Research

Open Access

Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish

Cornelia Stein[*], Mario Caccamo[†], Gavin Laird[†] and Maria Leptin[*†]

Addresses: [*]Institute for Genetics, University of Cologne, Zuelpicher Str. 47, 50674 Cologne, Germany. [†]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK.

Correspondence: Maria Leptin. Email: mleptin@uni-koeln.de

Published: 27 November 2007

Genome ***Biology*** 2007, **8:**R251 (doi:10.1186/gb-2007-8-11-r251)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/11/R251

Received: 20 April 2007
Revised: 30 October 2007
Accepted: 27 November 2007

Abstract

Background: The zebrafish has become a widely used model to study disease resistance and immunity. Although the genes encoding many components of immune signaling pathways have been found in teleost fish, it is not clear whether all components are present or whether the complexity of the signaling mechanisms employed by mammals is similar in fish.

Results: We searched the genomes of the zebrafish *Danio rerio* and two pufferfish for genes encoding components of the Toll-like receptor and interferon signaling pathways, the NLR (NACHT-domain and leucine rich repeat containing) protein family, and related proteins. We find that most of the components known in mammals are also present in fish, with clearly recognizable orthologous relationships. The class II cytokines and their receptors have diverged extensively, obscuring orthologies, but the number of receptors is similar in all species analyzed. In the family of the NLR proteins, the canonical members are conserved. We also found a conserved NACHT-domain protein with WD40 repeats that had previously not been described in mammals. Additionally, we have identified in each of the three fish a large species-specific subgroup of NLR proteins that contain a novel amino-terminal domain that is not found in mammalian genomes.

Conclusion: The main innate immune signaling pathways are conserved in mammals and teleost fish. Whereas the components that act downstream of the receptors are highly conserved, with orthologous sets of genes in mammals and teleosts, components that are known or assumed to interact with pathogens are more divergent and have undergone lineage-specific expansions.

Background

With the sequence of the zebrafish genome as well as the sequences of two pufferfish genomes nearly completed, and in view of the widespread use of the zebrafish as a model to study immunity [1], it is both pertinent and feasible to determine which of the genes that encode components of the mammalian immune system are also found in fish. In addition to being a prerequisite for using the zebrafish as a model system for the genetic analysis of human immunity, knowledge of components of immune defense systems in the zebrafish would also aid our understanding of the evolution of immunity.

Zebrafish are a member of the large group of teleost fish that, together with a small nonteleost sister group, constitute the ray-finned fishes. The ray-finned fishes diverged from the common ancestor of other bony vertebrates, which include tetrapods as well as lungfishes and coelacanths, 450 million years ago. They appear to have undergone a massive radiation about 235 million years ago, resulting in as many teleost species as there are species represented by all other vertebrates together (approximately 24,000 species in each case). One genetic event that has been regarded to be associated with the radiation of the teleosts in particular is a whole genome duplication event early in the teleost lineage. Although some genes or regions of the genome, most notably the *Hox* gene clusters, have been maintained in multiple copies, others have undergone re-diploidization. The availability of additional gene copies has been proposed to have facilitated the evolution of the high level of diversity in morphology and behavior in the teleost fish [2,3].

Components of the adaptive immune system have been studied intensively in many fish species and have been analyzed molecularly and genetically (for review [4]). Unlike the adaptive immune system, some of the systems that contribute to innate immunity are conserved throughout the animal kingdom. The presence of genes encoding components of these systems in the zebrafish and other fish was therefore not unexpected. In addition to the well studied adaptive immune genes, protein and gene families involved in innate immune mechanisms that have been analyzed in detail include the complement gene family (for review [5]), the Toll-like receptors (TLRs) [6,7], and two sets of receptor genes that encode proteins structurally similar to the immunoglobulin-type and C-type lectin domain-type of mammalian NK (natural killer cell) receptors [8-11]. Similarly, genes encoding tumor necrosis factors (TNF), ILs, IFNs, and their respective receptors have been identified in various fish species [12-18]. Together with studies on subsets of intracellular signaling molecules [19-23], these findings indicate that many components of innate immune signaling pathways known from mammals are conserved in the teleost fish. However, it is not clear whether all components are present or whether, in general, the complexity of the signaling mechanisms employed by mammals is similar in fish. For example, whereas some members of the TLR family exhibit orthologous relationships between zebrafish and mammals, there are also expansions within the TLR gene family that are specific for the zebrafish or the mammals [6,7]. Similarly, the novel immune-type receptors, which share several common features with mammalian immunoglobulin-type natural killer cell receptors, exhibit species-specific expansions and diversifications [8,10].

This report concentrates on identifying those molecules known from mammalian innate immune signaling systems that are conserved between teleost fish and mammals. The study is restricted to the pathways that have not been extensively studied by others previously. It is likely that there are also nonconserved defense systems associated with the characteristic physiologies of fish and mammals (for example, skin defense peptides), and future genetic research may well reveal additional fish-specific molecules and mechanisms.

To be able to judge orthologous relationships properly, we also included protein family members that have not been shown to have immune signaling functions, in particular because it cannot be excluded that these may have as yet unidentified roles in immune signaling, as has recently been discovered for TNF-receptor associated factor (Traf)3 [22]. We find that the families of intracellular signaling adaptors and enzymes are largely conserved. By contrast, the class II cytokines and their receptors have diverged significantly, and the NLR (NACHT-domain and leucine rich repeat containing) proteins exhibit extensive, species-specific gene amplification and diversification.

Results and discussion

As the basis for our search, we first assembled a set of sequences of mammalian genes that encode components of the TNF, IFN, and TLR pathways, and the NLR proteins in mice and humans (Figure 1). We then identified homologs of these genes in the zebrafish genome. We first checked whether Ensembl [24] or ZFIN [25] listed potential homologs and added these to our list. In cases in which putative homologs were not found in Ensembl or ZFIN, we used TBLASTN [26] to screen unfinished clones from the genome sequencing project and trace sequences from the whole genome shotgun project. If matching sequences were found, they were analyzed in detail in their genomic context and were manually annotated to generate a gene prediction, using the available mammalian sequences and any existing expressed sequence tags (ESTs) as evidence. Where gene predictions were available from the *Tetraodon nigroviridis* or *Takifugu rubripes* genomes, we also included these in our analyses, but we did not make any assemblies or annotations ourselves. A complete list of all sequences used in this study is provided in Additional data files 1-9.

We used MEGA software [27] to compare the encoded fish proteins with their mammalian counterparts. For some proteins, the annotated sequences were not complete and could not be completed because the available DNA sequence was not sufficiently reliable or had gaps. We therefore point out that the phylogenetic trees we present show relationships, but are not intended to show precise evolutionary distances.

For most of the core signal transduction components of each pathway we found clear orthologous relationships between the mammalian and the zebrafish genes (see Gene families with largely orthologous relationships between teleosts and mammals, below), as illustrated for example by the branches for Tollip (Toll-interacting protein) or Tab (Tak1-binding protein)3 in Figure 2. These branches reflect the known evolutionary relationships between the five species. Mouse and human exhibit the highest level of similarity, the two

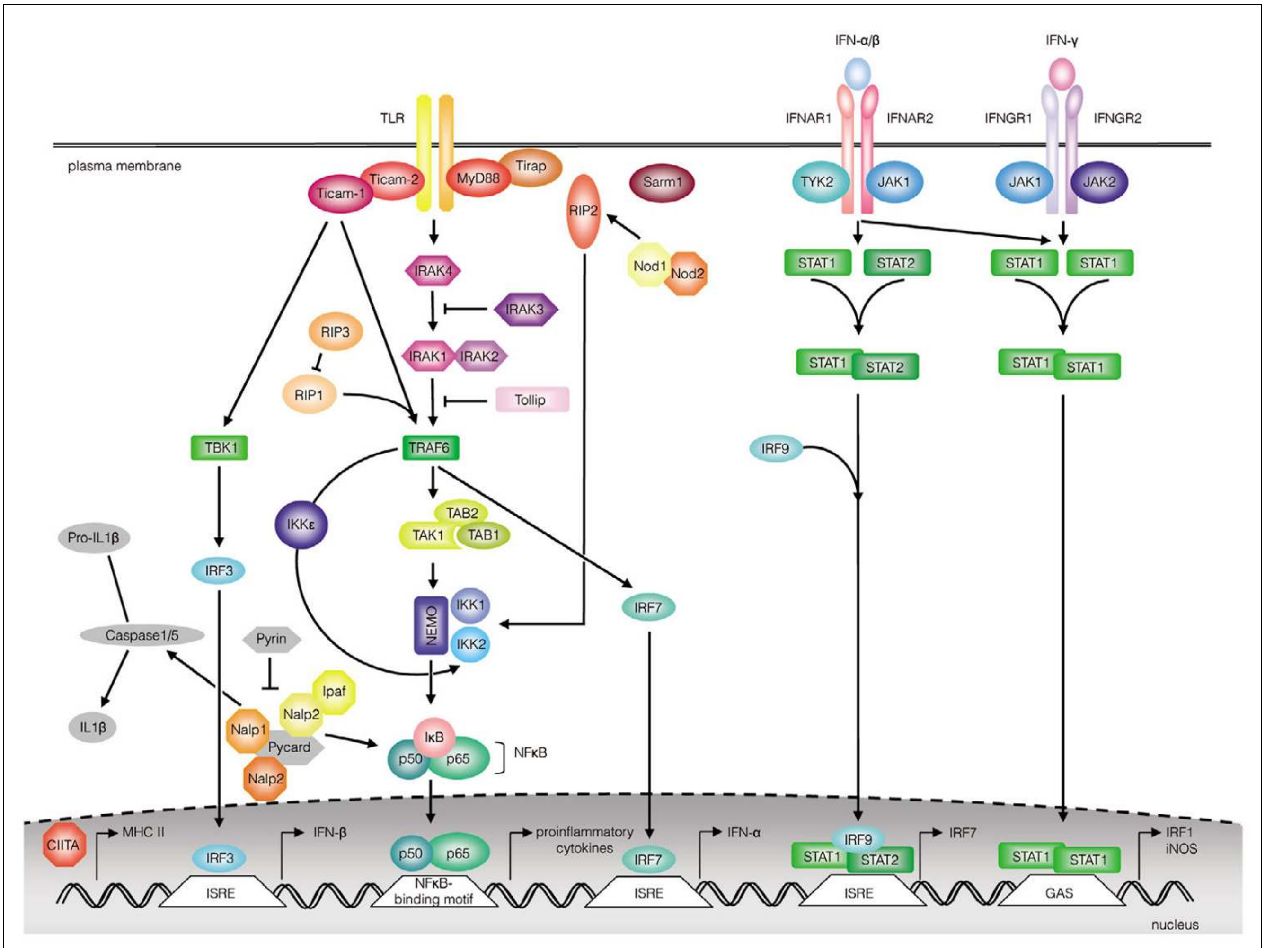

Figure 1

Components of the TLR and IFN signaling pathways and intracellular pattern recognition receptors. The molecules analyzed in this study are shown in color. For simplicity, not all members of each protein family are shown. IFN, interferon; TLR, Toll-like receptor.

pufferfish are closely related to each other, and the zebrafish is more closely related to the pufferfish than to mammals and therefore shares a branch with the pufferfish on the phylogenetic tree. In several cases, for example Tab1 and Tab2, the *Tetraodon* sequences do not group with their counterparts from *Takifugu*. In most of these cases this is due to internal deletions or insertions, or terminal deletions or extensions in the *Tetraodon* genes, which are most easily explained by unreliable predictions for these genes based on faulty assembly of the genome (see below for specific cases). We have not investigated these cases further.

For the class II cytokine receptor family the orthology was less clear (see Class II cytokines and their receptors, below) or nonexistent, as has previously been noted [17]. For one group of proteins, those containing NLRs, our comparison reveals extensive, species-specific expansion of subfamilies (see Intracellular pathogen sensors: the NACHT-domain family, below).

Each of these groups of proteins is discussed individually below.

Gene families with largely orthologous relationships between teleosts and mammals

In the protein families of the immune kinases, the adaptors in the TLR signaling pathway, the interferon response factors (IRFs), the signal transducers and activators of transcription (Stats), and the Trafs we found orthologous genes in fish for almost all of the mammalian genes. This is summarized in Figures 2 to 6. However, there were also occasional duplications or losses either in the fish or in the mammalian lineage. The findings are briefly summarized below and in the figure legends.

Kinases

The kinases were the family that exhibited the most apparent orthologies between fish and mammals. For all of the essential kinases involved in signal transduction mediated by TLR,

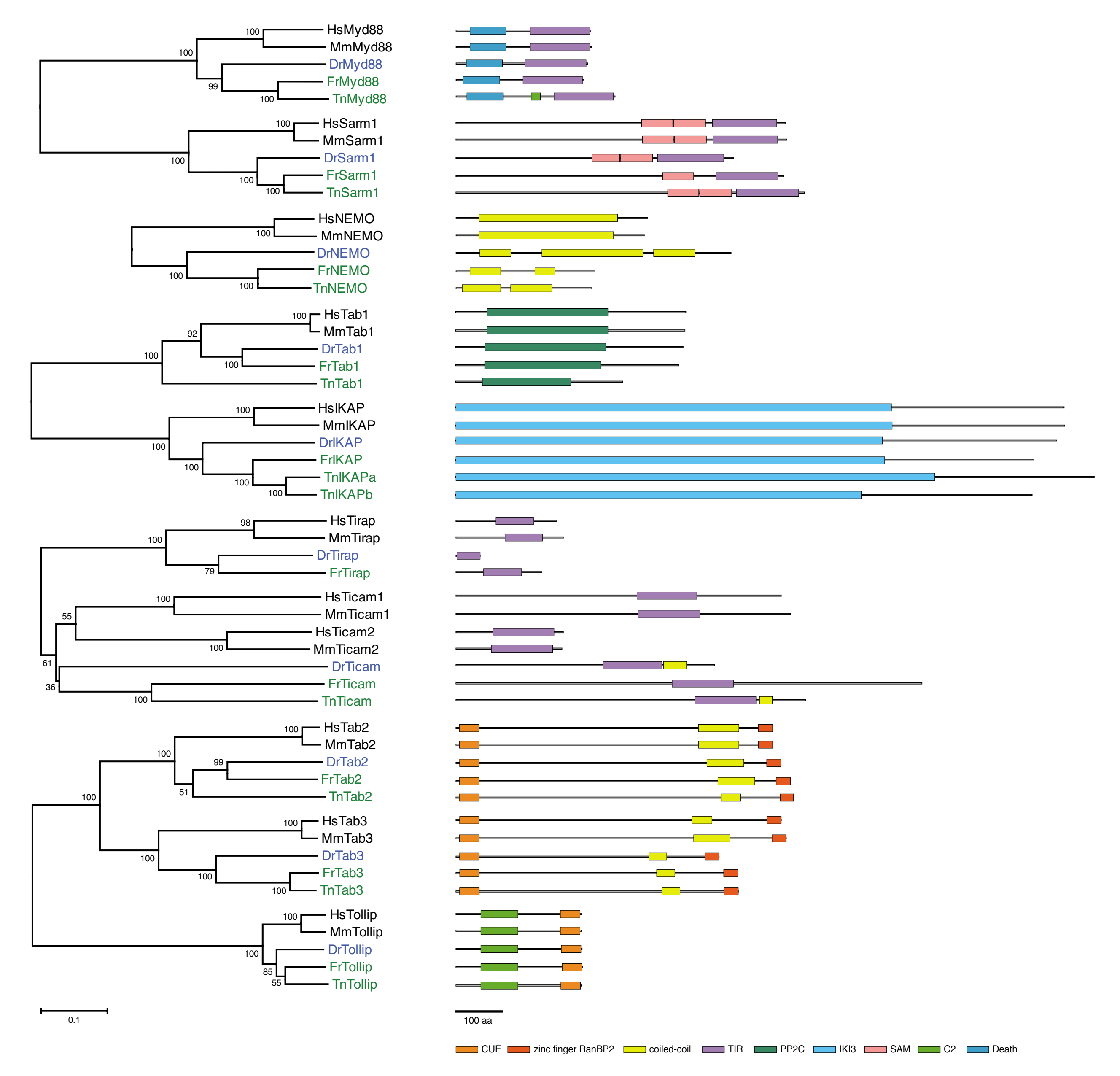

Figure 2

Phylogenetic trees of the innate immune signaling adaptors and diagrams of their protein structures. The fish protein names are highlighted in blue (Dr [*Danio rerio*]) or green (Fr [*Takifugu rubripes*] and Tn [*Tetraodon nigroviridis*]). The numbers in the tree indicate the bootstrap values. Scale: interval of 0.1 amino acid substitutions. Hs, *Homo sapiens*, Mm, *Mus musculus*. Protein domains are shown as boxes based on identification by Pfam [55] or Smart [56]. Some domains were not recognized by these programs, although manual inspection indicated clear conservation of the domain within the protein family. These domains are also shown as boxes in the diagrams. The identities of the domains are listed at the bottom. Scale bar = 100 amino acids. The *Tetraodon* version of the Ikap (IKK [Inhibitor of nuclear factor-κB kinase] complex associated protein) gene contains two full repeats of the IKI3 domain. It is not clear whether this prediction is due to an error in the genome assembly or whether the gene does indeed contain an internal duplication covering the whole length of the gene found in other species. The two halves of the predicted gene were treated as separate peptides in the phylogenetic tree and the diagram.

