Evolutionary redesign of the Atlantic cod (*Gadus morhua* L.) Toll-like receptor repertoire by gene losses and expansions

Monica H. Solbakken¹, Ole K. Tørresen¹, Alexander J. Nederbragt¹-⁵, Marit Seppola³, Tone F. Gregers¹, Kjetil S. Jakobsen¹ & Sissel Jentoft¹,⁴

Genome sequencing of the teleost Atlantic cod demonstrated loss of the Major Histocompatibility Complex (MHC) class II, an extreme gene expansion of MHC class I and gene expansions and losses in the innate pattern recognition receptor (PRR) family of Toll-like receptors (TLR). In a comparative genomic setting, using an improved version of the genome, we characterize PRRs in Atlantic cod with emphasis on TLRs demonstrating the loss of TLR1/6, TLR2 and TLR5 and expansion of TLR7, TLR8, TLR9, TLR22 and TLR25. We find that Atlantic cod TLR expansions are strongly influenced by diversifying selection likely to increase the detectable ligand repertoire through neo- and subfunctionalization. Using RNAseq we find that Atlantic cod TLRs display likely tissue or developmental stage-specific expression patterns. In a broader perspective, a comprehensive vertebrate TLR phylogeny reveals that the Atlantic cod TLR repertoire is extreme with regards to losses and expansions compared to other teleosts. In addition we identify a substantial shift in TLR repertoires following the evolutionary transition from an aquatic vertebrate (fish) to a terrestrial (tetrapod) lifestyle. Collectively, our findings provide new insight into the function and evolution of TLRs in Atlantic cod as well as the evolutionary history of vertebrate innate immunity.

Functional understanding of teleost immunity and its diversity is still in its infancy. Homologs of both mammalian innate and adaptive immune genes have been detected in teleost genomes, however, teleosts display greater genetic diversity as well as some functional discrepancies - for examples see references¹-³. Central to innate immunity are pattern recognition receptors (PRRs) that detect pathogen associated molecular patterns (PAMPs) and initiate various features of the host's immune system - see⁴ and references therein. One of the largest PRR families is the Toll-like receptors (TLRs). Upon ligand interaction, TLRs initiate the production of cytokines, anti-viral components and co-stimulatory molecules via the TLR signalling pathway - see⁵ and references therein. The diversity of TLR repertoires among multicellular organisms is substantial. The invertebrate TLR repertoire spans from several hundred genes in the sea urchin (*Strongylocentrotus purpuratus*) to only two genes in the ascidian *Ciona intestinalis*⁶. This is in stark contrast to the less extensive vertebrate repertoire that generally display between 10–13 TLR genes - overview in⁷-⁹.

Currently, there are ~20 known vertebrate TLRs (TLR1-26, the annotation used for individual genomes varies) where mammals display TLR1-13 in contrast to fish which also display TLR14–26. Vertebrate TLRs form six families: TLR1, TLR3, TLR4, TLR5, TLR7 and TLR11 and individual species generally harbours at least one member from each family⁸. However, some exceptions are known such as the lack of TLR11 -family representatives in mammals. Teleosts display greater genetic diversity of TLRs but functional studies on mammalian TLR homologs overall report identical protein function - see⁹.

In contrast to the genetic diversity found within the innate immune system the adaptive immune system is shown to display an intra-genetic polymorphic nature, i.e. to enable adaptation of the immune response towards
specific targets\textsuperscript{10}. Large structural or functional alterations affecting acquired immunity have been perceived as less likely. During the last decade, however, several alternative immune strategies have been identified in vertebrate species – for details see\textsuperscript{11,12}. Atlantic cod (\textit{Gadus morhua}) is a particularly interesting case as genome sequencing revealed complete loss of the MHC-II pathway accompanied by an extreme gene expansion of MHC-I and gene losses and expansions within the TLRs\textsuperscript{13-15}. By taking advantage of a new and substantially improved genome assembly combined with large scale genomic analyses we here perform a deep characterization of the major innate immune gene families in Atlantic cod, with emphasis on TLRs. Our phylogenetic analysis shows that the gene losses and expansions in Atlantic cod are extreme compared to other vertebrate lineages, including other teleosts. Comparative gene syntenies firmly establish the loss of TLR1/6, TLR2 and TLR5 and expansion of TLR7, TLR8, TLR9, TLR22 and TLR25. Further, we are also able to more accurately determine TLR copy number, characterize TLRs not found in the earlier version of the genome and perform multiple selection analyses. We detect varying numbers of sites under diversifying selection within the TLR expansions most likely increasing the detectable ligand repertoire through neo- and subfunctionalization. Protein structure modelling and phylogenetic analysis suggest that TLR losses do not reduce the available genetic toolkit to detect pathogens. Furthermore, our transcriptome profiling of Atlantic cod TLRs show a likely tissue specific paralog usage. Finally, a comprehensive vertebrate TLR phylogeny demonstrates that there is a shift in TLR repertoires following the transition from aquatic to terrestrial life styles mirroring different selective pressures in the two environments.

