Genomic Analysis of the Immune Gene Repertoire of Amphioxus Reveals Extraordinary Innate Complexity and Diversity

Supplementary A

Content

1 TLR system

Birth-and-death evolution of the vertebrate TLR11 lineage

Although the vertebrate TLR11 family can be tracked back to bonyfishes, not an orthologous lineage can extend from bonyfishes to mammals, which suggests high turnover rate of the TLR11 lineage. To the extreme, mouse has 3 divergent TLR11 members (TLR11, 12 and 13), whereas the only human TLR11 member has become a pseudogene (Roach et al. 2005).

Stable paraphyletic relationship between 33 amphioxus V-TLRs and the vertebrate TLR11 family

The low bootstrap value in Figure 1 for the cluster containing 33 amphioxus V-TLRs and the vertebrate TLR11 family is largely due to two highly divergent insect V-TLRs. Insect V-TLRs are incorporated into the tree in Figure 1 in order to provide more information for the evolution of V-TLRs. The clustering of 33 amphioxus V-TLRs and the vertebrate TLR11 family is actually quite stable, because it is supported by both sequence similarity and phylogenetic analysis. Firstly, most amphioxus V-TLRs share 40-50% aa identity with members of the vertebrate TLR11 family, much higher than with members of other vertebrate TLR families; among them, Bf68489, Bf68417, Bf142546, even share over 50% aa identity with their vertebrate TLR11 counterparts. Secondly, both phylogenetic analyses by using minimum-evolution method in Figure 1 and Figure S3 support the relation and excluding insect V-TLRs from the analysis (Figure 3) greatly improves the bootstrap value. Finally, phylogenetic analyses by using other methods, like maximum-likelihood and neighbor-joining, recover the same topology.

The selection force on the amphioxus TLRs of the SC75 lineage

Of 19 members of the SC75 lineage, 12 are not affected by unsequenced regions (gaps), frame shift or stop codon mutations. Hence, 12 full length sequences are used to generate an alignment (Figure S4) and put to selection tests. As stated in the paper, pairwise amino acid identities of the TIR domains of SC75 TLRs are more than 85%, whereas the LRR regions are full of munations, small indels and large portions of deletions and insertions (Figure S4). So intuitively these LRR regions may be controlled by diversifying selection. We then used the MEGA3 software for simple selection tests, the overall mean d_N/d_S values of 12 sequences are given as followed:

LRR region:

complete deletion, Nei-Gojobori method (p-distance) d_N/d_S =0.317/0.461=0.688

complete deletion, Nei-Gojobori method (Jukes-Cantor) d_N/d_S =0.419/0.816=0.513

TIR domain:

complete deletion, Nei-Gojobori(p-distance) d_N/d_S =0.044/0.160=0.275

complete deletion, Nei-Gojobori method (Jukes-Cantor) d_N/d_S =0.046/0.195=0.236

LRRNT+LRRCT+TM+TIR:

complete deletion, Nei-Gojobori (p-distance) d_N/d_S =0.098/0.258=0.380

complete deletion, Nei-Gojobori method (Jukes-Cantor) d_N/d_S =0.106/0.334=0.317

Apparently the d_N, d_S and d_N/d_S are all elevated in the LRR regions. Since the Nei-Gojobori method (p-distance) for selection test is rather conservative, d_N/d_S = 0.688 may suggest a fraction of tested sites under positive selection. Prompted by this, we performed an advanced selection test with the PAML v3.15 package (Yang et al. 2005). The results are as followed:

Codon frequency is estimated using F1X4 option.

All indels are deleted.

One-ratio model M0, Site-specific model M1a (nearly neutral) and M2a (positive selection) are used to fit the data:

M0 model: lnL=-19792.89, np=1.

M1a model: lnL=-19031.81, np=2.

M2a model: lnL=-18820.50, np=4.

The likelihood ratio tests (LRT) indicate that M1a model fits the data significantly better than M0 (Probability<0.001), and M2a model fits the data significantly better than M1a models (Probability<0.001), hence suggesting a fraction of tested sites under positive selection. The Bayes empirical Bayes (BEB) approach (Yang et al. 2005) is then used to identify all possible sites under positive selection. It shows that all sites under positive selection are located in the LRR regions (unpublished data). Therefore, our conclusion is that the LRR regions of SC75 lineage should be dominated by diversifying selection.

2 NLR system

The overall architecture of amphioxus NLR proteins

As showed in Figure S6, most of the non-typical NLR structures apparently derive from the typical NLR structure (DEATH/CARD-NACHT-LRR). As for DLRs (DFD-LRR), because their DFD and LRR amino acid sequences are similar to other NLRs, DLRs are considered as NLRs with a missing NACHT domain.

There are 14 gene models (DFD-NACHT) containing no LRR regions. As a routine (see Materials and Methods), when we found a NLR model without LRR regions, we analyzed 20kb C-terminal sequences beyond the gene model, with this procedure we have recovered missing LRR regions for many NLR models (see Supplementary B, the model structure is marked by "unpredictedLRR"). However, we failed to detect LRRs for these 14 models. It has been reported that the LRR regions of NLRs of thevertebrate and the sea urchin are encoded in complex exon structures, but it should not prevent us from finding the LRR fragments, unless the intron between LRR and NACHT domains spans more than 20kb (this scenario is very unlikely). Considering the architecture occurs 14 times (including alleles), it is unlikely to be a computational artifact. On the other hand, we have cloned an NLR cDNA with complete 3'-UTR and the DEATH-NACHT structure from *B. japonicum* (Accession: EU183367).

There are 30 NLR models (NACHT-LRR) containing no clear DFD N-terminal domains (we have analyzed 20kb N/C-terminal sequences flanking the models), which is either complete lost or is substituted by other domains. There are several ESTs (with 5'UTR) supporting this type of models, including Bf69066 (supported by BW703366), Bf78182 (supported by BW864684), Bf120153 (supported by BW772443, BW785759), Bf121225 (supported by BW953997, BW895355, BW745463, BW913163), Bf89727 (supported by BW802409).

There are 21 NLR models containing neither detectable LRRs nor clear DFD domains. Despite not EST evidence for them, considering that both DFD and LRR can be absent, the presence of NACHT-only genes is reasonable. Nevertheless, we have analyzed 20kb N/C-terminal sequences flanking the models before we reach this conclusion.

There are 22 DLR models (DFD-LRR). According to our analysis, no detectable NACHT domain resides in the sequence between the DFD and the LRRs. There are two ESTs (Accession: BW794253, BW781035) supporting the existence of a CARD-LRR (Bf132252).

The N-terninal domain structure of amphioxus NLR proteins

Amphioxus NLRs with DEATH or CARD domains are present in abundance and usually have simple and compact exon structures. They also have EST evidence: EU183367, EU183368 for DEATH-NACHT combination and EU183366 for CARD-NACHT combination (regardless of the presence or absence of LRR).

The DED-NACHT-(LRR) structure occurs 5 times (models), but has no EST support so far. However, DEDs are adjacent to NACHTs, in other words, only short interval sequences (<1-1.5kb, intron included) between them and there is no interruption by other domains.

When we extended the analysis to the 20kb region before the N-terminal of the NLR models, it yielded be some new domain combination, like CARD-CARD, DED-DEATH and DFD-nonDFD. However, because of the complex genomic structure and the obvious lack of EST evidence, these novel domain combinations are questionable. Among them, the model Bf97362 may be an exception, for its DEATH-DEATH structure is adjacent to the NACHT domain.

As for the DLR models, there are various domain combinations can be found (Supplementary B), but due to the complex exon structures, these combinations require experimental evidence. However, the DFD domain right next to the NACHT domain appears valid because of close adjacency. So far, only model Bf132252 (CARD-LRR) has EST supports (Accession: BW794253, BW781035).

3 LRRIG genes

A typical LRR and IGcam containing protein (LRRIG) consists of N-terminal LRRs, one or more central IGcam domains, a transmembrane region and a C-terminal cytoplasmic tail. There are approximately 30 vertebrate LRRIG proteins, including AMIGO, NGL-1, LINGO-1, NLRRs and LRRIGs. These genes are usually expressed on neural cells, mediating cell adhesion, signal transduction and therefore associate with the development, maintenance and regeneration of the nervous system (Chen et al. 2006b). In *D. melanogaster* there are several LRRIG proteins, of which Kek1 can inhibit EGFR activity during eye development (Layalle et al. 2004). The sea urchin genome encodes approximately 20 LRRIG gene models according to our analysis. The amphioxus draft genome contains 240 LRRIG gene models (approximate 190 genes, Table 2), most of which contain single IGcam. There are at least 113 of them containing predicted transmembrane regions. There are 194 LRRIG models encoding LRR and IGcam in the same exon, which is believed to favor rapid duplication and diversification (Figure S7). The number of LRR motifs of vertebrate LRRIGs varies from 5 to 15 in different families. Analysis of 131 well-predicted amphioxus LRRIG

models indicates that more than half of them have 8-11 LRR motifs, and this number ranges from 4 to 24 if all 131 LRRIGs are taken in account.

The immunological relevance of LRRIGs is not determined, but both LRR motifs and the IGcam domain are competent immune recognition modules. As for the IGcam, insect hemolins can mediate anti-bacterial response by recognizing lipopolysaccharide through their IGcam domains (Schmidt et al. 1993). Insect and vertebrate DSCAM proteins carry multiple IGcam domains and function in neuron development. Recent studies further showed that insect *DSCAM* can produce 38016 alternative-spliced mRNA isoforms and the derived proteins can act as diversified receptors in both immunity and neuron development (Watson et al. 2005; Chen et al. 2006a). Reminiscent of the saying “the brain and the immune system speak a common biochemical language” (Boulanger and Shatz 2004; Du Pasquier 2005), it is of interest to quest whether the expanded amphioxus LRRIG repertoire has a role in host defense.

