PhD in Aquatic Biosciences // No. 34 - 2019

Exploration of an anglerfish genome

Arseny Dubin

FACULTY OF BIOSCIENCES AND AQUACULTURE

NORD University

www.nord.no

Exploration of an anglerfish genome

Arseny Dubin

A thesis for the degree of
Philosophiae Doctor (PhD)

PhD in Aquatic Biosciences no. 34 (2019)
Faculty of Biosciences and Aquaculture

PhD in Aquatic Biosciences no. 34 (2019)

Arseny Dubin
Exploration of an anglerfish genome

ISBN: 978-82-93165-33-0

Print: Trykkeriet NORD

Nord University
N-8049 Bodø
Tel: +47 75 51 72 00
www.nord.no

Preface

This dissertation is submitted in fulfilment of the requirements for the degree of Philosophiae Doctor (PhD) at the Faculty of Biosciences and Aquaculture, Nord University. All research presented was original and performed during the Stipendiatprogram Nordland, with financial support from Nord University.

The PhD project team consisted of the following members:

Arseny Dubin: MSc, FBA, Nord University: PhD Student

Steinar Daae Johansen: Professor, FBA, Nord University: Primary Supervisor

Lars Martin Jakt: Researcher, FBA, Nord University: Co-supervisor

NORD universitet

Dubin Arseny

Bodø, Norway 2019

Acknowledgements

I've been involved in a project that many can only dream about: An entire genome project all for myself, with unlimited possibilities for investigation. And not just the genome project, I got to work with an extraordinary group or organisms - anglerfishes!

Because I was lucky with my data and species choice the only limit was time, and I feel extremely fortunate for that.

I was most of all fortunate to have met and worked with many good people during this project.

First, I would like to thank my main supervisor – Steinar Daae Johansen. I took this PhD in large part because of you. I really appreciate all the freedom I had in this project. You are always full of ideas, yet you never force them on your students. You are always up for fun projects with uncertain outcomes, yet I always felt safe about my PhD and certain that I could finish on time. You are a fantastic supervisor and I'm glad that I had an opportunity to work with you again!

I knew most of the team when I started my PhD, except for you, Lars Martin. Since I didn't know you when I learned that you would be my co-supervisor, I decided to ask the opinions of some master and bachelor students who had taken your courses. They said: "He asks too many questions", "He is too hardcore maaan..." or "He is very smart, but quite harsh and asks too many questions". Most were afraid of you and that got me worried as well. They were wrong! You are an extremely interested person and exactly because of your endless curiosity, questions, discussions, I've learned so much! Often, I would just come by for a chat about things I've done recently, then end up with a completely different view on a thing I was doing and a dozen new ideas to test. Our discussions also made me feel calm at times of stress and confident in my project. You are also exceptionally selfless. You help people at the expense of your time, and never ask anything in return. Without your help, not only me but many other students would not have been able to finish. Thank you for everything, Lars Martin!!

I would also like to thank Tor Erik Jørgensen. Without you there would not have been data for me to analyse!

Thank you Truls for your critical reading of the manuscripts and discussions!

I thank my fellow PhD students that kept inviting me for hikes and beer nights even though I kept rejecting most of the invitations. I still appreciate the invitation.

I thank the best ever office mate and a friend - Kyle! We shared an office for 3 years yet the number of times we've met there could be counted on one or two hands.

I thank my parents and friends for their support.

And finally, I would like to thank my girlfriend Uliana for being so understanding of my late work hours even though our time together was so limited, and for your willingness to travel to Norway so many times, for always being there for me, and for your love and support!

Table of contents

List of figures and tables

List of papers

Paper I.

Dubin A, Jørgensen TE, Jakt LM, Moum T, Johansen SD (2017)

The Mitochondrial Genome of the European Anglerfish *Lophius piscatorius* Express Low-Level Substitution Heteroplasmy. Ann Mar Biol Res 4(1): 1019.

Paper II.

Dubin A, Jørgensen TE, Jakt LM, Johansen SD

The mitochondrial transcriptome of the anglerfish *Lophius piscatorius*

Submitted to BMC Research Notes

Paper III.

Dubin A, Jørgensen TE, Moum T, Johansen SD, Jakt LM (2019)

Complete loss of the MHC II pathway in an anglerfish, *Lophius piscatorius*

Biol. Lett. 20190594. http://dx.doi.org/10.1098/rsbl.2019.0594

Article in press

Paper IV.

Dubin A, Jørgensen TE, Moum T, Johansen SD, Jakt LM

Assembly and annotation of an anglerfish genome

Manuscript

List of abbreviations

OxPhos - Oxidative Phosphorylation

MHC – Major Histocompatibility Complex

WGD – Whole Genome Duplication

bp – base pairs

kb – kilo base pairs

Mb – mega base pairs

Gb – giga base pairs

Abstract

The anglerfishes comprise an extremely diverse order of teleosts with unique adaptations. The most notable is sexually parasitism of reproduction where the male attaches to the female. This can result in fusion of two genetically distinct organisms, which would in most vertebrate species result in an immune rejection. However, in sexually parasitic anglerfish fusion occurs with no immune rejection. The mechanisms that have allowed the evolution of such adaptations are of interest not just to evolutionary biology, but perhaps also to biomedical research related to the prevention of allogenic rejection after transplantation. Nevertheless, anglerfishes remain poorly understood. In this project we have produced the first chromosome level assembly of an anglerfish (*Lophius piscatorius*). We also provide an annotation of this genome based on orthology inference and believe that this will provide a comprehensive genetic resource for the study of anglerfish biology facilitating research addressing the evolution of anglerfish specific properties.

As part of an analysis of the initial contig level assembly we characterized the *L. piscatorius* mitochondrial genome and transcriptome. This identified low-level heteroplasmic sites, a species-specific control region indel, as well as a novel long non-coding RNA derived from the Cytochrome Oxidase I locus. Furthermore, we observed a remarkable sequence conservation of the mitochondrial-derived peptide Humanin. These findings contribute to our understanding of mitochondrial regulation and function, and are of interest not only to anglerfish research.

It is thought that sexual parasitism has evolved independently multiple times within the Ceratioidei suborder, suggesting that they may share a common genetic predisposition that facilitates sexual parasitism. As the removal of immune rejection is a requirement for the fusion of two individuals it is possible that this predisposition arises from a modified immune system that may be shared with the non-parasitic anglerfish taxons. Given that two teleost taxons (Gadiformes and *Syngnathus*) have previously been reported to lack the MHC II arm of the adaptive immune system we made use of the initial draft genome assemblies to establish the absence or presence of

MHC II in *L. piscatorius*. Surprisingly we found an absence of exactly the same five (of 30 assayed) genes absent in Gadiformes. This observation implies that these five genes (CD4, CD74 A/B, MHC II α/β) comprise a core set of MHC II genes that have no essential functions external to MHC II, and suggests the possibility that loss of MHC II may have been one of the events that enabled the development of sexual parasitism in anglerfish.

To annotate the final chromosome level assemblies, we made use of in silico gene predictions supported by evidence from RNA followed by an orthology based functional annotation. An analysis of the resulting annotation confirmed that *L. piscatorius* has a fairly typical teleost genome in terms of genome size, global gene repertoire and gene feature composition. We also observed a chromosomal orthology with several teleost species that argues that the scaffolds reported here do indeed represent physical chromosomes. These analyses also revealed a teleost specific bimodality in intron length distribution that could be correlated to genome size within the teleosts, suggesting a coupling between the mechanisms governing intron and genome size in teleosts.

The work presented in this thesis not only provides new genome resources that should facilitate further research into the weird and wonderful world of the anglerfishes, but also confirms an unexpected plasticity in teleost adaptive immunity. Surprisingly we were also able to observe fundamental genome properties related to intron size that have not previously been reported. Our work thus touches not only on the specifics of teleost immunology but also on general mechanisms underlying genome evolution in the teleosts.

1 Introduction

1.1 The mitochondrial genome

Mitochondrial genes and genomes have been, and are still widely used as the basis for phylogenetic analyses as homologous sequences are relatively easy to obtain and hence available from a wide range of taxa from protists, to plants, fungi and animals (Harrison 1989; Boore 1999; Galtier et al. 2009; Jiang et al. 2016; Xin et al. 2017; Bronstein et al. 2018). This makes any newly sequenced mitochondrial genome a useful addition to an enormous collection of existing full mitochondrial genome sequences. Mitochondrial genomes often appear as a "by-catch" in next generation sequencing data, as they assemble relatively well even with a low coverage sequencing due to their high copy number, distinct sequence and organisation compared to nuclear DNA.

The vertebrate mitochondrial genome is circular and has a highly conserved organisation and sequence. It contains 13 conventional protein coding genes, 2 ribosomal RNAs, and 22 transfer RNA genes (Figure 1) (Boore 1999; Jørgensen and Johansen 2018). These 13 proteins are essential parts of the oxidative phosphorylation (OxPhos) pathway, the major part of which is encoded in the nuclear genome (72 genes) (Kühlbrandt 2015). The vertebrate mitochondrial genome has coding potential on both strands denoted as the Heavy and Light (H and L) strands, originally separated based on their buoyancy in caesium chloride density gradient centrifugation (Barroso Lima and Prosdocimi 2018). The H-strand encodes 2 ribosomal subunits, 14 tRNAs and 12 mitochondrial proteins. The L-strand encodes 8 tRNAs and one protein (reviewed in (Jørgensen and Johansen 2018)). In addition to the conventional OxPhos proteins, the vertebrate mitochondrial genome codes for several unconventional peptides. One of these peptides, Humanin, appears to be highly conserved in many species (Lee et al. 2013; Jørgensen and Johansen 2018). The largest non-coding part of the mitochondrial genome is the Control Region (CR). It is located between the genes of tRNA Pro and tRNA Phe and contains the displacement loop (D-loop), transcription initiation sites for both strands, and the origin of replication for H-strand. The L-strand origin of replication is located between the genes of tRNA Asp and tRNA Cys and is functionally conserved in most vertebrates (Jørgensen and Johansen 2018).

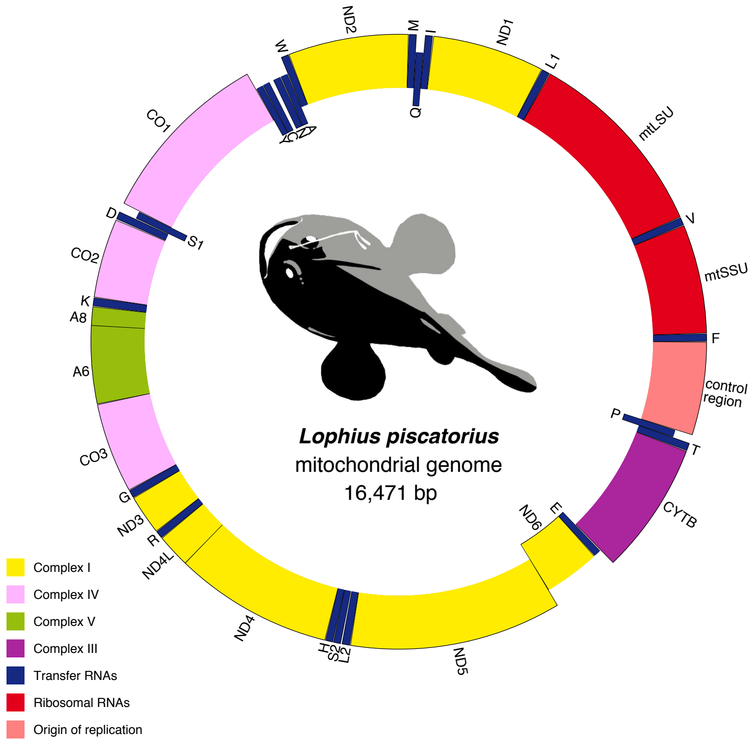


Figure 1. Mitochondrial genome organization of L. piscatorius

Features derived from the heavy (H) and light (L) strands are shown as outer and inner arc segments respectively. tRNAs, rRNAs and the control region are indicated by blue, salmon pink and red segments respectively. Coding regions are coloured by function. Gene abbreviations: mtSSU and mtLSU, mitochondrial small- and large-subunit ribosomal RNA; ND1-6, NADH dehydrogenase subunit 1 to 6 (yellow); COI-III, cytochrome oxidase subunit I to III (ligth purple); A6 and A8, ATPase subunit 6 and 8 (green); CYTB, cytochrome b (purple); tRNA genes are indicated by the standard amino acids code (blue).

1.2 Variation in nuclear genome size

It is reasonable to assume that with the increase in organisational complexity arising during evolution, over time the number of genes and the associated genome size would increase. Originally, genome size estimations were based on the amount of nuclear DNA in the cell. For such studies, the term C-value (C - stands for constant) was introduced by H. Swift (Swift 1950; Greilhuber 2005) and in general it refers to a haploid nuclear DNA content. The results of these studies were surprising, as though there seemed to be a general trend for a genome size increase with the increase in complexity (eg. Prokaryotes-to-Eukaryotes), there was no correlation between organismal complexity and DNA content (figure2).

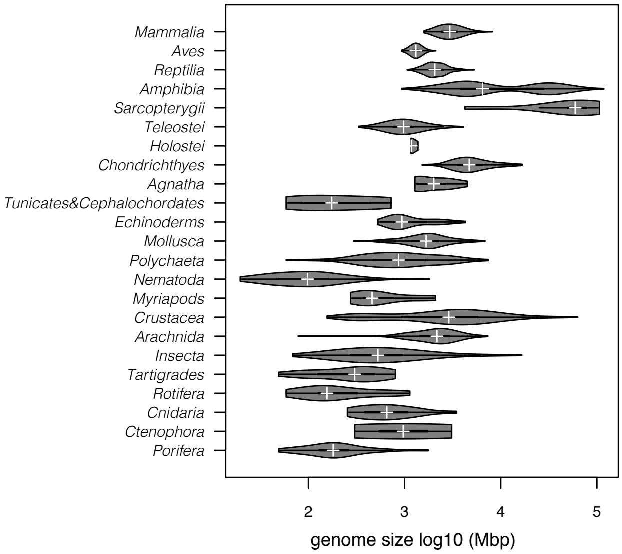


Figure 2. C-value based genome size estimations

The distributions of log transformed (base 10) genome sizes for different taxonomic groups from the Porifera (sponges) to mammals are shown as violin plots, with the median size indicated by white crosses. The units of the axis are log10 transformed genome sizes in mega base pairs; hence 3 indicates a genome size of 1 Gb. C-values were obtained from the Animal Genome Size Database, Gregory, T.R. (2019) (http://www.genomesize.com). Picograms were converted to base pairs (bp) by: bp = mass in pg x 0.978 e9.

At that time the amount of DNA was assumed to be directly related to the number of protein coding genes, hence the discrepancy between expectation and observation. This conflict was termed the 'C-value paradox' or later the 'C-value enigma' (Gregory 2001, 2005; Elliott and Gregory 2015). As more genomes were sequenced, it became clear that the genome size variation in Eukaryotes is not directly related to the variation in the number of protein coding genes and that the number of genes only weakly, if at all, correlated with the organismal complexity. There is, however, a positive correlation between the genome size and the number of genes in some Eukaryotes (Gregory 2005; Elliott and Gregory 2015).

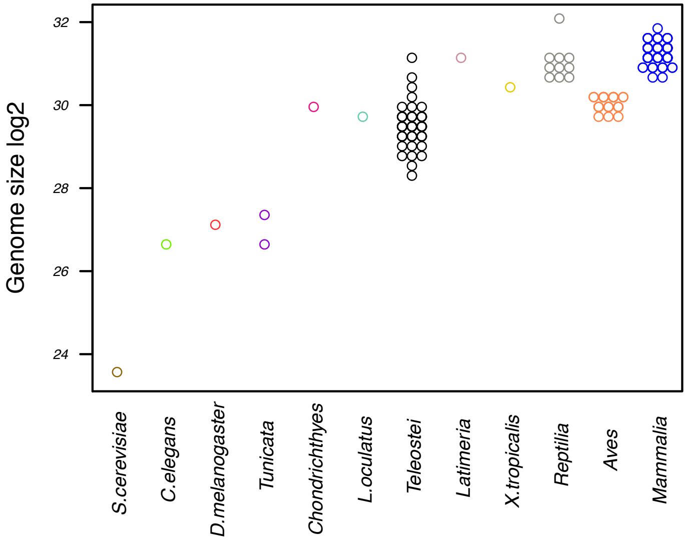


Figure 3. Genome sizes of sequenced genomes

Genome size estimates of sequenced genomes obtained from Ensembl. Each point represents the size of a single genome. The points are grouped along the x-axis by taxonomic groups as indicated. The genome size has been log transformed using base 2 such that difference of one unit indicates a doubling of genome size. Most teleosts have genome sizes less than 1Gbp ($2^{29.9}$) whereas mammals generally have a genome size around 3 Gb ($2^{31.5}$). Reptilia represents all Ensembl species that belong to the clade Sauria with the exception of birds, which are shown as a separate group: Aves. Overlapping points were collapsed using the R package Beeswarm using the center method.

