Evolved for success in novel environments: The round goby genome

Authors

Irene Adrian-Kalchhauser[1], Anders Blomberg[2*], Tomas Larsson[3*], Zuzana Musilova[4*], Claire R Peart[5*], Martin Pippel[6*], Monica Hongroe Solbakken[7*], Jaanus Suurväli[8*], Jean-Claude Walser[9*], Joanna Yvonne Wilson[10*], Magnus Alm Rosenblad[2,11§], Demian Burguera[4§], Silvia Gutnik[12§], Nico Michiels[13§], Mats Töpel[2§], Kirill Pankov[10§], Siegfried Schloissnig[14§], Sylke Winkler[6§]

Author contributions

* equal contribution, section lead authors, listed alphabetically

§ co-authors with equal contribution, listed alphabetically

Affiliations

[1] Program Man-Society-Environment, Department of Environmental Sciences, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland

[2] Department of Chemistry and Molecular Biology, University of Gothenburg, Medicinaregatan 9C, 41390 Gothenburg, Sweden

[3] Department of Marine Sciences, University of Gothenburg, Medicinaregatan 9C, 41390 Gothenburg, Sweden

[4] Department of Zoology, Charles University, Vinicna 7, CZ-128 44 Prague, Czech Republic

[5] Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany

[6] Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany

[7] Centre for Ecological and Evolutionary Synthesis, University of Oslo, Blindernveien 31, 0371 Oslo, Norway

[8] Institute for Genetics, University of Cologne, Zülpicher Strasse 47a, D-50674 Köln, Germany

[9] Genetic Diversity Centre, ETH, Universitätsstrasse 16, 8092 Zurich, Switzerland

[10] Department of Biology, McMaster University, 1280 Main Street West, Hamilton, ON, Canada

[11] NBIS Bioinformatics Infrastructure for Life Sciences, University of Gothenburg, Medicinaregatan 9C, 41390 Gothenburg, Sweden

[12] Biocenter, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland

[13] Institute of Evolution and Ecology, University of Tuebingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany

[14] Research Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), 1030 Vienna, Austria.

Corresponding author

Irene Adrian-Kalchhauser

email: irene.adrian-kalchhauser@unibas.ch

mail: University of Basel, Vesalgasse 1, 4051 Basel

phone: +41 61 2070410

Keywords

PacBio, *Neogobius melanostomus,* invasive species, fish, genomics, evolution, adaptation, gene duplication, vision, olfaction, innate immunity, detoxification, osmoregulation, epigenetics

Abstract

Since the beginning of global trade, hundreds of species have colonized territories outside of their native range. Some of these species proliferate at the expense of native ecosystems, i.e., have become invasive. Invasive species constitute powerful *in situ* experimental systems to study fast adaptation and directional selection on short ecological timescales. They also present promising case studies for ecological and evolutionary success in novel environments.

We seize this unique opportunity to study genomic substrates for ecological success and adaptability to novel environments in a vertebrate. We report a highly contiguous long-read based genome assembly for the most successful temperate invasive fish, the benthic round goby (*Neogobius melanostomus)*, and analyse gene families that may promote its impressive ecological success.

Our approach provides novel insights from the large evolutionary scale to the small species-specific scale. We describe expansions in specific cytochrome P450 enzymes, a remarkably diverse innate immune system, an ancient duplication in red light vision accompanied by red skin fluorescence, evolutionary patterns in epigenetic regulators, and the presence of genes that may have contributed to the round goby's capacity to invade cold and salty waters.

A recurring theme across all analyzed gene families are gene expansions. This suggests that gene duplications may promote ecological flexibility, superior performance in novel environments, and underlie the impressive colonization success of the round goby. *Gobiidae* generally feature fascinating adaptations and are excellent colonizers. Further long-read genome approaches across the goby family may reveal whether the ability to conquer new habitats relates more generally to gene copy number expansions.

Introduction

Since the beginning of global trade and the colonial period, hundreds of species have colonized territories outside their native range. A fraction of those species proliferates at the expense of native species and ecosystems, i.e., they are invasive. While invasive species present challenges for biodiversity and ecosystem conservation, they also constitute exciting eco-evolutionary models for adaptation and ecological success in novel or changing environments (1–4).

The benthic round goby *Neogobius melanostomus* (**Figure 1A**) is one of the most widespread invasive fish species. Since 1990, round gobies have been detected in over 20 countries outside their native Ponto-Caspian range. In some regions of Europe and North America, they have become the most common fish species (5–7) (**Figure 1B**). Lasting impacts on biodiversity and on ecosystems have been observed (see (8) for a summary of the impacts). In recent years, the round goby has therefore become a novel model for ecology, behavior and evolution (**Figure 1C**).

The round goby effortlessly outcompetes native species with similar ecology, and is therefore a promising candidate to study fundamental questions on long-term ecological and evolutionary success. Round goby sequence data are presently restricted to a handful of phylogenetic markers (9–15). However, genome analyses have previously provided significant insights into fish ecology and evolution. Examples are genome compaction (16), the transition from fin to limb (17), loss of major parts of adaptive immunity (18), or effects of genome duplication (19). We therefore expect relevant insights into round goby biology and into success in novel environments from the round goby genome sequence.

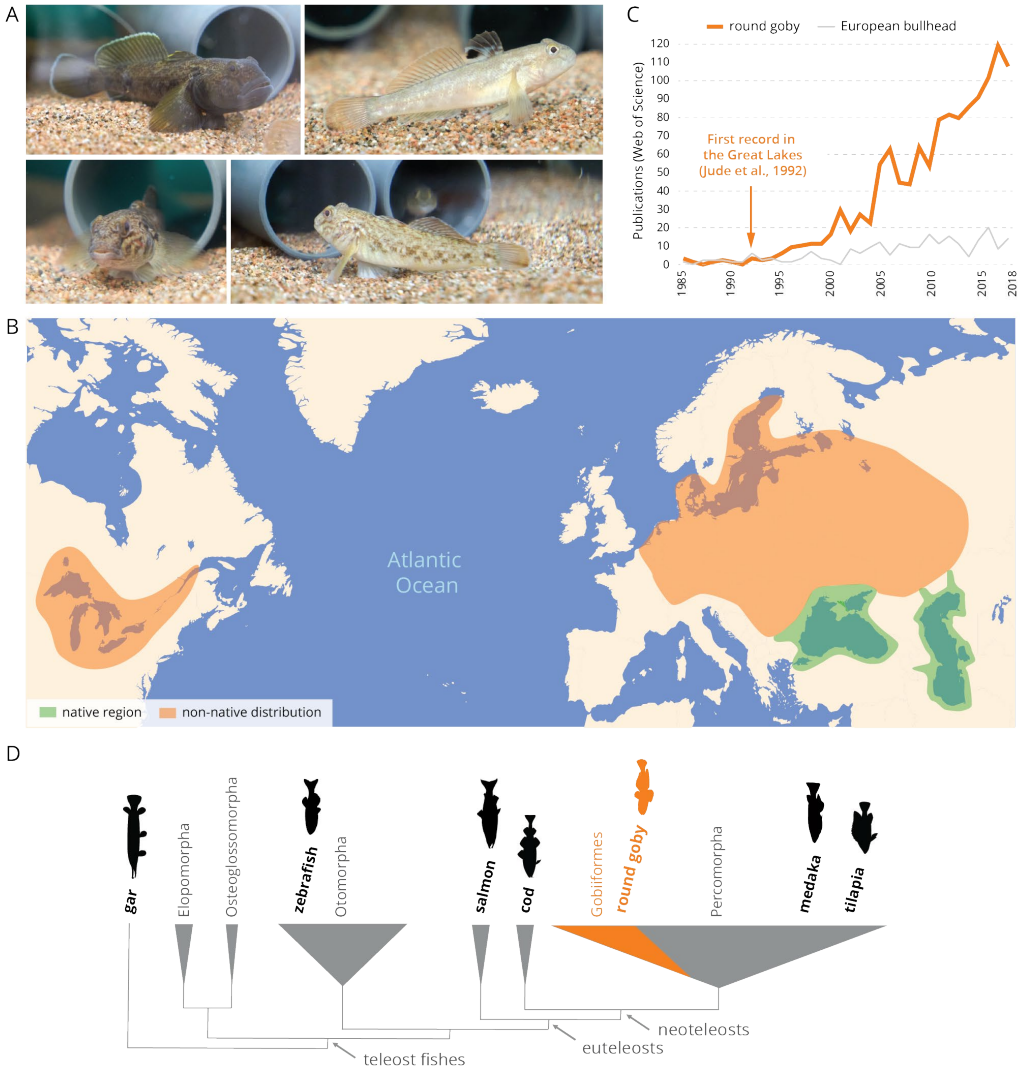


Figure 1

The round goby, a benthic invasive fish species. A *Wild-caught round goby in aquaria. Individuals are usually brightly colored or spotted with a characteristic black dot on the first dorsal fin. During the reproductive season, territorial males develop a black body color (first panel).* ***B*** *The growing scientific relevance of the species is reflected by records in Web of Science (orange) when compared to a non-invasive fish with similar ecology (European bullhead, Cottus gobio; grey).* ***C*** *Current distribution of round goby. The round goby has spread from its native region (green) to many European rivers and lakes, the Baltic Sea, the Great Lakes and their tributaries (orange).* ***D*** *Phylogenetic position of the round goby among fishes.*

The survival of an individual in a novel environment depends on how well it can perceive, react to, and accommodate to its new surroundings. In this study, we therefore explore a high quality and contiguous genome assembly of the round goby for genes related to three categories: environmental perception, reaction to environmental conditions, and long-term accommodation to novel environments. We focus on gene families that have been hypothesized to play a role in the colonization of novel environments and on gene families that may relate to round goby invasion ecology.

For environmental perception, we investigated genes responsible for sensory perception in fishes. We specifically focused on the opsin genes for visual perception, as well as on the olfactory receptors for odor perception. Vision in fishes is often specifically adapted to environmental conditions, such as darkness in deep water (20), modified color spectrum in turbid water (21, 22), habitat color (23), or specific light regimes or light compositions (24–26). The overall spectral sensitivity range of teleost fishes exceeds the human visual range and, in many cases, includes the UV (23) and far-red (27) spectrum. Similarly, olfaction is an essential chemoreception sense for fish, allowing for fast responses to predators and alarm cues as well as for intra-species communication. Pheromones have play an important role in the round goby (28–30), and males attract females into their nests by releasing them (31). A particularly specialized sense of smell therefore may provide an advantage during initial population establishment in novel environments.

We further investigated genes that may mediate responses to novel environments, namely genes involved in detoxification, ion transport and the immune system. The round goby occurs even in chemically contaminated harbors (32–34) and appears to tolerate xenobiotic compounds well. This suggests that the round goby may be particularly well equipped to degrade and eliminate chemical pollutants. We therefore analyze the cytochrome P450 gene superfamily, which is a particularly important and conserved part of the xenobiotic response (35). The round goby also tolerates a wide range of salinities (0 to 25 PSU / ‰) and temperatures (0°C-30°C) and occurs at latitudes ranging from <40° N in the Ponto-Caspian region to >60° N in the Baltic Sea. Most fishes tolerate only a narrow range of salinities (36); the round goby however belongs to a specialized group, the euryhaline fish species, which thrive in fresh and brackish environments and includes estuarine species and migratory species such as salmon. We study the genetic basis of osmoregulation and osmolyte

production in round goby to gain insights into the evolution of salinity and cold tolerance, and to possibly predict future range expansions. Invasive species encounter an array of previously unknown pathogens when they colonize a habitat, and invasion success may be related to a species' ability to tackle novel immune challenges (37). Intriguingly, the round goby displays a low parasite load at the invasion front (38). We therefore characterize key factors of the innate and the adaptive immune system.

We also investigated conserved gene regulators because such genes might be involved in long-term adaptation to a novel environment. Mechanisms such as DNA methylation and histone modifications promote long- and short-term gene expression regulation and therefore mediate adaptations to altered conditions at the cellular level (39), but also regulate genome-scale evolutionary processes such as the distribution of meiotic recombination events (40) or transposon activity (41), and provide stochastic variability as basis for selection (42). Epigenetic variants have been proposed to cause fitness-relevant differences in gene expression and phenotype (43, 44). The ecological flexibility of the round goby has been linked to enhanced gene expression plasticity in response to environmental stimuli (45) and to their ability to pass information on water temperature to their offspring through maternal RNA (46). To understand the features of core epigenetic regulators in the round goby, we focused on two widely conserved and well characterized parts of the epigenetic machinery: the histone-methylating PRC2 complex and the DNA methylases. Both mechanisms are thought to restrict developmental plasticity, downregulate gene expression (at least in mammals), and have been linked to plastic responses, behavioral changes, and environmental memory (47–51).

Finally, we take advantage of the high genome contiguity to investigate sex determination using RAD sequencing data. Fish display a wide variety of sex determination mechanisms, ranging from sex chromosomes to multilocus genetic sex determination to environmental sex determination (52), and sex determination in the round goby has not previously been investigated.

Results

1. The round goby genome

The round goby genome assembly ("RGoby_Basel_V2", BioProject accession PRJNA549924, BioSample SAMN12099445, GenBank genome accession VHKM00000000, release date July 22 2019) consists of 1364 contigs with a total length of 1.00 Gb (1'003'738'563 bp), which is within the expected size range (53–55). It is assembled to high contiguity (NG50 at 1'660'458 bp and N50 at 2'817'412 bp). GC content is 41.60%. An automated Maker gene annotation predicts a total of 38,773 genes and 39,166 proteins, of which 30,698 are longer than 100 amino acids (**Table 1**; annotation track available as **Supplemental_Material_S1**). The genome does not appear to contain a sex chromosome or a large sex determining region, since a RAD-tag dataset from 40 females and 40 males with an estimated resolution of 25,000 – 45,000 bp does not contain any sex-specific loci.

Approximately 47% of the genome assembly is masked as various types of repetitive sequences by RepeatMasker in the Maker annotation pipeline. The genome consists of approximately 9% predicted interspersed repeats (**Supplemental_Table_S1**), which is much lower than for zebrafish (*Danio rerio*, total genome size 1427.3Mb, 46% predicted as interspersed repeats) but higher than for the more closely related three-spined stickleback (*Gasterosteus aculeatus*, total genome size 446.6Mb, 3.2% predicted as interspersed repeats). Among interspersed repeats, the long terminal repeat (LTR) retrotransposon family is the most common in many species including fish (Repbase, https://www.girinst.org/repbase/). RepeatMasker identifies 0.9% LTR retrotransposons in the round goby genome, but separately run *de novo* predictions with LTRfinder and LTRharvest (**Supplemental_Table_S1)** indicate an underestimation of LTR retrotransposons and interspersed repeats by this approach. The latter approaches estimate that the proportion of LTR retrotransposons in the round goby genome is 11.2% (3.8% LTRs with target-site-repeats; LTRfinder) or 4.9% (LTRharvest), respectively.

In addition to the genome sequence, we provide raw short read sequencing data from various published and ongoing projects. They include RNA sequencing data from early cleavage embryos

(46), DNA methylation capture data from adult male brains (47), as well as RAD tags from two local Swiss populations and ATAC seq reads from brain and liver (unpublished; **Table 1**).

Table 1. Round goby genome assembly and annotation statistics.

Assembly	
Number of contigs	1364
Total genome length (bp)	1,003,738,563
Longest contig (bp)	19,396,355
Smallest contig (bp)	21,178
N50 contig length (bp)	2,817,412
Annotation	
Number of genes	38,773
Genomic repeat content (%)	47
G + C (%)	41.60
LTR retrotransposons (%)	4.9 - 11.2
Accession	NCBI BioProject PRJNA549924 Accession VHKM00000000
Additional sequencing data	
RNA (Adrian-Kalchhauser 2018)	Embryonic transcriptome (1-32 cell stages) from 16 clutches NCBI BioProject PRJNA547711 NCBI SRA SRR9317352 - SRR9317366
DNAme (Somerville 2019)	Brain DNA methylation data from 15 males NCBI BioProject PRJNA515617 NCBI SRA SRR8450505 - SRR8450528
RADseq (unpublished)	RAD Seq data from 120 individuals NCBI BioProject PRJNA547536 NCBI SRA SRR9214152 - SRR9214154
ATACseq (unpublished)	ATAC Seq data of liver and brain from 50 individuals NCBI BioProject PRJNA551348 NCBI SRA SRR9714857 - SRR9714901

2. Sensory perception genes: Vision

Vertebrates perceive color with cone cells expressing one of four types of opsin proteins (usually sensitive to the red, green, blue, and ultraviolet part of the spectrum) and dim light with rod cells expressing the rod opsin. The UV and blue light is detected by the short-wavelength sensitive SWS1 and SWS2 opsins, the green part of the spectrum is perceived mostly by the rhodopsin-like RH2 opsins, and the red color by the long-wavelength sensitive LWS opsins. Rod cells are active in the dim-light conditions and contain the rod opsin RH1 (56). Gene duplications and losses of the opsin genes during fish evolution correlate to certain extent with adaptations to specific environments (20, 57).

We identified two cone opsin gene duplications in the round goby genome. Firstly, the genome features a recent duplication of the green-sensitive RH2 gene. RH2 duplications are a common phenomenon in fish (**Figure 2**). Secondly, the genome features an ancient duplication of the long-wave red-sensitive LWS gene. The event can be traced most likely to the ancestor of all teleosts, or possibly even to the ancestor of Neopterygii (**Figure 2**; see **Supplemental_Fig_S1** for full tree). As expected, the round goby genome further contains one dim-light rod opsin (RH1) gene, two blue-sensitive SWS2 genes (57), and as previously reported for gobies, lacks the UV/violet-sensitive SWS1 gene (20, 25, 57).

The proposed ancestral position of the red opsin gene duplication is supported by three lines of evidence. First, the monophyly of all other teleost + gar LWS genes is strongly supported by the Bayesian analysis (Bayesian posterior probability value = 1). Second, the distant phylogenetic position is supported by trees based on individual exons, which indicate a low probability of a compromised phylogenetic signal, e.g. due to the partial gene conversion. Three of four exons cluster at the same position as the whole gene, while the fourth exon (exon 4) cluster with the genes resulting from a more recent teleost-specific LWS duplication specific to *Astyanax* and *Scleropages* (58) (**Supplemental_Fig_S2**). Third, the choice of outgroup (parietopsin or pinopsin) does not affect the position of the LWS2 gene. Together, these analyses suggest either (1) the presence of an ancient gene duplication event of the LWS gene in the ancestor of teleost and holostean fishes (i.e. *Neopterygii*) which was retained only in the goby family, or (2) a teleost-specific event, possibly identical to that reported for characins and bony tongues (58), with a subsequent concerted goby-specific sequence diversification in exons 2, 3 and 5.

The spectral sensitivity of photopigments, i.e. their excitation wavelength can be modified by substitutions in certain key amino acids (59). We find that round goby LWS1 and LWS2 differ in the key spectral tuning site at amino acid 277 (position 261 of bovine rhodopsin, **Table 2)** suggesting a sensitivity shift of 10 nm.

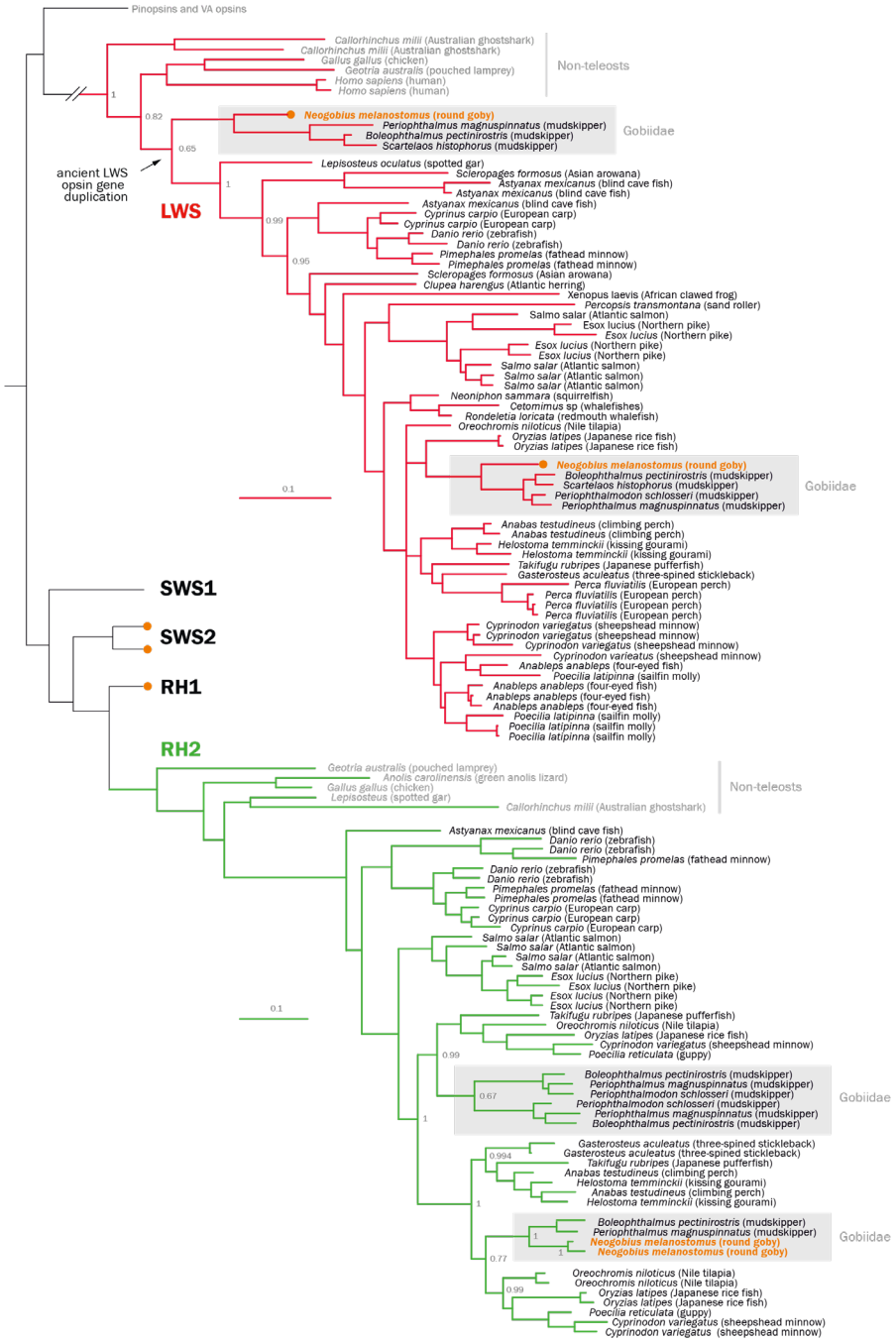


Figure 2

Phylogenetic tree of vertebrate opsin gene sequences. *Maximum-likelihood phylogenetic tree based on the cone and rod visual opsins and using VA opsins and pinopsins as outgroup. The round goby genome contains two LWS gene copies, which seem to be the results of an ancient gene duplication event, and two more recently duplicated RH2 gene copies. Round goby is indicated in orange. Red opsin branches are indicated in red. Green opsin branches are indicated in green. Non-teleost species and the outgroup (VA opsins and pinopsins) are indicated in grey. Grey boxes highlight Gobiidae.*

To find a possible link to the ecological significance of the red opsin duplication, we checked for the presence of red skin fluorescence in the round goby. Interestingly, round goby individuals of both sexes and of all sizes (n=10) feature weakly red fluorescent crescents above the eyes (**Figure 3**). Whether such pattern has any relevance for the putatively enhanced vision in the red spectrum remains elusive.

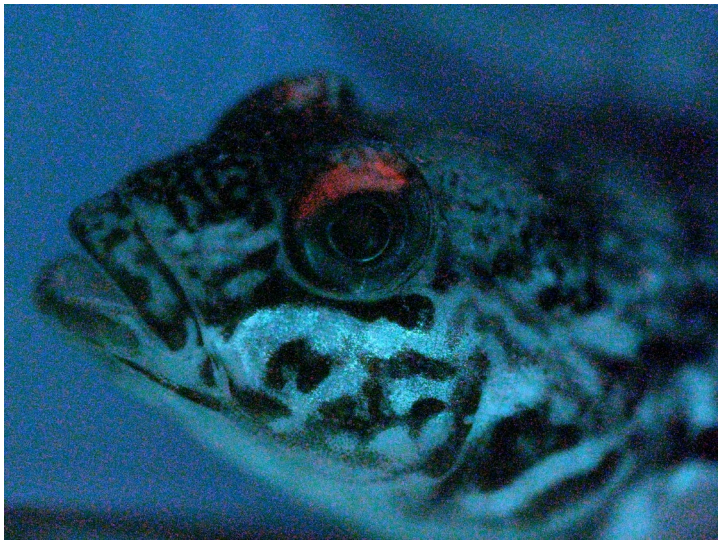


Figure 3

Red fluorescence in the round goby. *Round gobies exhibit red fluorescence above the eyes when exposed to green light.*

Table 2. Amino acid analysis of key tuning sites in Gobiidae red opsins proteins.

species	ecology	gene	key tuning amino acid site in round goby					max. spectral sensitivity (wavelength)	reference
			180	197	277	285	308		
		bovine rhodopsin equivalent site:	*164*	*181*	*261*	*269*	*292*		
Boleophthalmus pectinirostris	terrestrial mudskipper	LWS1	A	H	Y	T	A	553 nm	(25)
		LWS2	A	H	F	A	A	531 nm	
Periophthalmus magnuspinnatus	terrestrial mudskipper	LWS1	S	H	Y	T	A	560 nm	
		LWS2	A	H	F	T	A	546 nm	
Neogobius melanostomus	freshwater temperate rivers and lakes	LWS1	S	H	Y	T	A	560 nm	this study
		LWS2	S	H	F	T	A	550 nm*	this study
* = predicted by the key tuning sites, and Y261F shift of 10 nm; Yokoyama, 2008.									

