# Linking species habitat and past paleoclimatic events to evolution of the teleost innate immune system

Monica Hongrø Solbakken[1*], Kjetil Lysne Voje[1], Kjetill Sigurd Jakobsen[1], Sissel Jentoft[1,2]*

[1] Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway

[2] Department of Natural Sciences, University of Agder, Kristiansand, Norway

*Corresponding authors:

Monica Hongrø Solbakken. Department of Biosciences, Centre for Ecological and Evolutionary Synthesis, P.O. Box 1066, Blindern, 0316 Oslo, Norway. m.h.solbakken@ibv.uio.no, +47 22 84 41 64

Sissel Jentoft. Department of Biosciences, Centre for Ecological and Evolutionary Synthesis, P.O. Box 1066, Blindern, 0316 Oslo, Norway. sissel.jentoft@ibv.uio.no, +47 22 85 72 39

## Abstract

Host-intrinsic factors as well as environmental changes are known to be strong evolutionary drivers defining the genetic foundation of immunity. Using a novel set of teleost genomes and a time-calibrated phylogeny, we here investigate the family of Toll-like receptors (*TLRs*) and address the underlying evolutionary processes shaping the diversity of the first line defence. Our findings reveal remarkable flexibility within the evolutionary design of teleost innate immunity characterized by prominent *TLR* gene losses and expansions. In the order of Gadiformes, expansions correlate with the loss of Major Histocompatibility Complex class II (*MHCII*) and diversifying selection analyses support that this has fostered new immunological innovations in *TLR* within this lineage. In teleosts overall, *TLR* expansions correlate with species latitudinal distributions and maximum depth. In contrast, lineage-specific gene losses overlap with well-described changes in paleoclimate (global ocean anoxia) and past Atlantic Ocean geography. In conclusion, we suggest that the evolvability of the teleost immune system has most likely played a prominent role in the survival and successful radiation of this lineage.

Keywords: Adaptive evolution, innate immunity, Toll-like receptors, gene loss, gene expansion, past climatic change.

## Background

The evolutionary success of ray-finned fish (class Actinopterygii) is characterized by large species radiations [1]. Actinopterygii comprises an exceptionally diverse group of fish with species inhabiting numerous aquatic habitats spanning from Arctic to Antarctic oceans, deep-sea benthos to the shore, along coastlines and rivers as well as freshwater systems. Moreover, the high degree of diversity is mirrored in the array of life history strategies, morphological varieties,

distinct migratory behaviour and reproductive strategies displayed [2, 3 - and references therein]. The teleost lineage is the largest within the class of ray-finned fish [4]. Genome sequencing efforts of non-model organisms have provided new insight into the extreme diversity of the teleost lineage including evidence for several alternate immunological strategies. The discoveries of the genetic loss of the Major Histocompatibility (*MHC*) class II pathway in Atlantic cod (*Gadus morhua*) as well as the functional loss in the more distant broadnosed pipefish (*Syngnathus typhle)* [5, 6], show that *MHCII* is not crucial for the defence against pathogens and survival in some fish species. These findings are further supported in a recent study by Malmstrøm et al., which demonstrated that the loss of *MHCII* is shared by the entire Gadiformes lineage [7]. Accompanying the loss of *MHCII,* highly variable *MHCI* copy number within the Gadiformes was reported, with several species having more than 40 copies including Atlantic cod found to have 80-100 copies [7, 8]. Furthermore, it was hypothesized that the expanded repertoire of *MHCI* had undergone sub- or neofunctionalization as a possible adaptation to the *MHCII* loss. However, Malmstrøm et al. also identified large numbers of *MHCI* in many Percomorphaceae lineages (all containing *MHCII*) demonstrating an extreme evolutionary plasticity of teleost immunity, and that it most likely is influenced by species habitat. Additional analyses revealed a correlation between high *MHCI* copy number and elevated speciation rates, and thus being a key to the success of this group of fishes [7].

The teleost immune system also displays important strategies with respect to the innate immune system such as the alternative set of *Toll-like receptors (TLR)* compared to other vertebrates [9-11]. Again, Atlantic cod is reported to be divergent compared to other investigated teleosts. In a recent study, the *TLR* repertoire in Atlantic cod was characterized and compared to that of other genome-sequenced fish species, revealing that Atlantic cod displays large gene expansions and

several gene losses. These findings were attributed to the loss of *MHCII,* which may have boosted evolutionary innovation leading to a more complex *TLR* repertoire [12].

In general, it is the genetic basis of teleost alternative immunological strategies that has been investigated and studies beyond this tend to focus on genes related to the adaptive immune system. However, the underlying selective mechanisms driving the variety of immunological strategies observed and why they arose are poorly understood – especially for the innate immune system. Using genome assemblies from 66 teleost species our aim was to characterise teleost *TLRs* with emphasis on the Gadiformes lineage and thereby investigate the possible link between the loss of *MHCII*, past and present environmental conditions and the genetic architecture of the innate immune system. We show that the teleost *TLR* repertoire contains an array of lineage-specific losses and expansions, with the Gadiformes lineage as an extreme outlier. Importantly, within the Gadiformes we discovered expansions of *TLR* genes to be correlated with the loss of *MHCII* and to display different patterns of selection. Furthermore, in teleosts overall, we found that *TLR* copy number variation correlated with species latitudinal distribution in teleosts overall. In contrast, a weak correlation was found with species maximum depth for *TLR9* and *TLR22*. This suggests that there is a strong on-going selection of the innate immune system linked to specific environmental and host intrinsic factors. Furthermore, timing of the lineage-specific losses overlaps with well-described changes in paleoclimate and continental drift, and hence unveils past adaptive signatures driving the genetic change within the teleost immune system. Our study reveals a remarkable evolutionary flexibility of teleost innate immunity, which has played an essential role in the survival and radiation of the teleost lineage.