4 Other LRR-containing models

Leucine-rich repeat (LRR) modules of 20–29 amino acids are present in more than 8,000 proteins from viruses, bacteria, archaea and eukaryotes. LRR-containing proteins participate in nearly all known biological functions (Pancer and Cooper 2006), and a large part of them are involved in host defense of both animals and plants. In plants there are about 1% of the genes of whole genome encode disease resistance factors that contain LRRs (Nurnberger et al. 2004). Besides, LRR modules are the building blocks of the rearranged antigen receptors of lamprey and hagfish (Pancer and Cooper 2006). The above-mentioned TLR and NLR are major defense molecules in echinoderms, protochordates and jawless vertebrates. Amphioxus LRRIG proteins also have undergone large expansion but their role in immunity is not clear. In this section, we focused on the other LRR-containing proteins encoded in the amphioxus genome, many of which may have a role in immunity.

LRR-TM-DEATH proteins

We found 3 LRR-TM-DEATH models with signal peptides, transmembrane regions and cytoplasmic DEATH domains, hence consisting of.a novel class of membrane receptors not reported previously. These gene models encode LRR and DEATH in a single exon or in adjacent exons (intron <1kb), hence this domain structure is unlikely caused by faulty prediction. These genes may function as receptors which probably activate downstream signal pathway through interaction with cytoplasmic DEATH adaptors.

Models containing both LRR and other domains

In addition to TLRs, NLRs, LRRIGs, LRR-TM-DEATHs, there are 185 gene models containing both LRR and other domains in the genome. The most abundant domain structures include 24 Fbox-LRR models (which has 20 homologs in humans), 37 human MFHAS1-like gene models (LRR-GTPase or LRR-GTPase-DEATH), 13 LRR-DEATH-Kinase or LRR-DEATH models (which have no homolog in vertebrates and no EST evidence but are encoded in clear exon structures). Notably, the human *MFHAS1* is a candidate oncogene found in a B-cell lymphoma cell line (Tagawa et al. 2004). In the rest 111 models, various domains can be found, some of which have homologs in vertebrates, and most of which lack EST evidence.

Models containing only LRRs

We have identified a total of 1589 LRR-containing models in the amphioxus genome. In addition to TLRs，NLRs, LRRIGs and models clearly containing LRR and other domains, there are 1178 models left, which contain only LRR and hence termed LRR-only models (Table 3). BLASTP analysis indicates that 230 LRR-only models are apparent fragments of other LRR-containing genes，hence there are still 948 LRR-only models left. Since LRR genes are often incorrectly captured by gene prediction programs, an LRR gene may be broken down into 2 or more models. To address this issue, we examined the genomic distance between LRR-only models and all LRR-containing models (including LRR-only). We found that numbers of adjacent models (LRR-only vs. LRR-only or other LRR-containing) with distance smaller than 3kb, 5kb and 10 kb are 59, 89 and 123, respectively, which means that most of 948 LRR-only models should represent distinct genes because few amphioxus introns can span over 10 kb.

However, due to the nature of the current draft genome, not all LRR-only models really represent genes containing only LRR. We have found that some LRR-only models likely contain other domain that either fails to be captured or is too diverged to be detected by our methods (data not shown). Nevertheless, 948 distinct LRR-containing models comprises a huge LRR arsenal.

Our genomic survey indicates that 78 scaffolds account for more than half of 948 LRR-only models. There are 11 scaffolds containing 10-17 LRR-only models. Many LRR-only models are encoded in single exon (some might be artifacts because LRR domain in separate exons are difficult to be correctly captured by gene prediction programs). There are 266 LRR-only models containing transmembrane regions, but this number is greatly underestimated because a large fraction of LRR-only models are not correctly predicted and we did not perform manual corrections because of the

large number and the inability of current gene prediction programs.

We also calculated the number of LRR repeats of 161 well-predicted LRR-only models carrying clear LRRNT and LRRCT capping motifs. The average LRR numbers for each LRR-only gene is approximately 10, and is ranged from 3 to 30. More than half the models (83 models) contain 5-12 LRR motifs. Since LRR domains are highly variable in primary sequence and length, conventional methods of phylogenetic analysis (i.e. molecular tree reconstruction) can not be applied. So we used the BLASTCLUST program, which can cluster sequences into different groups according to their sequence similarity. With 50% sequence coverage and 65% aa identity as thresholds, BLASTCLUST identifies 28 groups that contain at least 5 LRR-only models, of which two largest groups include 56 and 37 members (the third largest had only 12 members). If the identity threshold is relaxed to 60%, member numbers of two groups are expanded to 112 and 49, respectively. These facts suggests that the amphioxus LRR-only repertoire may also have undergone the same evolutionary history as amphioxus TLRs and NLRs, namely, lineage-specific duplications and diversification.

5 Domain combinations in amphioxus C-type lectins

Twelve non-CTLD domains present in both vertebrate and amphioxus CTL proteins are COL, CUB, EGF, CCP, LDLa, VWF, PKD (polycystic kiney disease), WSC (yeast cell wall integrity and stress response component protein), Ig-like, REJ (Receptor for Egg Jelly domain), Recin, fibronetin. Domains present in vertebrates but absent in amphioxus CTL models are PSI (domain found in Plexins, Semaphorins and Integrins), alpha-helix, SCP (sterol carrier protein), Calx-beta, Link domain and CSPG repeat (chondroitin sulfate proteoglycan core protein). More information about these domains and corresponding CTL architectures were detailed previously (Zelensky and Gready 2005).

More than 200 amphioxus CTL models have complex exon structures. Many domains found in these amphioxus CTL models are absent in vertebrate CTLs. The popular ones include LY (Low-density lipoprotein-receptor YWTD domain), TSP1 (Thrombospondin,type I), GPS (G-protein-coupled receptor proteolytic site domain), MAM (Domain in meprin, A5, receptor protein tyrosine phosphatase mu), NIDO (Nidogen,extracellular region), Kringle domain, FA58C (Coagulation factor 5/8 type, C-terminal), etc. However, only FA58C (BW697762), NIDO (BW795887, BW870375, BW839447, BW815452 and BW882828) and LY (Yu and Xu. unpublished data) have EST evidence so far.

References

Bell JK, Mullen GE, Leifer CA, Mazzoni A, Davies DR et al. (2003) Leucine-rich repeats and pathogen recognition in Toll-like receptors. Trends Immunol 24(10): 528-533.

Boulanger LM, Shatz CJ (2004) Immune signalling in neural development, synaptic plasticity and disease. Nat Rev Neurosci 5(7): 521-531.

Chen BE, Kondo M, Garnier A, Watson FL, Puettmann-Holgado R et al. (2006a) The molecular diversity of Dscam is functionally required for neuronal wiring specificity in Drosophila. Cell 125(3): 607-620.

Chen Y, Aulia S, Li L, Tang BL (2006b) AMIGO and friends: an emerging family of brain-enriched, neuronal growth modulating, type I transmembrane proteins with leucine-rich repeats (LRR) and cell adhesion molecule motifs. Brain Res Brain Res Rev 51(2): 265-274.

Du Pasquier L (2005) Immunology. Insects diversify one molecule to serve two systems. Science 309(5742): 1826-1827.

Layalle S, Ragone G, Giangrande A, Ghysen A, Dambly-Chaudiere C (2004) Control of bract formation in Drosophila: poxn, kek1, and the EGF-R pathway. Genesis 39(4): 246-255.

Nurnberger T, Brunner F, Kemmerling B, Piater L (2004) Innate immunity in plants and animals: striking similarities and obvious differences. Immunol Rev 198: 249-266.

Pancer Z, Cooper MD (2006) The evolution of adaptive immunity. Annu Rev Immunol 24: 497-518.

Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK et al. (2005) The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci U S A 102(27): 9577-9582.

Schmidt O, Faye I, Lindstrom-Dinnetz I, Sun SC (1993) Specific immune recognition of insect hemolin. Dev Comp Immunol 17(3): 195-200.

Tagawa H, Karnan S, Kasugai Y, Tuzuki S, Suzuki R et al. (2004) MASL1, a candidate oncogene found in amplification at 8p23.1, is translocated in immunoblastic B-cell lymphoma cell line OCI-LY8. Oncogene 23(14): 2576-2581.

Watson FL, Puttmann-Holgado R, Thomas F, Lamar DL, Hughes M et al. (2005) Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science 309(5742): 1874-1878.

Yang, Z., W.S. Wong, and R. Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22:** 1107-1118.

Zelensky AN, Gready JE (2005) The C-type lectin-like domain superfamily. Febs J 272(24): 6179-6217.

Table S1. Cross-species comparison of the immune-related protein domains.

	fly	urchin	amphioxus	zebrafish	human	notes
domains involved in pathogen recognition and clearance						
AAA(NACHT)	299	476	509	296	339	ologomerizaton of NLR
SR	14	1966	479	183	129	SRCR receptors
CLTLD	39	346	1316	388	175	C-type lectins
LDLa	329	743	1389	276	272	low density lipoprotein receptors class A
IG	1027	1080	1675	1679	2290	immunoglobulin
LRR	565	1150	5485	1687	1128	present in NLR, TLR, LRRIG
CCP	84	741	1675	307	592	complement system, complement control proteins
TSP1	77	328	796	232	382	complement system, Thrombospondin type 1 repeats
VWA	10	93	350	144	226	complement system,von Willebrand factor type A
FBG	18	100	399	43	37	complement system, fibrinogen-C-terminal domain
C1q	0	8	79	56	43	complement system
CUB	88	608	669	154	222	complement system, Domains in C1r, C1s
MACPF	2	28	50	20	17	membrane-attack complex/perforin
CASc	8	48	65	25	31	caspases
domains of cytokines and their receptors						
TNFR	1	8	66	25	81	tumor necrosis factor receptor
TNF	2	2	31	13	23	tumor necrosis factor
EGF	344	1907	1164	391	437	Epidermal growth factor domain
domains mediating protein interactions						
CARD	0	24	158	15	24	death fold domain
DEATH	11	91	491	29	45	death fold domain
DED	0	14	132	12	22	death fold domain
TIR	13	337	163	18	28	Toll/IL-1 receptor domain
TRAF	6	20	38	30	18	tumor necrosis factor receptor assciated factors
ANK	672	12635	818	1346	1461	ankyrin repeats
TPR	216	520	4074	421	466	Tetratricopeptide repeats
WD40	1072	2063	1895	1537	1540	WD40 repeats, present in Apaf-1
SAM	63	109	111	119	126	Sterile alpha motif, present in adaptor SARM1
SPRY	33	29	69	248	146	domains in butyrophilin/marenostrin/pyrin
domains involved in signal transduction						
SH2	77	106	49	164	161	Src homology 2 domains
SH3	208	236	299	371	426	Src homology 3 domains
PDZ	204	226	210	397	445	
PH	152	210	185	346	421	pleckstrin, inositol phosphate binding
PI3Kc	14	17	22	20	24	Phosphoinositide 3-kinase
S_TKc	413	509	511	883	598	Serine/Threonine protein kinases
TyrKc	77	148	426	153	156	Tyrosine kinase
RAS	18	32	32	25	13	small GTPase
RAS-RAB	3	66	51	32	24	small GTPase
RAS-RAN	73	117	113	125	117	small GTPase
RAS-RHO	8	27	13	32	14	small GTPase
PTPc	47	83	80	82	114	Protein tyrosine phosphatase, catalytic domain
PKC-C1	53	47	60	107	120	protein kinase C, C1 domain
PKC-C2	119	195	215	281	296	protein kinase C, C2 domain