3. Sensory perception genes: Olfaction

Olfactory receptors (OR) in vertebrates are 7-transmembrane-domain G-protein coupled transmembrane proteins. They are expressed in neurons embedded in membranes of the olfactory lamellae. Mammals usually have several hundred OR genes that cluster in two major types (~400 in human, and ~1000 genes in mouse) (60). Teleost fishes possess fewer OR genes but feature a higher diversity (5 kinds of type 2 ORs in teleosts as compared to 2 kinds of type 2 ORs in mammals) (61). The binding properties of individual ORs, especially in fishes, are virtually unexplored.

We identified 112 putative olfactory receptor genes in the round goby genome. To put this result into evolutionary context, all analyses were carried out in comparison with two Gobiidae species (blue-spotted mudskipper and giant mudskipper) and two percomorph species (three-spined stickleback, *Gasterosteus aculeatus* and Nile tilapia, *Oreochromis niloticus*; **Figure 4A**). The round goby presented similar number of ORs (112) to the giant mudskipper (106) and stickleback (115), notably less than the blue-spotted mudskipper (154) and near half the amount compared to Nile tilapia (214). We find that all ORs belong to one of two transmembrane domain subtypes according to the Pfam database (7tm4 or 7tm1; **Figure 4B**; **Supplemental_Fig_S3**). This matches a previous large-scale phylogenetic analysis which identified two main types of olfactory receptor genes in vertebrates (61). The functional differences between the domain subtypes are unclear, but their different consensus sequences may confer distinct biochemical properties.

Our analyses identify several cases of clade-specific gene expansions. Certain OR genes are expanded in parallel in several lineages (**Figure 4C**). Likely, those expansion events are the result of clade-restricted gene duplications, although a secondary role for gene conversion after species divergence cannot be ruled out. While the Nile tilapia features the greatest overall amount of expansions, the round goby presents the highest number of genes and expansions within the 7tm1 subfamily, a trend that is consistent in the other Gobiidae species (**Figure 4D**).

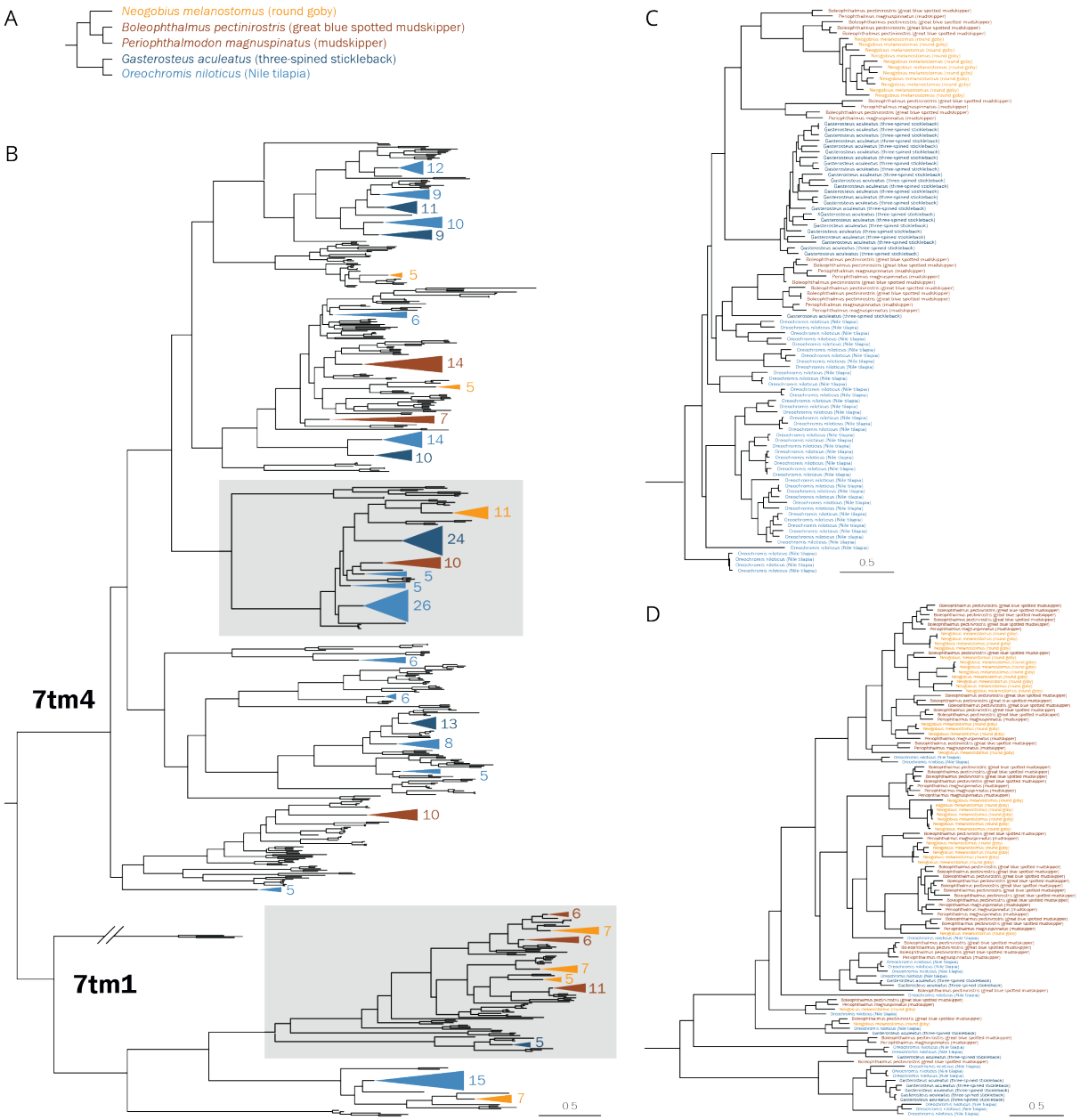


Figure 4

Phylogenetic tree of percomorph olfactory receptor protein sequences. A *Phylogenetic relationship among five analyzed percomorph species, i.e. three gobiids, one cichlid and one stickleback.* ***B*** *Maximum-likelihood phylogenetic tree constructed with adrenergic receptors as outgroup. Sequences were identified de novo except for nile tilapia (blue). Branches magnified in panels C and D are highlighted with grey boxes.* ***C*** *Branch of the 7tm4 family featuring large independent expansions in all species analyzed.* ***D*** *Branch of the 7tm1 family featuring several expansions in Gobiidae (red, orange) that are not paralleled in other percomorph species (blue).*

4. Response to the environment: Detoxification

The CYP gene superfamily is an essential part of the defensome, a collection of genes that provide protection against harmful chemicals (35). Vertebrate genomes contain between 50-100 CYP genes. The genomes of fugu and zebrafish, for example, encode 54 (62) and 94 (63) CYP genes respectively. Expansions of individual CYP families occur in both mammals and fish. For example, zebrafish have three times as many CYP2 family members (40) as most other vertebrate species (13-15), and similar expansions of CYP2 genes have been observed in mice and rats (64).

We find that the round goby genome contains few CYP genes. We identify 25 complete or partial CYP genes, as well as 21 gene fragments. Pseudogenes are common for CYP genes (62, 63, 65), which is why strict annotation criteria are applied first before smaller fragments are considered. In total, the genome contains approximately 50 CYP genes (**Supplemental_Table_S2**).

When including gene fragments, all expected CYP families are present in the round goby, and the phylogenetic analyses show the expected relationships between gene families and between vertebrates (**Figure 5**). Fish and most vertebrates have CYP genes from 17 families (CYP 1-5, 7, 8, 11, 17, 19, 20, 21, 24, 26, 27, 46 and 51) (62), while the CYP39 family occurs in humans and zebrafish, but not in fugu (62, 63). In the round goby, the complete or partial genes could be assigned to 9 CYP families (CYP 1- 4, 8, 19, 26, 27 and 51). The families CYP7, CYP11, CYP17 and CYP21 were present among the sequence fragments.

Contrary to expectations, the classical detoxification families CYP1-3 were not expanded (**Figure 5**). CYP1, 2, 3 and to a lesser extent CYP4 proteins are responsible for the oxidative metabolism of xenobiotic compounds (pollutants, drugs, etc.). In rodents and humans, the CYP1 family metabolizes planar cyclic aromatic hydrocarbon compounds (reviewed in (66)), the CYP2 family metabolizes structurally diverse drugs, steroids and carcinogens, the CYP4 family catalyzes the ω-hydroxylation of the terminal carbon of fatty acids and xenobiotics, and CYP3 genes metabolize a range of structurally different compounds in the liver and intestines. Over 50% of all pharmaceutical compounds are metabolized by CYP3A genes in human. The goby genome contains three or four CYP1 genes: one CYP1B gene, two CYP1C genes, and one CYP1A fragment. The latter lacks two main characteristics

(I- and K-helix) and could therefore be a pseudogene. As expected for a vertebrate (64), the genome contains many CYP2 genes. The most important fish CYP2 families were represented, including CYP2J, CYP2N, CYP2Y and CYP2AD. Finally, the round goby had a single CYP3A gene and a potential CYP3A fragment. This is somewhat unusual because fish often feature species-specific CYP3 subfamilies in addition to CYP3A. For example, medaka also contains CYP3B genes, zebrafish CYP3C genes, and Acanthopterygii fish CYP3D genes (67).

We find that the round goby genome contains six CYP8 genes, which is more than expected based on observations from the other gobies. The closely related large blue-spotted mudskipper has only two CYP8 genes (XM_020924471 and XM_020919000.1; about 73-85% identity); no sequences were found in other mudskipper species. Accordingly, we assume that the CYP8B genes have undergone species-specific tandem duplications in the round goby, as is also known for the subfamilies CYP2AA, CYP2X and CYP2K in zebrafish (64). Five round goby CYP8 genes locate to the same contig with high sequence similarity (~90%), which is similar to zebrafish CYP8B1-3 that also colocalize on the same chromosome (63). Misidentification of closely related CYP7, CYP8, and CYP39 genes as CYP8 is unlikely given the colocalization and high sequence similarity. The function of the expansion is presently unclear, although expression patterns in zebrafish suggest a role in the early embryo (63). In humans, CYP8 genes act as prostacyclin synthases that mediate steroid metabolic pathways in bile acid production or prostaglandin synthesis (68). Based on structural similarities with yeast proteins, CYP8 genes might also have E3 ubiquitin ligase activity. The almost identical crystal structures of zebrafish and human CYP8A1 suggest similar functions in fish and mammals (69).

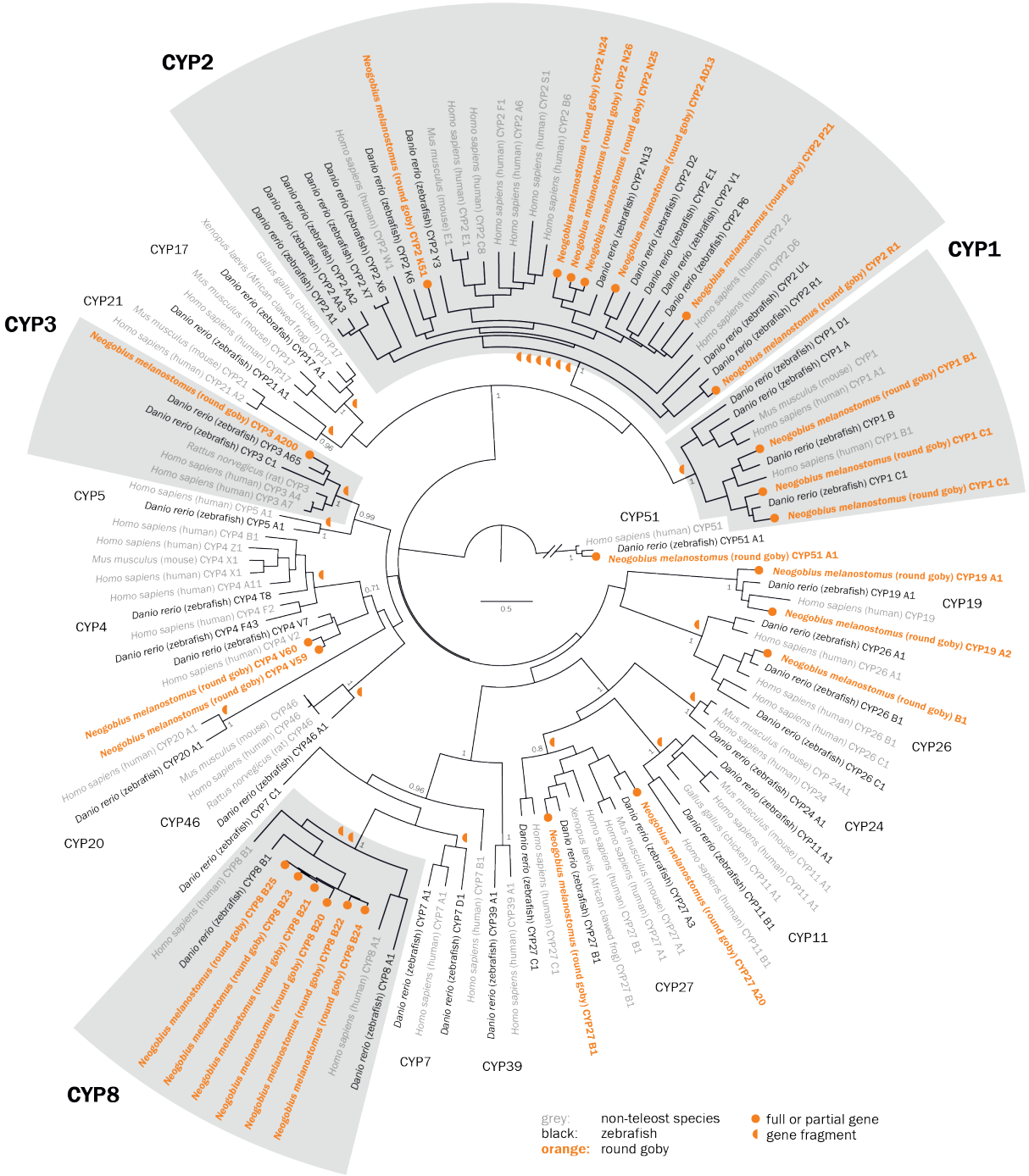


Figure 5.

Phylogenetic tree of vertebrate CYP protein sequences. *Maximum likelihood phylogenetic tree with 100 bootstraps, rooted with the CYP51 family. Detoxification genes CYP1-3 do not feature unusual expansions, while the CYP8 family is expanded to six members (grey boxes). Non-fish vertebrates are printed in grey. Fragments too short for tree building but attributable to a certain family are indicated by orange half circles next to the root of the respective family.*

5. Response to the environment: Osmoregulation

Osmotic homeostasis depends on passive ion and water uptake through cell membranes and the intercellular space, on the active uptake or excretion of ions, and on the production and accumulation of osmolytes. To understand the ability of round goby to tolerate a wide range of salinities, we therefore compared the round goby repertoire of osmoregulatory genes to those of a stenohaline freshwater species (zebrafish) and of euryhaline species (Nile tilapia, blue-spotted mudskipper and three-spine stickleback).

Passive ion and water transport across membranes (transcellular permeability) depends on the superfamily of aquaporin proteins. Aquaporins transport water (classical aquaporins), water and glycerol (aquaglyceroporins), ammonia (aquaammoniaporins), or additional undescribed molecules (unorthodox aquaporins; **Figure 6**). Primary sequences are only moderately conserved between the classes (approximately 30% identity), but all aquaporins share six membrane-spanning segments and five connecting loops. We find 15 aquaporin genes in the round goby, which compares to the number in human (n=13) or zebrafish (n=20) and is lower than in the euryhaline Atlantic salmon (n=42) (70, 71). With 5 classical water aquaporins, 6 aquaglyceroporins, 2 aquaammonioporins, and 2 unorthodox aquaporins, the round goby features the same types of aquaporins as freshwater stenohaline fish (e.g., zebrafish) and highly euryhaline fish (e.g., tilapia; **Figure 6**).

Ion and water flow between cells in epithelia (paracellular permeability) is regulated by tight junctions, of which claudin and occludin proteins are the most important components. Mammalian genomes contain ~ 20 claudin genes, invertebrates such as *Caenorhabditis elegans* or *Drosophila melanogaster* contain 4-5 genes, and fish often feature large expansions. For example, the fugu genome contains 56 claudins, of which some occur in clusters of > 10 genes (72). The round goby genome features 40 claudin paralogues, which is in line with numbers known from other fish. All human claudin genes were represented as homologues (**Supplemental_Fig_S4**), and the round goby genome contains one occludin gene in each of the two known subclades of the protein family (**Supplemental_Fig_S5**).

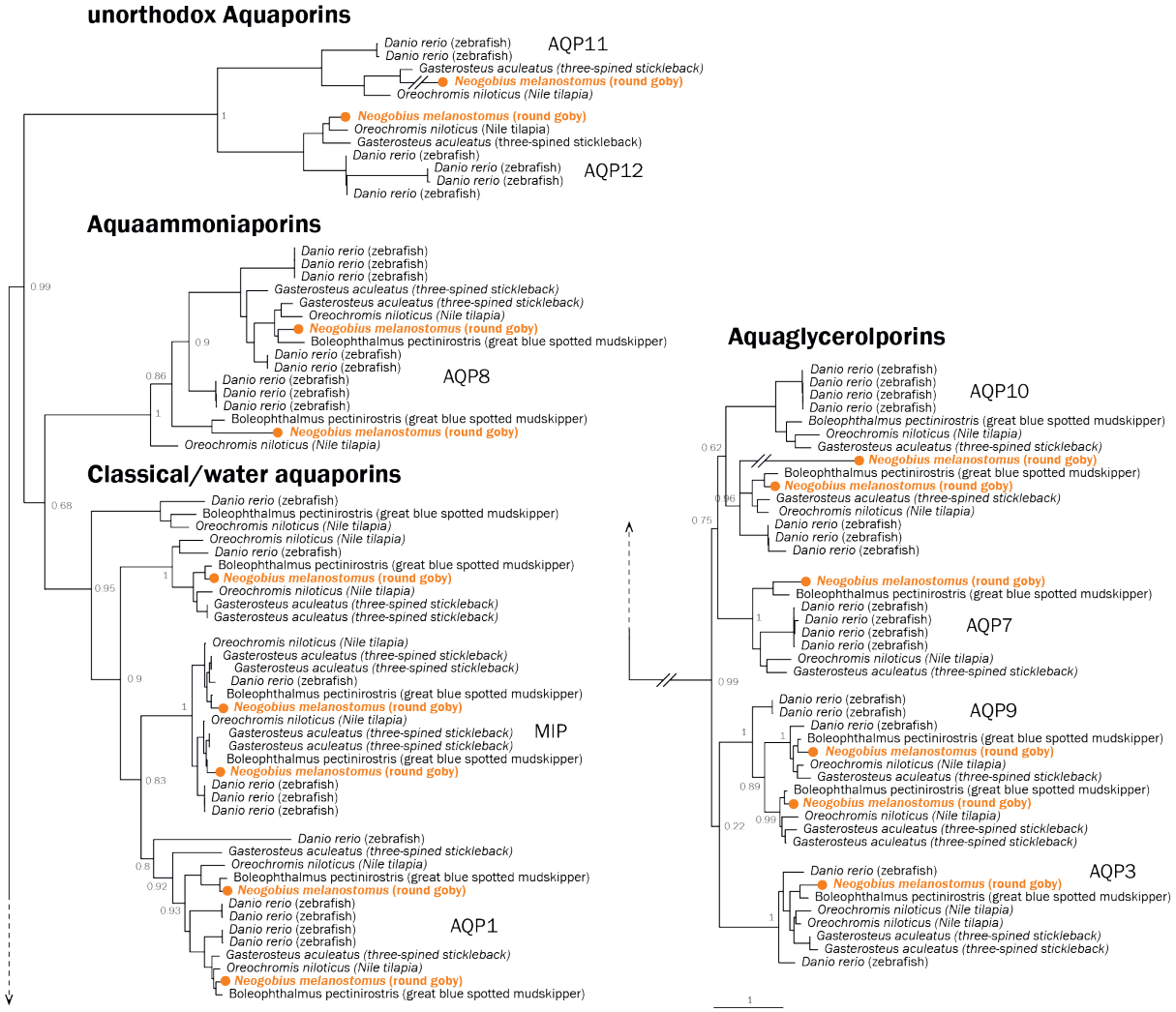
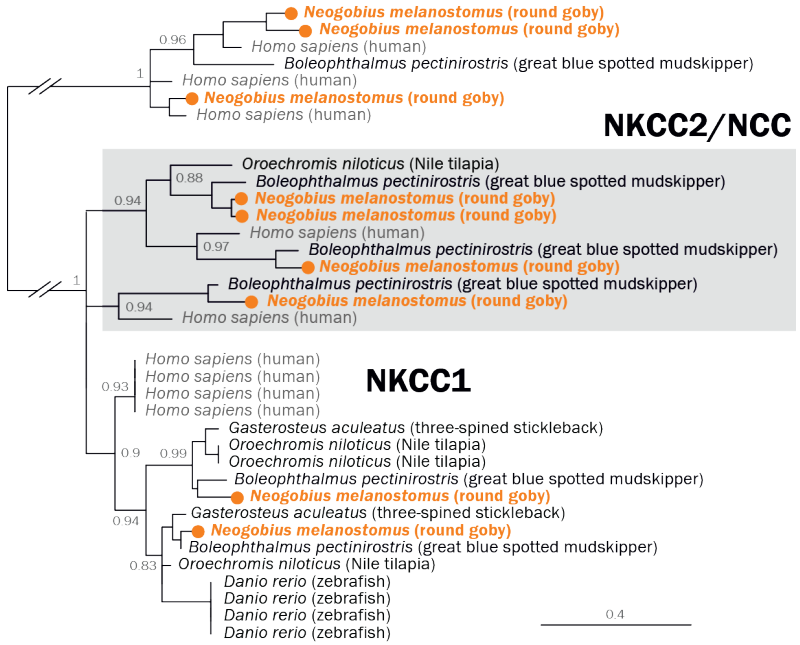


Figure 6

Phylogenetic tree of fish aquaporin proteins. *Maximum-likelihood tree with 100 bootstraps of round goby (Neogobius melanostomus, orange) in relation to cyprinid zebrafish (Danio rerio) and percomorph three spine stickleback (Gasterosteus aculeatus), nile tilapia (Oreochromis niloticus), and great blue-spotted mudskipper (Boleophthalmus pectinirostris). Zebrafish was used as outgroup in each aquaporin subfamiliy. The main classes of aquaporins are labeled with human genes names.*

In the kidney, intestine and gills, fish use active ion transport (mostly sodium transporters) to maintain osmotic balance. Mechanisms mediating sodium uptake include electroneutral Na+/H+ exchange via the NHE3b protein, Na+/Cl- cotransport via the NCC protein, and coupling of Na+ absorption with H+ secretion by a V-type H+-ATPase (73). We find 12 Na+/H+ exchanger genes, 5 Na+-K+-ATPase catalytic alpha subunits and 6 Na+-K+-ATPase regulatory beta subunits in the round goby genome.

396 The round goby thus contains the same types of genes, but less copies, than either zebrafish or tilapia
397 (**Supplemental_Fig_S6**). We find that round goby, and also mudskippers, feature an interesting
398 distribution of Na+/Cl- co-transporters to subgroups; while most zebrafish and tilapia Na+/Cl- co-
399 transporters belong to the NKCC1 subgroup, Gobiidae feature more genes in the NKCC2 subgroup
400 (**Figure 7**).



