Developmental and Comparative Immunology 36 (2012) 521–533

Contents lists available at SciVerse ScienceDirect

Developmental and Comparative Immunology

journal homepage: www.elsevier.com/locate/dci

Shark class II invariant chain reveals ancient conserved relationships with cathepsins and MHC class II

Michael F. Criscitiello [a,*], Yuko Ohta [b], Matthew D. Graham [b,1], Jeannine O. Eubanks [a], Patricia L. Chen [a], Martin F. Flajnik [b]

[a] Department of Veterinary Pathobiology, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA
[b] Department of Microbiology and Immunology, School of Medicine, University of Maryland at Baltimore, Baltimore, MD 21201, USA

ARTICLE INFO

Article history:
Received 11 September 2011
Revised 16 September 2011
Accepted 16 September 2011
Available online 1 October 2011

Keywords:
MHC
Antigen processing
Invariant chain
Evolution
Shark

ABSTRACT

The invariant chain (Ii) is the critical third chain required for the MHC class II heterodimer to be properly guided through the cell, loaded with peptide, and expressed on the surface of antigen presenting cells. Here, we report the isolation of the nurse shark Ii gene, and the comparative analysis of Ii splice variants, expression, genomic organization, predicted structure, and function throughout vertebrate evolution. Alternative splicing to yield Ii with and without the putative protease-protective, thyroglobulin-like domain is as ancient as the MHC-based adaptive immune system, as our analyses in shark and lizard further show conservation of this mechanism in all vertebrate classes except bony fish. Remarkable coordinate expression of Ii and class II was found in shark tissues. Conserved Ii residues and cathepsin L orthologs suggest their long co-evolution in the antigen presentation pathway, and genomic analyses suggest 450 million years of conserved Ii exon/intron structure. Other than an extended linker preceding the thyroglobulin-like domain in cartilaginous fish, the Ii gene and protein are predicted to have largely similar physiology from shark to man. Duplicated Ii genes found only in teleosts appear to have become sub-functionalized, as one form is predicted to play the same role as that mediated by Ii mRNA alternative splicing in all other vertebrate classes. No Ii homologs or potential ancestors of any of the functional Ii domains were found in the jawless fish or lower chordates.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The hallmark molecular components of the adaptive immune response have been found in the oldest group of living jawed vertebrates, the cartilaginous fish. Multiple IgH (Flajnik, 2002), IgL (Criscitiello and Flajnik, 2007), and TCR (Criscitiello et al., 2010; Rast et al., 1997) are diversified by RAG (Bernstein et al., 1994) and AID (Conticello et al., 2005) in sharks and rays. Furthermore, these animals are in the oldest extant group of vertebrates having a polymorphic, polygenic major histocompatibility complex (MHC) (Kasahara et al., 1992).

MHC class II glycoproteins present peptides to $CD4^+$ T cells. Newly synthesized class II α and β chains assemble in the endoplasmic reticulum (ER) together with a Type II glycoprotein called the invariant chain (Ii) (Jones et al., 1979). Ii is a chaperone ensuring the correct folding and trafficking of MHC class II proteins (Bikoff et al., 1993). Ii first trimerizes before the sequential addition of three class II α/β dimers (Lamb and Cresswell, 1992). In this nine-chain complex, each Ii blocks the peptide binding groove of one of the three class II heterodimers with a peptide called CLIP (class II-associated invariant chain peptide) (Freisewinkel et al., 1993), which prevents loading of class II with ER-derived proteins and peptides and provides the groove occupancy required for the stability of class II heterodimers (Zhong et al., 1996). In the trans-Golgi the αβIi complex is diverted from the secretory pathway to the endocytic pathway via a conserved motif in the Ii cytoplasmic tail. The complex is transported to an acidified post-Golgi vesicle where first the membrane distal portion of Ii is proteolytically degraded to leave the LIP (leupeptin-induced peptide), which still blocks the peptide binding cleft and retains the Ii transmembrane and cytoplasmic segments that continue to target the complex to the endosomal MHC class II compartment (MIIC) (Blum and Cresswell, 1988). The low pH of MIIC activates proteases to further cleave the membrane-proximal portion of Ii, leaving only CLIP in the peptide binding cleft. Blockade of this progressive cleavage of the Ii results in accumulation of Ii intermediates and reduced class II surface expression (Neefjes and Ploegh, 1992). DM can then bind class

* Corresponding author. Address: Texas A&M University, Mailstop 4467, College Station, TX 77843, USA. Tel.: +1 979 845 4207, mobile: +1 305 299 2522; fax: +1 979 862 1088.
E-mail addresses: mcriscitiello@cvm.tamu.edu (M.F. Criscitiello), yota@som.umaryland.edu (Y. Ohta), jeubanks@cvm.tamu.edu (J.O. Eubanks), pchen@cvm.tamu.edu (P.L. Chen), mflajnik@som.umaryland.edu (M.F. Flajnik).
[1] Deceased.

0145-305X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dci.2011.09.008

II and release LIP or CLIP, facilitating the exchange for endosomal peptides before transport to the cell surface (Schafer et al., 1996). Maturation of phagosomes containing non-self cargo (versus apoptotic self-cargo) may be enhanced by toll-like receptor-mediated vesicular signaling, marking those phagosomes with pathogenic contents for fusion into the MIIC compartment (Blander and Medzhitov, 2006).

The enzymes that cleave Ii are related to papain and known as cathepsins, which are the same proteases that degrade lysosomal contents for antigen loading (Riese and Chapman, 2000). The activation of cathepsins requires an acidic environment, and they can be divided into four categories depending on the critical component of their active sites: cysteine, aspartate, serine, or metal ions (Turk et al., 2001). Cysteine cathepsins are primarily involved in antigen processing, specifically cathepsins L and S are dedicated to this function. Cathepsin activity is regulated by several protein inhibitors, including cystatins, thyropins and even one domain of Ii itself. The thyroglobulin-like (Tg) domain of longer Ii isoforms is a strong inhibitor of cathepsin L but not cathepsin S (Bevec et al., 1996).

The mammalian Ii gene has an exon organization that largely corresponds to its structural protein domains. The first exon encodes the amino-terminal cytoplasmic tail including the endosomal targeting motifs, the second exon encodes the transmembrane domain, and the third exon encodes a linker between the membrane and the trimerization domain that includes CLIP. The fourth, fifth and sixth exons contribute to the three alpha helices and connecting strands of the trimerization domain. The seventh exon, alternatively spliced out in short isoforms, encodes the Tg domain that presumably inhibits cathepsin proteolytic action. The eighth exon nearly encodes the entire carboxy-terminal end, which is often rich in charged residues but has unclear function. The ninth protein-coding exon encodes the final amino acid or two and contains the stop codon.

Unlike MHC genes, Ii genes do not display high allelic polymorphism, but four variants of the protein are found in human as p33, p35, p41 and p43 (O'Sullivan et al., 1987). Use of an alternative start codon accounts for the small molecular weight differences between the (predominant) p33 and p41, and p35 and p43 forms. The p35 and p43 forms contain an ER-retention motif lacking in the shorter forms from the alternative initiation site; this signal is concealed upon $\alpha\beta$ binding and allows transport of the nonamer to the Golgi (Schutze et al., 1994). The 10 kDa distinction between the p33/p35 and p41/p43 forms results from the alternatively spliced Tg domain exon, mentioned above. The Tg domain is a structural motif found in several functionally unrelated proteins (e.g., testican, equistatin, thyroglobulin) and sometimes functions as an inhibitor of cysteine proteases, often with higher target specificity than the better studied cystatins (Mihelic and Turk, 2007).

Ii knockout mice show impeded class II transport and surface expression. Class II found on the surface of Ii-deficient cells has an unstable conformation due to the lack of endogenously processed peptide, but the dimers can bind peptide added to the medium. Accordingly, cells from these mice do not present whole exogenous antigen well and the animals have greatly reduced numbers of thymic and peripheral $CD4^+$ T cells (Bikoff et al., 1993; Viville et al., 1993). These transgenic mice studies suggest that Ii prevents class II from binding floppy, incompletely folded proteins in the ER (rather than preventing the binding of peptides transported into the ER by TAP) and stabilize the heterodimer. Ii knockout mice with restoration of either p31 or p41 (containing the Tg domain) have shown that both forms participate in class II folding and assembly, can reconstitute the $CD4^+$ T cell population, and rescue immune responses to protein antigen (Shachar et al., 1995; Takaesu et al., 1995). The complete functions of each isoform are not known; however p41 has been shown to be necessary for airway hyperresponsiveness and IgE responses in the lung (Ye et al., 2003).

Although crucial to class II antigen presentation, Ii and cathepsins are encoded outside of the MHC (Long et al., 1983). However, cathepsins S and L are found in MHC paralogous regions (Flajnik and Kasahara, 2010), one of many linkages that contribute to hypotheses of an ancestral "pre-adaptive immune complex" encoding antigen receptors, NK receptors and antigen processing and presentation components (Ohta et al., 2011). Ii and homologous genes have been identified in several divergent vertebrate model species, although such reports are few in comparison to class II α/β chains. Amongst poikilothermic vertebrates, annotated Ii sequences have been submitted to public databases from reptiles and amphibians and studies have been conducted on Ii from bony fish species. The cloning of the first Ii from lower vertebrates was done in zebrafish, and this work confirmed that, like in mammals, fish Ii-like transcripts exist in multiple forms (Yoder et al., 1999). Work in rainbow trout also found two Ii products (Dijkstra et al., 2003) that are encoded by two paralogous genes (Fujiki et al., 2003) as opposed to alternative splicing. Structure of sea bass Ii was modeled more recently with analysis of potential interactions with class II and cathepsins (Silva et al., 2007). Here, we report the first description of Ii from the cartilaginous fish, the oldest vertebrate group with MHC-based adaptive immunity. We set out to determine whether the gene and its expression were phylogenetically conserved, and attempted to find genes related to precursors of Ii and cathepsins in jawless vertebrates and lower deuterostomes.

2. Methods

2.1. Cloning of nurse shark Ii chain and cathepsins

A *Ginglymostoma cirratum* (nurse shark) spleen/pancreas cDNA library was constructed in the pDONR222 vector using the Gateway cloning system (Invitrogen). From this an ~8000 clone expressed sequence tag (EST) database was created after removing known housekeeping, MHC, immunoglobulin and TCR clones by subtractive colony hybridization using 137 mm Magna membranes (Osmonics) for probing and high stringency washing techniques described previously (Criscitiello et al., 2004). DNA was prepared with 96-Turbo plasmid miniprep kits (Qiagen) or TempliPhi rolling circle DNA amplification (GE Healthcare) and single dye-terminator based sequencing runs were performed at the University of Maryland Biopolymer Core Facility using the universal M13 reverse primer. ESTs were used as queries against the non-redundant protein sequence database with blastx (NCBI). Gene-specific primers (Supplemental Table 1) were designed to complete sequencing of clones with high identity to Ii and cathepsins. Additional cDNA libraries from shark lymphoid tissues were assayed by 5′ and 3′ rapid amplification of cDNA ends (RACE) PCR with gene-specific Ii primers to identify all expressed splice variants. These were cloned and sequenced as above or with Zyppy plasmid DNA miniprep kit (Zymo Research), extended with BigDye XTerminator (Applied Biosystems), purified and sequenced by the Texas A&M DNA Technologies Core Laboratory.

2.2. Blotting

Total RNA was prepared for northern blotting as described (Bartl et al., 1997), and 10 μg was loaded in each lane. The nurse shark nucleotide diphosphate kinase (NDPK) probe used as a loading control was amplified with primers NDPKF and NDPKR (Kasahara et al., 1992) (Supplemental Table 1). A probe for nurse shark Ii was amplified from primers NSIiF1 and NSIiR1 which generate a probe from cDNA encoding the endosomal targeting sequence of the cytoplasmic tail to CLIP. Northern blotting and

probing for nurse shark class IIA has been described previously (Kasahara et al., 1992; Ohta et al., 2004). A putatively single exon probe amplified from primers NSIiF2 and NSIiF8 was used in genomic Southern blotting of DNA from shark erythrocytes as previously described (Criscitiello et al., 2006). Blots were probed from five related sharks digested with five different enzymes as well as single enzyme blots of families (mother and pups) of analyzed MHC paternities by restriction fragment length polymorphism (RFLP) (Ohta et al., 2002).

2.3. Database mining and structural prediction

Portions of the nurse shark Ii sequences described here were used to query the database of the elephant shark genome project (http://esharkgenome.imcb.a-star.edu.sg/) by blastn and tblastn. Two scaffolds were identified in this cartilaginous fish containing exons predicted to encode Ii by visual inspection for GT/AG intron boundaries and comparison of predicted protein sequence with other vertebrates.

CD74 (nomenclature for surface expressed Ii (Koch et al., 1991)) is annotated on scaffold 29 of the *Anolis carolinus* genome (AnoCar1.0) but only two exons are marked. Identification of the entire Ii locus in this reptile was accomplished by first finding anole ESTs (e.g., FG760756 and FG750983) with homology to caiman and other vertebrate Ii sequences, these were used to identify the remaining exons with the exception of the first. The first exon was predicted based on visual scrutiny of ten kilobases 5′ of the second exon.

Annotated and partially annotated genomic sequences of Ii loci of human, chicken and the frog *Xenopus tropicalis* as well as ICLP-1 and ICLP-2 loci of zebrafish genomic sequences were checked against available cDNA data to tabulate and compare exon/intron sizes and phases. These loci were studied using genome browsers at NCBI, Ensemble and UCSC websites where open reading frames 5′ and 3′ to the Ii ortholog were analyzed for conserved synteny of the region amongst vertebrates.

Similar BLAST approaches were used to identify additional lower vertebrate Ii and cartilaginous fish cathepsin EST sequences using nurse shark and other vertebrate sequences as query (Supplemental Table 2) and to search for orthologs of Ii in lower chordates.

2.4. Phylogenetic analysis

Amino acid alignments of Ii homologs were initially made in Bioedit with ClustalW employing gap opening penalties of 10 and gap extension penalties of 0.1 for pairwise alignments then 0.2 for multiple alignments and the protein-weighting matrix of Gonnett or Blossum (Fig. 2 and Supplementary Fig. 1) (Hall, 1999; Tamura et al., 2007). These alignments were then heavily modified by hand. MEGA was used to infer the phylogenetic relationships of Ii homologs. Evolutionary distances were computed using the Dayhoff matrix (Schwarz and Dayhoff, 1979) and 387 column positions in the 37 selected sequences. A Neighbor-Joining tree was made from 1000 bootstrap replicates, using pairwise deletion.

3. Results

3.1. Isolation and identification of nurse shark Ii chain

Generation of a nurse shark EST database from spleen and pancreas yielded an Ii clone (clone 104D3, comprising 1802 bp, Fig. 1) that showed high identity to many Ii sequences of bony fish and tetrapods (highest protein match to the pike *Esox lucius*, expect 6*e*−19, 35% amino acid identity over 196 positions). This sequence was used to design primers (Supplemental Table 1) for 5′ and 3′RACE PCR. Many clones were sequenced from several primer combinations amplified from peripheral blood and spleen to obtain all expressed genes, but only three other reproducible coding sequence variants were isolated, all exemplified by clone D2 (Fig. 1). D2 is a long splice variant of EST clone 104D3 with a 195 bp insertion and also contains two single nucleotide polymorphisms (SNP) and two small deletions (indel). The indel creating serine 35 (104D3 aa numbering) and the point mutation exchanging glutamic acid for glutamine at 211 both appear to be allelic polymorphisms, since each occur independently in both the long and short splice isoforms.

3.2. Mining of other Ii sequences from poikilothermic vertebrates

Isolation of the nurse shark Ii sequence prompted database searches for Ii in other species. We found ESTs of complete and partial Ii from dogfish shark (*Squalus acanthus*) and Pacific electric ray (*Torpedo californica*) and performed partial genomic analysis of Ii from the more primitive Holocephalin, the elephant shark (*Callorhinchus milii*). Other bony fish Ii ESTs were found, full genomic annotation was completed for the green anole lizard (*Anolis carolina*) and the Ii exon/intron structure was manually acquired for the two zebrafish ICLPs, *X. tropicalis*, chicken and human Ii genes. Accession numbers are shown in Supplemental Table 2 and all sequences analyzed are aligned in Supplemental Fig. 1.

3.3. Sequence analysis of Ii from cartilaginous fish

The primary functional domains that have been described in Ii of higher vertebrates (transmembrane, trimerization, CLIP, Tg; all colored domains in Fig. 1) are present in Ii of both major radiations of elasmobranchs (modern sharks (Selachi) and skates and rays (Batoidea), Fig. 2). Additionally, the elephant shark Ii confirms that at least the trimerization domain and Tg are present in the more primitive holocephalians. Protein identity to the complete nurse shark long isoform (D2 adding N-terminal MSADEQQNALL) in pairwise alignments range from 33% with trout S25-7 Ii, to 29% with caiman to 26% with chicken and human. The long connecting linker domain between the trimerization domain and Tg seems to be a common feature in the cartilaginous fish lineage.

3.3.1. Cytoplasmic tail and transmembrane domain

As a Type II transmembrane protein, Ii has an amino terminal cytoplasmic tail. No evidence was found by 5′RACE for use of an alternative start codon in nurse shark or in the databases for any other ectothermic vertebrate. Database searches only found clear evidence from the rhesus macaque (XM_001099491) and a new world marmoset (XR_087115) of alternative start codon use similar to human, and the giant panda Ii has a putative 15aa alternative amino terminal peptide similar in sequence to the primate sequences. The Arg-Arg-Ser-Arg ER localization signal in this longer alternative initiation product of human Ii is how such signals were discovered (Bakke and Dobberstein, 1990; Schutze et al., 1994), yet this signal cannot be detected in the cytoplasmic tails of other vertebrate Ii. Di-leucine like motifs have been identified in the mammalian Ii cytoplasmic region which serve as endosomal targeting motifs (Pieters et al., 1993) and mediate Ii interaction with the clathrin adaptor proteins AP1 and AP2 (Kongsvik et al., 2002). One such candidate sequence is seen in the nurse shark Ii at position 21–22 (Fig. 2). Acidic residues preceding the di-leucine are conserved in the shark motif and have proven necessary for sorting to large endosomes (Pond et al., 1995). The EST included in Fig. 2 for Ii of the electric ray has the longest cytoplasmic domain of any Ii studied that does not use an alternative start methionine.

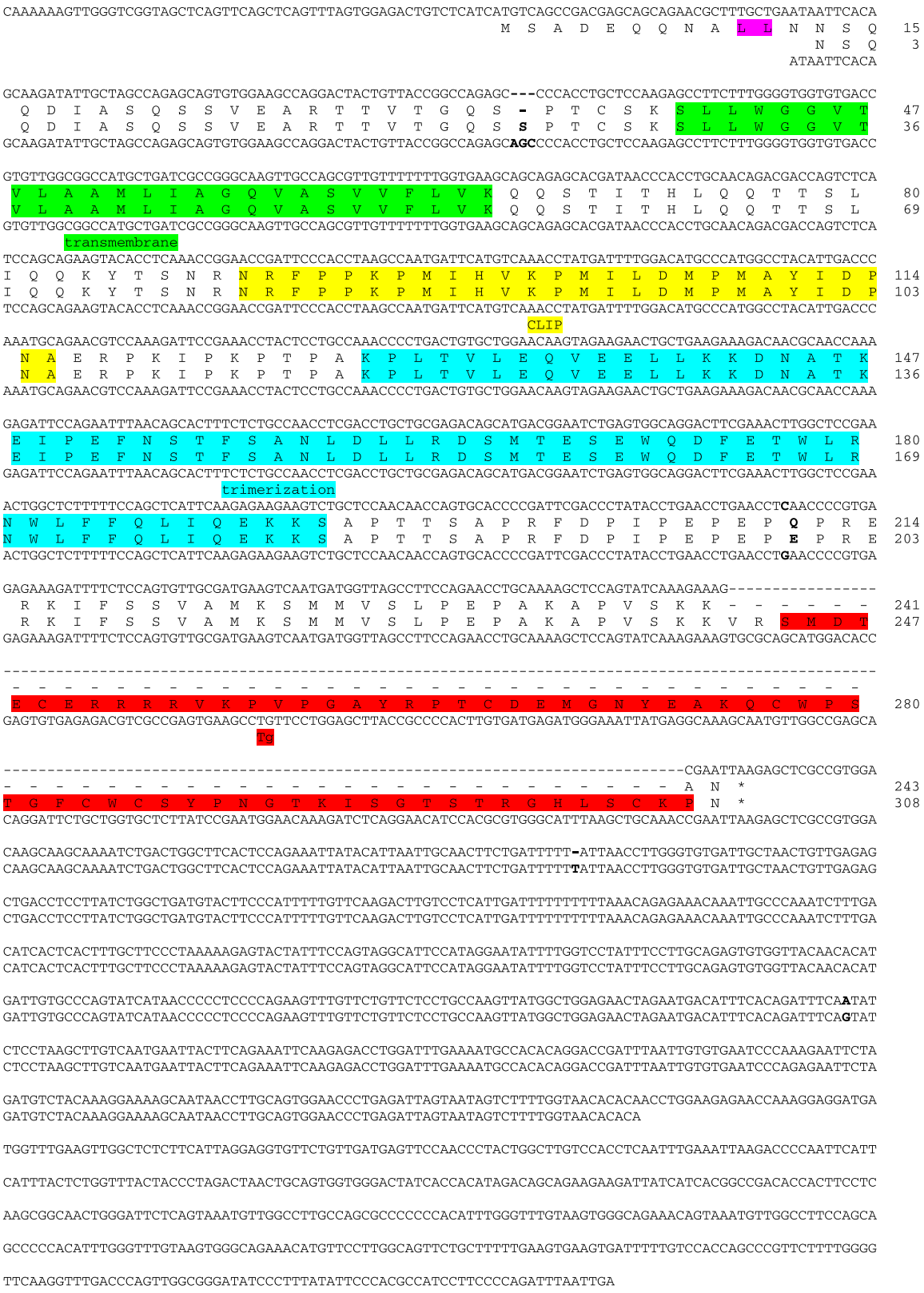


Fig. 1. Nurse shark cDNAs with homology to Ii chain. Nucleotide and putative amino acid sequence of nurse shark cDNA clones 104D3 (top) and D2 (bottom). Gaps introduced for alignment are shown as dashes. Single base point mutations and insertion/deletions are highlighted. Amino acids of the endosomal targeting motif are highlighted in pink, the transmembrane domain in green, CLIP yellow, trimerization domain blue and Tg domain red. NCBI has given these sequences Accession No. JF507710 and JF507711. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