TNF, and nucleotide oligomerization domain containing protein (Nod), we find orthologs in zebrafish and in most cases also in pufferfish. IL-1 receptor associated kinase (IRAK)2, which is thought to serve as an accessory protein in combination with IRAK1, was not found in any of the three fish. This suggests that it has arisen from a duplication event that occurred only within the mammalian lineage (Figure 3). The alternative, loss of IRAK2, for example in the teleost lineage

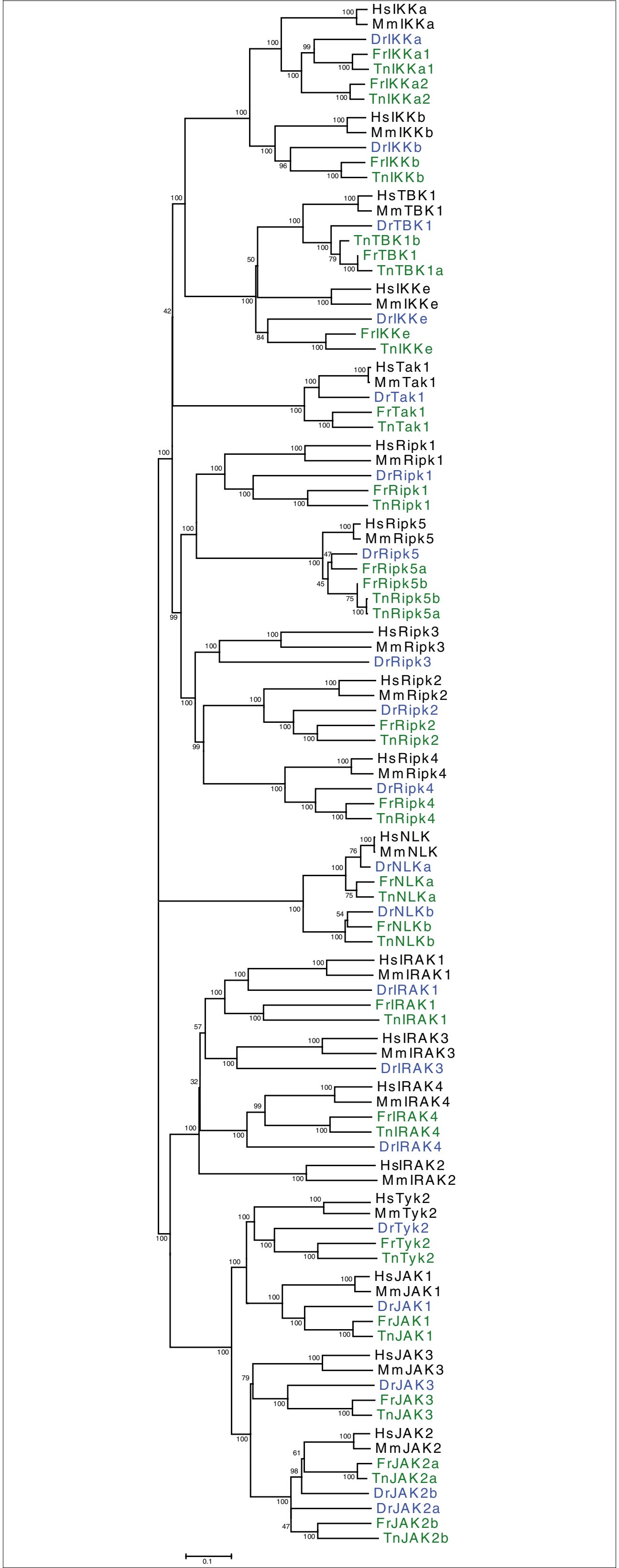

Figure 3
Phylogenetic tree of the kinases. Details of the tree are as in Figure 2.

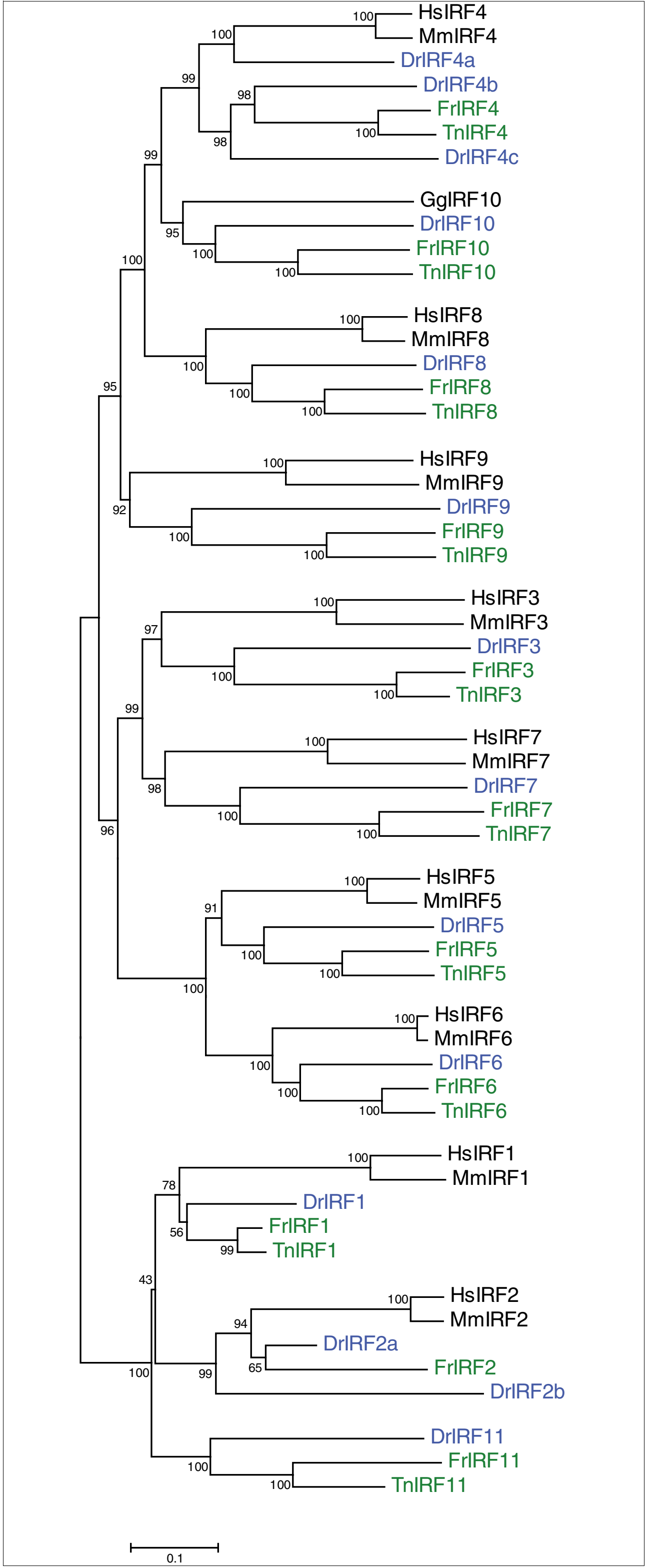

Figure 4
Phylogenetic tree of the interferon response factors. Details are as in Figure 2. The chicken (Gg [*Gallus gallus*]) IRF10 was included to show its relationship to fish IRF10, because no ortholog for this gene is found in mammals. IRF, interferon response factor.

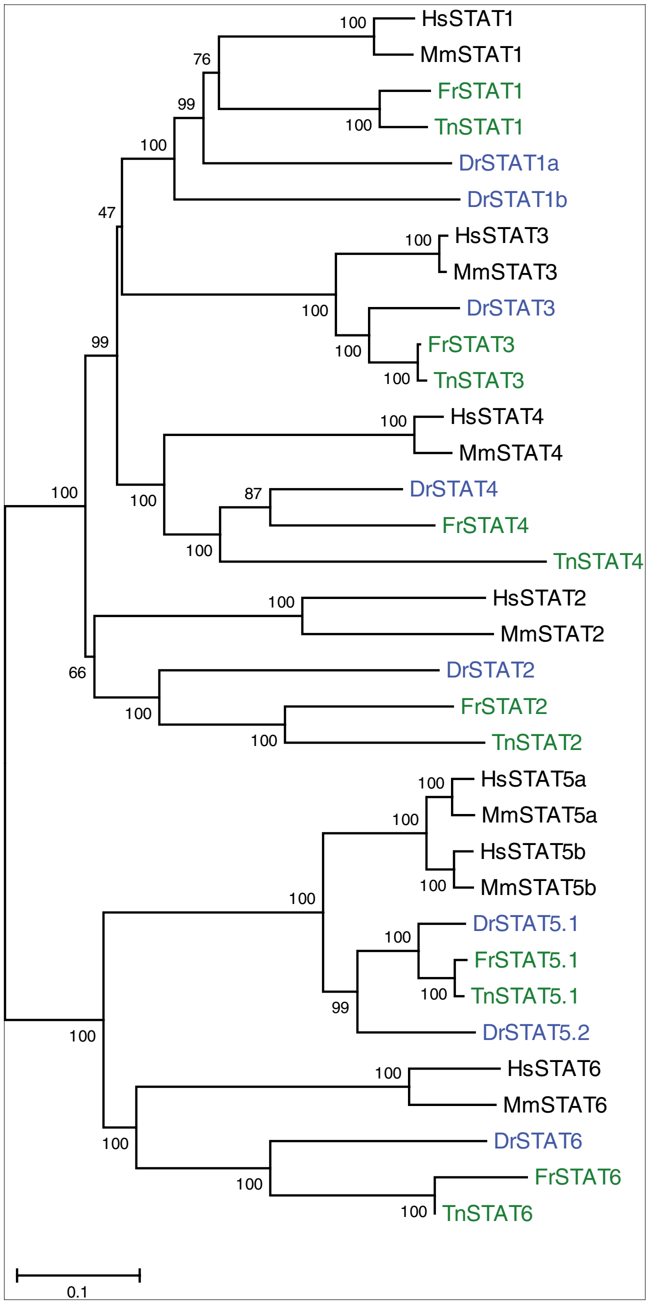

Figure 5

Phylogenetic tree of the STAT proteins. Details are as in Figure 2. STAT, signal transducer and activator of transcription.

(it is also absent in *Medaka* and stickleback), is less likely because a search of the ray and shark genomes did not identify any sequences for IRAK2. Conversely, we find duplications in the fish lineage for Jak2 (Janus kinase 2) and NLK (nuclear factor-κB [NF-κB] essential modulator-like kinase), and duplications in both pufferfish for IKKa (inhibitor of NF-κB kinase) and Ripk5 (receptor-interacting protein kinase 5).

Adaptors

The adaptors that are involved in innate immune signaling cascades are well conserved in fish, as was previously observed for those interacting with the TLRs [6,7,23]. We find orthologous genes in each of the three fish species for Myd88 (myeloid differentiation factor 88), Sarm1 (sterile α and HEAT/armadillo motif containing protein 1), Tollip, IKAP (IKK complex associated protein), NEMO (NF-κB essential modulator), Tab1, Tab2, and Tab3, and in the zebrafish and *Takifugu* for Tirap (Toll/IL-1 receptor associated protein). For the mammalian Ticam (Toll-like receptor adaptor molecule)1 and Ticam2 (also named TRIF and TRAM) genes, there is only one homologous gene in each of the three fish, which is equally distant to Ticam1 and Ticam2, indicating a duplication of an ancestral gene in the mammalian lineage and subsequent divergence of the two copies (Figure 2). The alternative interpretation, that Ticam2 was lost specifically in the teleost lineage, does not fit with the fact that it is also not present in the genomes of *Xenopus* and chicken [28]. An apparent contradiction to our observation is a report of both Ticam1 and Ticam2 in *Hydra* [29]. However, cnidarians too have only one Ticam, because the gene cited as Tram is in fact not the TRAM (TRIF-related adaptor molecule) that is synonymous with Ticam2, but encodes an unrelated protein, the translocation-associated membrane protein, which has the same acronym.

IFN response factors

For IRF1, IRF3, and IRF5 to IRF9, clear orthologous relationships are found between mammals and fish. In each fish species we also find an additional gene, which we call IRF11 and which is equally distant to both IRF1 and IRF2. DrIRF4b, which is most closely related to the IRF4s found in the pufferfish, maps to a region of the genome that is syntenic with the region containing IRF4 in mammals and in the two pufferfish, indicating that these are orthologous genes. In addition to the homologs of the IRFs in mammals, we find an additional IRF in each of the fish, which we named IRF10, because it groups with a similar gene from chicken. It appears that this gene has been lost in mammals (Figure 4).

Signal transducers and activators of transcription

Mammalian Stat2, Stat3, Stat4, and Stat6 have clear orthologs in all three fish species (Figure 5). Stat5 has been independently duplicated in mammals and in zebrafish [21]. The group of Stat1 genes contains one gene from each pufferfish with a good match to mammalian Stat1, but two genes from zebrafish that are surprisingly divergent but still resemble Stat1 more than the other Stats. The duplication event that led to this situation is recognizable in the genome, because the whole region containing the gene is duplicated and syntenic with the same region in human (Figure 7). The positions of flanking genes in the human and zebrafish genome are indications of a number of rearrangements. On chromosome 9 in the zebrafish these have been associated with a further

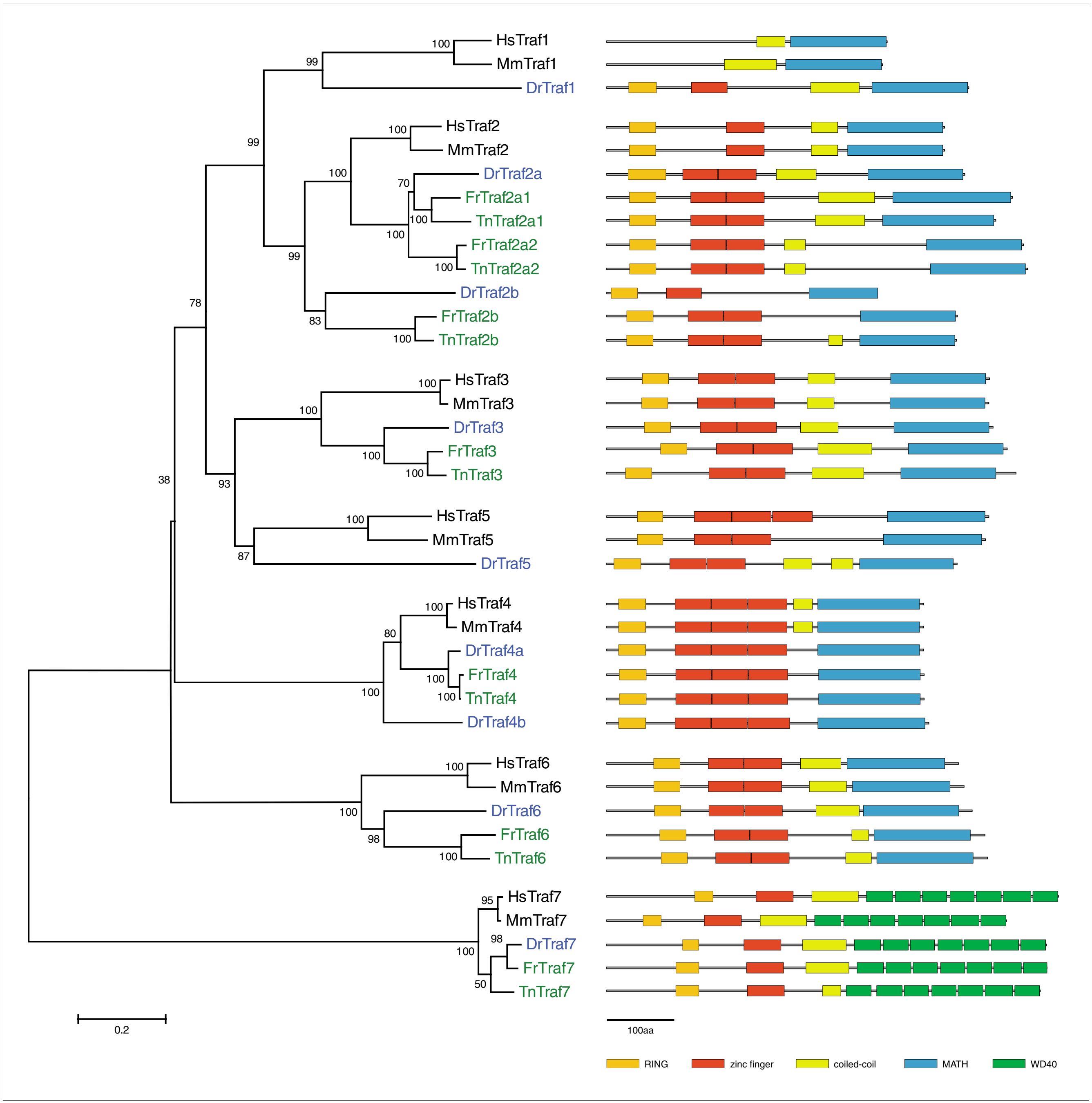

Figure 6

Phylogenetic tree of the TRAFs and diagrams of their protein domain structure. Details are as in Figure 2, except that the scale shows 0.2 amino acid substitutions. TRAF, tumor necrosis factor receptor-associated factor.

duplication of part of the Stat1 gene, resulting in a pseudogene (ENSDARG00000040710; STAT1Ψ in Figure 7).