Results

Atlantic cod PRR gene families – the deviating TLRs. We have investigated all major PRR gene families in Atlantic cod using the new and improved genome assembly (for details see method section “Genome assembly”). The TLR repertoire in Atlantic cod is clearly different compared to the other investigated teleosts and vertebrates. Within the collectin, pentraxin, retinoic acid-inducible (RIG) 1-like and nucleotide-binding oligomerization domain (NOD)-like families no clear differences were found – except for two genes: Atlantic cod has no evident homolog of NOD2 and AIM2 (Supplementary Tables 1–3). We have therefore focused on the TLR repertoire in the following investigations.

Gene syntenies verify TLR gene losses and expansions. We performed gene synteny analyses on all genomic regions in the assembly containing complete TLRs in Atlantic cod against the genomes of medaka (\textit{Oryzias latipes}), fugu (\textit{Takifugu rubripes}), tetraodon (\textit{Tetraodon nigroviridis}), zebrafish (\textit{Danio rerio}) and stickleback (\textit{Gasterosteus aculeatus}). We found conserved gene organization up- and downstream of TLR1/6, TLR2 and TLR5 proving their absence from the Atlantic cod genome. Comparatively, each species contained some genomic reshuffling and additional open reading frames – particularly prominent in zebrafish (Fig. 1). We find that TLR7, TLR8, TLR9, TLR22 and TLR25 are expanded in Atlantic cod and that the gene copies display both tandem and non-tandem organization in numerous contigs (Fig. 2). The TLR8 and TLR22 expansions are the most numerous with twelve copies each. The three TLR7 copies are interspersed among the twelve TLR8 copies. They are present in three different contigs where two have partial gene synteny compared to the other investigated teleosts (Fig. 2). Again, zebrafish display the most deviating local genomic architecture (Fig. 2). The five copies of TLR9 are tandemly organized on a single contig that display general conserved synteny with the other species, however with some minor gene shuffling (Fig. 2). The twelve copies of TLR22 are found in eight contigs. Three of these contigs have tandem organization of the TLR22 copies, but most contigs are short and only contain a single gene. In only two contigs could synteny with flanking genes be determined (Fig. 2). The TLR22 synteny also reveals that zebrafish has lost TLR22. This species also harbours a local inversion involving four genes downstream of the predicted TLR22 region and display several additional open reading frames upstream compared to the other investigated species (Fig. 2). Finally, TLR25 consists of seven copies in Atlantic cod found in three contigs. Two of the contigs demonstrate partial synteny and contigs with several TLR25 copies display tandem organization. Medaka was the only other species containing TLR25 and no local synteny directly downstream of the TLR25 genomic region was evident for this species (Fig. 2). The single copy Atlantic cod TLRs, TLR3, TLR14, TLR21 and TLR23 were also located to genomic regions displaying conserved local synteny compared to the other investigated species (data not shown).

TLR expression patterns using RNAseq. To investigate TLR expression patterns in Atlantic cod we performed RNAseq using the spleen/head kidney of healthy juvenile cod where the resulting reads were mapped towards all full-length TLRs found in the new Atlantic cod genome assembly. Most of the 43 full-length TLRs had detectable expression levels; however, four TLRs (two TLR8 and two TLR25) had very low to no detectable expression. For the remaining TLRs, substantial variation in expression levels was observed (Fig. 3). The four genes with the lowest expression levels also displayed poor sequence quality resulting in protein translations containing framen-shifts and stop codons possibly indicating pseudogenes. This was also the case for an additional six TLRs. In total 10 full-length TLR genes were excluded from further analysis (Supplementary Table 4).

Endolysosomal sorting signals in Atlantic cod. We compared known endolysosomal sorting signals from mammalian TLRs in the transmembrane, linker and cytosolic region against the corresponding regions of Atlantic cod TLRs. We found that the sorting signal in TLR3 and TLR8 were well conserved across all investigated species with the exception of TLR3 in lamprey (Fig. 4A). We also searched for similar signals in the remaining TLRs: TLR7, TLR8, TLR14, TLR21, TLR23 and TLR25. For TLR25 a putative sorting signal was found (Fig. 4B), but for the other TLRs no clear conserved signalling motifs could be discerned (data not shown).

Protein structure modelling and diversifying selection. We modelled the 3D protein structure of all full-length TLRs in Atlantic cod (excluding those in Supplementary Table 4) onto the mammalian TLR5 structure (Fig. 5, Supplementary Figs 1, 2 and 3) as the overall structure of the TLR protein is central to TLR function. All
modelled TLRs conformed to the overall TLR structure with a solenoid ecto-domain, transmembrane domain, linker and Toll/interleukin-1 receptor (TIR) domain. TLR3, TLR7, TLR8, TLR9, TLR21, TLR22 and TLR23 displayed a longer solenoid ecto-domain structure (Fig. 5, Supplementary Figs 1 and 2). TLR14 and TLR25 demonstrated a somewhat shorter structure with loops modelled in their ecto-domains - more similar to the structure of other plasma membrane TLRs in mammals (Supplementary Figs 2 and 3).

The expanded Atlantic cod TLRs, with the exception of TLR7 due to low copy number, were analyzed for sites under selection using three phylogeny-guided methods; SLAC, FEL and REL (see methods for details and Table 1). TLR22 appears to have the most sites under diversifying selection and TLR25 the least. Sites common between two or more selection analyses were mapped onto one of the modelled protein structures for each of the TLR8, TLR9, TLR22 and TLR25 gene expansions demonstrating that the sites are mainly located to loops interspersed between the leucine-rich repeat elements in the TLRs ecto-domains (Fig. 5A–D).
Figure 2. Gene synteny comparison of genomic regions in Atlantic cod towards genomic regions in stickleback, medaka, fugu, tetraodon and zebrafish containing TLR7, TLR8, TLR9, TLR22 and TLR25. 