NOTES:

1. Protein sets of human, zebrafish, and *D. melanogaster* are downloaded from NCBI FTP site. Notably, the NCBI predicted protein sets are not non-redundant and no procedure is performed to exclude those redundant protein entries. Proteins of sea urchin is also used the NCBI predicted protein set, where many domains are represented by two alleles. Proteins of amphioxus is used the JGI predicted protein set, where 75% loci are represented by two haplotypes. Taken together, this comparison analysis is APPROXIMATE because the estimation of the domain number is APPROXIMATE.
2. HMMER2.0 and SMART domain set is used to perform this domain estimation at the e-value cutoff < 0.01.

Table S2. Information of 927 amphioxus CTL gene models containing single CTLD domain.

Large family ID	Family (1)	Gene model numbers	small CTL	without sugar binding motifs (2)	Collectin structure (3)	major domain content	Notes, EST evidence and expression
A	A01	189	Yes		2 models	single CTLD	
	A02	19	Yes			single CTLD	
	A03	7	Yes		Yes	COL-CTLD	
	A04	4	Yes			single CTLD	
	A05	5	Yes		Yes	COL-CTLD	
	A06	14	Yes		Yes	COL-CTLD	
	A07	6	Yes	Yes		single CTLD	
	A08	13	Yes			single CTLD, CCP-CTLD	
	A09	5	Yes			single CTLD	
	A10	6	Yes			single CTLD	
	A11	8	Yes			single CTLD	
	A12	4	Yes	Yes		single CTLD	
	A13	6	Yes		Yes	COL-CTLD	
	A14	4	Yes		Yes	COL-CTLD	
	A15	4	Yes		Yes	COL-CTLD	
	B01	15	Yes			single CTLD	EST, secreted?, gut/skin, EU183372
	B02	9	Yes			single CTLD	
	B03	8	Yes			single CTLD	
	B04	9	Yes			single CTLD	
B	C01	36	Yes			single CTLD	
C	C02	108	Yes			single CTLD	
	D01	6		Yes		NIDO-CTLD	EST, secreted
	D02	4		Yes		uncertain	
	D03	7		Yes		EGF,VWF,CCP	
	D04	8	Yes			single CTLD	
	D05	8	Yes			single CTLD	
	D06	6				uncertain	
	D07	4				uncertain	contain FA58C, etc
	D08	7				uncertain	
	D09	4		Yes		uncertain	
	D10	4	Yes			single CTLD	
D	D11	100	Yes		1 models	single CTLD, EGF-CTLD	a loose subfamily with diverged members; EST, secreted, gut/skin, EU183370
	D15	4		Yes		CUB-CTLD	
	E01	6	Yes			single CTLD	
	E02	11	Yes			single CTLD	EST, Secreted?, gut/skin, EU183371
	E03	18	Yes		2 models	single CTLD	
	F01	4	Yes			single CTLD	
	F02	10	Yes			single CTLD	
	G01	12				uncertain	contain EGF
	G02	12		Yes		uncertain	contain EGF
	G03	4	Yes	Yes		EGF-CTLD	
	G04	9				uncertain	Contain VWF, CCP, etc
E	G05	40	Yes	Yes		EGF-CTLD-EGF	EST, secreted, gut, EU183373~EU183375
	others	160	n/a	n/a	21 models	n/a	
	total	**927**	**692**	**91**	**66**		

(1) Only 43 subfamilies that have at least 4 members are shown.

(2) There are 483 out of 927 CTLD containing EPN and QPD. There are more "unusual" patterns could be viewed as derivatives of EPN or QPD, like QPS, EPS, EPK, EPE, QPN, EPD, etc. All these add up to no less than 650 CTLDs, but the number depends on how we define what are EPD/QPD-derived patterns. However, members of 10 subfamilies completely lack these motifs (marked by "Yes").

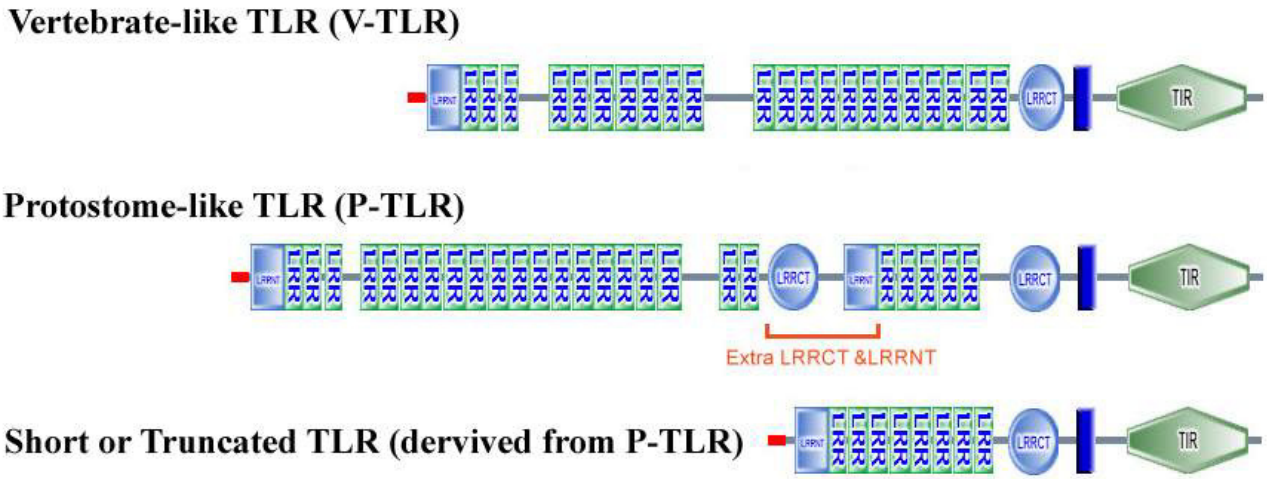
(3) These gene models contain similar structure (COL-CTLD) to vertebrate collectins, but they are not necessary to have similar primary sequences to collectins. There are 6 subfamilies of COL-CTLD structure. There are also some COL-CTLDs dispersed in other subfamilies, despite no EST evidence for them at present.

Table S3. Grouping of the amphioxus DFD gene models based on their architectures.

DFD gene groups	Domain architecture	number of gene models
ANK-CARD		13
ANK-DD		9
Apaf1-like	CARD/DD/DED+NBARC+WD40/TPR	20
CARD-3DD		5
CARD-DD-NRF		2
CARD-FB3	CARD+FB3+weakNACHT	4
CARD-TIR		18
CARD-X2	X2 is an unknown domain	2
CARD-ZnF1		11
CASP2-like	CARD+CASP	11
CASP7L		4
CASP8	DED+DED+CASP	2
CRADD-like	CARD+DD	4
DAPK	STK+DD	2
DD-CARD		8
DD-DD		2
DD-FB3		3
DD-ZnF		2
DED-CARD		4
DEDD		2
DED-GBP		4
DED-SPRY		1
DED-TPR		4
DLR	CARD/DD/DED+LRR	22
DR	TNFR+DD	19
FADD-like	DED+DD	7
FB3-CARD		3
GLY	DED+DEATH+OTUB+X3+Gly TPR+X3+weakTIR+DEATH+Gly CARD+RAS+DEATH+Gly	38
IGFN-TM-CARD		11
IRAK-like	DEATH+STK	5
LRR-DD-STK		17
LRR-DD-TIR		2
LRR-RAS-CARD		1
LRR-RAS-DD		9
LRR-TM-DD		3
multiDD		5
multiDED		1
MyD88-like	DD+TIR	12
NHL-DD		8
NHL-DED		55
NHL-DED-DD		12
NLR/NLAA	CARD/DD/DED/TIR+NACHT+LRR	50
PEA15	DED	2
PIDD-like	LRR+DEATH	4
RIG-I-like	DED/TIR/DD/CARD+helicase	6
RIPK-like	STK+DEATH	6
SPRY-DD		2
THOC1	Containing DEATH	2
tripleCARD		2
UNC5-like	Igcam+TSP1+TM+ZU5+DEATH	15
X1-CARD	X1 is an unknown domain	13
Orphan death-fold domains similar to genes listed above		52
Unknown orphan death-fold domains		111
Total gene models		**632**