Figure 3 shows that there is a considerable variation of genome sizes in some clades (eg. teleosts), whereas in others it is quite conserved (eg. birds) and it is not correlated with the intuitive complexity of the organisms. As another example, the genome sizes of a tarantula and a velvet spider are 5.8 Gb and 2.8 Gb, respectively (Sanggaard et al. 2014), while the human genome size is 3 Gb (GRCh38.p12, Ensembl Genome Browser). The genome sizes of a human parasite, the trematode *Clonorchis sinensis* is 516 Mb (Wang et al. 2011) whereas the dinoflagellate (single cell eukaryote) *Symbiodinium minutum* is approximately 1,5 Gb (Shoguchi et al. 2013). In comparison the genome of the teleosts *Takifugu rubripes* is 393 Mb (Kai et al. 2011) and *Gadus morhua* is 643 Mb (Tørresen et al. 2017).

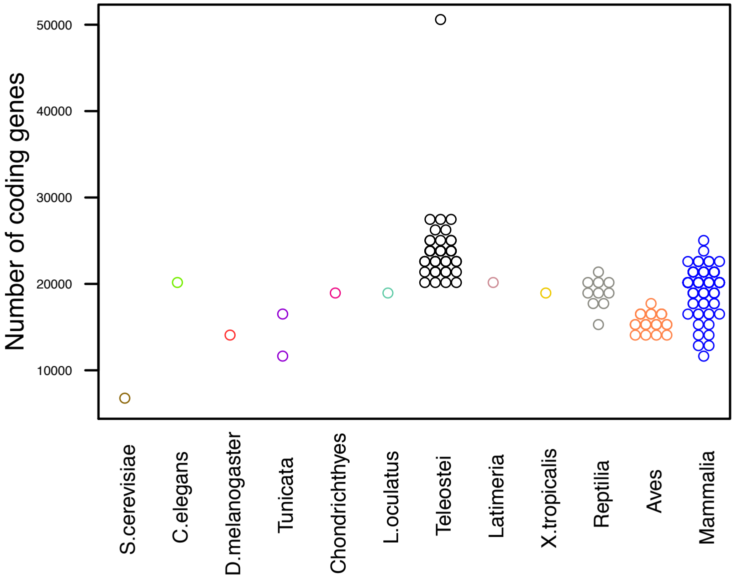


Figure 4. Number of annotated coding genes in Ensembl species

Each point represents the number of genes reported in a single assembly. The data is arranged as in Fig. 3. Vertebrates generally have around 20,000 annotated genes. It is unclear as to what extent the variance in gene number estimates represents biological variability or differences in the accuracy of assembly and annotation.

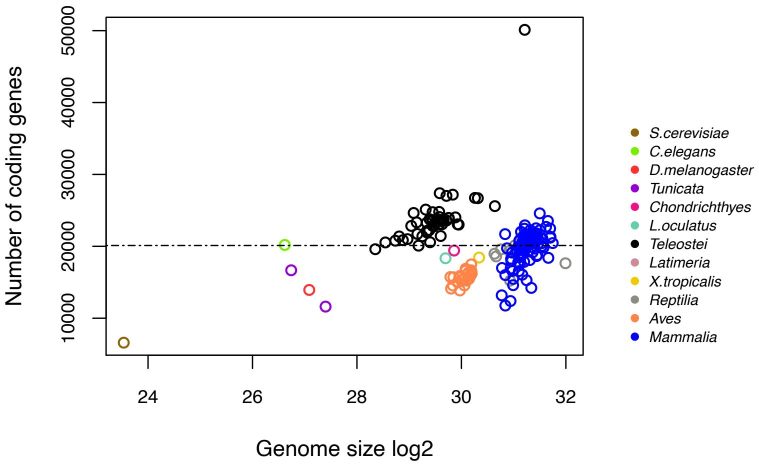


Figure 5. Genome size and gene numbers

(log2) Genome size plotted against the number of annotated coding genes reported by Ensembl. Each point represents a single genome assembly. Colours indicate taxonomic grouping as shown. Reptilia represents all Ensembl species that belong to the clade Sauria with exception of birds, which are shown as a separate group – Aves.

This considerable variation of genome sizes in all domains of life only partially comes down to the increase in the number of functional genes (figure 4,5). For example, the smallest genome of a self-replicating, free-living cellular organism is found in *Mycoplasma genitalium*. It is 580 kb long and contains approximately 475 genes (Fraser et al. 1995; Fookes et al. 2017). Though, the smallest genome in terms of physical size in a cellular organism, is shared between the obligate symbionts *Carsonella ruddii* (endosymbiont of certain gamma-proteobacteria) whose genome size is 159 kb (Tamames et al. 2007), and *Nanoarchaeum equitans* ("ecto"symbiont of the Crenarchaeota genus *Ignicoccus*) which has a genome of 490 kb that contains 536 genes and has a coding efficiency of 95% (Waters et al. 2003). Thus, the difference in the genome sizes between *Mycoplasma* and *Nanoarchaeum*, at least partially comes down to differences in the number and length of noncoding regions.

The largest sequenced genomes belong to loblolly pine (*Pinus taeda*; 23.2 Gb) (Neale et al. 2014) and axolotl (*Ambystoma mexicanum*; 32 Gb) (Nowoshilow et al. 2018). While the estimated largest genome belongs to a single cell eukaryote *Polychaos dubium* (~670 Gb); however, these estimates are in general considered untrustworthy (Friz 1968; Elliott and Gregory 2015). If *P. dubium* measurements are discounted, the estimated largest genomes belong to the *Paris japonica* (~152 Gb) and marbled lungfish (*Protopterus aethiopicus*; ~130 Gb) (Pellicer et al. 2010). Interestingly, using deep mRNA sequencing from 22 tissues authors identified 23,251 protein-coding genes in the *A. mexicanum* genome, which is similar to other vertebrates with much smaller genomes (Nowoshilow et al. 2018). Such drastic differences in the number of genes and genome sizes raises questions about the source of this diversity, the possible minimal number of genes and the role of a non-coding part of the genome (figure 5).

1.3 Gene duplications and their role in adaptive evolution

There are two ways in which new sequences can be added to a genome: by duplication events (domain/gene/genome) or via lateral gene transfers from another organism (Chen et al. 2013; Brown 2017). Duplication events are the most common way of acquiring new genes across all domains of life, while gene transfers are much more prevalent among prokaryotes (Chen et al. 2013; Brown 2017). These processes do not necessarily result in the appearance of new functionality but provide a new genetic material for selection to act upon.

Gene duplications usually result from abnormal recombination events (e.g. unequal crossing over and unequal exchange between sister chromatids) or retrotranspositions. In the latter case host mRNA is reverse transcribed into DNA and subsequently integrated into the genome (Pink et al. 2011; Chen et al. 2013; Brown 2017). At the duplication site the copies are identical (except for retrotransposed genes) and will have to have the same functionality unless the regulatory context of the two duplicates is sufficiently different to cause differences in gene expression. Due to the selective pressure acting on differences resulting from mutation and recombination events the duplicated genes will either stay the same, diverge into two different genes with similar or completely new functions, or one copy will degrade and become a pseudogene (Chen et al. 2013; Brown 2017).

Gene duplicates keep the original function when the sequences or regulatory context is the same. In this case the benefit of having two copies of the same gene could come from the increase in the rate of gene product synthesis. A good example is the ribosomal RNA genes which are present in multiple copies in most organisms (from ~350 copies in the human genome and ~4000 in the pea genome) (Brown 2017). Comparative genomics studies in vertebrates and invertebrates show that duplicates of genes having important functions in development often (than expected by random chance) remain intact and preserve the original function (Chen et al. 2013; Van de Peer et al. 2017). It is hypothesised that the danger of having abnormal products of such important genes keeps copies under strict purifying selection and thus preserving the duplicates (dominant-negative hypothesis) (Van de Peer et al. 2017). In rainbow trout, 52% of the gene duplicates that originated in the salmonid-specific genome duplication event were lost, and the remaining 48% remained in duplicates. Genes that are involved in embryonic development, neuronal synapse development, and transcription factors are among those 48% (Berthelot et al. 2014).

When benefits of having two copies is negligible, random mutations will start to accumulate and may cause pseudogenization or, more rarely, neofunctionalization in the gene function (Chen et al. 2013; Brown 2017). It is often considered that pseudogenes have lost their use for the organism and will slowly degrade with time. Surprisingly, it has been shown that some pseudogenes can play important parts of the RNAi system and regulate gene expression or become decoys for miRNAs preventing tumorigenesis (Pink et al. 2011).

Occasionally, mutations will change the duplicate gene product in a way that is beneficial to the organism. Duplication/divergence cycles can happen multiple times with any of the copies. Speciation events can cause descendants of the original gene to diverge even further due to differences in selective pressure. These processes give rise to gene families and super families (Chen et al. 2013; Brown 2017). Genes that perform the same function, but split due to a speciation event, are called orthologues. Descendants of the ancestral gene that underwent a duplication event are called paralogues. Ohnologues - are paralogues that originated as a result of a whole-genome duplication event (Brown 2017; Van de Peer et al. 2017).

New gene functions can also arise from domain duplication and domain shuffling, causing structural domains to be repeated and making genes longer. Over time, duplicated domains may diverge and change the original gene activity or function, or can remain the same if the increase in the number of structural domains is beneficial (e.g. making the product more stable). Domain shuffling is the process where domains from different genes are duplicated and combined resulting in a gene that might have a completely novel function (Chen et al. 2013; Brown 2017). There are several other ways in which novel genes and gene products can originate, including transposon-domestication, reading frame shifts, gene fusion and fission, or even de-novo from non-coding regions as in the case of parts of the antifreeze peptides of icefishes (Logsdon and Doolittle 1997; Chen et al. 2013; Kim et al. 2019).

1.4 Whole genome duplication events

Even though gene and domain duplications are more frequent, whole genome duplications (WGD) are by far the greatest sources of new genetic material for selection (Brown 2017; Van de Peer et al. 2017). With the increase in the number of sequenced genomes, it is possible to trace multiple genome duplication events back in time, and assess the frequency of WGDs and their consequences. WGDs are considered to be a common cause of speciation in plants and some animals (Leggatt and Iwama 2003; Comai 2005; Van de Peer et al. 2017). Nonetheless, polyploidy is considerably more frequent in plants than in animals, which is hypothesised to be related to the instability caused by the presence of multiple sex chromosome pairs (Orr 1990; Comai 2005; Van de Peer et al. 2017).

In animals, polyploidy seems to be more common among ectotherms. Polyploidy has played an especially important role in the evolution of amphibians and fish. In these lineages it has been linked to periods of environmental fluctuations. The propensity for genome duplications is thought to be greatly affected by temperature, and production of unreduced gametes with temperature shock has been demonstrated in experimental studies (Leggatt and Iwama 2003; Van de Peer et al. 2017).

The identification of genes that arose due to ancient WGDs using comparative genomics and molecular evolution models have revealed that, during the course of evolution the vertebrate lineage underwent two genome duplication events. The teleost lineage is defined by an additional WGD and contains taxons which arose from further WGD events (Glasauer and Neuhauss 2014; Lien et al. 2016; Varadharajan et al. 2018).

Polyploidy is the state of having more than two sets of homologous chromosomes (Van de Peer et al. 2017). Compared to smaller scale duplication events, whole genome duplications are much more drastic and result in either extinction of the lineage or gradual re-diploidization through non-homologous recombination, deletions and pseudogenization (Levasseur and Pontarotti 2011).

While polyploidy can provide an adaptive advantage in the short term (heterosis), it seems unclear what causes it to be established in the long term. One hypothesis suggests that genome duplications become fixed during global catastrophic events, like mass extinctions, or times of environmental instability (Van de Peer et al. 2017). This idea is supported by the fact that ancient genome duplication occurrences tend to cluster around times of mass extinction, glaciation periods and dramatic changes in the environment. At these unstable times, polyploids are thought to be able to outcompete their diploid ancestors due to their increased adaptive potential (Van de Peer et al. 2017). Furthermore, environmental fluctuations and stress have been demonstrated to cause the production of unreduced gametes in many organisms (Orr 1990; Leggatt and Iwama 2003; Van de Peer et al. 2017). Interestingly, analysis of ancient WGD events in plants and animals showed that the increase in speciation rates after duplication is not immediate and can be delayed by several million years, and that many of the duplicated genes shortly return to their singleton status resulting in the gradual rediploidization (Comai 2005; Van de Peer et al. 2017).

1.5 Lophiiformes

Anglerfishes comprise the morphologically diverse teleost order Lophiiformes with at least 321 living species. Because of their lifestyle, hard-to-reach habitat, and the lack of knowledge for many taxa, new members of the order are being described relatively frequently (Shedlock et al. 2004; Pietsch et al. 2009; Miya et al. 2010; Ho et al. 2013; Pietsch and Sutton 2015; Ho 2016; Ho and Ma 2016; Rajeeshkumar et al. 2017; Betancur-R et al. 2017; Arnold and Pietsch 2018).

The order is divided into 5 suborders and 18 families:

1. Lophioidei or goosefishes (1 family) (Caruso and Bullis Jr 1976; Caruso and Suttkus 1979; Caruso 1981, 1983, 1985; Shedlock et al. 2004; Miya et al. 2010; Betancur-R et al. 2017)

2. Antennarioidei or frogfishes (4 families) (Pietsch 1981; Shedlock et al. 2004; Pietsch et al. 2009; Last and Gledhill 2009; Miya et al. 2010; Arnold and Pietsch 2012; Betancur-R et al. 2017)

3. Chaunacoidei or toadfishes (1 family) (Caruso 1989; Shedlock et al. 2004; Miya et al. 2010; Betancur-R et al. 2017)

4. Ogcocephaloidei or batfishes (1 family) (Ochiai and Mitani 1956; Bradbury 1967; Miya et al. 2010; Ho et al. 2013; Betancur-R et al. 2017)

5. Ceratioidei or seadevils, which contains 11 families and almost half of the total number of anglerfish species (Pietsch and Orr 2007; Pietsch 2009; Miya et al. 2010; Betancur-R et al. 2017)

The first four suborders are shallow/deep water benthic ambush predators. Ceratioidei, on the other hand, are meso/bathypelagic and abyssal-benthic dwellers (Pietsch and Orr 2007; Pietsch 2009; Miya et al. 2010).

Anglerfishes possess a number of notable adaptations, but perhaps one of their most recognisable characteristics is their luring apparatus or illicium which is derived from the first dorsal fin spine and located above the snout in most anglerfish species (Pietsch and

Orr 2007; Pietsch 2009; Miya et al. 2010). In Lophioidei and Antennarioidei, illicial movements can range from a simple flicking to complex patterns mimicking prey. Compared to some Chaunacoidei and Ogcocephaloidei members, and most Ceratioidei females illicium is relatively immobile. In these taxa, the lure is retractable by a movement of the illicial pterygiophore (base of the dorsal fin) (Schultz 1957; Bradbury 1967; Pietsch and Grobecker 1978; Thangstad 2006; Pietsch 2009). The illicium itself usually ends with a fleshy outgrowth or esca which serves as bait. The esca can appear as a simple outgrowth of skin, like in *Lophius piscatorius*, or can include complex structures to aid with luring. The Ceratioidei esca, in most species, includes a photophore - an organ filled with bioluminescent bacteria (Hulet and Musil 1968; O'day 1974; Pietsch and Orr 2007; Pietsch 2009). In some batfish species the esca is known to release a chemical compound which attracts marine gastropods (Nagareda and Shenker 2009).

Another unique anglerfish feature - extreme sexual dimorphism coupled with male attachment is restricted to the deep-water Ceratioidei suborder (Regan 1925; Pietsch 1976, 2005, 2009; Pietsch and Orr 2007; Miya et al. 2010; Vieira et al. 2013). In these species, male bodies are only a fraction of the female's bodyweight and size. For example, males of *Ceratias holboelli* can be 1/60 of the length and half a million times lighter than a female (Regan 1925; Pietsch 1976, 2005, 2009; Quigley et al. 2005; Pietsch and Orr 2007). In some *Linophrynidae* species, males are only 6-10 mm in length and are one of the smallest sexually mature vertebrates known. Dwarfed ceratioid males lack the illicium, though most have well-developed eyes and or enlarged nostrils to find females (Pietsch 1976, 2005, 2009; Pietsch and Orr 2007).