401

402 **Figure 7**

403 ***Phylogenetic tree of human and fish sodium/potassium/chloride co-transporter proteins***
404 ***(NKCC).*** *Maximum likelihood tree with 100 bootstraps of round goby (Neogobius melanostomus,*
405 *orange), zebrafish (Danio rerio), three spine stickleback (Gasterosteus aculeatus), nile tilapia*
406 *(Oreochromis niloticus), great blue-spotted mudskipper (Boleophthalmus pectinirostris) and as*
407 *contrast humans (Homo sapiens, grey). Potassium/chloride co-transporters (KCC) are used as*
408 *outgroup.*

409

410 Finally, fish produce osmolytes to actively take up and retain water. In particular, the cyclic polyol myo-
411 inositol is used by euryhaline teleosts to acclimate to high salinity. Two enzymes are required for its
412 production: myo-D inositol 3-phosphate synthase (MIPS) and inositol monophosphatase (IMPA). In
413 addition, some fish actively accumulate myo-inositol with a sodium/myo-inositol cotransporter (SMIT)
414 (74, 75). This transporter is of particular importance for marine fish exposed to high salt concentrations
415 (76, 77), while freshwater fish lack a SMIT gene (e.g. the freshwater stenohaline zebrafish lacks the

SMIT gene). The presence of SMIT has therefore been proposed to be a critical prerequisite for high salinity tolerance in fish (78). We find that the round goby genome contains MIPS and IMPA, and importantly, also a SMIT gene (**Supplemental_Fig_S7**).

6. Response to the environment: Immune System

It has been speculated that invasion success may relate to the ability to fight novel immune challenges (37). We therefore characterized key genes related to the immune system, focusing on genes that span both the innate and adaptive immune system such as pattern recognition receptors, selected cytokines and chemokines, antigen presentation, T-cell surface receptors and antibodies (**Supplemental_Table_S3; Supplemental_Table_S4**).

We find that the round goby genome features a classical adaptive immunity setup (**Table 3**). Vertebrate adaptive immunity is characterized by the Major Histocompatibility Complex (MHC) class I and MHC class II proteins and their regulators. MHCI presents antigens derived from a cell's intracellular environment, while MHCII presents antigens derived from material engulfed by macrophages, B-cells or dendritic cells (79). We find 26 full length MHCI sequences from the classic U-lineage and one sequence from the teleost-specific Z-lineage (80) (**Supplemental_Table_S5**). MHCII is represented by 8 alpha (2 fragments) and 9 beta copies (**Supplemental_Table_S6)**. The uneven numbers may be attributed to assembly issues, but also, additional small fragments were not further investigated (data not shown). We also identify the key MHC-supporting peptides Beta-2-Microglobulin, *CD74*, *TAP1/2* and *tapasin*. Beta-2-Microglobulin (*B2M*) is present in two copies, one of which contains several indels, a diverged region, and no stop codon and thus may be a pseudogene. The round goby has two copies of *TAP2*, which promotes the delivery of peptides to MHCI (annotated as *TAP2* and *TAP2T;* **Supplemental_Table_S4; Supplemental_Fig_S8**). Two *TAP2* genes have also been described in zebrafish, and our results thus suggest this is conserved feature among teleosts (81). In addition, we identify the MHC transcriptional regulators *CIITA* and *NLRC5* (**Supplemental_Table_S3**). The presence of the thymus transcription factor *AIRE* and the T-cell receptors *CD4* and *CD8* confirms the presence of helper T cells and cytotoxic T cells in the round goby.

Table 3. Overview of manually annotated key adaptive immune genes

Gene	NEME annotation	Contig annotation	Start	End	Strand
CIITA	NEME_493	Contig_2585	3 985 719	3 993 128	Antisense
AICDA	NEME_58	Contig_447	597 424	599 014	sense
AIRE	NEME_9	Contig_79	14 106 230	14 113 573	antisense
B2M	NEME_421	Contig_2242	363 050	363 352	antisense
B2M_pseudo	NEME_421	Contig_2242	368 352	368 721	antisense
CD4	NEME_213	Contig_1334	340 445	348 248	sense
CD74	NEME_71	Contig_593	791 743	796 652	antisense
CD8a	NEME_729	Contig_3231	634 222	648 487	antisense
CD8b	NEME_729	Contig_3231	656 030	660 462	antisense
RAG1	NEME_106	Contig_787	4 690 414	4 695 142	sense
RAG2	NEME_106	Contig_787	4 699 042	4 700 651	antisense
TAP1	NEME_582	Contig_2864	694 776	722 339	sense
TAP2	NEME_387	Contig_2107	2 987 106	2 993 287	antisense
TAP2T	NEME_299	Contig_1786	3 697 645	3 704 089	sense
Tapasin	NEME_387	Contig_2107	3 111 989	3 119 308	sense

Similarly, the humoral adaptive immune response (also termed the B-cell mediated immune response) is intact in the round goby. Humoral immunity in fish is characterized by three antibody isotypes consisting of immunoglobulin heavy chains delta (IgD), mu (IgM), and tau (IgT). We identify a contig-spanning immunoglobulin heavy chain locus (**Supplemental_Fig_S9**) containing 8 delta constant domains, and 4 constant mu domains, as well as genes responsible for heavy chain recombination and immunoglobulin hypermutation (*RAG1/2* and *AID(AICDA)*; **Table 3**; **Supplemental_Table_S3**). There is no evidence for the presence of immunoglobulin tau constant domains, which are commonly found in carps and salmonids (82).

While round goby adaptive immunity conforms to vertebrate standards, its innate immune repertoire displays remarkable and unusual features. We find that all components of the inflammasome (a signaling pathway involved in inflammatory responses; **Figure 8**) are expanded. Inflammasome assembly is activated through pathogen pattern recognition receptors (83), and ultimately triggers a local or systemic acute phase response by producing IL-1 family cytokines (83, 84) and/or promotes cell death via pyroptosis (84). In the round goby genome, components of the entire cascade (pattern recognition receptors, ASC adaptor proteins, IL-1, and acute phase proteins) are present in unexpectedly large numbers (**Figure 8; Supplemental_Table_S8**). In the following, our findings are described step-by-step from the cell surface down to the acute phase response.

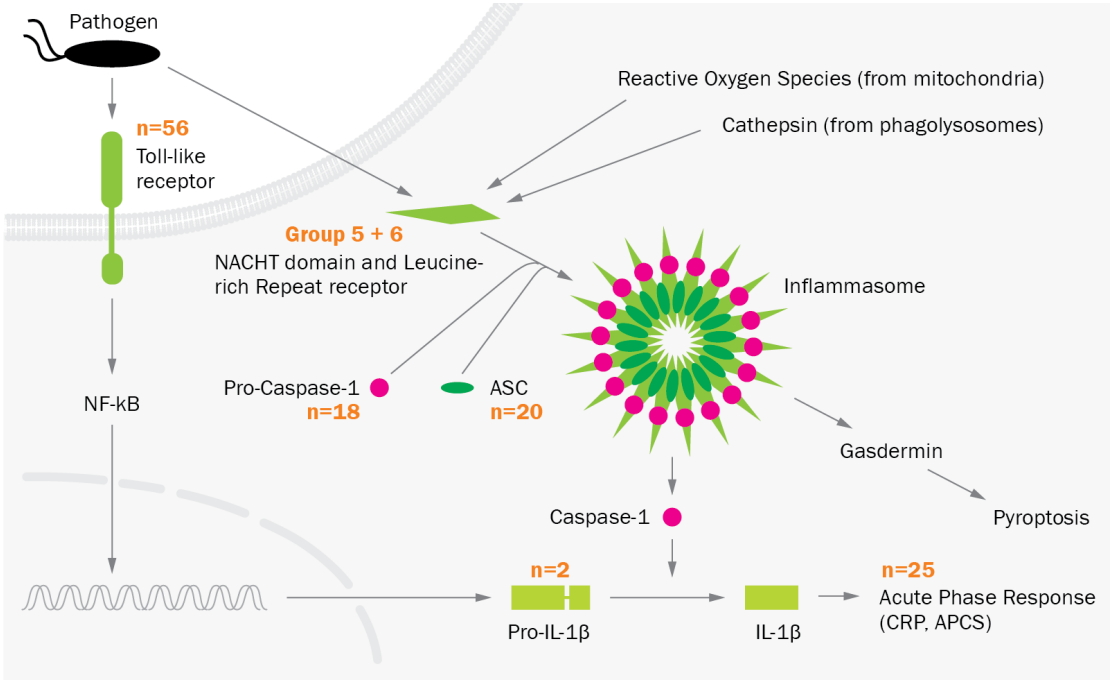


Figure 8.

The inflammasome pathway. *Several components of the pathway are expanded in the round goby (gene numbers in round goby, or novel groups for NLRs, are indicated in orange). Pathogen-associated patterns are recognized by pattern recognition receptors such as Toll-like receptors at the cell surface, or NLRs in the cytoplasm. This interaction triggers the transcription of cytokine precursors via NFkB, and the activation and assembly of inflammasome components (NLRs, Pro-Caspase-1, and ASC). Inflammasome-activated Caspase-1 then initiates the maturation of cytokines and an acute phase inflammatory response (CRP, APCS proteins), and / or pyroptosis through gasdermin.*

Perhaps the best studied pattern recognition receptors are the Toll-like Receptors (TLRs), pathogen-recognizing molecules that are generally expressed either at the plasma membrane or on the endosomal membranes. Sixteen TLR types with slightly differing ligand binding activities are conserved across vertebrates, and most vertebrate genomes contain one to three copies of each type. As expected for a teleost, the round goby genome does not contain the LPS-detecting TLR4 genes. However, in total we find 56 TLRs, of which 40 appear to originate from an expansion of Toll-Like Receptor 23-like genes (**Figure 9**). Small expansions of specific *TLRs* are somewhat common in fish (85), and indeed, we find minor *TLR22* and *TLR23* expansions to 6-13 copies in the genomes of other *Gobiidae*. However, the extent of the expansion of *TLR23* exceeds even what is observed for *TLR22*

in the relatives of cod (*Gadiformes*) (86). Phylogenetically, the identified TLR23 sequences form three clades, of which two are specific to *Gobiidae*, while the third contains TLR23 sequences from other teleosts as well (**Supplemental_Fig_S10**). In terms of genomic location, round goby TLRs 22 and 23 were distributed across several contigs with some copies arranged in tandem, which suggests several independent duplication events.

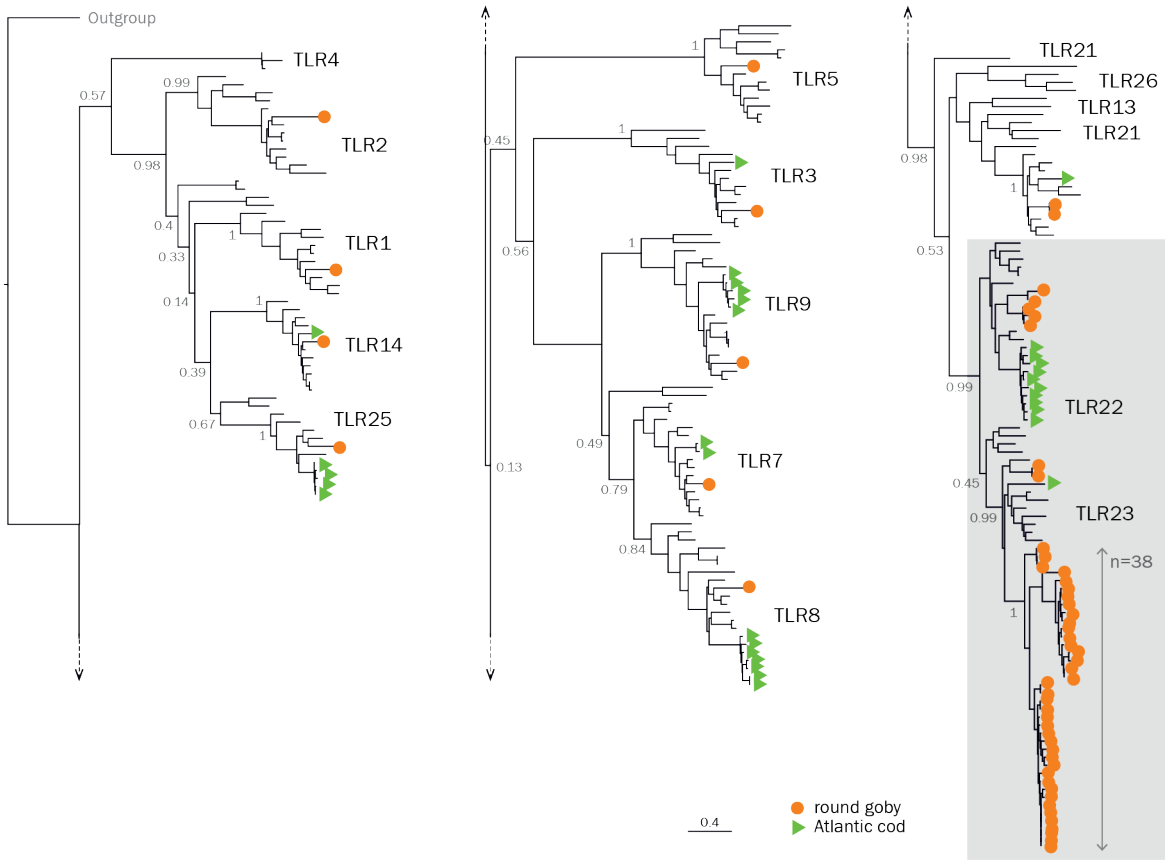


Figure 9

Phylogenetic tree of teleost Toll Like Receptor protein sequences. *A maximum likelihood phylogenetic tree run with the JTT substitution model and 500 bootstrap replicates on the transmembrane, linker and TIR domain of all TLRs found in a selected set of teleosts in the Ensembl database, the Atlantic cod genome version 2, and all manually investigated Gobiiformes. A TLR sequence from the lancelet Branchiostoma belcheri was used as an outgroup and the root was placed upon its corresponding branch. Green triangles, Atlantic cod. Orange circles, round goby. Grey box, TLR22 and TLR23.*

For intracellular pathogen recognition receptors of the NACHT domain and Leucine-rich Repeat containing receptor (NLR) family, we identify two new, previously undescribed families (Group 5 and 6) present in the round goby and also in the mudskipper *B. pectinirostris* (**Figure 10**).

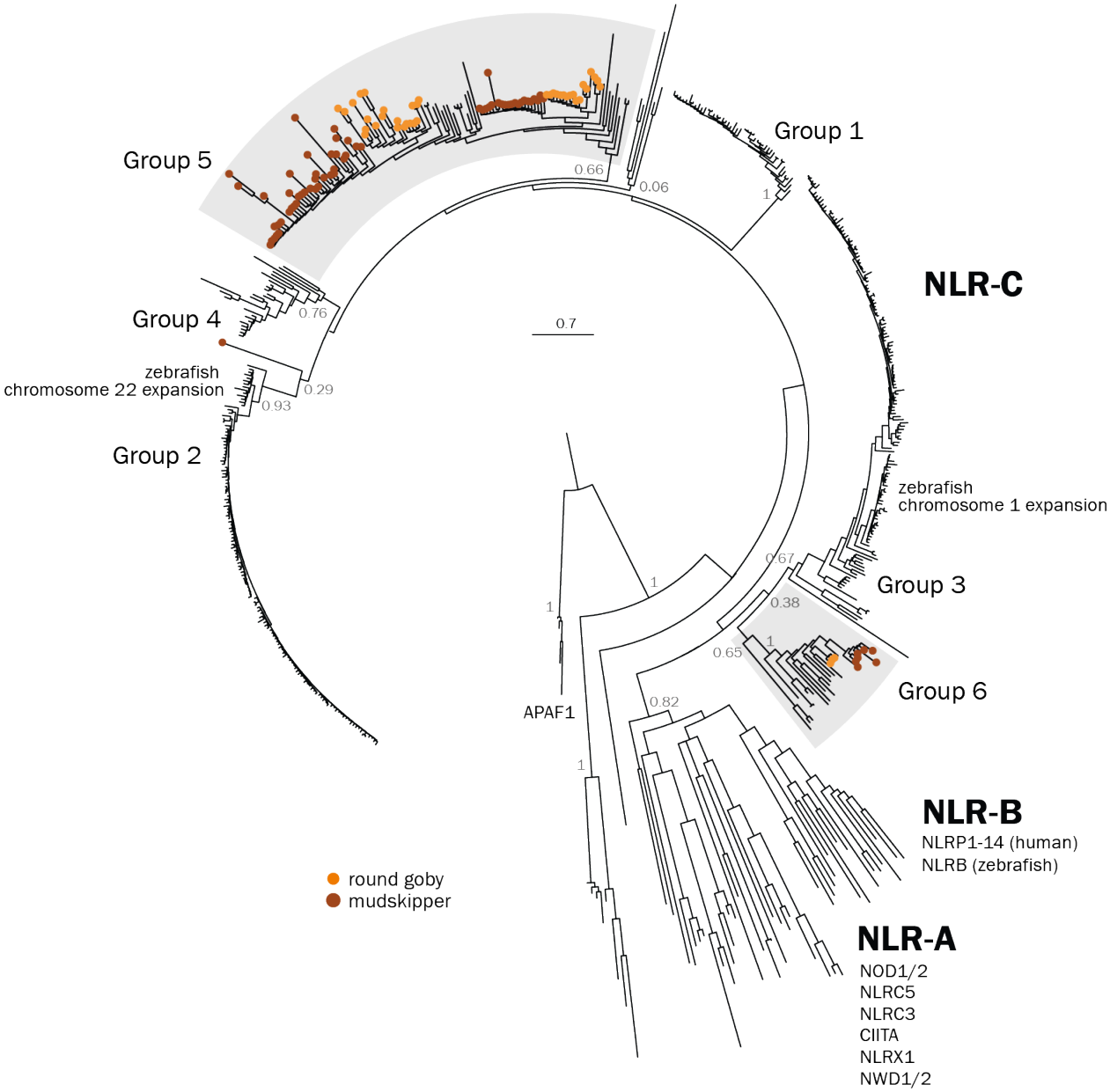


Figure 10

Phylogenetic tree of the NLR nucleotide-binding domain sequences *in round goby, great blue dotted mudskipper, zebrafish and human. Maximum Likelihood phylogenetic tree with 500 bootstraps rooted at the split between NB-ARC (found in APAF) and NACHT domains (present in all the other NLRs). NB-ARC domains from APAF1 orthologs were used as an outgroup. Bootstrap values are shown for nodes that determine an entire cluster. The tree resolves all three major classes of vertebrate NLRs (NLR-A, NLR-B, NLR-C). NLR-A genes were conserved in all analyzed species, no*

513 *NLR-B genes were found from the gobies. Six groups of NLR-C genes were identified, four of which*
514 *are exclusive to zebrafish (Group 1-4) and two contain only sequences from gobies (Groups 5 and 6).*
515 *Within the goby-specific groups, lineage-specific expansions can be seen for both round goby*
516 *(orange) and mudskipper (brown).*
517
518 NLRs have diverse roles from direct pathogen recognition to transcriptional regulation of the MHC
519 (NLRs CIITA and NLRC5) and contribute to inflammasome activation. Mammalian genomes display
520 20-40 NLRs in families NLR-A and NLR-B, while fish also feature a fish-specific subfamily (NLR-C)
521 (87) and a much expanded NLR repertoire (e.g. 405 NLR-C genes in zebrafish) (88, 89). The round
522 goby genome contains at least 353 NLRs (**Supplemental_Table_S8**), which include 9 highly
523 conserved vertebrate NLRs (*NOD1*, *NOD2*, *NLRC3*, *NLRC5*, *NLRX1*, *NWD1*, *NWD2*, *APAF1*, *CIITA*)
524 as well as 344 NLR-C genes. Fish NLRs cluster into 6 groups of which 2 represent novel NLR-C
525 clades (groups 5 and 6, **Figure 10**). The novel groups are supported by phylogenetic analyses as well
526 as motif presence/absence (**Table 4)**. NLR-C groups are characterized by highly conserved versions
527 of the sequence motif Walker A. The most common sequence for Walker A observed in both goby
528 NLR-C groups, GVAGVGKT, is not associated with any of the four major NLR-C groups in zebrafish
529 (88). Also, NLR subtypes often carry group-specific combinations of the protein-protein interaction
530 domain PYD and/or B30.2 domain. This holds true for *Gobiidae* NLR-C groups, since only group 5
531 NLRs can carry an N-terminal PYD domain and/or a C-terminal B30.2 domain (88), similar to the
532 zebrafish Group 1 and 2 NLRs (**Table 4**). In contrast, some group 6 NLRs have C-terminal CARD
533 domains, which in both human and zebrafish are attached to specific inflammasome-associated NLR-
534 B genes (90). The round goby C-terminal CARD-containing NLRs are found on the same few scaffolds
535 and share a high degree of sequence similarity, indicative of a recent expansion. This expansion is
536 absent from mudskipper and thus restricted to the round goby lineage. Many other Group 6 NLRs are
537 fragmented, with large insertions in the middle of their conserved 2 kb exon
538 (**Supplemental_Table_S8**).
539
540

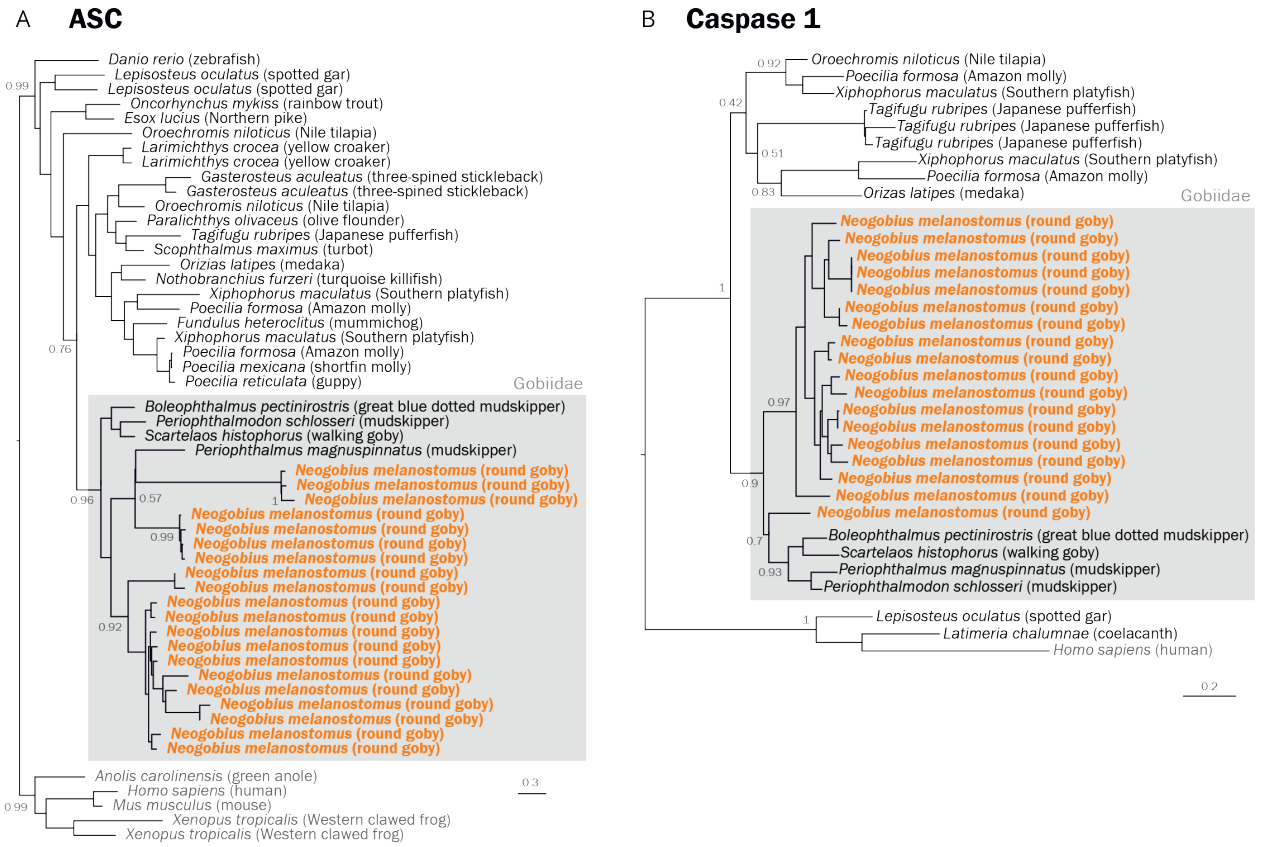
Table 4. Key features of each of the six NLR-C subgroups. x denotes a variable amino acid.

