11 zebrafish MHCI molecules, while the majority (~62% found across 34 positions) of these zebrafish polymorphic residues are also shared with HLA-A and/or HLA–B. Shared polymorphisms at each position are highlighted in bold. Polymorphisms are shared between zebrafish and humans at all positions, with the exception of position 84, where the Y84R substitution is found throughout non-mammalian vertebrates. This overlap in substitutions across species indicates either shared ancestry for these alleles preserved across species, or continued exploration through a highly conserved stereo-chemical space. Additional residues are shown in Table S10.

HLA-A Specif.	**F**V	IK	H	**R**Q EP	V				V	RM	A**Q** S	E**S**	LP	PW	R	DI
HLA-B Specif.	W	NS T**W** **V**		N**DK**	L	C	IPS	G**L**		QT	**D**E **T**	T	CT	M	L	
HLA Common	**IL**	RM	CF S**Y**	H	N**D** HF S**Y**	**Y**	**T**	**W**	AP	**A**E VW	**R**L **W**	APV	**Y**	AR **E**G L**T**	GS**W**	HT**Y**
Position	95	97	99	114	116	118	143	147	150	152	156	158	159	163	167	171
Uaa	**F**	**W**	**Y**	M	**Y**	**Y**	**T**	**W**	T	**A**	**Q**	N	**Y**	**T**	**W**	**Y**
Uba	**F**	F	**Y**	W	I	**Y**	**T**	**W**	T	**A**	**W**	G	**Y**	**E**	**W**	**Y**
Uca	**L**	E	**Y**	R	**Y**	**Y**	**T**	**W**	D	G	**D**	N	**Y**	**E**	**W**	**Y**
Uda	**I**	E	**Y**	Y	**D**	**Y**	**T**	**W**	D	**A**	**Q**	N	**Y**	**E**	**W**	**Y**
Uea	**F**	Q	**Y**	**D**	I	**Y**	**T**	**W**	N	G	Y	G	**Y**	V	**W**	**Y**
Ufa	**F**	F	**V**	W	**Y**	**Y**	**T**	**W**	N	**A**	**R**	Q	**Y**	Q	**W**	**Y**
Uga	**F**	**V**	**Y**	W	**Y**	**Y**	**T**	**W**	D	**A**	**T**	**S**	**Y**	**E**	**W**	**Y**
Uha	N	R	**Y**	F	**D**	**Y**	**T**	**L**	D	**A**	Y	N	**Y**	**E**	**W**	**Y**
Uia	V	N	**Y**	**D**	Y	**Y**	**T**	**W**	N	**A**	Y	N	**Y**	**E**	**W**	**Y**
Uja	V	A	**Y**	**K**	F	**Y**	**T**	**W**	D	G	**D**	N	**Y**	V	**W**	**Y**
Uka	**F**	E	**Y**	**R**	Y	**Y**	**T**	**W**	D	G	**D**	N	**Y**	**E**	**W**	**Y**

Table S10. Polymorphic peptide binding residues conserved between zebrafish and human MHCI molecules (continued). MHC Class I (MHCI) sequence polymorphisms predicted to be in contact peptides in the binding cleft are shown for human HLA-A and HLA-B molecules. Residues are listed as specific to either HLA-A or HLA-B alleles, or common to both, as described (6). Columns provide MHCI residue numbering according to IMGT nomenclature. Amino acid polymorphism at these sites is pervasive for the 11 zebrafish MHCI molecules, while the majority (>62% found across 34 positions) of these zebrafish polymorphic residues are also shared with HLA-A and/or HLA–B. Shared polymorphisms are highlighted in bold. Polymorphisms are shared between zebrafish and humans at all positions, with the exception of position 150. This overlap in substitutions across species indicates either shared ancestry for these alleles preserved across species, or continued exploration through a highly conserved stereo-chemical space. Additional residues are shown in Table S9.