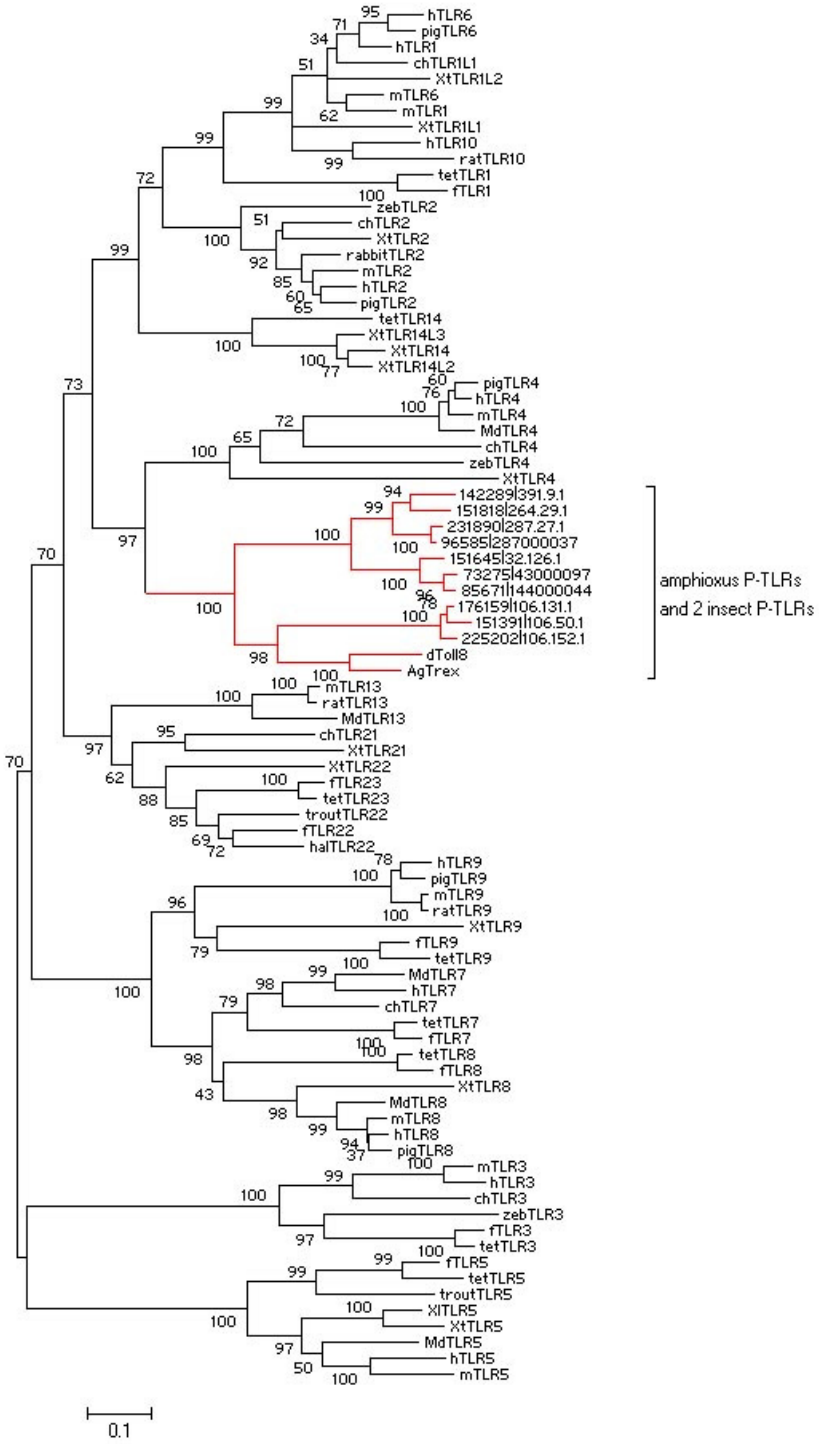
Abbreviations not explained elsewhere. ANK=ankyrin; DD=DEATH; NRF=Nose Resistant to Fluoxetine-4; FB3=fribronetin type 3; ZnF=zinc finger; CASP=caspase; STK=serine/threonine kinase; GBP= Guanylate-binding protein; SPRY=domain first identified in splA and ryanodine receptor; Gly= Glycosyl transferase; IGFN=Ig and fibronectin; RAS=small G protein ras; TM=transmembrane; NHL=first identified in NCL-1, HT2A, LIN-41; NLAA=NLR without DFD or LRR domain; ZU5=Domain present in ZO-1 and Unc5-like netrin receptors.

Figure S1. Two structural types of TLR.



Schematic of two TLR structures: 1) vertebrate-like TLR (without extra LRRCT-LRRNT motif), 2) protostome-like TLR (with extra LRRCT-LRRNT motif) and short TLR derived from protostome-like TLR (having a cytoplasmic TIR highly similar to that of protostome-like TLR). Figures is produced by SMART tools (http://smart.embl-heidelberg.de/).

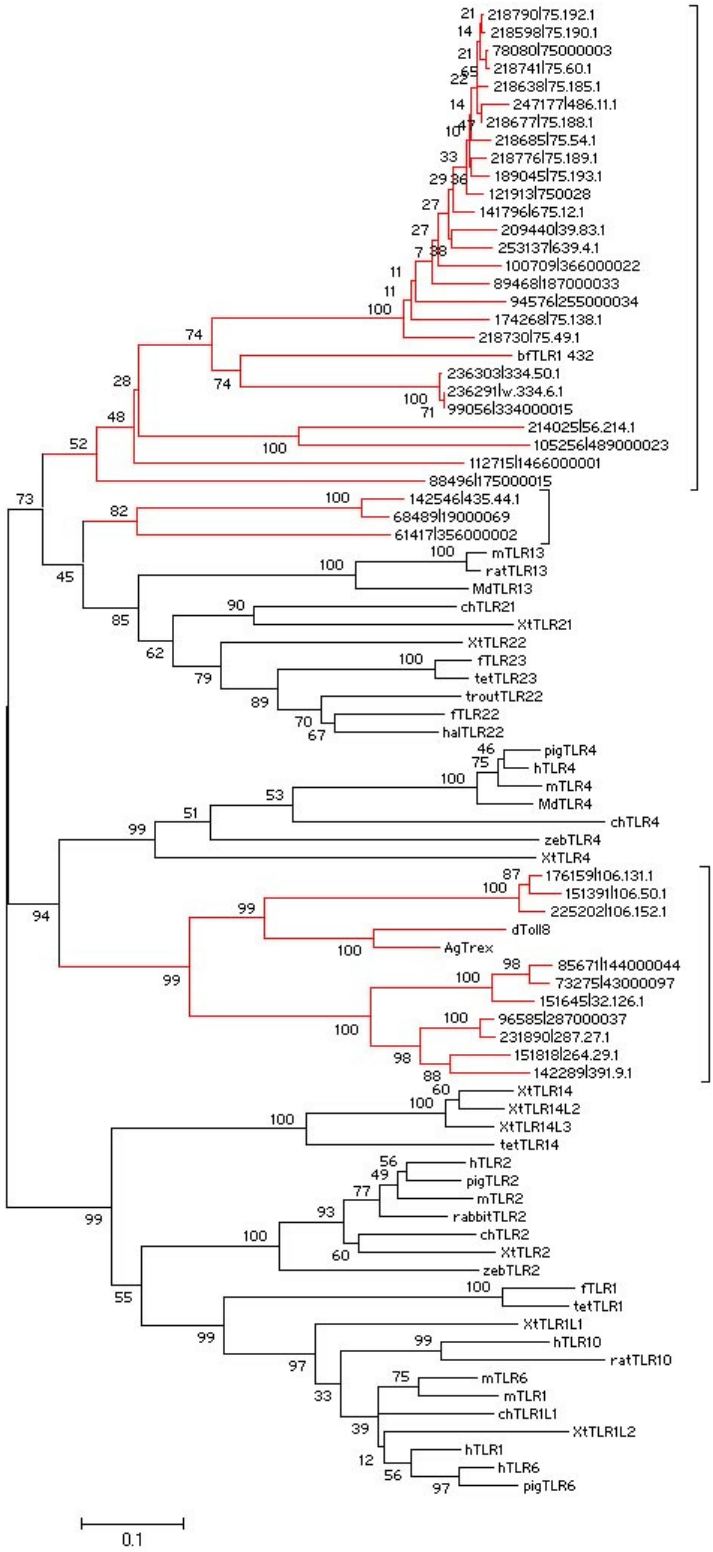
Figure S2. Phylogenetic analysis of amphioxus P-TLRs and all vertebrate TLR families.



A minimum-evolution tree of amphioxus P-TLRs and all six vertebrate TLR families based on TIR domain. The amphioxus P-TLRs are red colored. This tree shows that amphioxus P-TLRs form a stable clade with vertebrate TLR4 family even in the presence of those highly divergent vertebrate TLR lineages (TLR7, TLR3 and TLR5). This pattern may be caused by long-branch attraction. However, vertebrate TLR3/5/7 lineages are much “longer” branches, so, this pattern may also reflect that vertebrate TLR4 lineage derived from an ancient P-TLR lineage by the losing of typical P-TLR structure (an extra LRRCT-LRRNT pair, see **Figure S1**).