Group	Identified in this study	Walker A	Last residues of the largest exon	PYD?	B30.2?
1		GIAGVGKT	L(I/M)PVVKNT(T/R)RA	+	+
2		GVAGIGKS	LSAVIKTSKRA	+	+
3		GIAGIGKT	L(IP/TA)AV(R/S)NC(RK/TR/RR)A	-	+
4		GVAGIGKT	LPV(I/V)xxxx(A/V)x	-	-
5	x	GVAG(V/A/I)GKT	(L/M)PV(V/I)KASxK(A/V)	+	+
6	x	GVAG(V/A)GKT	L(I/V)P(A/V)VRNCRKA	-	-

Once activated, some NLRs (including those with a C-terminal CARD) can oligomerize and form an inflammasome in order to activate specific caspases (usually Caspase 1, **Figure 8**). The interaction between NLRs and the caspase are mediated by the adaptor protein ASC (also known as PYCARD), which itself oligomerizes into large structures known as “specks” (91). Vertebrates have 1-2 copies of ASC, which are characterized by a characteristic combination of a single PYD and CARD domain. In the round goby genome, we find 20 cases of this domain combination. Since the genomes of other gobies contain 1-2 PYD-ASC combinations, the expansion appears to be specific to the round goby (**Figure 11A**). The effector protein Caspase 1 is present as one gene in humans and as two genes in zebrafish. We find that the round goby genome features an expansion to 18 copies. Interestingly, some of those genes appear to contain a CARD domain (as seen in mammals and several species of fish) while others have PYD (as seen in zebrafish). This suggests that a caspase with both domains may have existed in the common ancestor of fish and tetrapods, with most lineages having retained only one of the two. However, phylogenetic analyses reveal that all round goby Caspase 1 genes are the result of a single expansion event specific to this species (**Figure 11B**). An alternative explanation for the presence of both PYD- and CARD-caspases 1 genes would be a recurrent acquisition of PYD in different lineages. In any case, in addition to Caspase 1 genes, caspase 3 (a key component of apoptosis which may be activated by Caspase 1) is also expanded to 5 copies. Caspase 4 and 5, on the other hand, appear to be absent.

Finally, we find that genes encoding for two peptides produced in the course of inflammation, the acute phase reactants C-reactive protein (CRP) and serum amyloid component P (APCS), are expanded to a total of 25 copies (compared to 2-7 in fish, and 5-19 in the other *Gobiidae*). In fish, CRP and APCS are closely related and cannot be distinguished based on BLAST scores or phylogeny. As

566 seen in other fish species, all investigated CRP/APCS sequences resolve into two major phylogenetic
567 clades, with the mammalian sequences in a third (**Supplemental_Figure_S11**).
568



569

570 **Figure 11**

571 ***A Phylogenetic tree of gnathostome ASC protein sequences.*** *Maximum Likelihood phylogenetic*
572 *tree with 500 bootstraps rooted at the split between tetrapods and ray-finned fish. Tetrapods were*
573 *used as outgroup. Round goby is indicated in orange. Gobiidae are highlighted with a grey box. The*
574 *goby sequences form a clear separate cluster (marked with the box), with a large expansion apparent*
575 *in the round goby.* ***B Phylogenetic tree of gnathostome Caspase 1 protein sequences*** *The*
576 *Caspase 1 tree comprises all protein sequences annotated as CASP1 in the investigated Gobiiformes*
577 *genomes aligned together with reference sequences from Ensembl and GenBank. The root was*
578 *placed on the branch containing the mammalian sequences.*
579
580

7. Adaptation to novel environments: Epigenetic regulators

The PRC2 complex establishes and maintains gene repression (92) and thus represents a plasticity-restricting mechanism. The complex mediates di- and trimethylation of lysine 27 on histone H3 and contains four proteins: a catalytic subunit (either *enhancer of zeste* EZH1 or EZH2), *suppressor of zeste* SUZ12, *embryonic ectoderm development* EED and *RB Binding Protein 4* RBBP4 (50). In mammals, the alternative catalytic subunits EZH1 and EZH2 have partially complementary roles (93, 94), and requirements for the two alternative catalytic subunits differ between species – in contrast to mammals, zebrafish develop in the absence of either catalytic subunit (95, 96). We find that the round goby genome contains the usual complement of PRC2 components: two copies of SUZ12 (of which one appears quite diverged), one copy of EED, one copy of RBBP4, and two copies of EZH (with multiple isoforms determined by RACE experiments). For SUZ12, EED, and RBBP4, sequence-based identification was straightforward, and phylogenetic analyses followed the known phylogenetic relationships of fish, mammals, and other vertebrates (**Supplemental_Fig_S12**). The catalytically active subunits EZH1 and EZH2 do cluster with the closest species in the phylogeny, the mudskipper *B. pectinirostris* (**Figure 12**), but the deeper relationships within EZH2 are poorly supported and may suggest a complex evolutionary history.

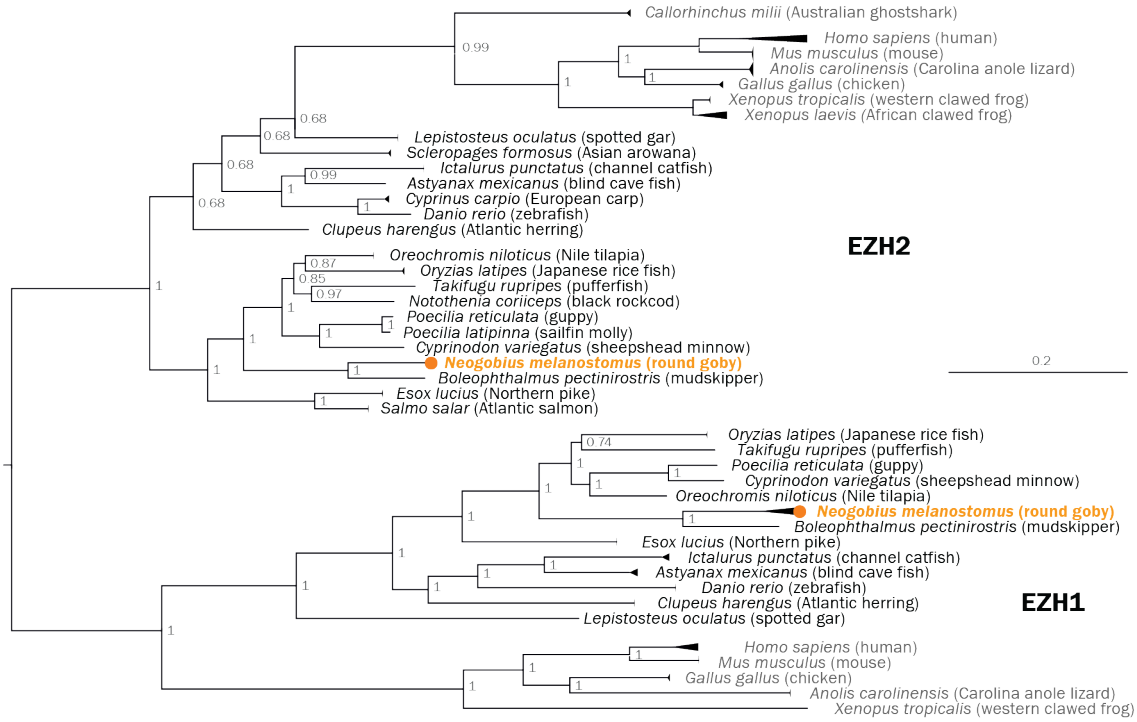


Figure 12

Phylogenetic tree of vertebrate EZH proteins. *Midpoint-rooted Bayesian phylogenetic tree. Note the position of the Australian ghost shark (potential outgroup) within the poorly supported EZH2 branch. When rooting with Australian ghost shark, teleost EZH2 genes cluster with EZH1 (data not shown). Round goby is indicated in orange.*

Methylation marks similarly regulate gene expression and are deposited by conserved enzymes called DNA methyltransferases (DNMTs). Mammals feature two types of DNMTs, DNMT3 (three genes A, B, and L) and DNMT1 (one gene). The two types perform both *de novo* and maintenance methylation, respectively, in a dynamic division of labor (97). Interestingly, fish feature a variable repertoire of DNMT3 genes. Medaka, fugu, zebrafish, and carp have three, five, six, and twelve DNMT3 genes, respectively (98). We find that the round goby genome features one DNMT1 that follows the expected phylogenies, and five DNMT3 genes, of which two cluster with vertebrate DNMT3A sequences, and three with vertebrate DNMT3B sequences (**Figure 13**). The number of DNMT3 genes in round goby corresponds to that seen in in stickleback, fugu and tilapia (99). In general, the DNMT3 phylogeny is not well supported, which limits conclusions about the evolution of specific DNMT3 genes.

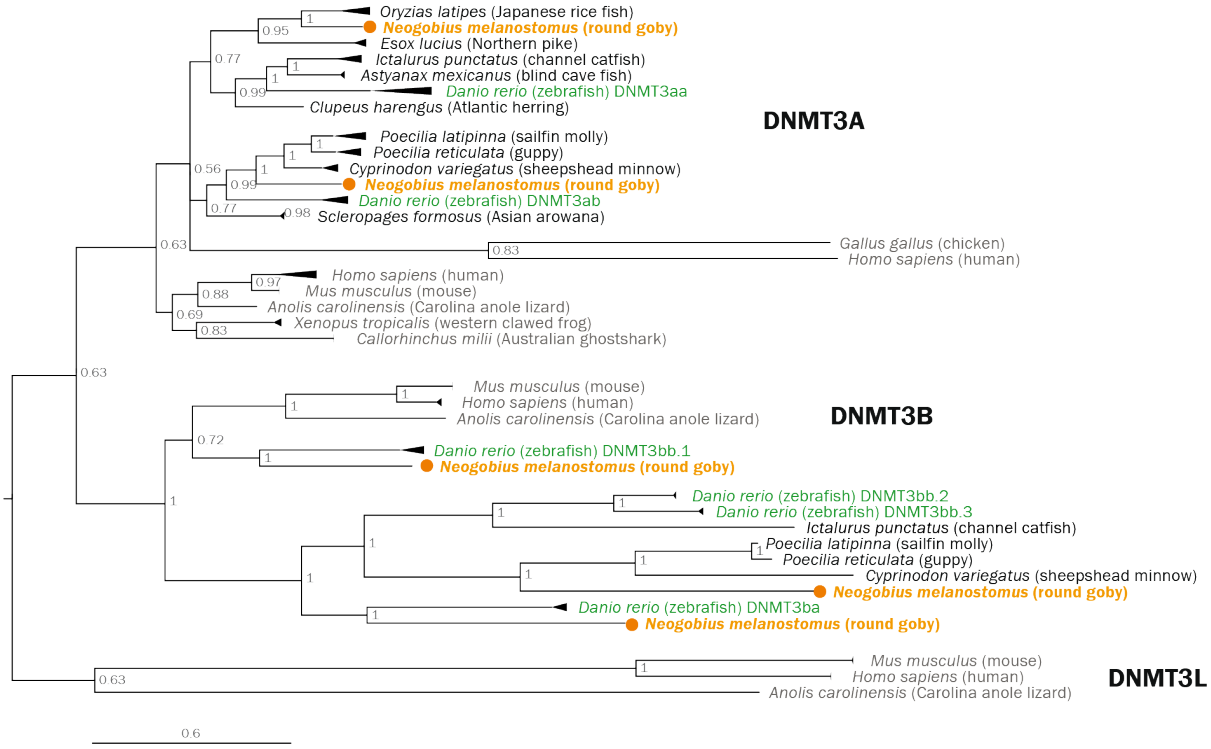


Figure 13

Phylogenetic tree of vertebrate DNMT3 proteins. *Midpoint-rooted Bayesian phylogenetic tree. The Australian ghost shark (potential outgroup) is positioned among DNMT3A genes. Round goby is indicated in orange. Zebrafish, the only other fish with well-annotated DNMT3 genes, is indicated in green.*

Discussion

General observations

Our analyses depict a genome that, in many respects, is similar to other teleost genomes. There is no evidence for recent genome duplications, and genome size, gene content and GC content are within the ordinary range. Transposable elements can create genetic variation and mediate ecological success (100), but repeat analyses do not reveal unusual transposon activities in the round goby. Small genome size has been proposed to foster invasiveness (101), but the round goby genome is not particularly small. Phylogenetic analyses reveal that many of the analyzed gene families conform to expectations. For example, green opsin gene duplications and the loss of the UV opsin are observed in many fish lineages (20). Similarly, the expected gene families and overall gene complements are found for olfactory receptors, cytochrome P450, and osmoregulatory proteins, for adaptive immunity and epigenetic regulators. Multilocus sex determination has previously been suggested for freshwater gobies (102), and indeed our data suggest a multigenic and/or environmental sex determination system for the round goby, rather than a large sex-determining region or a sex chromosome. Overall, these findings support the validity of the sequencing and assembly approach, and suggest that selected findings of interest are not artefacts. In addition, the round goby genome sequence also reveals several novel and interesting findings of which some pertain to teleost genomes in general, some to *Gobiidae*, and some to specific gene families, with possible implications for invasive potential.

Environmental perception

We find that the visual system of *Gobiidae* may be more efficient in the red parts of the light spectrum. This is intriguing considering the benthic life style of gobies and their occurrence in turbid areas. In clear waters, red light from the sun is the least abundant part of the spectrum (and virtually absent

651 below 15m of depth) because red light penetrates least through water, but many organisms convert
652 the deeply penetrating green and blue wavelengths into red. Indeed, the eyes of gammarids, a
653 common prey of round goby, strongly reflect red light (103). An enhanced red perception through an
654 additional red opsin gene may thus be relevant for round goby predation success. In turbid waters, red
655 is the most common part of the light spectrum because long wavelengths experience least scattering
656 (21). Round gobies readily establish populations in turbid environments. The retention of two red opsin
657 genes may thus possibly relate to the remarkable ability of the round goby to colonize turbid habitats.
658 Our predictions based on the key amino-acid substitution suggest that LWS1 is expected to be most
659 sensitive at 560 nm (same as one of the mudskipper gobies) (25), while LWS2 is expected to be most
660 sensitive at 550 nm (59). Similar small differences in the sensitivity maximum can indeed result in
661 functionally different spectral tuning palettes (e.g. during development or in different environmental
662 conditions) (104).
663
664 The presence of red fluorescence on top of the eye in round goby is the first unequivocal description
665 of fluorescence in a freshwater fish and might be interpreted as being associated with the ability to
666 discriminate different shades of red colors. However, the fluorescence in the specimens investigated
667 was quite weak. Unless fluorescence expression is stronger under natural conditions or in the
668 ancestral population from which the invading populations stem, a visual function of the weak
669 fluorescence observed here seems unlikely (see warnings by (105)). Fluorescence is, however,
670 widespread and stronger among several marine gobies (106). Although the fluorescent "eyebrows" of
671 the round goby show a striking similarity to those of some marine gobies, their function will remain
672 unclear until properly tested. Social functions are possible – for example, in sand gobies, dark eyes
673 indicate female readiness to spawn (107). Alternatively, they may simply provide camouflage for
674 individuals buried in bright sand up to the eyes. Functional hypotheses for fluorescence, such as
675 communication, camouflage and improved prey detection have been extensively reviewed by (108).
676 The genetic tools now available for the round goby may allow for experimental manipulation of
677 fluorescence expression, once the actual fluorophores that produce the fluorescent signal have been
678 identified.
679
680 **Response to the environment**
681

With respect to ecological and physiological aspects of success and invasiveness, some findings on CYP genes, on osmoregulation, and on innate immunity call for further attention. The mostly minimal complement of cytochrome P450 proteins present in the round goby is unexpected considering the occurrence of round goby in polluted areas (109, 110). The CYP1-3 gene complement for xenobiotic metabolism is similar to other teleost genomes, and the ability of the round goby to survive in contaminated environments must therefore have other reasons. Round goby may cope with contaminations at the level of gene expression, either through higher basal expression values or by a particularly rapid response to exposure (45). Alternatively, this species may have peculiarities in other, not yet analyzed areas of the defensome (e.g. transporters). Analyses of the tissue expression of CYP families 1, 2 and 3, and also the study of other defensome gene families, including the nuclear receptors regulating CYP gene expression, transporters and conjugating enzyme families, may be useful in this respect.

Another potentially relevant finding is the ability of the round goby to not only produce, but also accumulate osmolytes. Species distribution constraints often arise from physiological limitations. The round goby is one of the most geographically wide-ranging invasive fish species in Europe and North America, and the ability to accumulate osmolytes may impact its range expansion in three ways. Firstly, 0-25 PSU (common for coastal waters, but lower than the ocean) is the species' current limit for unperturbed osmoregulation (111). However, the round goby's repertoire of key-genes in myo-inositol production and accumulation might bestow the species with the potential to eventually tolerate higher salinities, for example through the evolution of altered gene regulation patterns, and colonize higher PSUs. Secondly, osmolytes improve water retention and thus desiccation tolerance. In this context, myo-inositol accumulation may have contributed to overland dispersal. Overland dispersal of eggs or larvae with boats or fishing gear involves air exposure, and indeed, round goby eggs withstand desiccation for up to 48 hours (112). Finally, osmolytes essentially act as anti-freeze agents and molecular chaperones, and contribute to cryoprotection in diverse organisms from bacteria (113) to flies (114). Osmolytes may thus enable the round goby to combat a number of environmental conditions and to colonize new areas. The surprising and unexpected ability of the round goby to colonize cold areas well below its temperature optimum of 22°C, such as the Northern Baltic Sea, may be linked to osmolyte production.

Lastly, the "strike fast, strike hard" innate immune system and the impressively large inflammation machinery of the round goby may enhance the species' colonization potential. Fish immunity appears to be quite plastic. For example, cod have disposed of some core adaptive immunity components (18), yellow croaker feature an expanded TNF repertoire (115), and channel catfish retain a high number of recent duplications and SNPs in immune genes (116). Meanwhile, in salmonids, genes specifically retained after the 4th whole genome duplication are not immune genes (117).

We find that the round goby genome contains multiple copies of genes for inflammasome assembly, activation, and function. This is interesting because the fish inflammasome complex is much more poorly characterized than that of mammals. Maturation of IL-1 by inflammasome-activated Caspase 1 cleavage in fish is a matter of debate, since teleost IL-1 proteins lack the conserved caspase cleavage site present in mammalian IL-1b and IL-18 (118). However, as has been shown for zebrafish, Caspase 1 can also utilize an alternative site to cleave and mature IL-1 (90, 119). In any case, the caspases also mediate cell death via pyroptosis and the presence of other components such as ASC, caspases and pro-IL1 and pro-IL18 supports a role for inflammasomes in fish. Zebrafish ASC oligomerize and form "specks" as seen in mammals (90). The molecular dynamics of inflammasome activation therefore represent a potential future research avenue in the round goby.

In terms of ecological success, the round goby's expanded repertoire of pathogen recognition receptors may broaden the scope of its immune response and increase the range of detectable ligands and pathogens. The expanded acute phase repertoire may also contribute to a fast response, or inversely, may limit excessive cell damage. In humans, the acute phase protein CRP contains inflammation as part of a negative feedback loop (120). Thus, the round goby may re-enter homeostasis faster compared to other fish species with smaller CRP/APCS repertoires. The larger acute phase repertoire may also function to limit the cellular damage caused by the potentially large amount of inflammasome combinations the round goby can generate. In this context, we suggest systematic investigations into a potential relation between inflammasome expansions and invasiveness in *Gobiidae*, in combination with immune challenge experiments.

Long-term adaptation

744 We identify a potentially interesting evolutionary history for the conserved PRC2 component EZH in
745 fish, and add to the previous observation that the conserved *de novo* DNA methylation machinery
746 features a surprising diversity in fish. These results underscore the need for in-depth investigations
747 into the role and relevance of epigenetic regulation and transgenerational inheritance in teleosts. Our
748 findings support the emerging idea that epigenetic regulation in fish follows somewhat different rules
749 than in mammals. For the histone methylating complex PRC2, our results suggest interesting
750 phylogenetic relationships of EZH proteins in fish. EZH proteins act in tissue specific complexes
751 comprised of core SUZ12, EED, and RBBP4, but also AEBP2, PCL proteins and JARID2. These
752 proteins enhance PRC2 efficiency, contribute to recruitment to target sites, or inhibit the complex (50,
753 92). Small sequence changes can have strong effects on the entire complex, since the precise
754 interactions among the components and with other gene regulators impact its function and localization
755 (121–124). For example, species-specific insertions (125) are thought to regulate PRC2 recruitment
756 and/or exclusion from target genes (126). We suggest that the future incorporation of more sequences
757 of both EZH1 and EZH2 from a greater range of taxa and the inclusion of currently unannotated
758 versions of the genes associated with both the teleost specific whole genome duplication and lineage
759 specific duplications (96) would aid understanding of the evolutionary history of the entire complex. We
760 expect that studying PRC2 in non-mammalian vertebrates may reveal ancestral or less abundant
761 interactions, functions or also complex associations of PRC2.
762
763 Similarly, our results warrant an in-depth exploration of DNA methylation in fish. Originally, DNA
764 methylation evolved to distinguish own (methylated) DNA from foreign (non-methylated) DNA such as
765 introduced by viruses. Therefore, cytosines in CG base contexts are by default methylated. In
766 mammals, DNA methylation in CG dense regions (CG islands) is associated with gene repression.
767 However, DNA methylation also features species- and taxon-specific differences, even among
768 vertebrates, which are still greatly underappreciated. For example, non-methylated genome regions in
769 fish are unexpectedly CG-poor (127), fish differ from mammals with respect to the distribution of
770 methylated CpGs in the genome (128), algorithms developed on mammals fail to identify CpG islands
771 in fish (129), genome-wide CpG island predictions in cold-blooded animals consist primarily of false
772 positives (130), and fish CG methylation occurs mainly in coding regions, where it correlates positively
773 with gene expression levels (131). These curious differences are further enhanced by the seemingly
774 random copy number variation in the *de novo* DNA methyltransferase DNMT3 in teleosts, which do not

reflect genome duplication events in teleosts (99). DNMT3 genes display highly spatiotemporal expression patterns particularly during development (132–135), and an in-depth and species-aware exploration of the role of DNA methylation in fish is clearly warranted.

Gene expansions

A general theme across several of the analyzed gene families is gene expansions. Gene expansions have been linked to invasive potential before (136) and are recurrent in fish genomes, both within (117, 137) and outside (116, 138, 139) the context of whole genome duplications. Many duplicated genes are known to experience rapid neofunctionalization rather than subfunctionalization (137), and have the potential to compensate against mutation even after divergence (140). The round goby and its relatives are definitely strong candidates for further and systematic investigation of a link between gene expansions and colonization or invasion potential. The *Benthophilinae* group is recently diversified crowd of fish with many members inexplicably on the move (141), and *Gobiidae* in general share a remarkable colonization potential (10, 142). Importantly, recent gene expansions can be difficult to resolve with short reads, and genomes based on long read sequencing (as presented here) will be instrumental in this regard.

Among the receptor families analyzed, the NLRs, TLRs, and olfactory receptors, we identify a couple of particularly beautiful case studies for recent expansions and repeated radiations. Our identification of two previously undescribed NLR-C gene families (88), here termed group 5 and group 6, indicates substantial diversification of NLRs in fish. Different teleost lineages appear to feature different NLR-C subfamilies with large lineage-specific expansions reminiscent of olfactory receptor repertoires. Similarly, we identify interesting cases of parallel expansions across families, and also family-specific expansions, among olfactory receptors. Both cases warrant investigations into the evolution of ligand binding repertoires. For example, 7tm1 subfamily members may be involved in the detection of distinctive types of odors relevant for round goby, and possibly, *Gobiidae* ecology (28–30). Which types of odorants are detected by parallel expanded ORs, and whether these expansions serve to detect similar or different types of odorant molecules in different species, remains to be studied. Finally, the massively expanded TLR22 and TLR23 families warrant an exploration of their ligand binding properties. TLR22 and TLR23 have been suggested to recognize nucleic acid ligands (85), but

806 some also react to protein or lipid pathogen-associated patterns (143–145), and their role in fish is
807 currently unclear.
808
809 In summary, this work provides a solid basis for future research on the genomic, genetic, and
810 epigenetic basis of ecological success. Clearly, many more gene families or pathways may contribute
811 to the round goby's invasion success. For example, the presented analyses barely scratch the surface
812 of epigenetic regulation, innate immunity and transporters (e.g. of toxins). We did not investigate
813 endocrine pathways (which govern growth and reproductive success) nor antimicrobial peptides
814 (which contribute to innate immune defense), areas which may yield fruitful information of the success
815 of this invader. We welcome future research using this novel genomic resource, and encourage
816 experts on those pathways to contribute their knowledge.

Methods

A relevant note upfront is that this manuscript is the product of a long-standing collaboration of leading experts in their respective fields. The gene families analyzed differ widely with regard to sequence conservation, the number and similarity of genes within and between species, the scope of questions in the field, etc. Compare, for example, the *de novo* identification of hundreds of virtually identical NLR receptors with the manual annotation of a handful of extremely conserved DNA methyltransferases, or the phylogenetic analysis of the conserved vertebrate CYP gene family with a fish-centered comparison of osmotic balance regulators. Accordingly, each collaborator applied methods that were suited for the respective situation. As a common theme, genes were identified by blast, sequences were extracted and aligned with other fish and/or other vertebrates, trees were constructed with either bayesian or maximum-likelihood methods, and findings were always verified against the mudskipper genomes.