	Allele 1 Accession	Allele 2 Accession	Allelic Diver.	EST or TSA allele support	EST or TSA allele support	Species Identifier
Psmb8						
Human	NP_683720	AHW47925.1	0.4%	BQ053596.1	BM919976.1	H. sapiens
Xenopus	NP_001080028.1	NP_001084323.1	13.4%	JZ824716.1	DC123145.1	X. laevis
Shark	AAL59862.1	AAL59861.1	22.4%	-	-	H. francisci
Salmon	AAG43439.1	ACI66984.1	27.9%	DY720790.1	GO056345.1	S. salar
Zebrafish	NP_571467.3	NP_001017791.1	30.8%	EH453180.1	CO813836.1	D. rerio
Psmb9						
Human	AAV38527.1	CAA44603.1	0.9%	CR997314.1	BQ051470.1	H. sapiens
Xenopus	NP_001003660.1	XP_004920717.2	4.2%	DN061107.1	EL726920.1	X. tropicalis
Shark	AAL59852.1	-	NA	-	-	C. milii
Salmon	NP_001117174.1	ACI69781.1	0.5%	EG867804.1	EG797054.1	S. salar
Zebrafish	NP_571466.1	NP_571753.1	13.6%	DN894394.1	AW076691.1	D. rerio
Psmb13						
Human	-	-	NA	-	-	H. sapiens
Xenopus	-	-	NA	-	-	X. tropicalis
Shark	XP_007882653.1	AFK10751.1	1.5%	JK959871.1	JK931281.1	C. milii
Salmon	XP_013997221.1	ABQ59681.1	7.2%	DY734168.1	DY740375.1	S. salar
Zebrafish	NP_571752.1	GDQH01002062.1	29.5%	EH507326.1	GDQH01002062.1	D. rerio
Tap2						
Human	NP_000535.3	AAP36912.1	0.1%	BM549545.1	BG285267.1	H. sapiens
Xenopus	NP_001081860.1	NP_001085260.1	29.5%	EB467320.1	CD363178.1	X. laevis
Shark	AGQ17917.1	AAL59859.1	0.4%	-	-	G. cirratum
Salmon	NP_001133546.1	NP_001117161.1	9.1%	DW580394.1	GE796559.1	S. salar
Zebrafish	GDQH01014094.1	CAD58764.1	49.7%	GDQH01014094.1	EB986911.1	D. rerio
MHCI						
Human	NP_001229687.1	NP_002108.4	20.8%	BQ900919.1	AL540488.3	H. sapiens
Xenopus	NP_001079871.1	AAF03407.1	18.9%	CF284001.1	CD329590.1	X. laevis
Shark	AAB97322.1	AAB97346.1	16.3%	-	-	T. scyllium
Salmon	AAN75113.1	AAN75119.1	29.1%	GO056509.1	GE771800.1	S. salar
Zebrafish	NP_571546.1	Q8HWF3	52.0%	EH497205.1	EH452240.1	D. rerio

Table S11. Divergent alleles encoding proteasome, TAP, and MHCI molecules within different vertebrates. Divergent alleles were identified using the BLAST algorithm with 'EST', 'TSA', and 'nr' databases across representative vertebrate species. Highest divergence levels for predicted amino acid sequences ('Allelic Diver.') were calculated using BLAST, with Divergence(%) = 100(%) – Identity(%). Levels are likely to underestimate sequence diversity found among species, considering limitations such as under-sampling. No sequence found is indicated by '-', and 'NA' means not applicable (as no percentage could be calculated).

	Allele 1 Accession	Allele 2 Accession	Allelic Diver.	EST or TSA allele support	EST or TSA allele support	Species Identifier
Psmb8						
Mouse	NP_034854.2	AAA75035.1	0.4%	CX233063.1	BI412709.1	M. musculus
Chicken	-	-	NA	-	-	G. gallus
Coelacanth	XP_006001021.1	-	31.4%	GAPS01019864.1	GAPS01046775.1	L. menadoe.
Gar	XP_015195220.1	XP_006643413.2	8.5%	-	-	L. oculatus
Medaka	NP_001171881.1	BAD93264.1	18.5%	BJ880842.1	BJ918390.1	O. latipes
Fugu	XP_003978904.1	-	1.5%	CA591128.1		T. rubripes
Psmb9						
Mouse	BAA40680.1	BAE31463.1	2.0%	BI854415.1	BY708333.1	M. musculus
Chicken	-	-	NA	-	-	G. gallus
Coelacanth	XP_006001022.1	-	0.9%	GAPS01037885.1	-	L. chalamnae
Gar	-	-	NA	-	-	L. oculatus
Medaka	NP_001265756.1	BAA19766.1	0.5%	DC274529.1	BJ508286.1	O. latipes
Fugu	CAC13120.1	-	NA	-	-	T. rubripes
Psmb13						
Mouse	-	-	NA	-	-	M. musculus
Chicken	-	-	NA	-	-	G. gallus
Coelacanth	-	-	NA	-	-	L. chalamnae
Gar	-	-	NA	-	-	L. oculatus
Medaka	NP_001171882.1	BAH29626.1	9.1%	DC251640.1	DC249921.1	O. latipes
Fugu	XP_003978905.1	-	1.0%	BU805709.1	-	T. rubripes
Tap2						
Mouse	AIC84017.1	BAE31870.1	1.1%	BI659951.1	CX227308.1	M. musculus
Chicken	NP_001092827.1	AEE25622.1	1.9%	CD217563.1	BU304487.1	G. gallus
Coelacanth	XP_006001020.1	-	0.0%	GAPS01015709.1		L. chalamnae
Gar	XP_006001020.1	-	NA	-	-	L. oculatus
Medaka	NP_001265816.1	BAB84549.1	1.8%	-	-	O. latipes
Fugu	XP_011615666.1	-	NA	-	-	T. rubripes
MHCI						
Mouse	NP_001001892.2	NP_034510.3	16.4%	CX226952.1	CX226363.1	M. musculus
Chicken	NP_001026509.1	BAD69566.1	23.0%	DR424122.1	DN930113.1	G. gallus
Coelacanth	AAA52345.1	AAA52352.1	14.7%	GAAA01059221.1	GAAA01003285.1	L. chalamnae
Gar	XP_015195141.1	XP_015195219.1	20.6%	-	-	L. oculatus
Medaka	NP_001265807.2	BAJ07296.1	22.7%	DK094922.1	FS531937.1	O. latipes
Fugu	XP_011607434.1	H2RYR0	20.3%	CA589387.1	CA589387.1	T. rubripes