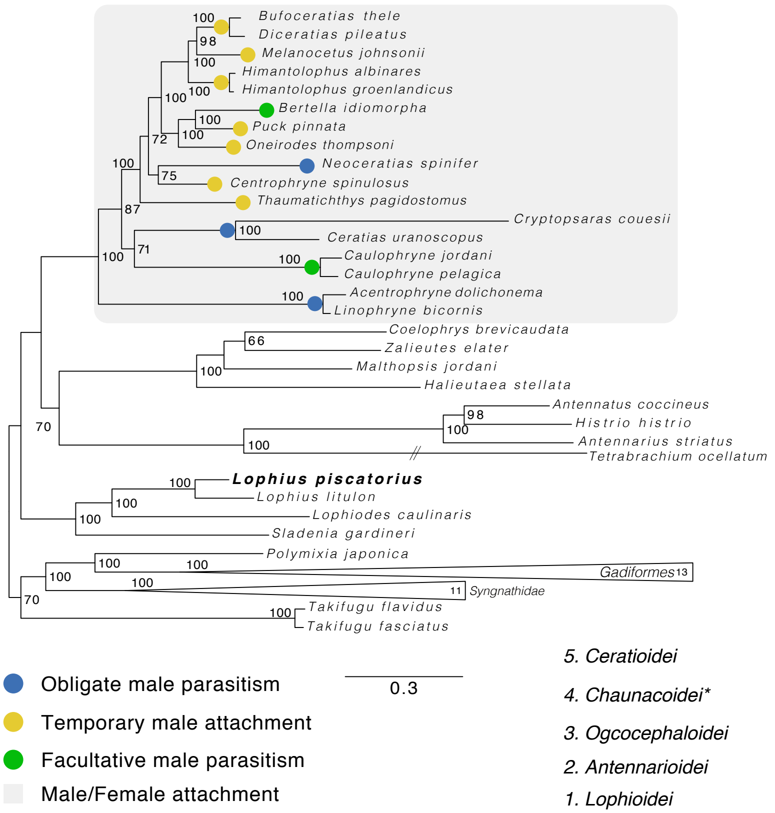


Figure 6. Phylogenetic tree of species belonging to the Lophiiformes and Gadiformes orders, the Syngnathidae family, the Takifugu genus and Polymixia japonica

The phylogenetic relationships were inferred from complete mitochondrial genome sequences. The scale indicates the number of substitutions per site. The *Tetrabrachium ocellatum* branch length has been halved due to its extreme length. Node support values are bootstrap probabilities based on 500 iterations. Phylogenetic relationships were inferred using a partitioned maximum likelihood analysis (with first, second and third codon positions, rRNA and tRNA as partitions) and a GTR GAMMA model as implemented in RaxML. Clades that exhibit male-female attachment behaviour are marked by grey box. The degree of male attachment is marked by coloured circles: Blue – males are obligate parasites, Yellow – males attach temporary, Green – males are facultative parasites. *Chaunacoidei was not included in the analysis.

In the Lophiiformes order, male attachment represents a spectrum (figure 6) (Pietsch 1976, 2005, 2009; Pietsch and Orr 2007; Miya et al. 2010). In the first four suborders, Lophioidei, Antennarioidei, Chaunacoidei and Ogcocephaloidei, males never attach to females. Male attachment is found only in the Ceratioidei suborder and can be either temporary, or permanent with fusion of two individuals (Pietsch 1976, 2005, 2009; Pietsch and Orr 2007). The classification of male attachment into temporary attachment, facultative and obligate male parasitism was shaped into a concise hypothesis by Pietsch (Pietsch 1976, 2005) (table 1).

Table 1. Male attachment types.

Non-parasitic attachment	Facultative parasitism	Obligatory parasitism
Male attachment seems rare and never involves fusion	Male attachment is rare If happens, always involves fusion	Male attachment is common If attached, are always fused
Gravid females observed without attached males	Gravid females observed with/without males attached	Gravid females never observed without male attached
Sexually mature free-living males are common	Sexually mature free-living males were observed	Free-living males always have underdeveloped gonads
Free-living males are able to feed	Free-living males are able to feed	Free-living male guts always found empty

*adapted from (Pietsch 2009)

In the taxa where males are considered to be obligate parasites (*Ceratiidae* and *Linophrynidae*), free-living males have modified pincer-like jaws that seem unsuitable for capturing prey and their gut is underdeveloped and is always found empty, suggesting a relatively short free-living stage. In these species, male attachment is common and always results in fusion. Reproductive organs in non-parasitized females and free-living males are always underdeveloped. Gravid females always have males attached to them (up to 8 males in some species) (Pietsch and Orr 2007), which suggests that male-female fusion is required for sexual maturation. According to Pietsch (2005): "Sexually mature or gravid individuals are assumed to be those whose gonads are obviously larger than those of other conspecific individuals of a similar standard length". Attached parasitic males often have reduced eyes, nostrils, digestive tract, sometimes obstructed gills and appear to be fully dependant on the female (Pietsch 1976, 2005, 2009; Pietsch and Orr 2007). Histological studies of the male/female attachment zone have found vascular plexuses where male and female circulatory systems join, although this is based on a preserved sample and more data is needed to confirm the actual blood exchange (Munk 2000). Interestingly, the proportion of parasitized females in species with obligatory parasitism is quite small, ranging from 6.2%-40% of all collected females. This suggests that the majority of individuals are not participating in reproduction at any given time of the year (Pietsch 2005).

Male parasitism is classified as facultative when free-living individuals of both sexes can be found with well-developed reproductive organs. In this case male attachment is rare, though it always results in fusion. Free-living male jaws appear to be suitable for feeding and their guts were found to contain food. Gravid females can be found with or without males attached (Pietsch 1976, 2005, 2009; Pietsch and Orr 2007).

In some ceratioid taxa, male attachment is non-parasitic and temporary. In such cases male attachment seems rare and if found attached, males are never fused with a female. Free-living males and females are found with well-developed reproductive organs. Free living males are able to feed, as their guts were found to contain food. Gravid females can be observed without male attachment. The distribution of male sexual parasitism within the Ceratioidei suborder, when overlaid on a phylogenetic tree, appears patchy with obligatory/facultative parasitism and temporary attachment scattered throughout the tree without apparent structure. The general consensus is that male parasitism independently evolved up to 7 times throughout the suborder (Pietsch 1976, 2005, 2009; Pietsch and Orr 2007; Miya et al. 2010).

Not unique to anglerfishes, but rare among fish in general, is the ability to "walk". The most notable walkers among anglerfishes are members of the Antennarioidei (frogfishes) and Ogcocephaloidei (batfishes) suborders. Their pectoral fins are often described as "limb-like" or "hand-like" which allow them (assisted by pelvic fins) to either jump in a manner similar to mudskippers, or "walk" on all four of their fins similar to tetrapods (Arnold and Pietsch 2012; Ho et al. 2013; Dickson and Pierce 2019). Both batfishes and frogfishes prefer to "walk" on the bottom and are ineffective swimmers, but compared to frogfishes, batfishes can swim in rapid bursts over short distances (Arnold and Pietsch 2012; Ho et al. 2013; Dickson and Pierce 2019).

1.6 *Lophius piscatorius*

The Lophioidei suborder represents a basal clade in the anglerfish taxonomy. It consists of one family - *Lophiidae*, 4 genera and includes 25 living species (Caruso and Bullis Jr 1976; Caruso and Suttkus 1979; Caruso 1981, 1983, 1985; Miya et al. 2010; Betancur-R et al. 2017).

Lophius piscatorius is a dorso-ventrally flattened bathydemersal fish, which can be found along the European continental shelf, from Gibraltar (including the Mediterranean and the Black Sea) to Barents Sea, Faroe Islands and Iceland. The *L. piscatorius* distribution range overlaps with a sister species - *L. budegassa* (Thangstad 2006; Farina et al. 2008). The species are very similar in appearance but are easy to separate by *L. budegassa's* dark coloration around the mouth and body wall (Thangstad 2006; Farina et al. 2008). In general, *Lophius* species prefer muddy and gravelly, sometimes rocky bottoms and depending on geographical location, they can be found at depths between 30 to 2600 m. Smaller, younger individuals tend to prefer shallower waters and large individuals move to the deep (Hislop et al. 2000; Piñeiro et al. 2001; Thangstad 2006; Farina et al. 2008).

It is known for *L. piscatorius* to perform vertical migrations, sometimes dramatic, changing depth from 118 to 20 m and then returning back to the bottom (Hislop et al. 2000; Thangstad 2006; Farina et al. 2008). As it lacks a gas bladder and apparently uses its large liver for vertical migrations instead, *L. piscatorius* appears to be resilient to sudden depth change (Thangstad 2006; Farina et al. 2008). Though uncommon, some mature individuals have been observed in the pelagic water layers of the North-east Atlantic (Hislop et al. 2000). Growth rate estimations in Norwegian waters report rates of ~11.5 cm/year before maturity and 8.4 cm after. Landa et al. (2008) suggests that growth rates for *L. piscatorius* are underestimated and reports overall growth rates from

~15 cm to ~6-7 cm, decreasing with age and length. The largest individuals caught were mostly females (Thangstad 2006; Farina et al. 2008).

L. piscatorius individuals reach maturity at 4-6 years. In Norway, mature fish are usually around 40-80 cm in length and 3-6 kg in weight. Males are smaller and tend reach maturity earlier. The primary means of estimating age in *Lophius* species are growth ring counts of saggital ottoliths and the illicium. Illicium growth rings are often more pronounced and easier to count (Thangstad 2006; Farina et al. 2008; Cañás et al. 2012). Depending on the geographical location, the spawning period lasts from late winter through summer. Eggs are released as one continuous gelatinous ribbon, light red to purple in colour, which can be longer than 10 m and up to 1 m wide. Spawning occurs in deep waters, close to the seabed. Hatching occurs as the egg ribbons rise to the surface. *Lophius* larvae use surface currents and can drift up to 120 days before they settle. As the fish matures, it moves to deeper waters (Thangstad 2006; Farina et al. 2008). *Lophius* members are opportunistic feeders and display low prey selectivity. Their diet consists of various bony marine fish, cephalopods and crustaceans. They are also known to consume seabirds (Thangstad 2006; Issac et al. 2017). Members of Lophioidei suborder are considered a delicacy (Farina et al. 2008).

1.7 Evolution of immune systems

Since the emergence of life in a form of replicator molecules, living systems had to deal with "cheaters" that would exploit common goods without producing anything in return. Both mathematical modelling of such simple replicator systems and in-vitro experiments show the inevitability of selfish element appearance (Mills et al. 1967; Szathmáry and Demeter 1987; Takeuchi and Hogeweg 2007; Koonin 2016; Iranzo et al. 2016). Without the means to protect themselves such systems would have been doomed, being eventually overturned by selfish elements. This means that the infamous evolutionary "arms race" was already present at a very early history of life.

The signs of past "battles" can be found in the genomes of species across all kingdoms:

1. Nearly all species host various selfish genetic elements both in their nuclear and, to lesser extent, organellar genomes (Bao et al. 2015; Koonin 2016; Iranzo et al. 2016)
2. The components of some very important systems are hypothesised to have selfish element origins. These include, but are not limited to:
    - Prp8 of the spliceosome - the largest protein of the spliceosomal ribonucleoprotein complex (responsible for mRNA processing) may have retro-element origins (Dlakic and Mushegian 2011)
    - Telomerase - another ribonucleoprotein required for the replication of linear chromosomes evolved from a non-LTR retrotransposon (de Lange 2015; Podlevsky and Chen 2016)
    - Recombination-activating genes (RAG1-RAG2 complex) of the vertebrate adaptive immune system, responsible for diversity of the T cell and B cell receptors, evolved from the Transib transposon (Koonin and Krupovic 2014)
    - CRISPR-Cas - the adaptive immune system of microorganisms have selfish element ancestry as well (Koonin and Krupovic 2014)

1.7.1 Immune needs and strategies

As life gradually increases complexity, selection also moves with it. For example, with the origin of multicellularity there will be a new and higher level of selection acting on an ensemble of cells as a whole, and not at the individual cell level. Reversion to the lower level (individual) of selection is detrimental to the higher (organism/cell aggregate) level (e.g. cancer) (Koonin 2016). Hence hosts have acquired a broad range of defensive strategies, often at a very high metabolic cost in order to maintain the organism's integrity and keep pathogens under control (Rimer et al. 2014; Iranzo et al.

2015). It also means that depending on the level of organisation, the life mode and the ecological niche, the immune needs and strategies vary extensively (Rimer et al. 2014).

These immune strategies can be roughly divided into three groups (Rimer et al. 2014; Koonin and Krupovic 2014; Brubaker et al. 2015; Iranzo et al. 2015):

1. Innate immunity - based on the recognition of the most conserved pathogen features with recognition molecules encoded in the genome
2. Adaptive immunity - highly efficient and pathogen-specific, with recognition molecules assembled through various somatic processes
3. Suicidal systems - cause programmed cell death in order to prevent pathogens from spreading

None of these is necessarily more ancient, nor is competing in its usefulness to the organism. These systems seem to differ not only in the underlying mechanisms of how they work but also functionally, complementing one another.

Innate immune mechanisms vary highly between species. But the key difference between innate and adaptive immune systems is that innate mechanisms utilise components that are encoded in the germ-line genome, cover broad range of pathogens and serve as an organism's immediate response to environmental challenges, whereas adaptive immune systems are assembled through somatic processes and refined by selection which allows them to be highly specific (Du Pasquier 2001; Cooper and Alder 2006; Rimer et al. 2014; Brubaker et al. 2015; Iranzo et al. 2015).

Innate immune strategies of most eukaryotes involve the use of antimicrobial peptides and Pattern Recognition Receptors (PPRs) that are pre-encoded in genome. These receptor proteins are used to detect either molecules characteristic for wide spectrum of pathogens (pathogen-associated molecular patterns or PAMPs) or the damage induced by these pathogens (damage-associated molecular patterns or DAMPs) (Seong and Matzinger 2004; Rubartelli and Lotze 2007; Rimer et al. 2014; Brubaker et al. 2015). In addition to the antimicrobial molecules and PPRs, animals have populations of specialised immune cells which employ various mechanisms to eliminate foreign or aberrant cells (mostly through phagocytosis) (Du Pasquier 2001). Jawed vertebrates (gnathostomes) also possess an additional lineage of cells which exist in-between the adaptive and innate immune systems - the natural killer cells (Du Pasquier 2001; Cooper and Alder 2006; Paust et al. 2010; Yoder and Litman 2011; Rimer et al. 2014). These are tightly interlinked with the gnathostome adaptive immune system and help to get rid of virus-infected or malfunctioning cells (Du Pasquier 2001; Paust et al. 2010; Yoder and Litman 2011; Rimer et al. 2014).

It is well accepted by now that probably every living organism has some sort of innate immune defence. Adaptive or acquired immune system on the other hand were believed to be exclusive to gnathostomes (jawed vertebrates) (Thompson 1995; Marchalonis et al. 1998; Du Pasquier 2001; Flajnik 2018). It seems unlikely, considering how big of an advantage adaptive immunity gives, that most life forms would rely only on innate immune components. Like jawed vertebrates, agnathans, microorganisms, plants, invertebrates have their own immune needs. They are faced with environmental changes and constantly evolving pathogens, against which they have to protect themselves and maintain organismal integrity.

1.7.2 Acquired/adaptive immune mechanisms

The adaptive/acquired immune system is highly specific and its final components are not pre-encoded in the genome. Instead it achieves its specificity through either somatic recombination processes or by utilising the pathogen's own DNA or RNA as template guiding molecules to target this same pathogen afterwards (Cooper and Alder 2006; Flajnik and Kasahara 2010; Rimer et al. 2014). Adaptive immune systems often include mechanisms that maintain the ability to produce specific recognition molecules that have been found useful in prior defence reactions. This is referred to as an, 'immunological memory' and is often stated as a requirement for an immune system to considered adaptive.

Though it is debated whether to call RNA-guided RNA interference (RNAi) an actual adaptive immune system, it is widespread among eukaryotes and serves not only as a defence system but also as a way to regulate gene expression. The disagreement regarding classification of RNAi as an adaptive or innate immune system stems from the definition of the adaptive immune system itself. Some consider RNAi an adaptive immune system (Voinnet 2001; Bergstrom and Antia 2006; Rimer et al. 2014) because it is able to mount a response specific to a particular pathogen using components not encoded in the genome (specificity/adaptability argument). Other authors consider RNAi an innate immune system due to the lack of immune memory (Obbard et al. 2009; Koonin and Krupovic 2014), and others again call it a "semi-adaptive innate immune defence" due to its specificity, but lack of memory (Zambon et al. 2006). Plants, and most invertebrates (e.g. nematodes and arthropods), rely heavily on RNAi as a protection against transposons and viruses (Voinnet 2001; Zamore 2002; Baulcombe 2004; Lu et al. 2005; Wang et al. 2006; Zambon et al. 2006; Obbard et al. 2009; Rechavi et al. 2011; Rimer et al. 2014).

1.7.3 Gnathostome adaptive immunity

An adaptive immune system that acquires its diversity through recombination of genome segments was thought to have evolved first in gnathostomes. This was a major advance that allowed for the potential recognition of almost any possible pathogen that might be encountered followed by an immunological memory (Thompson 1995; Matsunaga and Rahman 1998; Marchalonis et al. 1998; Du Pasquier 2001; Litman et al. 2010; Flajnik 2018).