## Materials and methods

### Sequencing and assembly summary

The 66 teleost genomes and species phylogeny were generated by Malmstrøm et al. [7, 13]. In short DNA was isolated from 66 teleost species and subjected to Illumina HiSeq sequencing (2 x 150 bp paired-end reads) which after trimming resulted in an overall coverage between 9 and 34X. The genomes were assembled using the Celera Assembler. For the phylogenetic reconstruction 9 reference fish species were added from Ensembl together with *Salmo salar*. An alignment of 71,418 bp was used as input for phylogenetic reconstruction with the Bayesian software BEAST [14]. The phylogeny was made using the Bayesian software BEAST combined with fossil time-calibration. Note: all timings derived from the phylogeny presented in this study includes the confidence interval to illustrate the uncertainty underlying the time fossil calibration performed by Malmstrøm et al. [7], and thus spans a longer time period than the branches depicted in the phylogeny (Figure 1).

### Gene searches

Protein query TIR (Toll/interleukin-1 receptor) domain sequences from Atlantic cod [12], all fish genomes available at Ensembl [15] and channel catfish [16-18], collectively representing all known vertebrate *TLR* genes to date, were used for TBLASTN searches towards the 66 fish genomes supplied by Malmstrøm et al (see below for parameters). *TLR* copy numbers for the Ensembl species were taken from [12]. The NCBI BLAST tool was used to search the *Salmo salar* genome (ICSASG_v2, GCA_000233375.4) with default settings using the same query sequences. TBLASTN from BLAST+ 2.2.26 [19] was used with an e-value cut-off at 1e-10 (and in some cases lower, to capture the largest expansions), otherwise default settings. The number of detected TIR domains was counted for each *TLR* gene. Due to the fragmented nature of the

genomes, conservative estimates of copy numbers were used and are shown in Supplementary table 1. These copy numbers form the foundation for the *TLR* repertoires depicted in Figure 1.

Note on gene annotation: *TLR* gene annotation varies greatly between species. In this study the following annotations are used (similar to that of [12]): *TLR1*, *TLR1/6* (in cases where annotation has not been provided and phylogeny cannot determine stronger homology towards *TLR1* or *TLR6*), *TLR2*, *TLR3*, *TLR4*, *TLR5*, *TLR6*, *TLR7*, *TLR8*, *TLR9*, *TLR10*, *TLR11*, *TLR12*, *TLR13*, *TLR14*, *TLR15*, *TLR16*, *TLR18* is by phylogeny determined to be *TLR14*, *TLR15*, *TLR16*, *TLR19* is by phylogeny determined to be *TLR26*, *TLR20* is by phylogeny determined to be *TLR26*, *TLR21*, *TLR22*, *TLR23*, *TLR25* and *TLR26*.

**_TLR_, _MHC_, latitude and depth correlations using SLOUCH**

For genes displaying more than four different gene copy numbers (*TLR8*, *TLR9*, *TLR22*, *TLR23*, *TLR25*) we ran SLOUCH — Stochastic Linear Ornstein-Uhlenbeck Models for Comparative Hypotheses. This is a phylogenetic comparative method designed to study adaptive evolution of a trait along a phylogeny implemented in the R program SLOUCH [20-22]. The output of models analysed in SLOUCH can be summarized by a regression, which includes information on whether the analysed traits are evolving towards the estimated optima, how fast (or slow) this evolution is, and how much of the trait variation that is explained by evolution towards these optima. We used SLOUCH to test whether *TLR* copy numbers have evolved towards optima that are influenced by the species' latitudinal distribution (values obtained from Fishbase.org [23]), species maximum depth (values obtained from Fishbase.org [23]) and evolutionary loss of the *MHCII* complex. We defined 6 latitudinal categories: 75, 50, 25, 0 (equator), -25 and -50. If a species' latitudinal distribution includes or crosses one of these categories it was assigned to that respective category (multiple assignments are possible). Some species were not included in any

126 of the categories due failure to cross the defined latitudes. Similarly, where data on depth was
127 unavailable, species were excluded from the phylogeny resulting in a reduced tree used as input
128 for SLOUCH.

129 The model of evolution in SLOUCH is based on an Ornstein–Uhlenbeck process and assumes
130 that a trait (e.g. gene copy number) has a tendency to evolve towards a 'primary' optimum $\Theta$. We
131 assume that average copy number in a lineage can take any non-negative real number (i.e.,
132 intraspecies variation in copy numbers exist). A primary optimum is defined as the average
133 optimal state that species will reach in a given environment when ancestral constraints have
134 disappeared [20], at a rate proportional to a parameter $\alpha$. As an example, in some of our analyses,
135 we investigated whether species sharing the same latitudinal distribution have a tendency to
136 evolve similar copy numbers for a given *TLR* locus. Lag in adaptation towards primary optima is
137 quantified by a half-life parameter, $t_{1/2} = \ln(2)/\alpha$, which can be interpreted as the average time it
138 takes a species to evolve half the distance from the ancestral (copy number) state towards the
139 predicted optimal (copy number) state. For example, a half-life of zero signifies immediate
140 adaptation of the trait to any change in the optimum for every lineage present in the phylogeny. A
141 half-life above zero indicates adaptation is not immediate, with the amount of constrained
142 evolution increasing with an increasing half-life. The model of evolution used in SLOUCH also
143 includes a stochastic component with standard deviation $\sigma$, which can be interpreted as
144 evolutionary changes in the trait (e.g. copy numbers) due to unmeasured selective forces and
145 genetic drift. This component of the model is reported as $v_y = \sigma^2/2\alpha$, and can be interpreted as the
146 expected residual variance when adaptation and stochastic changes have come to an equilibrium.

147 Our latitudinal categories, maximum depth and evolutionary losses of *MHCII* represent 'niches'
148 and the model estimates one primary optimum for each niche included in any particular model.

149 The different states of niches (e.g. presence and absence of *MCHII*) are known for all extant
150 species in our phylogeny, but are unobserved for internal branches in the tree. We therefore
151 mapped a separate state called *ancestral* to all internal nodes in the phylogeny to avoid having to
152 infer uncertain primary optima. The method uses generalized least squares for estimation of the
153 regression parameters (i.e., the influence of the predictor on the primary optimum) and maximum
154 likelihood for estimation of $\alpha$ and $\sigma^2$ in an iterative procedure. For a full description of the model
155 implemented in SLOUCH, see Hansen et al. 2008. All analyses were performed in R version 3.0
156 [22].