Genes with colored boxes were found in several of the investigated species whereas white boxes designated ORF represents open reading frames which are species-specific without certain annotation. Some genomic regions have been drawn in reversed order for visual purposes – designated "flip". TLRs in Atlantic cod removed from further analyses due to lacking expression and/or poor sequence quality listed in Supplementary Table S1 are written in black. TLR7 and TLR8 are located to the same genomic regions in the investigated fish species. Gene synteny is well conserved, however, zebrafish displays additional open reading frames of which some have proper annotation. Stickleback, tetraodon and zebrafish have two TLR7 whereas fugu and tetraodon lacks TLR8. Atlantic cod has three contigs containing both TLR7 and TLR8 copies interspersed. Two of these contigs have partial synteny towards the other fish species. TLR9 is also located to genomic regions with conserved synteny. Zebrafish displays less synteny downstream of its TLR9. Atlantic cod has five TLR9 copies tandemly organized on a single contig with well conserved synteny. Also TLR22 is located to a genomic region with relatively conserved synteny among the fish species. Medaka TLR22 is present on a scaffold with no other annotated genes present. No TLR22 was found in zebrafish and this species has a local inversion in the predicted TLR22 region. Atlantic cod has eight contigs with TLR22 gene copies present where two display partial synteny and tandem organization of the TLR22 copies. The remaining contigs are short and contains only that single gene. The predicted TLR25 regions have relatively well conserved synteny; however, synteny is absent downstream of medaka TLR25 and somewhat disturbed downstream in stickleback and upstream in zebrafish. TLR25 was only found in medaka and Atlantic cod. Atlantic cod TLR25 copies are present on three contigs of which two have partial synteny. Contigs with several TLR25 copies display tandem organization.
of the TLR1 species – see16. However, due to overlapping ligand profiles flagellin detection is likely covered by other PRR families in this plasma membrane associated TLR detecting flagellin7,8. However, no compensatory mechanism similar to that could be alleviated (Fig. 7). Our phylogeny and synteny analyses also describe the loss of TLR5 in Atlantic cod, a vertebrate TLR family. Our phylogeny suggests that Atlantic cod has other representatives within the TLR1/6 family – TLR14 and TLR25 – and thus any reduced ability to detect peptidoglycan/lipoprotein by TLRs could be alleviated (Fig. 7). Our phylogeny demonstrates that the TLR gene expansions in Atlantic cod are rather extreme compared to the relatively few duplicates, triplicates and a single quadruplet expansion (xenopus X. tropicalis)) seen in the other species. No expansions were found within the human TLR repertoire (Fig. 7, Table 2).

The TLR signalling pathway is intact in Atlantic cod. Using the mammalian TLR signalling network we searched for homologous genes in the new version of the Atlantic cod genome assembly (Supplementary Table 5). All components of the TLR signalling pathway were detected with the exception of TLR4 associated co-factors and some downstream T-cell/B-cell co-stimulatory molecules which were difficult to confirm due to distant sequence homology (Fig. 6). One downstream cytokine, interleukin-8 (IL8) showed substantial gene expansion: eight copies in total of which six were assembled to full-length (Supplementary Table 6). The translated sequences were subjected to a maximum likelihood (ML) protein sequence phylogenetic analysis together with IL8 from fugu, tetraodon, tilapia, stickleback, medaka and human. The phylogeny grouped Atlantic cod IL8s in two clades (Supplementary Fig. 4). Transcriptome profiling of IL8 (identical to that performed on Atlantic cod TLRs) did not resolve the paralogs sufficiently and thus the expression pattern of each clade or individual paralogs could not be further addressed (data not shown).

TLR annotation and vertebrate repertoires. We performed a multi-TLR, multi-species phylogenetic analysis using the translated sequence of the transmembrane, linker and TIR-domain regions of all TLR genes in selected vertebrate species with a main emphasis on teleosts (Supplementary Tables 2–4). The phylogeny resolved all six major TLR families, however, the TLR11 and TLR5 families display weaker support than the remaining families likely connected to the placement of TLR21, TLR26 and TLR13 (Fig. 7). Atlantic cod was the only species not harbouring any TLRs phylogenetically grouping within the TLR1/6 and the TLR2 clades of the TLR1-family. However, TLR14 and TLR25 are well supported within the TLR1-family clade. TLR14 was not found in chicken and human. TLR13 was present in the anole lizard (Anolis carolinensis), xenopus (Xenopus tropicalis) and coelacanth (Latimeria chalumnae). TLR25 and TLR26 were both sparsely found among the investigated fish species. Humans were the only species not displaying any members of the TLR11-family. The TLR5-family was not represented in either Atlantic cod or lamprey and the TLR4-family was only found in zebrafish, chicken (Gallus gandus), anole lizard and humans. Furthermore, the phylogeny demonstrates that the TLR gene expansions in Atlantic cod are rather extreme compared to the relatively few duplicates, triplicates and a single quadruplet expansion (xenopus TLR14) seen in the other species. No expansions were found within the human TLR repertoire (Fig. 7, Table 2).