Table S12. Divergent alleles encoding proteasome, TAP, and MHCI molecules within additional vertebrates. Divergent alleles were identified using the BLAST algorithm with 'EST', 'TSA', and 'nr' databases across representative vertebrate species. Highest divergence levels for predicted amino acid sequences ('Allelic Diver.') were calculated using BLAST, with Divergence(%) = 100(%) – Identity(%). Levels are likely to underestimate sequence diversity found among species, considering limitations such as under-sampling. No sequence found is indicated by '-', and 'NA' means not applicable (as no percentage could be calculated). Sequences highlighted in green were used to calculate divergence.

	Diver. Sum	Estimated Level
Human	22.2%	1
Xenopus	66.0%	3
Shark	40.6%	2
Salmon	73.8%	3
Zebrafish	175.6%	6
Mouse	19.9%	1
Chicken	24.9%	1
Coelacanth	47.0%	2
Gar	29.1%	2
Medaka	52.6%	2
Fugu	22.8%	1

Table S13. Combined allelic diversity estimates for proteasome, TAP, and MHCI molecules. Sequences provided in tables S11-12 were used to calculate highest divergence levels for alleles from each of the five MHC pathway genes across eleven vertebrate species. Cumulative levels of divergence were summed across the five genes (Diver. Sum) for individual species, and then used to estimate the overall level of lineage diversity throughout the MHC pathway for each species (Estimated Level), using a bin of 30 for levels.

For Datasets S1-5, sequence data are provided in fasta format.

Dataset S1. Accession numbers and sequence data for proteasome subunits. Sequences were derived from tBLASTn searches of ‘nr’, ‘WGS’, ‘TSA’, and ‘EST’ databases within NCBI querying vertebrate species of interest. Additional searches of the Ensembl and Uniprot databases were used to add additional sequences. The resulting sequences were used for multiple sequence alignments, phylogenetic analysis, and conserved synteny analysis to help identify sequence relationships.

Dataset S2. Accession numbers and sequence data for TAP subunits. Sequences were derived from tBLASTn searches of ‘nr’, ‘WGS’, ‘TSA’, and ‘EST’ databases within NCBI querying vertebrate species of interest. Additional searches of the Ensembl and Uniprot databases were used to add additional sequences. The resulting sequences were used for multiple sequence alignments, phylogenetic analysis, and conserved synteny analysis to help identify sequence relationships.

Dataset S3. Transcripts associated with zebrafish chromosome 19 core MHC haplotype D genes. Transcripts assembled from immune tissues of CG2 clonal zebrafish are provided, as listed in Supplemental Table 2, derived from the non-normalized Transcriptome Shotgun Assembly (TSA) which has been deposited under accession GDQH00000000 (the version described in this paper is the first version, GDQH01000000).

Dataset S4. Zebrafish antigen processing genes from the divergent chromosome 19 haplotype D. Deduced Proteasome and TAP amino acid sequences are provided from zebrafish core MHC haplotype D. Chromosome number and haplotype identifier are given in parentheses.

Dataset S5. Zebrafish chromosome 19 core MHC scaffold sequences. Genomic scaffolds from the CG2 clonal zebrafish assembly (core MHC haplotype D) are provided as illustrated in Figure 2 (‘rc’ indicates reverse complement), derived from the CG2v1.0 *de novo* assembly which has been deposited under Whole Genome Shotgun (WGS) accession LKPD00000000 (the version described in this paper is the first version, LKPD01000000).

Supplementary references

1. Tsukamoto, K, Miura F, Fujito N, Yoshizaki G, Nonaka M (2012) Long-lived dichotomous lineages of the proteasome subunit beta type 8 (PSMB8) gene surviving more than 500 million years as alleles or paralogs. *Mol Biol Evol* 29(10):3071–3079.

2. Huang C-H, Tanaka Y, Fujito NT, Nonaka M (2013) Dimorphisms of the proteasome subunit beta type 8 gene (PSMB8) of ectothermic tetrapods originated in multiple independent evolutionary events. *Immunogenetics* 65(11):811–821.

3. Sutoh Y, et al. (2012) Comparative genomic analysis of the proteasome β5t subunit gene: implications for the origin and evolution of thymoproteasomes. *Immunogenetics* 64(1):49–58.

4. Deverson EV, et al. (1998) Functional analysis by site-directed mutagenesis of the complex polymorphism in rat transporter associated with antigen processing. *J Immunol* 160(6):2767–2779.

5. Guillaume B, et al. (2010) Two abundant proteasome subtypes that uniquely process some antigens presented by HLA class I molecules. *Proc Natl Acad Sci* 107(43):18599–18604.

6. Nielsen M, et al. (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE* 2(8):e796.