157 We used SLOUCH to estimate the phylogenetic effect in the data. A phylogenetic effect indicates
158 that some part of the variation in the trait is explained by shared ancestry (i.e. phylogeny), which
159 means closely related species tend to have more similar trait values compared to more distantly
160 related species. The phylogenetic effect can be estimated in SLOUCH by running a model
161 without any predictor variables (i.e. no latitudinal categorical variables). The half-life parameter
162 in such a model will represent an estimate for how important shared history is in explaining the
163 distribution of trait means (average) on the phylogeny: A half-life of zero indicates that the trait
164 data is not phylogenetically structured, while a half-life > 0 indicates that there exists an
165 influence of phylogeny on the data. A phylogenetic effect can be due to slowness of adaptation,
166 adaptation towards phylogenetically structured optima, or a combination of both. To investigate
167 which of these scenarios we find support for, we contrasted the phylogenetic effect model with a
168 model run with predictor variables (e.g. latitudinal distribution or maximum depth) using the
169 bias-corrected Akaike Information Criterion (AICc), which balances goodness of fit (log-
170 likelihood) with the number of parameters in the model (model complexity). The model with the
171 lowest AICc value is the best supported. A better (lower) AICc value for a model including

predictor variables indicate evidence for a scenario where the traits in our models are evolving towards optima that are shared by species across niches (e.g. the same latitudinal section). $R^2$ was not used for assessing model support, but represents the amount of the total variation in the response trait (TLR genes copy number) that is explained by the optimal regression.

**Diversifying selection analysis using MEME and BSR**

As there were different degrees of *TLR* gene expansions throughout our dataset, and because expansions were more prominent within the Gadiformes order, we wanted to determine if any individual positions within the coding sequence or certain lineages have been affected by diversifying selection. Due to the fragmented nature of our dataset this analysis was not feasible unless we selected a set of species as well as a set of *TLRs*. We selected 9 species from the draft genome dataset: *Melanogrammus aeglefinus*, *Macrourus berglax* and *Muraenolepsis marmoratus* from the Gadiformes, *Stylephorus chordatus* which is a putative ancestral clade of Gadiformes, *Cyttopsis roseus* and *Zeus faber* from the Zeiformes (Gadiformes + *Stylephorus chordatus* sister clade), *Polymixia japonica* at the base of the Paracanthopterygii superorder, *Rondeletia loricata* and *Beryx splendens* as two closely related species outside the Paracanthopterygii. We also included *TLR* sequences from the second version of the Atlantic cod genome (GadMor2) as an additional Gadiformes representative [24]. Finally, we added the respective *TLR* sequences from fish species whose genomes are available through Ensembl [15]. Collectively, these species cover the entire range of the teleost phylogenetic tree obtained from Malmstrøm et al [7].

We selected three *TLR* genes for investigation: *TLR3* – a single copy gene present in all investigated teleosts, *TLR9* – expanded in most Gadiformes as well as present in all investigated teleosts and *TLR25* – mainly expanded in the C1 clade of the Gadiformes but also displaying both presence and absence patterns in our data. Collectively, these genes represent the range of

different patterns observed. Query *TLR* sequences were identical to those used for the overall *TLR* characterization described above except the full-length protein sequence was used in a TBLASTN with an e-value cut off = 1e-10 and otherwise default parameters towards the draft genomes and GadMor2. The target unitigs (draft genomes) and linkage group (GadMor2) regions were extracted and aligned towards the coding sequences obtained from Ensembl using CLUSTALW in MEGA5 [25]. The resulting alignment was manually curated to ensure that the reading frame was maintained. We chose to only investigate the ecto-domain of the TLR as the transmembrane and TIR domain are known to be under purifying selection. For all alignments the coverage of unitig sequence was variable. Therefore, the alignments were divided into sections to obtain alignments with the least amount of missing data. This resulted in one alignment for *TLR3*, two for *TLR9* and four for *TLR25*. The alignments are available in our GitHub repository.

These alignments were uploaded to www.datamonkey.org [26, 27] where we performed model selection analysis to find the best fitting model of nucleotide evolution for each of the alignments (reported in supplementary information). We then performed MEME (Mixed Effects Model of Evolution) analysis on all alignments as well as BSR (Branch-Site Random Effects Likelihood) analysis on *TLR9* and *TLR25* alignments allowing for the generation of gene trees based on the alignments. MEME is based on the ratio between non-synonymous to synonymous substitutions where this ratio can vary from site to site as well between lineages. In this way MEME can detect both pervasive and episodic positive (diversifying) selection. MEME compares its estimates to a null hypothesis for which all sites are evolving neutrally (worst case scenario) and thus the results given by MEME are conservative estimates. BSR is also based on the ratio between non-synonymous to synonymous substitutions. MEME implements this analysis for each individual site, but we also ran BSR alone to obtain on overall impression of any likely diversifying

selection affecting lineages or individual branches. In contrast, in BSR there is no need to define any branches a priori as neutral or under negative selection. Thus, detecting episodic diversifying selection in a few sites or in a few lineages becomes more reliable by using BSR [28, 29].

## Results

Mapping all the identified teleost *TLRs*— extracted from the 66 genome assemblies – onto the phylogeny of Malmstrøm et al [7] demonstrates the presence of comprehensive *TLR* repertoires in all investigated teleosts (Figure 1) similar to that found in other vertebrates [9, 11, 12]. However, most notable was the observation of three lineage-specific gene losses, several lineage-specific gene expansions and a substantial number of recorded species-specific repertoire variants (Figure 1). Specifically, *TLR1/2* are lost from the Gadinae (40-16 mya) in addition to being completely or partially lost in *Bregmaceros cantori*, *Benthosema glaciale*, *Stylephorus chordatus and Guentherus altivela*. *TLR5* is lost from the entire Paracanthopterygii superorder and the order Lampridiformes (175-130 mya) in addition to *Pseudochromis fuscus*. Further, we discovered a new *TLR*, here annotated as *TLR21beta* based on sequence homology, which is also absent in all Paracanthopterygiian species with the exception of *Polymixia japonica,* and Lampridiformes. However, in contrast to *TLR5*, the presence of *TLR21beta* does not follow any clear phylogenetic pattern outside Paracanthopterygii/Lampridiformes (Figure 1). The Gadinae is the only clade consistent with the recently reported alternative *TLR* repertoire in Atlantic cod [5, 12] due to the prominent gene losses of *TLR1/2*.