Discussion

Signs of compensatory mechanisms for lost TLRs. Our TLR phylogeny indicates that Atlantic cod is the only known species lacking TLR1/6 and TLR2 which is confirmed by gene synteny analysis (Figs 1 and 7). These TLRs, members of the TLR1-family, are known to recognize peptidoglycan/lipoproteins at the plasma membrane. Roach et al. have demonstrated a convincing link between phylogenetic relationships and function within vertebrate TLR families. Our TLR phylogeny suggests that Atlantic cod has other representatives within the TLR1-family – TLR14 and TLR25 – and thus any reduced ability to detect peptidoglycan/lipoprotein by TLRs could be alleviated (Fig. 7). Our phylogeny and synteny analyses also describe the loss of TLR5 in Atlantic cod, a plasma membrane associated TLR detecting flagellin. However, no compensatory mechanism similar to that of the TLR1-family was found as no other Atlantic cod TLR was placed within the TLR5-family (Figs 1 and 7). However, due to overlapping ligand profiles flagellin detection is likely covered by other PRR families in this species – see16.
Functional assessment of TLRs through comparative analyses. With the aim of inferring function on Atlantic cod TLRs we performed several comparative analyses based on sequence homology which we interpreted using established links between function and phylogenetic relationships, protein structure and sorting signals. For TLR3, TLR7, TLR8 and TLR9 our findings support earlier functional reports demonstrating nucleic acid ligands and intracellular localization identical to their mammalian counterparts (Figs 2, 4A, 5A, 5B and Supplementary Fig. 2)\(^{17}\). There are limited functional studies on non-mammalian TLRs (TLR11–26) of which TLR14–26 are present in teleosts. For TLR14 and TLR25 functional studies have so far not fully resolved ligand specificity. However, interesting results include transcriptional up-regulation of TLR14 after exposure to viable gram negative bacteria\(^{18}\) and transcriptional up-regulation of TLR25 in response to parasites\(^{19}\). We propose a TLR1-family-like function for TLR14 and TLR25 implying plasma membrane localization and peptidoglycan or lipopolysaccharide-like ligands. This is further supported by protein structure modelling resolving shorter disrupted solenoid structures (Supplementary Figs 2 and 3) – structures correlated with plasma membrane localization and non-nucleic acid ligands\(^{7,20}\). Furthermore, the presence of an intact TLR signalling pathway (Fig. 6) also supports the proposed function of TLR14 and TLR25. Otherwise one would expect a concurrent loss of adaptor proteins and co-factors specific for plasma membrane associated TLR proteins – in line with the observed loss of all TLR4-associated adapters in species lacking TLR4\(^{21}\). Lastly, our analysis revealed a putative endolysosomal sorting signal in TLR25 similar to that of mammalian TLR3 and TLR9 (Fig. 4B)\(^{22–25}\). For TLR21 reports suggest that it is an intracellular TLR with a nucleic acid ligand\(^{26,27}\). No firm conclusion can be drawn for TLR22; there are several incongruent reports indicating a cell surface location with a nucleic acid ligand as well as transcriptional response towards several non-nucleic acid stimulants like peptidoglycan and lipopolysaccharide\(^{28–32}\). The function of TLR23 is also not established\(^{29}\).

Figure 4. Edited amino acid alignments of the linker and transmembrane region of TLR3, TLR9 and TLR25 displaying known or putative tyrosine-containing endolysosomal sorting signals. (A) The known TLR3 endolysosomal sorting signal is well conserved across species (black box) with the exception of TLR3 in lamprey which has a phenylalanine in the tyrosine position and a tyrosine in the position before. For TLR9 the signal is conserved in all species (black box). (B) For TLR25 we propose an endolysosomal sorting signal in the linker region conserved across all species investigated that contain TLR25.
endosomal localization and that computational data supports a nucleic acid ligand for TLR22, our findings suggest that this whole family of TLRs do have nucleic acid ligands and most like intracellular localization.

Functional implications of lost and expanded TLRs. We detected diversifying selection among paralogs within the expanded Atlantic cod TLRs: TLR8, TLR9, TLR22 and TLR25 (Table 1). TLR9 and TLR22 stand out with the highest number of sites reported. Upon PAMP recognition, TLRs form TLR-homodimer:ligand complexes. Vertebrates can further expand their detectable ligand repertoire by forming heterodimers within or between TLR families as have been demonstrated for TLR1/2, TLR2/6, TLR11/12 and TLR4/6. The number of sites under diversifying selection in the ecto-domain of TLR9 and TLR22 suggests that the Atlantic cod's innate immune strategy partly involves an increase in its detectable ligand repertoire relative to other investigated fish species through “heterodimerization” between paralogs or possibly heterodimerization of paralogs with other TLRs. For TLR8 and TLR25, the number of sites detected was much lower and somewhat inconsistent between the different methods (Table 1) suggesting that increased detectable ligand repertoire is not the main force maintaining these two gene expansions. We investigated the possibility of increased gene dosage by performing a