Although this system increased in complexity with gnathostome evolution, it is thought that the acquisition of all required cellular processes, tissues and genes happened relatively quickly as most components are present across all jawed vertebrates. T-cell receptors (TCR), B-cell receptors (BCR), and the Major Histocompatibility Complex (MHC) classes I and II, are all present across gnathostome lineages, from the Chondrichthyes to the bony aquatic and terrestrial vertebrates (Du Pasquier 2001; Litman et al. 2010; Flajnik 2018). The core peptide-binding region sequences of the MHC I and II are conserved even in sharks. Like in terrestrial vertebrates, shark MHC genes appear to be highly polymorphic, which in addition to the sequence conservation, suggests that both the antigen-presenting function and the selective pressure for polymorphism have stayed the same from the throughout the evolution of the gnathostomes (Okamura et al. 1997; Kurosawa and Hashimoto 1997). Although the specific sites of haematopoiesis vary, homologous tissues and organs including the thymus and spleen are also present across the gnathostomes (Orkin and Zon 2008; Neely and Flajnik 2016; Flajnik 2018).

The mechanisms behind the gnathostome immune system are well studied. There are two major lineages of adaptive immune cells - T lymphocytes and B lymphocytes. T - stands for thymus derived and B - for bone marrow derived (or bursa). During development, lymphocyte progenitor cells generate unique T cell or B cell receptor (TCRs and BCRs) antigen binding domains by rearranging variable (V), diversity (D), and joining (J) genome segments. Nucleotides can be added to the joints during the assembly, further increasing diversity. In addition, Activation-induced cytidine deaminase (AID) in the activated B-cells can add point mutations to already assembled immunoglobulin genes increasing variance and potentially producing higher affinity antibodies. This process is called somatic hypermutation. While B cell receptors recognise exposed parts (epitopes) of intact molecules, T cell receptors require antigen processing first, and recognise epitopes presented to them in a digested form by antigen presenting cells on top of the Major Histocompatibility Complex (MHC) molecules. Antigen binding triggers intracellular signalling pathways required for a proper immune response (Cooper and Alder 2006; Litman et al. 2010; Flajnik 2018) (figure 7).

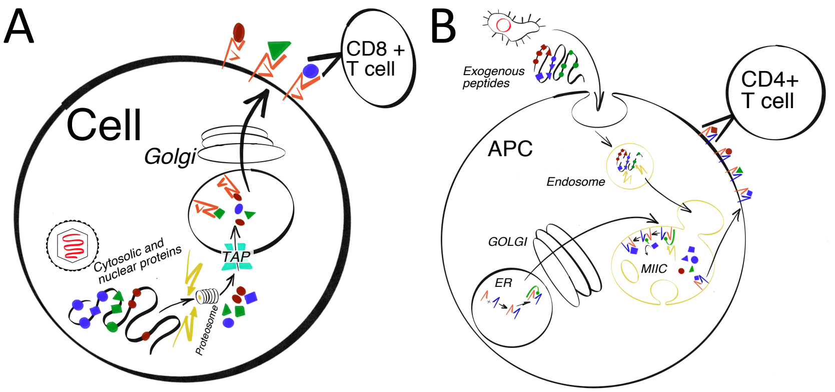


Figure 7. The MHC I and MHC II presentation pathways

Both the MHC I (A) and MHC II (B) pathways are required for the presentation of processed antigens to T-cells. In the MHC I pathway, intracellular peptides are degraded in proteasomes and the resulting peptides are transported into the endoplasmic reticulum (ER) where they are loaded onto MHC I molecules which are then transported to the plasma membrane and presented to CD8+ T cells. The MHC I complex is active in all cell types. In the MHC II pathway, exogenous peptides are imported into the cell by phagocytosis and degraded in endosomes. MHC II molecules (α and β chains) are assembled in the endoplasmic reticulum (ER) where the peptide binding groove is blocked by CD74 (invariant chain). The complex is then transported into the MHC class II compartment (MIIC) via the golgi. Here the CD74 chain is degraded freeing the peptide binding groove, and MHC II is loaded with a processed antigen. The loaded MHC II molecules are then exported to the plasma membrane where they present antigens to CD4 + T cells. TAP: transporter associated with antigen processing, APC: antigen-¬presenting cell, ER: endoplasmic reticulum, MIIC - the MHC class II compartment. Figure adapted from Neefjes et al. 2011

Interestingly, the enzyme complex that is responsible for the BCR and TCR combinatorial diversity - recombination activation genes (RAG) 1 and 2 originated from the Transib family of animal transposons. Cleavage mechanism and segment joining during V(D)J recombination are reminiscent of transposition reactions. Target site duplications generated by RAG1/2 and Transib transposons, terminal inverted repeats

(TIRs) of transposons and recombination signal sequences (RSSs) of RAG complex share high similarity as well. This could serve as an evidence that the RAG-dependent adaptive immune system originated as a result of an insertion of DNA transposon into an immunoglobulin or a TCR V region-like gene (Fugmann et al. 2006; Koonin and Krupovic 2014; Kapitonov and Koonin 2015).

Despite the previous belief that only jawed vertebrates possess highly sophisticated adaptive immunity, studies done in jawless fish show that this is not true (Uinuk-ool et al. 2002). Lampreys and hagfishes are the only living representatives of jawless fish (agnathans). Agnathans were found to possess cells that are morphologically reminiscent of gnathostome lymphocytes. Furthermore, many lymphocyte-specific genes were discovered in lamprey and hagfish as well, and they seem to be expressed. Despite that, key elements of the conventional immune system - rearranging TCRs and BCRs, RAG1/RAG2 complex and MHC, are lacking. Instead, upon activation, lamprey lymphoblasts were found to express leucine-rich-repeat (LRR) proteins that were highly variable in their amino acid sequence. The diversity of these variable lymphocyte receptors (VLRs) is also generated by recombination of LRR segments (Uinuk-ool et al. 2002; Pancer et al. 2004; Cooper and Alder 2006; Rast and Buckley 2013; Flajnik 2018).

1.7.4 Diversity of fish immune systems

Whole genome and transcriptome studies done in marine fish have revealed that Atlantic cod, pipefish, elephant shark, and coelacanth possess alternative adaptive immune strategies. Cod and pipefish genomes lack important genes coding for the parts of the adaptive immune system including MHC II, CD4, CD74 (only cod), and CIITA (only pipefish) (Star et al. 2011; Haase et al. 2013). Whereas the elephant shark has no T-helper cells nor CD4 receptor genes, and the coelacanth genome lacks genes coding for IgM (Amemiya et al. 2013; Venkatesh et al. 2014). A 2016 study, based on low-coverage genome sequences (9-39 times) of 66 teleost species, uncovered that not only cod, but the entire Gadiformes order has lost MHC II and related pathway genes, and that some gadiform species seem to have compensated for the loss with MHC I expansion

(Malmstrøm et al. 2016). Additionally, the cod genome shows expansion of Toll-like receptor genes (TLR) (Solbakken et al. 2016). Despite the lack of one arm of the classical adaptive immune pathway, these fish do not seem to show increased vulnerability to bacterial infections and parasites. However, it is still unknown what caused these adaptations or what exactly the alternative immune pathways are and how they function. Although, most of the adaptive immune components are conserved among jawed vertebrates, these reports question our views on the permanence of their immune systems.

1.7.5 Major histocompatibility complex in fish

Due to its combinatorial diversity, almost any antigen can be recognised by the gnathostome adaptive immune system. While BCRs can detect a plethora of various epitopes in their native form, TCRs require the antigen to be digested and presented to T-cells in an MHC-bound form (Murphy and Weaver 2017).

In fish, the MHC was first described in carp in 1990 (Hashimoto et al. 1990; Grimholt 2016), and like in any other vertebrates it is divided into three classes - MHC class I is expressed on most cells of the body and helps to protect an organism against viral infections and aberrant cells; MHC class II is expressed by the antigen-presenting cells (dendritic cells, macrophages, and B cells) and helps to protect against extracellular threats. MHC class III is comprised of other immune genes like heat shock proteins and tumor necrosis factor (Neefjes et al. 2011; Dijkstra et al. 2013; Grimholt et al. 2015; Grimholt 2016; Wilson 2017).

The two major classes are divided into classical and non-classical molecules. For MHC class I molecule to be defined as classical, it must bind peptides, be expressed on various tissues and cell types, and it must have many alleles per locus (Grimholt et al. 2015; Grimholt 2016; Wilson 2017). The classical MHC class I molecule consists of alpha and beta-microglobulin chains, where alpha 1 and alpha 2 domains of the alpha chain make up the peptide binding groove. They present endogenous peptides to CD8+ T cells (Grimholt et al. 2015; Grimholt 2016; Wilson 2017).

For MHC class II molecules to be defined as classical, they should bind peptides, be highly polymorphic and be expressed on antigen presenting cells. They consist of alpha and beta chains, where each is made up of two domains. The peptide binding groove is composed of the alpha 1 and beta 1 domains and presents exogenous peptides to CD4+ T cells (Dijkstra et al. 2013; Grimholt 2016; Wilson 2017).

Some non-classical molecules either have an unusual or no peptide binding groove, and therefore they cannot bind peptides. Some have very specific expression patterns or are present in a low copy numbers in the genome (Dijkstra et al. 2013; Grimholt et al. 2015; Grimholt 2016). It is hypothesised that these non-classical molecules either perform different immune-related functions (e.g. binding non-peptide ligands, assist peptide loading, or interaction with other cellular receptors), or have been repurposed for completely different tasks (Grimholt 2016). In humans there are other molecules that share alpha 1 and 2 domains with the classical MHC I molecules, and at the same time perform different functions. To make this distinction between classical and non-classical molecules species must be studied in great detail and very few of the fish species are (Dijkstra et al. 2013; Grimholt et al. 2015; Grimholt 2016; Wilson 2017).

In fish, 5 MHC I and 3 MHC II lineages have been identified: MHC I - U, Z, S, L, P; and MHC II - A, B and E. To date, all of the teleost classical MHC I genes belong to the U lineage. U lineage ancient alpha 1 and alpha 2 domain lineages are shared between distantly related species. The MHC I Z lineage though being non-classical, seem to be present in all studied species and displays very high conservation of the binding domain throughout the gnathostome lineage (Grimholt et al. 2015; Grimholt 2016). In teleosts, MHC class II classical molecules all belong to the A lineage and seem to be present in all species (with few exceptions) (Dijkstra et al. 2013; Grimholt 2016).

2 Aims

The overall goal of this thesis was to produce a high quality nuclear and organellar genome sequence, assembly and annotation of a local anglerfish species *Lophius piscatorius*. This to act as a resource for molecular, phylogenetic and population studies, and to encourage further research on genome evolution in this teleost clade.

As a focus point we chose to characterise *L. piscatorius* adaptive immune system gene repertoire, in particular the MHC I and II pathways.

Objectives:

1. Investigate features of the *L. piscatorius* mitochondrial genome and corresponding transcriptome (Papers I and II)
2. Characterize features of the *L. piscatorius* adaptive immune system in comparison to other teleosts with a focus on the MHC pathway genes (Paper III)
3. Produce a high quality draft genome assembly and annotation of the *L. piscatorius* nuclear genome (Paper IV)

3 Main results

Molecular features of the *L. piscatorius* mitochondrial genome and corresponding transcriptome

Objective 1. Contributions from Papers I and II

In this study, presented as two papers, we characterised the mitochondrial genome (mitogenome) and transcriptome of *L. piscatorius*.

In Paper I we confirm the standard vertebrate organisation of *L. piscatorius* mitogenome. By aligning the newly sequenced mitogenome to related *Lophius* species we show that *L. piscatorius* (Europe) and *L. americanus* (North America) can be distinguished from *L. litulon* (Asia) by a 40 bp indel in the control region. Subsequent phylogenetic analysis showed that *L. litulon* was indeed the most diverged of the *Lophius* species included in this analysis. *Spladenia gardineri* was placed at the base of the Lophioidei suborder followed by *Lophoides caulinaris*, *Lophiomus setigerus*, and finally the *Lophius* species, which is in concordance with previous reports.

The high depth of sequencing (2227 times coverage), combined with the use of the highly accurate SOLiD sequencing approach, allowed us to investigate low-level heteroplasmic features of the *L. piscatorius* mitogenome. We identified seven heteroplasmic sites in total: one in the COIII gene that resulted in an amino acid change, four in the LSU rRNA gene, and two in the control region.

In Paper II we sequenced and analysed the mitogenome and transcriptome of another *L. piscatorius* individual. We identified nine polymorphic sites in the mitogenome from the two analysed individuals, seven of which fell in the open reading frames of protein coding genes; 4 were non-synonymous. All canonical genes were expressed with the cytochrome oxidase transcripts being the most abundant followed by NADH dehydrogenase and ATPase subunit transcripts (Paper II, Figure 1 B). Such differences in abundance are likely to be related to variance in mRNA stability. Consistent with previous reports in fish, we noted that most mRNAs contained no or very short untranslated regions (UTRs). One exception is the 75 bp 3'UTR of the COI transcript, which we hypothesise may contribute to mRNA stability. Like in other species, most mitochondrial transcripts (coding and non-coding) were found polyadenylated. Among the non-canonical transcripts, we detected two previously described (in Atlantic cod and human) long noncoding RNAs (lncRNAs) in the control region. We report a 178 nt antisense RNA to the 5' region of the COI gene that represents a previously undescribed long non-coding RNA in vertebrate mitochondria (lncCOI). Finally, we noted sequence conservation of one of the mitochondrial-derived peptides - Humanin - across the Lophioidei members, zebrafish, gadiform fishes and mammals, and speculate on its regulatory function in mitochondria.

Characterization of the *L. piscatorius* adaptive immune system with a focus on the MHC pathways

Objective 2. Contributions from Paper III

Some anglerfish species exhibit a unique mode of reproduction - male sexual parasitism. This adaptation involves male-female attachment and sometimes results in a fusion of two or more individuals. Due to the patchy distribution of male-female attachment within the anglerfish order, it has been suggested that this adaptation arose independently multiple times during the course of anglerfish evolution. Hence, we hypothesise that there is a common predisposition shared among all anglerfish. Considering the unique reproductive strategy within the anglerfish clade, and its implications for the immune system, we decided to focus on the *L. piscatorius* adaptive immune system.

As a starting point we chose to investigate the presence of the MHC I and II pathways genes due to their role in immune rejection and recent reports of MHC II pathway loss in two teleost clades. We first sequenced and assembled genomes of two *L. piscatorius* individuals. In order for our study to be comparable with a recently published immune pathway investigation of 66 teleost genomes, we used the same set of immune system genes and followed the described methodology with only a few changes.

We were unable to identify genes coding for MHC II α and β, CD74 A/B and CD4 in *L. piscatorius* - the same set of immune genes lost in within the Gadiformes order and, with the exception of CD74, in the members *Syngnathus* genus. We hypothesise that in teleosts these genes (+/- CD74) represent an independent evolutionary module and have no external essential function outside of the MHC II pathway. Being a third reported taxon that lost MHC II, our finding corroborates the dispensability of this pathway in teleosts.

L. piscatorius belongs to the most basal suborder in the anglerfish taxonomy. Interestingly, a prior study of 66 teleost genomes included another anglerfish, *Antennarius striatus*, that belongs to the adjacent suborder Antennarioidei. This species appears to have an intact adaptive immune system. If the current taxonomy is correct, MHC II loss is likely to be restricted to the *Lophius* genus or Lophioidei suborder. However, by conducting a phylogenetic analysis based on the complete mitogenome sequences and including/excluding certain outgroup clades we observed a shift in two branches placing either the Lophioidei or Antennarioidei suborders at the base of the anglerfish tree. Thus, our observations suggest interesting questions regarding the accuracy of the current taxonomy and its implications for evolution of sexual parasitism.

Genome assembly and annotation

Objective 3. Contributions from Paper IV

The Paper IV manuscript discusses features of the chromosome-level assembly of *L. piscatorius* and its preliminary annotation. The first part describes the general assembly statistics and some challenges that we encountered on the way. In brief, we found that approximately 90% of the genome is contained within chromosome-level scaffolds, while the remaining 10% is contained within contigs of ~160 kb or less. By performing a gene-space completeness analysis we found 95% of the conserved actinopterygii orthologues to be complete within our assembly. Further examination of the assembly revealed the presence of sequences microsporidian in origin, which are most likely from *Spraguea lophii* an intracellular parasite infecting specifically members of the *Lophius* genus.

We discovered that ~75% of the non-*Lophius* sequences are located within the unscaffolded contigs, most of which were less than 10kb long, and probably explaining why they were not incorporated into scaffolds. The filtering of such reads is difficult even with the parasite genome available, due to the fact that the parasite genome itself can be contaminated with host sequences. In fact, we found that many of the chromosomal loci that matched with the parasite genome, can also be aligned with a 75% identity to other teleost species.

In order to produce the preliminary annotation of the genome we used the MAKER 2 pipeline. As one of the measures of annotation quality we compared the distribution of gene, exon/intron lengths, and the exon number per gene to a selection of teleost species, spotted gar (*Lepisosteus oculatus*) and the ascidian *Ciona intestinalis*. We found that the exon lengths and the exon number per gene to be conserved for all examined species. Interestingly, we noted that the intron length in teleosts follows a bimodal distribution (short and long introns), compared to the unimodal distribution in gar and *C. intestinalis*. We hypothesise that the smaller intron peak could be a consequence of the same processes that resulted in small genome sizes of many teleost species.