Three *TLRs* are found in all species; *TLR3*, *TLR14* and *TLR21*, the latter with the exception of *Benthosema glaciale*. Within the Gadiformes we find gene expansions for *TLR7*, *TLR8*, *TLR9*, *TLR22*, *TLR23* and *TLR25*, especially within the C1 clade (the Gadiformes segregate into two distinct clades here named C1 and C2, see Figure 1). Outside the Gadiformes the presence of

*TLR25* displays no obvious phylogenetic pattern. This is in contrast to *TLR7*, *TLR8* and *TLR9* which are present in all species with the exception of a single *TLR8* loss in *Guentherus altivela*. *TLR22* and *TLR23* are found in all Gadiformes except in *Bregmaceros cantori* and show a substantial degree of gene expansion within the Gadiformes lineage – particularly for *TLR22*. Outside the Gadiformes, the expansion of *TLR22* is less pronounced whereas, in contrast, *TLR23* is frequently expanded. However, *TLR22* and *TLR23* display phylogenetically non-structured patterns of presence and gene loss outside the Gadiformes order (Figure 1, Supplementary table 1). Finally, there are two rare teleost *TLRs*, i.e. – *TLR4* and *TLR26*. *TLR4* is found in the Holocentriformes and in 3 out of 4 Beryciformes species in addition to *Danio rerio*, *Polymixia japonica* and *Guentherus altivela*. *TLR26* is mainly found in species basal to the Gadiformes and in two Beryciformes: *Rondeletia loricata* and *Beryx splendens* (Figure 1, Supplementary table 1).

To identify episodic diversifying selection, MEME and BSR selection analyses were performed on the ecto-domain of three *TLR* representatives – *TLR3*, *TLR9* and *TLR25*. MEME reported 19 sites for *TLR3*, 35 sites for *TLR9* and 18 sites for *TLR25* likely to have experienced diversifying selection (Figure 2). The BSR analysis identified multiple nodes and branches encompassing most *TLR9* paralogs in the Gadiformes (mainly Gadinae) subject to diversifying selection. Diversifying selection was also detected in one of the *TLR25* alignments at one node and on one branch encompassing some of the Gadinae *TLR25* paralogs (Supplementary figure 1 and 2).

Associations between specific *TLR* expansions, species latitudinal distributions, species maximum depth as well as the absence of *MHCII* – specific for the Gadiformes lineage (Figure 1) – were further investigated using Stochastic Linear Ornstein-Uhlenbeck Models for Comparative Hypotheses (SLOUCH) [21]. Models using the specified latitudinal categories as predictor variables showed that latitude explained 19-32 % of the *TLR* copy number variation for *TLR8*,

*TLR9*, *TLR22* and *TLR25* (Table 1) whereas species maximum depth explained 4-10 % of the variation seen in *TLR9* and *TLR22* (Supplementary information). Especially northern latitudinal categories were found to be associated with higher copy numbers of *TLR8*, *TLR22* and *TLR25*, while increased copy numbers of *TLR9* were associated with more tropical latitudes – particularly in the equatorial region (Table 1, Supplementary table 1). However, for *TLR23* there was no indication that the copy number has evolved as a consequence of changes in latitude or depth (Table 1, data not shown for depth correlation). Moreover, within the Gadiformes lineage we found strong support for scenarios where *TLR8*, *TLR9*, *TLR22* and *TLR25* have evolved additional gene copies with the loss of *MHCII* explaining between 14-27 % of the copy number variation (Table 2). The explained variation in copy numbers was 3-6 % larger (compared to latitude alone) and 3-16 % larger (compared to *MHCII* loss alone) when we ran models where copy numbers of *TLR8*, *TLR22* and *TLR25* evolved towards optima jointly defined by latitudinal categories and presence/absence of *MHCII*. This indicates that both latitude and loss of *MHCII* have contributed to the expansion of these *TLRs*. However, we were not able to distinguish the relative contribution of *MHCII* and latitude, respectively. This is contrary to the striking result obtained for *TLR9* where the combination of latitude and loss of *MHCII* explained 50 % of the copy number variation – compared to 20 % and 22 % for latitude and *MHCII* loss separately (Table 2).

## Discussion

Overall, vertebrate and teleost genome duplications may explain some of the teleost *TLR* repertoire variation demonstrated here with respect to gene expansions. However, the extreme numbers seen for some of the *TLR* expansions within the Gadiformes indicate that these genes have undergone additional lineage-specific duplication events — a phenomenon also seen for

other genes in teleost species [30]. Gene duplicates preserved after a duplication event commonly undergo neo- or subfunctionalization [31 and references therein]. In Atlantic cod, we have previously demonstrated that the *TLR* expansions and their paralogs show signs of diversifying selection. For some expansions, this was indicative of neofunctionalization due to high numbers of sites under selection in putative dimerization and ligand-interacting regions. For other expansions it was more indicative of subfunctionalization due to fewer sites under selection combined with tissue-specific expression patterns [12]. The selection analyses on the chosen *TLR* representatives demonstrated that *TLR3* and *TLR25* display similar amounts of sites subject to diversifying selection despite their highly different patterns in our dataset (single copy present in all species versus expanded in Gadiformes combined with both presence and absence in the remaining species). In contrast, *TLR9* displayed almost double the number of sites reported as under diversifying selection (Figure 2). In the human system, and by proxy in teleosts, the TLR3 protein is located to the endosomal membranes and signals for an antiviral response upon interaction with double-stranded RNA (dsRNA) [32]. It has recently been demonstrated that mammalian TLR3 also can detect structured RNAs [33]. This could explain the presence of sites under diversifying selection in fish *TLR3* adapting the protein towards different structured RNAs or other possible ligands not presently known.