Genomic DNA library preparation and PacBio sequencing

Genomic DNA was extracted from the liver of one male individual of round goby caught in Basel, Switzerland (47° 35′ 18″ N, 7° 35′ 26″ E). At the Genome Center Dresden, Germany, 300 mg of liver tissue were ground by mortar and pestle in liquid nitrogen and lysed in Qiagen G2 lysis buffer with Proteinase K. RNA was digested by RNase A treatment. Proteins and fat were removed with two cycles of phenol-chloroform extraction and two cycles of chloroform extraction. Then, DNA was precipitated in 100% ice cold ethanol, spooled onto a glass hook, eluted in 1x TE buffer, and stored at 4 °C. 10 µg of DNA was cleaned using AMPure beads. From this DNA, five long insert libraries were prepared for PacBio sequencing according to the manufacturer's protocols. Genomic DNA was sheared to 30-40 kb using the Megaruptor device. The PacBio libraries were size selected for fragments larger than 15-17.5 kb using the BluePippin device. PacBio SMRT sequencing was performed with the P6/C4 chemistry using 240 min sequencing runs. Average read length was 11-12 kb. In total, 86 SMRT cells were sequenced on the PacBio RSII instrument resulting in 46 gigabases (Gb; an estimated 46x coverage for a putative ~1 Gb genome) polymerase reads.

Assembly of the round goby genome

The round goby genome was assembled at the Heidelberg Institute for Theoretical Studies HITS gGmbH. Raw PacBio reads were assembled using the Marvel (146, 147) assembler with default parameters unless mentioned otherwise. Marvel consisted of three major steps, namely the setup phase, patch phase and the assembly phase. In the setup phase, reads were filtered by choosing only the best read of each Zero-Mode Waveguide as defined by the H5dextract tool (146) and requiring subsequently a minimum read length of 4k. The resulting 3.2 million reads were stored in an internal Marvel database. The patch phase detected and fixed read artefacts including missed adapters, polymerase strand jumps, chimeric reads and long low-quality segments that were the primary impediments to long contiguous assemblies (146). To better resolve those artefacts only low complexity regions were masked (DBdust) and no further repeat masking was done. The resulting patched reads longer than 3k (41x coverage) were then used for the final assembly phase. The assembly phase stitched short alignment artefacts from bad sequencing segments within overlapping read pairs. This step was followed by repeat annotation and the generation of the overlap graph, which was subsequently toured in order to generate the final contigs. By using an alignment-based approach, the final contigs were separated into a primary set and an alternative set containing bubbles and spurs in an overlap graph. To correct base errors, we first used the correction module of Marvel, which made use of the final overlap graph and corrected only the reads that were used to build the contigs. After tracking the raw reads to contigs, PacBio's Quiver (148) algorithm was applied twice to further polish contigs as previously described (146).

Automated annotation of the round goby genome

The round goby genome assembly was annotated using Maker v2.31.8 (149, 150). Two iterations were run with assembled transcripts from round goby embryonic tissue (46) and data from eleven other actinopterygian species available in the ENSEMBL database (downloaded the 15th February 2016, http://www.ensembl.org, see **Table 5**) as well as the SwissProt protein set from the uniprot database as evidence (downloaded March 2, 2016; https://www.uniprot.org/downloads). In addition, an initial set of reference sequences obtained from a closely related species, the sand goby *Pomatoschistus minutus*, sequenced by the

879 CeMEB consortium at University of Gothenburg, Sweden (https://cemeb.science.gu.se), was
880 included. The second maker iteration was run after first training the gene modeler SNAP
881 version 2006-07-28 (151) based on the results from the first run. Augustus v3.2.2 (152) was
882 run with initial parameter settings from Zebrafish. Repeat regions in the genome were masked
883 using RepeatMasker known elements (153) and repeat libraries from Repbase (154) as well as
884 *de novo* identified repeats from the round goby genome assembly obtained from a
885 RepeatModeler analysis (153).
886
887 **Table 5**. Summary of reference data from Ensembl used for the annotation.

Reference species	Number of protein sequences	Assembly version from ENSEMBL (downloaded 15th Feb 2016)
Astyanax mexicanus	23698	AstMex102
Danio rerio	44487	GRCz10
Gadus morhua	22100	gadMor1
Gasterosteus aculeatus	27576	BROADS1
Lepisosteus oculatus	22483	LepOcu1
Oreochromis niloticus	26763	Orenil1.0
Oryzias latipes	24674	MEDAKA1
Poecilia formosa	30898	PoeFor_5.1.2
Takifugu rubripes	47841	FUGU4
Tetraodon nigroviridis	23118	TETRAODON8
Xiphophorus maculatus	20454	Xipmac4.4.2

888
889 In order to ensure the completeness and quality of the current assembly and the associated gene
890 models, the assembly and the predicted protein sequences were run against reference sets at two
891 different taxonomical levels (303 eukaryotic and 4584 actinopterygian single copy orthologues) using
892 the BUSCO pipeline v2.0 (155, 156).
893
894 The maker annotation results were used to generate a database for JBrowse/Webapollo using the
895 script “maker2jbrowse” included with JBrowse (157, 158). Predicted protein and transcript sequences
896 were used to query the uniprot database, using blastp and blastn respectively, and the best hit
897 descriptions were transferred to the fasta headers with scripts bundled with Maker as described in
898 (150). The annotated genome is currently hosted on a WebApollo genome browser and Blast server at
899 the University of Gothenburg, Sweden at http://albiorix.bioenv.gu.se/.
900 Our analyses reveal that some degree of care is warranted regarding gene models. *De novo*
901 annotation without transcriptome data tends to be biased towards known and conserved genes,
902 homopolymer sequencing errors may cause annotation errors, and fish proteins have diverged faster

903 than mammalian homologs (159). For example, 25% of human genes cannot be identified in the
904 pufferfish (16). Even in the well-characterized zebrafish, targeted approaches have the potential to
905 reveal additional novel genes (160). We therefore encourage researchers to consider genome-wide
906 blast searches in addition to a consultation of round goby gene models, and hope that extensive RNA
907 sequencing data can be generated in the future to improve the predictions.
908
909 **Sex determining regions**
910
911 To investigate whether the round goby genome features large sex determining regions, we analyzed
912 own available RAD sequencing data. We prepared restriction site–associated DNA (RAD) (161)
913 libraries following the protocol used by (162, 163), which is largely based on (166). In short, we used
914 the Sbf1 enzyme on DNA extracted from 57 females, 56 males, and 5 juveniles caught in Basel,
915 Switzerland, and pooled 39-40 individuals per library for SR 100bp sequencing with Illumina. 45
916 females and 47 males retained sufficient numbers of reads (>150000) per sample after cleaning and
917 demultiplexing, were processed with the Stacks pipeline using the genome independent approach
918 (164), and were analyzed for sex-specific loci present exclusively in males or females. Considering a
919 genome size of ~1GB, the presence of 23 chromosomes (165), and a calling success of 21877 loci in
920 95 or 96 individuals (49220 loci in at least 40 individuals), we expected an average density of one
921 RAD locus every 45710 (20316) bp and an average number of 951 (2140) markers for an average
922 sized chromosome. The presence of a sex chromosome should thus be indicated by hundreds of sex-
923 specific RAD loci, while a contiguous sex determining region larger than 45000 bp would be indicated
924 by one or more sex specific RAD loci. Read numbers per locus for each sample were extracted from
925 the *.matches.tsv file output from Stacks and analyzed for sex-specific loci with standard R table
926 manipulations.
927
928 **Vision**
929
930 Opsin genes were extracted from the genome assembly using the Geneious software
931 (http://www.geneious.com) (167) by mapping the genomic scaffolds (Medium Sensitivity, 70% identity
932 threshold) against individual opsin exons of Nile tilapia (*Oreochromis niloticus*; GenBank Acc. no.:
933 MKQE00000000.1). This led to capturing of all scaffolds containing any visual opsin. The genes were

then annotated by mapping back of the single exons of tilapia against each scaffold separately (High Sensitivity; 50% identity threshold) combined with the Live Annotate & Predict function as implemented in Geneious, based on the Nile tilapia and mudskipper (25) opsin gene annotation. All regions upstream and downstream from every opsin gene, as well as the intergenic regions were separately tested for presence of any further opsin gene or its fragment (pseudogene). The annotated genes were checked for the reading frame and the putative protein product was predicted.

We next performed phylogenetic analysis on the visual opsin genes (i.e. SWS1, SWS2, RH2, RH1 and LWS opsins) across vertebrates, with focus on selected model species of teleost fishes. We further specifically focused on the LWS genes from the fish species or lineages known to possess multiple LWS copies, such as livebearers and pupfishes (Cyprinodontiformes) (168), zebrafish (*Danio rerio*) (169), salmon (*Salmo salar*) (170), common carp (*Cyprinus carpio*) (170), cavefish (*Astyanax mexicanus*) (171), Northern pike (*Esox lucius*) (170), labyrinth fishes (*Anabas testudineus*) (20), Asian arowana (*Scleropages* formosus) (170) as well as other gobies, such as mudskippers (25) and reef gobies (20). The opsin gene sequences from round goby and other fish species, including outgroup of non-visual opsins (pinopsin, parietopsin, vertebrate-ancestral opsins and opn3 opsin; **Supplemental_Material_S2**) were aligned using the MAFFT (172) plugin (v1.3.5) under the L-ins-i algorithm as implemented in Geneious. Exon 5 (exon 6 in case of LWS) and part of exon 1 (or entire exon 1 in case of LWS), which provided ambiguous alignment due to their higher variability, were discarded. We estimated the model parameters by jModeltest 2.1.6 (173, 174), and subsequently used the bayesian inference to calculate single-gene phylogeny using the MrBayes 3.2.6 (175) software as implemented on the CIPRES Science gateway (176).

Olfaction

Olfactory receptor (OR) peptide sequences to be used as a reference were extracted from a publicly available *Oreochromis niloticus* protein dataset (177). Those references were blasted (tblastn) against the genomes of the round goby (*Neogobius melanostomus*), the blue-spotted mudskipper (*Boleophthalmus pectinirostris*) (25), the giant mudskipper (*Periophthalmodon magnuspinatus*) (25) and the three-spined stickleback (*Gasterosteus aculeatus*) (178), using an e-value threshold of $10e^{-50}$. Only the hit with highest bit-score for each genomic position with more than one alignment was

965 employed in subsequent steps. Mapped hits belonging to contiguous positions of the protein
966 (maximum overlap of 15 aminoacids) and with a genomic distance smaller than 10kb were joined as
967 exons of the same CDS-gene model. Obtained sequences were translated to proteins using
968 TransDecoder (http://transdecoder.github.io), filtering all models that produce peptides smaller than
969 250 aminoacids. While many ORs are usually around 300 aminoacids long in total, 250 is close to the
970 average size of their main transmembrane domain, which is centrally located in the protein and more
971 suitable to interspecific alignment compared to N-terminal and C-terminal ends. We acknowledge that
972 this method might introduce a reduced proportion of recent pseudogenes that could lead to a small
973 overestimation of OR genes with coding capacity, although all species should be affected equivalently.
974
975 Next, an hmmscan (http://hmmer.org/) was produced against Pfam database to identify the domain
976 with highest score for each obtained protein sequence. We also filtered against false positive detection
977 using blast against confident OR and non-OR protein datasets. For phylogenetic analysis, sequences
978 (**Supplemental_Material_S3**) were aligned with MAFFT (https://mafft.cbrc.jp/alignment/server/) and a
979 Maximum Likelihood methodology was employed to build the tree using W-IQ-TREE software (179)
980 with standard parameters and Ultrafast bootstrap (180). Four adrenergic receptor sequences from
981 *Oreochromis niloticus* were used as an outgroup. Monophyletic groups formed by five or more genes
982 of the same species were considered as lineage-specific gene expansions. Because of the
983 phylogenetic proximity of the two mudskippers and the differences in their genome assembly statistics,
984 only *B. pectinirostris* was considered and *P. magnuspinatus* sequences were allowed to be included in
985 their lineage-specific expansion groups.
986
987 **Detoxification**
988
989 The Basic Local Alignment Search Tool (BLAST, v. 2.2.31) (181) was used to identify local alignments
990 between the round goby genome and a query including all annotated CYPs in humans and zebrafish
991 (vertebrate) and the most dissimilar invertebrate CYPs from *Drosophila melanogaster* (arthropod),
992 *Caenorhabditis elegans* (nematode) and *Capitella teleta* (annelid; **Supplemental_Material_S4**). Only
993 BLAST high scoring pairs with Expect values of 1.0 x 10-10 or smaller were considered significant.
994

The JBrowse genome viewer (v1.12.1) was used to manually annotate the significant regions of each genome from the BLAST search, identifying start (ATG) and stop (TGA/TAA/TAG) codons, exon number, and splice site signals (GT/AG) at intron-exon boundaries. The lengths of the potential CYPs were identified and considered full length at ~500 amino acid residues long. Potential genes were matched to the well-curated cytochrome P450 HMM in the Pfam protein family database (182) to confirm identity. The ScanProsite tool (183) was used to verify the presence of four largely conserved CYP motifs: the I-helix, K-helix, meander coil and heme loop. Each gene was classified as 'complete' (proper length with start and stop codon, all motifs present, and match to the HMM) or 'partial' (presence of at least the entire ~120 amino acid region that contains all motifs but clearly less than full length). Any potential CYP that was missing at least one of the motifs was considered a gene 'fragment' (**Supplemental_Table_S9**).

All of the 'complete' and 'partial' round goby CYPs (**Supplemental_Table_S9**) were included in further analyses. Clustal Omega (v1.2.4) (184) was used to generate a multiple sequence alignment of the round goby sequences and a variety of well-known vertebrate CYPs from humans, *Danio rerio*, *Mus musculus*, *Xenopus laevis*, *Gallus gallus*, and *Rattus norvegicus* (125 sequences in total; **Supplemental_Material_S5**). Mesquite (v3.10) (185) was utilized to trim the alignment, especially at the termini of the protein sequences where significant variation is typically observed, leaving only the portion of the alignment representative of the homology of the sequences. The final 'masked' alignment (**Supplemental_Material_S6**) was used as input for the Randomized Axelerated Maximum Likelihood program (RAxML v8.2.10) (186). 100 bootstrap trees were generated with the rapid generation algorithm (-x) and a gamma distribution. The JTT substitution matrix with empirical frequencies was implemented in tree generation. The final maximum likelihood phylogenetic tree was visualized with Figtree (v1.4.3) (187) and rooted with the CYP51 family of enzymes.

Osmoregulation

Protein sequences for aquaporins, tight junction proteins, ion transporters, and enzymes in osmolyte production pathways were retrieved from the round goby genome by BLASTing well-characterized proteins from zebrafish, downloaded from Uniprot (March 2018), against the round goby gene models/proteins. Only round goby gene-models/proteins for which the predicted protein covered at

least 70%, with a sequence identity of at least 40% and with E-value < 10^{-20} of the corresponding protein in zebrafish were used for the phylogenetic analyses. Well-established paralogues belonging to different subclasses of the respective protein family, based on either literature search or from initial phylogenetic analysis of that particular protein family, were used as additional query sequences to minimize the risk of missing relevant round goby sequences. Osmoregulatory genes from human and zebrafish were used for overall classification of clades in the respective protein family. Some modifications were made to the retrieved round goby sequences before analysis: i) For NHE ion transporters, a 780 aa long non-homologous N-terminus from one of the *Neogobius* sequences was removed before the phylogenetic analysis. ii) Some of the claudin genes were subjected to manual curation of Maker predicted proteins. The claudin genes in fish consist of several tandem arrays, which in some cases results in merging of 2-4 claudin genes by the Maker software. Claudins have a typical trans-membrane (TM) pattern with four distinct TM domains. All manually curated claudin genes from round goby were examined to have the expected four TM domains by TMHMM searches. Round goby protein sequences after manual curation are available in the supplement (**Supplemental_Material_S7**).

No myo-inositol phosphate synthase (MIPS) and sodium/inositol cotransporter (SMIT) proteins from zebrafish was found in Uniprot. To confirm that there are truly no MIPS and SMIT genes in zebrafish, the zebrafish genome at NCBI was also searched for homologies using blastp and tblastn using as query the MIPS and SMIT protein sequences from tilapia as query, and no hits were found. Thus, in the case of MIPS and SMIT, tilapia sequences were used for searching for round goby homologues. For the phylogenetic analyses, protein sequences from zebrafish (*Danio rerio*), three spine stickleback (*Gasterosteus aculeatus*), tilapia (*Oreochromis niloticus*), mudskipper (*Boleophthalmus pectinirostris*) and *Homo sapiens* (exception for human NKA-beta) were used in comparison to round goby, and were obtained from Uniprot (zebrafish, stickleback, tilapia, human) or RefSeq (mudskipper; **Supplemental_Material_S7**). Phylogenetic analyses of osmoregulatory proteins in round goby were performed using maximum likelihood with PhyML v3.0 with 100 bootstraps and using Gblocks to eliminate poorly aligned positions and highly divergent regions. PhyML analyses were performed at the Phylogeny.fr website (http://www.phylogeny.fr) using default settings.

Immune system

To perform an overall characterization of key genes related to the immune system, protein queries representing core components of innate and acquired immunity from several fish species as well as mammalian reference sequences were downloaded from UniProt and Ensembl. The protein queries were aligned prior to usage to ensure sequence homology. We also added previously extracted protein sequences from the Toll-like receptor family, reported by (85), and MHCI sequences reported by (80). All queries are listed in **Supplemental_Table_S4**. To enable comparative analyses between sequenced Gobiiformes, the genomes of *Periophthalmodon schlosseri* (GCA_000787095.1), *Periophthalmus magnuspinatus* (GCA_000787105.1), *Scartelaos histophorus* (GCA_000787155.1) and *Boleophthalmus pectinirostris* (GCA_000788275.1) were additionally downloaded from NCBI.

All protein queries were used in a tblastn (blast+ v. 2.6.0) towards the round goby genome assembly using default parameters and a e-value cutoff of 1e-10 (188). Some queries (*caspase-1*, *TLRs*, *IL1* and *IL8*) were also used in an identical tblastn towards the other Gobiiformes genomes. Genomic hit regions were extracted using BEDtools (v. 2.17.0) extending both up- and downstream as needed to obtain full length gene sequences (189). The extracted genomic regions were imported into MEGA7, the reading frame was adjusted for each exon and aligned as proteins to the corresponding translated coding sequence of queries using MUSCLE with default parameters. Intronic sequences were removed leaving an in-frame coding sequence (190, 191). All alignments were subjected to manual evaluation before subsequent analysis.

To generate phylogenetic trees, protein alignments were made and model tested using the ProtTest3 server (http://darwin.uvigo.es/software/prottest_server.html) specifying BIC and no tree optimization (server has been disabled but ProtTest is available for download from GitHub) (192). All alignments reported the JTT model as best hit. Maximum likelihood trees were produced by using RAxML-PTHREADS (v 8.0.26), PROTCATJTT, rapid bootstrap and 500 bootstrap replicates (193). The final trees were imported into FigTree (http://tree.bio.ed.ac.uk/software/figtree/), and subsequently Adobe Illustrator, for presentation purposes.

In order to identify members of the large multigenic family of fish-specific NACHT and Leucine-Rich Repeats containing genes (NLRs; the fish-specific subset is also known as NLR-C) (87), an alignment

of 368 zebrafish NLR-C proteins was obtained from (88). A combination of tblastn, HMMER3 searches (194) and alignments with MAFFT v7.310 (195) was used to generate first an initial list of "candidate regions" potentially containing an NLR (**Supplemental_Material_S8**) and then an annotation of the characteristic domains in round goby NLR-C family members (see **Supplemental_Material_S9)** and **Supplemental_Table_S8** for details), consisting of 25 PYRIN , 1 N-terminal CARD, 12 C-terminal CARD, 343 FISNA-NACHT and 178 B30.2 domains. Custom HMM models for major NLR exons (FISNA-NACHT, and PRY-SPRY/B30.2) were generated and utilized during this process (Supplementary Methods, **Supplemental_Material_S10**). The majority of identified FISNA-NACHT exons contained frameshifts or a large insertion, indicating either pseudogenization, acquisition of new introns, problems with the assembly, or a combination of the three (196). In any case, for the subsequent phylogenetic analysis, only the 61 clearly intact NLRs were used. These were aligned with NLRs from human, zebrafish and the mudskipper goby using MAFFT (**Supplemental_Material_S9**; **Supplemental_Material_S11**); Maximum Likelihood trees were produced with RAxML-PTHREADS, PROTCATJTT, rapid bootstrap and 500 bootstrap replicates (193). The final trees were imported into FigTree (http://tree.bio.ed.ac.uk/software/figtree/), and subsequently Adobe Illustrator. The alignments were inspected manually for presence of the conserved Walker A motifs and sequence logos for these were generated with WebLogo (197). Finally, we performed a survey of the PYD domains, Peptidase_C14 domains (Caspases) and CARD. All cases of a PYD domain followed by an adjacent CARD in the round goby (putative apoptosis-associated speck-like protein containing a CARD (ASC), also known as PYD-CARD or PYCARD) were identified from the HMMER3 dataset. The open reading frames containing these were translated, concatenated, and aligned with similarly structured proteins from human, mouse, lizard, frog and all the fish in Ensembl, and with PYD-CARDs identified from the other available goby assemblies (**Supplemental_Material_S12**). A phylogenetic tree was generated as described above. The annotation for NLR-C genes consists of predicted positions for all of the major conserved NLR-associated domains (PYD, CARD, FISNA-NACHT-helices, LRRs, B30.2; **Supplemental_Table_S8**).

Epigenetic regulators

We focused on two gene expression regulators which are conserved among all eukaryotes: the Polycomb Repressive Complex 2 (PRC2), which deposits repressive histone methylation marks, and

the DNA methylases, which methylate cytosine in CpG contexts. The presence of both marks is commonly associated with a downregulation of gene expression. The protein sequences of zebrafish orthologues of PRC2 components RBBP2, EED, EZH1-2, and SUZ12 (50) and of DNA methylases DNMT1 and DNMT3 (198) were blasted against the round goby genome using default parameters of the Albiorix Blast server. The protein sequence of predicted proteins at the hit site was extracted manually in the round goby genome browser and aligned with mouse, human, and zebrafish protein sequences. When the first and/or last exon sequences as predicted in the round goby genome differed significantly from the mouse, human, and zebrafish sequences, we attempted confirmation by 3' and 5'RACE on RNA extracted from whole juvenile animals (see **Supplemental_Material_S13** for primer sequences and PCR conditions). A putative CDS was combined from automated annotation and RACE results, and aligned to sequences extracted from a variety of fish taxa, shark, chicken, frog, lizard, and human (**Supplemental_Material_S14**). Given the high conservation of these proteins in eukaryotes, and the absence of major unexpected differences between round goby and other vertebrates, additional Gobiidae were not included in the analyses. In order to perform codon aware alignment MACSE (199) was used. The model and partitioning scheme used for each phylogenetic analysis was estimated using PartitionFinder2 (200) using PhyML (201) with corrected AIC scores (AICc) used for model selection. Phylogenetic analyses were performed using MrBayes 3.2.6 (175, 202) with three independent runs for each gene. Analyses were run for 2,000,000 generations or until the standard deviation of split frequencies was below 0.01 up to a maximum of 20,000,000 generations. In order to aid convergence in the EZH analyses the temperature parameter was set to 0.05.

Transposable elements

A number of different applications were used for the repeat annotation of the genome. They are described in the repeat annotation report (**Supplemental_Material_S15**). In summary, in addition to the identification of repeats with RepeatModeler (as described above), we used TRF (203) to predict tandem repeats. RepeatMasker (153), a homology-based approach was used to produce a genome-wide overview of interspersed repeats. LTR Finder (204) and LTRharvest (205) in combination with LTRdigest (206), both de novo approaches, were used to predict LTRs.

Availability of data and materials

The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VHKM00000000. The version described in this paper is version VHKM01000000.

Various Illumina reads are available under the accessions indicated in Table 1.

Other datasets supporting the conclusions of this article are included within the article and its additional files.