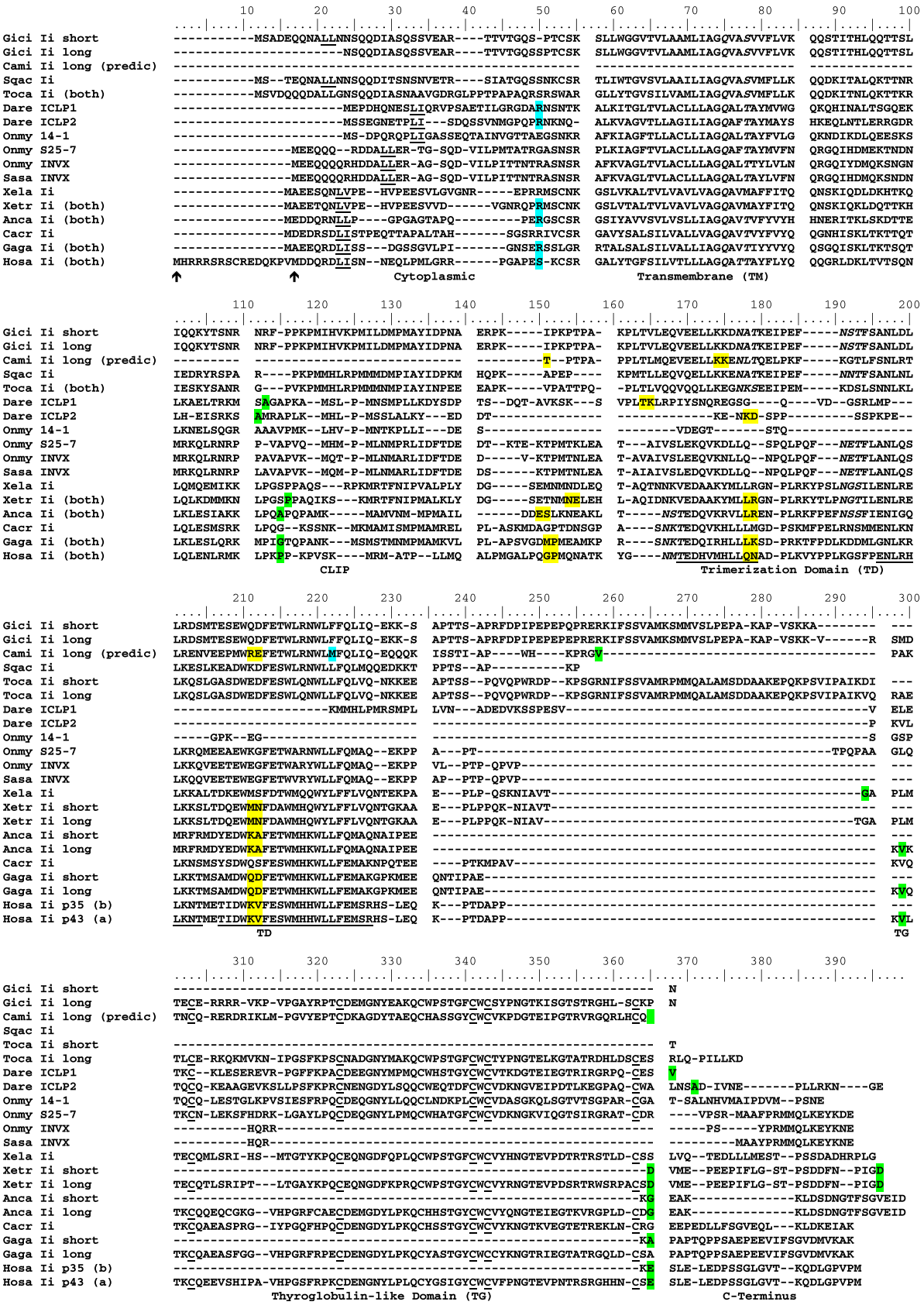


Fig. 2. Amino acid alignment identifies nurse shark and other chondrichthian Ii when compared to other vertebrate sequences. Yellow highlighting (spanning two residues) indicates phase zero, green highlighting indicates phase one and blue highlighting indicates phase two intron positioning. Genbank Accession numbers are shown in Supplemental Table 2. Endosomal targeting motifs (L–L/I/V) are underlined, as are the six conserved cysteines of Tg-like domains and (in the human sequences) the three alpha helices of the trimerization domain. Asparagine-linked glycosylation motifs (N–X–S/T, X ≠ P) are in italics. Also in italics are polar residues corresponding to the hydrophilic patch of the transmembrane domain implicated in Ii trimerization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The transmembrane region of Ii is much more conserved among vertebrates than any other part of the molecule, typically rich in hydrophobic residues. The transmembrane region of cartilaginous fish Ii maintains two of three polar residues that in human form a hydrophilic patch important for trimerization (Ashman and Miller, 1999). The nurse shark sequence has QvAS at positions 75–78 where the human employs QaTT.

3.3.2. CLIP (class II associated peptide)

CLIP has been shown in mammals to be crucial for class II folding, transport, and peptide groove occupancy (Romagnoli and Germain, 1994), yet has proven difficult to compare amongst vertebrates as evidenced by very different published alignments that include sequences from bony fish (Dijkstra et al., 2003; Fujiki et al., 2003; Silva et al., 2007). Indeed, assigning different gap penalties and inclusion or exclusion of different sequences can drastically change alignments of this region performed by programs such as Clustal and Muscle. Silva et al. noted striking conservation of large and small CLIP side chains in a teleost Ii that bind in conserved pockets of human MHC (Ghosh et al., 1995), although this required inserting a three amino acid gap in the mammalian CLIP core region that occupies the MHC class II groove (Silva et al., 2007). When cartilaginous fish, amphibian and reptile sequences are included in the alignment such conservation patterns are not as evident. The regions between the transmembrane domain and CLIP, and CLIP and the carboxy-proximal trimerization domain, show similarly poor conservation between shark and other vertebrate groups. Similarly, the cytoplasmic region showed some divergence between the bony fish and other vertebrate sequences, CLIP and the region linking the CLIP with the trimerization domain show significant differences in teleost Ii in comparison to other vertebrates.

3.3.3. Trimerization domain and linker region

Carboxy-proximal to CLIP is a trimerization domain containing three conserved alpha helices (underlined in Fig. 2). Many residues in these helices of human Ii that are important for nonamer formation are evolutionarily conserved (Jasanoff et al., 1998). Four leucines (165 and 166 of helix a, 195 and 198 of helix b) are well conserved, and several other residues (Trp210, Phe213, Glu214, Trp216, Trp220, and Phe223) are nearly invariant. Amino acids making the crucial bonds required for packing of the third (c) alpha helix in the Ii trimer (Bijlmakers et al., 1994) are conserved, and three residues (Phe208, Trp211, Trp215) are invariant over evolutionary time (see also Supplemental Fig. 1). In cartilaginous fish Ii, one (elephant shark and electric ray) or two (nurse and dogfish shark) putative N-linked glycosylation sites are found in or flanking alpha helix A, as has been seen in other vertebrates.

3.3.4. Tg domain and carboxy terminus

The longer isoform (D2) of nurse shark Ii includes a Tg domain. This isoform has been shown in mammals to be generated by insertion of a 195bp exon (Strubin et al., 1986). Tg domains display inhibitory activity against cysteine- or cation-dependent proteases and are called thyropins (Lenarcic and Bevec, 1998). The Tg domain of Ii from shark also is a likely thyropin. This cysteine-rich domain in mammals interacts with the active site of some cathepsins in an inhibitory fashion, and can discriminate cathepsins L from S (Guncar et al., 1999). All six cysteines that in human Ii form three disulfide bonds that stabilize the tertiary structure of the domain are conserved in shark Ii, suggesting that the shark Ii Tg domain assumes a two sub-domain structure seen in human Ii. Indeed, these six cysteines (and a tryptophan between the fourth and fifth cysteine) are found in all vertebrate Ii having a Tg domain (Fig. 2 and Supplemental Fig. 1).

Like in most vertebrates, some short teleost Ii clones were found. These Tg-less fish Ii were first identified from the trout (INVX) but this Ii homolog was encoded by a gene distinct from the Tg-containing trout Ii chains, rather than the products of alternative splicing (Fujiki et al., 2003). Sequences with a predicted domain structure like INVX have also been submitted from Atlantic salmon (Leong et al., 2010). This may be due to the teleost-specific genome duplication, generating two Ii loci.

The Ii carboxy terminal portion adjacent to the Tg domain is much shorter in nurse shark than other vertebrates, except the cyprinid ICLP-1's in which it is also short. Little is known about this portion of Ii but it could be involved in CD74 signaling of macrophage inhibitory factor with CD44 (Shi et al., 2006). This region is rich in charged residues in bony vertebrates, and teleost Ii (except the ICLP-1's) have a distinctive stretch of aliphatic and other hydrophobic residues preceding the region of glutamic and aspartic acids, lysines and arginines (Fig. 2 and Supplemental Fig. 1).

3.4. Comparative domain structure

Measurement of predicted domains and intervening protein linker sequence showed general conservation between Ii of different classes of jawed vertebrates (Fig. 3). The most conspicuous deviation is shared by Ii from cartilaginous fish and ICLP-1 of bony fish, where an extended linker between the trimerization domain (or A alpha helix of trimerization domain in the case of ICLP-1) and Tg domain is coincident with a shorter carboxy terminal tail than is found in tetrapod and other bony fish Ii forms. As the ICLP-1 and nurse shark linkers are dissimilar to each other (and anything else in the databases) it is doubtful that there is rescue in this region of any function missing in their short carboxy terminus.

As stated above, some teleost Ii forms only appear without certain domains. Splice forms are never found of either zebrafish ICLP or trout 14-1 with a complete trimerization domain and trout and salmon INVX does not contain the Tg domain (Dijkstra et al., 2003; Fujiki et al., 2003; Yoder et al., 1999).

3.5. Genomic organization of the Ii locus in vertebrates

We compared protein coding exon/intron lengths and splice sites where available from shark to man (Fig. 4, annotation of green anole lizard and elephant shark shown in Supplemental Figs. 2 and 3, respectively). Similar intron positions and phases were found in frog, lizard, chicken and man (Fig. 2). Although intron sizes were much longer in *X. tropicalis* and much shorter in chicken (typical for chickens), the relative size of these introns to those of other higher mammals is consistent with expansions and contractions of these loci. A separate small exon for the stop codon was not found in the anole lizard as it was in other tetrapods.

The two zebrafish ICLP loci analyzed showed that ICLP-1 on chromosome 14 was twice as long as ICLP-2 on chromosome 12, including an additional exon for the linker between the partial trimerization domain and Tg. As might be expected considering the sequence divergence, the partial trimerization domains of the ICLPs and subsequent linker showed the most diversity in exon/intron organization. This region in shark contains a phase two intron not consistent with other vertebrates, and the ICLPs contained a phase zero intron the position of which could not be aligned with precision to the introns of other vertebrate Ii. The three domains encoding the elephant shark's trimerization domain were found together on one scaffold, and the Tg domain was found on another.

Southern blotting with genomic DNA of a family of nurse sharks usually resulted in two bands hybridizing to an Ii transmembrane-CLIP probe (Supplemental Fig. 4) consistent with a single Ii locus in

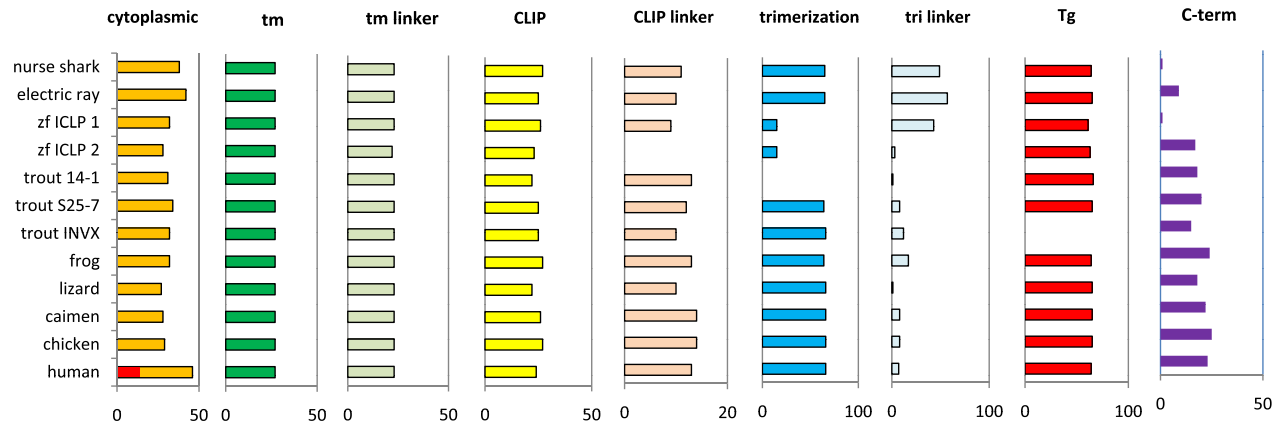


Fig. 3. Invariant chain domain conservation in cartilaginous fish and tetrapods. Representative predicted protein sequences were chosen to compare the entire longest splice forms of invariant chain homologs found. *Xenopus laevis* was used for frog. Human is shown with earlier alternative start site that gives 16 additional amino acids to the cytoplasmic tail, shown in red. Length is measured in amino acids. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

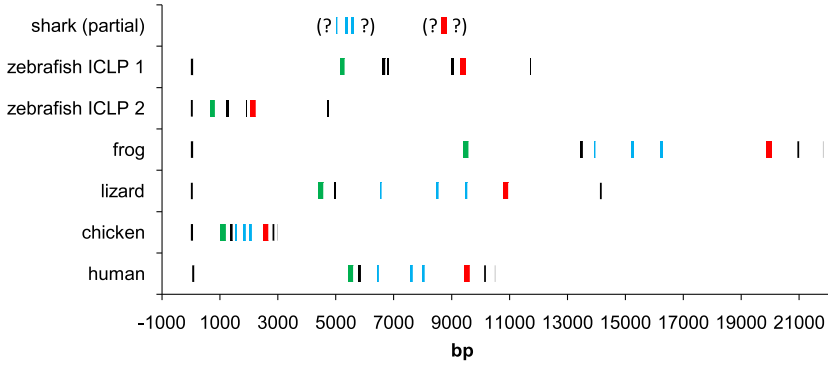


Fig. 4. Exon-intron organization and relative size similar from shark to man. Exons encoding the transmembrane domain are shown in green, the exons approximately encoding the three alpha-helices of the trimerization domain are shown in blue, and the Tg domain in red. Question marks denote missing data from elephant shark scaffolds, exon content in each set of parentheses is on one unmapped scaffold. Drawn to scale, distance in base pairs is shown at bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

shark, rather than multiple loci as in bony fish. RFLP were compared with known patterns for MHC (Ohta et al., 2002) and shark Ii was not MHC-linked (data not shown).

3.6. Tissue expression

Northern blotting on RNA from many nurse shark tissues demonstrated coordinate regulation of Ii transcripts with those of MHC class II (Fig. 5). Expression was highest in gill, spleen and spiral valve (shark intestine), with lower but obvious expression also in peripheral blood leukocytes, thymus and brain. Consistent with what is known in mammals, such synchronized expression of these two genes suggests shared transcriptional regulation early in vertebrate adaptive immunity. In mammals both genes' promoters (and that of HLA-DM) depend on the class II transactivator (CIITA) to coordinate upregulation of transcription in response to interferon γ ((Brown et al., 1993), reviewed in (Ting and Trowsdale, 2002)). CIITA has yet to be unambiguously identified below bony fish but it appears likely that it or an analogous system regulates their expression in shark. We did identify a candidate CIITA partially encoded on an elephant shark scaffold (AAVX01120910.1) which shares 42% identity and 66% similarity ($e = 4e{-}49$) with CIITA of zebra finch (Fig. 6).

Multiple 5′ and 3′RACE experiments yielded Ii cDNA clones yielding transcripts of 1.8B (D2) and 2.3B (104D3) (Fig. 1) based on alternative polyadenylation sites in the 3′ untranslated region differing by 497B (assuming 200B poly-A tails (Wahle and Keller,

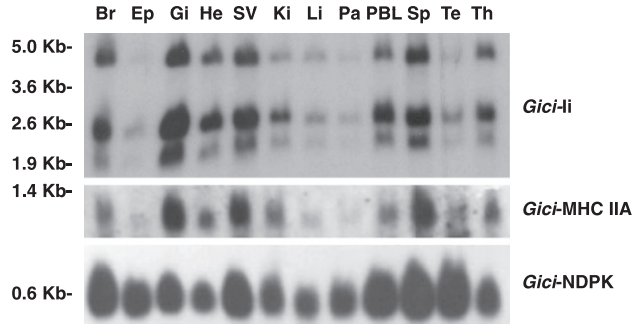


Fig. 5. Tissue expression of shark Ii mRNA. Northern blot hybridization of nurse shark tissues (Br, brain; Ep, epigonal; Gi, gills; He, SV, spiral valve; Ki, kidney; Li, liver; Pa, pancreas; PBL, peripheral blood leukocytes; Sp, spleen; Te, testis; and Th, thymus) with the *Gici*-Ii probe demonstrates that three transcripts for Ii are expressed at similar levels relative to MHCIIA. Size in bases of the transcripts is shown on the left.

1992)). Each of these has an alternative splice isoform possible without the 195b exon encoding the Tg domain. The northern blotting confirmed predominant bands migrating near the 1.6–1.8 KB and 2.1–2.3 KB sequence lengths of each form with and without the Tg domain exon, but the slightly larger transcripts were higher than predicted possibly do to longer poly-A tails or longer 5′ untranslated regions. There was also a larger 4.8KB band that we could not identify from cDNA. This large transcript displayed

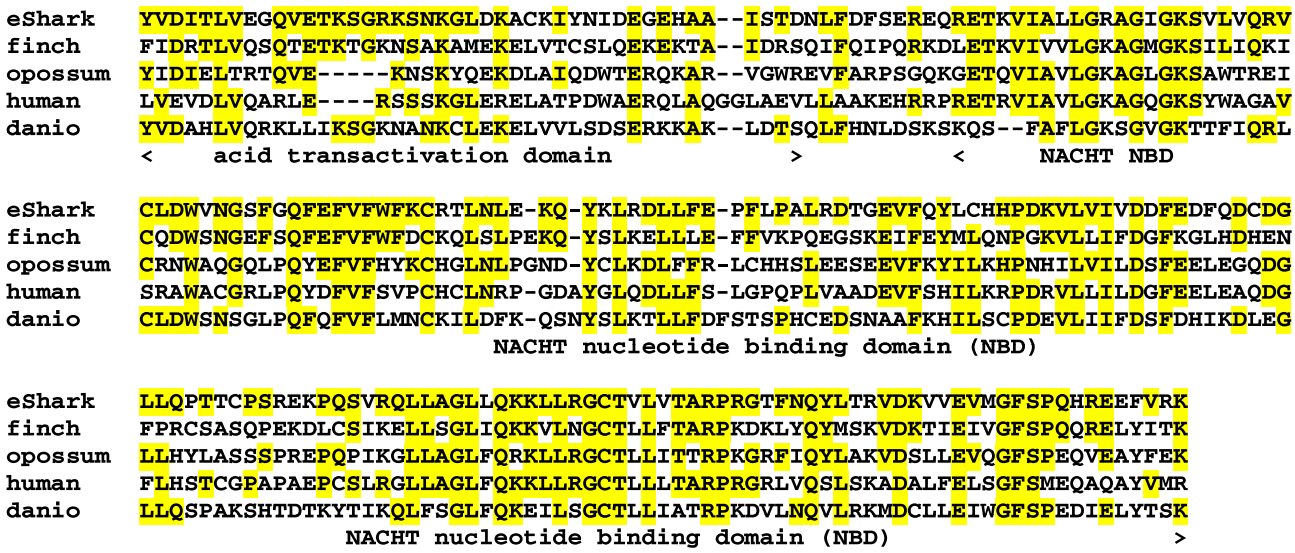


Fig. 6. CIITA of cartilaginous fish. Amino acid alignment of partial putative elephant shark (eShark) ortholog of CIITA. Yellow highlighting indicates identity with other vertebrate CIITA proteins. Predicted domains of protein shown under alignment. Other sequences included in the alignment: finch XP002195062, opossum XM001376433, zebra danio XP001343072, and human NP000237. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

parallel tissue expression levels as the lower bands. Therefore there is likely an additional polyadenylation variant or exon splice isoform that has yet to be identified. We can exclude the possibility of a second Ii locus that eluded the Southern probing as it would have hybridized to the same probe on the northern blot (Supplemental Fig. 4).

3.7. Cathepsin degradation and evolution of Ii regulation

Since the Tg-like domain is a putative component of Ii of all jawed vertebrate groups back to shark, we searched for possible interactions with cathepsins. Several cathepsin ESTs were found from nurse shark, two of which were more similar to the ancient L cathepsin lineage, implicated in Ii proteolysis, than to the primeval B lineage that are not (Fig. 7, Supplemental Fig. 5) (Uinuk-Ool et al., 2003). Additional putative cathepsin L/S EST sequences were mined from the little skate (*Leucoraja erinacea*). Based on the crystal structure of the mammalian Ii Tg domain with cathepsin L we modeled similar interactions between the Ii and cathepsin orthologs from cartilaginous fish (Guncar et al., 1999).

As described above, the Tg-like domain of mammalian Ii forms a wedge-shaped conformation of three loops stabilized by three disulfide bonds between conserved cysteines (Turk et al., 1999). The inhibited papain-like cysteine proteases including cathepsin L share a common fold of two domains, which separate on the top in a "V" shaped active site cleft (Coulombe et al., 1996). Several interactions identified in the solved mammalian cathepsin L–Tg domain structure could be maintained by residues found in the Tg domains and putative cathepsin L in cartilaginous fish.

3.8. Ii evolution

Besides the structural data described above, two additional lines of evidence from phylogenetic and syntenic analyses suggested that the canonical functions of Ii arose at the origins of RAG/AID/MHC-based adaptive immunity.