TNF-receptor associated factors

All of the Traf protein family members Traf1 to Traf7 are represented in fish (Figure 6). For Traf3, Traf6, and Traf7 we find one gene in each of the three fish species, in all cases with the same protein structure and a high degree of similarity. Traf1 and Traf5 are present in zebrafish, but no predictions exist for these genes in the pufferfish genomes. It is interesting that zebrafish Traf1 differs from mammalian Traf1 in that, like the other family members, it contains a Ring finger and a zinc finger (Figure 6), indicating that the absence of these domains in mammalian Traf1 is due to a loss that occurred specifically in

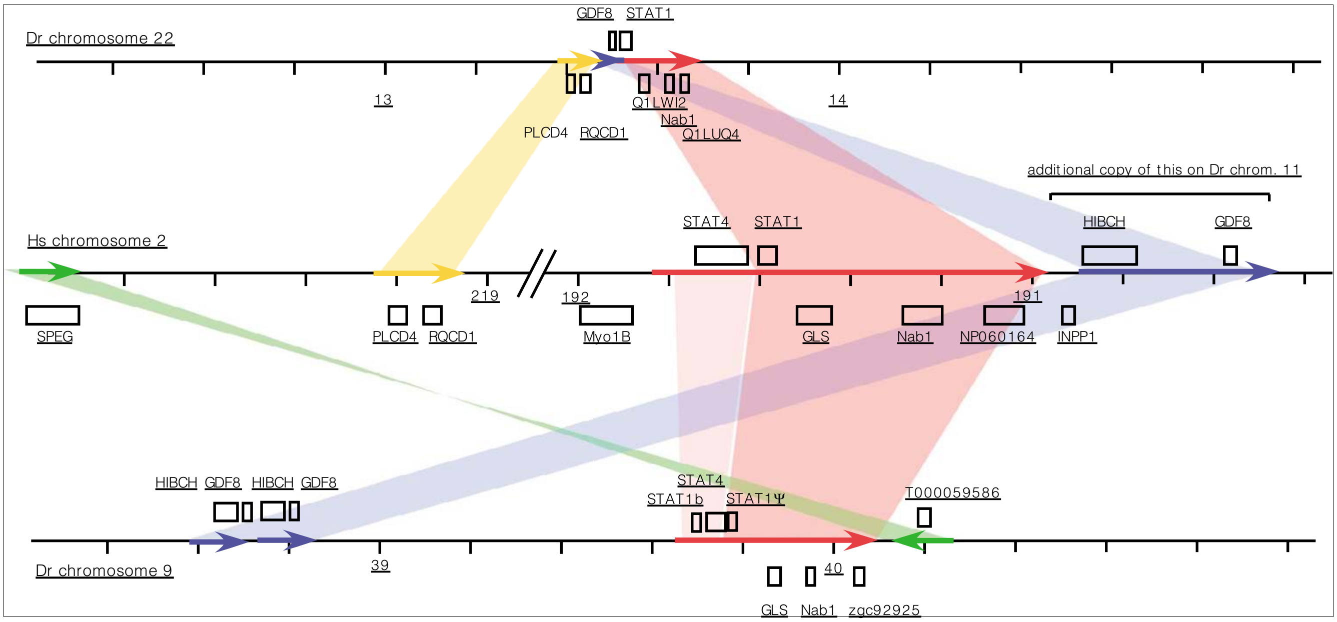

Figure 7

Synteny between regions containing STAT4 and STAT1 genes on human chromosome 2 and *Danio rerio* chromosomes 22 and 9. Genes transcribed on the top or bottom strands are shown above and below the lines representing the chromosomes. Homologous regions are shown by colored arrows. A further duplication of the region containing HIBCH and GDF8 is found on *Danio rerio* chromosome 11. Numbers represent nucleotide positions in the genome in megabases based on the Zv6 assembly. Gene names are Swissprot, Zebrafish Information Network, or Ensembl identifiers.

the mammalian lineage. Traf4 is duplicated only in zebrafish [30], whereas there have been several duplication events in the fish lineage for Traf2.

In summary, for the families described thus far, clear orthologies exist between the teleost and mammalian lineages, with a few duplications for some of the gene family members.

Class II cytokines and their receptors

Class II cytokine receptors

Mammals have two distinct, heterodimeric receptors for type I and type II IFNs, as well as a set of closely related receptors for other class II helical cytokines. Although a large group of this type is found in fish, there are no simple orthologies between the receptors of this class in mammals and teleost fish [16,17,31]. A previous analysis identified 11 genes in *Tetraodon*, named cytokine receptor family B (CRFB)1 to CRFB11 [17]. The authors found that the genomic region containing IFN-α receptor (IFNAR)chain 2, IL-10 receptor (IL10R)chain 2, IFNAR1, and IFN-γ receptor (IFNGR)chain 2 in mammals is syntenic with a region containing six class II cytokine receptor genes in *Tetraodon* [17] (see Figure 8). However, sequence comparison allowed no clear assignment of the fish genes to their mammalian counterparts, with the exception of the genes encoding tissue factor (TF), which is duplicated in *Tetraodon* (TF1 and TF2). A subsequent study [31], which included all available sequences throughout the animal kingdom, came to a slightly different conclusion regarding the phylogenetic relationships. In this study the authors subdivided the genes into groups encoding ligand-binding and non-ligand-binding chains before conducting their phylogenetic analysis. However, the justification for the assignment of particular fish genes that have no clear orthologs in mammals to one or other group is not obvious, especially because no sequence data were given in this study that unambiguously identify the genes analyzed. We therefore revisited the phylogeny of class II cytokine receptors in teleosts and mammals.

The family is defined by the presence of the D200 domain, which consists of two immunoglobulin domain-like subdomains of the fibronectin type III class, SD100A and SD100B. As has previously been pointed out [17], the bioinformatic identification of class II cytokine receptor genes is not trivial, and it is therefore unsurprising that Ensembl [24] contained predictions for only ten such genes in zebrafish. Three of these do not encode class II cytokine receptors but for thrombopoietin and titins, which have similar domains. To identify further receptor genes we searched the zebrafish genome and all available zebrafish ESTs for the subdomains SD100A and SD100B (see Materials and methods, below).

We identified 22 candidates, of which seven had incomplete D200 domains or exhibited only spurious resemblance to D200 domains. These and the three genes encoding the D200-containing proteins thrombopoietin and titin were eliminated from further analysis. Gene predictions were available for eight of the remaining 12 genes. Of the four genes

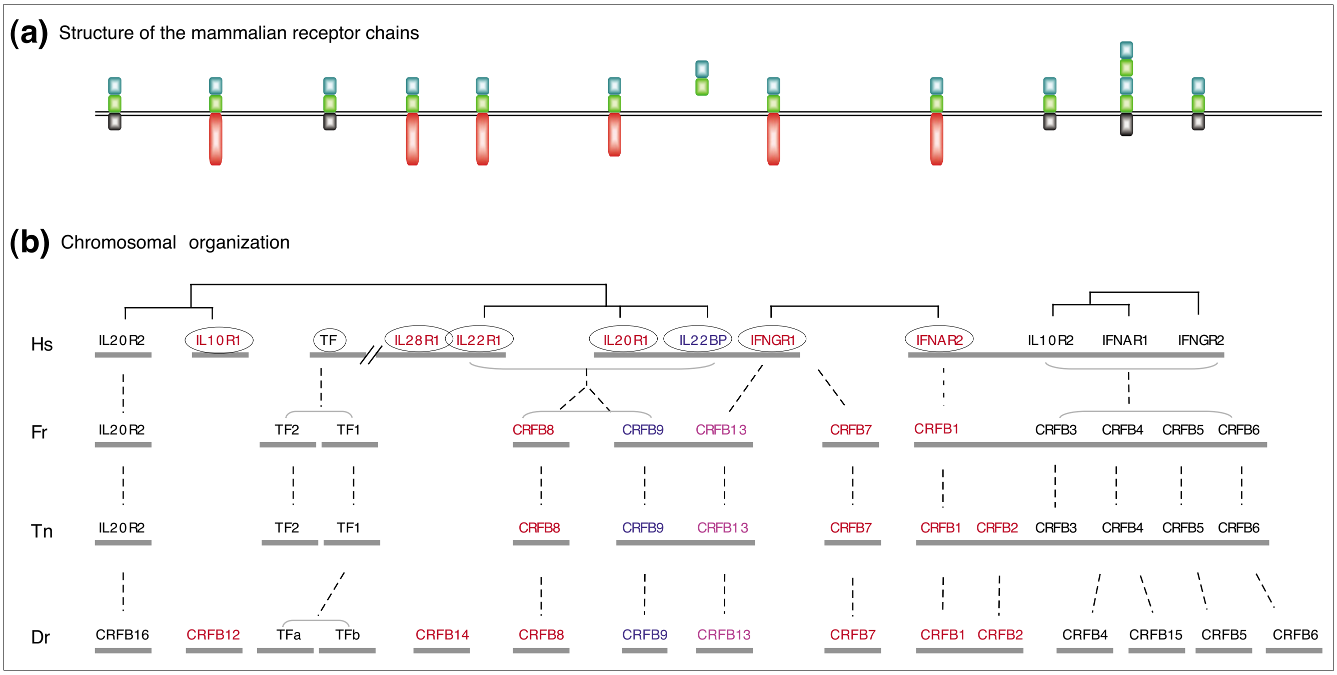

Figure 8

Syntenic organization of classII cytokine receptor genes. **(a)** Diagram of the structures of the mammalian receptor chains, with the blue and green rectangles representing the S100A and S100B domains, the red rectangle the intracellular domain of the ligand binding chains, and the gray rectangle the intracellular domain of the non-ligand-binding chains and TF (after Renauld [57]). **(b)** Synteny between regions containing class II cytokine receptor genes in mammals and fish. Fat horizontal lines indicate chromosomes in the four species. The brackets above the human genes show evolutionary relationships between the paralogs. Vertical broken lines indicate suggested evolutionary relationships between the genes in the different species, based on the tree in Figure 9. Color coding of names: red = long intracellular domain; black = short intracellular domain; blue = no intracellular domain; and pink = intermediate length intracellular domain. Circled names indicate ligand binding chains. Round brackets denote groups of genes in cases where there are no clear orthologous relationships of individual members with genes in the other species.

that had not been predicted by automated annotation tools, two (CRFB15 and CRFB16) were found only in the as yet unplaced whole genome shotgun sequences. We re-annotated all 12 genes using the known gene structure of class II cytokine receptor genes and homology to known class II receptor genes as support. We used these sequences for a phylogenetic analysis, which, in addition to the mouse and human sequences, also included *Takifugu rubripes* and *Tetraodon nigroviridis* CRFB1 to CRFB11 and IL20R2, as well as an additional gene, the product of which we shall call CRFB13 (Ensembl: NEWSINFRUG00000164405 and GSTENG0003154300). A set of recently described zebrafish class II cytokine receptor genes included two genes not identified by us (DrCRFB2 and DrCRFB6), which we have added to our analysis [18]. Finally, DrCRFB14 was found by Georges Lutfalla, who generously contributed its sequence for inclusion in this analysis.

The phylogram of the class II cytokine receptors (Figure 9) corroborates previous conclusions that this gene family has undergone independent gene duplications and divergence in teleost fish and mammals. Some of the fish genes cannot be matched to likely orthologs in mammals, and *vice versa*, with four exceptions in which high bootstrap values justify the interpretation of the genes sharing direct common ancestors. The genes encoding TF in mammals cluster with two genes from each fish. The phylogeny indicates independent duplication events in the pufferfish and zebrafish lineage. The other set of genes that reliably group together are those encoding IL20R1, IL20R2, and IL-22 binding protein (IL22BP), with one representative in each of the mammals and fish.

For the other relationships between mammalian and fish genes the bootstrap values are so low that the relationships discussed below must be considered with caution. Several mammalian genes have no plausible orthologs in the three fish genomes analyzed here, and others have more than one.

We therefore sought further evidence for evolutionary relationships by analyzing the genomic context of the genes. A summary is shown in Figure 8. Two sets of genes are linked both in mammals and in the two pufferfish. The first is the IFNAR2, IL10R2, IFNAR1, and IFNGR2 complex and its syntenic complex described by Lutfalla and colleagues [17] for *Tetraodon*. This synteny is also maintained in *Takifugu* and in all three cases continues outside the class II cytokine recep-

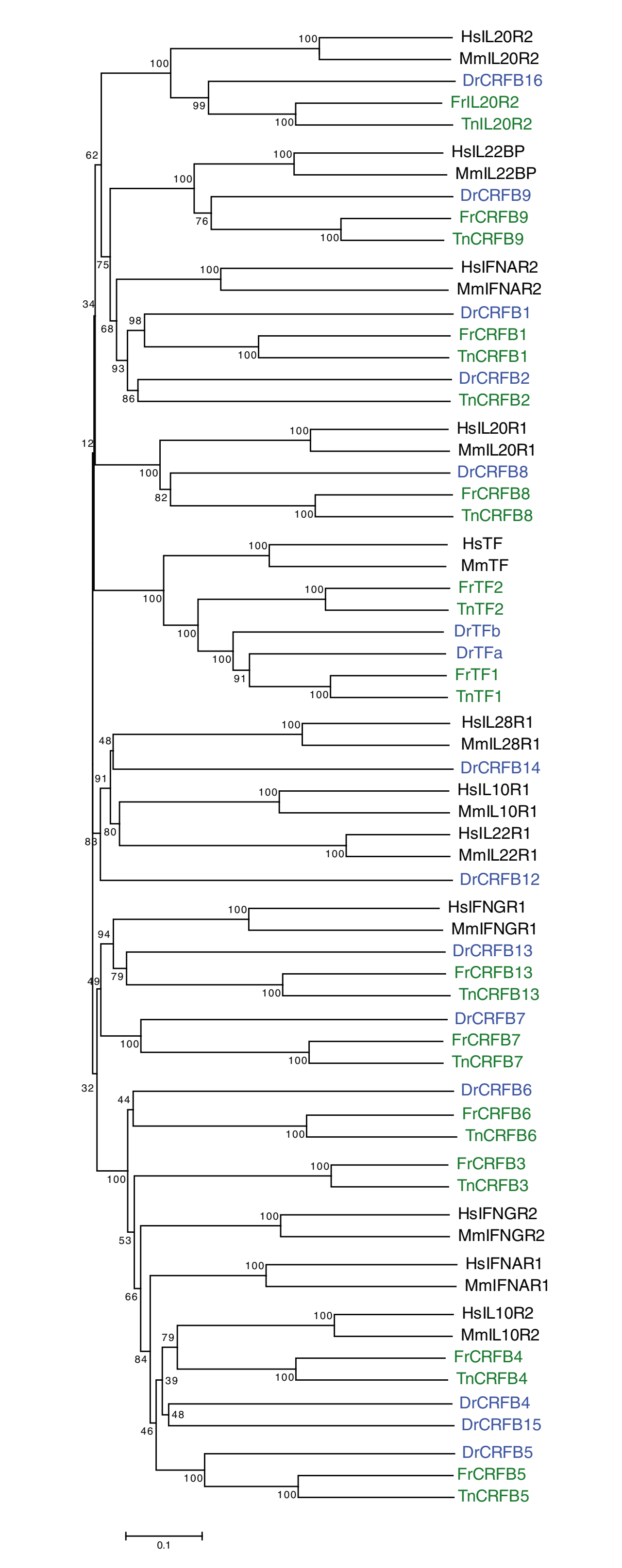

Figure 9
Phylogenetic tree of the class II cytokine receptors. Details are as in figure 2.

tor complex, in that the gene neighboring IFNGR2 is Tm50b in all cases, followed by Nnp1. However, the corresponding genes in the zebrafish are no longer linked (although they all lie on the same chromosome).