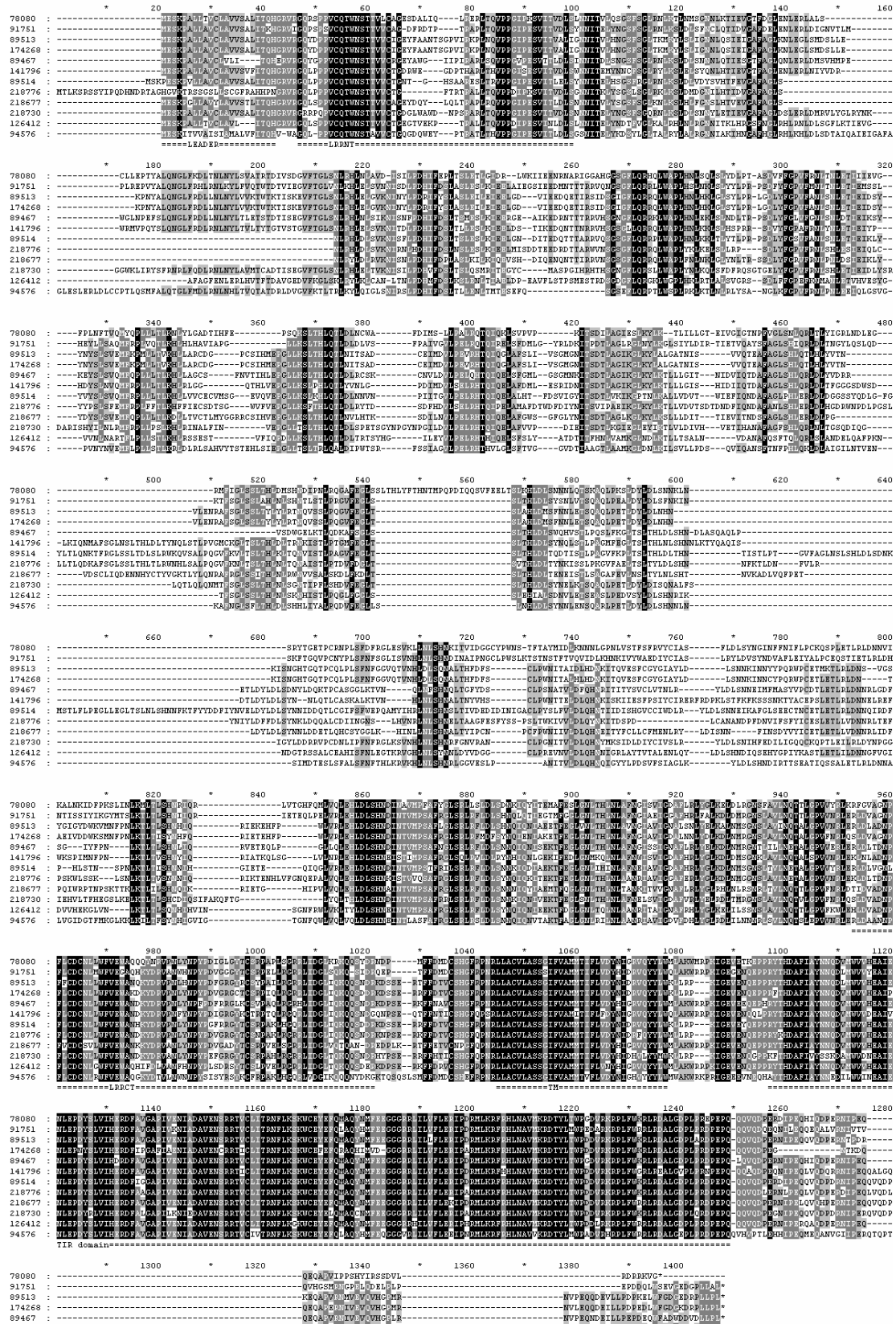
Nevertheless, FEW sites in the alignment that can clearly separated amphioxus P-TLRs and vertebrate TLR4 lineage from other TLR sequences (data not shown). Hence, this branching pattern remains dubious.

Figure S3. Phylogenetic analysis of amphioxus TLRs and vertebrate TLR1/4/11 lineages.



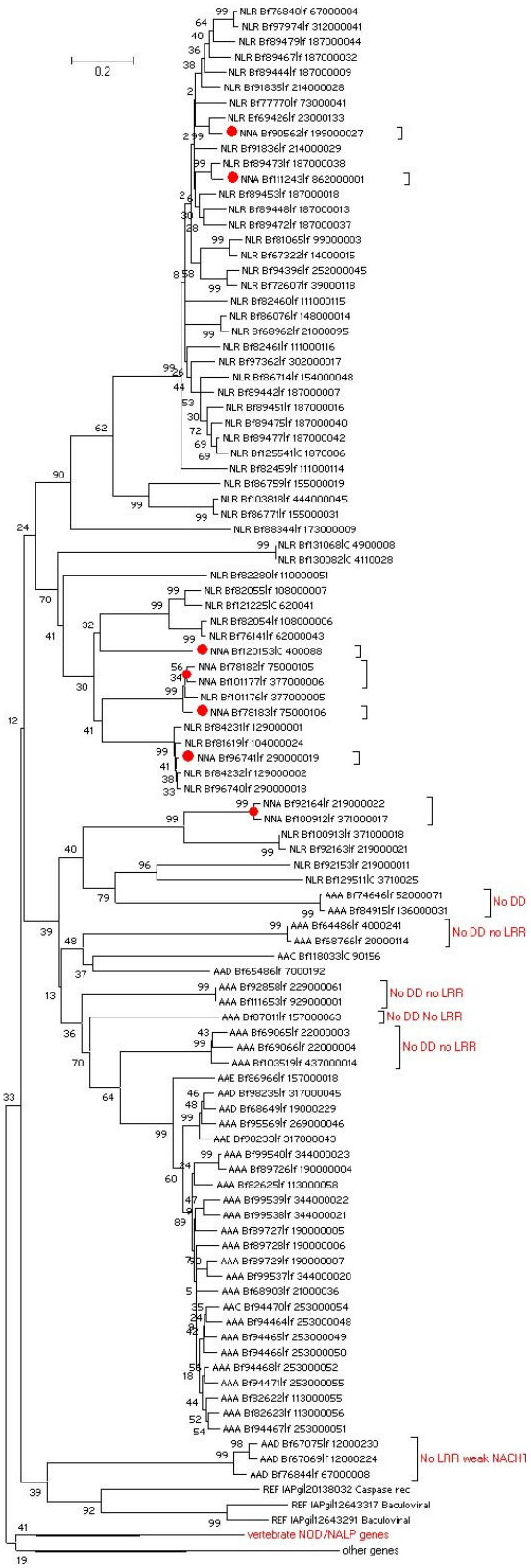
This is a minimum-evolution tree of amphioxus TLRs and vertebrate TLR1, TLR11 and TLR4 lineages. This tree is different from the tree in **Figure 1** in that it excludes insect V-TLRs and other divergent amphioxus and vertebrate TLR sequences because they are too divergent to affect the significance of the tree. As the tree shows, it gains more statistic significance after deleting those divergent sequences.

Figure S4. Alignment of 12 V-TLRs of the amphioxus SC75 lineage.



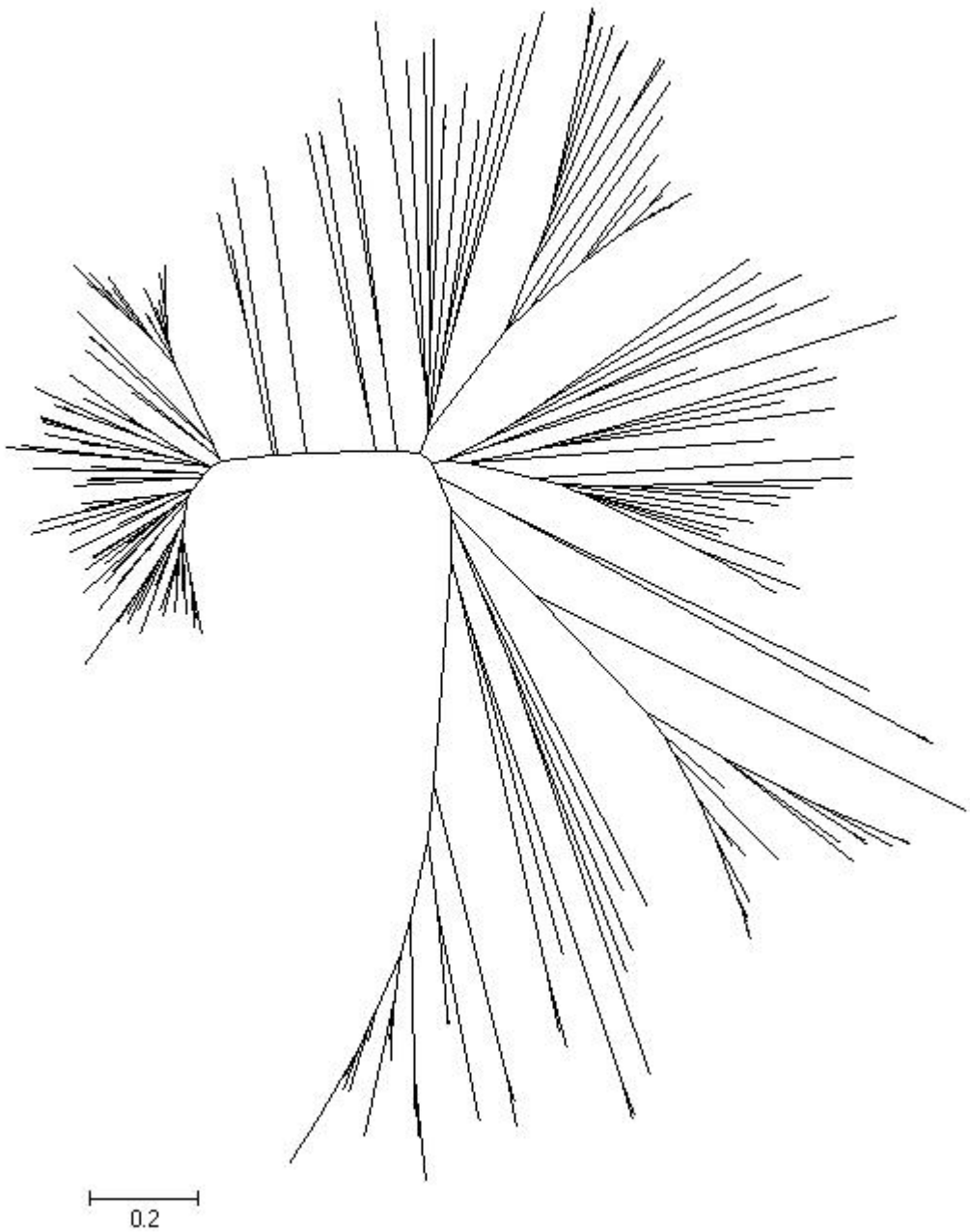
This alignment is produced with 12 correctly predicted, high-quality TLR protein sequences of the amphioxus SC75 lineage. It shows that the SC75 TLR lineage is highly conserved in TIR domain and highly diversified in LRR region. The LRR region is located between LRRNT and LRRCT.