3.8.1. Phylogenetic analysis

We performed many phylogenetic analyses with different alignments, excluding various domains, and with several matrix- and tree-building algorithms. The tree with the most support at significant nodes includes the entire longest form (Tg domain encoding exon spliced in) of the proteins (Fig. 8). Ii from chondrichthyes clustered basal to Ii and Ii paralogs from bony vertebrates. The incomplete sequence from the dogfish shark behaved erratically with different tree building methods, whereas the incomplete elephant shark repeatedly fell basal to all the other vertebrate Ii, as expected for the primitive Holocephalian. As suggested by sequence analysis, domain structure, and intron splice sites; cardinal Ii emerged in the cartilaginous fish.

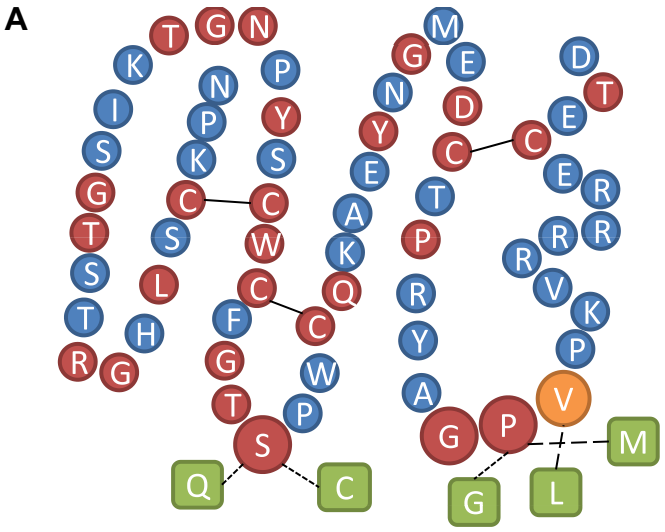
On the other hand, teleost sequences fell into two well supported clades: those ICLP-like sequences lacking the trimerization domains and those more typical of other vertebrate Ii. A duplication leading to the ICLP genes likely occurred in the ancestors of protacanthopterygii (including samonids) and ostariophysi (including catfish and cyprinids). The INVX duplication generating Tg-less bony fish Ii may have occurred more recently in the salmonid lineage, as they have thus far only been found in trout and salmon.

3.8.2. Genomic syntenies

We used available genome projects to study the Ii locus. The amniota include the (paraphyletic) reptiles, birds and mammals, showed conservation of syntenic genes up- and downstream of the Ii locus (Supplemental Fig. 6). However, we identified only one neighboring gene (*Tcof1*, encoding the treacle protein associated with Treacher Collins Syndrome) that was conserved between the amphibian *Xenopus* and the amniotes. The gene distal to Ii on the other side of *Tcof1* in frog was *csf1r* (colony stimulating factor receptor), that linked to the flanking regions of zebrafish ICLPs. The additional genome-wide duplication in bony fish likely allowed for much divergence, but clear conservation of synteny was found among frog, lizard, bird, mammals, and one of the bony fish forms. Surprisingly, the cod has lost Ii as well as class II and functional CD4 (Star et al., 2011).

4. Discussion

In characterizing the expressed Ii gene in nurse shark and other cartilaginous fish we found general conservation of sequence, splice variants, expression, genomic organization, predicted structure, and function with Ii from tetrapods. Evidence was also found of an ancient relationship between the specialized cathepsin L and



B

N Shark catL 97F2	PVKNQGSCGSCWAFSSTGALEGQTFKKTGRLIPLSEQNLVDCSQAQGNYGCGGGWMENA
N Shark catL 48B1	PVKNQRHCGSCWAFSAVGALEGQTFKKTRKLISLSEQNLVDCSNAEGNHGCNGGLMEYA
Skate catL 598970	PVKNQGSCGSCWAFSSTGALEGQTFKMTGKLISLSEQNLVDCSTSQGNSGCSGGWMENA
Skate catL 592412	PVKDQGQCGSCWAFSTTGALEGQQFRKSGKLVSLSEQNLVDCSKPEGNEGCNGGLMDQA
Skate catL 051409	GVKNQGRCGSCWAFSAVGALEGQTFRKTGKLISLSEQNLVDCSNAEGNHGCNGGLMQYA
Human catL	PVKNQGQCGSCWAFSATGALEGQMFRKTGRLISLSEQNLVDCSGPQGNEGCNGGLMDYA
Human catS	EVKYQGSCGACWAFSAVGALEAQLKLKTGKLVSLSAQNLVDCSTEKYGNKGCNGGFMTT

C

E Shark	PAKTNCQ-RERDRIKLMPGVYEPTCDKAGDYTAEQCHASSGYCWCVKPDGTEIPGTRVRGQRLHCQ-
N Shark	SMDTECE-RRRRVKPV-PGAYRPTCDEMGNYEAKQCWPSTGFCWCSYPNGTKISGTSTRGHL-SCKP
Ray	RAETLCERKQKMVKNI-PGSFKPSCNADGNYMAKQCWPSTGFCWCTYPNGTELKGTATRDHLDSCES
Frog	PLMTEC-QMLSRIHSM-TGTYKPQCEQNGDFQPLQCWPSTGFCWCVYHNGTEVPDTRTRSTLD-CSS
Lizard	KVKTKCQQEQCGKG-VHPGRFCAECDEMGDYLPKQCHHSTGYCWCVYQNGTEIEGTKVRGPLD-CDG
Chicken	KVQTKCQAEASFGG-VHPGRFRPECDENGDYLPKQCYASTGYCWCCYKNGTRIEGTATRGQLD-CSA
Human	PVLTKCQEEVSHIPAVHPGSFRPKCDENGNYLPLQCYGSIGYCWCVFPNGTEVPNTRSRGHHN-CSE

Fig. 7. Predicted cathepsin L inhibition by Tg domain in cartilaginous fish. (A) Bead amino acid schematic of Tg-like domain of nurse shark. Amino acids of Ii predicted from mammalian crystal structures to be important contacts with cathepsin L and conserved in shark are shown as larger circles. Conserved cathepsin amino acids that are predicted to interact with Ii are shown as green boxes, dotted lines show hydrogen bonds and electrostatic interactions, dashed lines show hydrophobic interactions. Disulfide bonded cysteines are shown joined by a line. (B) Five cathepsins from elasmobranchs having similarities to cathepsin L aligned with cathepsin L and S from human. Residues predicted to form key conserved interactions are in green. (C) Amino acid alignment of Tg-like domain from elephant shark, nurse shark, Pacific electric ray, frog (*X. laevis*), anole lizard, chicken and human. Residues conserved in at least five of the seven aligned sequences are shown in red in panel (C) and red beads in panel (A). Nurse shark Val255 and orthologous residues are highlighted in orange as this position was usually maintained as an aliphatic, hydrophobic residue. Underlined amino acids make key contacts between Ii Tg-like domain and cathepsin L in human. Structure interactions adapted from Guncar and Turk (Guncar et al., 1999). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the Ii as well as the regulation of the former's action by the Ii's own Tg domain. Lockstep expression of Ii and class II was observed and a putative CIITA was identified from cartilaginous fish. Genomic syntenies, exon-intron structure and Southern blotting suggest that one Ii locus emerged early in the Ig/TCR/MHC based adaptive immune system of jawed vertebrates and was maintained in all major groups. Additional genome wide duplication(s) afforded bony fish multiple Ii loci encoding different domain structures, presumably with distinct functions, one of which is found in Ii splice forms of all other vertebrates.

Class II groove occupancy by CLIP is a hallmark property of the invariant chain, yet we found little strict conservation of this sequence between vertebrate classes. However methionine residues were common in shark as other vertebrate CLIP regions. Two human Ii CLIP methionine residues (in position127 and 138 of Fig. 2) are the most important occupiers of conserved class II pockets (Ghosh et al., 1995), and meta comparative analyses may reveal a broader "super-motif" among vertebrate class II of deep pockets that recognize these CLIP methionine residues (Malcherek et al., 1995). Other residues such as alanine, arginine, proline and serine also are common in CLIP. We suspect that the high positive selection exerted on the class II peptide binding groove over the half-billion years of Ii evolution has forced CLIP to change accordingly. Additionally, temperature may pose very different requirements on the CLIP-class II interaction in endothermic birds and mammals versus poikilothermic vertebrates. CLIP and the region linking CLIP with the trimerization domain are where the teleost Ii homologs show significant differences from other vertebrate Ii, suggestive of distinct physiology.

All current data suggest that the trimerization domain is a fixed component of all Ii from all groups except teleost fish, as in the cyprinid ICLPs and trout 14-1 sequence. In some of these fish sequences an extended region is substituted for the trimerization domain, most evident as positions 222–253 of zebrafish ICLP-1

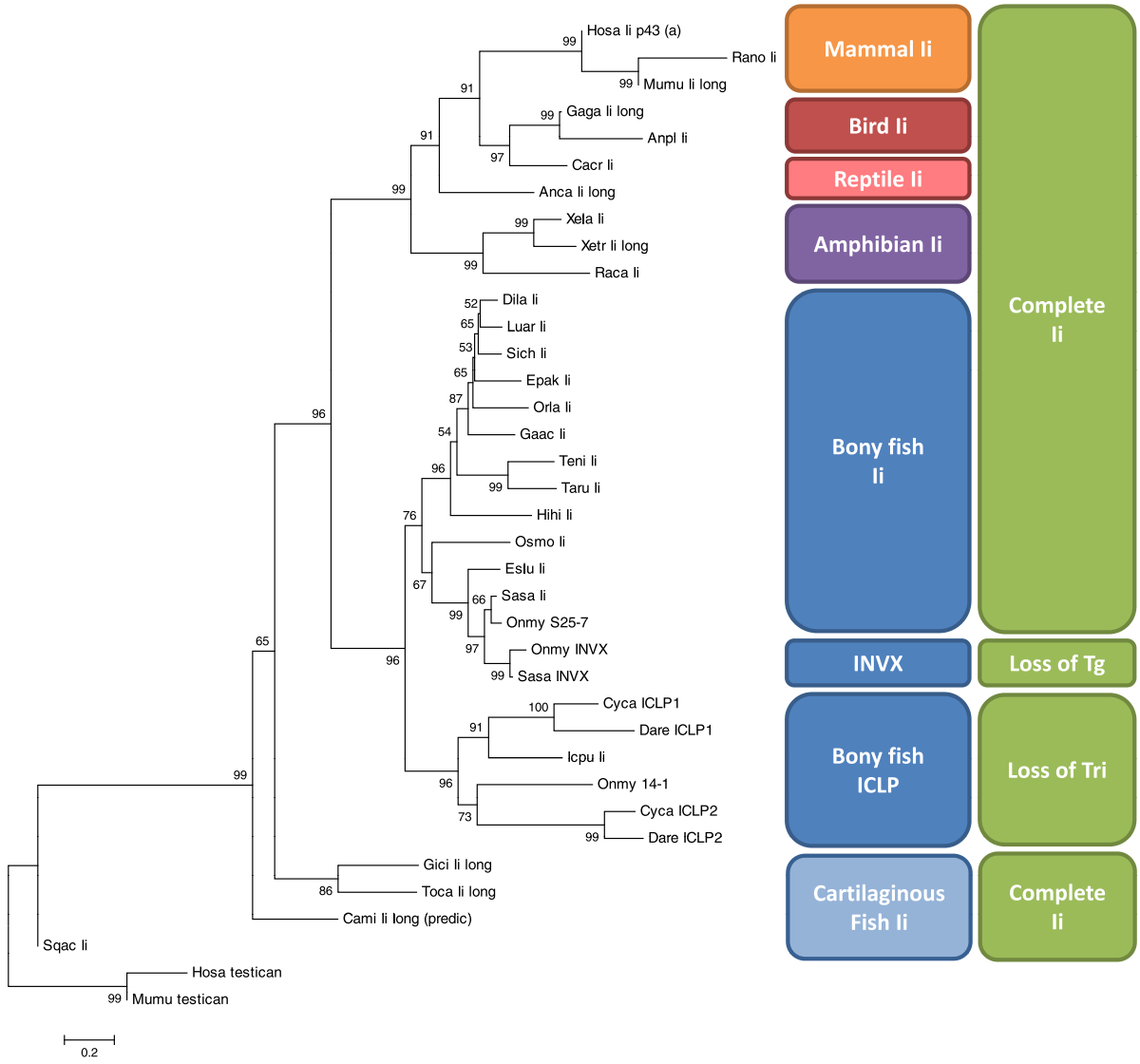


Fig. 8. Cartilaginous fish Ii groups with other Ii in phylogenetic analysis. Neighbor joining tree of Ii amino acid sequences aligned and analyzed in MEGA. Alignment (with sequences not included in this tree) is found in Supplemental Fig. 1, sequence accession numbers and names given in Supplemental Table 2. Bootstrap values at nodes inferred from 1000 replicates. Evolutionary distance is shown with the scale bar in the units of amino acid substitutions per site. Boxes to the center-right highlight different vertebrate phylogenetic clusters of Ii and far right mark loss of Tg and trimerization domains in some teleost Ii forms, compared to the complete Ii.

(Fig. 2). It is noteworthy that nurse shark and electric ray Ii have extended regions following their trimerization domains. Although the teleost sequence replacing the trimerization domain in ICLP-1 may be similar to prokaryotic catalase (Dijkstra et al., 2003), it is possible this was due to untrimmed vector deposited in the databases. These unique portions of the bony fish sequences and the extended region after the trimerization domain of cartilaginous fish are similar neither to each other nor any to any other protein.

Interestingly, cathepsin S, the cathepsin free of Tg domain inhibition, could not be identified from cartilaginous fish, suggesting that it may have evolved after the emergence of the class II pathway. Cathepsin S is pH independent and is upregulated by interferon γ (Chapman, 1998), characteristics of a highly specialized adaptive cathepsin. In mammals the regulation afforded by cathepsin S and L (or S, L and V (Sevenich et al., 2010)) may be used for modulation of T cell selection in the thymus (Lombardi et al., 2005), as cathepsin L is necessary for optimal positive selection by cortical thymic epithelial cells (Nakagawa et al., 1998). Perhaps this was coincident with the neofunctionalization of cathepsin K in osteoclasts. Other components of the class II pathway arose after the emergence of adaptive immunity, DM in amphibians and DO in mammals.

At least twice in teleost Ii evolution new Ii loci emerged that have been specialized by loss of exons: trout and salmon INVX have no Tg domain (like alternative splicing in other vertebrates) and the ICLPs (including trout 14-1 and a catfish cDNA) have no trimerization domain. Loss of the trimerization domain in the ICLPs would be expected to alter the class II processing pathway significantly. Trimerization is necessary for release of the Ii/class II complex from the ER and protection from rapid degradation, yet the Ii TM domain is capable of trimerizing as well (Dixon et al., 2006). The self-affinity of Ii in the membrane is attributed to a glutamine and two threonine residues, of which the glutamine is well conserved from shark to man (Fig. 2). The teleost ICLPs that lack the trimerization domain do have one of the less conserved threonines as well as the glutamine. Future work must address how MHCII/Ii trimers (instead of nonamers) articulate with calnexin and are transported to the MIIC. Interestingly, some bony fish appear to

make multiple Ii forms with diverse domain configurations (trout having full length S25-7, INVX lacking the Tg and 14-1 lacking the trimerization), while other (perhaps most) fish seem restricted to just one or two variants. The two genes in one fish species are often very similar, apparently without subfunctionalization. But maintenance of two ICLP forms in the cyprinids suggests the possibility of distinct physiology for ICLP-1 and ICLP2.

A wealth of circumstantial evidence suggests the Tg domain is resistant to proteolysis which shifts the control to the regulation of cathepsin S which is exempt from the Tg domain's inhibition. But certainly the smaller forms must be favored in some circumstances? Evidence for this is not easy to find in the literature, but Ii p31 (without Tg) predominates in B cells and the longer form enhances antigen processing in macrophages and dendritic cells (Ye et al., 2003). At a gross tissue level, nurse shark certainly seems to express the two isoforms defined here at similar levels (Fig. 5) as neither the middle nor the lower "bands" representing differential polyadenylation sites is tight, signifying that both contain the isoforms with and without the Tg domain.

This Tg domain near the carboxy-terminus of Ii (Katunuma et al., 1994) is a member of the cystatin superfamily (Brown and Dziegielewska, 1997) and this suggests that Ii could have evolved from a stefin or cystatin for chaperone function. As CLIP became more specialized for the class II peptide binding groove, the Tg domain could have coevolved with a specialized cathepsin. Work in Ii/H2-M$^{-/-}$ mice with restoration of p43 suggested that, like DM, the Ii Tg domain may augment peptide-CLIP exchange (Bikoff et al., 1998), which might be the ancestral method of facilitating peptide/CLIP exchange prior to the emergence of DM in amphibians.

Several search strategies failed to identify an Ii ortholog in jawless chordates (Supplemental Table 3). To obtain the remnant genes containing functional domains in Ii, the Tg and trimerization domains of nurse shark and electric ray, as well as the 'transmembrane to CLIP' and carboxy terminus from nurse shark were used as bait in BLAST searches against lamprey genomic scaffolds, lamprey ESTs, hagfish ESTs and genomic scaffolds from the urochordate *Ciona intestinalis* and cephalochordate *Branchiostoma floridae*. That no likely candidates were revealed may indicate great sequence divergence between the gnathostome Ii and the fragmented loci that were assembled for the genesis of Ii; such ancestral loci would likely have encoded genes with quite different functions before the emergence of the class II antigen presentation system.

However, syntenic clusters of genes may be maintained in cartilaginous fish and even jawless vertebrates to the cystatin family member or other gene(s) that later gave rise to the Ii. We used these identified syntenic genes (*Arsi*, *Tcof1*, *Rps14*, *Ndst1*, *Synpo* and *Csf1r*) from more recent vertebrate groups to search Version 3.0 of the lamprey (*Petromyzon marinus*) genome project supercontigs. Despite finding good candidates for human *Tcof1* (lamprey contig 40760), mouse *Arsi* (lamprey contig 7255), zebrafish *Csf1r* (amphioxus unassigned chromosomal scaffold 617826) and human RPS14 (lamprey contig 47654, and many lesser candidates for these and other syntenics in lamprey) we were unable to identify a neighboring gene on these short contigs that bore domains or characteristics likely to be co-opted into the Ii (trimerization, Tg-like, Type II transmembrane with di-leucine motifs). We do not think this effort will remain futile; completion of the lamprey genome as well as those of other vertebrate genomes may yet yield an Ii antecedent. The location of the *CD74* gene in humans (5q11–23) is intriguing, while not on an MHC paralogon it is near a region that seems to have broken off from the MHC paralogous region on chromosome 9 (Lundin et al., 2003).

In considering the evolution of the class II system, we should note that many functions have been described for Ii. As mentioned, macrophage migration inhibitory factor (MIF) binds the extracellular portion of Ii and then signals by the complex associating with CD44 (Leng et al., 2003). Both mouse Ii isoforms in the lung can mediate allergen-induced lung inflammation and eosinophilia, but the p41 is necessary for IgE response and airway hyper responsiveness (Ye et al., 2003). The free cytoplasmic domain of Ii has been shown to induce B cell differentiation via NF-κB (Matza et al., 2002a,b). These or other functions of Ii suggest an alternative physiology that may have preceded its more famed roles as chaperone and peptide cleft occupier, and may also expose the ancestral loci. Cathepsins are ancient and were clearly co-opted for class II antigen generation. Sea urchin cathepsins (L and B) are up-regulated in LPS activated coelomocytes (Nair et al., 2005). Cathepsin S cleavage of the Ii cytoplasmic tail even can regulate the motility of dendritic cells via the tail's interactions with myosin II. There are many leads to follow in many model organisms.

5. Conclusion

Now that the third chain of MHC class II has been identified in sharks, focus can turn to how this antigen presentation system arose to play a major role in adaptive immunity. As comparative genetic data continues to grow, so should our ability test evolutionary hypotheses. Exciting advances in the jawless agnathans suggest their variable lymphocyte receptor (VLR) system uses VLR-B for free receptors in a humoral response and VLR-A (and perhaps VLR-C) in cellular responses (Pancer et al., 2005; Rogozin et al., 2007). These T cell-like VLR-A bearing lymphocytes develop in a "thymoid" region of the lamprey pharynx (Bajoghli et al., 2011). Apparently lacking MHC and Ii, are these T cell analogs of the jawless fish lacking a pathway of processing antigen from the endosomal pathway that only evolved in the jawed gnathostomes? A similar question can now be asked of the adaptive immune system of the Atlantic cod, whose genome lacks MHC class II, Ii and functional CD4 (Star et al., 2011). Much remains to be learned of the adaptive immune system of this fish, but it is clear that this is a derived loss of the class II/Ii/CD4 arm of adaptive immunity in a subset of teleosts.

What genomic loci in animals with only innate or innate and a VLR based adaptive system (reviewed in (Saha et al., 2010)) were exploited by the fledgling adaptive system that uses MHC restricted T cells and RAG mediated rearrangement of Ig superfamily genes? The peptide binding regions of MHC and the Ii had definitive ancestors, as did immunoglobulins and T cell receptors. Many cathepsin genes are encoded in MHC paralogous regions which are thought to have originated from two rounds of full genome duplication early in vertebrate evolution (Flajnik and Kasahara, 2010). Several other recent studies point to earlier genomic concentrations of genes that are not linked in mammals, such as Ig-like variable domains in T cell receptors (Criscitiello et al., 2010; Parra et al., 2010) and B_2-microglobulin in the MHC (Ohta et al., 2011). With the tightly coordinated expression of class II and Ii and the finding of a partial CIITA locus in shark reported here, there are new reasons to investigate the origins of the antigen presenting cell. Large scale genomics and RNAi screens are beginning to elucidate how CIITA is regulated, casting spotlights on the RMND5B and MAPK1 (Paul et al., 2011). Lower vertebrates may tell us if and how CIITA was pirated from an innate NLR (NOD, LRR receptor) locus early in the gnathostomes to serve as a master regulator in early antigen presenting cells to allow their presentation to, and activation of, helper T cells.

Acknowlegements

This work was supported by the NIH through grants to MFC (AI56963) and MFF (AI027877), and is dedicated to the memory of Matt Graham.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.dci.2011.09.008.

References

Ashman, J.B., Miller, J., 1999. A role for the transmembrane domain in the trimerization of the MHC class II-associated invariant chain. J. Immunol. 163, 2704–2712.