The synteny is roughly reflected in the sequence similarities, in that IFNAR2 is most similar to CRFB1 and CRFB2 and that the IL10R2/IFNAR1/IFNGR2 group clusters with the CRFB3/4/5/6/15 group. In particular, the IL10R2/IFNAR1/IFNGR2 and CRFB3/4/5/6/15 genes encode receptors with short cytoplasmic domains, whereas IFNAR2 and CRFB1 and CRFB2 have long cytoplasmic tails. However, within the group orthologies are not clear. It is therefore not possible to conclude whether the ancestral complex that existed before the split of the teleosts and tetrapods contained two genes (a precursor for IFNAR2 and a precursor of the IL10R2/IFNAR1/IFNGR2 group) with subsequent independent duplications in teleosts and mammals, or four genes, with fast divergence in the IL10R2/IFNAR1/IFNGR2 and the CRFB3/4/5/6/15 groups obscuring their common origin.

The second region in which a syntenic arrangement of genes is retained is the one containing IFNGR1, IL20R1 and IL22BP in mammals, and CRFB9 and the previously undetected CRFB13 in *Tetraodon* and *Takifugu*. Again, the closest relatives of these genes (CRFB9 and CRFB13, respectively) are not syntenic in zebrafish. Notably, fish CRFB9 proteins share the absence of a transmembrane domain with the mammalian IL22BPs. In view of this and the syntenic arrangement, the most reasonable interpretation is a homology of IFNGR1/CRFB13 and IL22BP/CRFB9.

In summary, teleost fish have approximately the same number of class II cytokine receptors as mammals, but the genes have evolved rapidly and independently since the separation of the species. We shall leave the discussion at this point, because the current set of data does not support further speculation. A statement about which of these receptors are functionally equivalent will have to await experimental analysis, as has been conducted for two of the zebrafish CRFBs [18]. It will be interesting to determine whether fish distinguish between viral and bacterial induced IFN signaling pathways in the same way as mammals.

Class II cytokines

IFNs have been reported in several fish species with an ambiguous nomenclature [32-40]. We find ten class II cytokine genes in zebrafish, and five in each pufferfish (Figure 10). The large group of mammalian type I IFNs cluster together on one branch of the phylogenetic tree that does not include any fish cytokines. This fits with the view that the generally intronless type I IFN genes are the product of a retrotransposition event [17], which occurred after the split of teleosts and tetrapods. Apart from the clear fish orthologs of the mammalian type II IFNs, the remaining fish class II cytokines are more similar to the mammalian ILs and type III

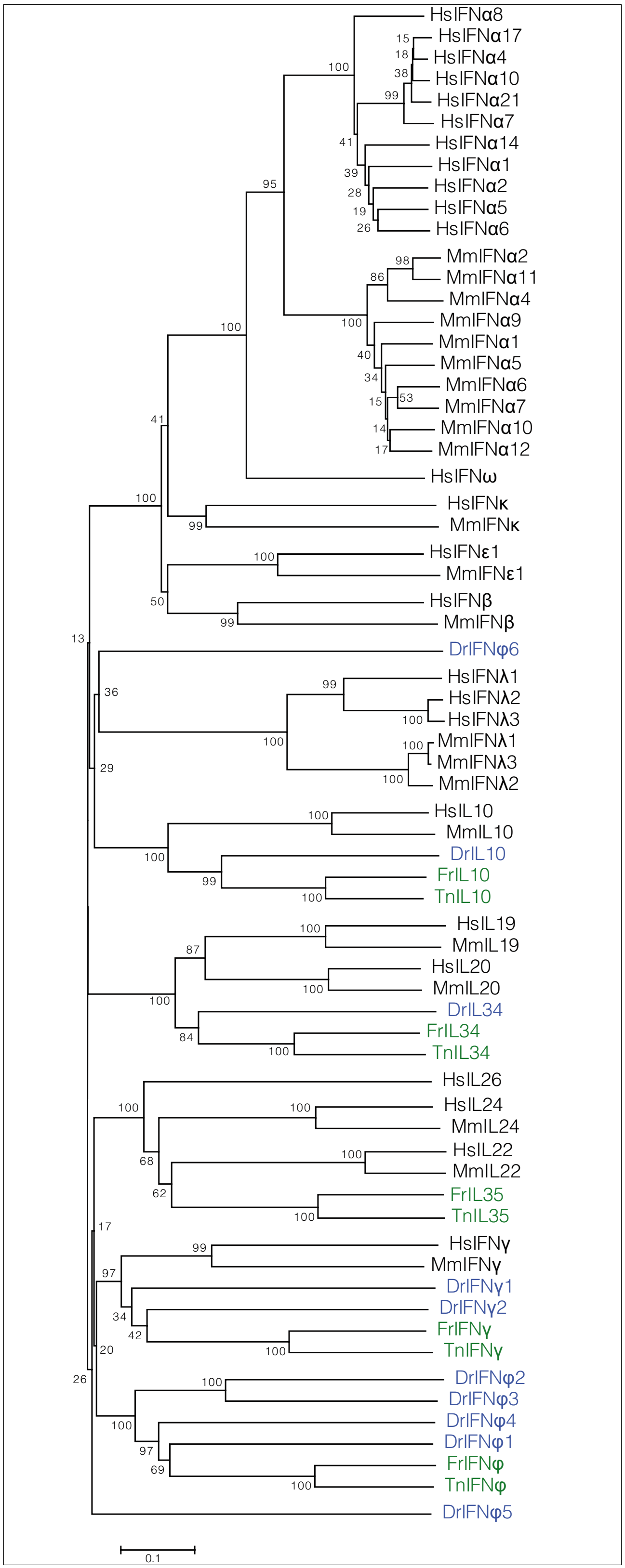

Figure 10

Phylogenetic tree for the classII cytokines. Details are as in Figure 2. See text for gene names.

IFNs. Like these, they are mostly encoded by genes with four phase 0 introns, supporting the view that this constitutes the gene structure of the ancestral class II cytokine gene.

Among these class II cytokines, IL-10 exhibits an apparent orthology between fish and mammals [41,42]. This is also supported by the genomic locations of the IL-10 genes, which are situated adjacent to and on the opposite strand of the Mapkap2 genes in all five species (Figure 11). The genes that had been annotated as IL-20 in the zebrafish (Refseq: NP_001076424.1) and *Tetraodon* (Uniprot: Q7SX60), and initially as IL-19 and then changed to IL-24 in *Takifugu* (Ensembl: SINFRUG00000154816) are equally related to mammalian IL-19 and IL-20. The previous automated naming of the fish genes should therefore be amended. In concordance with the nomenclature rules for vertebrate gene families, this gene has therefore been given the next available number in the IL series (IL-34). The fish IL-34 genes and the mammalian IL-19, IL-20, and IL-24 genes are located in the vicinity of the IL-10 genes (in the zebrafish this gene has not yet been placed on a chromosome), but duplications and inversions have broken up the syntenic relationships downstream of IL-10. The phylogenetic tree argues for a common precursor for these genes that has duplicated in mammals, yielding IL-19 and IL-20. Whether IL-24 is the product of a second local duplication or of an older duplication of a larger segment of the genome is not clear, but it shows a higher degree of similarity to the class II cytokine genes found in a complex on a different chromosome in all five species (Figure 11).

A second group of class II cytokines exhibiting high sequence similarity are the mammalian IL-22, IL-24 and IL-26, and two pufferfish interleukins annotated as 'IL-24' in *Tetraodon* (Uniprot: Q7SX82) and 'homologous to IL-24' in *Takifugu* (Ensembl: SINFRUG00000156387). Again, the phylogram shows that this name is problematic, because if anything these proteins are more similar to IL-22, and their genes exhibit the same syntenic relation to the flanking MDM1 gene as the IL-22 genes do in mammals (Figure 11). However, the zebrafish gene in the same position (RefSeq: NP_001018628), annotated as IL-22 [33], is highly divergent in sequence. Because frequent duplications and loss of genes as well as rapid sequence divergence appear to operate within this family, originally orthologous genes may no longer be recognizable. This is further illustrated by the flanking IL-26 gene in the human genome. The mouse genome has lost this gene; in the zebrafish a class II cytokine gene described as IL-26 [33] is present in this position, but it does not cluster with the IL-22/24/26 group. Although the IL genes between MDM1 and IFN-γ are in apparently orthologous positions in all five species, there is no indication that the mammalian arrangement MDM1/IL-22/IL-26/IFN-γ represents the ancestral cluster, rather than the IL genes having arisen by independent duplications in mammals and teleosts. Because the names given to the fish cytokines of this group are

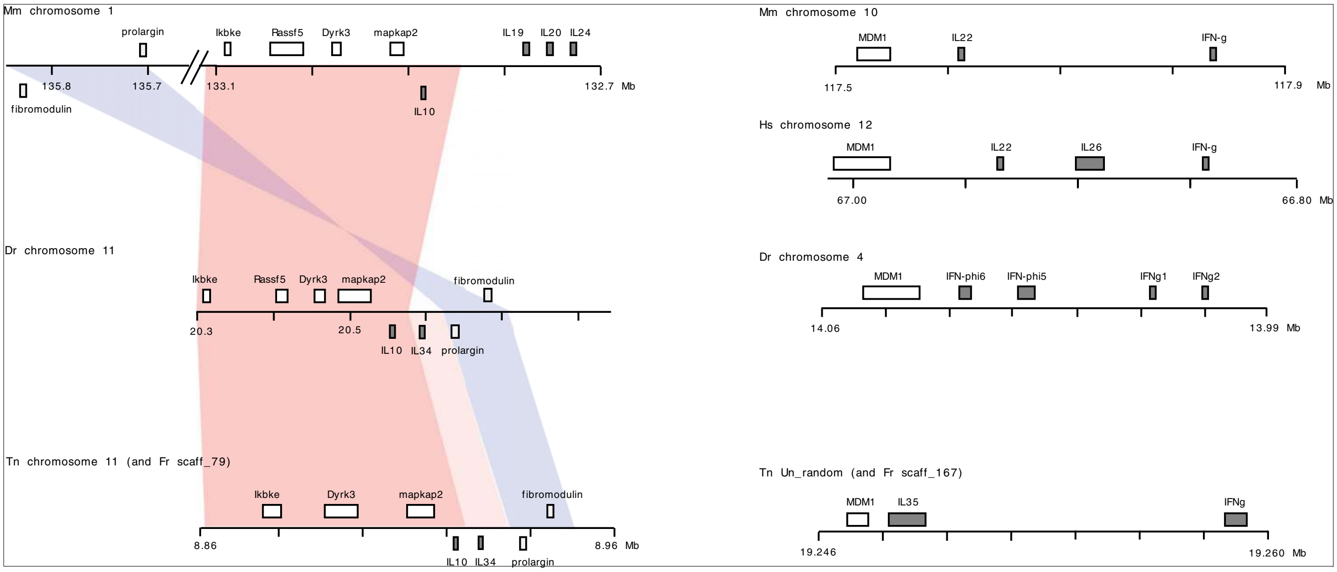

Figure 11

Genomic organization of two class II cytokine gene clusters. Chromosomes are shown as lines with the positions of the region marked in megabase pairs underneath. Genes transcribed on the top strand are shown above the line, and those transcribed in the opposite direction are shown below. Class II cytokine encoding genes are shaded in gray. In the left diagram the syntenic regions and duplications, and inversions surrounding the IL-10 locus are shaded in red and blue. The human IL-10 gene is located on chromosome 1 and the region shows the same arrangement as in the mouse. The current zebrafish genome assembly Zv7 does not yet contain the recently sequenced clone CU459075, which places IL-34 into the interval between IL-10 and prolargin (IL-34 is included in Zv7 on the unplaced contig Zv7_NA1656). There are therefore no coordinates for the right end of the interval. The two pufferfish show the same arrangement both for the region around IL-10 and for the MDM1/cytokine/IFN-γ region. The names for the fish genes are explained in the text. IFN, interferon; IL, interleukin.

extremely confusing and suggest relationships for which there is no evidence, we again propose a new nomenclature, as shown in Figures 10 and 11 (IFN-φ6 for zebrafish IL-22, IFN-φ5 for zebrafish IL-26, and IL-35 for the pufferfish IL-24).

Four of the remaining fish class II cytokine genes cluster with the mammalian INF-γ genes and the rest do not group with any of the mammalian genes. The pufferfish each have one IFN-γ gene, whereas the zebrafish has two, namely IFN-γ1 and IFN-γ2 [33,34], which lie in tandem in a position in the genome that has retained its synteny between mammals and teleosts (Figure 11).

Finally, a group of teleost class II cytokines, some of which had previously been called IFN-λ, cluster on a branch without mammalian cytokines. Because they are not more related to mammalian IFN-λ than to other cytokines, we call them IFN-φ1 to IFN-φ4. IFN-φ1 has previously been described as 'zebrafish interferon', 'IFNab', and 'IFN-λ' [17,18,32], and IFN-φ2 and IFN-φ3 as 'type I IFN 2' and 'type I IFN 3' [43]. Only one gene of this type, most closely related to the zebrafish IFN-φ1 gene, is found in the two pufferfish. This may be due to the difficulty in identifying these genes, and it would not be surprising if further class II cytokine genes were found in the pufferfish genomes.

In summary, like the receptors, the class II cytokine genes have duplicated and diverged independently in fish and mammals. It remains to be tested experimentally which class II cytokines are responsible for which immune function.

Intracellular pathogen sensors: the NACHT-domain family

A large family of cytoplasmic proteins, characterized by the presence of a nucleotide-binding domain, the NACHT domain [44,45] or the closely related NB-ARC domain [46], has been implicated in inflammation and innate immune signaling in animals and plants. Some of them have been shown to recognize intracellular pathogen-associated molecular patterns through their carboxyl-terminal leucine-rich repeats (LRRs). They differ in their amino-terminal effector domains (for example, CARD or pyrin domains), which mediate signal transduction to downstream targets, leading to the activation of NF-κB or the apoptotic pathway.