Additional files

Supplemental Figure S1: .pdf; full opsin phylogenetic tree

Supplemental Figure S2: .pdf; opsin tree constructed with individual exons

Supplemental Figure S3: .pdf; full olfactory receptor phylogenetic tree

Supplemental Figure S4: .pdf; claudin phylogenetic tree

Supplemental Figure S5: .pdf; occludin phylogenetic tree

Supplemental Figure S6: .pdf; sodium transporter phylogenetic trees

Supplemental Figure S7: .pdf; myo-inositol production and accumulation gene phylogenetic trees

Supplemental Figure S8: .pdf; TAP gene phylogenetic tree

Supplemental Figure S9: .pdf; Ig locus schematic

Supplemental Figure S10: .pdf; TLR phylogenetic tree of Gobiidae

Supplemental Figure S11: .pdf; CRP / APCS phylogenetic tree

Supplemental Figure S12: .pdf; SUZ12, EED, and RBBP4 phylogenetic trees

Supplemental Material S1: .gz; annotation tracks for genome

Supplemental Material S2: .txt; opsin sequences used for tree building

Supplemental Material S3: .txt; olfactory receptor sequences used for tree building

Supplemental Material S4: .txt; CYP sequences used as query

Supplemental Material S5: .txt; CYP sequences used for tree building

Supplemental Material S6: .phy; CYP alignment

Supplemental Material S7: .doc; sequences for osmoregulatory proteins

Supplemental Material S8: .fas; NLR candidate regions

Supplemental Material S9: .doc; detailed methods for NLR annotation

Supplemental Material S10: zip of .hmm; hmm models used to identify NLRs

Supplemental Material S11: .fas; NLR sequences used for vertebrate tree building

Supplemental Material S12: .fas; NLR sequences used for Gobiidae tree building

Supplemental Material S13: .doc; detailed methods for RACE

Supplemental Material S14: zip of .nex; dnmt1, dnmt3, eed, ezh, rbbp4, and suz12 alignments

Supplemental Material S15: .txt; detailed methods for repeat annotation

Supplemental Table S1: .xls; overview and results for repeat sequences analysed

Supplemental Table S2: .xls; results for CYP genes (location, sequence, length)

Supplemental Table S3: .doc; overview table of immune genes analysed

Supplemental Table S4: .xls; immune gene sequences used as query

Supplemental Table S5: .doc; results for MHCI

Supplemental Table S6: .doc; results for MHCII

Supplemental Table S7: .txt; results for other immune genes analysed

Supplemental Table S8: .xls; results for NLRs

Supplemental Table S9: .xls; results for CYPs

Funding

ZM was funded by the Czech Science Foundation (16-09784Y) and the Swiss National Science Foundation (PROMYS - 166550). DB was funded by CZ.02.2.69/0.0/0.0/16_027/0008495 - International Mobility of Researchers at Charles University. Genome sequencing was funded with a contribution of the Freiwillige Akademische Gesellschaft Basel to IAK. KP was funded by an Undergraduate Student Research Award and a Discovery grant RGPIN5767-16 (to JYW) from the Natural Sciences and Engineering Research Council of Canada. MT, TL and MAR were funded by the Center for Marine Evolutionary Biology. JS was funded by grants LE 546/9-1 and WI 3081/5-1 within the Deutsche Forschungsgemeinschaft (DFG) - funded Priority Programme SPP1819. MHS was funded by the Norwegian Research Council (grant numbers 199806/S40 and 222378/F20). AB was funded by the Swedish Research Council (VR; #2017-04559).

Acknowledgements

We are grateful to Prof. Patricia Burkhardt-Holm for her continuous support and encouragement. We thank Bernd Egger, Astrid Böhne, Philipp Hirsch and Patricia Burkhardt-Holm for critically reading the manuscript. We thank Fabio Cortesi for his insightful comments and the Center for Marine Evolutionary Biology for hosting a Blast server and a genome browser. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures". We thank Maria Leptin for her helpful comments and advice during annotation of the inflammasome components.

Author contributions

Sylke Winkler isolated DNA and generated PacBio reads, Martin Pippel and Siegfried Schloissnig assembled the genome sequence, and Tomas Larsson, Mats Tölpel and Magnus Alm Rosenblad performed automated annotation and provided the genome browser and Blast server.

Jean-Claude Walser provided transposable element analyses, Silvia Gutnik, Claire Peart, and Irene Adrian-Kalchhauser provided DNA methyltransferase and PRC2 analyses, Anders Blomberg provided osmoregulation analyses, Monica Hongroe Solbakken and Jaanus Suurväli provided immune gene analyses, Zuzana Musilova and Demian Burguera provided vision and olfaction analyses, Joanna Yvonne Wilson and Kirill Pankov provided CYP gene analyses, Nico Michiels investigated red fluorescence.

Irene Adrian-Kalchhauser initiated, designed, and supervised the project, acquired the necessary funding, coordinated annotation efforts, compiled the manuscript and handled the submission process.

Permissions

Fish used in this work were caught in accordance with permission 2-3-6-4-1 from the Cantonal Office for Environment and Energy, Basel Stadt.

References

1. Bock DG *et al. (*2014) What we still don't know about invasion genetics. *Molecular Ecology*.
2. Prentis PJ, Wilson JRU, Dormontt EE, RICHARDSON DM, Lowe AJ (2008) Adaptive evolution in invasive species. *Trends in Plant Science* 13:288–294.
3. Tsutsui ND, Suarez AV, Holway DA, Case TJ (2000) Reduced genetic variation and the success of an invasive species. *Proceedings of the National Academy of Sciences* 97:5948.
4. Lee CE (2002) Evolutionary genetics of invasive species. *TRENDS IN ECOLOGY & EVOLUTION* 17:386–391.
5. Jude DJ, Reider RH, Smith GR (1992) Establishment of Gobiidae in the Great Lakes Basin. *Can. J. Fish. Aquat. Sci.* 49:416–421.
6. Michalek M, Puntila R, Strake S, Werner M (2012) *HELCOM Baltic Sea Environment Fact Sheet 2012*.
7. Roche KF, Janac M, Jurajda P (2013) A review of Gobiid expansion along the Danube-Rhine corridor - geopolitical change as a driver for invasion. *KNOWLEDGE AND MANAGEMENT OF AQUATIC ECOSYSTEMS*.
8. Hirsch PE, N'Guyen A, Adrian-Kalchhauser I, Burkhardt-Holm P (2015) What do we really know about the impacts of one of the 100 worst invaders in Europe? A reality check. *Ambio*.
9. Dufour BA, Hogan TM, Heath DD (2007) Ten polymorphic microsatellite markers in the invasive round goby (Neogobius melanostomus) and cross-species amplification. *MOLECULAR ECOLOGY NOTES* 7:1205–1207.
10. Adrian-Kalchhauser I *et al. (*2017) The mitochondrial genome sequences of the round goby and the sand goby reveal patterns of recent evolution in gobiid fish. *BMC GENOMICS* 18:177.
11. Feldheim KA *et al. (*2009) Microsatellite loci for Ponto-Caspian gobies: markers for assessing exotic invasions. *Molecular Ecology Resources* 9:639–644.
12. Neilson ME, Stepien CA (2009) Escape from the Ponto-Caspian: Evolution and biogeography of an endemic goby species flock (Benthophilinae: Gobiidae: Teleostei). *Molecular Phylogenetics and Evolution* 52:84–102.

13. Bowley LA, Alam F, Marentette JR, Balshine S, Wilson JY (2010) Characterization of vitellogenin gene expression in round goby (Neogobius melanostomus) using a quantitative polymerase chain reaction assay. *Environmental toxicology and chemistry / SETAC* 29:2751–2760.

14. Thacker CE, Roje DM (2011) Phylogeny of Gobiidae and identification of gobiid lineages. *Systematics and Biodiversity* 9:329–347.

15. Thacker CE, Thompson AR, Roje DM (2011) Phylogeny and evolution of Indo-Pacific shrimp-associated gobies (Gobiiformes: Gobiidae). *Molecular Phylogenetics and Evolution* 59:168–176.

16. Aparicio S *et al. (*2002) Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. *Science* 297:1301.

17. Amemiya CT *et al. (*2013) The African coelacanth genome provides insights into tetrapod evolution. *NATURE* 496:311 EP -.

18. Star B *et al. (*2011) The genome sequence of Atlantic cod reveals a unique immune system. *NATURE* 477:207–210.

19. Li J-T *et al. (*2015) The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (Cyprinus carpio). *Scientific reports* 5:8199 EP -.

20. Musilova Z *et al. (*2019) Vision using multiple distinct rod opsins in deep-sea fishes. *Science* 364:588.

21. Seehausen O, van Alphen JJM, Witte F (1997) Cichlid Fish Diversity Threatened by Eutrophication That Curbs Sexual Selection. *Science* 277:1808.

22. Seehausen O *et al. (*2008) Speciation through sensory drive in cichlid fish. *NATURE* 455:620-U23.

23. Barth FG; Schmid A; Douglas RH, eds (2001) *The Ecology of Teleost Fish Visual Pigments: a Good Example of Sensory Adaptation to the Environment? Ecology of Sensing* (Springer Berlin Heidelberg).

24. Hornsby MAW, Sabbah S, Robertson RM, Hawryshyn CW (2013) Modulation of environmental light alters reception and production of visual signals in Nile tilapia. *JOURNAL OF EXPERIMENTAL BIOLOGY* 216:3110–3122.

25. You X *et al. (*2014) Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *NATURE COMMUNICATIONS* 5.

1296 26. Busserolles F de *et al. (*2017) Pushing the limits of photoreception in twilight conditions. The rod-
1297 like cone retina of the deep-sea pearlsides. *SCIENCE ADVANCES* 3.

1298 27. Kenaley CP, Devaney SC, Fjeran TT (2014) The complex evolutionary history of seeing red.
1299 Molecular phylogeny and the evolution of an adaptive visual system in deep-sea dragonfishes
1300 (Stomiiformes: Stomiidae). *Evolution; international journal of organic evolution* 68:996–1013.

1301 28. Corkum LD *et al. (*2006) Evidence of a Male Sex Pheromone in the Round Goby (Neogobius
1302 melanostomus). *Biol Invasions* 8:105–112.

1303 29. Farwell M *et al. (*2017) Differential female preference for individual components of a reproductive
1304 male round goby (Neogobius melanostomus) pheromone. *JOURNAL OF GREAT LAKES*
1305 *RESEARCH* 43:379–386.

1306 30. Tierney KB *et al. (*2012) Invasive male round gobies (Neogobius melanostomus) release
1307 pheromones in their urine to attract females. *Can. J. Fish. Aquat. Sci.* 70:393–400.

1308 31. Laframboise AJ, Katare Y, Scott AP, Zielinski BS (2011) The Effect of Elevated Steroids Released
1309 by Reproductive Male Round Gobies, Neogobius melanostomus, on Olfactory Responses in
1310 Females. *JOURNAL OF CHEMICAL ECOLOGY* 37:260–262.

1311 32. Marentette JR *et al. (*2010) Signatures of contamination in invasive round gobies (Neogobius
1312 melanostomus): A double strike for ecosystem health? *ECOTOXICOLOGY AND*
1313 *ENVIRONMENTAL SAFETY* 73:1755–1764.

1314 33. Marentette JR, Balshine S (2012) Altered Prey Responses in Round Goby from Contaminated
1315 Sites. *ETHOLOGY* 118:812–820.

1316 34. McCallum ES *et al. (*2014) Persistence of an invasive fish (Neogobius melanostomus) in a
1317 contaminated ecosystem. *BIOLOGICAL INVASIONS* 16:2449–2461.

1318 35. Goldstone JV *et al. (*2006) The chemical defensome. Environmental sensing and response genes
1319 in the Strongylocentrotus purpuratus genome. *DEVELOPMENTAL BIOLOGY* 300:366–384.

1320 36. Whitfield AK (2015) Why are there so few freshwater fish species in most estuaries? *J Fish Biol*
1321 86:1227–1250.

1322 37. Lee KA, Klasing KC (2004) A role for immunology in invasion biology. *TRENDS IN ECOLOGY &*
1323 *EVOLUTION* 19:523–529.

1324 38. David GM *et al. (*2018) A minimalist macroparasite diversity in the round goby of the Upper Rhine
1325 reduced to an exotic acanthocephalan lineage. *Parasitology* 145:1020–1026.

1326 39. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression. How the genome integrates
1327 intrinsic and environmental signals. *Nature genetics* 33:245 EP -.

1328 40. Zamudio N *et al. (*2015) DNA methylation restrains transposons from adopting a chromatin
1329 signature permissive for meiotic recombination. *Genes & development* 29:1256–1270.

1330 41. Choi J, Lyons DB, Kim Y, Moore JD, Zilberman D (2019) *DNA methylation and histone H1*
1331 *cooperatively repress transposable elements and aberrant intragenic transcripts. Supplemental*
1332 *Information*.

1333 42. Feinberg AP, Irizarry RA (2010) Stochastic epigenetic variation as a driving force of development,
1334 evolutionary adaptation, and disease. *PROCEEDINGS OF THE NATIONAL ACADEMY OF*
1335 *SCIENCES OF THE UNITED STATES OF AMERICA* 107:1757–1764.

1336 43. Herman JJ, Sultan SE (2016) DNA methylation mediates genetic variation for adaptive
1337 transgenerational plasticity. *Proceedings of the Royal Society B: Biological Sciences*
1338 283:20160988.

1339 44. Cortijo S *et al. (*2014) Mapping the Epigenetic Basis of Complex Traits. *Science* 343:1145.

1340 45. Wellband KW, Heath DD (2017) Plasticity in gene transcription explains the differential
1341 performance of two invasive fish species. *Evol Appl* 10:563–576.

1342 46. Adrian-Kalchhauser I, Walser J-C, Schwaiger M, Burkhardt-Holm P (2018) RNA sequencing of
1343 early round goby embryos reveals that maternal experiences can shape the maternal RNA
1344 contribution in a wild vertebrate. *BMC EVOLUTIONARY BIOLOGY* 18:34.

1345 47. Somerville V *et al. (*2019) *DNA Methylation Patterns in the Round Goby Hypothalamus Support an*
1346 *On-The-Spot Decision Scenario for Territorial Behavior*.

1347 48. Grimm SA *et al. (*2019) DNA methylation in mice is influenced by genetics as well as sex and life
1348 experience. *NATURE COMMUNICATIONS* 10:305.

1349 49. Weyrich A *et al. (*2016) Paternal heat exposure causes DNA methylation and gene expression
1350 changes of in Wild guinea pig sons. *Ecology and evolution*.

1351 50. Margueron R, Reinberg D (2011) The Polycomb complex PRC2 and its mark in life. *NATURE*
1352 469:343–349.

1353 51. Gibbs DJ *et al. (*2018) Oxygen-dependent proteolysis regulates the stability of angiosperm
1354 polycomb repressive complex 2 subunit VERNALIZATION 2. *NATURE COMMUNICATIONS*
1355 9:5438.

1356 52. Martinez P *et al. (*2014) Genetic architecture of sex determination in fish. Applications to sex ratio
1357 control in aquaculture. *Frontiers in genetics* 5.

1358 53. Hardie DC, Hebert PDN (2003) The nucleotypic effects of cellular DNA content in cartilaginous
1359 and ray-finned fishes. *GENOME* 46:683–706.

1360 54. Hardie DC, Hebert PDN (2004) Genome-size evolution in fishes. *Can. J. Fish. Aquat. Sci.*
1361 61:1636–1646.

1362 55. Gregory TR (2019) Animal Genome Size Database.

1363 56. Bowmaker JK, Hunt DM (2006) Evolution of vertebrate visual pigments. *Current Biology* 16:R484-
1364 R489.

1365 57. Cortesi F *et al. (*2015) Ancestral duplications and highly dynamic opsin gene evolution in
1366 percomorph fishes. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE*
1367 *UNITED STATES OF AMERICA* 112:1493–1498.

1368 58. Liu D-W *et al. (*2019) The Cone Opsin Repertoire of Osteoglossomorph Fishes. Gene Loss in
1369 Mormyrid Electric Fish and a Long Wavelength-Sensitive Cone Opsin That Survived 3R.
1370 *MOLECULAR BIOLOGY AND EVOLUTION* 36:447–457.

1371 59. Yokoyama S (2008) in *Linkage disequilibrium and association mapping*, ed Weir BS, pp 259–282.

1372 60. Niimura Y (2012) Olfactory receptor multigene family in vertebrates. From the viewpoint of
1373 evolutionary genomics. *Current genomics* 13:103–114.

1374 61. Niimura Y (2009) On the Origin and Evolution of Vertebrate Olfactory Receptor Genes.
1375 Comparative Genome Analysis Among 23 Chordate Species. *GENOME BIOLOGY AND*
1376 *EVOLUTION* 1:34–44.

1377 62. Nelson DR (2003) Comparison of P450s from human and fugu. 420 million years of vertebrate
1378 P450 evolution. *Archives of Biochemistry and Biophysics* 409:18–24.

1379 63. Goldstone JV *et al. (*2010) Identification and developmental expression of the full complement of
1380 Cytochrome P450 genes in Zebrafish. *BMC GENOMICS* 11:643.

1381 64. Kirischian N, McArthur AG, Jesuthasan C, Krattenmacher B, Wilson JY (2011) Phylogenetic and
1382 functional analysis of the vertebrate cytochrome p450 2 family. *JOURNAL OF MOLECULAR*
1383 *EVOLUTION* 72:56–71.

1384 65. Dejong CA, Wilson JY (2014) The Cytochrome P450 Superfamily Complement (CYPome) in the
1385 Annelid Capitella teleta. *PLoS ONE* 9:e107728.

1386 66. Luch A, Baird WM in *The carcinogenic effects of polycyclic aromatic hydrocarbons*, pp 19–96.

1387 67. Yan J, Cai Z (2010) Molecular evolution and functional divergence of the cytochrome P450 3
1388 (CYP3) Family in Actinopterygii (ray-finned fish). *PLoS ONE* 5:e14276.

1389 68. Yokoyama C *et al. (*1996) Human Gene Encoding Prostacyclin Synthase (PTGIS). Genomic
1390 Organization, Chromosomal Localization, and Promoter Activity. *Genomics* 36:296–304.

1391 69. Li Y-C *et al. (*2008) Structures of prostacyclin synthase and its complexes with substrate analog
1392 and inhibitor reveal a ligand-specific heme conformation change. *JOURNAL OF BIOLOGICAL*
1393 *CHEMISTRY* 283:2917–2926.

1394 70. Finn RN, Cerdà J (2011) Aquaporin evolution in fishes. *Frontiers in physiology* 2:44.

1395 71. Finn RN, Chauvigné F, Hlidberg JB, Cutler CP, Cerdà J (2014) The Lineage-Specific Evolution of
1396 Aquaporin Gene Clusters Facilitated Tetrapod Terrestrial Adaptation. *PLoS ONE* 9:e113686.

1397 72. Loh YH, Christoffels A, Brenner S, Hunziker W, Venkatesh B (2004) Extensive Expansion of the
1398 Claudin Gene Family in the Teleost Fish, Fugu rubripes. *GENOME RESEARCH* 14:1248–1257.

1399 73. Hwang P-P, Chou M-Y (2013) Zebrafish as an animal model to study ion homeostasis. *Pflügers*
1400 *Archiv - European Journal of Physiology* 465:1233–1247.

1401 74. Ronkin D, Seroussi E, Nitzan T, Doron-Faigenboim A, Cnaani A (2015) Intestinal transcriptome
1402 analysis revealed differential salinity adaptation between two tilapiine species. *Comparative*
1403 *biochemistry and physiology. Part D, Genomics & proteomics* 13:35–43.

1404 75. Rim JS *et al. (*1998) Transcription of the sodium/myo-inositol cotransporter gene is regulated by
1405 multiple tonicity-responsive enhancers spread over 50 kilobase pairs in the 5 '-flanking region.
1406 *JOURNAL OF BIOLOGICAL CHEMISTRY* 273:20615–20621.

1407 76. Wang X, Kültz D (2017) Osmolality/salinity-responsive enhancers (OSREs) control induction of
1408 osmoprotective genes in euryhaline fish. *PROCEEDINGS OF THE NATIONAL ACADEMY OF*
1409 *SCIENCES OF THE UNITED STATES OF AMERICA* 114:E2729-E2738.

1410 77. Sacchi R, Gardell AM, Chang N, Kültz D (2014) Osmotic regulation and tissue localization of the
1411 myo-inositol biosynthesis pathway in tilapia (Oreochromis mossambicus) larvae. *J. Exp. Zool.*
1412 321:457–466.

1413 78. Sacchi R, Li J, Villarreal F, Gardell AM, Kültz D (2013) Salinity-induced regulation of the
1414 myo-inositol biosynthesis pathway in tilapia gill epithelium. *J. Exp. Biol.*
1415 216:4626.

1416 79. Flajnik MF (2018) A cold-blooded view of adaptive immunity. *Nature reviews. Immunology* 18:438–
1417 453.

1418 80. Grimholt U *et al. (*2015) A comprehensive analysis of teleost MHC class I sequences. *BMC*
1419 *EVOLUTIONARY BIOLOGY* 15:32.

1420 81. McConnell SC *et al. (*2016) Alternative haplotypes of antigen processing genes in zebrafish
1421 diverged early in vertebrate evolution. *PROCEEDINGS OF THE NATIONAL ACADEMY OF*
1422 *SCIENCES OF THE UNITED STATES OF AMERICA* 113:E5014-E5023.

1423 82. Mashoof S, Criscitiello MF (2016) Fish Immunoglobulins. *Biology* 5:45.

1424 83. Riera Romo M, Perez-Martinez D, Castillo Ferrer C (2016) Innate immunity in vertebrates. An
1425 overview. *Immunology* 148:125–139.

1426 84. Guo H, Callaway JB, Ting JP-Y (2015) Inflammasomes. Mechanism of action, role in disease, and
1427 therapeutics. *NATURE MEDICINE* 21:677–687.

1428 85. Solbakken MH *et al. (*2016) Evolutionary redesign of the Atlantic cod (Gadus morhua L.) Toll-like
1429 receptor repertoire by gene losses and expansions. *Scientific reports* 6:25211 EP -.

1430 86. Solbakken MH, Voje KL, Jakobsen KS, Jentoft S (2017) Linking species habitat and past
1431 palaeoclimatic events to evolution of the teleost innate immune system. *Proceedings. Biological*
1432 *sciences* 284.

1433 87. Laing KJ, Purcell MK, Winton JR, Hansen JD (2008) A genomic view of the NOD-like receptor
1434 family in teleost fish. Identification of a novel NLR subfamily in zebrafish. *BMC EVOLUTIONARY*
1435 *BIOLOGY* 8.

1436 88. Howe K *et al. (*2016) Structure and evolutionary history of a large family of NLR proteins in the
1437 zebrafish. *Open biology* 6:160009.

1438 89. Tørresen OK *et al. (*2018) Genomic architecture of haddock (Melanogrammus aeglefinus) shows
1439 expansions of innate immune genes and short tandem repeats. *BMC GENOMICS* 19:240.

1440 90. Li J-Y *et al. (*2018) Characterization of an NLRP1 Inflammasome from Zebrafish Reveals a Unique
1441 Sequential Activation Mechanism Underlying Inflammatory Caspases in Ancient Vertebrates.
1442 *Journal of immunology (Baltimore, Md. : 1950)* 201:1946–1966.

1443 91. Kuri P *et al. (*2017) Dynamics of in vivo ASC speck formation. *The Journal of Cell Biology*
1444 216:2891–2909.

1445 92. Schwartz YB, Pirrotta V (2013) A new world of Polycombs. Unexpected partnerships and
1446 emerging functions. *Nature Reviews Genetics* 14:853 EP -.

1447 93. Mu W, Starmer J, Shibata Y, Della Yee, Magnuson T (2017) EZH1 in germ cells safeguards the
1448 function of PRC2 during spermatogenesis. *DEVELOPMENTAL BIOLOGY* 424:198–207.

1449 94. Xu J *et al. (*2015) Developmental control of polycomb subunit composition by GATA factors
1450 mediates a switch to non-canonical functions. *Molecular cell* 57:304–316.

1451 95. San B *et al. (*2016) Normal formation of a vertebrate body plan and loss of tissue maintenance in
1452 the absence of ezh2. *Scientific reports* 6:24658.

1453 96. Völkel P *et al. (*2019) Ezh1 arises from Ezh2 gene duplication but its function is not required for
1454 zebrafish development. *Scientific reports* 9:4319.

1455 97. Jeltsch A, Jurkowska RZ (2014) New concepts in DNA methylation. *Trends in Biochemical*
1456 *Sciences* 39:310–318.

1457 98. Ponger L, Li W-H (2005) Evolutionary Diversification of DNA Methyltransferases in Eukaryotic
1458 Genomes. *MOLECULAR BIOLOGY AND EVOLUTION* 22:1119–1128.

1459 99. Wang F-L *et al. (*2018) Genome-wide identification, evolution of DNA methyltransferases and their
1460 expression during gonadal development in Nile tilapia. *COMPARATIVE BIOCHEMISTRY AND*
1461 *PHYSIOLOGY B-BIOCHEMISTRY & MOLECULAR BIOLOGY* 226:73–84.

1462 100. Stapley J, Santure AW, Dennis SR (2015) Transposable elements as agents of rapid
1463 adaptation may explain the genetic paradox of invasive species. *Mol Ecol* 24:2241–2252.

1464 101. Pysek P *et al. (*2018) Small genome separates native and invasive populations in an
1465 ecologically important cosmopolitan grass. *Ecology* 99:79–90.

1466 102. Pezold FL (1984) Evidence for multiple sex-chromosomes in THE FRESH-WATER GOBY,
1467 GOBIONELLUS-SHUFELDTI (PISCES, GOBIIDAE). *COPEIA*:235–238.

1468 103. Bitton P-P, Christmann SAY, Santon M, Harant UK, Michiels NK (2018) Visual modelling
1469 validates prey detection by means of diurnal active photolocation in a small cryptobenthic fish.
1470 *bioRxiv*:338640.