Bajoghli, B., Guo, P., Aghaallaei, N., Hirano, M., Strohmeier, C., McCurley, N., Bockman, D.E., Schorpp, M., Cooper, M.D., Boehm, T., 2011. A thymus candidate in lampreys. Nature 470, 90–94.

Bakke, O., Dobberstein, B., 1990. MHC class II-associated invariant chain contains a sorting signal for endosomal compartments. Cell 63, 707–716.

Bartl, S., Baish, M.A., Flajnik, M.F., Ohta, Y., 1997. Identification of class I genes in cartilaginous fish, the most ancient group of vertebrates displaying an adaptive immune response. J. Immunol. 159, 6097–6104.

Bernstein, R.M., Schluter, S.F., Lake, D.F., Marchalonis, J.J., 1994. Evolutionary conservation and molecular-cloning of the recombinase activating gene-1. Biochem. Biophys. Res. Co. 205, 687–692.

Bevec, T., Stoka, V., Pungercic, G., Dolenc, I., Turk, V., 1996. Major histocompatibility complex class II-associated p41 invariant chain fragment is a strong inhibitor of lysosomal cathepsin L. J. Exp. Med. 183, 1331–1338.

Bijlmakers, M.J., Benaroch, P., Ploegh, H.L., 1994. Mapping functional regions in the lumenal domain of the class II-associated invariant chain. J. Exp. Med. 180, 623–629.

Bikoff, E.K., Huang, L.Y., Episkopou, V., van Meerwijk, J., Germain, R.N., Robertson, E.J., 1993. Defective major histocompatibility complex class II assembly, transport, peptide acquisition, and CD^{4+} T cell selection in mice lacking invariant chain expression. J. Exp. Med. 177, 1699–1712.

Bikoff, E.K., Kenty, G., Van Kaer, L., 1998. Distinct peptide loading pathways for MHC class II molecules associated with alternative Ii chain isoforms. J. Immunol. 160, 3101–3110.

Blander, J.M., Medzhitov, R., 2006. On regulation of phagosome maturation and antigen presentation. Nat. Immunol. 7, 1029–1035.

Blum, J.S., Cresswell, P., 1988. Role for intracellular proteases in the processing and transport of class II HLA antigens. Proc. Natl. Acad. Sci. USA 85, 3975–3979.

Brown, A.M., Wright, K.L., Ting, J.P., 1993. Human major histocompatibility complex class II-associated invariant chain gene promoter. Functional analysis and in vivo protein/DNA interactions of constitutive and IFN-gamma-induced expression. J. Biol. Chem. 268, 26328–26333.

Brown, W.M., Dziegielewska, K.M., 1997. Friends and relations of the cystatin superfamily – new members and their evolution. Protein Sci. 6, 5–12.

Chapman, H.A., 1998. Endosomal proteolysis and MHC class II function. Curr. Opin. Immunol. 10, 93–102.

Conticello, S.G., Thomas, C.J., Petersen-Mahrt, S.K., Neuberger, M.S., 2005. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol. Biol. Evol. 22, 367–377.

Coulombe, R., Grochulski, P., Sivaraman, J., Menard, R., Mort, J.S., Cygler, M., 1996. Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. EMBO J. 15, 5492–5503.

Criscitiello, M.F., Flajnik, M.F., 2007. Four primordial immunoglobulin light chain isotypes, including lambda and kappa, identified in the most primitive living jawed vertebrates. Eur. J. Immunol. 37, 2683–2694.

Criscitiello, M.F., Kamper, S.M., McKinney, E.C., 2004. Allelic polymorphism of TCRalpha chain constant domain genes in the bicolor damselfish. Dev. Comp. Immunol. 28, 781–792.

Criscitiello, M.F., Ohta, Y., Saltis, M., McKinney, E.C., Flajnik, M.F., 2010. Evolutionarily conserved TCR binding sites, identification of T cells in primary lymphoid tissues, and surprising trans-rearrangements in nurse shark. J. Immunol. 184, 6950–6960.

Criscitiello, M.F., Saltis, M., Flajnik, M.F., 2006. An evolutionarily mobile antigen receptor variable region gene: doubly rearranging NAR-TcR genes in sharks. Proc. Natl. Acad. Sci. USA 103, 5036–5041.

Dijkstra, J.M., Kiryu, I., Kollner, B., Yoshiura, Y., Ototake, M., 2003. MHC class II invariant chain homologues in rainbow trout (*Oncorhynchus mykiss*). Fish Shellfish Immunol. 15, 91–105.

Dixon, A.M., Stanley, B.J., Matthews, E.E., Dawson, J.P., Engelman, D.M., 2006. Invariant chain transmembrane domain trimerization: a step in MHC class II assembly. Biochemistry 45, 5228–5234.

Flajnik, M.F., 2002. Comparative analyses of immunoglobulin genes: surprises and portents. Nat. Rev. Immunol. 2, 688–698.

Flajnik, M.F., Kasahara, M., 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat. Rev. Genet. 11, 47–59.

Freisewinkel, I.M., Schenck, K., Koch, N., 1993. The segment of invariant chain that is critical for association with major histocompatibility complex class II molecules contains the sequence of a peptide eluted from class II polypeptides. Proc. Natl. Acad. Sci. USA 90, 9703–9706.

Fujiki, K., Smith, C.M., Liu, L., Sundick, R.S., Dixon, B., 2003. Alternate forms of MHC class II-associated invariant chain are not produced by alternative splicing in rainbow trout (*Oncorhynchus mykiss*) but are encoded by separate genes. Dev. Comp. Immunol. 27, 377–391.

Ghosh, P., Amaya, M., Mellins, E., Wiley, D.C., 1995. The structure of an intermediate in class II MHC maturation: CLIP bound to HLA-DR3. Nature 378, 457–462.

Guncar, G., Pungercic, G., Klemencic, I., Turk, V., Turk, D., 1999. Crystal structure of MHC class II-associated p41 Ii fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S. EMBO J. 18, 793–803.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acid Symp. Ser. 41, 95–98.

Jasanoff, A., Wagner, G., Wiley, D.C., 1998. Structure of a trimeric domain of the MHC class II-associated chaperonin and targeting protein Ii. EMBO J. 17, 6812–6818.

Jones, P.P., Murphy, D.B., Hewgill, D., McDevitt, H.O., 1979. Detection of a common polypeptide chain in I–A and I–E sub-region immunoprecipitates. Mol. Immunol. 16, 51–60.

Kasahara, M., Vazquez, M., Sato, K., McKinney, E.C., Flajnik, M.F., 1992. Evolution of the major histocompatibility complex: isolation of class II A cDNA clones from the cartilaginous fish. Proc. Natl. Acad. Sci. USA 89, 6688–6692.

Katunuma, N., Kakegawa, H., Matsunaga, Y., Saibara, T., 1994. Immunological significances of invariant chain from the aspect of its structural homology with the cystatin family. FEBS Lett. 349, 265–269.

Koch, N., Moldenhauer, G., Hofmann, W.J., Moller, P., 1991. Rapid intracellular pathway gives rise to cell surface expression of the MHC class II-associated invariant chain (CD74). J. Immunol. 147, 2643–2651.

Kongsvik, T.L., Honing, S., Bakke, O., Rodionov, D.G., 2002. Mechanism of interaction between leucine-based sorting signals from the invariant chain and clathrin-associated adaptor protein complexes AP1 and AP2. J. Biol. Chem. 277, 16484–16488.

Lamb, C.A., Cresswell, P., 1992. Assembly and transport properties of invariant chain trimers and HLA-DR-invariant chain complexes. J. Immunol. 148, 3478–3482.

Lenarcic, B., Bevec, T., 1998. Thyropins – new structurally related proteinase inhibitors. Biol. Chem. 379, 105–111.

Leng, L., Metz, C.N., Fang, Y., Xu, J., Donnelly, S., Baugh, J., Delohery, T., Chen, Y., Mitchell, R.A., Bucala, R., 2003. MIF signal transduction initiated by binding to CD74. J. Exp. Med. 197, 1467–1476.

Leong, J.S., Jantzen, S.G., von Schalburg, K.R., Cooper, G.A., Messmer, A.M., Liao, N.Y., Munro, S., Moore, R., Holt, R.A., Jones, S.J., Davidson, W.S., Koop, B.F., 2010. *Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. BMC Genom. 11, 279.

Lombardi, G., Burzyn, D., Mundinano, J., Berguer, P., Bekinschtein, P., Costa, H., Castillo, L.F., Goldman, A., Meiss, R., Piazzon, I., Nepomnaschy, I., 2005. Cathepsin-L influences the expression of extracellular matrix in lymphoid organs and plays a role in the regulation of thymic output and of peripheral T cell number. J. Immunol. 174, 7022–7032.

Long, E.O., Strubin, M., Wake, C.T., Gross, N., Carrel, S., Goodfellow, P., Accolla, R.S., Mach, B., 1983. Isolation of cDNA clones for the p33 invariant chain associated with HLA-DR antigens. Proc. Natl. Acad. Sci. USA 80, 5714–5718.

Lundin, L.G., Larhammar, D., Hallbook, F., 2003. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. J. Struct. Func. Genom. 3, 53–63.

Malcherek, G., Gnau, V., Jung, G., Rammensee, H.G., Melms, A., 1995. Supermotifs enable natural invariant chain-derived peptides to interact with many major histocompatibility complex-class II molecules. J. Exp. Med. 181, 527–536.

Matza, D., Kerem, A., Medvedovsky, H., Lantner, F., Shachar, I., 2002a. Invariant chain-induced B cell differentiation requires intramembrane proteolytic release of the cytosolic domain. Immunity 17, 549–560.

Matza, D., Lantner, F., Bogoch, Y., Flaishon, L., Hershkoviz, R., Shachar, I., 2002b. Invariant chain induces B cell maturation in a process that is independent of its chaperonic activity. Proc. Natl. Acad. Sci. USA 99, 3018–3023.

Mihelic, M., Turk, D., 2007. Two decades of thyroglobulin type-1 domain research. Biol. Chem. 388, 1123–1130.

Nair, S.V., Del Valle, H., Gross, P.S., Terwilliger, D.P., Smith, L.C., 2005. Macroarray analysis of coelomocyte gene expression in response to LPS in the sea urchin. Identification of unexpected immune diversity in an invertebrate. Physiol. Genom. 22, 33–47.

Nakagawa, T., Roth, W., Wong, P., Nelson, A., Farr, A., Deussing, J., Villadangos, J.A., Ploegh, H., Peters, C., Rudensky, A.Y., 1998. Cathepsin L: critical role in Ii degradation and CD4 T cell selection in the thymus. Science 280, 450–453.

Neefjes, J.J., Ploegh, H.L., 1992. Inhibition of endosomal proteolytic activity by leupeptin blocks surface expression of MHC class II molecules and their conversion to SDS resistance alpha beta heterodimers in endosomes. EMBO J. 11, 411–416.

O'Sullivan, D.M., Noonan, D., Quaranta, V., 1987. Four Ia invariant chain forms derive from a single gene by alternate splicing and alternate initiation of transcription/translation. J. Exp. Med. 166, 444–460.

Ohta, Y., Landis, E., Boulay, T., Phillips, R.B., Collet, B., Secombes, C.J., Flajnik, M.F., Hansen, J.D., 2004. Homologs of CD83 from elasmobranch and teleost fish. J. Immunol. 173, 4553–4560.

Ohta, Y., McKinney, E.C., Criscitiello, M.F., Flajnik, M.F., 2002. Proteasome, transporter associated with antigen processing, and class I genes in the nurse shark *Ginglymostoma cirratum*: evidence for a stable class I region and MHC haplotype lineages. J. Immunol. 168, 771–781.

Ohta, Y., Shiina, T., Lohr, R.L., Hosomichi, K., Pollin, T.I., Heist, E.J., Suzuki, S., Inoko, H., Flajnik, M.F., 2011. Primordial linkage of beta2-microglobulin to the MHC. J. Immunol. 186, 3563–3571.

Pancer, Saha, N.R., Kasamatsu, J., Suzuki, T., Amemiya, C.T., Kasahara, M., Cooper, M.D.. Variable lymphocyte receptors in hagfish. Proc. Natl. Acad. Sci. USA 102 (26), 9224–9229.

Parra, Z.E., Ohta, Y., Criscitiello, M.F., Flajnik, M.F., Miller, R.D., 2010. The dynamic TCRdelta: TCRdelta chains in the amphibian *Xenopus tropicalis* utilize antibody-like V genes. Eur. J. Immunol. 40, 2319–2329.

Paul, P., van den Hoorn, T., Jongsma, M.L., Bakker, M.J., Hengeveld, R., Janssen, L., Cresswell, P., Egan, D.A., van Ham, M., Ten Brinke, A., Ovaa, H., Beijersbergen, R.L., Kuijl, C., Neefjes, J., 2011. A Genome-wide multidimensional RNAi screen reveals pathways controlling MHC class II antigen presentation. Cell 145, 268–283.

Pieters, J., Bakke, O., Dobberstein, B., 1993. The MHC class II-associated invariant chain contains two endosomal targeting signals within its cytoplasmic tail. J. Cell Sci. 106 (Pt 3), 831–846.

Pond, L., Kuhn, L.A., Teyton, L., Schutze, M.P., Tainer, J.A., Jackson, M.R., Peterson, P.A., 1995. A role for acidic residues in di-leucine motif-based targeting to the endocytic pathway. J. Biol. Chem. 270, 19989–19997.

Rast, J.P., Anderson, M.K., Strong, S.J., Luer, C., Litman, R.T., Litman, G.W., 1997. Alpha, beta, gamma, and delta T cell antigen receptor genes arose early in vertebrate phylogeny. Immunity 6, 1–11.

Riese, R.J., Chapman, H.A., 2000. Cathepsins and compartmentalization in antigen presentation. Curr. Opin. Immunol. 12, 107–113.

Rogozin, I.B., Iyer, L.M., Liang, L., Glazko, G.V., Liston, V.G., Pavlov, Y.I., Aravind, L., Pancer, Z., 2007. Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. Nat. Immunol. 8, 647–656.

Romagnoli, P., Germain, R.N., 1994. The CLIP region of invariant chain plays a critical role in regulating major histocompatibility complex class II folding, transport, and peptide occupancy. J. Exp. Med. 180, 1107–1113.

Saha, N.R., Smith, J., Amemiya, C.T., 2010. Evolution of adaptive immune recognition in jawless vertebrates. Semin. Immunol. 22, 25–33.

Schafer, P.H., Green, J.M., Malapati, S., Gu, L., Pierce, S.K., 1996. HLA-DM is present in one-fifth the amount of HLA-DR in the class II peptide-loading compartment where it associates with leupeptin-induced peptide (LIP)-HLA-DR complexes. J. Immunol. 157, 5487–5495.

Schutze, M.P., Peterson, P.A., Jackson, M.R., 1994. An N-terminal double-arginine motif maintains type II membrane proteins in the endoplasmic reticulum. EMBO J. 13, 1696–1705.

Schwarz, R., Dayhoff, M., 1979. Matrices for detecting distant relationships. In: M, D. (Ed.), Atlas of Protein Sequences. National Biomedical Research Foundation, pp. 353–358.

Sevenich, L., Hagemann, S., Stoeckle, C., Tolosa, E., Peters, C., Reinheckel, T., 2010. Expression of human cathepsin L or human cathepsin V in mouse thymus mediates positive selection of T helper cells in cathepsin L knock-out mice. Biochimie 92, 1674–1680.

Shachar, I., Elliott, E.A., Chasnoff, B., Grewal, I.S., Flavell, R.A., 1995. Reconstitution of invariant chain function in transgenic mice in vivo by individual p31 and p41 isoforms. Immunity 3, 373–383.

Shi, X., Leng, L., Wang, T., Wang, W., Du, X., Li, J., McDonald, C., Chen, Z., Murphy, J.W., Lolis, E., Noble, P., Knudson, W., Bucala, R., 2006. CD44 is the signaling component of the macrophage migration inhibitory factor-CD74 receptor complex. Immunity 25, 595–606.

Silva, D.S., Reis, M.I., Nascimento, D.S., do Vale, A., Pereira, P.J., dos Santos, N.M., 2007. Sea bass (*Dicentrarchus labrax*) invariant chain and class II major histocompatibility complex: sequencing and structural analysis using 3D homology modelling. Mol. Immunol. 44, 3758–3776.

Star, Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrom, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., Wetten, O.F., Lanzen, A., Winer, R., Knight, J., Vogel, J.H., Aken, B., Andersen, O., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R.B., Tina, K.G., Espelund, M., Nepal, C., Previti, C., Karlsen, B.O., Moum, T., Skage, M., Berg, P.R., Gjoen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S.D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S.W., Stenseth, N.C., Jakobsen, K.S.. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477 (7363), 207–210.

Strubin, M., Berte, C., Mach, B., 1986. Alternative splicing and alternative initiation of translation explain the four forms of the Ia antigen-associated invariant chain. EMBO J. 5, 3483–3488.

Takaesu, N.T., Lower, J.A., Robertson, E.J., Bikoff, E.K., 1995. Major histocompatibility class II peptide occupancy, antigen presentation, and CD4+ T cell function in mice lacking the p41 isoform of invariant chain. Immunity 3, 385–396.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

Ting, J.P., Trowsdale, J., 2002. Genetic control of MHC class II expression. Cell 109 (Suppl.), S21–S33.

Turk, D., Guncar, G., Turk, V., 1999. The p41 fragment story. IUBMB Life 48, 7–12.

Turk, V., Turk, B., Turk, D., 2001. Lysosomal cysteine proteases: facts and opportunities. EMBO J. 20, 4629–4633.

Uinuk-Ool, T.S., Takezaki, N., Kuroda, N., Figueroa, F., Sato, A., Samonte, I.E., Mayer, W.E., Klein, J., 2003. Phylogeny of antigen-processing enzymes: cathepsins of a cephalochordate, an agnathan and a bony fish. Scand. J. Immunol. 58, 436–448.

Viville, S., Neefjes, J., Lotteau, V., Dierich, A., Lemeur, M., Ploegh, H., Benoist, C., Mathis, D., 1993. Mice lacking the MHC class II-associated invariant chain. Cell 72, 635–648.

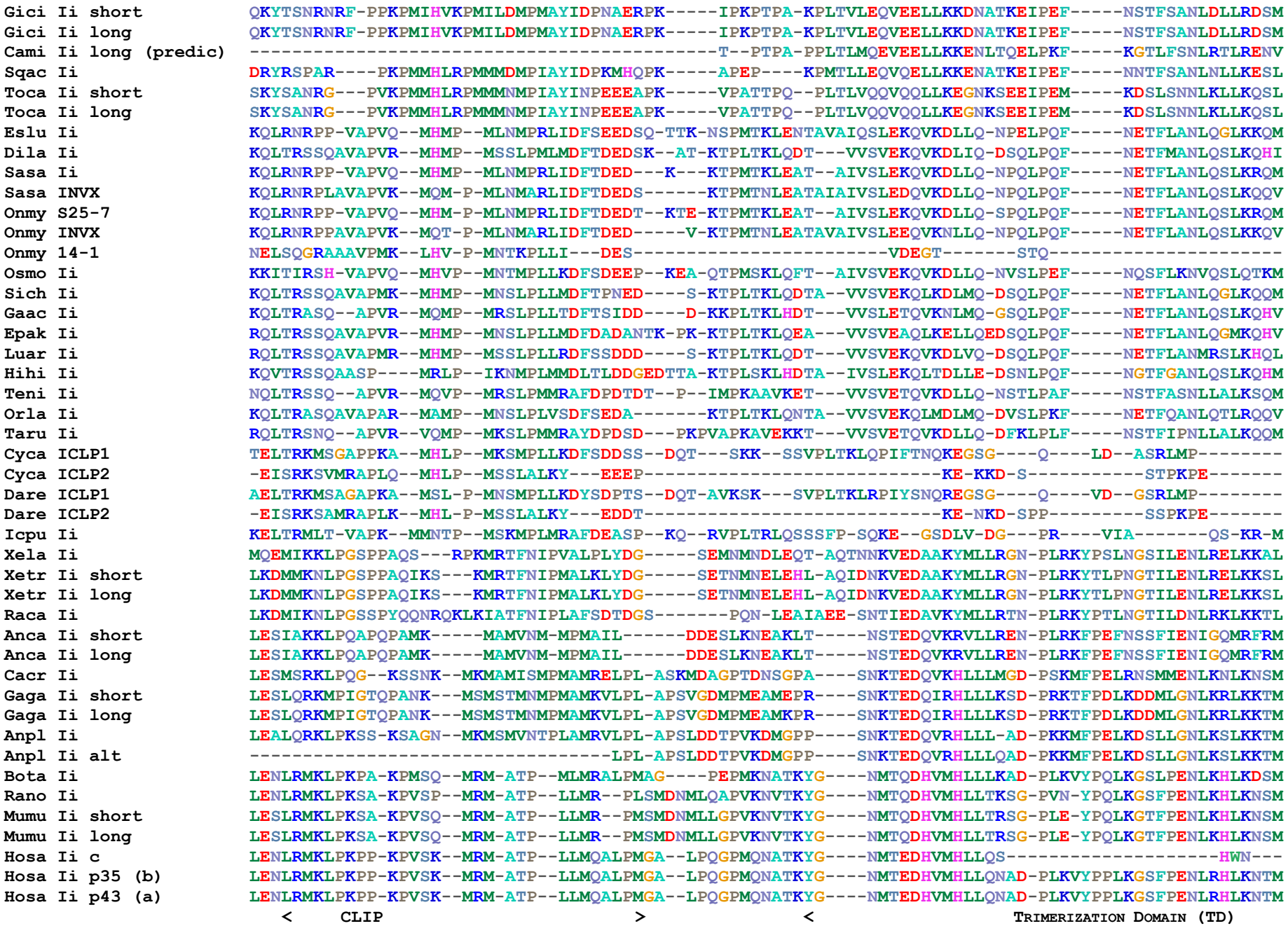
Wahle, E., Keller, W., 1992. The biochemistry of 3′-end cleavage and polyadenylation of messenger RNA precursors. Ann. Rev. Biochem. 61, 419–440.

Ye, Q., Finn, P.W., Sweeney, R., Bikoff, E.K., Riese, R.J., 2003. MHC class II-associated invariant chain isoforms regulate pulmonary immune responses. J. Immunol. 170, 1473–1480.