Figure S5. Phylogenetic analysis of the NACHT domain of all amphioxus typical NLRs.



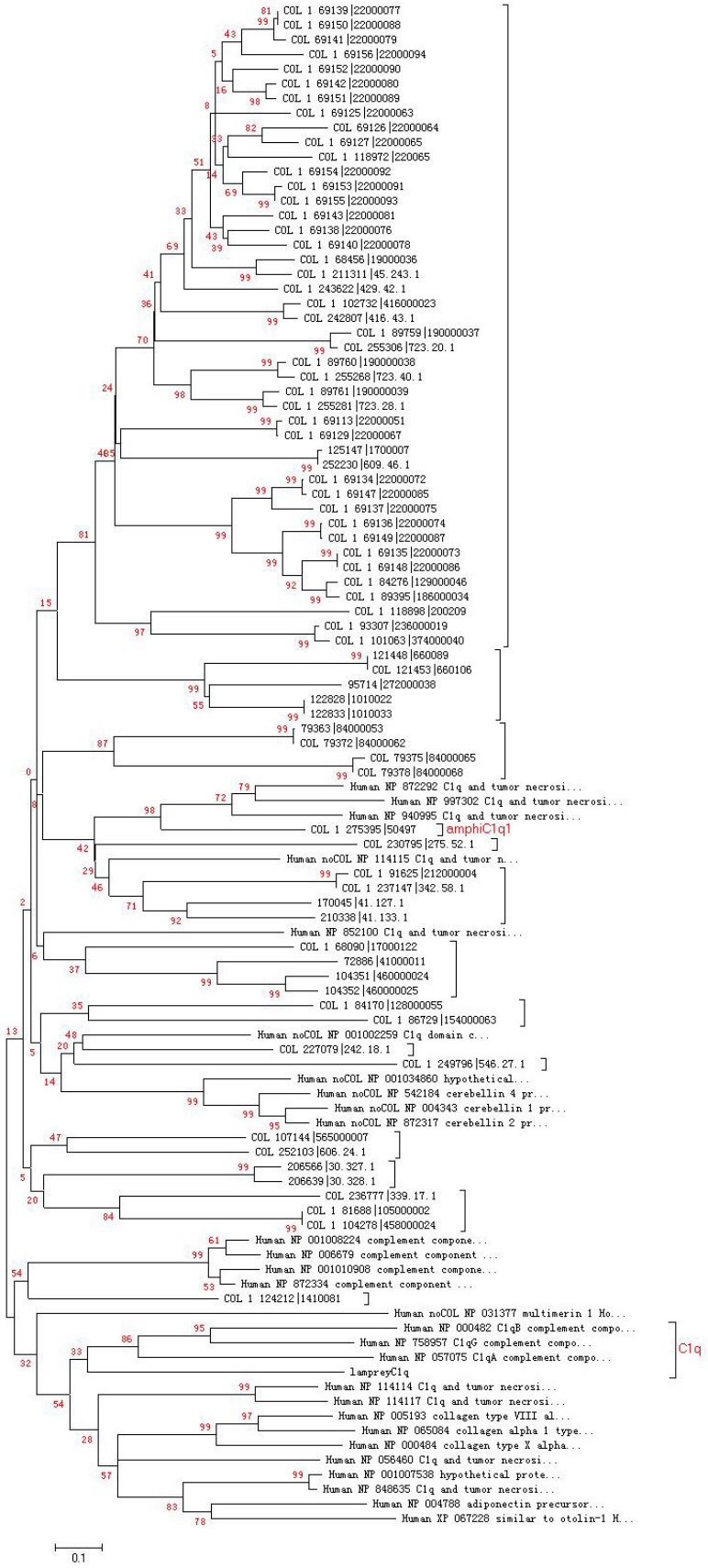
A minimum-evolution tree of the NACHT domain of 96 NACHT-containing NLR gene models. Those NLRs without NACHT domains (termed DLR) are not included in this tree. Abbreviations: NLR=NLR containing NACHT and LRR; AAA~AAD=NLR containing no detectable LRR regions.

Figure S6. Phylogenetic analysis of the first IGcam domain of amphioxus LRRIG models.



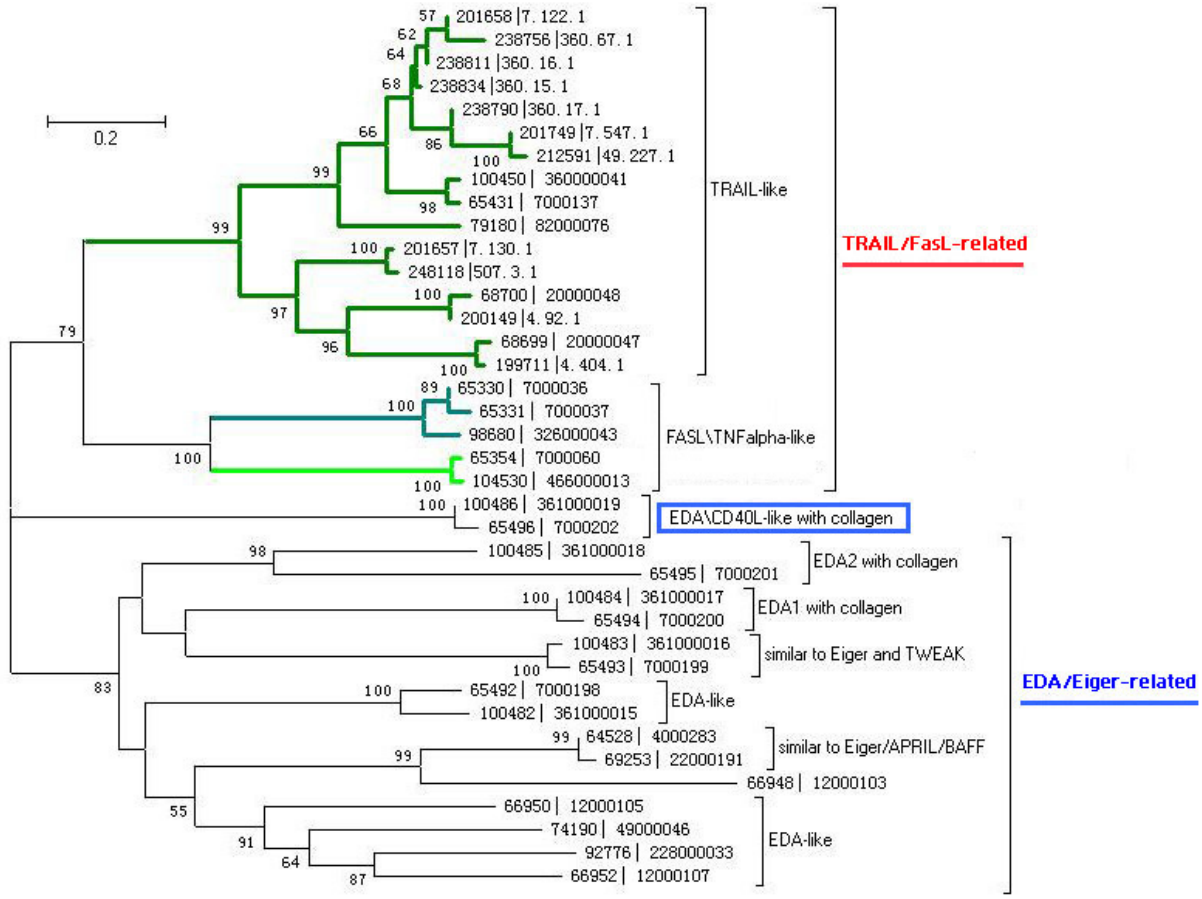
A minimum-evolution tree of the first IGcam domain of 229 amphioxus LRRIG gene models.

Figure S7. Phylogenetic analysis of the C1q domain of all amphioxus C1q-like models.



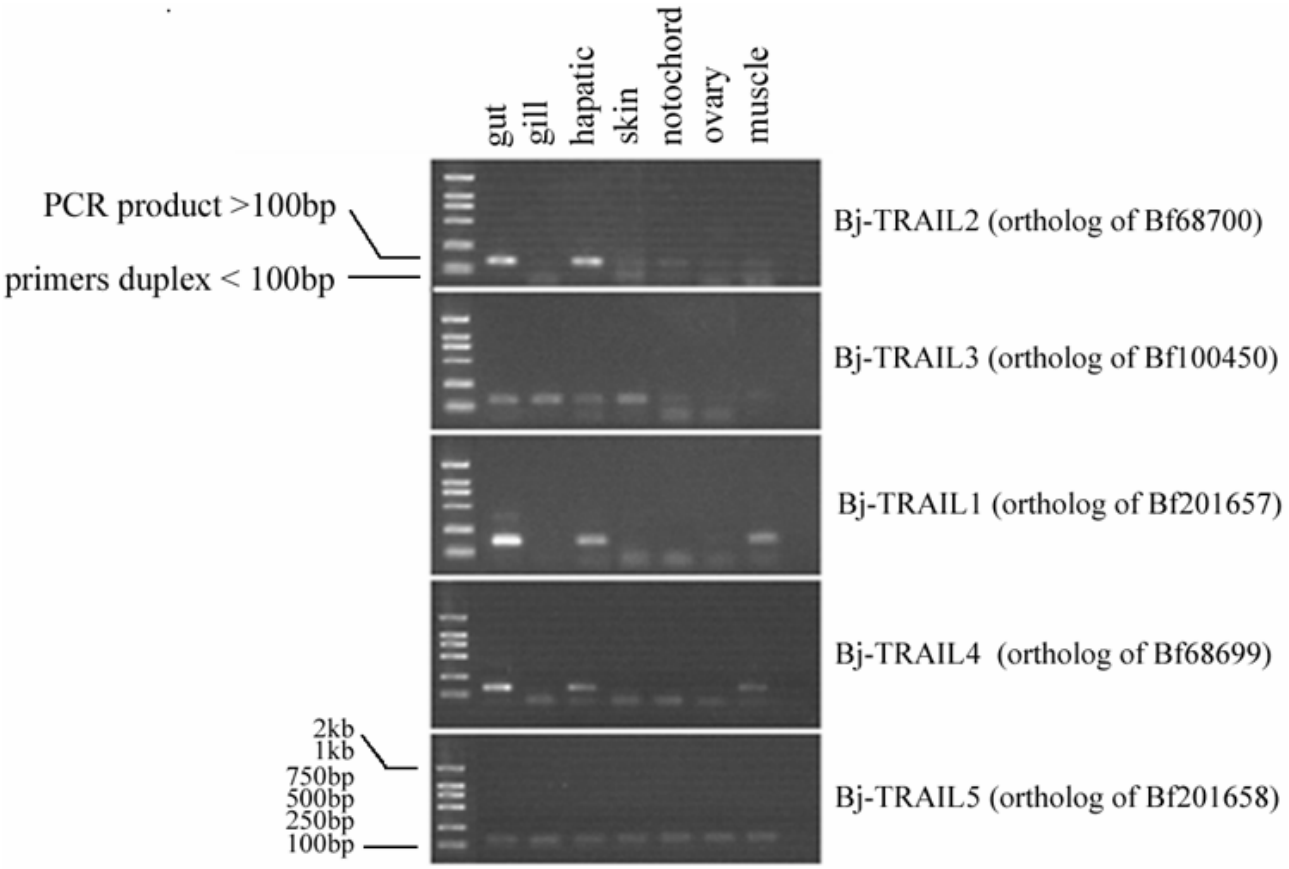
A minimum-evolution tree of all amphioxus C1q-like models, all human C1q-like genes and the lamprey C1q. Human C1q-like genes and the lamprey C1q are indicated. The amphioxus C1q-like models with N-terminal COL (collagen) domain are marked by “COL”.