1471 104. Carleton KL, Dalton BE, Escobar-Camacho D, Nandamuri SP (2016) Proximate and ultimate
1472 causes of variable visual sensitivities. Insights from cichlid fish radiations. *genesis* 54:299–325.

1473 105. Marshall J, Johnsen S (2017) Fluorescence as a means of colour signal enhancement.
1474 *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.

1475 106. Michiels NK *et al. (*2008) Red fluorescence in reef fish. A novel signalling mechanism? *BMC*
1476 *ECOLOGY* 8:14pp.

1477 107. Olsson KH *et al. (*2017) Dark eyes in female sand gobies indicate readiness to spawn. *PLoS*
1478 *ONE* 12:e0177714.

1479 108. Anthes N, Theobald J, Gerlach T, Meadows MG, Michiels NK (2016) Diversity and Ecological
1480 Correlates of Red Fluorescence in Marine Fishes. *FRONTIERS IN ECOLOGY AND EVOLUTION*
1481 4.

1482 109. Vélez-Espino LA, Koops MA, Balshine S (2010) Invasion dynamics of round goby (Neogobius
1483 melanostomus) in Hamilton Harbour, Lake Ontario. *BIOLOGICAL INVASIONS* 12:3861–3875.

1484 110. Young JAM *et al. (*2010) Demography and substrate affinity of the round goby (Neogobius
1485 melanostomus) in Hamilton Harbour. *JOURNAL OF GREAT LAKES RESEARCH* 36:115–122.

1486 111. Behrens JW, van Deurs M, Christensen EAF (2017) Evaluating dispersal potential of an
1487 invasive fish by the use of aerobic scope and osmoregulation capacity. *PLoS ONE* 12:e0176038.

1488 112. Hirsch PE *et al. (*2016) A tough egg to crack: recreational boats as vectors for invasive goby
1489 eggs and transdisciplinary management approaches. *Ecology and evolution* 6:707–715.

1490 113. Miladi H, Elabed H, Ben Slama R, Rhim A, Bakhrouf A (2017) Molecular analysis of the role of
1491 osmolyte transporters opuCA and betL in Listeria monocytogenes after cold and freezing stress.
1492 *Archives of microbiology* 199:259–265.

1493 114. Vigoder FM *et al. (*2016) Inducing Cold-Sensitivity in the Frigophilic Fly Drosophila montana by
1494 RNAi. *PLoS ONE* 11:e0165724.

1495 115. Wu C *et al. (*2014) The draft genome of the large yellow croaker reveals well-developed innate
1496 immunity. *NATURE COMMUNICATIONS* 5.

1497 116. Liu Z *et al. (*2016) The channel catfish genome sequence provides insights into the evolution
1498 of scale formation in teleosts. *NATURE COMMUNICATIONS* 7.

1499 117. Berthelot C *et al. (*2014) The rainbow trout genome provides novel insights into evolution after
1500 whole-genome duplication in vertebrates. *NATURE COMMUNICATIONS* 5:3657.

1501 118. Reis MIR, do Vale A, Pereira PJB, Azevedo JE, dos Santos NMS (2012) Caspase-1 and IL-1
1502 beta Processing in a Teleost Fish. *PLoS ONE* 7.

1503 119. Vojtech LN, Scharping N, Woodson JC, Hansen JD (2012) Roles of inflammatory caspases
1504 during processing of zebrafish interleukin-1β in Francisella noatunensis infection. *INFECTION*
1505 *AND IMMUNITY* 80:2878–2885.

1506 120. Richter K *et al. (*2018) C-Reactive Protein Stimulates Nicotinic Acetylcholine Receptors to
1507 Control ATP-Mediated Monocytic Inflammasome Activation. *Frontiers in immunology* 9:1604.

1508 121. Cao R, Zhang Y (2004) SUZ12 is required for both the histone methyltransferase activity and
1509 the silencing function of the EED-EZH2 complex. *Molecular cell* 15:57–67.

1510 122. Ciferri C *et al. (*2012) Molecular architecture of human polycomb repressive complex 2. *eLife*
1511 1:e00005.

1512 123. Chittock EC, Latwiel S, Miller TCR, Müller CW (2017) Molecular architecture of polycomb
1513 repressive complexes. *Biochemical Society transactions* 45:193–205.

124. Cao Q *et al. (*2014) The central role of EED in the orchestration of polycomb group complexes. *NATURE COMMUNICATIONS* 5:3127.

125. Liu X, Yang J, Wu N, Song R, Zhu H (2015) Evolution and Coevolution of PRC2 Genes in Vertebrates and Mammals. *Advances in protein chemistry and structural biology* 101:125–148.

126. Davidovich C, Cech TR (2015) The recruitment of chromatin modifiers by long noncoding RNAs. Lessons from PRC2. *RNA (New York, N.Y.)* 21:2007–2022.

127. Cross S, Kovarik P, Schmidtke J, Bird A (1991) Non-methylated islands in fish genomes are GC-poor. *NUCLEIC ACIDS RESEARCH* 19:1469–1474.

128. Jiang N *et al. (*2014) Conserved and Divergent Patterns of DNA Methylation in Higher Vertebrates. *GENOME BIOLOGY AND EVOLUTION* 6:2998–3014.

129. Han L, Zhao Z (2008) Comparative analysis of CpG islands in four fish genomes. *Comparative and functional genomics*:565631.

130. Huska M, Vingron M (2016) Improved Prediction of Non-methylated Islands in Vertebrates Highlights Different Characteristic Sequence Patterns. *PLOS COMPUTATIONAL BIOLOGY* 12.

131. McGaughey DM, Abaan HO, Miller RM, Kropp PA, Brody LC (2014) Genomics of CpG methylation in developing and developed zebrafish. *G3 (Bethesda, Md.)* 4:861–869.

132. Campos C, Valente LMP, Fernandes JMO (2012) Molecular evolution of zebrafish dnmt3 genes and thermal plasticity of their expression during embryonic development. *GENE* 500:93–100.

133. Takayama K, Shimoda N, Takanaga S, Hozumi S, Kikuchi Y (2014) Expression patterns of dnmt3aa, dnmt3ab, and dnmt4 during development and fin regeneration in zebrafish. *GENE EXPRESSION PATTERNS* 14:105–110.

134. Firmino J *et al. (*2017) Phylogeny, expression patterns and regulation of DNA Methyltransferases in early development of the flatfish, Solea senegalensis. *BMC DEVELOPMENTAL BIOLOGY* 17:11.

135. Wood RK, Crowley E, Martyniuk CJ (2016) Developmental profiles and expression of the DNA methyltransferase genes in the fathead minnow (Pimephales promelas) following exposure to di-2-ethylhexyl phthalate. *Fish Physiol Biochem* 42:7–18.

1542 136. Wu N *et al. (*2019) Fall webworm genomes yield insights into rapid adaptation of invasive
1543 species. *Nature ecology & evolution* 3:105–115.

1544 137. Lien S *et al. (*2016) The Atlantic salmon genome provides insights into rediploidization.
1545 *NATURE* 533:200 EP -.

1546 138. Kim B-M *et al. (*2019) Antarctic blackfin icefish genome reveals adaptations to extreme
1547 environments. *Nature ecology & evolution* 3:469–478.

1548 139. Mu Y *et al. (*2018) An improved genome assembly for Larimichthys crocea reveals hepcidin
1549 gene expansion with diversified regulation and function. *Communications biology* 1:195.

1550 140. El-Brolosy MA *et al. (*2019) Genetic compensation triggered by mutant mRNA degradation.
1551 *NATURE* 568:193–197.

1552 141. Roche K *et al. (*2015) A newly established round goby (Neogobius melanostomus)
1553 population in the upper stretch of the river Elbe. *Knowl. Manag. Aquat. Ecosyst.*:33.

1554 142. Patzner RA; VanTassel J.L.; Kovačić M; Kapoor BG, eds (2011) *The biology of gobies*
1555 (Science Publishers, Enfield, NH).

1556 143. Xing J, Zhou X, Tang X, Sheng X, Zhan W (2017) Characterization of Toll-like receptor 22 in
1557 turbot (Scophthalmus maximus). *FISH & SHELLFISH IMMUNOLOGY* 66:156–162.

1558 144. Paria A, Makesh M, Chaudhari A, Purushothaman CS, Rajendran KV (2018) Toll-like receptor
1559 (TLR) 22, a non-mammalian TLR in Asian seabass, Lates calcarifer. Characterisation, ontogeny
1560 and inductive expression upon exposure with bacteria and ligands. *Developmental & Comparative*
1561 *Immunology* 81:180–186.

1562 145. Qi Z *et al. (*2018) Molecular characterization of three toll-like receptors (TLR21, TLR22, and
1563 TLR25) from a primitive ray-finned fish Dabry's sturgeon (Acipenser dabryanus). *FISH &*
1564 *SHELLFISH IMMUNOLOGY* 82:200–211.

1565 146. Nowoshilow S *et al. (*2018) The axolotl genome and the evolution of key tissue formation
1566 regulators. *NATURE* 554:50–55.

1567 147. Grohme MA *et al. (*2018) The genome of Schmidtea mediterranea and the evolution of core
1568 cellular mechanisms. *NATURE* 554:56–61.

148. Chin C-S *et al. (*2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *NATURE METHODS* 10:563.

149. Cantarel BL *et al. (*2008) MAKER. An easy-to-use annotation pipeline designed for emerging model organism genomes. *GENOME RESEARCH* 18:188–196.

150. Campbell MS, Holt C, Moore B, Yandell M (2014) Genome Annotation and Curation Using MAKER and MAKER-P. *Current protocols in bioinformatics* 48:4.11.1-39.

151. Korf I (2004) Gene finding in novel genomes. *BMC BIOINFORMATICS* 5.

152. Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics (Oxford, England)* 24:637–644.

153. Smit AFA, Hubley R, Green P (2013-2015) *RepeatMasker Open-4.0*.

154. Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *MOBILE DNA* 6.

155. Waterhouse RM *et al. (*2017) BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *MOLECULAR BIOLOGY AND EVOLUTION* 35:543–548.

156. Kriventseva EV, Zdobnov EM, Simão FA, Ioannidis P, Waterhouse RM (2015) BUSCO. Assessing genome assembly and annotation completeness with single-copy orthologs. *BIOINFORMATICS* 31:3210–3212.

157. Dunn NA *et al.* GMOD/Apollo: 2.2.0 JB#1.15.4-release.

158. Lee E *et al. (*2013) Web Apollo. A web-based genomic annotation editing platform. *Genome biology* 14:R93.

159. Jaillon O *et al. (*2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *NATURE* 431:946–957.

160. Pauli A *et al. (*2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *GENOME RESEARCH* 22:577–591.

161. Baird NA *et al. (*2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 3:e3376.

1595 162. Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary
1596 diversification as revealed in replicate lake-stream stickleback population pairs. *Mol Ecol* 21:2852–
1597 2862.

1598 163. Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in
1599 threespine stickleback fish. *NATURE COMMUNICATIONS* 6:8767.

1600 164. Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using
1601 Stacks. *NATURE PROTOCOLS* 12:2640–2659.

1602 165. Ocalewicz K, Sapota M (2011) Cytogenetic characteristics of the round goby Neogobius
1603 melanostomus (Pallas, 1814) (Teleostei. Gobiidae: Benthophilinae). *Marine Biology Research*
1604 7:195–201.

1605 166. Hohenlohe PA *et al. (*2010) Population Genomics of Parallel Adaptation in Threespine
1606 Stickleback using Sequenced RAD Tags. *PLOS GENETICS* 6:e1000862.

1607 167. Kearse M *et al. (*2012) Geneious Basic. An integrated and extendable desktop software
1608 platform for the organization and analysis of sequence data. *BIOINFORMATICS* 28:1647–1649.

1609 168. Ward MN *et al. (*2008) The molecular basis of color vision in colorful fish. Four long wave-
1610 sensitive (LWS) opsins in guppies (Poecilia reticulata) are defined by amino acid substitutions at
1611 key functional sites. *BMC EVOLUTIONARY BIOLOGY* 8:210.

1612 169. Rennison DJ, Owens GL, Taylor JS (2012) Opsin gene duplication and divergence in ray-
1613 finned fish. *Molecular Phylogenetics and Evolution* 62:986–1008.

1614 170. Lin J-J, Wang F-Y, Li W-H, Wang T-Y (2017) The rises and falls of opsin genes in 59 ray-
1615 finned fish genomes and their implications for environmental adaptation. *Scientific reports* 7.

1616 171. Register EA, Yokoyama R, Yokoyama S (1994) Multiple origins of the green-sensitive opsin
1617 genes in fish. *JOURNAL OF MOLECULAR EVOLUTION* 39:268–273.

1618 172. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5. Improvement in accuracy of
1619 multiple sequence alignment. *NUCLEIC ACIDS RESEARCH* 33:511–518.

1620 173. Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2. More models, new
1621 heuristics and parallel computing. *NATURE METHODS* 9:772.

1622 174. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large
1623 phylogenies by maximum likelihood. *Systematic biology* 52:696–704.

1624 175. Ronquist F, Huelsenbeck JP (2003) MrBayes 3. Bayesian phylogenetic inference under mixed
1625 models. *BIOINFORMATICS* 19:1572–1574.

1626 176. Hajkova P *et al. (*2010) Genome-Wide Reprogramming in the Mouse Germ Line Entails the
1627 Base Excision Repair Pathway. *Science* 329:78.

1628 177. Brawand D *et al. (*2014) The genomic substrate for adaptive radiation in African cichlid fish.
1629 *NATURE* 513:375–381.

1630 178. Peichel CL, Sullivan ST, Liachko I, White MA (2017) Improvement of the Threespine
1631 Stickleback Genome Using a Hi-C-Based Proximity-Guided Assembly. *The Journal of heredity*
1632 108:693–700.

1633 179. Trifinopoulos J, Nguyen L-T, Haeseler A von, Minh BQ (2016) W-IQ-TREE. A fast online
1634 phylogenetic tool for maximum likelihood analysis. *NUCLEIC ACIDS RESEARCH* 44:W232-5.

1635 180. Hoang DT, Chernomor O, Haeseler A von, Minh BQ, Le Vinh S (2018) UFBoot2. Improving
1636 the Ultrafast Bootstrap Approximation. *MOLECULAR BIOLOGY AND EVOLUTION* 35:518–522.

1637 181. Altschul S (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215:403–
1638 410.

1639 182. Finn RD *et al. (*2010) The Pfam protein families database. *NUCLEIC ACIDS RESEARCH*
1640 38:D211-22.

1641 183. Sigrist CJA *et al. (*2010) PROSITE, a protein domain database for functional characterization
1642 and annotation. *NUCLEIC ACIDS RESEARCH* 38:D161-6.

1643 184. Sievers F *et al. (*2011) Fast, scalable generation of high-quality protein multiple sequence
1644 alignments using Clustal Omega. *MOLECULAR SYSTEMS BIOLOGY* 7:539.

1645 185. Maddison WP, Maddison MP (2016) Mesquite: a modular system for evolutionary analysis.
1646 Version 3.10.

1647 186. Stamatakis A (2014) RAxML version 8. A tool for phylogenetic analysis and post-analysis of
1648 large phylogenies. *Bioinformatics (Oxford, England)* 30:1312–1313.

1649 187. Rambaut A (2016) Figtree v1.4.3: Tree figure drawing tool.

188. Camacho C *et al. (*2009) BLAST+. Architecture and applications. *BMC BIOINFORMATICS* 10:421.

189. Quinlan AR, Hall IM (2010) BEDTools. A flexible suite of utilities for comparing genomic features. *BIOINFORMATICS* 26:841–842.

190. Edgar RC (2004) MUSCLE. A multiple sequence alignment method with reduced time and space complexity. *BMC BIOINFORMATICS* 5:113.

191. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *MOLECULAR BIOLOGY AND EVOLUTION*.

192. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3. Fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* 27:1164–1165.

193. Stamatakis A (2006) RAxML-VI-HPC. Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)* 22:2688–2690.

194. Eddy SR (2011) Accelerated Profile HMM Searches. *PLOS COMPUTATIONAL BIOLOGY* 7.

195. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7. Improvements in performance and usability. *MOLECULAR BIOLOGY AND EVOLUTION* 30:772–780.

196. Wang P *et al. (*2018) Factors Influencing Gene Family Size Variation Among Related Species in a Plant Family, Solanaceae. *GENOME BIOLOGY AND EVOLUTION* 10:2596–2613.

197. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo. A sequence logo generator. *GENOME RESEARCH* 14:1188–1190.

198. Edwards JR, Yarychkivska O, Boulard M, Bestor TH (2017) DNA methylation and DNA methyltransferases. *EPIGENETICS & CHROMATIN* 10:23.

199. Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE. Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS ONE* 6:e22594.

200. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B (2017) PartitionFinder 2. New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *MOLECULAR BIOLOGY AND EVOLUTION* 34:772–773.

1677 201. Guindon S *et al. (*2010) New algorithms and methods to estimate maximum-likelihood
1678 phylogenies. Assessing the performance of PhyML 3.0. *Systematic biology* 59:307–321.

1679 202. Huelsenbeck JP, Ronquist F (2001) MRBAYES. Bayesian inference of phylogenetic trees.
1680 *BIOINFORMATICS* 17:754–755.

1681 203. Benson G (1999) Tandem repeats finder. A program to analyze DNA sequences. *NUCLEIC*
1682 *ACIDS RESEARCH* 27:573–580.

1683 204. Xu Z, Wang H (2007) LTR_FINDER. An efficient tool for the prediction of full-length LTR
1684 retrotransposons. *NUCLEIC ACIDS RESEARCH* 35:W265-W268.

1685 205. Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de
1686 novo detection of LTR retrotransposons. *BMC BIOINFORMATICS* 9:18.

1687 206. Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification
1688 of de novo predicted LTR retrotransposons. *NUCLEIC ACIDS RESEARCH* 37:7002–7013.

1689

Supplemental_Fig_S1

The round goby genome

VA opsins and pinopsin (outgroup)

LWS (red)

SWS1 (UV)

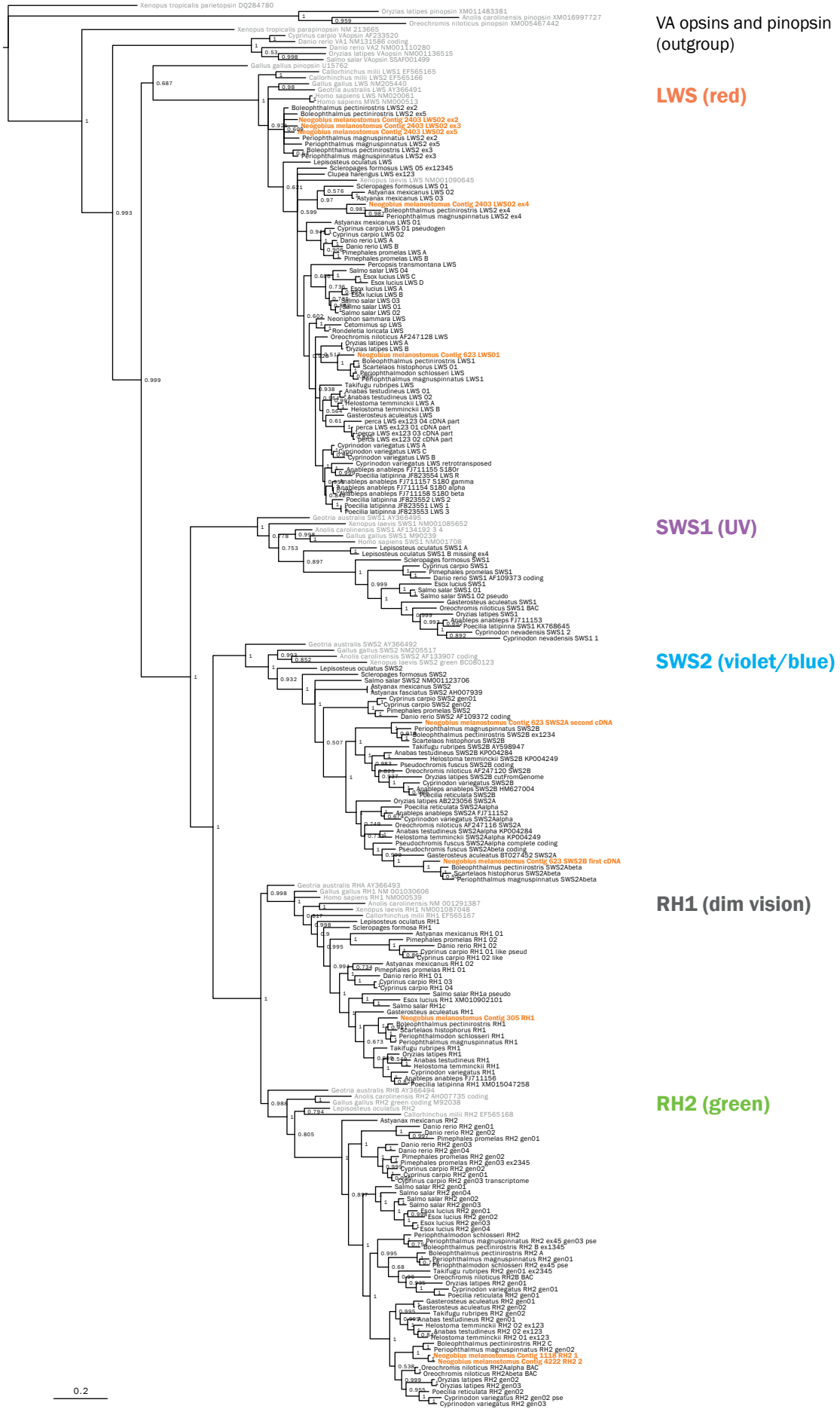
SWS2 (violet/blue)

RH1 (dim vision)

RH2 (green)

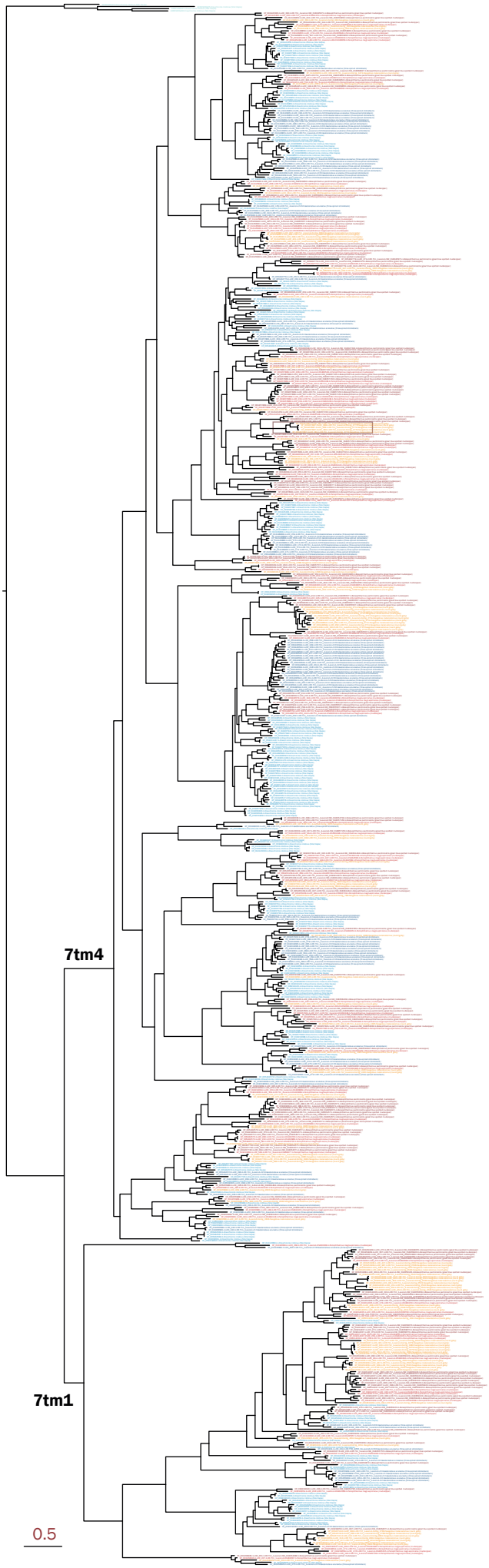
Phylogenetic tree of vertebrate opsin protein sequences. Maximum-likelihood phylogenetic tree with VA opsins and pinopbsins as outgroup. Round goby is indicated in orange. Non-teleost species and the outgroup (VA opsins and pinopsins) are indicated in grey.

Supplemental_Fig_S2
The round goby genome



Phylogenetic tree of vertebrate opsin protein sequences, using single exons. Maximum-likelihood phylogenetic tree with VA opsins and pinopbsins as outgroup. Round goby is indicated in orange. Non-teleost species and the outgroup (VA opsins and pinopsins) are indicated in grey.