Yoder, J.A., Haire, R.N., Litman, G.W., 1999. Cloning of two zebrafish cDNAs that share domains with the MHC class II-associated invariant chain. Immunogenetics 50, 84–88.

Zhong, G., Castellino, F., Romagnoli, P., Germain, R.N., 1996. Evidence that binding site occupancy is necessary and sufficient for effective major histocompatibility complex (MHC) class II transport through the secretory pathway redefines the primary function of class II-associated invariant chain peptides (CLIP). J. Exp. Med. 184, 2061–2066.

Gici Ii short
Gici Ii long
Cami Ii long (predic)
Sqac Ii
Toca Ii short
Toca Ii long
Eslu Ii
Dila Ii
Sasa Ii
Sasa INVX
Onmy S25-7
Onmy INVX
Onmy 14-1
Osmo Ii
Sich Ii
Gaac Ii
Epak Ii
Luar Ii
Hihi Ii
Teni Ii
Orla Ii
Taru Ii
Cyca ICLP1
Cyca ICLP2
Dare ICLP1
Dare ICLP2
Icpu Ii
Xela Ii
Xetr Ii short
Xetr Ii long
Raca Ii
Anca Ii short
Anca Ii long
Cacr Ii
Gaga Ii short
Gaga Ii long
Anpl Ii
Anpl Ii alt
Bota Ii
Rano Ii
Mumu Ii short
Mumu Ii long
Hosa Ii c
Hosa Ii p35 (b)
Hosa Ii p43 (a)
< CYTOPLASMIC >< TRANSMEMBRANE >


Gici Ii short
Gici Ii long
Cami Ii long (predic)
Sqac Ii
Toca Ii short
Toca Ii long
Eslu Ii
Dila Ii
Sasa Ii
Sasa INVX
Onmy S25-7
Onmy INVX
Onmy 14-1
Osmo Ii
Sich Ii
Gaac Ii
Epak Ii
Luar Ii
Hihi Ii
Teni Ii
Orla Ii
Taru Ii
Cyca ICLP1
Cyca ICLP2
Dare ICLP1
Dare ICLP2
Icpu Ii
Xela Ii
Xetr Ii short
Xetr Ii long
Raca Ii
Anca Ii short
Anca Ii long
Cacr Ii
Gaga Ii short
Gaga Ii long
Anpl Ii
Anpl Ii alt
Bota Ii
Rano Ii
Mumu Ii short
Mumu Ii long
Hosa Ii c
Hosa Ii p35 (b)
Hosa Ii p43 (a)
< CLIP
>
<
TRIMERIZATION DOMAIN (TD)

Gici Ii short
Gici Ii long
Cami Ii long (predic)
Sqac Ii
Toca Ii short
Toca Ii long
Eslu Ii
Dila Ii
Sasa Ii
Sasa INVX
Onmy S25-7
Onmy INVX
Onmy 14-1
Osmo Ii
Sich Ii
Gaac Ii
Epak Ii
Luar Ii
Hihi Ii
Teni Ii
Orla Ii
Taru Ii
Cyca ICLP1
Cyca ICLP2
Dare ICLP1
Dare ICLP2
Icpu Ii
Xela Ii
Xetr Ii short
Xetr Ii long
Raca Ii
Anca Ii short
Anca Ii long
Cacr Ii
Gaga Ii short
Gaga Ii long
Anpl Ii
Anpl Ii alt
Bota Ii
Rano Ii
Mumu Ii short
Mumu Ii long
Hosa Ii c
Hosa Ii p35 (b)
Hosa Ii p43 (a)
TD
>
<
TG

Supplemental Figure 1. Extensive amino acid alignment of ectothermic vertebrate Ii homologs with select birds and mammals. Putative domain boundaries are shown below the alignment. Database accession numbers of sequences and full species names in Supplemental Table 2.

```
DNA: ATACAGAACTCTGCCTGGCAACCGATGATGCCATTTCCAGCGTCCAATCAG534338

DNA: ATTTCTTGAAAGATTAAACCTGGGCGATTGTGGCCCCCTTTGAGTCCTGTA

DNA: CAGGACATAGAAGTGGGAGGAGGTAGGAAGGGTATTTCCAAATCTGTGAGT

DNA: CTGGCAACTTTCACTTCCCATTTCCATTGTTCAAGAAGCTTCTGAAGAACT

DNA: GAGCCCTTATAGAATCAGTGATCCTGTTGGGCAAAGCAGTTGCTTTCTCAG

DNA: CCTCCAGGCACAGCAGCAGCAGAAGGCAGCAGACCAAGGCCATGGAGGATG
 +3:                                          M  E  D  Dexon 1

DNA: ACCAGAGGAACCTTCTCCCCGGGCCAGGGGCTGGCACTGCCCCACAGCCAG
                              ............................  EST FG760756.1
 +3:   Q  R  N  L  L  P  G  P  G  A  G  T  A  P  Q  P  E

DNA: AACGGTATTTGCATTTCAGCAAGCTTCCTCCTCTATTGATTCCACATTTTC
     ....                                                  EST FG760756.1

DNA: ACTGCTTGGCCAGGGTAGGCGAGGGGGTACAGCGACATGAATGTTTTCCTC

DNA: TGTGCCACATATGAAGAATGAGAAATTTTGTAGTTTTGTGGCACTTTTAAA

DNA: ATGTAACCTTTAACATAGAACCCATTAAAAAGTTACCTTTTGGCATTACAA

DNA: CACTCAGAATCCTCAGCCAGCAAATTTTACACCATTAAGGCAACCTTTCCG

DNA: ATCTCGGTTTGTAGTCGGCTGATTATTTCCTGGTGTGGCGAGCAGGGCTCA

DNA: GGGTGTATCTCTAAACCAGGCATGGGCAAACTTCTGCTCTCCAGGTGTTTT

DNA: GGACTTCTGGCTGTTAGGAATTGTGGGAGTTGAAAGCCAAAACAACTGGGG

DNA: AGCTAAAGTTTGCCCATGCCTGCTCCAAACTATACCAAAAGTGCTACTTCG

DNA: AGGCAAGGTTTAATGTCCCCTCACTTCTTGTCTAAGGGTGCATCTACATTG

DNA: TCAAATTAATGCAGTTTGACACCACTTTAATTGCCATAGCTCAAAGCTATA

DNA: GAATCCTGGGATTTGTAGTTTGGTGAGGCACCAACCCTCTTAGGCGGAGAA

DNA: GACCTTGTGAAAATACAGCTCCCATGATTCTGTGCCACTGAGCCATGACAG

DNA: TTAAAGTGGGCAAAAAGTTGTTTTGGACTTCAACTCCCACAATTCCTAACA

DNA: GCCGGTTTGTGGGAGTTGGAATCCAAAACACCTGGAAGGCCAAAGTTTGCC

DNA: CACGCCTGGTGTAGATGCACTCTTAGAGAAGATCCGAGGGGACCAGGGTTT

DNA: AAATTACGTCTTGCCCAGAAGCACCTGTTTTAGGATAAAGTTCCTAATTTC

DNA: CAAGCACAAAAGAAGGTTTATTTTAGGAAACGTTGTACCCGAAAGGGATAG

DNA: GGAGAAAGTAAACGGCAAATAATCTGCTGTCTTGGGTGTGGGGAGAGATGA

DNA: GTAAAGCTTTAATCCTGCACAGACCTACCTAGAAGTTAGATGCTTTCTACA
```