An initial search in the fish genomes for homologs of the known mammalian NLR proteins of the Nod subfamily found homologs for Nod3 and Nod9 in all three fish species: Nod2 in zebrafish and *Takifugu*, and Nod1 in *Takifugu*. Three genes in zebrafish, two in *Takifugu*, and one in *Tetraodon* were annotated as 'Nalps' (NACHT, leucine rich repeat and PYD containing proteins) but did not group with the mamma-

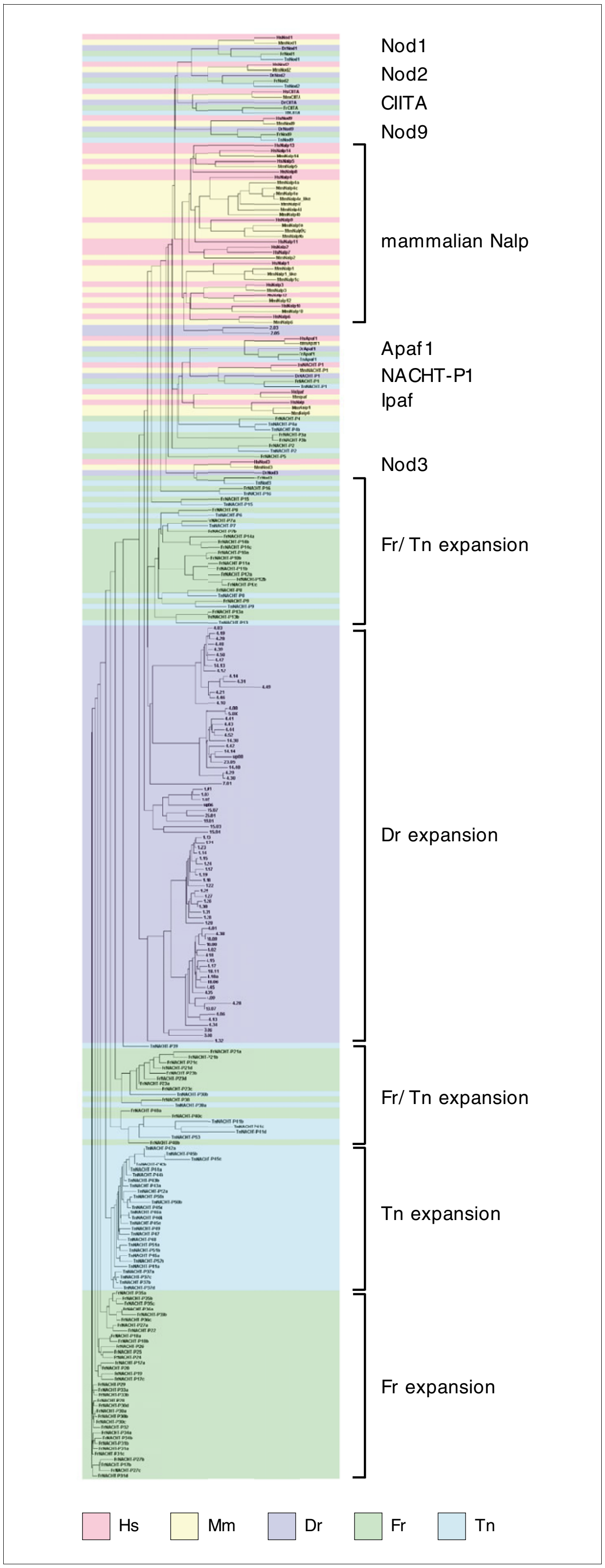

Figure 12

Figure 12
Overview of a phylogenetic tree of 277 NLR proteins. Each sequence is assigned a background color to illustrate species relationships: pink = human, yellow = mouse, blue = zebrafish, green = *Takifugu*, and turquoise = *Tetraodon*. The 'canonical' proteins Nod1, Nod2, Nod3, Nod9, CIITA, and Apaf, which show clear homologous relationships between the five species, cluster at the top (rainbow colors). The mammalian Nalp proteins cluster together (pink/yellow region). Each fish has a large group of species-specific proteins (blue, green, and turquoise regions). In addition, *Takifugu* and *Tetraodon* share several apparently orthologous gene pairs (green and turquoise region). Apaf, apoptotic protease activating factor; CIITA, major histocompatibility complex class II, transactivator; Nalp, NACHT, leucine rich repeat and PYD containing protein; NLR, nucleotide-binding domain/NACHT domain and leucine rich repeat containing family; Nod, nucleotide oligomerization domain containing protein.

lian Nalps on a phylogenetic tree. We found no homologs for any of the mammalian Nalps in fish. We therefore screened the whole zebrafish genome for sequences encoding NACHT-domains. This revealed a large number of additional sequences encoding NACHT domains. Most of these were not within genes found by the automated gene prediction algorithms, because the number of and similarity between the genes was so high that they had been masked as repeats. We therefore annotated these genes manually using ESTs as guides and identified a large set of novel NACHT-domain containing genes. After we had completed our initial annotations, automated predictions for 205 NACHT-domain encoding genes were deposited at the National Center for Biotechnology Information (NCBI). These showed only a partial overlap with our sequences. Many were incomplete or contained two NACHT domains, indicating incorrect annotations. We therefore re-screened and re-annotated the zebrafish genome and have found more than 200 genes of this class (the complete list is given in Additional data file 9). These are numbered sequentially by chromosome number and by their order on the chromosome. We have not been able to produce perfect gene models for all of them. As discussed below, they have novel amino-terminal sequences, and in the absence of sufficient EST evidence we were unable in all cases to draw reliable conclusions regarding the 5' end of the gene. Similarly, the LRRs in the carboxyl-terminal region are difficult to predict reliably. Extensive experimental work will be needed to characterize these genes. For our analysis here we have selected a set of 70 representative sequences.

We also searched the two pufferfish genomes for members of this gene family to find out whether the group we found in zebrafish was specific to this species, or whether the massive gene duplication had occurred early in the fish lineage. We found 70 members of this family among the annotated genes in the genome of *Takifugu rubripes*. A large number of matches found in the *Tetraodon* genome were not parts of predicted or annotated genes, as had been the case in the zebrafish. Again, these sequences had been masked as repeats. We manually assembled a set of sequences using homology to the zebrafish and *Takifugu* sequences as guides. It is striking that the majority of the members of this gene family (40/49) are located within incompletely assembled contigs/scaffolds that have not been assigned to chromosomes (the 'Un_random' set). Initially, our searches for NACHT-domain encoding genes resulted in a number of predictions that spanned separate contigs, but which had additional fragments of genes of this family interspersed within their predicted introns. This suggests that these predictions were not correct, but were due to accidental occurrence of apparently spliceable gene fragments in neighboring contigs of this assembly that are in fact not located next to each other in the genome. This view is supported by the finding that three sequences, which are very closely related to consecutive parts of the other fish Nod2 genes, were positioned on widely separated contigs in the Un_random assembly. We have combined these three fragments into one sequence, which we call TnNod2. The high proportion of genes from this family in the nonassembled part of the genome might be an indication that the proper assembly of these contigs is made difficult or impossible precisely because of the repetitive nature of this family.

Phylogenetic relationships of NLR protein families in mammals and fish

A phylogenetic tree of all NLR containing predicted peptides from human, mouse, and the three fish species reveals the following relationships (Figure 13). The canonical Nod proteins Nod1, Nod2, Nod3 (recently renamed as Nlrc3) and Nod9 (recently renamed as NlrIX), as well as Apaf1 (apoptotic protease activating factor 1) and CIITA (major histocompatibility complex class II, transactivator), are present in all five species and exhibit clear orthologous relationships (Figure 14). The Nalp proteins (which have recently been renamed Nlrp) form a separate branch, representing a mammalian expansion of NLR proteins. For most of the genes on this branch, there are closely related pairs of mouse and human genes, but several cases of mouse-specific or human-specific duplications can also be found, notably the mouse Nalp4 genes. Two zebrafish sequences that cluster with this group, 2.03 and 2.05, encode only a NACHT domain with a divergent P-loop and should therefore not be considered Nalp-like proteins.

Most strikingly, the large groups of newly identified fish sequences lie on mostly species-specific branches. The majority of the zebrafish genes form a branch of their own, which includes no genes from either of the two pufferfish. Consistent with the closer relationship between the two pufferfish, the genes from these two species are less clearly separated. Whereas one branch contains exclusively a subset of genes from *Takifugu*, the branch that contains the majority of *Tetraodon* genes also includes several *Takifugu* genes. There are two branches with several cases of apparent orthologies between *Takifugu* and *Tetraodon* (genes from the two species that are more similar to each other than to any other gene in their own species), indicating the existence of these genes before the split of the two species and suggesting conservation of their function. We note again that the *Tetraodon* gene predictions are less reliable and are often incomplete, leading to spurious homology assignments. The relationship of these sequences to the other fish sequences therefore represents an approximate picture that must be interpreted with caution.

Whereas most of the novel fish NLR proteins are more related to each other than to mammalian NLR proteins, there are exceptions (apart from the canonical proteins mentioned above). One group of new fish proteins, which we named NACHT-P1, clustered with Apaf1. We wished to know whether this was a fish-specific NACHT protein and searched the mouse and human genomes for similar sequences. We found one ortholog in each case, neither of which had been characterized previously. Their amino-terminal parts contain no motifs known from other proteins. Like the Apaf proteins, these sequences contain WD40 repeats instead of LRRs.

FrNACHT-P2 and TnNACHT-P2 have an unusual amino-terminal addition, a filament domain. We found no other sequence in any organism that encodes a protein composed of a filament domain and a NACHT domain.

Fish-specific properties of novel fish NLR proteins

The large groups of novel, fish-specific NLR proteins are highly conserved in each species, indicating recent species-specific expansions (Figure 12 and additional file 10). Like other NLR proteins, they contain LRRs at the carboxyl-terminus, but the majority does not contain any of the amino-terminal effector domains that have been found in conjunction with NACHT-domains in mammals or plants (such as CARD, pyrin or TIR domains). However, the region immediately upstream of the NACHT domain is highly conserved in all of the fish proteins (Figure 14).

To find out whether this region corresponded to other known peptide motifs, we used a hidden Markov model built from the zebrafish sequences for a BLAST search of the mammalian genomes. No good matches were found. We then searched the three fish genomes. In the zebrafish and in *Takifugu* we found only those genes we had already identified via their NACHT domains. In the *Tetraodon* genome many but not all of the matches we found were upstream of NACHT domains or were part of our previous gene predictions. As the remaining ones were again located mainly in the Un-random set, we did not attempt to link them to the predictions for the NACHT domains, for the reasons discussed above. As in the other two fish genomes, none of the matches were within gene predictions for other (non-NACHT-domain) genes. This indicates that this domain, which we will call the Fisna (fish-specific NACHT associated) domain, has been recruited specifically by a common ancestor of the novel NLR proteins in the fish lineage. Confirming this view, a cursory search of other fish genomes showed highly similar sequences in cat-

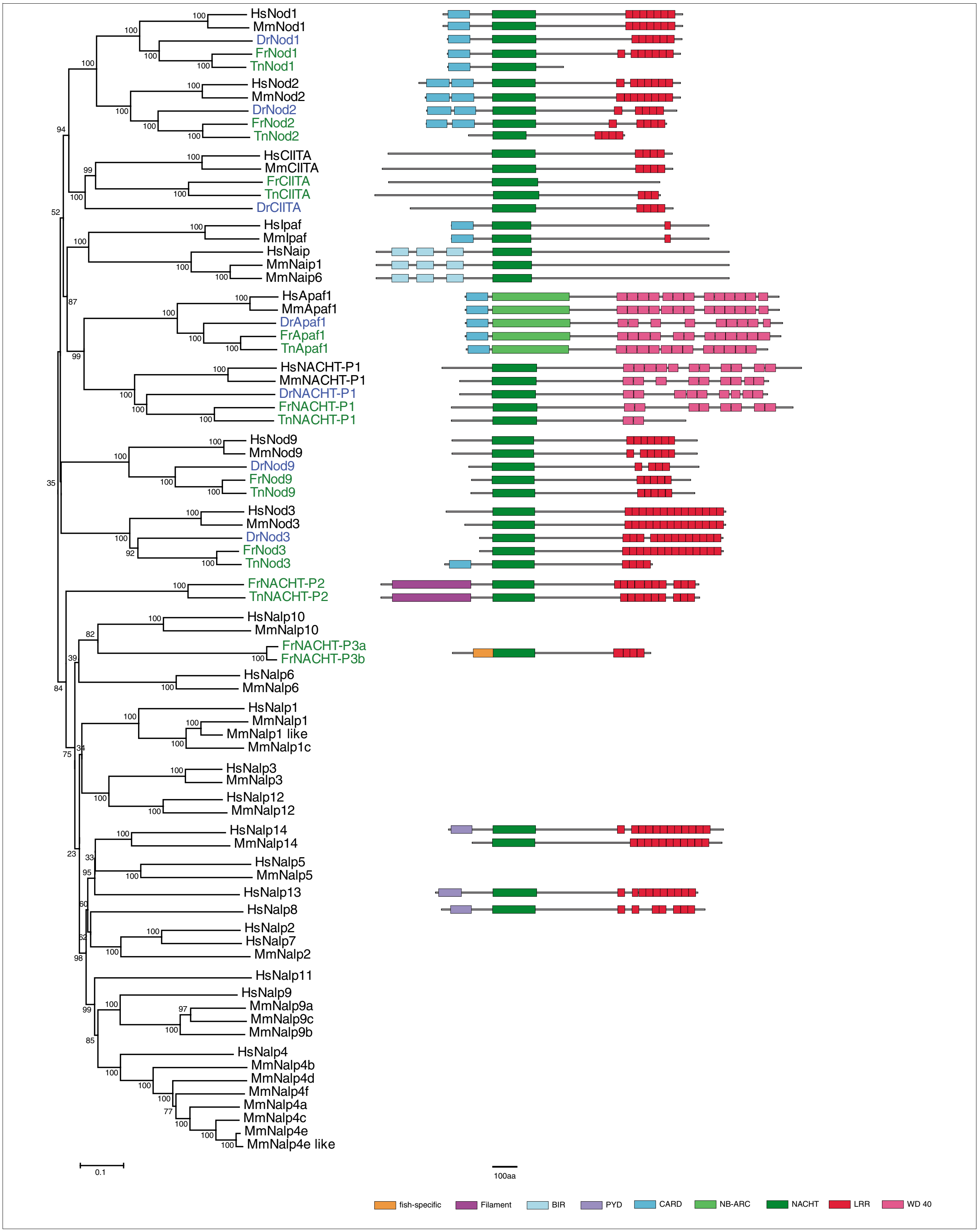

Figure 13 *(see legend on next page)*

Figure 13 *(see previous page)*
Phylogenetic tree of NACHT proteins shared by mammals and fish and diagram of their protein structures. In addition to the known proteins Nod1, Nod2, Nod3, Nod9, CIITA, and Apaf, this tree shows that a new protein is shared by all five species, which we have named NACHT-P1. The protein domain structure diagram shown next to NACHT-P3 is representative of the majority of the novel fish proteins. Apaf, apoptotic protease activating factor; CIITA, major histocompatibility complex class II, transactivator; Nod, nucleotide oligomerization domain containing protein.

fish and *Medaka*, also associated with NACHT-domain encoding genes, which we did not follow up further.

Although, as mentioned above, there is no evidence for the presence of this domain other than in fish, we noticed that a short peptide motif within this domain (LK/E/NQ/K/RYITE/D) is also found in mammalian Nod2 (LEDYITE), and another (LYIIEGESEGVNEEHEVLQ) just downstream of the first, in Nod3 (LLLVD/EGLSDLQQK/REHDLM/V/TQ). The region containing these sequences in Nod2 and Nod3 is neither part of the NACHT nor of the CARD domain and has not been assigned a cell biologic function. Their conservation in the new NLR gene families might indicate a shared origin and possibly shared functions.

A similar expansion of NLR-encoding genes was recently described in the sea urchin [47,48]. We compared the predicted sea urchin protein sequences with our sequences. In addition to sharing high similarity with the fish proteins in the NACHT domain and the LRRs, the sea urchin proteins also have a region upstream of the NACHT domain that is highly conserved among the sea urchin set of proteins, and includes sequence motifs similar to those in the fish proteins and in mammalian Nod2.

Peptide motifs in the amino-terminal part of the zebrafish NLR proteins

Further study of the amino-terminal regions of the new zebrafish NLR proteins showed that many of them contained considerable stretches of predicted peptide sequences upstream of the conserved fish domain, in some cases with multiple, related sequence repeats. Manual editing of the automated alignment created by ClustalW [49] revealed the following structure of the amino-terminal regions of this protein family (Figure 15).

Based on sequence similarity in the NACHT-domain, which is equally recognizable in the Fisna domain, the protein family can be subdivided into four groups (Figure 14). Each of these groups has further shared motifs upstream of the Fisna-domain (Figure 15). The amino-terminal sequences in group 1 are highly conserved and not found in any of the other families (darker green shading in Figure 15). A comparison with mammalian proteins showed that it has significant similarity with the pyrin-domain found in mammalian Nalp and MEFV (mediterranean fever)proteins. Group 2 has a 101 amino acid stretch upstream of the Fisna domain that is shared by all members of this group (lighter green shading in Figure 15). It shows a distant resemblance to the pyrin domain of group 1. The most amino-terminal sequences in this group contain motifs shared with members from groups 3 and 4. A motif shared by members from these three groups is a repeat (different hues of blue shading in Figure 15 indicate different versions of the repeat), which occurs in one, two, or three copies per protein, or in one case, in ten copies. Group 2 has a version of this repeat with a four-amino-acid insertion, which is also found in some members of group 3. These repeats are usually combined with a specific amino-terminal peptide of 14 amino acids (pink shading). Other conserved amino-terminal peptides (yellow or orange shading) are associated with a particular type of repeat. Group 4 is the least homogeneous, showing divergence both within the group and in comparison with the other groups, in the repeats as well as in the Fisna and NACHT domains. No significant homologies to the repeat sequence are found in mammals.