Figure S8. Phylogenetic analysis of the TNF domain of all amphioxus TNF models.



A minimum-evolution tree of all amphioxus TNF gene models. This tree is different from the tree in **Figure 5** in that it contains only amphioxus TNF models and hence it clearly indicates the evolution of TNF family within the amphioxus lineage. The tree also indicates TNF models with N-terminal collagen (COL) domains.

Figure S9. Expression profiles of five amphioxus TRAILs.

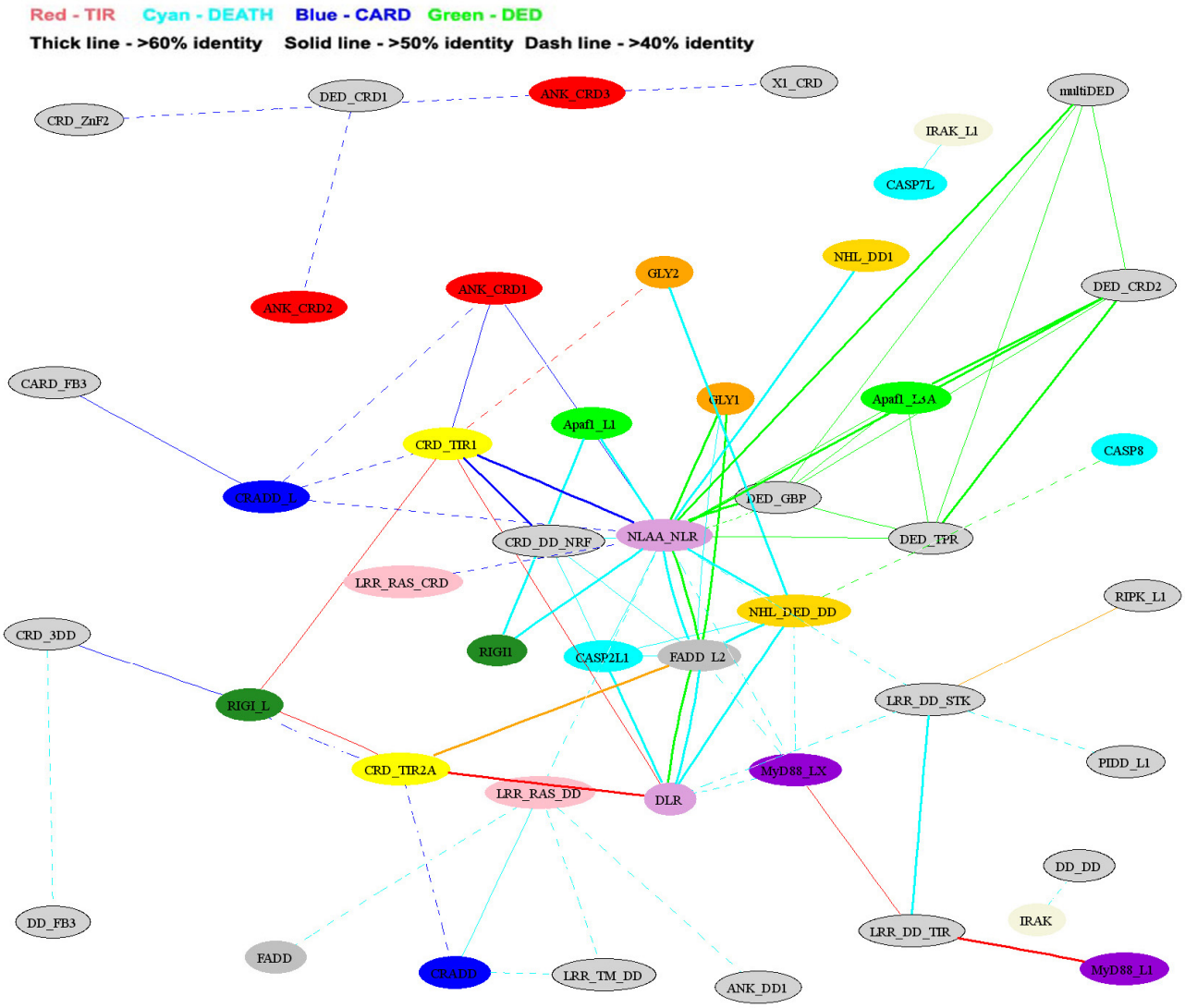


This is the semi-quantitative RT-PCR result of five amphiTRAILs (from *B. japonicum*), which indicates that different amphiTRAILs have different expression pattern. The result is confirmed by real-time RT-PCR (unpublished data). Bj means *B. japonicum*.

The corresponding experimental procedure (in brief):

1) 300ul stock reaction solution contains dNTP, buffer, *Hot start* LA Taq (TAKARA corporation) and 20ul 1ST strand cDNA from a certain tissue (synthesized by Invitrogen supperscriptIII).
2) Separate stock reaction solution into five tubes, 48 ul per tube. Each tube adds in 2 ul specific primer pair designed for a amphiTRAIL.
3) Repeat 1) and 2) for the other four amphiTRAILs.
4) 30~35 cycles PCR amplification, then electrophoresis with 5 ul PCR products.
5) Repeat three times.

Figure S10. Schematic illustration of the domain similarity between different DFD architectures.



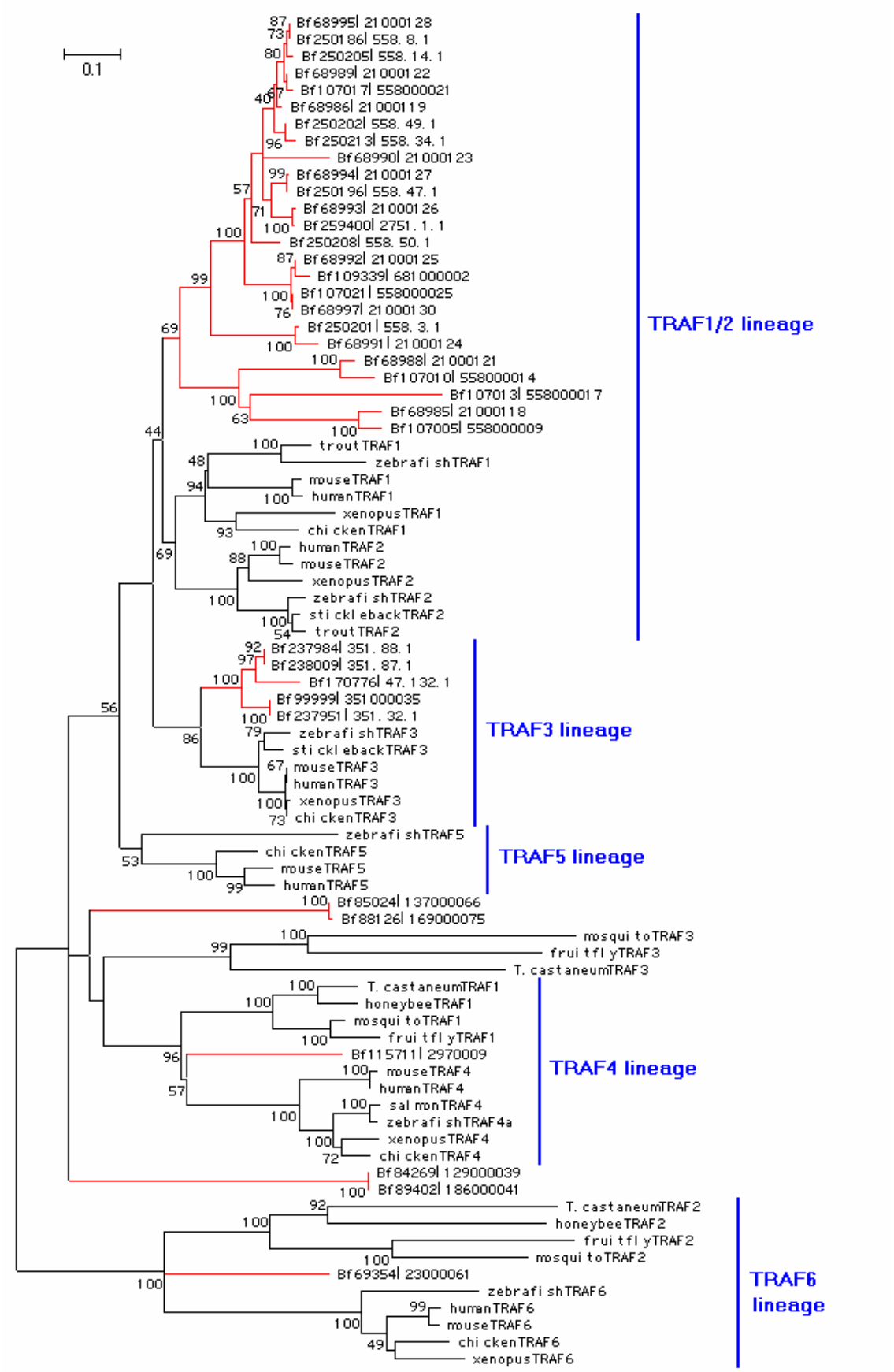
This analysis supports that dynamic domain reshuffling plays a role in shaping the huge DFD gene repertoire of amphioxus.