DNA: CTGTGTGGCTTGCTTCTGAGTAAACATGGCTAAGATTGTGCAGCACAGGAA535715

DNA: TTTCCTGCCCTTTTCCTGCTGTTTTGCCCTTGTGCAACTGGAAAATAAGTA

DNA: TAGGCTTATGTGTTTCATGCCCCTTCATATTTTAATGCCCTAGTGGTAAAT

DNA: TTCTATCCAATCTGCCAGTAAATATACAATAATTTTACAGAATTCTCAGCC

DNA: TAGCTGGAAGAAAGCAAAGTTAAGGCAAATTGCACTTTATTGAACTACATA

DNA: TGTCATTTTGTATCTCTTCTTATGTTGAATATCAGATTTTGTATGCTTGTC

DNA: CCCAAAATCTATAACCACATCTAGCTTATGCATTAGGACCAAGGCTATTTT

DNA: TCATTTTTTGTCATAAATGAACGAATGTTGCACATTTCTGTTTTTCTGTTT

DNA: TTCTCTGAAAACCTACCTCACTCATAAGGAAGGCGGACAAAGAGGTATTAT

DNA: CACTTTCTCTTTTCATACATACACTTCATTGCGATGTGTGTGCCCAACTTT

DNA: CCTTCAAGAAGCATCAAAATAGATCATCATCCGGGCGTCCCCTGGGCAGTG

DNA: TCCTTGCAGACGGCCAATTCTCTCACACCAGAAGTGACTTGCAGTTTCTCA

DNA: AGTTGCTCCTGACACGAAAAAAATTATTCTCACAACAACCCTGTAGGGTAG

DNA: TTCAGGTTAAAAGTAACTTCCTTAAGATGACCTGGTAAGTATCATAGAGGA

DNA: ACAGGGATTTGAACCAGGGTTCCCACACTTTCATGTCCCTCCGCTGCATTA

DNA: TTTCTTAGTGTGCCTCAGTTTTTCCTTTTTTTTAAATGCTATTAGGCTCTT

DNA: TGTCTTCTTGATTTGATACAGCAAATAATAATCTTTAGCAATAGTGACCAT

DNA: TTCAGGGAATCACAGAGCTGAAATGCCTGCAACTGGCCATAGATGCTAAGT

DNA: CAAAGCCAAATAGATTATAGAAGAGTGATTTTTTTTGTGTGTGTCAAGAGC

DNA: GACTTGAGAAACTTAAAGTCACTTCTGGTGTGAGAGAATTGGCCATCTGCA

DNA: AGGACGTTGCCCAGGGGACGCCCGATATTTTGATGTTTTGCCATCCTTGTG

DNA: GGAGGCTTCTCTCATGCCCCCACATGGGGAGCTGGAACTGGCAGAGGGAGC

DNA: TCATCCACACTCTCCCCAGGTTGGATTCGAACCTGCAACCTTCAGGTCAGC

DNA: AACCCAACCTTCAAGTCAGCAGTCCTGCCAGCACAGGAGTTTAACCCGTTC

DNA: CGCCACCGGGGCAATCTTAGGGATGCTAAATCAAAAAGCCAATCAGATTAT

DNA: AGAAGAGTGATGACATCATTTGTTACTAGGAAGAAAGTTGTCCAAAAACTT

DNA: CTCATCCGCACTTGAACATTACTTACAGGCTCATGTTACATTTAGACATGT

DNA: TTTGTTGTTGTGTGCTTTCAAGTCATTTCTAGGTTGTTGTGACCCTCAGGC

DNA: AACTTTATCATGATTTTCTTGTCAAGTTTTGTTCAGAGAGGCTTTTCCTTT537143

DNA: CCTCTGAGTCTGAGAGAGTGCGAGTTGCCTATGGTTATCCAGTGGATTTCT

DNA: ATGGCTGAGCACAGATTCAAAACCCTGGTCTCCACAGTCTTAGTCCAACGC

DNA: TTTAACCACTGCATCACACTGGCTAAAGTACTATTTTTTTCCTGAAGAGTT

DNA: CTTTTAAATCAACAAGTGTTATGATCCACTTATGAATCCGATCCAAAACAA

DNA: TATAAAACAAATAGCTCTAACATGTGCCTTTTAATCTTGAATTAATCTGAA

DNA: GGTAAGTAGTTAATTAATTGGGAATAAATAGTAGGATTTCATTGTTCAGTA

DNA: CAATAAAAAAGCATTTTAATAAAAGTGCCCTGAATAATGAGGCAGTGATTT

DNA: TTAAAATGTTAATTACCTCTAATTTTGTTACAGAATCAGGGGAGAGAATCA

DNA: ATTATGAATGACAATTAAAAAGCTTTAGAAAAGACATTTCTCTCACACAAA

DNA: GAATTCCTCTGGTTTATTTTTGGTATTTCAAGGCTATTTCAGGACTATGCC

DNA: CTACAAAATGAGGTTGCATCTACACTATGGAATTAATGCAGTTTGACCCCA

DNA: CTTTAACTGCCACTTTAACCGAACTGTGGGAGTTGTAGCTTTACAAGGTCT

DNA: TCAGCCTTCTCTGTCAAAGAGTGCTCACCAAATTGCAAATCCCAGGATTTT

DNA: ATGTCAGTTAAAGTGGGATCAAGCTGCTTTAATCCTTCAGTGTGGATGCAG

DNA: CCAAGGTCACCGCCAGACACAGCAACTGTAACAGTTATAATAGGGAGCAAG

DNA: CAGAATACTTTTCAAAGCTTCTTGGAACACAAAAACCATAGGCAAATTGAG

DNA: TTGTTGGCAGCTGTAATCAGAACGCTCCCTTTAATGCCTGCGTGGAAGAAC

DNA: CAGATTGTGAAGGAGGGACATTTTTGAACAAGCTGCCTGCTGCTAACACAC

DNA: ACATACATAAGTACGTCCTTGAATGAAGAACATTTTCTCCAGATGAGCAGA

DNA: CTTATGCAGCAATAACTGAATAAGTCGAATAAAAATCCTACAGTTGATGAA

DNA: TCTACACTGTAGAATTCAAACTTGTTTTCTGCTGCCTTGGAGACTAGGATG

DNA: GGGAACATTGGCTTCAAACGACAGGAAAGGAGATTCCACCTGAACATCAGG

DNA: AAGAACTTCCTCACTGTGAGAGCTGTTCAGCAGTGGAACTCTCTGCCCCAG

DNA: AGTGTGTTGGAGGCTCCTTCTTTGGAGACTTTTAAGCAGATGACCAACTGT

DNA: TGGGGGTGCATTGAATGCGATTTTCCTTCTTCTTGGCAGGGGGTTGGACTG

DNA: GATGGCCCACAAGGTCTCTTTCAACTTTATGATTCTATGATTTTATGCAGT

DNA: TAGATATCATTTTAACTGCCATGGCTCAATGCAATGGGAGGTGTAGTTTGG

DNA: TGAGGCACCAGGCTTCTTTGGCAGAGAAGGCTAAAGGCTTTGTAAAGCTTT

DNA: GGCTCCCATGATTCCATACCATTGAGCCATAGCTGCTAGAGTGTTGTCAAA538622

DNA: CTGCATTAATTTTACAATGTAGATGCACCCCAAAAGTCCCTTTAAGCAGGT

DNA: TTTTTTTTCGTGTCAGGAGCAACCGGAGTTGCTTCTTGAGTGAGAGAATTG

DNA: GCTGTCTGCAAGGACGTTGCCCAGAGGACGCCCGGATGTTTTGATGTTTTA

DNA: CCATCCTTGTGAGAAGCTTCTCTCATGTCCCCGCATAGAGCTAGAGCTGAT

DNA: AGAGGGAGCTCAACTGCGCTCTACCCGGGTGGGATTCGAACCTGGCAGCCA

DNA: GAGCCGGCCCTAGGAAATGTTCAAGTACAGGCGAACAGAATTTTGCCCCCC

DNA: CCCCCCCAAACCAATCACTGAAAAACAAAAGCGTTGGATAAGCAAAAATGT

DNA: TGGATAATAAGGAGGGATTAAGGAAAAGCCTAATAAACATCAAATTACGTT

DNA: ATGATTTTACAAATTAAGCACCAAAACATCATGTTATACAACAAATTAACA

DNA: GAAAAAGCAGTTCAATACATGGTAATGTTATGTAGGAATTACTATATTTGC

DNA: GAATTTAATACCAAACATTGAACTGGGATATAGGGCAGTGTGGACTTAGAT

DNA: AACCCAGTTCAAAGCAGATATTTTGGGTTATTCTGCCTTGATATTCTGGGT

DNA: TATATGGCTGTGTGACAGCTACTCATAAATTCCCACCCCCATGCGTCCTAC

DNA: ACATTCTGAGAATCTGCTGGAATTCTATGAAATGATCAAATGATGATGCAT

DNA: TACTTGAAATATGTAGTATTTGTGCTTTTCTAAATGTACACAGAATTATTT

DNA: TCCAGCAAAGATGCAATAGCAAGTGCACTTACCTACCAGGCATGGACAAAC

DNA: ATTGGCCCCCCAGGTGTTTTGGACTACAACTCCCACAATTCCTAACAGCCT

DNA: CAAGCCCCTTGCTTTCCCCCCTTAGCCATTTAAGATAAGGGGTAAAACCTG

DNA: GACACAGTATATTATTTGAGAACATAGAAATGCTGGACCACTCTCACAACC

DNA: ACCATGTCAGACAACAGAGAAGTCATTGAAATCCACAAGCATGTGGACAAT

DNA: TCCAACAGAAAGGAGGAATCCATGACAATGAAGAGTATCTGGCTACCAGTA

DNA: TTTAAAAACTCAGGACAGTAAATAAAGAACAACACACACAAAAACAGAGGA

DNA: ATTCCAGACATGAATCAATCAGGGGCAGCTCACACCTCCCAACAAAGGATT

DNA: CCCTCAAATAGGAAGCAGCCAGGCTTTGAAGCTGCAAGGCTATTCAATGCT

DNA: AATCAAGGTGGCCTATTGCAACATTCACACTTGCCTCCAGCAGACAAGAGT

DNA: TCTTTCTCTTCCAACCTGGACTTTCAACAAATAATTAAGCCCCACTTCCCT

DNA: AGTTTCGAACAGACCTCACAACCTCTGAGGATGCTTGCCATAGATGTGGGT

DNA: GAAACATCAGGAGAGATTGCTTCTGAAACATGGCAATACAGACTGGGAAAC540050

DNA: CCACAGCAAACCAGTGTTTCTGGCCATGAAAGCTTTTGACAACAGGTTCCA

DNA: GAAGTATTCTCTCCTGATTGATTAACACCTCCCAACAAAGGATTCCCTCAA

DNA: GCAGGAAACAGCCAGGCTTTGAAGCTGCAAGGCTTTTCAATGCTAATCAAG

DNA: GTGATTAATGGCAACAGTCACACTGGCCTCCAACAAACAAACAAGAGTTCT

DNA: TTCTCCCACCCTGAACATTCCACGGATATGTGTTATGAAATATTTTAATGG

DNA: TCGGAATCATTGGGTTGCTGTGAGTTTTCCAGGCTGTCTGGTCATGTTCCA

DNA: GAAGCATTCTCTCCTGACGTTTCACCCACATCTATGGCAAGCATCTTCAGA

DNA: GGTTGTGAGGTTTGTTGGAAACTAGGCAAGTGGGGTTTCTATATCTGTGGA

DNA: ATGTTCAGGGTGGGAGGAATAACTCTTTTCTGTTTGAGGCAGGTGTGAATA

DNA: TTTGAATTGGCCACCTTGATTTGCATGGAAAAGCCTTTCACCTTCAAGAAC

DNA: TGGCTGCTTCCTGCCTGGGGGAATCCCTTGTTTGGAAGGGTTAACTGGCCC

DNA: TGATTGTTTCTTGTCTTCTGAGTGTTGTTCTTTTTTTACTGTCCCAATTTT

DNA: AGCTTTTTTAATACTCAGATTTTGTTCATTTTCATGGTTTCTTCCTTTCTG

DNA: TTGAAATTCTACAGATACATAAACGCCACTTGCCTAGTTTCCAACAGATCT

DNA: CACAACCTCTGAGGGTAATATATATATATTTACTGCATTTTTACCCCGCCC

DNA: TATCTCAACCTCAAAGGGAACTCAGGGCGGCTTCCTGGTTGGCTGGAGCGG

DNA: TACCTATTGATCTACTCACATTTTGCACGTTTTTGAACTGCAAAGTTGGCA

DNA: GAATCAGAATATAATATAGAATATGATTACTGTGTTATGTTCTGAAAATCT

DNA: GTATTATTATTACGGTATCATTATGCTCTGCTAATATTGTGCGATGCTAAT

DNA: AATATAATACATTGTACGTAAATATGTCAATATTATTCGTAAACCGCTCTG

DNA: AGTCCCCTTTGGCGTGAGAAGGGCGGGATATAAATGTAGTAAATAAAATAA

DNA: AATAAATAAATAATGAATACACTCTCATTATTGTGTTTATTACATTAATAT

DNA: GTATGTCAATATGCATAGTATTGTGTTTATCACATTAACAGGTATATCAAT

DNA: ACGTAATAAACTCCCATTATAGTGTTTATTACATTAATATGTATGTCAGTA

DNA: TGCAGAGTATTGTGTTTATTACATTAATAGGTATATCAATATGTAATAAAC

DNA: TCCCATTATAGTGTTTATTACATTAATATATATATTAATATGCACAGTATT

DNA: GTGTTTATTACATTAATATGTATATTAATATGCACAGTAGTGTGTTTATTA

DNA: CATATATATATATATATATATATATATATATATATATATATCTCAATCTCT

DNA: TTAACCAGGGCTGGATTTTGTGGTTTTCTTCATGGGGTTGGACTAGATGTC541529

DNA: CTTTGGGGTTCCCCCTCTCCAATCTAAGATACTTTCAGGAGCTAATGCAAA

DNA: GCCTTTGGAGTCTTTTTGGGATTGTTTACATTCAAAAATGGGGGAATCTCC

DNA: ATCTCTCTTCCCATAGTAAATCCCAATATTATGTATATATGTGTGTATATG

DNA: TATGTATGTATACACACACACACACACATTGCAAACAAAAGGATCCCCCTC

DNA: ATTTATGTTGCAAAGGGAAAAGAGAAAGAGCAGCAGGTAGTAGGAATGGTT

DNA: GGGAGGAAAGAGAACAGAGAGAGGGCCCAAAGGAAAAAGCATCCAAGGTCC

DNA: CAAGGCCTGGTCCCTCTTCCAACAACAATAAACCCCGTTAGCAAAGGCCCC

DNA: CCGCCCCCCCCTCCCCAAGCAAAACAAGCCCCTTAAATGGACTCACAAATA

DNA: GAAAAATACTTACTAAAATGGGAAGGCAGTGAGTGGGTCACAGTGGCTTGG

DNA: CTTCTCCCCACACTCCCTCGCTCCTTTCCCCCTTCTCCCTTCTTCCTTTCC

DNA: CCCCAATCCCACTGCTTGCTTTACAGGGGAGGCAAAAGGATCAAAGGGACC

DNA: CAAGCCAGCTCCACGTGGAGCCTGATTCCTTCTCTTTGGCTGCTGCAAGGC

DNA: AAGTTGCAGCCAAGTGGGGGGGAGGTGCTGGGCTTCCCCGACCTCAAACTG

DNA: CCCAACATCCTCCTTCTCCAGAGGGAGCTGCTGGATGCTTTGCAGACTGGA

DNA: GAAGTGCTGTTTCACGTCTGCAAACGGCCCGACGCCTTCCTTCTAGGGAGG

DNA: AAGGCACTGGGCTCTGCAGAGCAGAGGCACAGCAGCGGAGGCAGGAGTTGC

DNA: CTTGATGGCGCCCCCTACAGGATGGCGCCACAGGCAAATGACTAGTTTGCC

DNA: TCTCGGTTGAACCGTCACTGCTGGCAGCCTTCAGGTCAGCAACCCAACCTT

DNA: CAAGTCACAAGGCTTTTATCCCCTAGGCCATCGGAGGTAACATCTTGAGGA

DNA: AATTACAGCTTGGGTGAGAAAGATGCTCTCTCTGCTTTTACTATCTAACCT

DNA: ATCTATGGAGTTGTTGCCTCAAAGTAAAACCATGTTCTCTTTTTTTAAAAT

DNA: TTAAATTCCAGAGTAATTAGATTCTAGAAAAAATGACATCTGATAAGAAGC

DNA: TGAAAAAAAGAGCACCATTGTATGCCTTCCTGTGGCATGGGTCCTCTCTGT

DNA: CTGGTACCATACTAATTGCTTTGCTAACTCTGGCTGATTATTTACCAAACT

DNA: AATAAAGGCATTTTACGGGGCGAGACGCATTGCAACAGGTTTAACCACTGG

DNA: TTATTCTGTTCTAAGAGAAACAAGCCTGCCCATAATTTTTGCCTGAGTGTG

DNA: AAGAGAGTCAAGGGCAATAGCTTTTCTCGTCTACTTTCAAGTTCCGTCTAC

```
DNA: CTAGCAACAAGTGATGTCACAAGATGATTAGAAAATTATTGGTCAGAATTC542957

DNA: AGCCATTTCCTCAAGCTATTTTTCATTTCAGTTATTATGTCTGAGGAGACA

DNA: TTTTAGTTCAGATAACACACAGTCACAGATTAGTAAAATAACAGAGAAGAG

DNA: GAAGGGAATATGAGCAAACAAATGTGGACATCCTCTTTTCTGACTTGTTAT

DNA: CTTGTTCAGCAATTAGTTTTTAGCCATAGTGTATGTTCTGTTTATCTTAGA

DNA: CTCAGGAGATATGCCACACCTGGAATATTTACTTAATAATGTCCAAATGAT

DNA: CTAATCCAAGTTAGTAAAAATTTCCAAACAGTCTTGTGACATCTGAAAAAC

DNA: TGTGGGGTTTCATTTGCCGTAAACATTCATAGACTCCAAGTGCATGGACTC

DNA: CAATACAGGTGGTTGCTGTCCACAAAAGCCTATGTCGGATTAAAATAAAAT

DNA: AAAATAAAAAGTAATTATTAGAGTGATGCTGGAGATAGGCACAAGGGCATC

DNA: GCGACCCATGAGAAATTTGCTCATCCCATATTTTCCTTAATTCCCTCAATT

DNA: TCTAGCATTGCAGTAGATTAAAATTTATTTGAACATTTAGAAACAGAATTT

DNA: AGGCACCTGAAGAAAAGGAGCAGTCAATGTGAACAAGAGATGGGATCACAG

DNA: AAAATAATAGTTTTAAGTGTGAATGTTTTTCAGTATCTCATTATCTGCAAT

DNA: GCTAGAAATTTGGGGACTGAAGGAAACTTTCATTTGGGTGGGGCAGAAGTC

DNA: TGGTTTTGGGGAGCAATGTCTGTATTTTGTCCTTACATCCCTGGCAAGACC

DNA: CCTCACTTGTTTATGTTGAGATGAAGTTTCTTACAGATTATTTTGATTCTC

DNA: TTCTGCAGCAGTGTATTCTCAGGTTGCATCTACATTGTAGAATGAATGTAG

DNA: TTTGATACCACTTTAACTTTCATGCCTCAATGCTATGGAATACTAGGATTT

DNA: GTAGTATGGTGAGGCACCAGCACTCTTTGGCAGAGAAGGCTACTGATGTTG

DNA: TGAAACTACAACTCCAAGGATTCCATAGCACTGAGGCATGGAAGTTAAAGT

DNA: GGTGTCAAACTGCATTCATTCTATAGTGGAGATGTACCCTCAGGTGCGAGG

DNA: AGATACAAGGGATAGTTGAAGGAGGACCCACATACATCTTTTCTCTCTTGC

DNA: AGGGGGTCCTGCAGCCGTGGATCCATCTATGCAGTTGTATCTGTCCTGGTG
.................................................  EST FG750983.1
.....................................  EST FG760756.1
 +1:  G  S  C  S  R  G  S  I  Y  A  V  V  S  V  L  Vexon 2

DNA: TCCTTGCTTATTGCCGGCCAGGCGGTCACTGTCTTCTATGTGTATCATCAC
.............A.....................................  EST FG750983.1
T
.................................................  EST FG760756.1
+1: S  L  L  I  A  G  Q  A  V  T  V  F  Y  V  Y  H  H
```

```
DNA: AACGAACGGATTACTAAGCTGAGCAAGGACACCACGGAGCTGAAGCTAGAA544232
.................................................  EST FG750983.1
     ....................................G..........  EST FG760756.1
A
+1: N  E  R  I  T  K  L  S  K  D  T  T  E  L  K  L  E

DNA: TCAATTGCCAAAAAGCTTCCTCAGGGTTAGTGGCTGCTGCTGCTCTTTCCT
.........................                          EST FG750983.1
     .........................                         EST FG760756.1
 +1: S  I  A  K  K  L  P  Q

DNA: TTTGACTTTCTTTACAATTTTAACTGCTGGGTGTACAATAGTAACCATGGA

DNA: TGGTTGCTTAAGAGCTCAGTGTCTCAGGACAGAGACCTCCTCCACTTCCCA
P
DNA: GGGGAAGGAATCTCAGTGCAAGAGTGCTATCTTGGAAAGTGTATATGCATT

DNA: TGTTAGAACTATTTTTATTTCAATGTCATACACTTAGAGTTTAAGAGTATG

DNA: TGTCAAGTATATCAGCAAAGGATTATAAAATATAAATTCACTTATTTTGTT

DNA: CCTGAACTTGTGATTCTCTGTTTAAAAAGCATGTTTTTGGAGGAAACTATT

DNA: TTTGTAAAAATTGTCCTGCATGTTTTATACCTGTTTCTCTTCTTATAGCCC
                                                     ...  EST FG750983.1
                                                     ...  EST FG760756.1
 +3:   Pexon 3

DNA: CCCAGCCAGCGATGAAGATGGCAATGGTCAATATGATGCCCATGGCTATTT
.................................................  EST FG750983.1
     .................................................  EST FG760756.1
 +3:   Q  P  A  M  K  M  A  M  V  N  M  M  P  M  A  I  L

DNA: TAGATGACGAGGTAAGCATGGCAGGAAGAGGTAGTTAAGGGAGGTTTCAGC
     ...........                                        EST FG750983.1
     ...........                                        EST FG760756.1
 +3:   D  D  E

DNA: GCTGAATACTTTGCTGTTCAATCCATAAAACTTTCCCACTACAATATTAAT

DNA: TTTTATGAAATGGCTATGGATTTAGGGAGGGCATTCGCCATATCTTTTCAT

DNA: GCCTTTCTTTGGTACAAAAACAAACAGCAGGAAACACTAGTTTGGTAATAC

DNA: GTTTGAAGCAACCCAGCCTCCTTCCTTGTCTCGTCTCTAATTAACCATCAT

DNA: GTGTGCCTTGCCACAATCTTCTTGAAGCTTAGGGAAAATCCATTATATATC

DNA: TCTCTATGATTGGATGCTTGTCTTGTTAGGACAGTCATTCCTTTCCTTCTA

DNA: AACCAGGGGTCCCCAAACTTTTCACAGTCCCTCAGACCGTTGGAGGGCTGG

DNA: ACTATAGTTTTAAAAACACTATGAATAAATTCCTATGCACACTGCACATAT

DNA: TGTATTTTGAAGTTAATAAACAAAATGGGAACAAATACAGCCTCAATATTA

DNA: ATAATCATAATCATAATAAAAATAATAAAGAGGGTTGGAAGAGACCCCTTG
```

```
DNA: GGCCATTTAGACTAACCCATTTCTGCCTTTGTGCACCAAAAGCACTAGCAA545303

DNA: AGCACCCCTTAACAATTAATTATTAATTAAATAATTATTAAAATACCATTA

DNA: TAAATACAAGCAAAGCTTTGGAAGGCATGCACCGGGTGAGGGAGGAGGCAA

DNA: GAACGGTGCACAACTTTCCCTCCTCCTTCCCGCCACACGCATCTTCTTCAG

DNA: CAGAGGAGGAGGGAGCCACTGCCATGGGAGCCAGATAAATAGATTCGATGG

DNA: GCCACATGTGGCCCACGGATCCCTGTTCTAAACTATCTTCCCCCACTCAAA

DNA: GTCATACTAGCACAATGGATAGGTCAGATGGCAATTACTAGGCATCTGCTT

DNA: CAGTACTGTCCTTTTTGAAGACCTTTTCCAGCTTCAGGAACAAACTGGAGG

DNA: CCTGTACATACATTCTCTCTGCAACAAAAGAGTAAGTTGTTCTTTAGTTTT

DNA: GATCTAGTCCAACATATCTGCTGCAGTATATGTGAAAGATCAGAGACAATT

DNA: CTTTAAGAAGACCAAGAGTGGATAACCTGTTGAACCATTGGCTACAATGGC

DNA: CAAGGCTGCTGGGAGTCATGATCTAAGAACTTCTGGAAGTCTAGGACCTGG

DNA: AGTAAGACCCTAGCAAGAAGTATTTCCACAAGGGCTCAGAATGAAGCAGGA

DNA: ATTTCCATCCTACAGATGGTGAATGTGCGACTTTCCAGGACTGGACTATAG

DNA: CGCCTATCAGCCATCTCAAACAATGAGCGACAATGGAATTTGTGGTTTGAT

DNA: GACATATTTAGTGCCATGTGTTCCTCATTCCTTGTTTGGTCTCTGTTCTAA

DNA: GCAATGTATGTTTGTAGCTTGTACATCCTTTGGGTCAACTGCCACCACCCT

DNA: ACATCAGATATATCCTGTTTGTAATAATGGAAACTGGAGATCTAATTGTCA

DNA: CTCTTCAGATTTTAACTTTTCTTTTTGTCCCATATGTAGTCTTTGAAAAAT
                   ...........CEST FG750983.1
                                            ...........  EST FG760756.1
 +1:                                        S  L  K  Nexon 4

DNA: GAAGCCAAGCTGACCAACAGCACTGAAGACCAAGTCAAGCGTGTCCTATTG
..C.............................................  EST FG750983.1
D
     ..................................................  EST FG760756.1
+1: E  A  K  L  T  N  S  T  E  D  Q  V  K  R  V  L  L

DNA: GTAAGAGTCTCATTCCCTCACCATGACAACTATGATTCTGCTTGATCATTT
`
DNA: GAAGACATAGATTCATTTTAGACTACTTGCAACTTAGGCTACTTGCACCAG

DNA: AGCTGTCAAATTAGTTAGACCTTTGATCTGATACATCAAGGCAAGTTTAGA

DNA: TGAATTGTTAATGTGTTCTTGCTGCACAGCCATCGTATTTGACTGGTATAT

DNA: GAGTGTTGTTAGAAATGAGAACTAAGACTGATGCTTTGTAATTTTTAGAAT
```

DNA: GTATATTTACAGTGGTCAATGACTGGATGTGACACCCCCCCCCCCACACAC546578

DNA: ACACACACACTTCCATCTTAAGCCACTGCCTGTTTCAAGATTTCTGAGGGC

DNA: ACTGGGAAATCATGATTGTGGCACCCTCCTGGCCATTGAGATGAGCAGCTG

DNA: TGACAGAGGCCTAAGGGGCCTCTTCAGCTGCCTGTCTCAGAATACCCTGAG

DNA: CATACCCAGTGCTCCGATTGGCAGCATCTTGAAAGCCCTTTGGTTGGGCCC

DNA: CTGCCTGGCTTAAGAGGCAGGTTTTGGGTGGTGAGAGGTAGAATTTGCAGA

DNA: TGTTAAAGCAATGTATTTATTTATTTATTTATTTATTTACAATTTTTCTGT

DNA: ATTTCTATCCCACCCTATCTCACCCTAAGGGAAACTCAGAGCAGTTTCAAC

DNA: TTGGCACAATTCGATGCCACATTGAAACATAGTGGAGTCAGCTGCATATAG

DNA: CTGCCAGTGCTTCATTGCTTTTATGTGACAGCTTGCAGTGTTCAATTGAAA

DNA: AGCACTTTGTTTCCTCCTTCTAATACAATCAGTAGGGAGCCTTCAAACTCA

DNA: AAAACACCAAAAGGATCATTGCTTCTATGCATGCAGTTTGTCCATGTCTGT

DNA: AGTTACAGTGTTTCTGGATATTTATCATCGAACGAGTCCTAAGTGCCATAC

DNA: TATTAATTGTGAAAATTGCACAGACTTGGAATTGTGAATTCTGTGGAAATC

DNA: CTCCCACATCCAAAGGACTCACATTAATTTTTTTCTATATGACGTTGGACT

DNA: ATATGGGCCACAAGAAGTTTAGTTTAGAGCAGAAGTTTGTCTGCTCTTTAT

DNA: GGGGAGCTATGGTAACAAATTGGACTAAGTTACTAAAGAACGTTCAACATT

DNA: AAAAATTCAGATTATGGGATTTGTCATACATACTTTGCATACTTATTATGA

DNA: AACATTAGGGATGACAGGTGCCTTTCTATCTCAGGGATCTCTCCTGTGATC

DNA: CTTGCATTAAGGTTCCTGTTGTTTTGTGTCTTGTGTTTTGAGTCATTCCCA

DNA: AAACATGTCAACCCTAAGACAAACCCATCACTGGGTTTTCTAGGCAAGATT

DNA: TGTTCAGGGGAGGGTGCTTTTTCCTCCCTCTGAAGCTGGGAGAATATGCCT

DNA: TTCTGAGGGATTAAAACCCTGGACTTCGGAGTCAGAATCCAGTGGTCAAAC

DNA: CACTATATTACATTAATTCTCCAGGGTGTTAGTCCAAGTTATGATGTGGAC

DNA: TGTAAATGTTCAAGAACTGAAAGTAGAATTATAGAGAGGGAAATGCTATCT

DNA: GTGTTTCCCCTCTGAAATCCCCTTGGCTCAGCTTGGGGAAAGCAATGTGAT

DNA: TCTCACTGTCAGTAACCTCTGATTGAATCACTGAAATGGGGTTGTCTCTTT

DNA: CATGTGCTGCCATTTCTTCAGACCATACTGTTCTCTGAGGATTGTTTCAAT

```
DNA: CTTCCTTGAACAACCAGAGCTTGTAGAACTTAATTTTGGATGACAGCTTCC548006

DNA: AGAATTCCCCCTCTTGACTGCAGGAGCCTGCGGGCTGTAATGCAAAAAAAG

DNA: TAAATGTTTCCAAGCTTTGTCCATATTAGAGATGACAAGGTGATCTTGCTG

DNA: TCCATTTGTTAATATGGTATATGCTTCCTCTACAGCGGGAAAACCCTCTAA
                                      .............  EST FG750983.1
                                      .............  EST FG760756.1
 +3:   R  E  N  P  L  Rexon 5

DNA: GAAAATTCCCTGAATTCAATAGCAGCTTCATTGAAAATATTGGTCAGATGA
.................................................  EST FG750983.1
     .................................................  EST FG760756.1
 +3:   K  F  P  E  F  N  S  S  F  I  E  N  I  G  Q  M  R

DNA: GATTTCGCATGGATTATGAAGACTGGAAGGTAAATTTTGCTATCCTTTCTA
     ....C....C.....................                    EST FG750983.1
T
     ............................                       EST FG760756.1
 +3:   F  R  M  D  Y  E  D  W  K

DNA: GAATTTGACAAGGTATTTTGGATAATCCTGTCTCCATTCAGTCTTATGATG

DNA: TGCATAGTGGATTTGTAGTGCAGAAGACAGCTCTTGTGTGAGAGAAGCTAC

DNA: TTCCCAGCAATATTGCTTTTTGAATGTTTTTCCTCCTGTGGGAAGAAACTT

DNA: CAAAGGATCAAGATTCAGTATAATCAGAAGCAGAGTCAGGGAGCTTTATGG

DNA: TGTTTCTGTGATCTAAGTCATATTTGCTACCTTAGTTGGGGTTTAGATGTT

DNA: TAATTGGACCCTACTTTGATAAACAGCAATAAACCCTACTCATCAGCTGCC

DNA: TGGGGTCGCACTGTAATAAATCACTATAAAGAGATGCTACAAAATGGCAAG

DNA: AATCTTATTAGTGATTCACAAGTTTAGGTCCATCTTATGCTTCTATTTTTT

DNA: GCCATTGTTGATATTTTTTTCTTCCCTCTTTAATGTTCAAAGAGCCCCTCC

DNA: CCTCCGCATTACTGTTATGATAGTCCCTAACAACCATTCTGTTGATGATGA

DNA: AAAGTGAGAGGAGGACCGGAGAAGCATCTGTCTATGTCTGTATGGCTGAGG

DNA: ACATCATCTGCTAAGATATTTGGTGGAATTTATCCTCAATTGGTATCTGAC

DNA: TCCCGTAGCCAGGTGCTTTTCTGATCCGTACTGTTCACTAGTAATGAAAAA

DNA: GCCATGGCATGTCTGTGTGCACACACATATATGTGCTATTGCAGCACTATG

DNA: GTGAGTCTTGAAGCTTGTGATATGGAGCCACAGTGGATTACCCCATGCAGA

DNA: TCTCAGATGCTGTGCTCAACCACATAGACTTTTGTCCCATTTCAACACAGA

DNA: CTTAGTTTGTCTCTACTGCTCTCACTGATCTCTGAAGCAAAGCAAATAGAA

DNA: ATCATATTTCCCTGCGACTCTCTTGCAGGCTTTTGAAACCTGGATGCACAA
```

```
                                  ......................  EST FG750983.1
                                  ......................  EST FG760756.1
+2:  A  F  E  T  W  M  H  Kexon 6

DNA: GTGGCTCCTCTTCCAAATGGCACAAAATGCGATCCCAGAAGAAAAAGGTAT549230
.............................................      EST FG750983.1
     ..............................................      EST FG760756.1
+2:  W  L  L  F  Q  M  A  Q  N  A  I  P  E  E  K

DNA: GGTGGAACGGCAGGACTGAAAAGGATTCTGCAGGAGACTAAAAGCATGTGT

DNA: TAACAATGCAATTTAAATTTAACCAACTAAAAATAAAAAACAAAAAACACA

DNA: AACCCTGGAATAAAATAAACAAAACAATGCAAAACCAACCAACCCCCTAAT

DNA: TCTCCCACCTCATAACTTTTCCCCACCCCTGTATCCCCTTCCTTTTTTAAA

DNA: AAAGTAAAAAATTAAAAAAATTTTTTTAAAACAATAACAGAATGTTAATTT

DNA: TTTAAATACATGCAAAGTGATTGGCAGAGCACAAATAATTTAAAATAGTAA

DNA: TTCTCTGAGGATGTGCCTTTGAGGAAAGATGGAAGACCTGCTGTTTGAATT

DNA: AAGACCCACTTTCAAGATGGATTTTTTTTTTTAAAAAGTTCAGTCCATTGA

DNA: GATATCTGAACAAGTTTAATAGCTTCCATCCCCCAATGCCCCATAAATGAT

DNA: TAATTAACTATCATGAACTGAACTTGCACACCAACACAGGCAATGACTTAT

DNA: CTGTGACAAATCTGGGCACAAACGTAATGTTAAAAAATAATGATCCCTCCT

DNA: CCTGTGATTTTAGCGGTTTCTGAAATAAAGCCTGAAGCTGGAATCTTCCTA

DNA: CCAATTAGCCACTCAGCAGTGGTCAGACACATCCCCATTAGCAGAATTTAT

DNA: ATTAATACTGCCTCTATTAAGAAGTGCTCATTAGCATTTGCTCTCATTAGG

DNA: CTAAATGAGTTCCCTGGGAGCCATCCCAATAAAATCGCTATCCTAATTAAA

DNA: CGGGCATATTCTTACACACATATTTTTCCCCTTGCCCCTTCCCTTTCCCAT

DNA: GGTGAGGCCACAATGGAACAACACATTTTCATGGGAAGGGAGAGCATAGAC

DNA: CATCATTCCAAAATAGATGGCCAAAACTGGTCAACTTGGAAATGCTTGTGT

DNA: AGATGAGCTTTATGAAGTCCCAGCCCACAATGGCCAAGGGGCGGAGATTCT

DNA: GGGAGTTGAAATATTCATTGAGAGATTGAGCGTCACTGTTTTTCACTATTA

DNA: TTGTGGAGTCTTCCTGTTATCGTTCCCATCATTTATGGCTCCAAACTATGC

DNA: TGACTGCAGCTGATATAAGGCAGTGTTTCCTTTGCACACTCACTCTTTCTT

DNA: CTTTAAGGAAGTCTGTTGACAAGTGTATGAAGCGATGCTGATAGCAGCTTG

DNA: CAATCAACTGTACATACTGAATTACTTTGGCTTTTGTTTTGTTTTAAAGTG
```

```
DNA: AAAACCAAGTGCCAGCAAGAACAGTGTGGTAAGGGTGTCCACCCAGGAAGA550505
+1: K  T  K  C  Q  Q  E  Q  C  G  K  G  V  H  P  G  Rexon 7

DNA: TTCTGTGCCGAGTGTGATGAAATGGGGGACTATTTGCCCAAACAGTGCCAC
+1: F  C  A  E  C  D  E  M  G  D  Y  L  P  K  Q  C  H

DNA: CACAGCACTGGGTATTGCTGGTGTGTCTACCAAAATGGCACAGAGATTGAG
                              ..........................EST FG782415.1
+1: H  S  T  G  Y  C  W  C  V  Y  Q  N  G  T  E  I  E

DNA: GGCACTAAAGTCCGTGGACCTCTGGATTGCGATGGTAAAGAACTTGTTTAT
     ................................                    EST FG782415.1
+1: G  T  K  V  R  G  P  L  D  C  D

DNA: CCTGGCTGATCATCAGGTGGGGAGAAGGCATTGAGGGGATCTTCTTTGCAG

DNA: TGTCAAGTCAGTTGGTTTGTCAGTAGCTAAGGATCAACATCAGAATGAACT

DNA: ACTCCTAGGAGCTAATGATTCTCATTATTTGCATTGAAGTCTAAAATGTTG

DNA: GCAGGGCTTGGAAAGCTACTTTTGGTGGCCATATTGGCTGGGAGATTTTAG

DNA: GAATTATAGCCCAAATAAGGTCCTATTTCTAGTCTTCTGCACAGAATGGTA

DNA: AGGAAAATGTCACATGAAGGACCTATTAAAAAAGCGTGTAAGGAGTAAATC

DNA: AGACTCGGGATGTGAGAGAGCCCCCAAACCAGAGCACTATAGCACAATATG

DNA: CCTTCTTCATTTCTTACCCCAAATCTTCAAGTTTTATAATGTGAGGTTCAC

DNA: ATAGTCAAGCTTATGCAAATAAAGCACTTTGACCTAAACTTGCATAAAATA

DNA: GCCTGGTGACAGTGTCTTGGTCATAGTAAGTGTTTCCTTTCTGGAACATAA

DNA: ATTCACTTCATTTTCTGCACATGCAACTTTAAAGCCTCTTGGCTCCTGTCA

DNA: TGGGGATTCATTGGCCAATACTTTTTGAAGCAATAAAAAGTAGAACAACAT

DNA: TTCTATTACATTAATTTGATGATGGAACTAGTCCTACCATCATTTGGCCCC

DNA: ATATGACCAAAGCAAGAAACCCTCTGTAACCATATTAAGGAACTTCCCACC

DNA: CTTGCATTTTGCTGGGCTGCCATTTCATAAGTTGTTTGTCACGACCCAGGC

DNA: TGCAGAGCACCAATAACCATACACAGAGGCCAGAATCTATCTAATATCTTT

DNA: ATTAAGGAAATATATAAGGTTAATAAAAGCAAGTGTAGGAAATAGTTCAGA

DNA: AGTAGACCTTTCAGGAAAGGTCAAAGTTAGTCCAAAGAAACAATGTCCAAT

DNA: ATGAAATATTAAGGTCCAAAGTTGTAATCCAATAACCGAAACACCCACTTT

DNA: GCCAAGCAAAGTGAGGGGAGATGACAAGGTTCTTAATCCAAGGAACTTGAT

DNA: GTGATGCTAGGAAGTAATCTTGATTCTTGGAACACAAGGCTTGATACAAGG
```