*TLR25* is a relatively newly identified fish-specific *TLR* where ligand and subcellular localization is yet to be determined [17]. We have earlier suggested that this TLR in Atlantic cod is located to the cell surface and interacts with ligands similar to other TLR1 family members (*TLR1*, *TLR2* and *TLR6*) – such as bacterial or parasitic lipoproteins [12]. In humans, TLR1, TLR2 and TLR6 form both homo- and heterodimers actively increasing their ligand repertoire [32]. Gadinae do not have *TLR1*, *TLR2* or *TLR6* (Figure 1) and thus, in their case *TLR14* and the expanded *TLR25*

may be a replacing set of TLR1 family members. In the Gadiformes, both *TLR1*, *TLR2* as well as *TLR14* and *TLR25* are present (Figure 1). The MEME analysis reported a similar number of sites under diversifying selection compared to *TLR3,* which could suggest that they are subjected to similar selective pressures. However, the BSR analysis indicated that nodes and branches representing only some of the Gadinae paralogs are subject to diversifying selection (Supplementary figure 2). Overall, this demonstrates that *TLR25* paralogs may be affected by different selection pressures within expansions whereas *TLR25* generally is adapted towards unknown species-specific factors.

In Atlantic cod, *TLR9* paralogs showed clear signs of diversifying selection and differences in expression patterns [12]. The MEME analysis reported a large amount of sites under diversifying selection and the BSR analysis strongly indicates diversifying selection on nodes and branches leading to different clades of Gadinae *TLR9* paralogs. In humans, TLR9 interacts with unmethylated single-stranded CpG DNA, both viral and bacterial, within the endosomal track in a highly sequence-dependent manner. However, dependent on the sequences, TLR9-ligand interaction can result in both antagonistic and agonistic signalling [34]. Diversification of *TLR9* paralogs could indicate adaptation towards lineage-specific pathogen loads or diversity within Gadiformes. Furthermore, in mammals TLR9 signalling can induce MHCI antigen cross-presentation [35] which overlap with the hypothesized subfunctionalization of some *MHCI* copies in Gadiformes [7]. Overall, our findings demonstrate that *TLR9* paralogs have experienced a different selection pressure compared to *TLR25* paralogs. Collectively, the gene expansions observed in Gadiformes, as well as in teleosts overall, are likely subject to different levels of neo- and subfunctionalization contributing to the further adaptation of the teleost innate immune system. Extreme northern or southern distributions are proxy indicators for temperature as these

regions are cooler but also have undergone a larger degree of paleoclimatic changes compared to the more tropical regions [36]. The observed expansions for *TLR7, TLR8*, *TLR9*, *TLR22, TLR23* and *TLR25*, especially within the Gadiformes, indicate selection towards higher copy number optima. This could potentially be explained by different pathogen loads or pathogen community compositions connected to highly variable paleoclimatic arctic environments. We found correlations between increased copy number of *TLR8*, *TLR22* and *TLR25* with more northern species distributions (Table 1). In contrast, *TLR9* showed higher optimal copy numbers in tropical regions – especially combined with the loss of *MHCII* (Table 1 and 2), most likely driven by the specific biotic or abiotic factors encountered in the tropics. Collectively, our findings indicate that, for the Gadiformes, both the paleogeographic distribution (reflecting the environments these species have inhabited through time) and the loss of *MHCII*, have been vital drivers for the expansion of *TLR8*, *TLR22*, *TLR25* and in particular *TLR9*.

By using a dated phylogeny we find that the successive alterations to the teleost immune system occurred in periods with substantial paleoclimatic fluctuations as well as oceanographic changes due to continental drift. Such events are often associated with periods of extinction followed by population diversification and subsequent speciation enabling the invasion of new niches [37, 38]. Our data suggests that the overall loss of *TLR5* (previously reported [39]) and *TLR21beta* (175-130 mya) overlap the Jurassic-Cretaceous (J-K) boundary (Figure 1).Although this transition between geological periods does not harbour any well-defined events, there is accumulating evidence of both species extinctions and radiations [40-44]. The loss of *TLR5* and *TLR21beta* may have occurred as adaptations to new habitats such as the expanding Central Atlantic Ocean . Even though both *TLR5* and *TLR21b* display lineage-specific loss, their presence/absence pattern

355 outside the Paracanthopterygii (Figure 1) indicate that they have experienced different selection
356 pressures before the J-K boundary.

357 Within the Gadiformes clade we find that the loss of *MHCII* coincides with the overall gene
358 expansion patterns of *TLR7*, *TLR8*, *TLR9*, *TLR22*, *TLR23* and *TLR25*, spanning a total interval
359 110-64 mya. This further overlaps with the early-late Cretaceous transition which includes one of
360 the late Cretaceous global anoxia events (95 mya). This anoxic environment, although likely
361 allowing a small degree of specialized adaptation, generally deprived the deep seas of species [45,
362 46]. Anoxic conditions led to higher extinction rates during this time period [47-50], fitting with
363 the metabolic cost scenario proposed to promote the loss of *MHCII* [51]. In this scenario, the
364 benefits of maintaining the MHCII system in some environments could not compensate for the
365 metabolic cost of expressing it. Coinciding with the anoxic event is the further northward opening
366 of the Central Atlantic Ocean [52] and the propagation of the South Atlantic Ocean to meet the
367 Central Atlantic Ocean [53-55]. The stress imposed by global ocean anoxia therefore appears
368 simultaneously with the appearance of new habitats. Further, this time period is associated with a
369 decrease in bony fish family richness, indirectly derived from fossil data [56], indicating that
370 these secondary changes to the Gadiformes immune system may have had slightly more adverse
371 effects here compared to the initial ones occurring at the J-K boundary. However, this likely had
372 a positive effect supporting species survival and radiation in the long term. The more recent loss
373 of *TLR1/2* from the Gadinae subfamily (40 – 16 mya) is likely a temperature-driven adaptation
374 caused by an abrupt cooling of global climate and loss of habitat due to the drastic decrease of
375 eustatic sea levels ~ 34 mya [50, 57, 58] overlapping with the opening of the North Atlantic
376 Ocean between Greenland and Norway [52].