A global synteny analysis showed that both chromosomal gene content and gene order were conserved among the examined teleosts (with the exception of *Danio rerio*) and showed the closest similarity of *L. piscatorius* to *Takifugu rubripes*. This verified the quality of our assembly and annotation and supported the current phylogenetic position of the Lophiiformes order.

4 General discussion

The primary goal of this work has been to produce a high quality annotated assembly of the *L. piscatorius* genome, which we present in our final paper (Paper IV). This assembly is both the first chromosome level, and the first annotated assembly of any anglerfish genome and we believe that it will serve as a useful resource for the further analysis of the many anglerfish specific adaptations and behaviours. During the construction of this final assembly we made use of the intermediate data to perform analyses pertaining to:

- The mitochondrial genome organisation and activity
- The gene repertoire of the adaptive immune system of *L. piscatorius* with an emphasis on the Major Histocompatibility Complex (MHC)

The analyses of mitochondrial sequences (Papers I and II) confirm that *L. piscatorius* contains a standard vertebrate mitochondrial genome. In addition, we analysed the transcriptional landscape and identified both known and unknown non-canonical transcripts. This demonstrates that although the mitochondrial genome is small and has been extensively studied, that there are still aspects of mitochondrial function that have been overlooked and that are coming to light through the use of high-throughput sequencing.

In Paper III we were able to confirm the absence of the MHC II arm of the adaptive immune system in *L. piscatorius*, making this the third teleost taxon where MHC II has been reported to be lost. This is interesting in its own right since both MHC I and MHC II are otherwise highly conserved across the jawed vertebrates (gnathostomes) and suggests something special about the dispensability of MHC II in teleosts. It is particularly interesting to see such an immune modification in anglerfish due to the presence of sexual parasitism in a number of anglerfish species.

Finally, we performed a number of analyses on our chromosome level assembly (Paper IV). These confirm the general correctness and completeness of our assembly and annotation. In addition, these analyses revealed a general property of teleost intron lengths which may be related to the observation that teleosts genomes are generally small and compact.

4.1 Mitochondrial genome and transcriptome (Papers I and II)

We confirmed that *L. piscatorius* has a typical vertebrate mitogenome organisation and content. As expected, during our preliminary phylogenetic analysis (Paper I) we confirmed that *L. piscatorius* groups with other members of *Lophius* genus. Interestingly, we noted that there is a 40 bp indel that distinguishes *L. piscatorius* and *L. americanus* from *L. litulon*. Though vertebrates generally have a conserved mitogenome gene order and high sequence conservation, such sequence variation, especially in the non-coding parts of the genome, are not uncommon (Satoh et al. 2016). Within the Lophiiformes both variation in the gene order (*Tetrabrachium ocellatum*, *Ceratias uranoscopus*, *Cryptopsaras couesii*) and insertions of long (>100 bp) non-coding intergenic sequences (*Caulophrynidae*, *Melanocetidae*, *Oneirodidae*, *Gigantactinidae*, *Linophrynidae*) have been observed (Miya et al. 2010).

Mitochondria are present in multiple copies within cells, and have variation in the sequence within one tissue, between different tissues of the same individual, and between individuals of the same species (Ameur et al. 2011; Emblem et al. 2012, 2014; Wallace and Chalkia 2013; Hedberg et al. 2019; Jørgensen et al. 2019). In *L. piscatorius* we identified seven low-level heteroplasmic sites within one individual (Paper I) and nine between the two individuals (Paper 2). Many mitochondrial SNPs in humans have been linked to disease, with some "dangerous variants" passed down the generations within one family (Stefano et al. 2017; Hedberg et al. 2019; Jørgensen et al. 2019).

Using RNA seq data we identified several mitochondrial long non-coding RNAs (lncRNA) in *L. piscatorius*. Two of these have been previously described and are transcribed from the Control Region (CR) origin (lncCR-L and lncCR-H) (Jørgensen et al. 2014, 2019). A further one has not previously been described and is transcribed from the antisense Cytochrome Oxidase I (COI) gene region and hence we refer to this as lncCOI. The function of these transcripts is currently unknown.

Finally, we decided to look into the sequence conservation of the humanin-like mitochondrial-derived peptide in anglerfish and compare it to other vertebrates. Most research on Humanin function, not surprisingly, has been done in humans and rats (Guo et al. 2003; Lee et al. 2013; Paharkova et al. 2015; Zárate et al. 2019). Though it was first described in 2001 (Hashimoto et al. 2001) there are still important questions to be asked about its cellular role. Studies in mammals indicate that Humanin is a circulating signal molecule involved in metabolism, apoptosis and stress resistance processes (Lee et al. 2013; Paharkova et al. 2015). Studies in birds and teleost fish indicate that Humanin is conserved in most, but not all, species (Jørgensen and Johansen 2018; Mortz et al. 2019). In this work, we confirm the presence of an Humanin open reading frame in *L. piscatorius*, suggesting that the Humanin sequence is conserved across most teleosts. Such high degree of sequence conservation throughout the vertebrate lineage indicates a common function of Humanin.

4.2 Loss of MHC II in *L. piscatorius* (Paper III)

Our initial interest in the *L. piscatorius* immune system was based on the disjunct distribution pattern of male parasitism within anglerfishes; which, according to Pietsch (2007) was first noted by Bertelsen et al. (1951). Pietsch then further suggested that male parasitism originated independently multiple times within the order (Pietsch 1976, 2005; Pietsch and Orr 2007). The multiple independent origin idea is supported by modern phylogenetic studies (Miya et al. 2010). Since male parasitism can include the fusion between male and female, which would normally result in an immune rejection, it seems likely that species which exhibit this mode of reproduction also possess a modified immune system. We hypothesized that this modification is shared across the anglerfishes because multiple independent origins of sexual parasitism imply a shared predisposition. Such modifications may include both the loss of genes and modified expression patterns. Hence, *Lophius piscatorius* - a local anglerfish which belongs to the most basal anglerfish suborder Lophioidei - seemed like a good choice to test this. We considered that in common with Atlantic cod (*Gadus morhua*) (Star et al. 2011) and pipefish (*Syngnathus typhle*) (Haase et al. 2013), the anglerfishes may have lost the MHC II arm of the adaptive immune system. Indeed, we were unable to identify MHC II orthologues in a preliminary analysis of a fragmented assembly based on low coverage sequencing data, supporting our suspicion of MHC II loss in *L. piscatorius*.

A study by Malmstrøm et al. (2016), released after this project was started, demonstrated the absence of the same immune pathway components in 27 cod-like species, suggesting that MHC II loss is shared by all Gadiformes members. This study also included the draft genome assembly of *Antennarius striatus* - an anglerfish belonging to the neighbouring clade of the Lophioidei clade. Unfortunately for our initial hypothesis, this study reported that *A. striatus* apparently contains an intact MHC II pathway. This observation could have three possible explanations:

1. Our initial hypothesis was wrong and MHC II loss is restricted to the *Lophius* genus or the Lophioidei suborder. MHC II loss in *L. piscatorius* has no relation to sexual parasitism

2. The current phylogeny is wrong and Antennarioidei, not Lophioidei is the basal anglerfish clade

3. *A. striatus* was sequenced as part of a high-throughput genome analysis project and it is possible that there are problems with the *A. striatus* sequences reported resulting in an inappropriate identification of MHC II genes

To test this, we devised the study which became the third paper of this thesis.

4.2.1 Confirming absence of MHC II pathway genes

Initially we chose to look for the presence of the same set of adaptive immune system genes as in Malmstrøm et al. (2016). This set includes a range of genes from both the MHC I and MHC II arms of the adaptive immune system. Hence it provides a set of positive control genes that can be used to determine how well we can identify genes. Using the same set of genes also makes it easy to compare the immune gene repertoire of *L. piscatorius* with the species included in Malmstrøm et al. (2016). Later we expanded the gene set to also include a collection of classical and non-classical MHC II sequences from (Dijkstra et al. 2013).

To identify the selected genes within the assemblies we first used an approach similar to that previously described (Malmstrøm et al. 2016), combining BLASTp with gene prediction and then using additional verification with reciprocal BLAST. However, we felt that the use of common 1e-10 threshold was not objective enough and instead developed our own scoring system based on BLAST bit scores and alignment lengths (Paper III) combined with a visualisation of the blast data that allowed us to select ORFs for further examination. That way we could potentially separate the real orthologues, from gene fragments and homologous but non-orthologous sequences. For most genes we could easily identify clear orthologues, but some did not quite fit our expectations. For these we manually examined the hits and discovered that the gene prediction software (Genscan) sometimes outputs chimeric sequences, combining immune gene protein with the adjacent unrelated sequence.

As a result, if one would just simply use the reciprocal BLAST scores or e-values as the only confirmation parameter, without manually examination, the genes would appear missing. We noted that often for these fusion ORFs, all reported BLAST hits would belong to the longest fusion partner, with no alignments to the shorter ORF. Whenever we saw such alignment patterns, with a large part of the ORF unidentified, we used the alignment coordinates to select the unaligned part and BLAST it against UniProt separately. With this approach we were able to identify all but seven genes in the dataset. Five belonged to the MHC II pathway (MHC II α and β, CD74 a/b and CD4) and the other two were ERAP2 and SEC61G.

The ERAP2 gene is a member of a family of similar genes and because of this it was difficult to identify a single gene as the ERAP2 orthologue. For statistical purposes we thus considered it missing. However, we believe that a more thorough approach would be able to identify the specific orthologue or orthologues.

SEC61G is notably reported as missing from a number of species by Malmstrøm et al. (2016) and it represents a special case. The gene prediction software failed to identify an ORF for this gene and we had to examine the BLAST output to find it. SEC61G is a highly conserved gene, however, part of its sequence consists of a low-complexity region. Scores arising from alignments to this region are adjusted by a low-complexity filter that is active by default in BLAST. This adjustment leads BLAST to not report alignments to this region of SEC61G and leads to the gene prediction software to missing SEC61G. However, using BLAST manually against our assembly, either with a very permissive e-value threshold or with the low-complexity filter deactivated, we could easily identify a SEC61G orthologue in *L. piscatorius*.

This problem is even clearer in the pipefish transcriptome investigation by Haase et al. (2013). In their study CD74 is reported as "non-functional" because it is missing 20 amino acids from the 3'-end. The seahorse (*Hippocampus comes*) seems to be missing the same part of the CD74 sequence but has the complete set of other MHC II pathway genes. Thus, it is difficult to say whether a gene truncation has resulted in a loss of function and this requires in-depth investigation of its sequence.

Synteny analysis is often used as an additional verification of gene loss. While this might be useful for genes with a highly conserved synteny, the MHC region in teleosts appears to be unlinked and lacking synteny (Grimholt 2016). Even if the gene would be missing from its conserved position, this can't serve as an evidence for its loss, especially considering how common the intra-chromosomal rearrangements are in teleosts.

4.2.2 Contamination-issue in the *A. striatus* data

The presence of MHC II in a related anglerfish species argues against a role of MHC II loss in the evolution of sexual parasitism. Hence, we wanted to test (Paper III) the possibility that there was a problem with the species identity of the *A. striatus* samples used in Malmstrøm et al. (2016). The previously published *A. striatus* mitogenome sequence (Miya et al. 2010) was compared to the sequences from Malmstrøm et al. (2016). A phylogenetic analysis of the dominant mitochondrial sequences showed that species used in the Malmstrøm-study was indeed *A. striatus*. However, in addition to the *A. striatus* mitogenome, we also found several mitochondrial sequences from a distantly related fish species suggesting the possibility of a misidentification of MHC II in *A. striatus*.

If the MHC II sequences reported by Malmstrøm et al. (2016) had been derived from a contamination the sequencing depth of the contigs containing these genes should be markedly lower than expected sequencing depth. We compared the coverages of the *A. striatus* mitogenome, contaminating mitogenomes, the MHC II-containing contigs and the overall coverage distribution. The results clearly show that MHC II pathway genes in *A. striatus* are not the result of cross-contamination, which confirm the conclusions in Malmstrøm et al. (2016).

4.2.3 State of current anglerfish phylogeny

While current morphology-based and molecular-based phylogeny (Miya et al. 2010) agree on the basal position of the Lophioidei suborder and the association of anglerfishes with Tetraodontiformes, we decided to re-investigate the phylogeny of the anglerfishes with the inclusion of our mitochondrial genome sequences.

We performed two separate phylogeny reconstructions using the same substitution models and alignment sets, but with the inclusion or exclusion of *Takifugu* species as an outgroup. With *Takifugu* included our results matched the conventional phylogeny. However, if we exclude this group from the analysis the positions of Antennarioidei and Lophioidei switch, making Antennarioidei the basal clade (figure 8). The basal position of Antennarioidei is required for the Lophiodei MHC II loss to be shared with species exhibiting sexual parasitism, and our analysis suggests that the currently accepted phylogeny should not be considered certain and that it is possible that the loss of MHC II occurred in a common ancestor of *Lophius* and the clades with sexual parasitism.

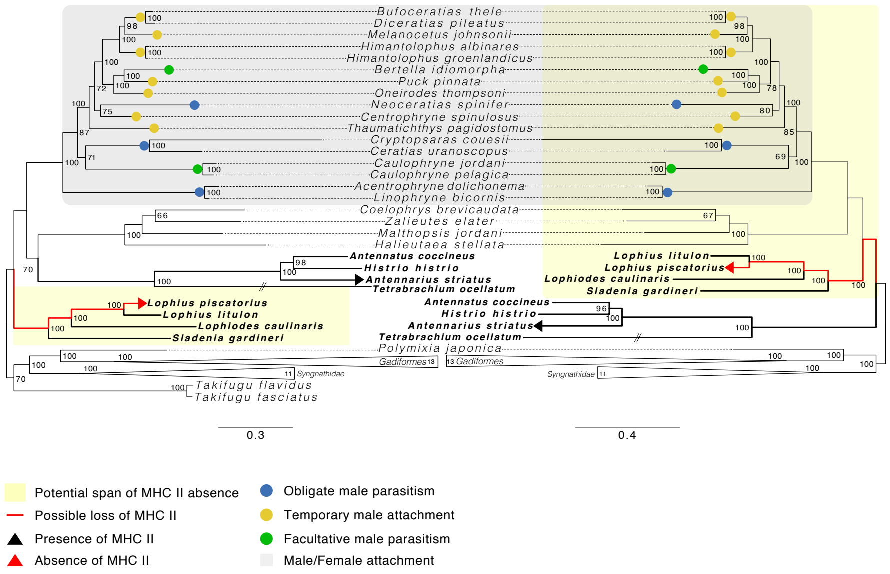


Figure 8. Sexual parasitism and MHC II loss in the Lophiiformes order

Two alternate Lophiiformes phylogenies produced using different outgroups. Right, the conventional tree that includes *Takifugu* species as an outgroup; left, an alternative tree which does not include *Takifugu* species. Species where the presence or absence of MHC II has been demonstrated are indicated by black (present) or red (absent) triangles. Since MHC II is present in *A. striatus*, the loss of MHC II observed in *L. piscatorius* must have happened after the divergence of their most common ancestor. The potential period in which this loss can have occurred is marked in red in the tree, and the set of species which may thus share this loss are indicated by a yellow background. Sexually parasitic species that exhibit male-female attachment are indicated in the figure. The conventional phylogeny excludes the possibility that sexually parasitic species share the MHC II loss observed in *L. piscatorius*. An absence of MHC II in these species would thus argue for the alternative phylogeny shown on the right. The scale indicates the number of substitutions per site. The *Tetrabrachium ocellatum* branch length has been halved due to its extreme length. Node support values are bootstrap probabilities based on 500 iterations. Phylogenetic relationships were inferred using a partitioned maximum likelihood analysis (with first, second and third codon positions, rRNA and tRNA as partitions) and a GTR GAMMA model as implemented in RaxML.

4.2.4 Why fish?

It seems that of all vertebrate genomes sequenced so far only those from a few cartilaginous and teleost fish lack important components of the adaptive immune system. The elephant shark is reported to lack CD4 and related transcription factors, but to retain polymorphic MHC II genes (Venkatesh et al. 2014). *Syngnathus typhle*, the entire Gadiformes order, and now *L. piscatorius* all appear to have a non-functional MHC II pathway (Star et al. 2011; Haase et al. 2013; Malmstrøm et al. 2016; Small et al. 2016) (Paper III). In addition, all teleosts appear to lack important accessory molecules involved in MHC II peptide loading (MHC II DM) (Dijkstra et al. 2013), and MHC I and II are unlinked in teleosts and lack defined synteny (Grimholt 2016; Wilson 2017).

Furthermore, fish in general appear to rely more on the innate immune system, with their adaptive immune systems being described as "sluggish" and greatly depend on temperature (Bly and William Clem 1991; Alcorn et al. 2002; Magnadóttir 2006; Bowden et al. 2007; Star and Jentoft 2012; Makrinos and Bowden 2016). Perhaps fish represent an early step in the evolution of the adaptive immune system with many of the cellular pathways that link the adaptive and innate and adaptive systems unformed. This could explain how loss of one arm of the adaptive immune system can occur without affecting the other.