DNA: CAACAAGGAACGAGGAACAAGGACAAGATCCGTGGTAAATACTTGGCAAGG551780

DNA: TCCGGAAAGCAAGGCAAGGTCCGGGAAGCAAGGCAAGGTTGGAAACGTGAA

DNA: CGAAGGCTTGGAAACAGGAACGAGGCTTGGAAACAGGAACGAAGGCTTGGA

DNA: AACGGGAGTAGCGCTGTCCACACACAATCTACTCCAGTTGCTGACGAATTG

DNA: ACTCCGCAAAATTCCTTCGGGGCAAGGAACCTAAATAGGGTCTCGTTTTCC

DNA: CACCAAAGAACACTTCTCTGGGGAATCAGAACCGAAAGTCAATCTCTGTGT

DNA: CCAGATGCATGACTCCCTAGAATTTCTCATGGAAGCAGTCTTAATCAGCTG

DNA: AATGCCTGGCTGCGATTCTCAGGCTTCTGCGATTAGCCTGCTGAACTCCTT

DNA: TTTGTTGTTGATAATAACTACGGCGAGAAAATGGGGGAGATTCTGACTCAG

DNA: GGCTTGTTTGGCAGATTTCTGGAAGGCAAACCTCTTGCAGGTGCAAGGGCT

DNA: CCAATTCAGGCTGGGAAAGTTCCAATTCTGACTGAAATGGTGGAAAACCCA

DNA: AGTTTTCCTCTTCATCTGTCACAACAGCGCTAGGAACGGGACTACATGGCC

DNA: CATGAGTCATCACATTGTTTGACTCATAACAACTTGAACTGCAAGATGACA

DNA: AGAGGGCTTAAAACACTTCATCTGTGTATAATTCAGTCAGACTTTCTTCTC

DNA: AAGACAACTTCATGACAAATAATTAAGAGTTCACTTAATGTGATTCCACAG

DNA: AGCATAGAAATTTCCCAATGCACCAGTATCTTCTATTTATTTATTTATTTA

DNA: TTTACTTACTTACTTACAGTATTTATTTTCCACCCTTCTCACCCCGAAGGG

DNA: GACTCAGAGCGGATCACAATTTACATATACATGGCAAACATTCAATGCCAT

DNA: TAGACATAGGACACACACAGAGACACAGAGGTTATTTAACATTTTCCAGCT

DNA: TCTGACTTCCTGAGGGTATGCTCGATTCCAGCCACAGGGTGAATTGCTGCT

DNA: TCATCATCCACTGTGACGCTGAGTCCTTGATGGATTACTTCCTAATCTTCC

DNA: AGCTCACACTGCTGGACATTTTTATGGTGATGTAAATTAGTTAAATTAGCC

DNA: TCCCAGCATAAGCGGTCCCTAAATTCCCTACTTGACAGATACAACTGTCTT

DNA: TCAGCTTGCTTAGGTAAACAACAAGCTGGGGCTATTTATGGTCAGGCACTC

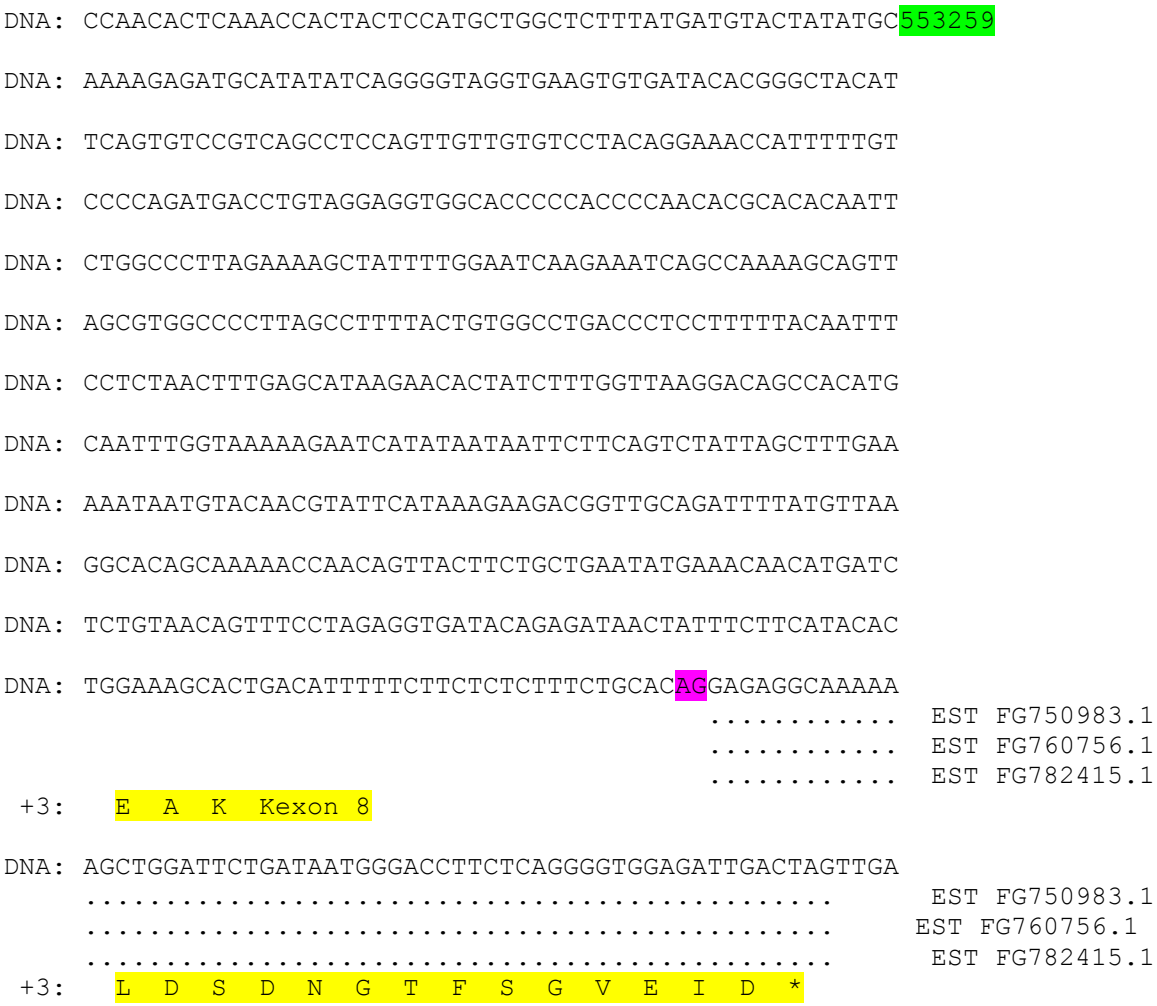
DNA: AATCCGACCCGAGCTTCGAACTCATGACCTCTTGGTCAATAGTGATTTATT

DNA: GCAGCTGGCTACTAACCAGCTGCGCCACAGCCCGGCCCATCATAATAATCA

DNA: TTGTGAGCTTGTGTTCATTCAAATTGTTTCCGATTTATGTTGATCCTAAAA

DNA: CAAATTTATGATTGGGGCTGAGAGGGTACGATTTTCTCAGGTTGACCCAGT

DNA: GAGTTTTCATGTCTGAGTGAAGATTTGAACCCTGGTCTCCAGAGTCATAGT



Supplemental Figure 2.*Anoliscarolinensis* Ii locus.19,635bp of scaffold 29 are shown. Homology to three anole Ii ESTs and the caimen Ii sequence were used to annotate the anole genomic sequence based on the AnoCar1.0 assembly accessed through Ensemble. Green highlighted numbers indicate the position on scaffold 29 of the base at the right end of that line. Predicted intron signals are highlighted in magenta, and dots indicate identity of ESTs to the genomic sequence. Amino acids fully coded by a predicted exon are highlighted in yellow. Nonsynonymous differences in EST and genomic sequences are highlighted in blue.

>AAVX01269494.1 1-1235:rev

DNA: GGAGTGGGGGGTGAAGGGGTATGGGAGGGGGTGGAGGAAACTCCCGTCTCT55

DNA: GTGTCCGCCTGCCTGGTTGCGCCATCCATCATCCGGGGAGGGGTGGGGGGG

DNA: CACGAAGGCACAACAGGTGAGACAAAGCGCGCTGGTTACCGTGATCTCCAC

DNA: CACCCCCGCCAACCCTGCCTGACCCCACCCTGCTCCACACTCAGCCAGTTA

DNA: TCATCTGTCGTTGTTTCACAGGAGAAGACCCCGACCCCAGCACCGCCCCTC
 +1:   T P T P A P P Lexon 4

DNA: ACGCTGATGCAAGAGGTCGAGGAGCTGCTGAAGGTGAGCGGGACAGGGGGC
+1: T L M Q E V E E L L K

DNA: AGGATCTGAGGGGGCCGGGACTGAGGGGGCAGGTGGGGGCTGTGCTGGGGT

DNA: TTGGGGGAAGGGCTGGGTCTGGGTGCTGGGGGGCTGGGGTTGGGGAGCTGA

DNA: GCCAGGGGATGGGGCTGGGGTTTGGGGGCCGTGTTAGGGCCAGGGTTGGGG

DNA: CTGGGGGTTGGTGCCGGAAGGGCTCAGGGCTGTGTTGGGGGGCTGGGTTTG

DNA: GGGTCTGCTGTTGTTGTCGCTTACTCTCTGTGTTTCTGCTCCTCAGAAAGA
 +2: G L LLL S L T L C V S A P Q K Eexon 5

DNA: GAACTTGACCCAGGAATTGCCGAAGTTCAAGGGGACCCTGTTCAGTAACCT
+2: N L T Q E L P K F K G T L F S N L

DNA: GAGGACCCTGCGTGAGAACGTGGAGGAGCCCATGTGGAGGGTGAGAGATGC
+2: R T L R E N V E EP M W R

DNA: GCTGCTGATGACAGTGAGGGCTCTGTCCCAACACCTTCCCCCGACCTCTCT

DNA: CTCTCTCTTCACAGACTTTAATGACCATTGTTTTCTGTGCAGGAGTTTGAG
 +1:   E F Eexon 6

DNA: ACTTGGCTGCGCAACTGGCTTATGTTCCAACTGATCCAGGAACAGCAGCAG
+1: T W L R N W L M F Q L I Q E Q QQ

DNA: AAAATCAGCTCCACCATCGCCCCATGGCACAAACCCCGAGGTAGGACATCT
+1: K I S S T I A P W H K P R

DNA: TACCACAGACACTCTTCCTCAGCCACACTCTCTTGGCCCTCTCACTCTCCC

DNA: CCTCGCTCTCACTCACCCCCCCACCTCCCCCAACATACCCTTACTCCCCTC

DNA: AGCCACACTCCCCTTCGGCCACACACACCCTCGCTCCCACTCCCCCCCTAC

DNA: CCCCACACCCTCACTCCTCTCAGCTACACACCCCCTCGCTCTCACCCCTCT

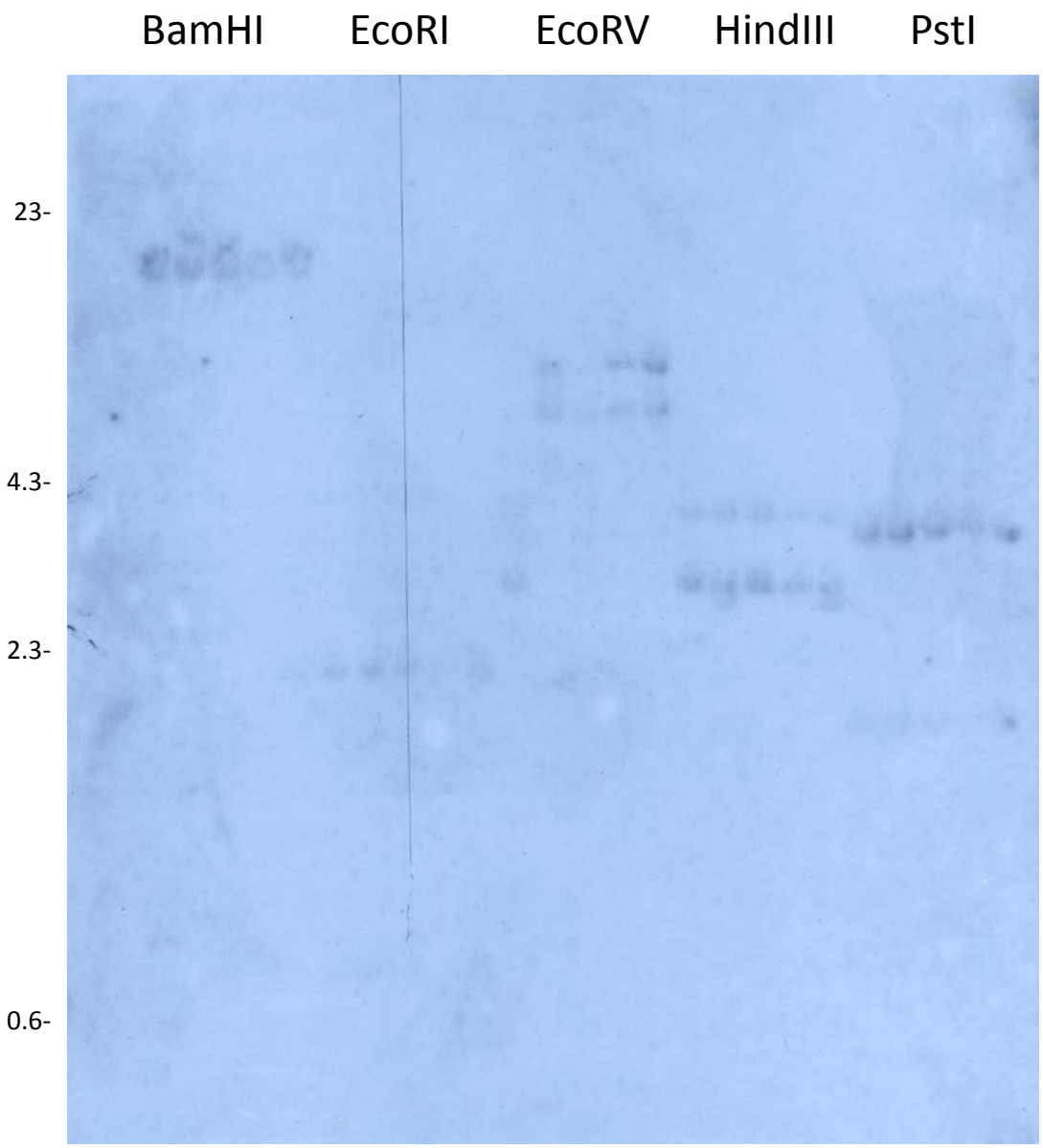
DNA: CTGCCACACTCCTTCCCCTTTCCCCCCACTATGAAGTGTGTGTCCGTGCAG

DNA: TGATTGCCTAACTAACTGAAGATGTGGGGTGTAGGTGGGTTGGGTTGTAGG

DNA: TGGCTTGGGGTGTAGGTGGGTTTTGGTGTGGGTGAGTTGGGGTGCAGGTGG

```
DNA: GTTAGGGGTGG1235


>AAVX01322053.1 1-467:fwd

DNA: TTTCGCCCACAAAATACACAGTCACTACACTTTACAGTATATGCTGTGAAG55

DNA: CGCTTTGAGACGTCTCAAAGACATCATTATTATTATTATTACACGTTGCTC

DNA: TCCGTGTCTCACACTCCCTCGCACGCCTCCTCTCCGCCCAGTGGTTCCCGC
 +2:  P  C  L  T  L  P  R  T  P  P  L  R  P  V  V  P  Aexon 8

DNA: CAAGACAAACTGCCAGCGAGAGCGTGATAGGATCAAGCTGATGCCGGGTGT
+2:  K  T  N  C  Q  R  E  R  D  R  I  K  L  M  P  G  V

DNA: TTATGAGCCGACCTGTGACAAGGCTGGTGACTACACGGCAGAGCAGTGCCA
+2:  Y  E  P  T  C  D  K  A  G  D  Y  T  A  E  Q  C  H

DNA: TGCCAGCTCCGGCTACTGCTGGTGTGTCAAGCCGGACGGCACTGAGATCCC
+2:  A  S  S  G  Y  C  W  C  V  K  P  D  G  T  E  I  P

DNA: TGGCACCCGCGTCCGCGGGCAACGGCTGCACTGCCAACGTAGGTCACAATC
+2:  G  T  R  V  R  G  Q  R  L  H  C  Q  R  R  S  Q  S

DNA: ATCACCACACAGTGAGGGGAGAGGGGGACAGGAGGAGGGGACGGAATAGGG

DNA: CAGGAGGAGGGGAAAAGGGGGGTCAGGAGGAGGGGTGAGGGGGGGCAGGAG

DNA: GAGTGGAC
```

Supplemental Figure 3. Partial predicted annotation of *Callorhinchusmilli* Ii locus. 1235bp of scaffold AAVX01269494 andAAVX01322053are shown. Homology to three elasmobranch as well as other vertebrate Ii sequences were used to annotate the elephant shark genomic sequence based on the scaffolds made available through the Elephant Shark Genome Project and GenBank. Green highlighted numbers indicate the position on scaffolds of the base at the right end of that line. Predicted intron signals are highlighted in magenta. Amino acids fully coded by a predicted exon are highlighted in yellow.