## Conclusions

Overall, our findings reveal unprecedented variability within the teleost innate immune system, particularly within the Gadiformes, characterized by significant gene expansions and losses. Intriguingly, we find that higher copy numbers of *TLRs* correlate with species latitudinal distribution and the loss of *MHCII*. Further evidence of diversifying selection indicates that the paralogs likely experience different selection pressures. The successive nature of these changes to the ancestral teleost immune system, combined with the extensive evolvability of the innate immune system described here, have likely contributed to the overall survival and successful radiation of this lineage.

## References


1. Eschmeyer, W.N. and R. Fricke, *CATALOG OF FISHES: GENERA, SPECIES, REFERENCES.* . 2015.
2. Faircloth, B.C., et al., *A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs).* PLoS One, 2013. **8**(6): p. e65923.
3. Sallan, L.C., *Major issues in the origins of ray-finned fish (Actinopterygii) biodiversity.* Biol Rev Camb Philos Soc, 2014. **89**(4): p. 950-71.
4. Volff, J.N., *Genome evolution and biodiversity in teleost fish.* Heredity (Edinb), 2005. **94**(3): p. 280-94.
5. Star, B., et al., *The genome sequence of Atlantic cod reveals a unique immune system.* Nature, 2011. **477**(7363): p. 207-10.
6. Haase, D., et al., *Absence of major histocompatibility complex class II mediated immunity in pipefish, Syngnathus typhle: evidence from deep transcriptome sequencing.* Biol Lett, 2013. **9**(2): p. 20130044.
7. Malmstrom, M., et al., *Evolution of the immune system influences speciation rates in teleost fishes.* Nat Genet, 2016. **48**(10): p. 1204-10.
8. Malmstrom, M., et al., *Unraveling the evolution of the Atlantic cod's (Gadus morhua L.) alternative immune strategy.* PLoS One, 2013. **8**(9): p. e74004.
9. Palti, Y., *Toll-like receptors in bony fish: from genomics to function.* Dev Comp Immunol, 2011. **35**(12): p. 1263-72.
10. Rebl, A., T. Goldammer, and H.M. Seyfert, *Toll-like receptor signaling in bony fish.* Vet Immunol Immunopathol, 2010. **134**(3-4): p. 139-50.
11. Roach, J.C., et al., *The evolution of vertebrate Toll-like receptors.* Proc Natl Acad Sci U S A, 2005. **102**(27): p. 9577-82.
12. 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, 2016. **6**: p. 25211.
13. Malmstrom, M., et al., *Whole genome sequencing data and de novo draft assemblies for 66 teleost species.* Sci Data, 2017. **4**: p. 160132.

14. Drummond, A.J., et al., *Bayesian phylogenetics with BEAUti and the BEAST 1.7.* Mol Biol Evol, 2012. **29**(8): p. 1969-73.
15. Cunningham, F., et al., *Ensembl 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D662-9.
16. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2013. **41**(Database issue): p. D36-42.
17. Quiniou, S.M., P. Boudinot, and E. Bengten, *Comprehensive survey and genomic characterization of Toll-like receptors (TLRs) in channel catfish, Ictalurus punctatus: identification of novel fish TLRs.* Immunogenetics, 2013. **65**(7): p. 511-30.
18. Liu, Z., et al., *The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts.* Nat Commun, 2016. **7**: p. 11757.
19. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
20. Hansen, T.F., *Stabilizing selection and the comparative analysis of adaptation.* Evolution, 1997. **51**(5): p. 1341-1351.
21. Hansen, T.F., J. Pienaar, and S.H. Orzack, *A comparative method for studying adaptation to a randomly evolving environment.* Evolution, 2008. **62**(8): p. 1965-77.
22. R-Core-Team. *R: A Language and Environment for Statistical Computing*. 2015; Available from: http://www.R-project.org.
23. Froese, R. and D. Pauly. 2015; Available from: www.fishbase.org.
24. Torresen, O.K., et al., *An improved genome assembly uncovers prolific tandem repeats in Atlantic cod.* BMC Genomics, 2017. **18**(1): p. 95.
25. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
26. Delport, W., et al., *Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology.* Bioinformatics, 2010. **26**(19): p. 2455-7.
27. Pond, S.L. and S.D. Frost, *Datamonkey: rapid detection of selective pressure on individual sites of codon alignments.* Bioinformatics, 2005. **21**(10): p. 2531-3.
28. Kosakovsky Pond, S.L., et al., *A random effects branch-site model for detecting episodic diversifying selection.* Mol Biol Evol, 2011. **28**(11): p. 3033-43.
29. Murrell, B., et al., *Detecting individual sites subject to episodic diversifying selection.* PLoS Genet, 2012. **8**(7): p. e1002764.
30. Yang, L., et al., *Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish.* BMC Genomics, 2013. **14**: p. 65.
31. Kassahn, K.S., et al., *Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates.* Genome Res, 2009. **19**(8): p. 1404-18.
32. Pietretti, D. and G.F. Wiegertjes, *Ligand specificities of Toll-like receptors in fish: indications from infection studies.* Dev Comp Immunol, 2014. **43**(2): p. 205-22.
33. Tatematsu, M., T. Seya, and M. Matsumoto, *Beyond dsRNA: Toll-like receptor 3 signalling in RNA-induced immune responses.* Biochem J, 2014. **458**(2): p. 195-201.
34. Vabret, N., N. Bhardwaj, and B.D. Greenbaum, *Sequence-Specific Sensing of Nucleic Acids.* Trends Immunol, 2017. **38**(1): p. 53-65.
35. Brutkiewicz, R.R., *Cell Signaling Pathways That Regulate Antigen Presentation.* J Immunol, 2016. **197**(8): p. 2971-2979.
36. Lawson, A.M. and J.T. Weir, *Latitudinal gradients in climatic-niche evolution accelerate trait evolution at high latitudes.* Ecol Lett, 2014. **17**(11): p. 1427-36.
37. Wellborn, G.A. and R.B. Langerhans, *Ecological opportunity and the adaptive diversification of lineages.* Ecol Evol, 2015. **5**(1): p. 176-95.
38. Simoes, M., et al., *The Evolving Theory of Evolutionary Radiations.* Trends Ecol Evol, 2016. **31**(1): p. 27-34.