Figure 5. Sites under diversifying selection mapped onto the protein modeled structures of one paralog from each of the gene expansions TLR8, TLR9, TLR22 and TLR25 in Atlantic cod. The transmembrane, linker and TIR domain is colored dark grey whereas the ecto-domain is colored light grey with its sheets in pale green and helices in light blue. Sheets overlap with leucine-rich repeats in the ecto-domain. Arrows pointing at bright blue/bright red/bright green represents sites under diversifying selection as reported in Table 1. (A) Five sites (blue) mapped onto the modeled structure of TLR8e. The five sites are located both within and on the surface of the ecto-domain. (B) Eight sites (blue) mapped onto TLR9b. The sites are mainly located to two clusters in the ecto-domain with one cluster right at the border towards the transmembrane domain and one cluster in the middle of the ecto-domain. The sites are located both within and on the surface of the structure. (C) One, three and four sites (green, red and blue, respectively) are mapped onto TLR22e. With the exception of one site at the tip of the ecto-domain, the sites are located to the first half of the ecto-domain, mainly on the outer surface of the ecto-domain surface. (D) Two sites (blue) mapped onto TLR25d located to the middle and within the ecto-domain.
transcriptome profiling of all TLRs expressed in the spleen/head kidney of healthy juvenile Atlantic cod. Here we found no evident need of increased gene dosage, however, it suggests more tissue-specific TLR and TLR paralog usage (Fig. 3). This is supported by TLR expression analyses by Sundaram et al. in Atlantic cod (including TLR22 paralogs) and by different expression levels of TLRs in various tissues in zebrafish and chicken.

---

**Table 1. Sites under diversifying selection as reported by SLAC, FEL and REL analyses.** Sites reported that are common between FEL and REL. Sites reported that are common between all, SLAC and FEL or FEL and REL respectively.

| Analysis | TLR8 | TLR9 | TLR22 | TLR25 |
|----------|------|------|-------|-------|
| SLAC     | 0    | 0    | 3     | 0     |
| FEL      | 5    | 9    | 27    | 2     |
| REL      | 0    | 44   | 7     | 0     |
| Common sites | 0 | 8' | 1/3/4'' | 0 |

---

**Figure 6.** The mammalian TLR signaling pathway as depicted in KEGG condensed and presented to fit the proposed situation in Atlantic cod. Ligands are: PG – peptidoglycan (gram positive bacteria), LP – lipoprotein, LA – lipoarabinomannan, Z – zymosan (yeast), LPS – lipopolysaccharide (gram negative bacteria), G – gram negative bacteria, F – flagellin, CpG – unmethylated CpG DNA from bacteria. TLR1/6, TLR2, TLR4 and TLR5 are not found in Atlantic cod (also see Figs. 1 and 7). The presence of CD14, LY96 and CD80/86 was difficult to determine and are thus marked as putative. TLR14, TLR21, TLR22, TLR23 and TLR25 have unknown signaling pathways, but are drawn at their most likely affiliated membranes with the exception of TLR22 drawn at the plasma membrane due to incongruent reports.
Teleost TLR repertoires are more diverse compared to other vertebrates. Our phylogenetic analysis of vertebrate TLRs revealed substantial variation in TLR repertoires. All investigated fish species, except zebrafish, lack representatives of the TLR4-family. TLR5 is not found in lamprey and Atlantic cod and TLR22 is lost in zebrafish (Figs 2 and 7 and Table 2). In contrast, certain TLRs are only present in a few species independent of phylogenetic relationships - i.e. TLR13, TLR23, TLR25 and TLR26. With regard to the gene expansions observed, duplications seems to be more frequent within teleosts and less frequently occurring in other vertebrate lineages (Fig. 7 and Table 2). This pattern may be connected to the teleost genome duplication event where a causal connection between gene/genome duplication and subsequent neofunctionalization of paralogs has been established in contrast to the usual reciprocal loss of gene duplicates43. This is also in line with the sites under diversifying selection detected in the Atlantic cod TLR expansions (Table 1). Our data also demonstrate that TLR14 is lost from birds and humans and that humans lack the entire TLR11-family. Notably, the TLR diversity and phylogeny suggest that life history strategies involving aquatic life stages require a different array of TLR11-family members and additional TLRs from the TLR1-family (Fig. 7 and Table 2). Thus, the transition from
an aquatic to a terrestrial lifestyle is associated with a shift in TLR repertoires – a shift that likely is linked to a highly different selection pressure on TLRs in the two environments.

**The birth-and-death of TLRs.** Multigene families connected to the immune system tend to follow a birth-and-death (BD) evolutionary model promoting diversification that manifests as general phylogenetic interspecific gene clustering patterns, the presence of pseudogenes and gene losses. Furthermore, gene expansions subjected to BD evolution and strong purifying selection undergo functional differentiation of the paralogs via sub- or neofunctionalization. TLRs in general and especially their TIR-domains and leucine-rich repeat elements are known to be under strong purifying selection. Our vertebrate TLR phylogeny demonstrates that gene losses and expansions are common in most lineages. However, the pattern is less pronounced in non-teleost lineages. Among teleosts, Atlantic cod shows the most pronounced loss and expansion pattern (Fig. 7 and Table 2). The BD model further supports our finding that sites under diversifying selection within TLR8 and TLR22 (and possibly TLR9 and TLR23) in Atlantic cod (Table 1) likely increase the detectable ligand repertoire in this species. Finally, the extreme case of Atlantic cod compared to other teleosts indicates that its TLR repertoire is associated with the loss of MHC-II, i.e. that the loss of such a major adaptive immune system component has boosted evolutionary innovation through interlinked gene losses and expansions leading to high complexity and greater relative dependence on the innate immune system in this species.