What this does not explain is a) why the loss of MHC II is apparently restricted to only three teleost taxa, and b) why loss of MHC I has not been observed. One of the possible explanations for the latter is that Natural Killer (NK) cells - components of the innate immune system, at least in mammals, interact and become activated by MHC I molecules (Yoder and Litman 2011; Fischer et al. 2013). Homologues of mammalian NK receptors that interact with MHC I have been described in cichlid fishes (Fischer et al. 2013). Therefore, if NK cell function in fish is already linked with MHC I, it would be hard to lose one without affecting the other.

It is difficult to say what the common selective pressure that could result in MHC II loss in the described taxa could be. Selective pressure on the immune system would depend not only on the general characteristics of the environment (cold, warm, etc.)

and abundance of parasites, but also on factors like ecological niche of the species in question, life strategy, population structure, parasite interactions, parental care, or mate choice strategies (Schulenburg et al. 2009; Hedrick 2017).

4.2.5 Ceratioidei reproductive strategies

Interestingly, the type of reproductive strategy where two individuals are required to fuse is among the vertebrates only found in anlgerfishes. A similar adaptation is known in members of the parasitic flat worm family *Diplozoidae* (which is unique among invertebrates) (Zurawski et al. 2003a, b; Pečínková et al. 2007; Avenant-Oldewage and Milne 2014). In these worms, two larvae called diporpa join together into a single organism. This triggers metamorphosis and sexual maturation, like in some Ceratioidei species. The joining of two genetically distinct *Diplozoidae* individuals involves the fusion of the nervous reproductive and digestive systems as well as the musculature (Zurawski et al. 2003a, b; Pečínková et al. 2007; Avenant-Oldewage and Milne 2014). However, fusion of two and sometimes more vertebrate individuals is more than puzzling, because of the higher order of organisation (tissues, organs, blood and lymphatic vessels) and sophisticated adaptive immune system (though little is known about Platyhelminthes immune system). It remains unknown how fused ceratioid males and females avoid immune rejection. However, the loss of MHC II constitutes a major immune system modification and it is possible that this has played a role in enabling the appearance of sexual parasitism in the anglerfishes.

4.3 Genome assembly and annotation (Paper IV)

In Paper IV we describe a chromosome level assembly and annotation of the *L. piscatorius* genome. We found that 90% of the final assembly was contained within 23 chromosome sized scaffolds. The remaining 10% of sequences appeared to be derived from either repetitive regions that were difficult to assemble or from an intracellular parasite that is commonly found within *L. piscatorius* cells. We annotated this assembly using a de-novo gene prediction followed by reciprocal BLAST to a set of well annotated teleost genomes. This identified around 20,000 orthologue candidates and allowed us to assign protein family identifiers to these genes. The numbers of family members were highly consistent with other teleost species arguing that we have identified the majority of protein coding genes in *L. piscatorius*.

As part of the validation of our genome we compared feature properties to that of other vertebrate genomes, confirming *L. piscatorius* to have a typical teleost genome in size and gene repertoire. In addition, these analyses revealed a fundamental property of teleost intron size that is likely to be related to the evolution of genome size in teleosts.

4.3.1 Assembly validation by genome annotation

Our preliminary annotation with a single MAKER 2 (Holt and Yandell 2011) run identified 45,552 candidate genes. After filtration by reciprocal BLAST this number was reduced approximately by half. Taking into account the average number of protein coding genes for vertebrate genome is ~20,000, this correlates well with the 96% completeness score we got from BUSCO.

We also considered the overall genome composition in terms of gene features (exons, introns and genes) and found that it was similar to the teleost genomes we used a comparison. We also found similar distributions of gene feature size and numbers in all the teleosts we looked at. These observations argue for the validity of our assembly and annotation process.

Since the chromosomal gene content and order is well conserved across teleosts it is also possible to validate assemblies by considering the global synteny of orthologous genes. We found that individual chromosomes of *L. piscatorius* could be mapped to orthologous chromosomes in all of the teleosts analysed. In addition, with the exception of *D. rerio* (zebrafish), we could also observe a conservation of gene order across large chromosomal regions. This serves as a strong argument that the scaffolds reported here correspond to physical chromosomes.

4.3.2 A teleost specific intron length distribution

Interestingly we observed a bimodal distribution of spliceosomal intron lengths in *L. piscatorius*. A similar distribution was found in all teleost species that we used as a comparison. However, in the two non-teleost species that we included in the analysis we observed a uni-modal distribution. Although the intron length distribution in *D. rerio* has previously been reported as bimodal (Moss et al. 2011), the same report describes the other teleost species inspected as having a uni-modal distribution.

The distribution of intron size is a fundamental property of any genome and it is surprising that our observations have not previously been reported. It is notable that we have considered the distribution of log transformed intron lengths, rather than the linear length distribution. Log transformation of values is appropriate when the underlying variance is likely to be affected by exponential processes; that is processes where the rate (or probability) of change is a function of the magnitude. That is, longer introns are more likely to change in size than shorter ones. Given that introns cannot shrink beyond a certain minimal size (50-100 bp), and the fact that processes that result in intron size increases are more likely to occur within larger introns we argue that intron size should be considered in log-space and that the observed bimodality is related to the rules that govern intron length.

We also observed a correlation between the intron length distribution and genome size within teleosts, with larger genomes having a smaller proportion of introns in the shorter peak. Although teleosts are descended from an ancestor that, compared to most vertebrates, underwent an additional genome duplication event they have in general smaller genome sizes compared to both mammal and avian species. This suggests both an evolutionary pressure and mechanism for genome size reduction in teleosts. Since teleost genomes appear to be characterised by large numbers of very short introns it suggests that this may be related to genome reduction. It is notable that we do not observe a similar distribution in *L. oculatus* (spotted gar). This further suggests that teleost genome sizes have arisen not simply as a result of evolutionary pressures but also because of a teleost specific mechanism for genome size reduction.

4.3.3 Traces of genome duplications

The most obvious difference in genome composition between the teleost and non-teleost species analysed here was found in the distributions of intron lengths. This may indeed be related to processes occurring as a result of the teleost specific genome duplication. We could see clear traces of these events when comparing the numbers of one-to-one orthologues between species (Paper IV, Fig. 6) with both *Ciona intestinalis* (a non-vertabrate chordate) and *L. oculatus* being smaller than that for the teleosts. In contrast these species both had larger numbers of one-to-many orthologues with *L. piscatorius* as would be expected after a genome duplication event.

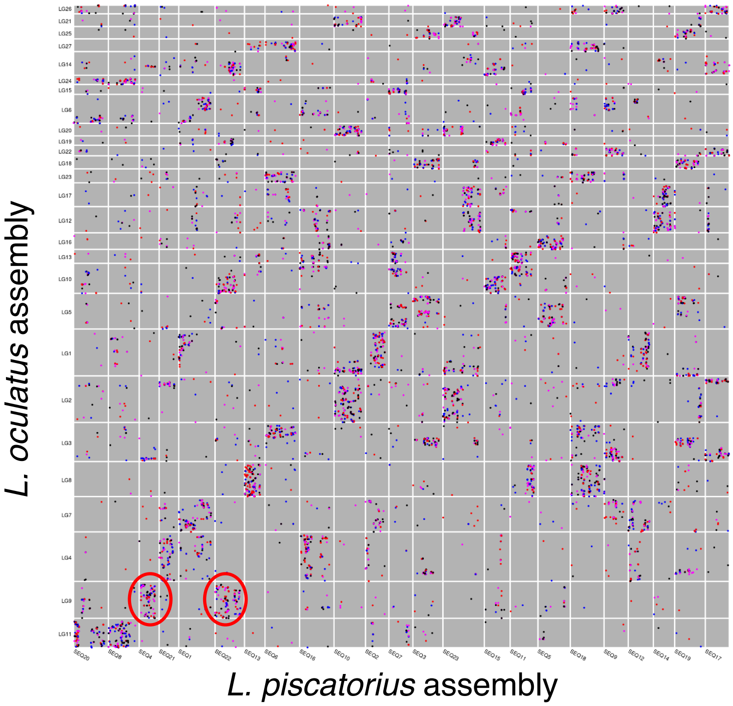


Figure 9. Orthology mapping between L. oculatus and L. piscatorius.

Many of the *L. oculatus* orthologues map to two *L. piscatorius* chromosomes representing remnants of a historical duplication event that happened in the common ancestor of the teleost lineage which includes Lophiiformes. Example of such duplicate mapping is marked with two red ellipses. Here *L. oculatus* orthologues from LG9 are mapping to SEQ4 and SEQ22 of *L. piscatorius*.

In general, we could easily identify orthologous chromosomes (chromosomes that contain the same set of genes) having a one-to-one relationship when comparing the gene locations of *L. piscatorius* with other teleosts. We were able to find a similar but weaker chromosomal orthology between *L. piscatorius* and *L. oculatus*, but here most chromosomes in *L. oculatus* mapped to two separate chromosomes in *L. piscatorius* (figure 9). This relationship is a clear indication of the historical teleost genome duplication and demonstrates how genome wide properties can reveal ancient events.

4.3.4 Spurious sequences in genome assemblies

During the genome annotation process, it became apparent that some of the annotated genes within our assembly belonged to a microsporidian parasite. Sequence comparisons confirmed that the presence of these genes was caused by the presence of *Spraguea lophii* in our samples. *S. lophii* is an intracellular microsporidian parasite which, as suggested by its name, specifically infects members of the *Lophius* genus. Somewhat fortunately, the whole genome of this parasite has been sequenced, and we were able to make use of this sequence to identify contigs containing parasite sequences. These analyses demonstrated that the vast majority of the parasite sequence had not been included within the chromosomal sized scaffolds, and hence, that the scaffolding process used here can, at least in favourable circumstances, exclude a large part of contaminating sequences from the final assembly. Nevertheless, we were able to find approximately 20,000 loci within our chromosomal scaffolds that could be aligned with a close to 100% identity to parasite contigs. However, these mapped to only 5 of the parasite contigs and represent a very small part of parasite assembly. Interestingly, one of these sequences (which could be found at around 19,000 chromosomal loci) could also be aligned to multiple loci within a number of other teleost species suggesting that this sequence may actually be a *L. piscatorius* sequence that has been mistakenly incorporated into the *S. lophii* assembly.

The observation of contaminating sequences appears as a trend across our work. Not only did we observe contaminating mitochondrial sequences in the *A. striatus* assembly in Paper III, but here (Paper IV) we observe what appears to be reciprocal contamination between *L. piscatorius* and *S. lophii*. Parasitic infections and the presence of closely associated symbiotic organisms are common across the animal kingdom and it is likely that a large number of published assemblies contain sequences from such contaminating organisms.

One of the common ways to filter contaminating reads prior to the assembly is to use GC content (Frazier et al. 2017; Karimi et al. 2018; Urbarova et al. 2019). However, we worried that with this approach we could over-deplete certain sequences in the host genome due to an overlap in the GC content. We found that by using a simple coverage threshold we were able to remove most of the parasite reads from the assembly. This approach is probably suitable only for projects that can sacrifice some sequencing data.

Another possible problem with filtering intracellular parasites or symbionts is that, even though the parasite genome assembly may be available, it itself can be "contaminated" with the host DNA as seems to appear to be the case for the *S. lophii* assembly. Hence care needs to be taken, both in identifying potential contamination and in efforts to eliminate such contamination.

5 Conclusions and future perspectives

Anglerfishes possess many remarkable adaptations, which are interesting not only from an evolutionary point of view but also for the understanding of the gnathostome immune system. The chromosome-level genome assembly and annotation presented in this thesis will greatly facilitate future anglerfish research, thus furthering the understanding of the evolution and phylogenetic relationships of this clade of teleost fish.

To resolve the mechanisms that have allowed sexual parasitism to evolve will require the sequencing of sexually parasitic species, but the *L. piscatorius* genome and annotation will provide a useful baseline genome for the Lophiiformes. Our assembly should also aid the assembly and annotation of further anglerfish species as the chromosome-level continuity can be used to guide the assembly process. The future comparison of this genome with genomes from species representing different levels of sexual parasitism offers fascinating prospect of identifying the genetic traits that may have enabled obligate parasitism.

The absence of MHC II pathway genes in *L. piscatorius*, even if shared with sexually parasitic species, offers only a hint at the potential mechanisms underlying the absence of immune rejection after allogenic fusion. The mechanisms behind allogenic rejection are complex and involve multiple components and pathways, and simply removing one arm of the MHC system may both be not sufficient to block immune rejection and at the same time have deleterious effects for the immune system. Clearly, it also is unlikely that it is possibly to directly translate what happens in anglerfish to other species including humans. However, the information of how the adaptive immune system can be modulated in the anglerfishes may help us to better understand the way in which the immune system can be modified without loss of function.

The availability of the *L. piscatorius* genome, and in the future additional anglerfish genomes will facilitate the identification of sequence variants that can be used to study the population structures of these clades. The presence of sexual parasitism suggests that finding mating partners is problematic which implies small effective populations containing little genetic variance, which in itself may have played a part in the development of allogenic immune tolerance. Population genetics can also be used to study how populations are affected by changing climate and other human influences. This is important given the increased interest for *Lophius* species and the development of specialized fisheries (Thangstad 2006; Farina et al. 2008).

6 References

Alcorn SW, Murray AL, Pascho RJ (2002) Effects of rearing temperature on immune functions in sockeye salmon (*Oncorhynchus nerka*). Fish Shellfish Immunol. 12:303–334. doi: 10.1006/fsim.2001.0373

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, et al (2013) The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311–316. doi: 10.1038/nature12027

Ameur A, Stewart JB, Freyer C, Hagström E, Ingman M, Larsson NG, et al (2011) Ultra-Deep Sequencing of Mouse Mitochondrial DNA: Mutational Patterns and Their Origins. PLoS Genet. 7:e1002028. doi: 10.1371/journal.pgen.1002028

Arnold RJ, Pietsch TW (2018) Fantastic Beasts and Where to Find Them: A New Species of the Frogfish Genus *Histiophryne* Gill (Lophiiformes: *Antennariidae*: *Histiophryninae*) from Western and South Australia, with a Revised Key to Congeners. Copeia 106:622–631. doi: 10.1643/CI-18-112

Arnold RJ, Pietsch TW (2012) Evolutionary history of frogfishes (Teleostei: Lophiiformes: *Antennariidae*): A molecular approach. Mol. Phylogenetics Evol. 62:117–129. doi: 10.1016/j.ympev.2011.09.012

Avenant-Oldewage A, Milne S (2014) Aspects of the morphology of the juvenile life stages of *Paradiplozoon ichthyoxanthon* Avenant-Oldewage, 2013 (Monogenea: *Diplozoidae*). Acta Parasitologica 59:247–254. doi: 10.2478/s11686-014-0235-1

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6:11. doi: 10.1186/s13100-015-0041-9

Barroso Lima NC, Prosdocimi F (2018) The heavy strand dilemma of vertebrate mitochondria on genome sequencing age: number of encoded genes or G + T content? Mitochondrial DNA Part A 29:300–302. doi: 10.1080/24701394.2016.1275603

Baulcombe D (2004) RNA silencing in plants. Nature 431:356–363

Bergstrom CT, Antia R (2006) How do adaptive immune systems control pathogens while avoiding autoimmunity? Trends Ecol. Evol. 21:22–28. doi: 10.1016/j.tree.2005.11.008

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. 5: 3657-3667. doi: 10.1038/ncomms4657

Betancur RR, Wiley EO, Arratia G, Acero A, Bailly N, Miya M, et al (2017) Phylogenetic classification of bony fishes. BMC Evol. Biol. 17:162. doi: 10.1186/s12862-017-0958-3

Bly JE, William Clem L (1991) Temperature-mediated processes in teleost immunity: In vitro immunosuppression induced by in vivo low temperature in channel catfish. Veterinary Immunology and Immunopathology 28:365–377. doi: 10.1016/0165-2427(91)90127-X

Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res. 27:1767–1780

Bowden TJ, Thompson KD, Morgan AL, Gratacap RML, Nikoskelainen S (2007) Seasonal variation and the immune response: A fish perspective. Fish & Shellfish Immunology 22:695–706. doi: 10.1016/j.fsi.2006.08.016

Bradbury MG (1967) The Genera of Batfishes (Family *Ogcocephalidae*). Copeia 1967: 399-422. doi: 10.2307/1442130

Bronstein O, Kroh A, Haring E (2018) Mind the gap! The mitochondrial control region and its power as a phylogenetic marker in echinoids. BMC Evol. Biol. 18:80. doi: 10.1186/s12862-018-1198-x

Brown TA (2017) Genomes 4, 4th edn. Garland Science, New York, NY

Brubaker SW, Bonham KS, Zanoni I, Kagan JC (2015) Innate Immune Pattern Recognition: A Cell Biological Perspective. Annu. Rev. Immunol. 33:257–290. doi: 10.1146/annurev-immunol-032414-112240