39. Solbakken, M.H., et al., *Successive Losses of Central Immune Genes Characterize the Gadiformes' Alternate Immunity.* Genome Biol Evol, 2016. **8**(11): p. 3508-3515.
40. Benson, R.B., et al., *Mesozoic marine tetrapod diversity: mass extinctions and temporal heterogeneity in geological megabiases affecting vertebrates.* Proc Biol Sci, 2010. **277**(1683): p. 829-34.
41. Benson, R.B. and P.S. Druckenmiller, *Faunal turnover of marine tetrapods during the Jurassic-Cretaceous transition.* Biol Rev Camb Philos Soc, 2014. **89**(1): p. 1-23.
42. Bambach, R.K., *Phanerozoic biodiversity mass extinctions.* Annual Review of Earth and Planetary Sciences, 2006. **34**: p. 127-155.
43. Alroy, J., *The shifting balance of diversity among major marine animal groups.* Science, 2010. **329**(5996): p. 1191-4.
44. Cavin, L., *Diversity of Mesozoic semionotiform fishes and the origin of gars (Lepisosteidae).* Naturwissenschaften, 2010. **97**(12): p. 1035-40.
45. Priede, I.G. and R. Froese, *Colonization of the deep sea by fishes.* Journal of Fish Biology, 2013. **83**(6): p. 1528-1550.
46. Rogers, A.D., *The role of the oceanic oxygen minima in generating biodiversity in the deep sea.* Deep-Sea Research Part Ii-Topical Studies in Oceanography, 2000. **47**(1-2): p. 119-148.
47. Takashima, R., et al., *Greenhouse World and the Mesozoic Ocean.* Oceanography, 2006. **19**.
48. Wilson, P.A. and R.D. Norris, *Warm tropical ocean surface and global anoxia during the mid-Cretaceous period.* Nature, 2001. **412**(6845): p. 425-9.
49. Sinninghe Damsté, J.S., et al., *A CO2 decrease-driven cooling and increased latitudinal temperature gradient during the mid-Cretaceous Oceanic Anoxic Event 2.* Earth and Planetary Science Letters, 2010. **293**(1–2): p. 97-103.
50. Harnik, P.G., et al., *Extinctions in ancient and modern seas.* Trends in Ecology & Evolution, 2012. **27**(11): p. 608-617.
51. Star, B. and S. Jentoft, *Why does the immune system of Atlantic cod lack MHC II?* Bioessays, 2012. **34**(8): p. 648-51.
52. Melankholina, E.N. and N.M. Sushchevskaya, *Development of passive volcanic margins of the Central Atlantic and initial opening of ocean.* Geotectonics, 2015. **49**(1): p. 75-92.
53. Granot, R. and J. Dyment, *The Cretaceous opening of the South Atlantic Ocean.* Earth and Planetary Science Letters, 2015. **414**: p. 156-163.
54. Voigt, S., et al., *Tectonically restricted deep-ocean circulation at the end of the Cretaceous greenhouse.* Earth and Planetary Science Letters, 2013. **369**: p. 169-177.
55. Murphy, D.P. and D.J. Thomas, *The evolution of Late Cretaceous deep-ocean circulation in the Atlantic basins: Neodymium isotope evidence from South Atlantic drill sites for tectonic controls.* Geochemistry Geophysics Geosystems, 2013. **14**(12): p. 5323-5340.
56. Guinot, G. and L. Cavin, *'Fish' (Actinopterygii and Elasmobranchii) diversification patterns through deep time.* Biol Rev Camb Philos Soc, 2016. **91**(4): p. 950-981.
57. Liu, Z., et al., *Global cooling during the eocene-oligocene climate transition.* Science, 2009. **323**(5918): p. 1187-90.
58. Goldner, A., N. Herold, and M. Huber, *Antarctic glaciation caused ocean circulation changes at the Eocene-Oligocene transition.* Nature, 2014. **511**(7511): p. 574-7.

# Declarations

## Acknowledgements

This work was supported by The Research Council of Norway. Some of the teleost genomes were assembled using the Abel Cluster, owned by the University of Oslo and the Norwegian metacenter 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 authors would like to thank Dr. Mark Ravinet and Helle Tessand Baalsrud for comments on the manuscript.

## Ethics approval and consent to participate

Not applicable.

## Funding

This study was supported by The Research Council of Norway (Grant numbers 222378/F20 and 199806/S40 to KSJ/SJ). The funding agency was not involved with any aspects of the project beyond providing funding.

## Availability of data and materials

All novel teleost sequence and genome resources are available at European Nucleotide Archive (ENA) and the Dryad digital repository, submitted by the Malmstrøm et al (2016): All raw data (sequencing reads) are available at ENA with study accession number PRJEB12469 (sample identifiers ERS1199874-ERS1199939). Genome assemblies, available at Dryad, exist in two versions (UTGs and scaffolds) under DOI: doi:10.5061/dryad.326r8. All additional resources needed to generate the findings presented here are available in our GitHub repository including, but not limited to, scripts, BLAST and SLOUCH output files:
https://github.com/MonicaSolbakken/TLR

## Consent for publication

Not applicable.

## Author contributions

MHS, KSJ and SJ conceived planned and oversaw the project. MHS generated all TLR related data based on BLAST searches towards the teleost genome resources as well as extracting information about latitude and depth from online databases. KLV performed all SLOUCH analyses. MHS made all figures/tables and wrote the overall text with significant aid of SJ and KSJ. KLV wrote all sections related to SLOUCH. All authors contributed with comments, edits and proofreading of the manuscript.

## Competing interests

The authors declare that they have no competing interests.

## Corresponding author

Correspondence and requests for materials should be addressed to MHS: m.h.solbakken@ibv.uio.no and SJ: sissel.jentoft@ibv.uio.no.