**Materials and Methods**

**Genome assembly.** The genome assembly used in this study is one of four assemblies used to produce a new release of the Atlantic cod genome (Tørresen & Nederbragt et al. in prep). In short, overlapping sequencing reads from Illumina (180 bp insert size, 100 nt read length) were merged with FLASH using default options. Meryl and merTrim were used to count and correct the reads, both programs from the Celera Assembler package 8.1. 454 reads used in Star et al. were converted from .sff files with sffToCA (also from Celera Assembler package) and corrected with merTrim, before trimmed with overlap based trimming (OBT, Celera Assembler program). Celera Assembler 8.2 alpha was used to trim subreads of PacBio sequencing reads. 20x of the merged Illumina 180 bp insert size reads, all paired 454 reads and the trimmed PacBio reads were used in an assembly with the Celera Assembler. The resulting genome assembly had some gaps closed with PBJelly and was polished by Pilon. Details are available upon request and later in Tørresen & Nederbragt et al. (in prep).

**Genome mining for PRRs.** We searched for PRR genes representing the major PRR families known in mammals listed in Supplementary Table 1 collected from Ensembl and UniProt. The search was performed using TBLASTN from the BLAST+ suite with an e-value cut-off of 1e−10. The low e-value was used to capture distant sequence homologs. Homologous relationships are described in Supplementary Table 1.

**Selection of full-length TLR genes for further analyses.** Annotated TLR sequences from selected species in Ensembl and GenBank covering all known TLR genes (listed in Supplementary Table 2) were compared towards the Atlantic cod genome using TBLASTN from the BLAST+ suite with an e-value cut-off of 1e−10 and otherwise default parameters. All putative contigs containing TLRs were loaded into MEGA5 where regions of interest in each scaffold were extracted. Only full-length TLRs containing a complete ecto-domain, transmembrane domain, linker and complete TIR-domain were evaluated further. We performed RNAseq to evaluate expression levels as some of the full-length TLRs extracted contained several insertions and deletions making poor translated protein sequences. All extracted full-length TLRs were used to make an Atlantic cod TLR index. The quality and adapter trimmed RNAseq sequences from six healthy juvenile Atlantic cod (see RNAseq method section) were mapped towards this database and raw counts extracted using the RSEM/Bowtie wrapper included in Trinity v2.0.6. These raw counts were normalized using the included edgeR scripts in Trinity to obtain TMM normalized FPKM counts. TLRs with large amounts of insertions/deletions, either alone or in combination with low read counts, were excluded from further analysis as the accuracy of the translate protein sequences was questionable (Supplementary Table 4). Count matrix is available in the GitHub repository (https://github.com/uio-cels/Solbakken_TLRs).

**Fish and totalRNA isolation for RNA sequencing.** Total RNA was isolated from the head kidney/spleen of six healthy juvenile Atlantic cod. These fish originate from the Norwegian cod breeding program and were reported to be healthy without any history of diseases. The use of live Atlantic cod was approved by the Norwegian Animal Research authority in Norway (FOTS id 1147) and all methods were in accordance with the approved guidelines. The fish were transported at approx. 2 g to 100 L tanks at the Aquaculture Research Station (Tromsø, Norway) for grow-out in seawater of 3.4% salinity at 10°C, 24-hour light and fed ad libitum with commercial feed (BioMar, Norway). The rates of water inflow were adjusted to an oxygen saturation of 90–100% in the outlet water. The tissue was stored on RNAlater (Life Technologies) and total RNA was isolated using Trizol (Life Technologies) according to protocol but using half the amount of tissue per volume Trizol recommended by the manufacturer. The complete laboratory protocol is available in the GitHub repository (https://github.com/uio-cels/Solbakken_TLRs). Sequencing libraries were produced according to the IlluminaTruSeq protocol (Illumina, Inc., San Diego, CA). Illumina HiSeq2000 100 bp paired-end sequencing services were provided by the Norwegian Sequencing Centre (http://www. sequencing.uio.no). Sequences were trimmed for adapters using Cutadapt v1.0 and trimmed on quality using Sickle using known Illumina adapter sequences, a Q threshold of 20 and otherwise default parameters.

**Synteny analyses.** The Ensembl genome browser v78 (unless otherwise stated) was used to chart annotated open reading frames around TLRs annotated in the selected fish species. Protein sequences from these genes...
were downloaded and used in a TBLASTN towards the Atlantic cod genome together with TLR representatives with an e-value cut-off of 1e\(^{-10}\). If a certain TLR was not annotated in one or several of the selected fish genomes in Ensembl we used the Ensembl BLAST tool with protein queries towards nucleic acid resources (TBLASTN) with default parameters to find the genomic region of interest. Some genome regions were reverse complemented for figure drawing purposes and this is noted in the respective figures (Figs 1 and 2).

**Endolysosomal sorting signals.** Characterized TLR sorting signals were obtained from the literature\(^{22,23}\). Protein sequence was obtained for all TLR3 and all TLR9 genes investigated in this study (Supplementary Table 2). These were aligned with default settings using MEGA5 and ClustalW (Fig. 4A)\(^{57}\). We also searched for similar tyrosine based signals in the linker region of the remaining Atlantic cod TLRs (TLR7, TLR8, TLR14, TLR21, TLR22, TLR23 and TLR25) (Fig. 4B).