Cañás L, Stransky C, Schlickeisen J, Sampedro MP, Fariña AC (2012) Use of the otolith shape analysis in stock identification of anglerfish (*Lophius piscatorius*) in the Northeast Atlantic. ICES J. Mar. Sci. 69:250–256. doi: 10.1093/icesjms/fss006

Caruso JH (1981) The Systematics and Distribution of the Lophiid Anglerfishes: I. A Revision of the Genus *Lophiodes* with the Description of Two New Species. Copeia 1981:522-549. doi: 10.2307/1444556

Caruso JH (1983) The Systematics and Distribution of the Lophiid Anglerfishes: II. Revisions of the Genera *Lophiomus* and *Lophius*. Copeia 1983:11-30. doi: 10.2307/1444694

Caruso JH (1985) The Systematics and Distribution of the Lophiid Anglerfishes: III. Intergeneric Relationships. Copeia 1985:870-875. doi: 10.2307/1445235

Caruso JH (1989) Systematics and Distribution of the Atlantic Chaunacid Anglerfishes (Pisces: Lophiiformes). Copeia 1989:153-165. doi: 10.2307/1445616

Caruso JH, Bullis Jr HR (1976) A review of the lophiid angler fish genus *Sladenia* with a description of a new species from the Caribbean Sea. Bull. Mar. Sci. 26:59–64

Caruso JH, Suttkus RD (1979) A new species of lophiid anglerfish from the Western North Atlantic. Bull. Mar. Sci. 29:491–496

Chen S, Krinsky BH, Long M (2013) New genes as drivers of phenotypic evolution. Nat. Rev. Genet. 14:645–660. doi: 10.1038/nrg3521

Comai L (2005) The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6:836–846. doi: 10.1038/nrg1711

Cooper MD, Alder MN (2006) The Evolution of Adaptive Immune Systems. Cell 124:815–822. doi: 10.1016/j.cell.2006.02.001

de Lange T (2015) A loopy view of telomere evolution. Front. Genet. 6:321. doi: 10.3389/fgene.2015.00321

Dickson BV, Pierce SE (2019) How (and why) fins turn into limbs: insights from anglerfish. Earth Environ. Sci. Trans. R. Soc. Edinb. 109:87–103. doi: 10.1017/S1755691018000415

Dijkstra JM, Grimholt U, Leong J, Koop BF, Hashimoto K (2013) Comprehensive analysis of MHC class II genes in teleost fish genomes reveals dispensability of the peptide-loading DM system in a large part of vertebrates. BMC Evol. Biol. 13:260. doi: 10.1186/1471-2148-13-260

Dlakic M, Mushegian A (2011) Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. RNA 17:799–808. doi: 10.1261/rna.2396011

Du Pasquier L (2001) The immune system of invertebrates and vertebrates. Comp. Biochem. Physiol. B 129:1–15. doi: 10.1016/S1096-4959(01)00306-2

Elliott TA, Gregory TR (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos. Trans. Royal Soc. B 370:20140331. doi: 10.1098/rstb.2014.0331

Emblem Å, Karlsen BO, Evertsen J, Miller DJ, Moum T, Johansen SD (2012) Mitogenome polymorphism in a single branch sample revealed by SOLiD deep sequencing of the *Lophelia pertusa* coral genome. Gene 506:344–349. doi: 10.1016/j.gene.2012.06.040

Emblem Å, Okkenhaug S, Weiss ES, Denver DR, Karlsen BO, Moum T (2014) Sea anemones possess dynamic mitogenome structures. Mol. Phylogenetics Evol. 75:184–193. doi: 10.1016/j.ympev.2014.02.016

Farina AC, Azevedo M, Landa J, Duarte R, Sampedro P, Costas G, et al (2008) Lophius in the world: a synthesis on the common features and life strategies. ICES J. Mar. Sci. 65:1272–1280

Fischer U, Koppang EO, Nakanishi T (2013) Teleost T and NK cell immunity. Fish & Shellfish Immunology 35:197–206. doi: 10.1016/j.fsi.2013.04.018

Flajnik MF (2018) A cold-blooded view of adaptive immunity. Nat. Rev. Immunol. 18:438–453. doi: 10.1038/s41577-018-0003-9

Flajnik MF, Kasahara M (2010) Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat. Rev. Genet. 11:47–59. doi: 10.1038/nrg2703

Fookes MC, Hadfield J, Harris S, Parmar S, Unemo M, Jensen JS, et al (2017) *Mycoplasma genitalium*: whole genome sequence analysis, recombination and population structure. BMC Genom. 18:993. doi: 10.1186/s12864-017-4399-6

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al (1995) The Minimal Gene Complement of *Mycoplasma genitalium*. Science 270:397–404. doi: 10.1126/science.270.5235.397

Frazier M, Helmkampf M, Bellinger MR, Geib SM, Takabayashi M (2017) De novo metatranscriptome assembly and coral gene expression profile of *Montipora capitata* with growth anomaly. BMC Genom. 18:710. doi: 10.1186/s12864-017-4090-y

Friz CT (1968) The biochemical composition of the free-living *Amoebae Chaos chaos*, *Amoeba dubia* and *Amoeba proteus*. Comp. Biochem. Physiol. 26:81–90. doi: 10.1016/0010-406X(68)90314-9

Fugmann SD, Messier C, Novack LA, Cameron RA, Rast JP (2006) An ancient evolutionary origin of the Rag1/2 gene locus. Proc. Natl. Acad. Sci. U.S.A. 103:3728–3733

Galtier N, Nabholz B, Glémin S, Hurst GDD (2009) Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18:4541–4550. doi: 10.1111/j.1365-294X.2009.04380.x

Glasauer SMK, Neuhauss SCF (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. Molecular Genetics and Genomics 289:1045–1060. doi: 10.1007/s00438-014-0889-2

Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biological Reviews of the Cambridge Philosophical Society 76:65–101. doi: 10.1017/S1464793100005595

Gregory TR (2005) Synergy between sequence and size in Large-scale genomics. Nat. Rev. Genet. 6:699–708. doi: 10.1038/nrg1674

Greilhuber J (2005) The Origin, Evolution and Proposed Stabilization of the Terms “Genome Size” and “C-Value” to Describe Nuclear DNA Contents. Ann. Bot. 95:255–260. doi: 10.1093/aob/mci019

Grimholt U (2016) MHC and Evolution in Teleosts. Biology 5:6. doi: 10.3390/biology5010006

Grimholt U, Tsukamoto K, Azuma T, Leong J, Koop BF, Dijkstra JM (2015) A comprehensive analysis of teleost MHC class I sequences. BMC Evol. Biol. 15:32. doi: 10.1186/s12862-015-0309-1

Guo B, Zhai D, Cabezas E, Welsh K, Nouraini S, Satterthwait AC, et al (2003) Humanin peptide suppresses apoptosis by interfering with Bax activation. Nature 423:456–461. doi: 10.1038/nature01627

Haase D, Roth O, Kalbe M, Schmiedeskamp G, Scharsack JP, Rosenstiel P, et al (2013) Absence of major histocompatibility complex class II mediated immunity in pipefish, *Syngnathus typhle*: evidence from deep transcriptome sequencing. Biol. Lett. 9:20130044. doi: 10.1098/rsbl.2013.0044

Harrison RG (1989) Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. Trends Ecol. Evol. 4:6–11. doi: 10.1016/0169-5347(89)90006-2

Hashimoto K, Nakanishi T, Kurosawa Y (1990) Isolation of carp genes encoding major histocompatibility complex antigens. Proc. Natl. Acad. Sci. U.S.A. 87:6863–6867. doi: 10.1073/pnas.87.17.6863

Hashimoto Y, Niikura T, Tajima H, Yasukawa T, Sudo H, Ito Y, et al (2001) A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. Proc Natl Acad Sci USA 98:6336–6341. doi: 10.1073/pnas.101133498

Hedberg A, Knutsen E, Løvhaugen AS, Jørgensen TE, Perander M, Johansen SD (2019) Cancer-specific SNPs originate from low-level heteroplasmic variants in human mitochondrial genomes of a matched cell line pair. Mitochondrial DNA Part A 30:82–91. doi: 10.1080/24701394.2018.1461852

Hedrick SM (2017) Understanding Immunity through the Lens of Disease Ecology. Trends Immunol. 38:888–903. doi: 10.1016/j.it.2017.08.001

Hislop JRG, Holst JC, Skagen D (2000) Near-surface captures of post-juvenile anglerfish in the North-east Atlantic-an unsolved mystery. J. Fish Biol. 57:1083–1087. doi: 10.1111/j.1095-8649.2000.tb02214.x

Ho H-C (2016) Records of deep-sea anglerfishes (Lophiiformes: Ceratioidei) from Indonesia, with descriptions of three new species. Zootaxa 4121:267. doi: 10.11646/zootaxa.4121.3.3

Ho H-C, Ma W-C (2016) Revision of southern African species of the anglerfish genus *Chaunax* (Lophiiformes: *Chaunacidae*), with descriptions of three new species. Zootaxa 4144:175. doi: 10.11646/zootaxa.4144.2.2

Ho H-C, Roberts CD, Shao K-T (2013) Revision of batfishes (Lophiiformes: *Ogcocephalidae*) of New Zealand and adjacent waters, with description of two new species of the genus *Malthopsis*. Zootaxa 3626:188–200. doi: 10.11646/zootaxa.3626.1.8

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:. doi: 10.1186/1471-2105-12-491

Hulet WH, Musil G (1968) Intracellular Bacteria in the Light Organ of the Deep Sea Angler Fish, *Melanocetus murrayi*. Copeia 1968:1321-1335. doi: 10.2307/1442019

Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV (2015) Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. BMC Evol. Biol. 15:. doi: 10.1186/s12862-015-0324-2

Iranzo J, Puigbò P, Lobkovsky AE, Wolf YI, Koonin EV (2016) Inevitability of Genetic Parasites. Genome Biol. Evol. 8:2856–2869. doi: 10.1093/gbe/evw193

Issac P, Robert M, Le Bris H, Rault J, Pawlowski L, Kopp D (2017) Investigating feeding ecology of two anglerfish species, *Lophius piscatorius* and *Lophius budegassa* in the Celtic Sea using gut content and isotopic analyses. Food Webs 13:33–37. doi: 10.1016/j.fooweb.2017.08.001

Jiang J, Yu J, Li J, Li P, Fan Z, Niu L, et al (2016) Mitochondrial Genome and Nuclear Markers Provide New Insight into the Evolutionary History of Macaques. PLOS ONE 11:e0154665. doi: 10.1371/journal.pone.0154665

Jørgensen TE, Bakke I, Ursvik A, Andreassen M, Moum T, Johansen SD (2014) An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA. BMC Evol. Biol. 14:182. doi: 10.1186/s12862-014-0182-3

Jørgensen TE, Johansen SD (2018) Expanding the Coding Potential of Vertebrate Mitochondrial Genomes: Lesson Learned from the Atlantic Cod. In: Seligmann H (ed) Mitochondrial DNA - New Insights. InTech

Jørgensen TE, Karlsen BO, Emblem Å, Jakt LM, Nordeide JT, Moum T, et al (2019) A mitochondrial long noncoding RNA in Atlantic cod harbors complex heteroplasmic tandem repeat motifs. Mitochondrial DNA Part A 30:307–311. doi: 10.1080/24701394.2018.1502281

Kai W, Kikuchi K, Tohari S, Chew AK, Tay A, Fujiwara A, et al (2011) Integration of the Genetic Map and Genome Assembly of Fugu Facilitates Insights into Distinct Features of Genome Evolution in Teleosts and Mammals. Genome Biol. Evol. 3:424–442. doi: 10.1093/gbe/evr041

Kapitonov VV, Koonin EV (2015) Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biol. Direct. 10:20. doi: 10.1186/s13062-015-0055-8

Karimi K, Wuitchik DM, Oldach MJ, Vize PD (2018) Distinguishing Species Using GC Contents in Mixed DNA or RNA Sequences. Evolutionary Bioinformatics 14:117693431878886. doi: 10.1177/1176934318788866

Kim BM, Amores A, Kang S, Ahn DH, Kim JH, Kim IC, et al (2019) Antarctic blackfin icefish genome reveals adaptations to extreme environments. Nat. Ecol. Evol. 3:469–478. doi: 10.1038/s41559-019-0812-7

Koonin EV (2016) Viruses and mobile elements as drivers of evolutionary transitions. Philos. Trans. Royal Soc. B 371:20150442. doi: 10.1098/rstb.2015.0442

Koonin EV, Krupovic M (2014) Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat. Rev. Genet. 16:184–192. doi: 10.1038/nrg3859

Kühlbrandt W (2015) Structure and function of mitochondrial membrane protein complexes. BMC Biol. 13:89. doi: 10.1186/s12915-015-0201-x

Kurosawa Y, Hashimoto K (1997) How did the primordial T cell receptor and MHC molecules function initially? Immunol. Cell Biol. 75:193–196. doi: 10.1038/icb.1997.28

Landa J, Duarte R, Quincoces I (2008) Growth of white anglerfish (*Lophius piscatorius*) tagged in the Northeast Atlantic, and a review of age studies on anglerfish. ICES J. Mar. Sci. 65:72–80. doi: 10.1093/icesjms/fsm170

Last PR, Gledhill DC (2009) A revision of the Australian handfishes (Lophiiformes: Brachionichthyidae), with descriptions of three new genera and nine new species. Zootaxa 2252:1–77. doi: 10.11646/zootaxa.2252.1.1

Lee C, Yen K, Cohen P (2013) Humanin: a harbinger of mitochondrial-derived peptides? Trends Endrocrinol. Metab. 24:222–228. doi: 10.1016/j.tem.2013.01.005

Leggatt RA, Iwama GK (2003) Occurrence of polyploidy in the fishes. Rev. Fish Biol. Fisher. 13:237–246. doi: 10.1023/B:RFBF.0000033049.00668.fe

Levasseur A, Pontarotti P (2011) The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. Biol. Direct. 6:11. doi: 10.1186/1745-6150-6-11

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al (2016) The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205. doi: 10.1038/nature17164

Litman GW, Rast JP, Fugmann SD (2010) The origins of vertebrate adaptive immunity. Nat. Rev. Immunol. 10:543–553. doi: 10.1038/nri2807

Logsdon JM, Doolittle WF (1997) Origin of antifreeze protein genes: A cool tale in molecular evolution. Proc. Natl. Acad. Sci. U.S.A. 94:3485–3487. doi: 10.1073/pnas.94.8.3485

Lu R, Maduro M, Li F, Li HW, Broitman-Maduro G, Li WX, et al (2005) Animal virus replication and RNAi-mediated antiviral silencing in Caenorhabditis elegans. Nature 436:1040–1043. doi: 10.1038/nature03870

Magnadóttir B (2006) Innate immunity of fish (overview). Fish & Shellfish Immunology 20:137–151. doi: 10.1016/j.fsi.2004.09.006

Makrinos DL, Bowden TJ (2016) Natural environmental impacts on teleost immune function. Fish & Shellfish Immunology 53:50–57. doi: 10.1016/j.fsi.2016.03.008

Malmstrøm M, Matschiner M, Tørresen OK, Star B, Snipen LG, Hansen TF, et al (2016) Evolution of the immune system influences speciation rates in teleost fishes. Nat. Genet. 48:1204–1210. doi: 10.1038/ng.3645

Marchalonis JJ, Schluter SF, Bernstein RM, Hohman VS (1998) Antibodies of sharks: revolution and evolution. Immunol. Rev. 166:103–122. doi: 10.1111/j.1600-065X.1998.tb01256.x

Matsunaga T, Rahman A (1998) What brought the adaptive immune system to vertebrates?-The jaw hypothesis and the seahorse. Immunol. Rev. 166:177–186

Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc. Natl. Acad. Sci. U.S.A. 58:217–224. doi: 10.1073/pnas.58.1.217

Miya M, Pietsch TW, Orr JW, Arnold RJ, Satoh TP, Shedlock AM, et al (2010) Evolutionary history of anglerfishes (Teleostei: Lophiiformes): a mitogenomic perspective. BMC Evol. Biol. 10:58

Mortz M, Dégletagne C, Romestaing C, Duchamp C (2019) Comparative genomic analysis identifies small open reading frames (sORFs) with peptide-encoding features in avian 16S rDNA. Genomics. doi: 10.1016/j.ygeno.2019.06.026

Moss SP, Joyce DA, Humphries S, Tindall KJ, Lunt DH (2011) Comparative Analysis of Teleost Genome Sequences Reveals an Ancient Intron Size Expansion in the Zebrafish Lineage. Genome Biol. Evol. 3:1187–1196. doi: 10.1093/gbe/evr090

Munk O (2000) Histology of the fusion area between the parasitic male and the female in the deep-sea anglerfish *Neoceratias spinifer* Pappenheim, 1914 (Teleostei, Ceratioidei). Acta Zool. 81:315–324

Murphy KM, Weaver C (2017) Janeway's immunobiology, 9th edition. Garland Science, Taylor & Francis Group, New York London