In summary, the amino-terminal parts of the novel NLR proteins contain up to three different motifs, two of which are found only in fish. The Fisna domain is found in all of the proteins and is located immediately upstream of the NACHT domain. It is specific for this protein family in fish. Groups 1 and 2 contain a pyrin-related domain upstream of the NACHT domain. Members of groups 2 to 4 can in addition contain one or more copies of a motif that is also specific for the novel fish NLR proteins. Members of groups 3 and 4 contain multiple variants of this motif but no pyrin-domain-like sequences.

Figure 14 *(see following page)*
The fish-specific domain upstream of the NACHT domain. **(a)** Alignment of a representative subset of the Fisna domain (the region upstream of the NACHT domain that is shared by all of the novel fish NLR proteins. The group names on the right refer to the subdivision of the *Danio rerio* groups according to similarities in the NACHT-domain and the Fisna domain (also see Figure 15) or indicate which species form the group. Peptide motifs with similarity to Nod2 and Nod3 are underlined. **(b)** Hidden Markov model (HMM) logo representing the consensus sequence of the Fisna domain in all three fish species. The logo has been generated using the software HMMER [58,59] and visualized using the HMM-Logo web server [60,61]. Peptide sequences from human Nod2 and Nod3 with similarity to short stretches of the Fisna consensus, color coded to highlight conserved residues, are listed underneath, as are stretches from the regions upstream of the NACHT domain present in 140 sea urchin NLR proteins. NLR, nucleotide-binding domain/NACHT domain and leucine rich repeat containing family; Nod, nucleotide oligomerization domain containing protein.

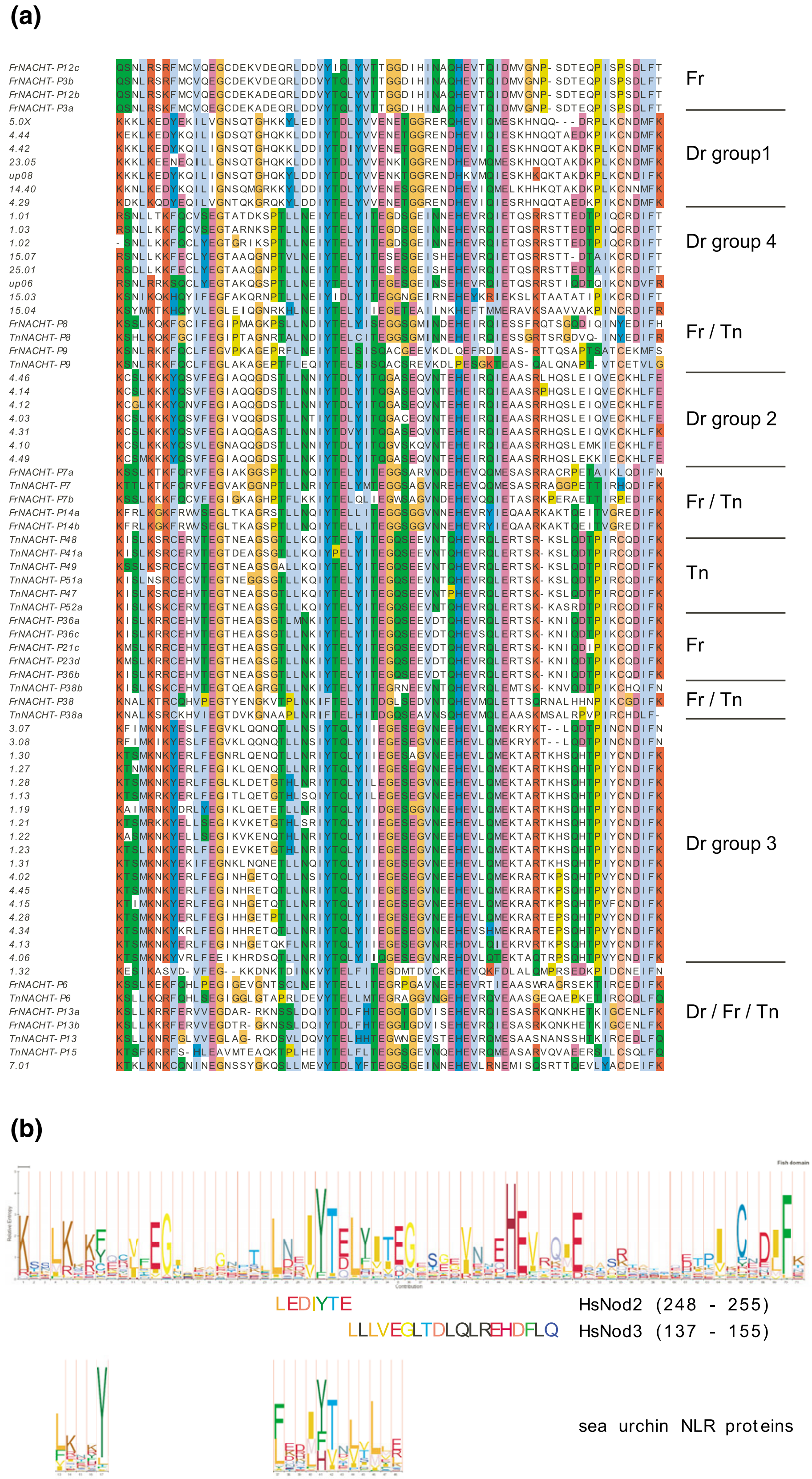
(a)
FrNACHT-P12c QSNLRSRFMCVQEGCDEKVDEQRLDDVYIQLYVTTGGDIHINAQHEVTQIDMVGNP-SDTEQPISPSDLFT
FrNACHT-P3b QSNLRSRFMCVQEGCDEKVDEQRLDDVYTQLYVTTGGDIHINAQHEVTQIDMVGNP-SDTEQPISPSDLFT
FrNACHT-P12b QSNLRSKFMCVQEGCDEKADEQRLDDVYTQLYVTTGGDIHINAQHEVTQIDMVGNP-SDTEQPISPSDLFT
FrNACHT-P3a QSNLRSKFMCVQEGCDEKADEQRLDDVYTQLYVTTGGDIHINAQHEVTQIDMVGNP-SDTEQPISPSDLFT
5.0X KKKLKEDYEKILVGNSQTGHKKYLEDIYTDLYVVENETGGRERQHEVIQMESKHNQQ---DRPLKCNDMFK
4.44 KEKLKEDYKQILIGDSQTGHQKKLDDIYTDLYVVENETGGRENDHEVIQIESKHNQQTAEDKPIKCNDMFK
4.42 KKKLKEDYKQILIGNSQTGHQKKLDDIYTDIYVVENETGGRENDHEVIQIESKHNQQTAKDKPLKCNDMFK
23.05 KKKLKEENEQILLGNSQTGHQKKLDDIYTDLYVVENKTGGRENDHEVMQMESKHNQQTAKDKPLKCNDMFK
up08 KKKLKEDYKQILIGNSQTGHQKYLDDIYTDLYVVENKTGGRENDHKVMQIESKHKQKTAKDKPLKCNDIFK
14.40 KKNLKEDYKQILIGNSQMGRKKYLDDIYTDLYVVENESGGRENDHEVIQMELKHHKQTAKDKPLKCNNMFK
4.29 KDKLKQDYEQILVGNTQKGRQKYLDDIYTELYVVENETGGRENDHEVIQIESRHNQQTAEDKPIKCNDMFK
1.01 RSNLLTKFQCVSEGTATDKSPTLLNEIYTELYITEGDSGEINNEHEVRQIETQSRRSTTEDTPIQCRDIFT
1.03 RSNLLKKFQCVSEGTARNKSPTLLNEIYTELYITEGDSGEINNEHEVRQIETQSRRSTTEDTPIQCRDIFT
1.02 -SNLLKKFQCLYEGTGRIKSPTLLNEIYTELYITEGDSGEINNEHEVRQIETQSRRSTTEDTPIQCRDIFT
15.07 RSNLLKKFECLYEGTAAQGNPTVLNEIYTELYITESESGEISHEHEVRQIETQSRRSTT-DTAIKCRDIFT
25.01 RSDLLKKFECLYEGTAAQGNPTLLNEIYTELYITESESGEISHEHEVRQIETQSRRSTTEDTAIKCRDIFT
up06 RSNLRRKSQCLYEGTAKQGSPTLLNEIYTELYITEGESGEINSEHEVRQIETQSRRSTIQDTQIKCNDVFR
15.03 KSNIKQKHQYIFEGFAKQRNPTLLNEIYIDLYITEGGNGEIRNEHEYKRIEKSLKTAATATIPIKCRDIFT
15.04 KSYMKTKHQYVLEGLEIQGNRKHLNEIYTELYIIEGETEAIINKHEFTMMERAVKSAAVVAKPINCRDIFR
FrNACHT-P8 KSSLKQKFGCIFEGIPMAGKPSLLNDIYTELYISEGGSGMINDEHEIRQIESSFRQTSGQDIQINYEDIFH
TnNACHT-P8 KSHLKQKFGCIFEGIPTAGNRTALNDIYTELCITEGGSGMINNEHEIRQIESSGRTSRGDVQ-INYEDIFR
FrNACHT-P9 KSNLRKKFQCLFEGVPKAGEPRFLNEIYTELSISQACGEEVKDLQEFRDIEAS-RTTQSAPTSATCEKMFS
TnNACHT-P9 KSNLRKKFQCLFEGLAKAGEPTFLEQIYTELSISQACSREVKDLPESGKTEAS-QALQNAPT-VTCETVLG
4.46 KCSLKKKYQSVFEGIAQQGDSTLLNNIYTDLYITQGASEQVNTEHEIRQIEAASRLHQSLEIQVECKHLFE
4.14 KCSLKKKYQSVFEGIAQQGDSTLLNNIYTDLYITQGASEQVNTEHEIRQIEAASRPHQSLEIQVECKHLFE
4.12 KCGLKKKYQNVFEGIAQQGDSTLLNNIYTDLYITQGASEQVNTEHEIRQIEAASRRHQSLEIQVECKHLFE
4.03 KCSLKKKYQSVFEGIVQQGDSTLLNTIYTDLYITQGACEQVNTEHEIRQIEAASRRHQSLEIQVECKHLFE
4.31 KCSLKKKYQSVFEGIAQQGASTLLNNIYTDVYITQGASEQVNTEHEIRQIEAASRRHQSLEIQVKCKHLFK
4.10 KCSLKKKYQSVLEGNAQQGDSTLLNNIYTDLYITQGVSKQVNTEHEIRQIEAASRRHQSLEMKIECKHLFE
4.49 KCSMKKKYQSVFEGIAQQGDSTLLNKIYTDLYITQGASEQVNTEHEIRQIEAASRRHQSLEKKIECKHLFE
FrNACHT-P7a KSSLKTKFQRVFEGIAKGGSPTLLNQIYTELYITEGGSARVNDEHEVQQMESASRRACRPETAIKLQDIFN
TnNACHT-P7 KTTLKTKFQRVFEGVAKGGNPTLLNRIYTELYMTEGGSAGVNREHEVQQMESASRRAGGPETTIRHQDIFK
FrNACHT-P7b KSSLKKKFQCVFEGIGKAGHPTFLKKIYTELQLIEGWSAGVNDEHEVQQIETASRKPERAETTIRPEDIFK
FrNACHT-P14a KFRLKGKFRWSEGLTKAGRSTLLNQIYTELLITEGGSGGVNNEHEVRYIEQAARKAKTQEITVGREDIFK
FrNACHT-P14b KFRLKGKFRWSEGLTKAGSPTLLNQIYTELLITEGGSGGVNNEHEVRYIEQAARKAKTQEITVGREDIFK
TnNACHT-P48 KISLKSRCERVTEGTNEAGSGTLLKQIYTELYITEGQSEEVNTQHEVRQLERTSR-KSLQDTPIRCQDIFK
TnNACHT-P41a KISLKSRCERVTEGTDEAGSGTLLKQIYPELYITEGQSEEVNTQHEVRQLERTSR-KSLQDTPIRCQDIFK
TnNACHT-P49 KSSLKSRCECVTEGTNEAGSGALLKQIYTELYITEGQSEEVNTQHEVRQLERTSK-KSLQDTPIRCQDIFK
TnNACHT-P51a KISLNSRCECVTEGTNEGGSGTLLKQIYTELYITEGQSEEVNTQHEVRQLERTSK-KSLQDTPIRCQDIFK
TnNACHT-P47 KISLKSRCEHVTEGTNEAGSGTLLKQIYTELYITEGQSEEVNTPHEVRQLERTSK-KSLQDTPIRCQDIFK
TnNACHT-P52a KISLKSKCERVTEGTNEAGSGTLLKQIYTELYITEGQSEEVNTQHEVRQLERTSK-KASRDTPIRCQDIFR
FrNACHT-P36a KISLKRRCEHVTEGTHEAGSGTLMNKIYTELYITEGQSEEVDTQHEVRQLERTSK-KNIQDTPIKCQDIFK
FrNACHT-P36c KISLKRRCEHVTEGTHEAGSGTLLNKIYTELYITEGQSEEVDTQHEVSQLERTSK-KNIQDTPIKCQDIFK
FrNACHT-P21c KMSLKRRCEHVTEGTHEAGSGTLLNKIYTELYITEGQSEEVDTQHEVRQLERTSK-KNIQDIPIKCQDIFK
FrNACHT-P23d KMSLKRRCEHVTEGTHEAGSGTLLNKIYTELYITEGQSEEVDTQHEVRQLERTSK-KNIQDTPIKCQDIFK
FrNACHT-P36b KISLKRRCEHVTEGTNEAGSGTLLNKIYTELYITEGQSEEVDTQHEVRQLERTSK-KNIQDTPIKCQDIFK
TnNACHT-P38b KISLKSKCEHVTEGTQEAGRGTLLNKIYTELYITEGRNEEVNTQHEVRQLEMTSK-KNVQDTPIKCHQIFN
FrNACHT-P38 KNALKTRCQHVPEGTYENGKVTPLNKIFTELYITDGLSEDVNTQHEVMQLETTSQRNALHHNPIKCGDIFK
TnNACHT-P38a KNALKSRCKHVIEGTDVKGNAAPLNRIFTELHITDGQSEAVNSQHEVMQLEAASKMSALRPVPIRCHDLF-
3.07 KFIMKNKYESLFEGVKLQQNQTLLNSIYTQLYIIEGESEGVNEEHEVLQMEKRYKT--LQDTPINCNDIFN
3.08 RFIMKIKYESLFEGVKLQQNQTLLNSIYTQLYIIEGESEGVNEEHEVLQMEKRYKT--LQDTPINCNDIFN
1.30 KTSMKNKYERLFEGNRLQENQTLLNRIYTQLYIIEGESAGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.27 KTNMKNKYERLFEGIKLQENQTLLNRIYTQLYIIEGESEGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.28 KTSMKNKYERLFEGLKLDETGTHLNRIYTQLYILEGESEGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.13 KTSMKRKYERLFEGITLQETGTHLSQIYTQLYILEGESEGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.19 KAIMRNKYDRLYEGIKLQETETLLNRIYTQLYIIDGESGGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.21 KTSMRKKYELLSEGIKVKETGTHLSRIYTQLYIIEGESEGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.22 KASMKNKYELLSEGIKVKENQTHLNRIYTQLYIIEGESEGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.23 KTSLKNKYERLFEGIEVKETGTHLNRIYTQLYIIEGESEGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
1.31 KTSMKNKYEKIFEGNKLNQNETLLNQIYTQLYIIEGESEGVNEEHEVLQMEKTARTKHSQHTPIYCNDIFK
4.02 KTSMKNKYERLFEGINHGETQTLLNSIYTQLYIIEGESEGVNEEHEVLQMEKRARTKPSQHTPVYCNDIFK
4.45 KTSMKNKYERLFEGINHRETQTLLNRIYTQLYIIEGESEGVNEEHEVLQMEKRARTKPSQHTPVYCNDIFK
4.15 KTIMKNKYERLFEGINHGETQTLLNRIYTQLYIIEGESEGVNEEHEVLQMEKRARTKPSQHTPVFCNDIFK
4.28 KTSMKNKYERLFEGIHHGETPTLLNRIYTQLYIIEGESEGVNEEHEVLQMEKRARTEPSQHTPVYCNDIFK
4.34 KTSMKNKYKRLFEGIHHRETQTLLNRIYTQLYIIEGESEGVNEEHEVSHMEKRARTEPSQHTPVYCNDIFK
4.13 KTSMKNKYERLFEGINHGETQKFLNRIYTQLYIIEGESEGVNREHDVLQIEKRVRTKPSQHTPVYCNDIFK
4.06 KTSMKNKYVRLFEEIKHRDSQTLLNRIYTQLYIIQGESEGVNREHDVLQTEKTAQTRPSQHTPVYCNDIFK
1.32 KESIKASVD-VFEG--KKDNKTDINKVYTELFITEGDMTDVCKEHEVQKFDLALQMPRSEDKPIDCNEIFN
FrNACHT-P6 KSSLKEKFQHLPEGIGEVGNTSCLNEIYTELLITEGRPGAVNEEHEVRTIEAASWRAGRSEKTIRCEDIFK
TnNACHT-P6 KSLLKQRFQHLSEGIGGLGTAPRLDEVYTELLMTEGRAGGVNGEHEVRQVEAASGEQAEPKETIRCQDLFQ
FrNACHT-P13a KSLLKRRFERVVEGDAR-RKNSSLDQIYTDLFHTEGGTGDVISEHEVRQIESASRKQNKHETKIGCENLFK
FrNACHT-P13b KSLLKKRFERVVEGDTR-GKNSSLDQIYTDLFHTEGGTGDVISEHEVRQIESASRKQNKHETKIGCENLFK
TnNACHT-P13 KSLLKNRFGLVVEGLAG-RKDSVLDQVYTELHHTEGWNGEVSTEHEVRQMESAASNANSSHTKIRCEDLFQ
TnNACHT-P15 KTSFKRRFS-HLEAVMTEAQKTPLHEIYTELFLTEGGSGEVNQEHEVRQMEASARVQVAEERSILCSQLFQ
7.01 KTKLKNKCQNINEGNSSYGKQSLLMEVYTDLYFTEGGSGEINNEHEVLRNEMISQSRTTQEVLYACDEIFK
Fr
Dr group1
Dr group 4
Fr / Tn
Dr group 2
Fr / Tn
Tn
Fr
Fr / Tn
Dr group 3
Dr / Fr / Tn
(b)
Fish domain
Relative Entropy
Contribution
LEDIYTE HsNod2 (248 - 255)
LLLVEGLTDLQLREHDFLQ HsNod3 (137 - 155)
sea urchin NLR proteins

Figure 14 *(see legend on previous page)*

Distribution in the genome

The genes encoding the novel proteins are distributed throughout the genome. Some chromosomes contain single genes, or a few, widely spaced genes, but many of the genes occur in large tandem clusters (Figure 16).