**TLR signalling pathway.** The mammalian TLR signalling pathway available through the KEGG database\(^62\) was used as a basis for mapping the pathway components in the Atlantic cod genome. The connected UniProt sequences for each pathway component were used in a TBLASTN search together with annotated homologs from fish species available at Ensembl or UniProt (Supplementary Table 5) towards the Atlantic cod genome with an e-value cut-off of 1e\(^{-15}\). The low e-value was used due to distant homology of sequences between fish and mammals. Genes that were difficult to verify are highlighted in Fig. 6.

**Protein structure prediction.** Translated Atlantic cod TLR sequences were submitted to the Phyre2 structure prediction server for modelling\(^63\). All sequences were modelled against TLR5. All TLRs from Homo sapiens (human), Petromyzon marinus (lamprey), Anolis carolinensis (lizard) and Oreochromis niloticus (tilapia) were also submitted to Phyre2 and modelled onto the human TLR5 crystal structure (Fold library id: c3j0aA). The structures were coloured for visualization purposes using Jmol\(^{64}\), differentiating between loops, sheets and helices.
as well as the transmembrane, linker and TIR-domain (Supplementary Table 2 and Supplementary Figs 1–3).

All Atlantic cod PDB files are available in the GitHub repository (https://github.com/uio-cels/Solbakken_TLRs).

Selection analyses. The expanded Atlantic cod TLRs with three or more full-length copies (TLR8, TLR9, TLR22 and TLR25) were analyzed using Datamonkey65. Nucleotide sequences were imported into MEGA5 for alignment using default ClustalW parameters. The alignment was then manually edited to ensure proper translation to amino acids. A maximum likelihood phylogeny was made using partial deletion, a Jukes-Cantor model of sequence evolution with gamma distributed rate heterogeneity57. The resulting phylogeny was submitted together with the nucleotide alignment to Datamonkey. For each TLR expansion a model test was first run. The proposed best model was used before running selection analyses with the SLAC, FEL and REL methods. These are codon based maximum likelihood methods estimating rates of nonsynonymous and synonymous changes at each site in an alignment to identify sites under positive or negative selection. These tests are originally designed to be run on interspecies alignments. Here, since the tests are run on intraspecies paralogs, we argue that the sites reported to be under positive selection actually are under diversifying selection. The term diversifying selection is thus used throughout this report. Fixed effects like-hood model (FEL) estimates the ratio of nonsynonymous to synonymous substitution rates for each site in a sequence alignment with fixed estimates for branch lengths and substitution rate bias parameters. Random effects like-hood model (REL) allows rate variation in both nonsynonymous and synonymous rates and a general underlying nucleotide substitution model. Single-likelihood ancestor counting (SLAC) model weights the nucleotide substitution biases which are estimated from the data and allow ambiguous codons in the data. Sites reported to be under diversifying selection in one or more tests are highlighted in one of the protein structure models made for each of the TLR8, TLR9, TLR22 and TLR25 expansions. In cases where only one test has reported sites it is noted in the Fig, legend (Table 1 and Fig. 5). Phylogenies and alignments are available in the GitHub repository (https://github.com/uio-cels/Solbakken_TLRs).

Vertebrate TLR phylogeny. Full-length protein sequences were not alignable due to large variations in the ecto-domain of the TLRs. Thus, the transmembrane region, linker and TIR-domain were used as basis for phylogenetic analysis after alignment and minor curation of the data using MEGA557. PROTTTEST66 was used for substitution model optimization with the Bayesian Information Criterion (BIC) model selection criterion and testing all seven models available. PROTTTEST suggested the JTT+I+G+F as the best substitution model. A maximum likelihood tree was produced using Randomized Axelerated Maximum Likelihood (RAxML) HPC-PTHREADS version 7. 2. 6 with the PROCATJTT model52. The rapid bootstrap/search for the best tree simultaneously option was used and the analysis was run with 500 bootstraps. The resulting phylogeny was used as the basis for the final TLR annotations of all sequences used and described in this study (Supplementary Table 2). The tree was imported into FigTree v.1.4.6 for cladogram transformation and then edited in Adobe Illustrator for improved Fig. visualization (Fig. 7). The alignment is available in the GitHub repository (https://github.com/uio-cels/Solbakken_TLRs).