This analysis compares the protein sequence similarity between cognate DFD domains from different DFD gene groups, where each group contains gene models with similar domain architectures (listed in Table S3).

Lines of different color represent different domain comparison: cyan lines for DEATH versus DEATH comparison, blue lines for CARD versus CARD, green lines for DED versus DED, red lines for TIR versus TIR. Thick, thin and dash lines represent identity >60%, >50%, >40%, respectively. Identity <40% is not shown by lines.

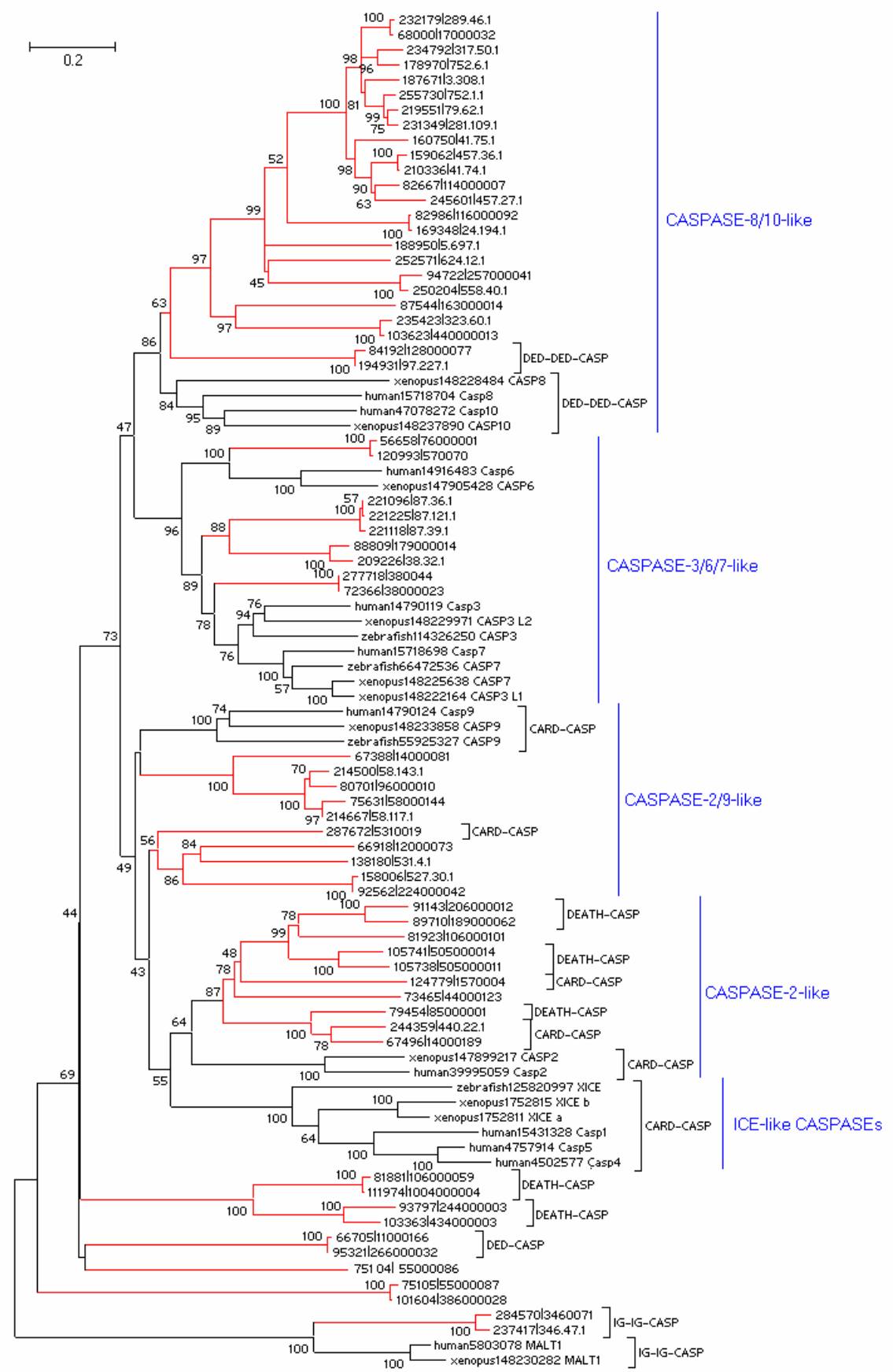
Nodes with the same color (except grey) represent sub-groups belonging to the same DFD gene group. For instance, CARD-TIR group (yellow colored) has two subgroups, CARD-TIR1 and CARD-TIR2. Although they have similar domain structure, CARDs from these two subgroups have little sequence identity (<40%), suggesting that CARDs of two groups may have different origins.

Figure S11. Phylogenetic analysis of all TRAF domains in amphioxus.



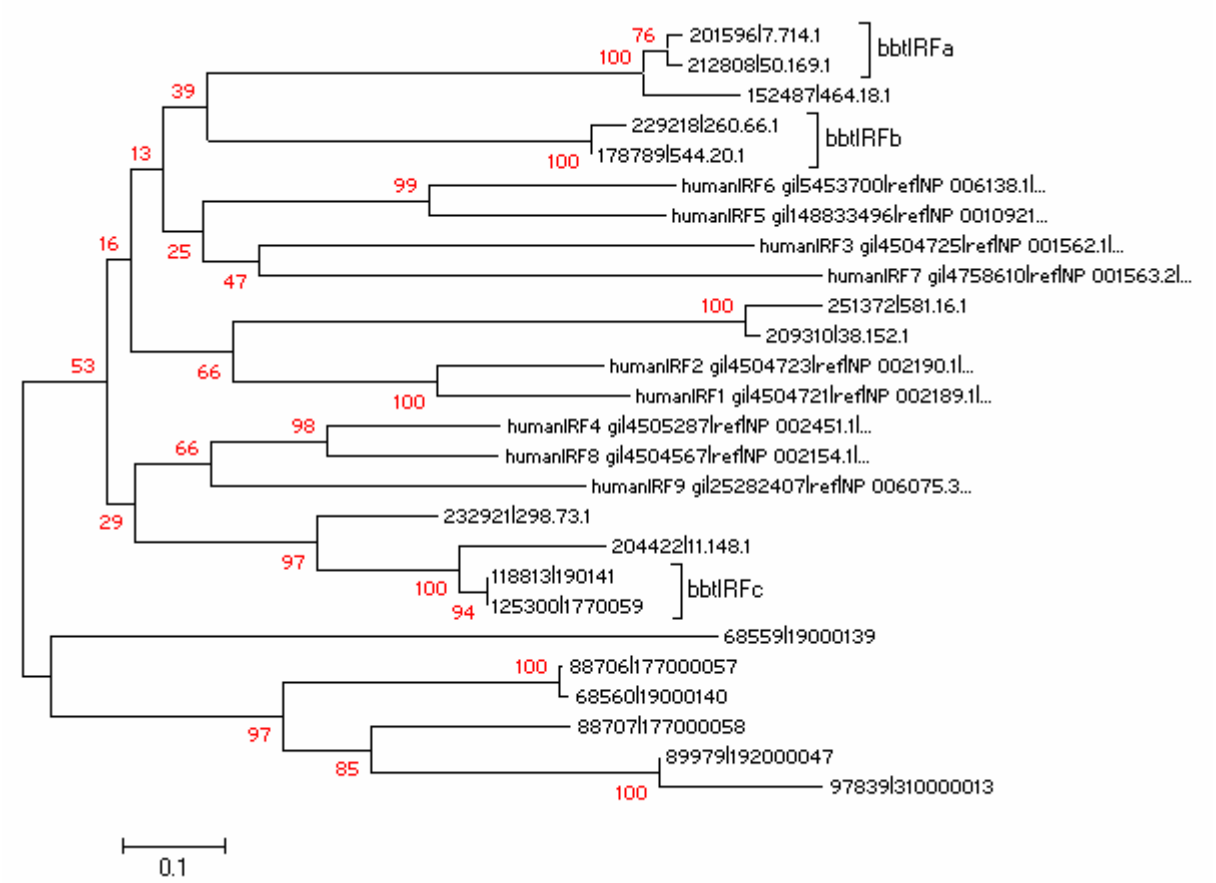
This is a minimum-evolution tree including all 36 amphioxus TRAF models and six vertebrate TRAFs and three insect TRAFs. The amphioxus TRAFs are indicated by red color.

Figure S12. Phylogenetic analysis of the caspase domain of all amphioxus caspase models.



This is a minimum-evolution tree including all amphioxus caspases and major vertebrate caspase lineage. Amphioxus caspases are indicated by red color. Domain combination of each caspase gene is also provided. If not specified, it means the gene contains no other domains or its structure is uncertain.

Figure S13. Phylogenetic analysis of all amphioxus IRF domains.



This is a minimum-evolution tree of the IRF domains of all amphioxus IRF models and human IRF genes. The human sequences are indicated and the others are amphioxus IRF sequences. We cloned three IRF full-length cDNAs from Chinese amphioxus (*B. japonicum*) and they are designated as bbtIRFa, bbtIRFb and bbtIRFc (as indicated on the tree). Expression analysis indicates that they are mainly expressed in the gut, the gill and the hepatytic diverticulum (data not shown).