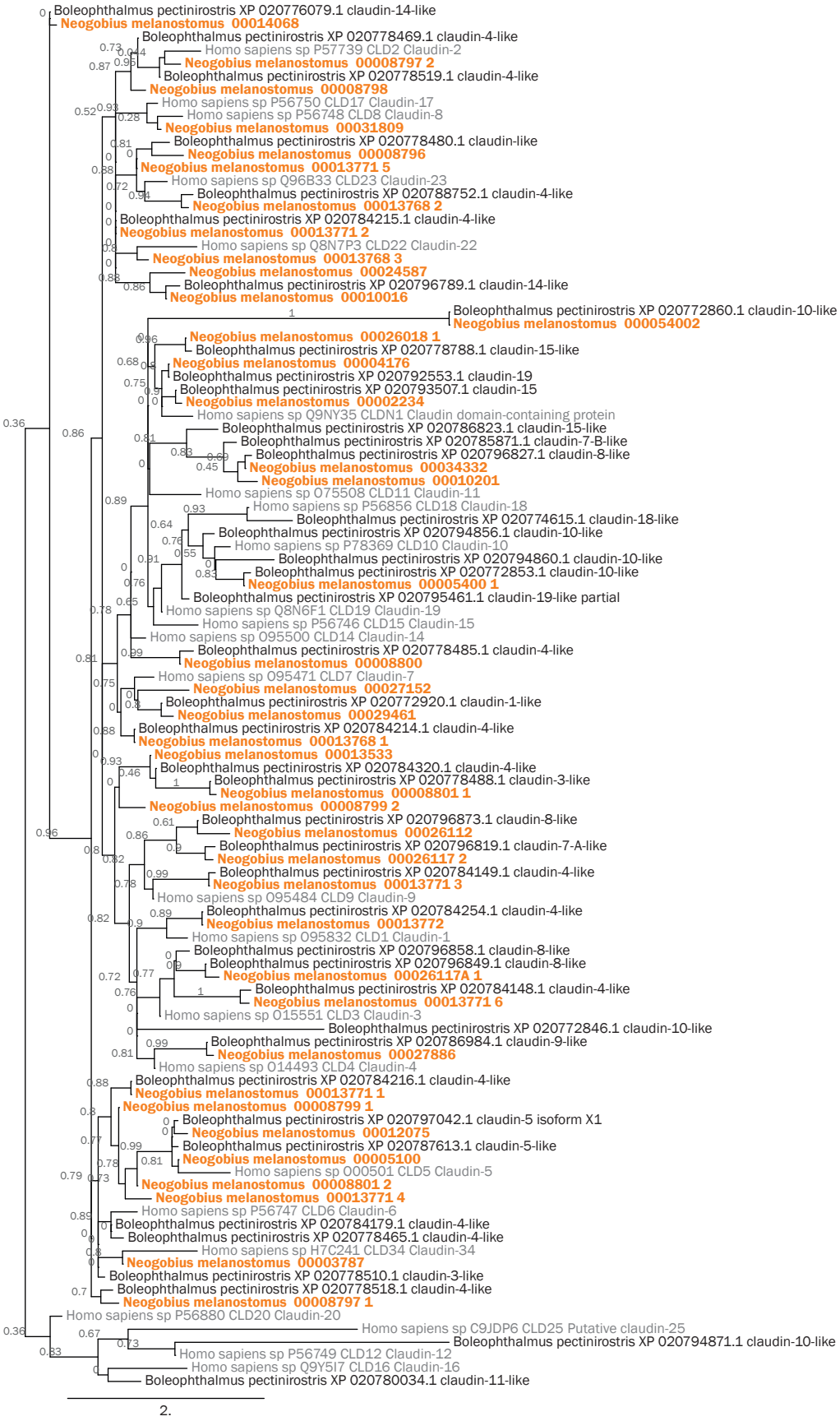
Supplemental_Fig_S3
The round goby genome



Maximum-likelihood phylogenetic tree of percomorph olfactory receptor protein sequences.

Supplemental_Fig_S4

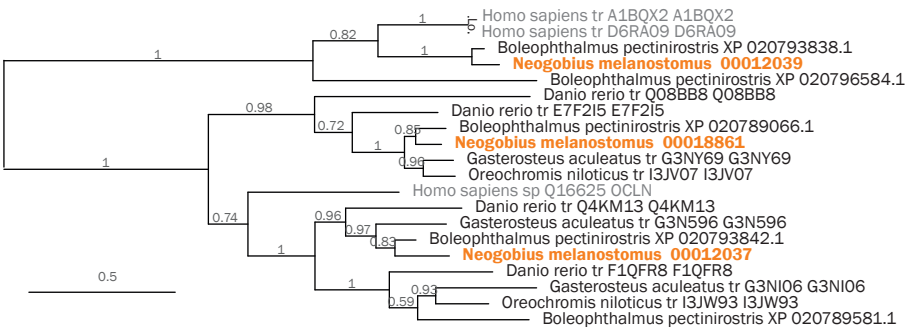
The round goby genome



Phylogenetic tree of vertebrate claudin genes. Maximum-likelihood tree with 100 bootstraps of round goby (Neogobius melanostomus, orange) in relation to great blue-spotted mudskipper (Boleophthalmus pectinirostris) and human (Homo sapiens, grey).

Supplemental_Fig_S5

The round goby genome



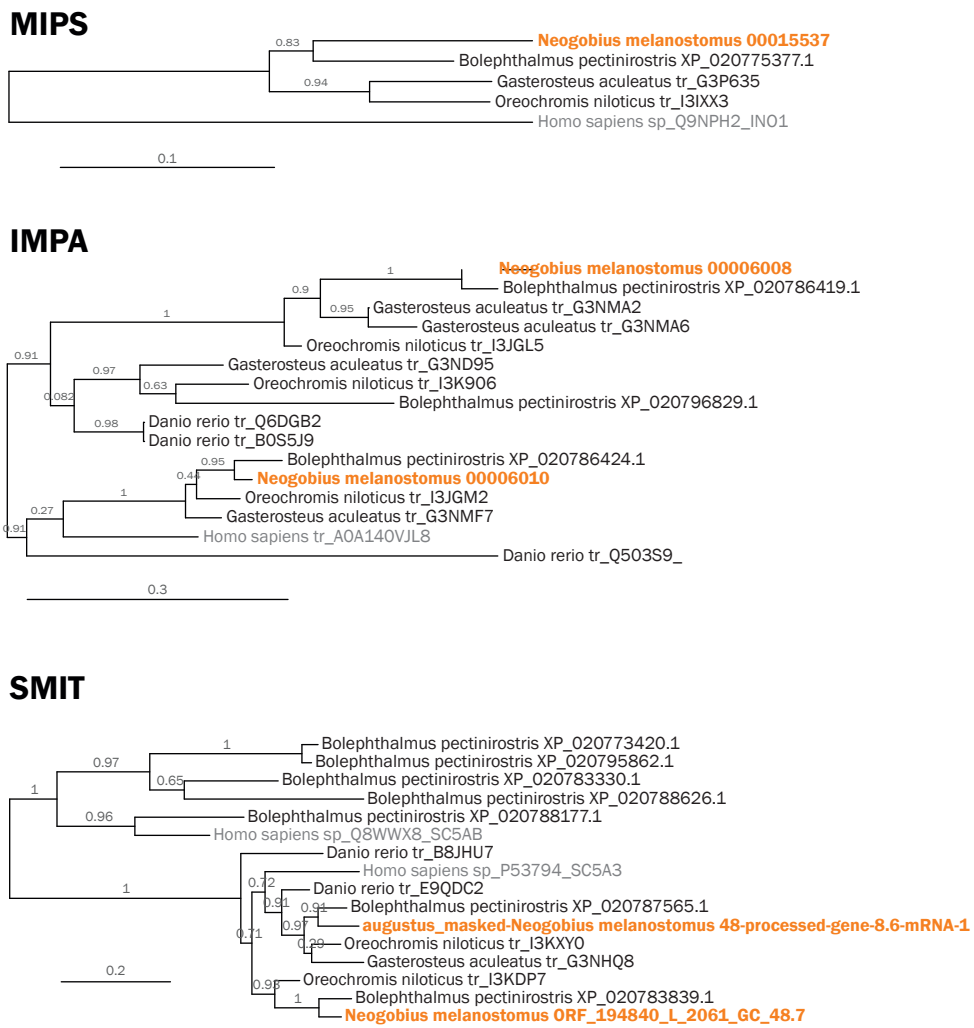
Phylogenetic tree of vertebrate occludin genes. Maximum-likelihood tree with 100 bootstraps of round goby (Neogobius melanostomus, orange) in relation to great blue-spotted mudskipper (Boleophthalmus pectinirostris), stickleback (Gasterosteus aculeatus), nile tilapia (Oreochromis niloticus), zebrafish (Danio rerio), and human (Homo sapiens, grey).

Supplemental_Fig_S6

The round goby genome

Phylogenetic tree of vertebrate ion transporters. Maximum-likelihood tree with 100 bootstraps of round goby (Neogobius melanostomus, orange) in relation to great blue-spotted mudskipper (Boleophthalmus pectinirostris), stickleback (Gasterosteus aculeatus), nile tilapia (Oreochromis niloticus), zebrafish (Danio rerio), and human (Homo sapiens, grey)

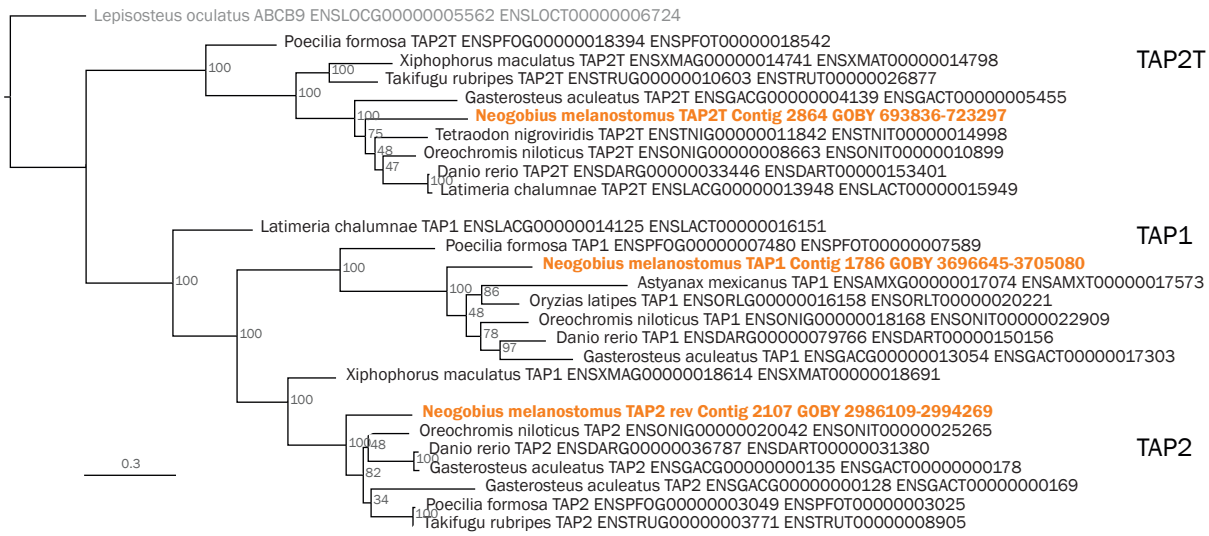
Supplemental_Fig_S7
The round goby genome



Phylogenetic tree of vertebrate genes promoting osmolyte production. Maximum-likelihood tree with 100 bootstraps of round goby (Neogobius melanostomus, orange) in relation to great blue-spotted mudskipper (Boleophthalmus pectinirostris), stickleback (Gasterosteus aculeatus), nile tilapia (Oreochromis niloticus), zebrafish (Danio rerio), and human (Homo sapiens, grey)

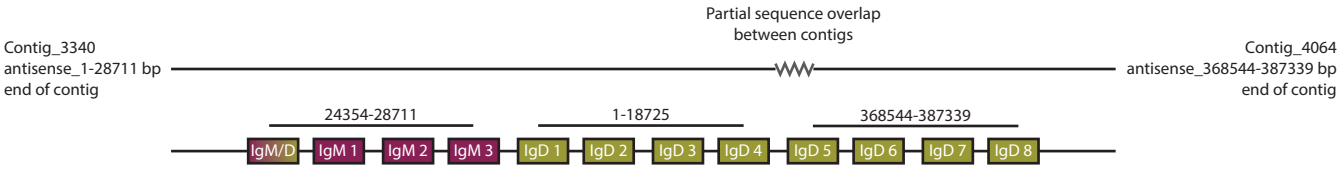
Supplemental_Fig_S8

The round goby genome



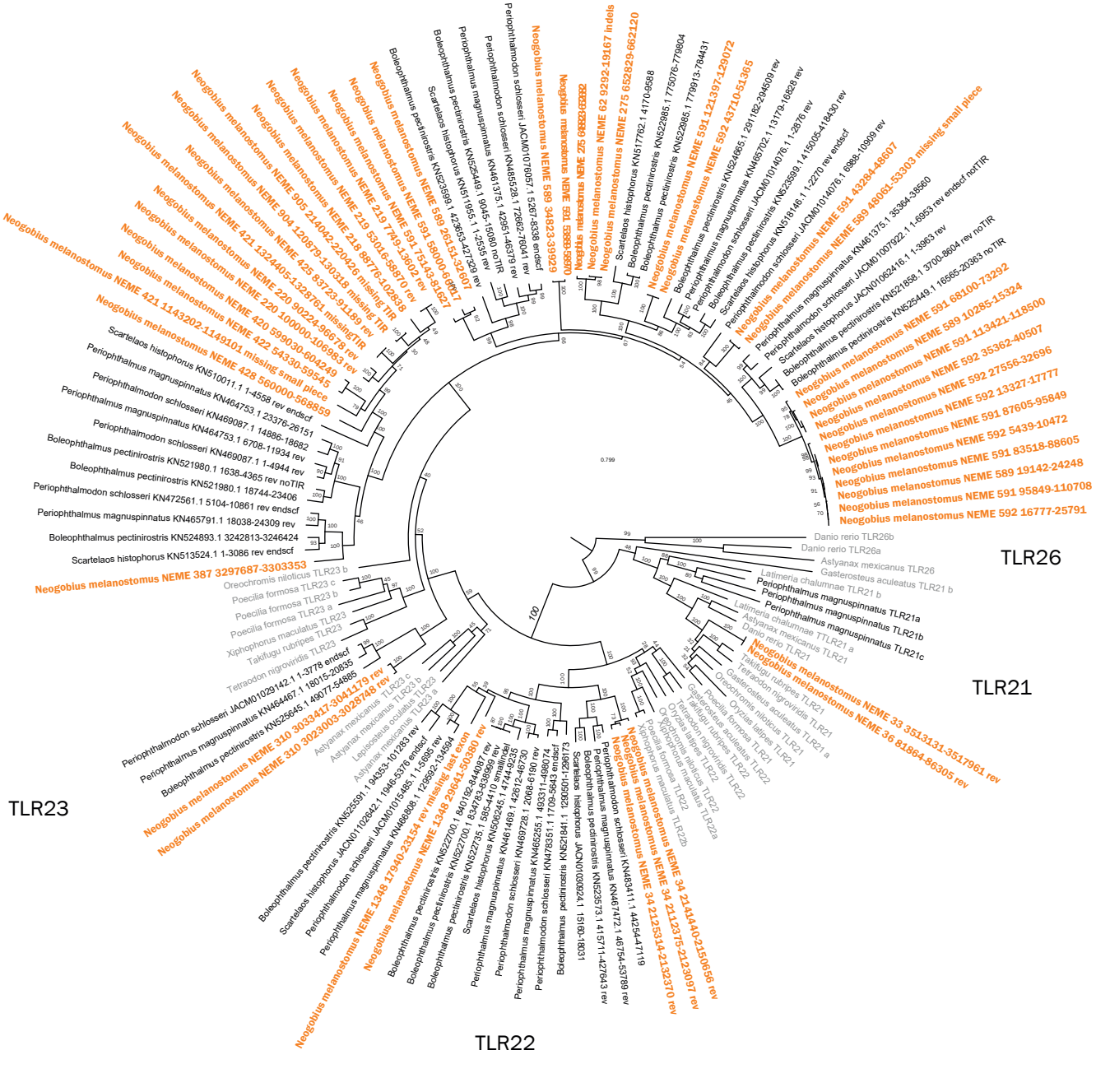
Phylogenetic tree of fish TAP genes. Maximum-likelihood tree with 500 bootstraps. Round goby (Neogobius melanostomus) is labeled orange. Outgroup: Lepistosteus oculatus ABC transporter.

Supplemental_Fig_S9
The round goby genome



The Ig locus spans two contigs and contains IgM and IgD domains.

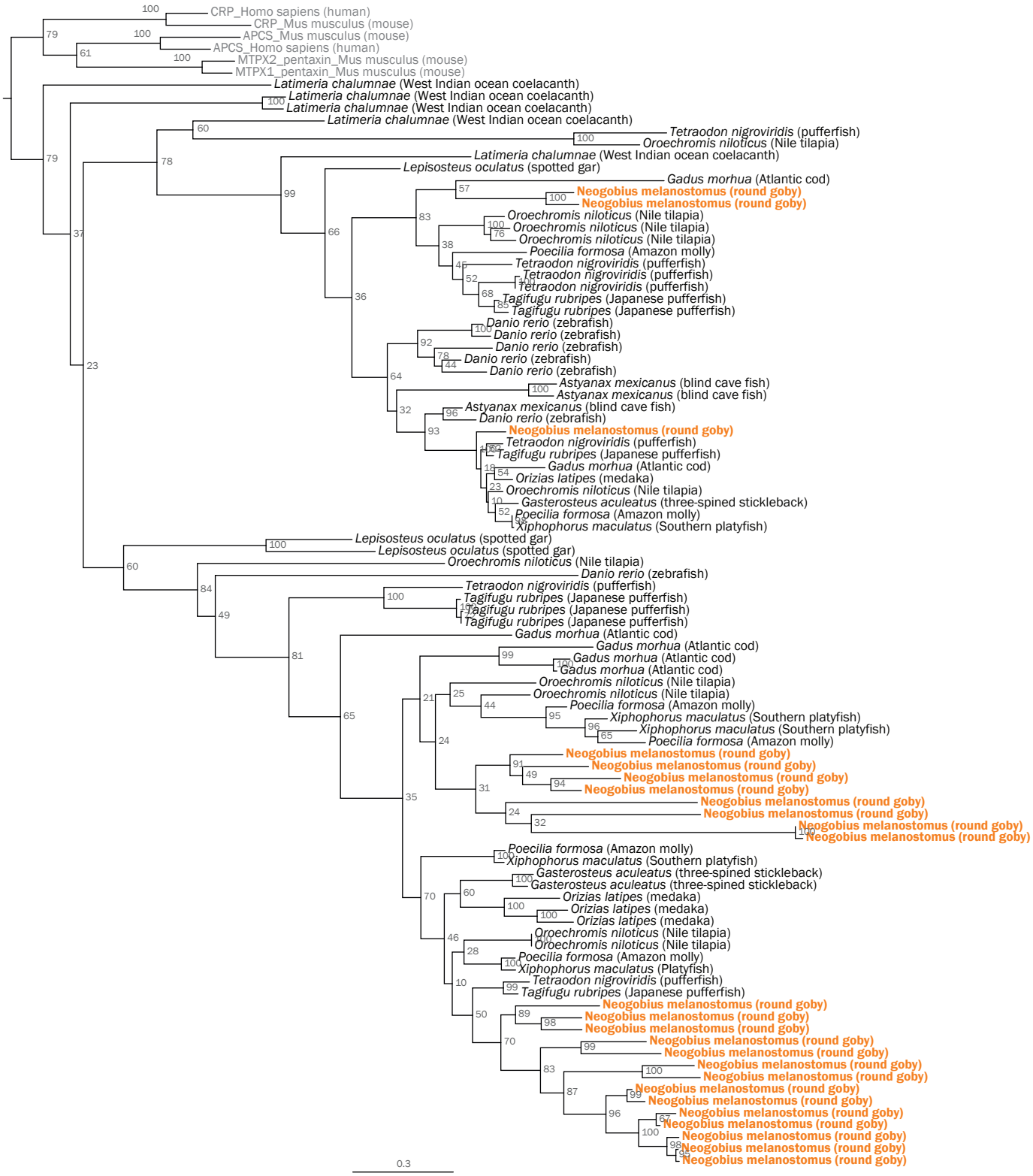
Supplemental_Fig_S10
The round goby genome



Phylogenetic tree of Gobiidae Toll Like Receptor protein sequences. A maximum likelihood phylogenetic tree run with the JTT substitution model and 500 bootstrap replicates on the transmembrane, linker and TIR domain of all TLRs found in Gobiiformes (Neogobius melanostomus, orange; other Gobiidae, black) and selected other fish (grey) for context.

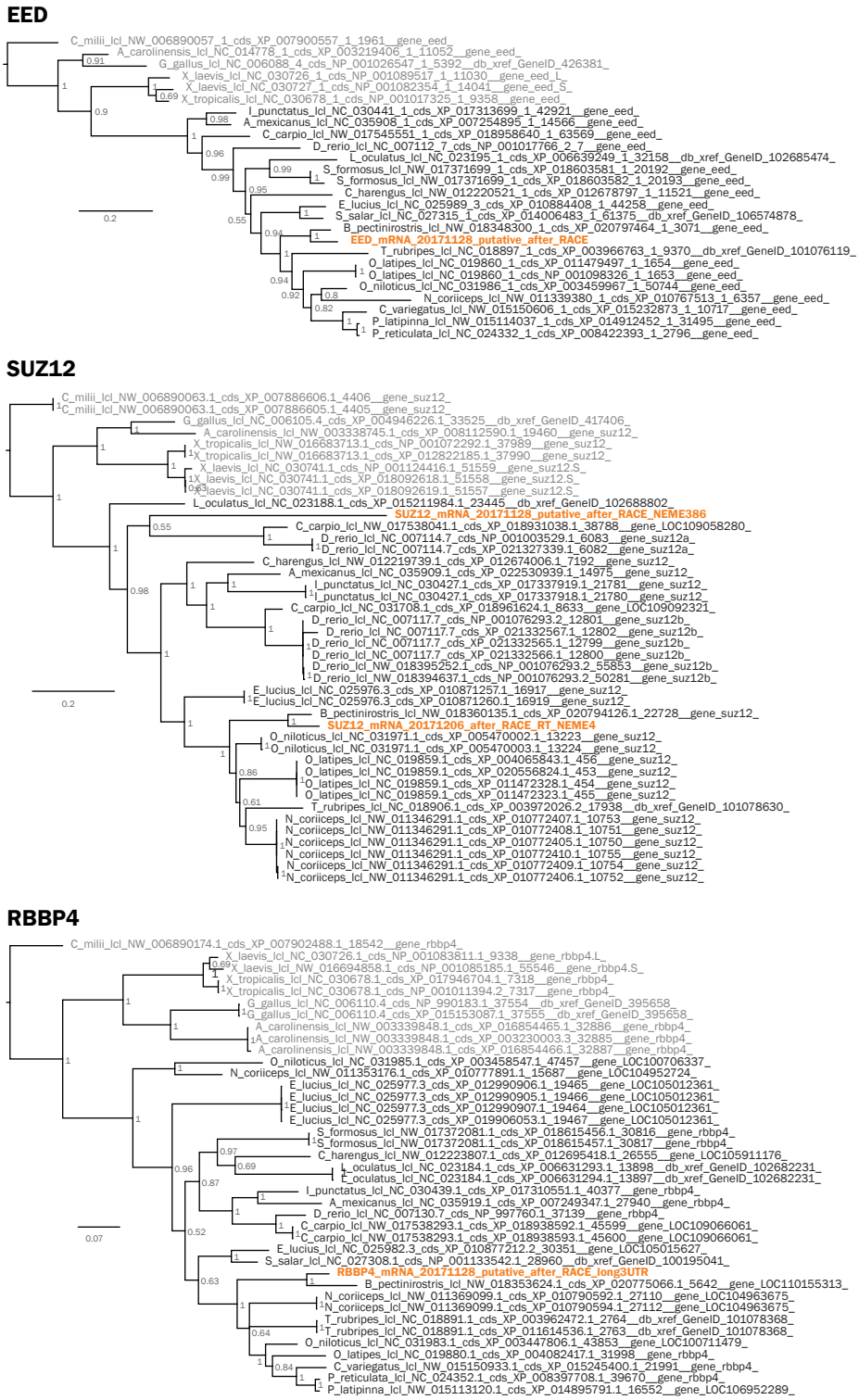
Supplemental_Fig_S11

The round goby genome



Phylogenetic tree of fish CRP/APCS sequences. Maximum Likelihood phylogenetic tree with 500 bootstraps rooted at the split between tetrapods and ray-finned fish. Tetrapods were used as outgroup and are indicated in grey. Round goby is indicated in orange.

Supplemental_Fig_S12
The round goby genome



Phylogenetic tree of vertebrate PRC2 components EED, SUZ12, and RBBP4. Bayesian phylogenetic tree rooted with Australian ghostshark (C. milii). Round goby is indicated in orange.