References
1. Buonocore, F. & Gerdol, M. Alternative adaptive immunity strategies: coelacanth, cod and shark immunity. Mol Immunol, doi: 10.1016/j.molimm.2015.09.003 (2015).
2. Zhu, L. Y., Nie, L., Zha, G., Xiang, I. X. & Shao, J. Z. Advances in research of fish immune-relevant genes: a comparative overview of innate and adaptive immunity in teleosts. Dev Comp Immunol 39, 39–62, doi: 10.1016/j.dci.2012.04.001 (2013).
3. Magnadottir, B. Innate immunity of fish (overview). Fish Shellfish Immunol 20, 137–151, doi: 10.1016/j.fsi.2004.09.006 (2006).
4. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. Cell 140, 805–820, doi: 10.1016/j.cell.2010.01.022 (2010).
5. Kawasaki, T. & Kawai, T. Toll-like receptor signaling pathways. Frontiers in immunology 5, 461, doi: 10.3389/fimmu.2014.00461 (2014).
6. Satake, H. & Sekiguchi, T. Toll-like receptors of deuterostome invertebrates. Frontiers in immunology 3, 34, doi: 10.3389/ fimmu.2012.00034 (2012).
7. Pati, Y. Toll-like receptors in bony fish: from genomics to function. Dev Comp Immunol 35, 1263–1272, doi: 10.1016/j.dci.2011.03.006 (2011).
8. Roach, J. C. et al. The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci USA 102, 9577–9582, doi: 10.1073/ pnas.0502272102 (2005).
9. Rauta, P. R., Samanta, M., Dash, H. R., Nayak, B. & Das, S. Toll-like receptors (TLRs) in aquatic animals: signaling pathways, expressions and immune responses. Immunol Lett 158, 14–24, doi: 10.1016/j.imlet.2013.11.013 (2014).
10. Boehm, T. & Swann, J. B. Origin and evolution of adaptive immunity. Annual review of animal biosciences 2, 259–283, doi: 10.1146/ annurev-animal-022513-114201 (2014).
11. Haase, D. et al. Absence of major histocompatibility complex class II mediated immunoresistance in pipefish, Syngnathus typhle: evidence from deep transcriptome sequencing. Biology letters 9, 20130044, doi: 10.1098/rsbl.2013.0044 (2013).
12. Star, B. et al. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477, 207–210, doi: 10.1038/nature10342 (2011).
13. Malmstrom, M., Jentoft, S., Gregers, T. F. & Jakobsen, K. S. Unraveling the evolution of the Atlantic cod’s (Gadus morhua L.) alternative immune strategy. PLoS One 8, e74004, doi: 10.1371/journal.pone.0074004 (2013).
14. Boehr, T. et al. The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci USA 102, 9577–9582, doi: 10.1073/ pnas.0502272102 (2005).
15. Malmstrom, M., Jentoft, S., Gregers, T. F. & Jakobsen, K. S. Unraveling the evolution of the Atlantic cod’s (Gadus morhua L.) alternative immune strategy. PLoS One 8, e74004, doi: 10.1371/journal.pone.0074004 (2013).
16. Star, R. & Jentoft, S. Why does the immune system of Atlantic cod lack MHC II? Bioessays 34, 648–651, doi: 10.1002/bies.201200005 (2012).
17. Saleh, M. The machinery of Nod-like receptors: refining the paths to immunity and cell death. Immunological Reviews 243, 235–246, doi: 10.1111/j.1600-065X.2011.01045.x (2011).
18. Zhang, J. et al. Toll-like receptor recognition of bacteria in fish: ligand specificity and signal pathways. Fish Shellfish Immunol 41, 380–388, doi: 10.1016/j.fsi.2014.09.022 (2014).
19. Hwang, S. D., Kondo, H., Hirono, I. & Aoki, T. Molecular cloning and characterization of Toll-like receptor 14 in Japanese flounder, Paralichthys olivaceus. Fish Shellfish Immunol 30, 425–429, doi: 10.1016/j.fsi.2010.08.005 (2011).
20. Zhao, F. et al. Expression profiles of toll-like receptors in channel catfish (Ictalurus punctatus) after infection with Ichthyophthirius multifiliis. Fish Shellfish Immunol 35, 993–997, doi: 10.1016/j.fsi.2013.05.023 (2013).
61. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Bioinformatics in Action* **17**, 10–12, doi: 10.1186/bte-11853772 (2012).
62. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
63. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**, 845–858, doi: 10.1038/nprot.2015.053 (2015).
64. Jmol: an open-source Java viewer for chemical structures in 3D, http://www.jmol.org/.
65. Delport, W., Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics (Oxford, England)* **26**, 2453–2457, doi: 10.1093/bioinformatics/btq429 (2010).
66. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* **21**, 2104–2105, doi: 10.1093/bioinformatics/bti263 (2005).
67. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)* **22**, 2688–2690, doi: 10.1093/bioinformatics/bti446 (2006).
68. FigTree http://tree.bio.ed.ac.uk/software/figtree/ (2015).

**Acknowledgements**

This work was supported by The Research Council of Norway (Grant number 199806/S40 to KSJ). The genome assembly was made using the Abel Cluster, owned by the University of Oslo and the Norwegian metacentre for High Performance Computing (NOTUR), and operated by the Department for Research Computing at USIT, the University of Oslo IT-department. http://www.hpc.uio.no/. The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities. We would like to thank Dr. Simon MacKenzie for valuable comments and discussions.

**Author Contributions**

M.H.S. performed all analyses. O.K.T. and A.J.N. made and provided the new version of the Atlantic cod genome. MS provided material for RNAseq. M.H.S., T.F.G., S.J. and K.S.J. interpreted the results. M.H.S. wrote the main text and prepared all figures with the assistance of T.F.G., S.J. and K.S.J. All authors contributed to review of the manuscript.

**Additional Information**

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Solbakken, M. H. et al. Evolutionary redesign of the Atlantic cod (Gadus morhua L.) Toll-like receptor repertoire by gene losses and expansions. *Sci. Rep.* **6**, 25211; doi: 10.1038/srep25211 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/