Conclusion

Our findings show that the components of the TLR and class II cytokine signaling systems known from mammals are also found in teleosts. Although all of the main constituents are present, there are differences in the degree to which the various functional groups are conserved. This is the case both for the divergence in sequence as well as for the creation of new genes by duplications.

The most highly conserved group of proteins are those involved in intracellular signal transduction downstream of the transmembrane receptors: the kinases, adaptors, Stats, Trafs and transcriptional regulators. They exhibit high sequence conservation and largely orthologous relationships, such that for each gene there is one copy in each species, and these genes are more closely related to each other than to other genes of the family. We see only a few cases of duplications. In some cases (Ticam-1, Ticam-2, and IRAK2) there appear to have been gene duplications only in mammals, but more often we find additional genes in the fish genomes. Additional copies of genes in the teleosts need not necessarily be generated by lineage-specific individual gene duplications, but may instead be remnants of the third whole genome duplication postulated for the teleost lineage [3]. We do not see as a general rule that for each mammalian gene there is more than one copy in the zebrafish genome. However, in the highly conserved gene groups we do in fact see more duplications of fish genes than of mammalian genes (additional copies for 12 genes in the case of teleosts, although not always in all three species, and only three duplicates in the two mammals). This suggests that at least some of these may indeed be remnants of the third whole genome duplication in teleosts, as is supported by the syntenic organization of the duplicated genes and the flanking genes in the case of the Stat genes.

The family of the class II cytokine receptors is neither highly conserved, nor does it exhibit species-specific expansions. The five species we compared have approximately the same number of receptor chain genes, but the divergence is so great that no reliable orthologies can be established. A similar lack of orthology is seen for the ligands. Apart from the lineage specific expansions of the type I IFNs, there are similar numbers of class II cytokine genes in the five species, but they cannot be assigned into orthologous groups (with the exception of IL-10 and IFN-γ). The strong divergence also prohibits speculations on which ligand might bind to which receptor in the zebrafish. For one pair this has recently been established experimentally; CRFB1 and CRFB5 are the receptor chains for INF-φ1 and are involved in defense against viruses [18]. Similar studies will be necessary to determine the functions of the remaining ligands and receptors. The rapid evolution of the gene families for the class II cytokines and their receptors probably reflects the fact that the IFN system is frequently subverted by pathogens, resulting in the need for compensatory mutations to escape inactivation. Significantly, the receptor family member that is not primarily associated with pathogen defense, TF, does not exhibit this high level of divergence.

The greatest divergence is found in the NLR protein family, with lineage-specific expansions in each organism, as has also been found for this type of protein in echinoderms [47,48]. Similar, if less extreme, situations are found for the TLRs [6,7] and the novel immune-type receptors [8-10], gene families that also have sets of orthologous receptors in fish and mammals as well as fish-specific expansions. Thus, the elements of the systems that are directly involved in interactions with pathogen components are those that are most likely to diversify by undergoing lineage-specific expansions. Indeed, a study that specifically tested the role of lineage-specific gene families in five eukaryotic species found that the genes that were particularly prone to such expansions included those involved in responses to pathogens [50]. Furthermore, our results are in concordance with recent findings from a comparison of three insect genomes that showed the following [51]: first, the genes associated with immune functions are on average more divergent than the rest of the genome; and second, that the divergence occurs primarily in those genes whose products interact with the pathogen. This study found that in addition to pathogen recognition proteins, this was also the case for the effectors, a set of proteins we have not analyzed in the zebrafish.

The expansion of gene families involved in pathogen recognition is likely to reflect adaptations of the species to new pathogen environments. We have not yet tested whether there is a particularly high level of sequence variability associated with particular parts of the NLR proteins. The number of LRRs varies greatly, but it will be necessary to validate the gene models for each gene before any reliable conclusions can be drawn. It will also be interesting to see whether the genes are more polymorphic than other genes in the genome. The fact that the few ESTs that are available, which are derived from a different strain of zebrafish, do not correspond 100% to any of the gene models is a hint that this might be the case. The function of the NLR genes and the significance of their species-specific expansion will be an exciting topic for experimental analysis.

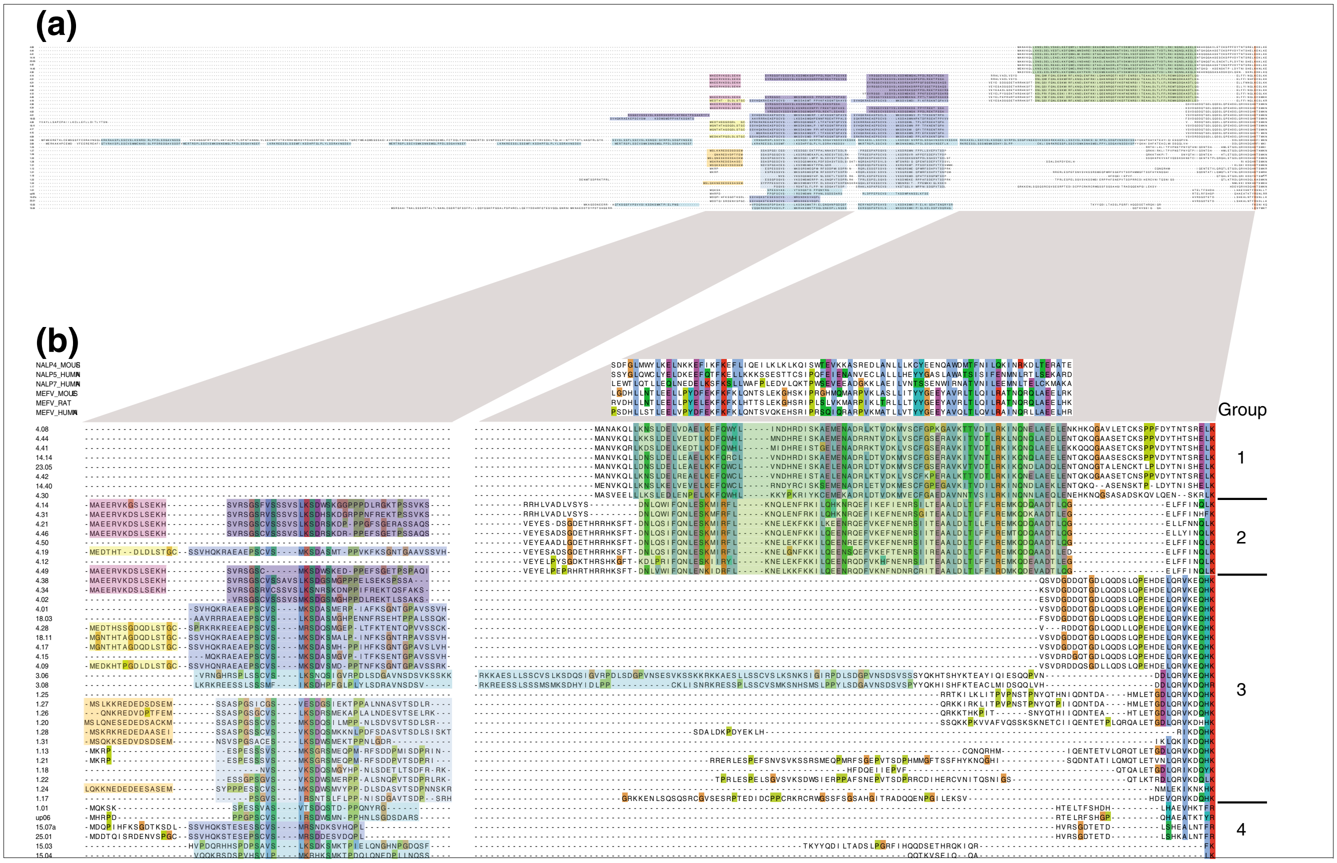

Figure 15

Structure of the amino-termini of the new zebrafish NLR proteins. ClustalW alignment of the set of 70 predicted NLR proteins was truncated four amino acids downstream of the start of the Fisna domain, and the alignment of the remaining amino-terminal sequences was edited manually using Jalview [62]. Sequences that did not extend significantly beyond the Fisna domain were deleted, as were some sequences in groups with many similar or identical sequences. The remaining sequences represent a set of characteristic compositions of motifs found in the amino-terminal part of this family of proteins. **(a)** Overview of the alignment with characteristic sequence motifs shaded in color: green = pyrin-like domain in groups 1 and 2; blue = repeated motif (different shades of blue mark different versions of the repeat); yellow/orange tones = conserved amino-terminal amino acids; and pink = specific amino-terminal peptide of 14 amino acids. **(b)** Details of the alignment in panel a in which amino acid similarities and identities are highlighted in ClustalW colors. A set of mammalian PYD domains are aligned above the zebrafish group 1 and group 2 pyrin-like domains to illustrate the similarity. NLR, nucleotide-binding domain/NACHT domain and leucine rich repeat containing family.

Materials and methods

Software

Standard web-based programs were used for sequence comparisons, alignments, and phylogenies. The phylogenetic trees in the figures were generated using the MEGA software package [27].

In all phylogenetic trees presented in this study complete sequences were used rather than only the conserved domains.

The alignments for generating the phylogenetic trees were performed with ClustalW using the Blosum matrix with standard parameters. For the phylogenetic reconstruction the neighbor-joining method [52] was used with a bootstrap test of 1,000 replicates. Gaps and missing data were treated as pair-wise deletions.

Manual annotations of genes were carried out by the Havana group at the Sanger Institute, in accordance with human annotation workshop guidelines [53].

Search for class II cytokine receptor genes

To identify class II cytokine receptor genes we searched the zebrafish genome and all available zebrafish ESTs for the subdomains SD100A and SD100B running the Prosite protein annotation [54] with the hidden Markov model matrices with accession numbers PS50299 (SD100A) and PS50300 (SD100B).

The screen of genomic sequences encoding SD100A or SD100B domains identified 12 genes, of which two encoded titins, one encoded thrombopoeitin, eight encoded cytokine class II receptor genes that previously were found to belong to the Interpro IPR000282 family, and one

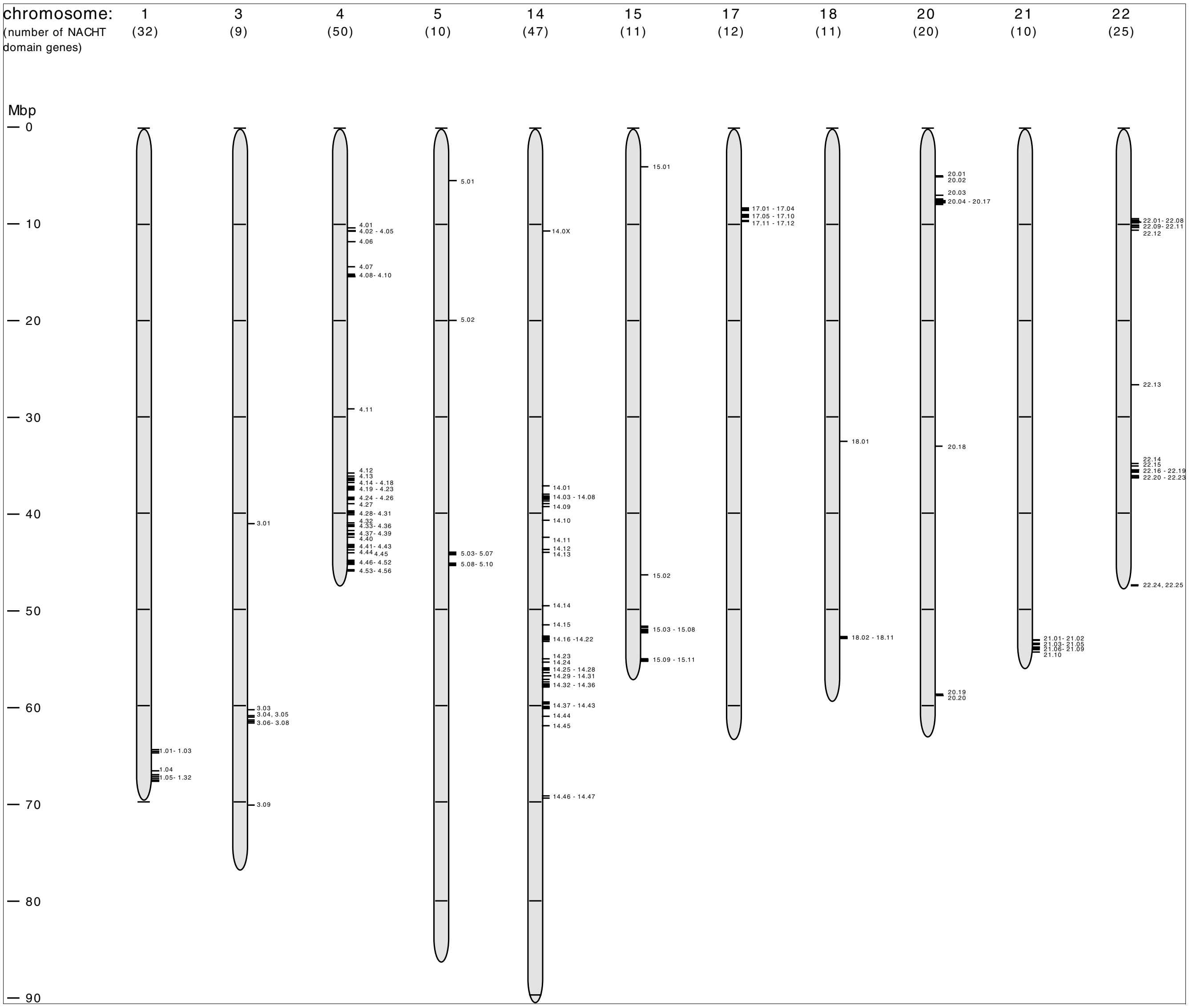

Figure 16

Chromosomal locations of zebrafish NLR proteins. The 11 chromosomes containing the main clusters of NLR genes are shown. The number of NLR genes on each chromosome is listed below the chromosome number. A further 42 genes) are distributed on 11 other chromosomes, and 20 genes are on as yet unplaced contigs. This list includes a compilation of all predictions (automated as well as manually annotated) and locations of hits from a TBLASTN search for NACHT domains. Future improvements of the genome assembly and further manual annotations will most likely result in minor changes of this map. Genes are denoted by lines on the right of the chromosome irrespective of orientation. NLR, nucleotide-binding domain/NACHT domain and leucine rich repeat containing family.

(GENSCAN00000036149) encoded a previously unidentified gene of this class.