Supplemental Figure 4. Nurse shark Ii is likely encoded by a single copy number gene. Enzyme used to digest each repeated set of five samples displayed at top. Marker (in KB) is shown on left.

```
>97F2 cathepsin L
DNA: TCTTGATGTGAAGGAGAGGCCAGGAGAAAAGCCATGATGATGTTCCTGCTG
 +1:                                             M  F  L  L

DNA: TTTCTGGGGACTTTGCAGATCGTGGCTGTGGCTATCAGTGCCTCACCATTG
 +1: F  L  G  T  L  Q  I  V  A  V  A  I  S  A  S  P  L

DNA: TTGTTTGACCCACTGTTGACTGAAGGCTGGGAGAAATGGAAGATGCTTCAC
 +1: L  F  D  P  L  L  T  E  G  W  E  K  W  K  M  L  H

DNA: CAGAAACAGTATGCAGAGAATGAAGAAGGTGTTCGGAGAATGGTGTGGGAG
 +1: Q  K  Q  Y  A  E  N  E  E  G  V  R  R  M  V  W  E

DNA: AAAAACTTTAGATTTGTGGAAAGTCACAACTTGGAATACTCGCTGGGTAAA
 +1: K  N  F  R  F  V  E  S  H  N  L  E  Y  S  L  G  K

DNA: CACGGCTTTAATCTTAAGATGAACCAGTTTGGAGACATGACCTTGGAGGAG
 +1: H  G  F  N  L  K  M  N  Q  F  G  D  M  T  L  E  E

DNA: TTCAATGAGCAGATGAATGGATTCCGTCTCTTTAAATCCAGGAACTCATCC
 +1: F  N  E  Q  M  N  G  F  R  L  F  K  S  R  N  S  S

DNA: CAGCCACTTGCAAGGATCCCTGAACTGGCTGAGATTCCAAAGAATGTGGAC
 +1: Q  P  L  A  R  I  P  E  L  A  E\/I  P  K  N  V  D

DNA: TGGCGCAATGAGGGATATGTCACTCCAGTGAAGAACCAGGGTAGCTGCGGC
 +1: W  R  N  E  G  Y  V  T  P  V  K  N  Q  G  S  C  G

DNA: TCATGCTGGGCTTTTAGCTCAACGGGAGCGTTGGAGGGACAGACCTTTAAA
 +1: S  C  W  A  F  S  S  T  G  A  L  E  G  Q  T  F  K

DNA: AAGACGGGAAGACTCATCCCATTGAGTGAGCAGAACCTGGTGGATTGTTCA
 +1: K  T  G  R  L  I  P  L  S  E  Q  N  L  V  D  C  S

DNA: CAAGCACAGGGGAATTATGGATGTGGTGGTGGATGGATGGAAAATGCCTTC
 +1: Q  A  Q  G  N  Y  G  C  G  G  G  W  M  E  N  A  F

DNA: TTGTATGTACATGAAAACAATGGGATTGATTCTGAGGCTGGGTACCCGTAC
 +1: L  Y  V  H  E  N  N  G  I  D  S  E  A  G  Y  P  Y

DNA: ACAGGCCAAGATGATCCCTGTAATTATGACGTC
 +1: T  G  Q  D  D  P  C  N  Y  D  V



>48B1 cathepsin L
DNA: TCAGAGAAAGGAGAGGTTATCCACGATGAAGACCTCCTTCTTTTCCCTCTG
 +2:                           M  K  T  S  F  F  S  L  C

DNA: CTTGGTGGTGCCCTTTCTGGCAGCTGCCTTTGCGCATACATTTGAGCCTTC
 +2:  L  V  V  P  F  L  A  A  A  F  A  H  T  F  E  P  S

DNA: GTTGGATGAAGCTTGGTTGAACTGGAAGTCATTTCACAATAAAGAGTACAC
 +2:  L  D  E  A  W  L  N  W  K  S  F  H  N  K  E  Y  T

DNA: CGGGGATGAAGATAATTATAGGAGGATGGTATGGGAGAAAAACTTAAAACA
 +2:  G  D  E  D  N  Y  R  R  M  V  W  E  K  N  L  K  Q
```

```
DNA: AATCCAGCTTCATAATCTTGAGCACTCGATGGGGAAGCACACATACCAGTT
 +2:  I  Q  L  H  N  L  E  H  S  M  G  K  H  T  Y  Q  L

DNA: GGGAATGAACCATTTTGGAGACCTTACAGCTGATGAATTCCAACAATTCAT
 +2:  G  M  N  H  F  G  D  L  T  A  D  E  F  Q  Q  F  M

DNA: GAATCAAATCCACCGGCTTGAAATGATGAACTCAACCAGGAAAAAGCCTGC
 +2:  N  Q  I  H  R  L  E  M  M  N  S  T  R  K  K  P  A

DNA: TTCAGCTGGGCCCAAACCATCCCAACTTCCTCAGAGCATTGACTGGAGGGA
 +2:  S  A  G  P  K  P  S  Q  L\/P  Q  S  I  D  W  R  D

DNA: CAAAGGTTATGTTACTCCAGTGAAAAACCAGCGACACTGTGGCTCATGCTG
 +2:  K  G  Y  V  T  P  V  K  N  Q  R  H  C  G  S  C  W

DNA: GGCTTTCAGTGCAGTTGGTGCCCTTGAGGGACAGACGTTCAAGAAGACAAG
 +2:  A  F  S  A  V  G  A  L  E  G  Q  T  F  K  K  T  R

DNA: AAAACTTATCTCTTTGAGTGAACAGAACTTGGTCGACTGCTCTAATGCAGA
 +2:  K  L  I  S  L  S  E  Q  N  L  V  D  C  S  N  A  E

DNA: GGGTAACCATGGATGTAATGGTGGGCTGATGGAATATGCCTTCAACTATGT
 +2:  G  N  H  G  C  N  G  G  L  M  E  Y  A  F  N  Y  V

DNA: GCAGTCTAATGGTGGGATTGACACAGAGGAATATTACCCATACACTGCAGA
 +2:  Q  S  N  G  G  I  D  T  E  E  Y  Y  P  Y  T  A  E

DNA: GGAAGGCACCTGTAAGTATAATCCCAATTTCAGTCCAACCACTTGCCATGG
 +2:  E  G  T  C  K  Y  N  P  N  F  S  P  T  T  C  H  G

DNA: CTCCAGGTTCGTAGCC
 +2:  S  R  F  V  A


>3A2 cathepsin Z
DNA: CTAGCTTTGAGCTGGGTGGCAGGTTTTTCTAGGATCTGCCTAAGTTGACGG

DNA: CTCTCACCATGTTTCACTGGCTGTTCACTCTATCTCTGCTGAGCTGCGTTT
 +3:        M  F  H  W  L  F  T  L  S  L  L  S  C  V  S

DNA: CCTCCGCCTTGCATTTTCGAGGCAGTCAGCCTTGCTACAAGATCTTGGCCG
 +3:  S  A  L  H  F  R  G  S  Q  P  C  Y  K  I  L  A  G

DNA: GCAAAAAGTTTCAAGGAGTCAGGTCTTACCCCCGGCCCCATGAATATCTTC
 +3:  K  K  F  Q  G  V  R  S  Y  P  R  P  H  E  Y  L  P

DNA: CCATAGCTGATTTGCCAAAAGTATGGGACTGGCGCAATGTGAATGGCACTA
 +3:  I  A  D\/L  P  K  V  W  D  W  R  N  V  N  G  T  N

DNA: ACTTTGCAAGCACAACCAGGAACCAGCACATCCCCCAGTACTGTGGATCAT
 +3:  F  A  S  T  T  R  N  Q  H  I  P  Q  Y  C  G  S  C

DNA: GCTGGGCCCATGGGAGCACCAGTGCTCTGGCAGATCGGATCAATATTAAGC
 +3:  W  A  H  G  S  T  S  A  L  A  D  R  I  N  I  K  R

DNA: GCAAAGGTGCTTGGCCATCTGCTTACCTCTCTGTGCAGCAAGTCATTGACT
```

```
 +3:   K  G  A  W  P  S  A  Y  L  S  V  Q  Q  V  I  D  C

DNA: GTGCCGATTCAGGCTCCTGCGAAGGGGGGGATCACATGGGAGTCTGGGAAT
 +3:   A  D  S  G  S  C  E  G  G  D  H  M  G  V  W  E  Y

DNA: ATGCACATAGACACGGCATTCCAGATGAAACCTGCAATAATTACCAGGCCA
 +3:   A  H  R  H  G  I  P  D  E  T  C  N  N  Y  Q  A  K

DNA: AGGATCAAGATTGTAAGCCTTTTAACCAATGCGGCACCTGTACAACCTTTA
 +3:   D  Q  D  C  K  P  F  N  Q  C  G  T  C  T  T  F  N

DNA: ACGTCTGCCACGTGGTCAAGAATTACACACTCTGGAAGGTTGGTGACTTTG
 +3:   V  C  H  V  V  K  N  Y  T  L  W  K  V  G  D  F  G

DNA: GCAGAGTCAACGGGCGAGAGAATATGATGGCAGAGATTTATGCCAACGGTC
 +3:   R  V  N  G  R  E  N  M  M  A  E  I  Y  A  N  G  P

DNA: CAATCAGTTGTGGCATTATGGCTACAAGAAAACTGGACGCCTACACTGGAG
 +3:   I  S  C  G  I  M  A  T  R  K  L  D  A  Y  T  G  G

DNA: GAGTGTATACGGAGTACCAAGCAGAGCCAGGGATAAACCATATCGTGTCAG
 +3:   V  Y  T  E  Y  Q  A  E  P  G  I  N  H  I  V  S  V

DNA: TGGCAGGCTGGGGTGTAGAGAATGGAACTGAATACTGGATTGTGCGCAACT
 +3:   A  G  W  G  V  E  N  G  T  E  Y  W  I  V  R  N  S

DNA: CTTGGGGCGATCCATGGGGGGAAAGAGGTTGGCTCCGAATCGTCACCAGTG
 +3:   W  G  D  P  W  G  E  R  G  W  L  R  I  V  T  S  A

DNA: CCTACATGGGAGGGAAAG
 +3:   Y  M  G  G  K
```

Supplemental Figure 5. Cartilaginous fish cathepsin L homologs. Putative leader peptide is highlighted in yellow, as are the conserved functionally important residues of the ERFNIN motif and catalytic site cysteines. Pro-protein cleavage site is marked by slashes (\/) between amino acids.

```
eShark    YVDITLVEGQVETKSGRKSNKGLDKACKIYNIDEGEHAA--ISTDNLFDFSEREQRETKVIALLGRAGIGKSVLVQRV
finch     FIDRTLVQSQTETKTGKNSAKAMEKELVTCSLQEKEKTA--IDRSQIFQIPQRKDLETKVIVVLGKAGMGKSILIQKI
opossum   YIDIELTRTQVE-----KNSKYQEKDLAIQDWTERQKAR--VGWREVFARPSGQKGETQVIAVLGKAGLGKSAWTREI
human     LVEVDLVQARLE----RSSSKGLERELATPDWAERQLAQGGLAEVLLAAKEHRRPRETRVIAVLGKAGQGKSYWAGAV
danio     YVDAHLVQRKLLIKSGKNANKCLEKELVVLSDSERKKAK--LDTSQLFHNLDSKSKQS--FAFLGKSGVGKTTFIQRL
          <    acid transactivation domain             >          <     NACHT NBD

eShark    CLDWVNGSFGQFEFVFWFKCRTLNLE-KQ-YKLRDLLFE-PFLPALRDTGEVFQYLCHHPDKVLVIVDDFEDFQDCDG
finch     CQDWSNGEFSQFEFVFWFDCKQLSLPEKQ-YSLKELLLE-FFVKPQEGSKEIFEYMLQNPGKVLLIFDGFKGLHDHEN
opossum   CRNWAQGQLPQYEFVFHYKCHGLNLPGND-YCLKDLFFR-LCHHSLEESEEVFKYILKHPNHILVILDSFEELEGQDG
human     SRAWACGRLPQYDFVFSVPCHCLNRP-GDAYGLQDLLFS-LGPQPLVAADEVFSHILKRPDRVLLILDGFEELEAQDG
danio     CLDWSNSGLPQFQFVFLMNCKILDFK-QSNYSLKTLLFDFSTSPHCEDSNAAFKHILSCPDEVLIIFDSFDHIKDLEG
                               NACHT nucleotide binding domain (NBD)

eShark    LLQPTTCPSREKPQSVRQLLAGLLQKKLLRGCTVLVTARPRGTFNQYLTRVDKVVEVMGFSPQHREEFVRK
finch     FPRCSASQPEKDLCSIKELLSGLIQKKVLNGCTLLFTARPKDKLYQYMSKVDKTIEIVGFSPQQRELYITK
opossum   LLHYLASSSPREPQPIKGLLAGLFQRKLLRGCTLLITTRPKGRFIQYLAKVDSLLEVQGFSPEQVEAYFEK
human     FLHSTCGPAPAEPCSLRGLLAGLFQKKLLRGCTLLLTARPRGRLVQSLSKADALFELSGFSMEQAQAYVMR
danio     LLQSPAKSHTDTKYTIKQLFSGLFQKEILSGCTLLIATRPKDVLNQVLRKMDCLLEIWGFSPEDIELYTSK
                        NACHT nucleotide binding domain (NBD)                    >
```

Supplemental Table I. Primers

Primer Name	For/Rev	Ii Domain	Sequence	Priming Site	Anneal Temp.
Invariant Chain-1R	R	CLIP	5′-AATGTAGGCCATGGGCATG-3′	DMPMAYI	57
Invariant Chain-1F	F	tri	5′-CCTACTCCTGCCAAACCCCTGACTGTG-3′	PTPAKPLTV	73
Invariant Chain-2R	R	Tg	5′-TCACATTCGGTGTCCATGCTGCGCAC-3′	VRSMDTEC	73
Invariant Chain-2F	F	tri	5′-CTCGACCTGCTGCGAGACAGCATGAC-3′	LDLLRDSM	73
Invariant Chain-3R	R	Tg	5′-CAGCTTAAATGCCCACGCGTGGATGTTCC-3′	GTSTRGHLS	73
NSIiF1	F	cyto	5′-GAATAATTCACAGCAAGATAT-3′	NNSQQDI	48
NSIiR1	R	CLIP	5′-AATCATTGGCTTAGGTGGGAATC-3′	RFPPKPMI	61
NSIiTMF	F	TM	5′-CTTTGGGGTGGTGTGACCGTGTTGG-3′	LWGGVTVLA	73
NSIiF2	F	tri	5′-CTTGGCTCCGAAACTGGCTC-3′	WLRNWL	62
NSIiR2	R	tri	5′-GAGCCAGTTTCGGAGCCAAG-3′	WLRNWL	62
NSIiR3	R	C-term	5′-CTTAATTCGCTTTCTTTGATACTG-3′	VSKKAN*	56
Ii104D3-R1	R	cyto	5′-CAAAGCGTTCTGCTGCTC-3′	EQQNAL	63
IiD2-F1	F	Tg	5′-ATCTCAGGAACATCCACGC-3′	ISGTST	63
IiD2-R2	R	Tg	5′- GAGATCTTTGTTCCATTCG-3′	PNGTKIS	57
NSIiR5	R	Tg	5′-ACCAGCAGAATCCTGTGCTC-3′	STGFCW	64
NSIiF4	F	tri link	5′-TCTCCAGTGTTGCGATGAAG-3′	FSSVAMK	64
NSIiR4	R	tri link	5′-GCTTTTGCAGGTTCTGGAAG-3′	LPEPAKA	64
NSIiF3	F	tri	5′-GCACTTTCTCTGCCAACCTC-3′	STFSANL	64
NSIi-R5	R	3′ UTR	5′-TTGACAAGCTTAGGAGATACTG-3′		58
NSIi-F5	F	cyto	5′-CAAGATATTGCTAGCCAGAGC-3′	QDIASQS	61
NSIi-F6	F	3′ UTR	5′-TTCCAGTAGGCATTCCATAG-3′		59
NSIi-R6	R	3′ UTR	5′-TTCTGCCCACTTACAAACC-3′		60
NSIi-F7	F	3′ UTR	5′-TTGGTAACACACACAACCTG-3′		59
NSIi-F8	F	TM link	5′-CAGCAGAGCACGATAACCCACC-3′	QQSTITH	70
NSCath48B1F1	F	cath	5′-CTTGGTGGTGCCCTTTCTGGC-3′	LVVPFLA	67
NSCath48B1R1	R	cath	5′-GCATTAGAGCAGTCGACCAAG-3′	LVDCSNA	58
NS NDPKF	F	NDPK	5'-GGTAACAAGGAACGAACCTTC-3'	GNKERTF	64
NS NDPKR	R	NDPK	5'-CTCATAGATCCAGTCTTGGGC-3'	AQDWIYE	64

Supplemental Table 2. Database accession numbers of sequences accessed or annotated for use in these analyses.

Species common name	Ii isoform or gene name	Species	Nucleotide accession or locus	Protein accession	Type of sequence	Tissue of origin
nurse shark	104D3/short	*Ginglymostomacirratum*	JF507710		EST	spleen/pancreas
nurse shark	D2/long	*Ginglymostomacirratum*	JF507711		cDNA	spleen
spiny dogfish		*Squalus acanthus*	EE049515		EST	rectal gland
Pacific electric ray	long	*Torpedo californica*	EW694352		EST	electric organ
Pacific electric ray	short	*Torpedo californica*	EW694337*		EST	electric organ
elephant shark	tri exons	*Callorhinchusmilii*	AAVX01269494		gDNA	testis
elephant shark	Tg exon	*Callorhinchusmilii*	AAVX01322053		gDNA	testis
northern pike		*Esoxlucious*	GH262596		EST	kidney/spleen/ heart/gill
rainbow trout	S25-7	*Oncorhynchusmykiss*	AY065836.1	AAL58576	cDNA	head kidney
rainbow trout	14-1	*Oncorhynchusmykiss*	AY081776.1	NP_001117913	cDNA	head kidney
rainbow trout	INVX	*Oncorhynchusmykiss*	AY065837	AAL58577	cDNA	head kidney
sea bass		*Dicentrarchuslabrax*	DQ821105	ABH09445	cDNA	head kidney
Atlantic salmon		*Salmosalar*	BT060361	ACN12721	cDNA	brain/kidney/ spleen
Atlantic salmon	INVX	*Salmosalar*	BT057821	ACM09693	cDNA	brain/kidney/ spleen
rainbow smelt		*Osmerusmordax*	BT075598	ACO10022	cDNA	brain/kidney/ spleen
Chinese perch		*Sinipercachuatsi*	AY395716	AAS77256.1	cDNA	
three-spined stickleback		*Gasterosteusaculeatus*	BT026899		EST	gills
Hong Kong grouper	ICLP	*Epinephelusakaara*	FJ358642	ACJ09264	cDNA	
medaka, ricefish		*Oryziaslatipes*		AU167149.1	EST	
spotted green pufferfish		*Tetraodonnigroviridis*	CR701322		cDNA	eyes
spotted green pufferfish	ICLP?	*Tetraodonnigroviridis*	CR691515		cDNA	eyes
fugu		*Takifugurubripes*	CA846646		EST	fin
fugu	INVX?	*Takifugurubripes*	BU806406		EST	gills
channel catfish		*Ictaluruspunctatus*	FD323417		EST	
mangrove red snapper		*Lutjanusargentimaculatus*	FJ772422	ACO82381	cDNA	
Atlantic halibut		*Hippoglossushippoglossus*	EU412489		EST	
spotted green pufferfish		*Tetraodonnigroviridis*	CR701322		EST	
common carp	ICLP1	*Cyprinuscarpio*	AB098609	BAC53767	cDNA	head kidney
common carp	ICLP2	*Cyprinuscarpio*	AB098610	BAC53768	cDNA	head kidney
zebrafish	ICLP1	*Daniorerio*	NM_131590	NP_571665	cDNA	spleen
zebrafish	ICLP2	*Daniorerio*	AF116539	AAD47423	cDNA	spleen
zebrafish	x	*Daniorerio*	NM_131372	NP_571447	cDNA	embryo
zebrafish	ICLP1	*Daniorerio*	Zv8 ch14		gDNA	
zebrafish	ICLP2	*Daniorerio*	Zv8 ch12		gDNA	
eastern tiger salamander		*Ambystomatiginum*	CN060626		EST	
bullfrog		*Ranacatesbeiana*	GO472684			
African clawed frog		*Xenopus laevis*	BC059976	AAH59976	cDNA	
western clawed frog	long and short	*Xenopus tropicalis*	XenTr4.2 scaffold 559		gDNA	
green anole	long and short	*Anoliscarolinensis*	Anocar1.0 scaffold 29		gDNA	
spectacled caiman		*Caiman crocodilus*	DQ235262	ABB22797	cDNA	
duck		*Anasplatyrhynchos*	AY905540	AAX89536	cDNA	
chicken		*Gallus gallus*	AY597053	AAT36345	cDNA	AAT36345

cDNA

chicken	long and short	*Gallus gallus*	Galgal2.1 ch13		gDNA
Norway rat		*Rattusnorvegicus*	X13044	CAA31450	cDNA
cow		*Bostaurus*	D83962	BAA12156	cDNA
mouse	long	*Musmusculus*		NP_001036070.1	cDNA
mouse	short	*Musmusculus*		NP_034675.1	cDNA
human	a/long/p43	*Homo sapiens*		NP_001020330.1	cDNA
human	b/short/p35	*Homo sapiens*		NP_004346.1	cDNA
human	long and short	*Homo sapiens*	ch5q32		gDNA

*Frameshift mutations in Pacific electric ray EST EW694337 were manually edited based on locally better sequence in EW694344 and EW694388)

Supplemental Table 3. Search results for Ii orthologs in jawless chordates.

query	against	high hit	position	score	E value	%ID	length	high hit matches
nurse shark Tg domain (274-342)	lamprey latest GP	contig 6853	8276-8401	143	2.3e -06	50%	42aa	preceding putative exon best matches testican 3 of birds and mammals
nurse shark Tg domain (274-342)	lamprey est's	nothing >50						
nurse shark Tg domain (274-342)	hagfish est's	nothing >50						
nurse shark Tg domain (274-342)	*Ciona* genome	scaffold 157	7622-7822	138	1.20E-06	35%	67aa	preceding putative exon best matches first vertebrate hits are CD22
electric ray Tg domain (274-343)	lamprey latest GP	contig 6853	8264-8425	144	1.20E-06	45%	56aa	preceding putative exon best matches testican of birds and mammals
nurse sharkTg domain (274-343)	amphioxus genome	no sig sim						
nurse shark tri domain (149-215)	lamprey latest GP	contig 319	2156-2369	80	5.7	29%	55aa	putative exon best match gap-pol polyprotein - like teleost
nurse shark tri domain (149-215)	lamprey est's	no sig sim						
nurse shark tri domain (149-215)	hagfish est's	no sig sim						
nurse shark tri domain (149-215)	*Ciona* genome	scaffold 1821	2760 to 2918	80	0.94	29%	56aa	exon best match hypothetical homologous to seminal vesicle secretion protein
electric ray tri domain (149-215)	lamprey latest GP	no sig to exons						
nurse shark tri domain (149-215)	amphioxus genome	no sig sim						
nurse shark TM and CLIP (56-137)	lamprey latest GP	no sig to exons						
nurse shark TM and CLIP (56-137)	lamprey est's	no sig sim						
nurse shark TM and CLIP (56-137)	hagfish est's	no sig sim						
nurse shark TM and CLIP (56-137)	*Ciona* genome	no sig to exons						
nurse shark TM and CLIP (56-137)	amphioxus genome	no sig sim						
nurse shark short c term (216-265)	lamprey latest GP	no hits						
nurse shark short c term (216-265)	lamprey est's	no hits						
nurse shark short c term (216-265)	hagfish est's	no hits						
nurse shark short c term (216-265)	*Ciona* genome	no hits						
nurse shark short c term (216-265)	amphioxus genome	no hits						