# Figures and tables

**Table 1 Phylogenetic comparative analyses of the evolution of *TLR* copy numbers in relation to species latitudinal distributions using SLOUCH.** For each model, we show the phylogenetically corrected $r^2$, and the AICc score. AICc balances goodness of fit (log-likelihood) with the number of parameters in the model (model complexity). The model with the lowest AICc value is the best supported.. $R^2$ represents the amount of the total variation that is explained by the model. Detailed output from each

546 model is given in supplementary information. The model called "phylogeny" does not include any
547 explanatory variables and is given as a reference point for comparison to models with predictor variables.

| | *TLR8* | | *TLR9* | | *TLR22* | | *TLR23* | | *TLR25* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Category** | **AICc** | **r2** | **AICc** | **r2** | **AICc** | **r2** | **AICc** | **r2** | **AICc** | **r2** |
| **Phylogeny** | 266.41 | 0.00 | 243.91 | 0.00 | 430.27 | 0.00 | 307.65 | 0.00 | 241.36 | 0.00 |
| **Group 75 latitude** | 260.29 | 18.32 | 239.07 | 18.91 | **418.86** | **24.63** | 311.72 | 0.96 | **226.61** | **32.26** |
| **Group 50 latitude** | **259.75** | **19.02** | 240.67 | 15.49 | 427.26 | 13.70 | 310.88 | 2.30 | 232.46 | 21.96 |
| **Group 25 latitude** | 259.98 | 18.72 | 240.34 | 17.24 | 429.86 | 8.86 | **307.22** | **7.91** | 233.31 | 20.89 |
| **Group 0 latitude** | 259.90 | 20.13 | **238.24** | **19.99** | 427.05 | 13.99 | 311.27 | 1.69 | 232.56 | 21.84 |
| **Group -25 latitude** | 260.06 | 18.63 | 239.78 | 16.69 | 429.34 | 9.62 | 309.78 | 4.00 | 233.38 | 20.80 |
| **Group -50 latitude** | 260.31 | 16.35 | 240.16 | 16.18 | 429.62 | 9.21 | 311.54 | 1.24 | 233.45 | 20.71 |

548

549 **Table 2 Phylogenetic comparative analyses of the evolution of *TLR* copy numbers in relation to**
550 **species latitudinal distributions and *MHCII* status using SLOUCH.** For each model, we show the
551 phylogenetically corrected $r^2$, and the AICc score. AICc balances goodness of fit (log-likelihood) with the
552 number of parameters in the model (model complexity). The model with the lowest AICc value is the best
553 supported.. $R^2$ represents the amount of the total variation that is explained by the model. Detailed output
554 from each model is given in supplementary information. The model called "phylogeny" in Table 1 does
555 not include any explanatory variables and is given as a reference point for comparison to models with
556 predictor variables.

| | *TLR8* | | *TLR9* | | *TLR22* | | *TLR23* | | *TLR25* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Category** | **AICc** | **r2** | **AICc** | **r2** | **AICc** | **r2** | **AICc** | **r2** | **AICc** | **r2** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Group MHCII** | 259.43 | 19.44 | 239.01 | 22.41 | 427.99 | 14.53 | 328.13 | 2.65 | 231.30 | 26.94 |
| **Group MHCII + Group 75 lat.** | 264.37 | 19.52 | 240.39 | 31.32 | **420.41** | **30.23** | 315.34 | 3.21 | **228.98** | **35.08** |
| **Group MHCII + Group 50 lat.** | **262.31** | **22.16** | 243.11 | 25.72 | 431.07 | 17.15 | 314.69 | 4.23 | 235.60 | 27.76 |
| **Group MHCII + Group 25 lat.** | 263.76 | 20.32 | 239.13 | 32.69 | 431.69 | 16.32 | **311.21** | **9.41** | 234.82 | 28.67 |
| **Group MHCII + Group 0 lat.** | **261.00** | **27.11** | **228.48** | **53.53** | 430.16 | 18.36 | 314.90 | 3.92 | 233.63 | 30.02 |
| **Group MHCII + Group -25 lat.** | 263.37 | 20.82 | 230.06 | 52.33 | 432.07 | 15.80 | 313.49 | 6.01 | 234.84 | 28.64 |
| **Group MHCII + Group -50 lat.** | 264.39 | 19.50 | 240.15 | 29.17 | 432.19 | 15.63 | 314.92 | 3.81 | 235.60 | 27.76 |

**Figure 1 The *TLR* repertoires of 76 teleosts mapped onto a time-calibrated species phylogeny.** All *TLRs* characterized in the new 66 teleost genomes as well as in 10 reference teleosts genomes (Ensembl and GenBank) mapped onto a species phylogeny generated by Malmstrøm et al. The phylogeny demonstrates the loss of *MHCII* 110-64 mya (branch range time, black star) reported by Malmstrøm et al. Lineage-specific *TLR* losses are marked by black circles (Gadinae *TLR1/2*, Paracanthopterygii *TLR5* and *TLR21beta*). The individual species' repertoires are depicted with boxes where the coloration represents the number of copies of each individual *TLR*. The six major *TLR* families: *TLR1-family*, *TLR3-family*, *TLR4-family*, *TLR5-family*, *TLR7-family* and *TLR11-family* are indicated with black bars underneath the *TLR* names. See Supplementary table 1 for copy number details. For *TLR1/2* a gradient-filled box indicates the presences of either *TLR1* or *TLR2*. The Paracanthopterygiian lineage, Gadiformes order and Gadinae family display shaded grey backgrounds.

**Figure 2 Overview of sites reported by the MEME analysis performed on *TLR3*, *TLR9* and *TLR25* in the selected species**. A schematic drawing of the *TLR3* (A), *TLR9* (B) and *TLR25* (C) protein domains with a black ecto-domain (dimerization and ligand interaction), an ochre transmembrane (TM) domain and a red TIR domain (signalling domain). Only the ecto-domain was subjected to selection analysis as the TM and TIR domains are known to be under purifying selection. Grey boxes indicates which parts of the ecto-domain that were included in the alignment and also shows how many alignments were generated per gene. The dark grey is a section overlap between TLR25 section 2 and 3 consisting of 22 codon positions. Blue arrows indicate sites reported by the MEME analysis. For site details see supplementary information.