To screen the ESTs, we first translated every EST sequence in the six possible frames and then searched for the subdomains. We followed a similar procedure with all the *ab initio* predictions (Genscan and Fgenesh) obtained in the analysis of the zebrafish Zv6 assembly [24].

From the EST analysis we obtained 69 different sequences, of which 14 encoded both subdomains. Comparison of the 69 sequences showed that they represented 20 different genes, for which we analyzed the known or predicted full-length sequences in more detail. One of the ESTs (accession CK692344) was not represented in the zebrafish genome (neither assembly Zv6 nor trace sequences) and turned out to correspond to a mouse gene. Three sequences had only spurious resemblances to SD100A or SD100B encoding sequences, often over very short stretches, and encoded known proteins with other functions. This left 16 potential candidates for cytokine class II receptor encoding genes, which we named zf1 to zf16. Six of these had also been identified by the genomic screen. Two candidates from the genomic screen were not in this group, because no ESTs exist for them. We named these candidates zf17 and zf18. We then assessed the annotations of zf1 to zf18, and annotated or re-annotated the sequences manually, if no annotations existed (zf1, zf2, zf6, and zf14) or the previous annotations appeared incomplete or incorrect. This analysis showed that twelve of the genes encoded proteins with the characteristics of class II cytokine receptors.

Search for new NLR proteins

For the manual annotation of NLR genes in the zebrafish genome, we initially used the ESTs with the accession numbers CF347458.1, CD284951.1, CO915312.1, CF266152.1, BM534859.1, and DT055906.1 as guides. The ESTs were not 100% identical to any of the genomic sequences we identified, which may be due to polymorphisms between the strains from which the genome sequence and the ESTs were derived. The NLR proteins were identified as follows. A TBLASTN search of the Ensembl zebrafish genome assembly Zv4 with the mammalian Nalp3 gene identified more than 200 sites in the genome encoding complete or partial NACHT domains. A collection of 170 NACHT-domain encoding zebrafish genes from the NCBI database, which only partly overlapped the set identified by TBLASTN, were also mapped onto the genome. The merged list of the two nonoverlapping sets of sites in the genome were sorted by chromosomal location, each site was given a number (chromosome number plus numerical ordering). The regions containing the potential genes were then further refined using available ESTs and gene predictions as guides. The resulting sequences were blasted against the finished and unfinished clone sequences and the hits on finished clones were finally manually annotated. For further refinement of annotations we also used the motifs identified in Figure 15 in particular to improve the predictions for the full amino-terminal extensions of the genes.

Abbreviations

Apaf1, apoptotic protease activating factor 1; CRFB, cytokine receptor family B; EST, expressed sequence tag; Fisna, fish-specific NACHT associated; IL, interleukin; IFN, interferon; IFNAR, interferon-α receptor; IFNGR, interferon-γ receptor; IL10R, interleukin-10 receptor; IL22BP, interleukin-22 binding protein; IRAK, interleukin-1 receptor associated kinase; IRF, interferon response factor; LRR, leucine-rich repeat; NF-κB, nuclear factor-κB; NLR, NACHT-domain and leucine rich repeat containing; Nalp, NACHT, leucine rich repeat and PYD containing protein; Nod, nucleotide oligomerization domain containing protein; Stat, signal transducer and activator of transcription; Tab, Tak1-binding protein; TF, tissue factor; Ticam, Toll-interleukin 1 receptor domain (TIR) containing adaptor molecule; TLR, Toll-like receptor; TNF, tumor necrosis factor; Traf, TNF-receptor associated factor.

Authors' contributions

CS and ML conducted BLAST searches, made alignments and phylogenetic trees, made the figures and wrote the text. MC identified the cytokine and cytokine receptor genes and analyzed their genomic contexts, GL made the manual annotations of the novel NLR genes and cytokine receptor genes.

Additional data files

The following additional data are available with the online version of this paper. Additional file 1 lists the kinase protein sequences in FASTA format. Additional file 2 lists the adaptor protein sequences in FASTA format. Additional file 3 lists the IRF protein sequences in FASTA format. Additional file 4 lists the Stat protein sequences in FASTA format. Additional file 5 lists the Traf protein sequences in FASTA format. Additional file 6 lists the class II cytokine receptor protein sequences in FASTA format. Additional file 7 lists the class II cytokine protein sequences in FASTA format. Additional data file 8 lists the NLR protein sequences in FASTA format, except for the zebrafish-specific NLRs. Additional data file 9 lists the zebrafish-specific NLR protein sequences in FASTA format. Additional data file 10 is a high resolution of the large phylogram of 277 NLRs presented in Figure 12.

Acknowledgements

This work was supported by the Wellcome Trust and the European Molecular Biology Organization. ML thanks Richard Durbin, Kerstin Jekosch, and staff at the Sanger Center for providing space and a stimulating sabbatical environment. We thank our colleagues, in particular Jonathan Howard, for discussions and suggestions, Dale Richardson for assembling the set of NCBI NACHT-domain predictions, and Jane Parker and Jeff Dangl for comments on the manuscript. Jonathan Rast very kindly provided a file with the sequences of the sea urchin NACHT domain proteins. We are especially thankful to Georges Lutfalla and Dina Aggad for sharing ideas and information, and for generously providing the sequences of DrCRFB14 and IFN-φ4.

References

1. Trede NS, Langenau DM, Traver D, Look AT, Zon LI: **The use of zebrafish to understand immunity.** *Immunity* 2004, **20:**367-379.
2. Venkatesh B: **Evolution and diversity of fish genomes.** *Curr Opin Genet Dev* 2003, **13:**588-592.
3. Volff JN: **Genome evolution and biodiversity in teleost fish.** *Heredity* 2005, **94:**280-294.
4. Traver D, Herbomel P, Patton EE, Murphey RD, Yoder JA, Litman GW, Catic A, Amemiya CT, Zon LI, Trede NS: **The zebrafish as a model organism to study development of the immune system.** *Adv Immunol* 2003, **81:**253-330.
5. Nonaka M, Kimura A: **Genomic view of the evolution of the complement system.** *Immunogenetics* 2006, **58:**701-713.
6. Meijer AH, Gabby Krens SF, Medina Rodriguez IA, He S, Bitter W, Ewa Snaar-Jagalska B, Spaink HP: **Expression analysis of the Toll-like receptor and TIR domain adaptor families of zebrafish.** *Mol Immunol* 2004, **40:**773-783.
7. Jault C, Pichon L, Chluba J: **Toll-like receptor gene family and TIR-domain adapters in *Danio rerio*.** *Mol Immunol* 2004, **40:**759-771.
8. Litman GW, Hawke NA, Yoder JA: **Novel immune-type receptor genes.** *Immunol Rev* 2001, **181:**250-259.
9. Yoder JA, Mueller MG, Wei S, Corliss BC, Prather DM, Willis T, Litman RT, Djeu JY, Litman GW: **Immune-type receptor genes in zebrafish share genetic and functional properties with genes encoded by the mammalian leukocyte receptor cluster.** *Proc Natl Acad Sci USA* 2001, **98:**6771-6776.

10. Yoder JA, Litman RT, Mueller MG, Desai S, Dobrinski KP, Montgomery JS, Buzzeo MP, Ota T, Amemiya CT, Trede NS, *et al.*: **Resolution of the novel immune-type receptor gene cluster in zebrafish.** *Proc Natl Acad Sci USA* 2004, **101:**15706-15711.
11. Panagos PG, Dobrinski KP, Chen X, Grant AW, Traver D, Djeu JY, Wei S, Yoder JA: **Immune-related, lectin-like receptors are differentially expressed in the myeloid and lymphoid lineages of zebrafish.** *Immunogenetics* 2006, **58:**31-40.
12. Bobe J, Goetz FW: **Molecular cloning and expression of a TNF receptor and two TNF ligands in the fish ovary.** *Comp Biochem Physiol B Biochem Mol Biol* 2001, **129:**475-481.
13. Engelsma MY, Huising MO, van Muiswinkel WB, Flik G, Kwang J, Savelkoul HF, Verburg-van Kemenade BM: **Neuroendocrine-immune interactions in fish: a role for interleukin-1.** *Vet Immunol Immunopathol* 2002, **87:**467-479.
14. Zou J, Secombes CJ, Long S, Miller N, Clem LW, Chinchar VG: **Molecular identification and expression analysis of tumor necrosis factor in channel catfish (*Ictalurus punctatus*).** *Dev Comp Immunol* 2003, **27:**845-858.
15. Huising MO, Stet RJ, Savelkoul HF, Verburg-van Kemenade BM: **The molecular evolution of the interleukin-1 family of cytokines; IL-18 in teleost fish.** *Dev Comp Immunol* 2004, **28:**395-413.
16. Reboul J, Gardiner K, Monneron D, Uze G, Lutfalla G: **Comparative genomic analysis of the interferon/interleukin-10 receptor gene cluster.** *Genome Res* 1999, **9:**242-250.
17. Lutfalla G, Roest Crollius H, Stange-Thomann N, Jaillon O, Mogensen K, Monneron D: **Comparative genomic analysis reveals independent expansion of a lineage-specific gene family in vertebrates: the class II cytokine receptors and their ligands in mammals and fish.** *BMC Genomics* 2003, **4:**29.
18. Levraud JP, Boudinot P, Colin I, Benmansour A, Peyrieras N, Herbomel P, Lutfalla G: **Identification of the zebrafish IFN receptor: implications for the origin of the vertebrate IFN system.** *J Immunol* 2007, **178:**4385-4394.
19. Baoprasertkul P, Peatman E, Somridhivej B, Liu Z: **Toll-like receptor 3 and TICAM genes in catfish: species-specific expression profiles following infection with *Edwardsiella ictaluri*.** *Immunogenetics* 2006, **58:**817-830.
20. Ben J, Jabs EW, Chong SS: **Genomic, cDNA and embryonic expression analysis of zebrafish IRF6, the gene mutated in the human oral clefting disorders Van der Woude and popliteal pterygium syndromes.** *Gene Expr Patterns* 2005, **5:**629-638.
21. Lewis RS, Ward AC: **Conservation, duplication and divergence of the zebrafish stat5 genes.** *Gene* 2004, **338:**65-74.
22. Oganesyan G, Saha SK, Guo B, He JQ, Shahangian A, Zarnegar B, Perry A, Cheng G: **Critical role of TRAF3 in the Toll-like receptor-dependent and -independent antiviral response.** *Nature* 2006, **439:**208-211.
23. van der Sar AM, Stockhammer OW, van der Laan C, Spaink HP, Bitter W, Meijer AH: **MyD88 innate immune function in a zebrafish embryo infection model.** *Infect Immun* 2006, **74:**2436-2441.
24. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, *et al.*: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34:**D556-561.
25. Sprague J, Doerry E, Douglas S, Westerfield M: **The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research.** *Nucleic Acids Res* 2001, **29:**87-90.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
27. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment.** *Brief Bioinform* 2004, **5:**150-163.
28. Sullivan C, Postlethwait JH, Lage CR, Millard PJ, Kim CH: **Evidence for evolving Toll-IL-1 receptor-containing adaptor molecule function in vertebrates.** *J Immunol* 2007, **178:**4517-4527.
29. Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, Agata K, Bosch TC: **The innate immune repertoire in cnidaria - ancestral complexity and stochastic gene loss.** *Genome Biol* 2007, **8:**R59.
30. Kedinger V, Alpy F, Tomasetto C, Thisse C, Thisse B, Rio MC: **Spatial and temporal distribution of the traf4 genes during zebrafish development.** *Gene Expr Patterns* 2005, **5:**545-552.
31. Krause CD, Pestka S: **Evolution of the Class 2 cytokines and receptors, and discovery of new friends and relatives.** *Pharmacol Ther* 2005, **106:**299-346.
32. Altmann SM, Mellon MT, Distel DL, Kim CH: **Molecular and functional analysis of an interferon gene from the zebrafish, *Danio rerio*.** *J Virol* 2003, **77:**1992-2002.
33. Igawa D, Sakai M, Savan R: **An unexpected discovery of two interferon gamma-like genes along with interleukin (IL)-22 and -26 from teleost: IL-22 and -26 genes have been described for the first time outside mammals.** *Mol Immunol* 2006, **43:**999-1009.
34. Chen JY, You YK, Chen JC, Huang TC, Kuo CM: **Organization and promoter analysis of the zebrafish (*Danio rerio*) interferon gene.** *DNA Cell Biol* 2005, **24:**641-650.
35. Milev-Milovanovic I, Long S, Wilson M, Bengten E, Miller NW, Chinchar VG: **Identification and expression analysis of interferon gamma genes in channel catfish.** *Immunogenetics* 2006, **58:**70-80.
36. Long S, Milev-Milovanovic I, Wilson M, Bengten E, Clem LW, Miller NW, Chinchar VG: **Identification and expression analysis of cDNAs encoding channel catfish type I interferons.** *Fish Shellfish Immunol* 2006, **21:**42-59.
37. Zou J, Carrington A, Collet B, Dijkstra JM, Yoshiura Y, Bols N, Secombes C: **Identification and bioactivities of IFN-gamma in rainbow trout *Oncorhynchus mykiss*: the first Th1-type cytokine characterized functionally in fish.** *J Immunol* 2005, **175:**2484-2494.
38. Zou J, Yoshiura Y, Dijkstra JM, Sakai M, Ototake M, Secombes C: **Identification of an interferon gamma homologue in *Fugu*, *Takifugu* rubripes.** *Fish Shellfish Immunol* 2004, **17:**403-409.
39. Long S, Wilson M, Bengten E, Bryan L, Clem LW, Miller NW, Chinchar VG: **Identification of a cDNA encoding channel catfish interferon.** *Dev Comp Immunol* 2004, **28:**97-111.
40. Robertsen B, Bergan V, Rokenes T, Larsen R, Albuquerque A: **Atlantic salmon interferon genes: cloning, sequence analysis, expression, and biological activity.** *J Interferon Cytokine Res* 2003, **23:**601-612.
41. Zhang DC, Shao YQ, Huang YQ, Jiang SG: **Cloning, characterization and expression analysis of interleukin-10 from the zebrafish (*Danio rerio*).** *J Biochem Mol Biol* 2005, **38:**571-576.
42. Zou J, Clark MS, Secombes CJ: **Characterisation, expression and promoter analysis of an interleukin 10 homologue in the puffer fish, *Fugu rubripes*.** *Immunogenetics* 2003, **55:**325-335.
43. Zou J, Tafalla C, Truckle J, Secombes CJ: **Identification of a second group of type I IFNs in fish sheds light on IFN evolution in vertebrates.** *J Immunol* 2007, **179:**3859-3871.
44. Damiano JS, Oliveira V, Welsh K, Reed JC: **Heterotypic interactions among NACHT domains: implications for regulation of innate immune responses.** *Biochem J* 2004, **381:**213-219.
45. Koonin EV, Aravind L: **The NACHT family: a new group of predicted NTPases implicated in apoptosis and MHC transcription activation.** *Trends Biochem Sci* 2000, **25:**223-224.
46. van der Biezen EA, Jones JD: **The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals.** *Curr Biol* 1998, **8:**R226-227.
47. Hibino T, Loza-Coll M, Messier C, Majeske AJ, Cohen AH, Terwilliger DP, Buckley KM, Brockton V, Nair SV, Berney K, *et al.*: **The immune gene repertoire encoded in the purple sea urchin genome.** *Dev Biol* 2006, **300:**349-365.
48. Rast JP, Smith LC, Loza-Coll M, Hibino T, Litman GW: **Genomic insights into the immune system of the sea urchin.** *Science* 2006, **314:**952-956.
49. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
50. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes.** *Genome Res* 2002, **12:**1048-1059.
51. Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, *et al.*: **Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes.** *Science* 2007, **316:**1738-1743.
52. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
53. **Havana** [http://www.sanger.ac.uk/HGP/havana/hawk.shtml]
54. **Prosite** [http://www.expasy.org/prosite/]
55. **Pfam** [http://www.sanger.ac.uk/Software/Pfam]
56. **Smart** [http://smart.embl-heidelberg.de]
57. Renauld JC: **Class II cytokine receptors and their ligands: key antiviral and inflammatory modulators.** *Nat Rev Immunol* 2003, **3:**667-676.

58. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
59. **HMMER** [http://bioweb.pasteur.fr/seqanal/interfaces/hmmbuild.html]
60. Schuster-Bockler B, Schultz J, Rahmann S: **HMM Logos for visualization of protein families.** *BMC Bioinformatics* 2004, **5:**7.
61. **HMM logo web server** [http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi]
62. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20:**426-427.