Nagareda BH, Shenker J (2009) Evidence for chemical luring in the polka-dot batfish *Ogcocephalus cubifrons* (Teleostei: Lophiiformes: *Ogcocephalidae*). Florida Scientist 72:11–17

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15:R59. doi: 10.1186/gb-2014-15-3-r59

Neefjes J, Jongsma MLM, Paul P, Bakke O (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat. Rev. Immunol. 11:823–836. doi: 10.1038/nri3084

Neely HR, Flajnik MF (2016) Emergence and Evolution of Secondary Lymphoid Organs. Annu. Rev. Cell Dev. Biol. 32:693–711. doi: 10.1146/annurev-cellbio-111315-125306

Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AWC, Pippel M, et al (2018) The axolotl genome and the evolution of key tissue formation regulators. Nature 554:50–55. doi: 10.1038/nature25458

Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM (2009) The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. Royal Soc. B 364:99–115. doi: 10.1098/rstb.2008.0168

Ochiai A, Mitani F (1956) A revision of the pediculate fishes of the genus *Malthopsis* found in the waters of Japan (family *Ogcocephalidae*). Pac. Sci. 10:271–285

O'day WT (1974) Bacterial luminescence in the deep-sea anglerfish *Oneirodes acanthias* (Gilbert, 1915). Contributions in science 255:1–12

Okamura K, Ototake M, Nakanishi T, et al (1997) The most primitive vertebrates with jaws possess highly polymorphic MHC class I genes comparable to those of humans. Immunity 7:777–790

Orkin SH, Zon LI (2008) Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. Cell 132:631–644. doi: 10.1016/j.cell.2008.01.025

Orr HA (1990) "Why Polyploidy is Rarer in Animals Than in Plants" Revisited. Am. Nat. 136:759–770

Paharkova V, Alvarez G, Nakamura H, Cohen P, Lee KW (2015) Rat Humanin is encoded and translated in mitochondria and is localized to the mitochondrial compartment where it regulates ROS production. Molecular and Cellular Endocrinology 413:96–100. doi: 10.1016/j.mce.2015.06.015

Pancer Z, Amemiya CT, Ehrhardt GRA, Ceitlin J, Larry Gartland G, Cooper MD (2004) Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature 430:174–180. doi: 10.1038/nature02740

Paust S, Senman B, Von Andrian UH (2010) Adaptive immune responses mediated by natural killer cells: Adaptive immune responses mediated by natural killers. Immunol. Rev. 235:286–296. doi: 10.1111/j.0105-2896.2010.00906.x

Pečínková M, Matějusová I, Koubková B, Gelnar M (2007) Investigation of *Paradiplozoon homoion* (*Monogenea*, *Diplozoidae*) life cycle under experimental conditions. Parasitology International 56:179–183. doi: 10.1016/j.parint.2007.01.010

Pellicer J, Fay MF, Leitch IJ (2010) The largest eukaryotic genome of them all? Bot. J. Linn. Soc. 164:10–15. doi: 10.1111/j.1095-8339.2010.01072.x

Pietsch TW, Arnold Rachel J, Hall David J (2009) A Bizarre New Species of Frogfish of the Genus *Histiophryne* (Lophiiformes: *Antennariidae*) from Ambon and Bali, Indonesia. Copeia 2009:37–45. doi: 10.1643/CI-08-129

Pietsch TW (1981) The osteology and relationships of the anglerfish genus *Tetrabrachium* with comments on lophiiform classification. Fish. Bull. 79:387–419

Pietsch TW (2009) Oceanic anglerfishes: extraordinary diversity in the deep sea. University of California Press, Berkeley

Pietsch TW (1976) Dimorphism, Parasitism and Sex: Reproductive Strategies among Deepsea Ceratioid Anglerfishes. Copeia 1976:781-793. doi: 10.2307/1443462

Pietsch TW (2005) Dimorphism, parasitism, and sex revisited: modes of reproduction among deep-sea ceratioid anglerfishes (Teleostei: Lophiiformes). Ichthyol. Res. 52:207–236. doi: 10.1007/s10228-005-0286-2

Pietsch TW, Grobecker DB (1978) The Compleat Angler: Aggressive Mimicry in an Antennariid Anglerfish. Science 201:369–370. doi: 10.1126/science.201.4353.369

Pietsch TW, Orr JW (2007) Phylogenetic Relationships of Deep-Sea Anglerfishes of the Suborder Ceratioidei (Teleostei: Lophiiformes) Based on Morphology. Copeia 2007:1–34

Pietsch TW, Sutton TT (2015) A New Species of the Ceratioid Anglerfish Genus *Lasiognathus Regan* (Lophiiformes: *Oneirodidae*) from the Northern Gulf of Mexico. Copeia 103:429–432. doi: 10.1643/CI-14-181

Piñeiro C, Casas M, Araujo H (2001) Results of exploratory deep-sea fishing survey in the Galician Bank: biological aspects on some of seamount-associated fish (ICES Division IXb). NAFO Scientific Council Research Documents 1:7

Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Francisco Carter DR (2011) Pseudogenes: Pseudo-functional or key regulators in health and disease? RNA 17:792–798. doi: 10.1261/rna.2658311

Podlevsky JD, Chen JJ-L (2016) Evolutionary perspectives of telomerase RNA structure and function. RNA Biol. 13:720–732. doi: 10.1080/15476286.2016.1205768

Quigley DTG, Flannery K, Patterson AMA (2005) Kroyer's Deep Sea Angler Fish *Ceratias holboelli* Kroyer, 1845 (Pisces, Lophiiformes, Ceratioidei, *Ceratiidae*) in Irish Waters. Ir. Nat.' J. 28:29–30

Rajeeshkumar MP, Meera KM, Hashim M (2017) A New Species of the Deep-Sea Ceratioid Anglerfish Genus *Oneirodes* (Lophiiformes: *Oneirodidae*) from the Western Indian Ocean. Copeia 105:82–84. doi: 10.1643/CI-16-467

Rast JP, Buckley KM (2013) Lamprey immunity is far from primitive. Proc. Natl. Acad. Sci. U.S.A. 110:5746–5747. doi: 10.1073/pnas.1303541110

Rechavi O, Minevich G, Hobert O (2011) Transgenerational Inheritance of an Acquired Small RNA-Based Antiviral Response in C. elegans. Cell 147:1248–1256. doi: 10.1016/j.cell.2011.10.042

Regan CT (1925) Dwarfed Males Parasitic on the Females in Oceanic AnglerFishes (Pediculati Ceratioidea). Proc. Royal Soc. B 97:386–400. doi: 10.1098/rspb.1925.0006

Rimer J, Cohen IR, Friedman N (2014) Do all creatures possess an acquired immune system of some sort?: Prospects & Overviews. BioEssays 36:273–281. doi: 10.1002/bies.201300124

Rubartelli A, Lotze MT (2007) Inside, outside, upside down: damage-associated molecular-pattern molecules (DAMPs) and redox. Trends Immunol. 28:429–436. doi: 10.1016/j.it.2007.08.004

Sanggaard KW, Bechsgaard JS, Fang X, et al (2014) Spider genomes provide insight into composition and evolution of venom and silk. Nat. Commun. 5:. doi: 10.1038/ncomms4765

Satoh TP, Miya M, Mabuchi K, Nishida M (2016) Structure and variation of the mitochondrial genome of fishes. BMC Genom. 17:719. doi: 10.1186/s12864-016-3054-y

Schulenburg H, Kurtz J, Moret Y, Siva-Jothy MT (2009) Introduction. Ecological immunology. Philos. Trans. Royal Soc. B 364:3–14. doi: 10.1098/rstb.2008.0249

Schultz LP (1957) The Frogfishes of the Family *Antennariidae*. Proc. USNM 107:47–105. doi: 10.5479/si.00963801.107-3383.47

Seong S-Y, Matzinger P (2004) Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses. Nat. Rev. Immunol. 4:469–478. doi: 10.1038/nri1372

Shedlock AM, Pietsch TW, Haygood MG, Bentzen P, Hasegawa M (2004) Molecular systematics and life history evolution of anglerfishes (Teleostei: Lophiiformes): Evidence from mitochondrial DNA. Steenstrupia 28:129–144

Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al (2013) Draft Assembly of the *Symbiodinium minutum* Nuclear Genome Reveals Dinoflagellate Gene Structure. Curr. Biol. 23:1399–1408. doi: 10.1016/j.cub.2013.05.062

Small CM, Bassham S, Catchen J, Amores A, Fuiten AM, Brown RS, et al (2016) The genome of the Gulf pipefish enables understanding of evolutionary innovations. Genome Biol. 17:258. doi: 10.1186/s13059-016-1126-6

Solbakken MH, Tørresen OK, Nederbragt AJ, Seppola M, Gregers TF, Jakobsen KS, et al (2016) Evolutionary redesign of the Atlantic cod (*Gadus morhua* L.) Toll-like receptor repertoire by gene losses and expansions. Sci. Rep. 6:25211. doi: 10.1038/srep25211

Star B, Jentoft S (2012) Why does the immune system of Atlantic cod lack MHC II? BioEssays 34:648–651. doi: 10.1002/bies.201200005

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, et al (2011) The genome sequence of Atlantic cod reveals a unique immune system. Nature 477:207–210. doi: 10.1038/nature10342

Stefano GB, Bjenning C, Wang F, Wang N, Kream RM (2017) Mitochondrial Heteroplasmy. In: Santulli G (ed) Mitochondrial Dynamics in Cardiovascular Medicine. Springer International Publishing, Cham, pp 577–594

Swift H (1950) The Constancy of Desoxyribose Nucleic Acid in Plant Nuclei. Proc. Natl. Acad. Sci. U.S.A. 36:643–654. doi: 10.1073/pnas.36.11.643

Szathmáry E, Demeter L (1987) Group selection of early replicators and the origin of life. J. Theor. Biol. 128:463–486

Takeuchi N, Hogeweg P (2007) The Role of Complex Formation and Deleterious Mutations for the Stability of RNA-Like Replicator Systems. J. Mol. Evol. 65:668–686. doi: 10.1007/s00239-007-9044-6

Tamames J, Gil R, Latorre A, Peretó J, Silva FJ, Moya A (2007) The frontier between cell and organelle: genome analysis of *Candidatus Carsonella ruddii*. BMC Evol. Biol. 7:181. doi: 10.1186/1471-2148-7-181

Thangstad T (2006) Anglerfish (Lophius spp) in Nordic waters. Nordic Council of Ministers

Thompson CB (1995) New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity 3:531–539. doi: 10.1016/1074-7613(95)90124-8

Tørresen OK, Star B, Jentoft S, Reinar WB, Grove H, Miller JR, et al (2017) An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genom. 18:95. doi: 10.1186/s12864-016-3448-x

Uinuk-ool T, Mayer WE, Sato A, Dongak R, Cooper MD, Klein J (2002) Lamprey lymphocyte-like cells express homologs of genes involved in immunologically relevant activities of mammalian lymphocytes. Proc. Natl. Acad. Sci. U.S.A. 99:14356–14361

Urbarova I, Forêt S, Dahl M, Emblem Å, Milazzo M, Hall-Spencer JM, et al (2019) Ocean acidification at a coastal CO2 vent induces expression of stress-related transcripts and transposable elements in the sea anemone *Anemonia viridis*. PLOS ONE 14:e0210358. doi: 10.1371/journal.pone.0210358

Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nat. Rev. Genet. 18:411–424. doi: 10.1038/nrg.2017.26

Varadharajan S, Sandve SR, Gillard GB, Tørresen OK, Mulugeta TD, Hvidsten TR, et al (2018) The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication. Genome Biol. Evol. 10:2785–2800. doi: 10.1093/gbe/evy201

Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, et al (2014) Elephant shark genome provides unique insights into gnathostome evolution. Nature 505:174–179. doi: 10.1038/nature12826

Vieira S, Biscoito M, Encarnação H, Delgado J, Pietsch TW (2013) Sexual Parasitism in the Deep-sea Ceratioid Anglerfish *Centrophryne spinulosa* Regan and Trewavas (Lophiiformes: *Centrophrynidae*). Copeia 2013:666–669. doi: 10.1643/CI-13-035

Voinnet O (2001) RNA silencing as a plant immune system against viruses. Trends Genet. 17:449–459

Wallace DC, Chalkia D (2013) Mitochondrial DNA Genetics and the Heteroplasmy Conundrum in Evolution and Disease. CSH Perspect. Biol. 5:a021220. doi: 10.1101/cshperspect.a021220

Wang X, Chen W, Huang Y, Sun J, Men J, Liu H, et al (2011) The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. Genome Biol. 12:R107. doi: 10.1186/gb-2011-12-10-r107

Wang XH, Aliyari R, Li WX, Li HW, Kim K, Carthew R, et al (2006) RNA Interference Directs Innate Immunity Against Viruses in Adult Drosophila. Science 312:452–454. doi: 10.1126/science.1125694

Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, et al (2003) The genome of *Nanoarchaeum equitans*: Insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad. Sci. U.S.A. 100:12984–12988. doi: 10.1073/pnas.1735403100

Wilson AB (2017) MHC and adaptive immunity in teleost fishes. Immunogenetics 69:521–528. doi: 10.1007/s00251-017-1009-3

Xin ZZ, Yu Liu, Zhu XY, Wang Y, Zhang HB, Zhang DZ, et al (2017) Mitochondrial Genomes of Two *Bombycoidea* Insects and Implications for Their Phylogeny. Sci. Rep. 7: 6544. doi: 10.1038/s41598-017-06930-5

Yoder JA, Litman GW (2011) The phylogenetic origins of natural killer receptors and recognition: relationships, possibilities, and realities. Immunogenetics 63:123–141. doi: 10.1007/s00251-010-0506-4

Zambon RA, Vakharia VN, Wu LP (2006) RNAi is an antiviral immune response against a dsRNA virus in Drosophila melanogaster. Cell. Microbiol. 8:880–889. doi: 10.1111/j.1462-5822.2006.00688.x

Zamore PD (2002) Ancient Pathways Programmed by Small RNAs. Science 296:1265–1269. doi: 10.1126/science.1072457

Zárate SC, Traetta ME, Codagnone MG, Seilicovich A, Reinés AG (2019) Humanin, a Mitochondrial-Derived Peptide Released by Astrocytes, Prevents Synapse Loss in Hippocampal Neurons. Frontiers in Aging Neuroscience 11:123. doi: 10.3389/fnagi.2019.00123

Zurawski TH, Mair GR, Maule AG, Gelnar M, Halton DW (2003a) Microscopical Evaluation of Neural Connectivity Between Paired Stages of *Eudiplozoon nipponicum* (Monogenea: *Diplozoidae*). Journal of Parasitology 89:198–200. doi: 10.1645/0022-3395(2003)089[0198:MEONCB]2.0.CO;2

Zurawski TH, Mousley A, Maule AG, Gelnar M, Halton DW (2003b) Cytochemical studies of the neuromuscular systems of the diporpa and juvenile stages of *Eudiplozoon nipponicum* (Monogenea: *Diplozoidae*). Parasitology 126:349–357. doi: 10.1017/S0031182002002871

Paper I

SciMedCentral®

Annals of Marine Biology and Research

Short Communication

The Mitochondrial Genome of the European Anglerfish *Lophius piscatorius* Express Low-Level Substitution Heteroplasmy

Arseny Dubin#, Tor Erik Jørgensen#, Lars Martin Jakt, Truls Moum, and Steinar D. Johansen*

Department of Biosciences and Aquaculture, Nord University, Norway
#These authors contributed equally to this work and should be considered as first authors

***Corresponding author**

Steinar D. Johansen, Genomics group, Faculty of Biosciences and Aquaculture, Nord University, N8049 Bodø, Norway, Tel: 4775517833; Email: steinar.d.johansen@nord.no

Submitted: 06 February 2017

Accepted: 20 March 2017

Published: 22 March 2017

Keywords

- Heteroplasmy
- Ion torrent sequencing
- *Lophius piscatorius*
- Monkfish
- Mitogenome
- Solid sequencing

Abstract

Lophius piscatorius is of increasing economic value in the Northern European fisheries as part of an effort to generate genome resources for *L. piscatorius* we deep sequenced and assembled the complete 16,472 bp mitochondrial genome at about 2500 times coverage by Ion Torrent PGM and SOLiD technologies. Gene content and organization was similar to that of the previously reported *L. americanus* and *L. litulon*. The highly accurate and abundant SOLiD sequence reads allowed us to identify seven low-level (1% to 2.6%) heteroplasmic substitution sites. *L. piscatorius* represents the first fish species where low-level mitochondrial heteroplasmy has been detected and reported.

INTRODUCTION

Anglerfishes (Lophiiformes) constitute a large and diverse order of marine fishes with unique distribution and behavioral features [1]. The genus *Lophius* represents one of the most basal families (Lophiidae) of the anglerfishes with seven species recognized worldwide [2]. Of these, the European anglerfish *L. piscatorius* is of increasing economic value to fisheries in Northern European waters. Mitochondrial genome sequences have contributed significantly in resolving evolutionary relationships among fishes at different taxonomic levels [3,4], including various Lophiiformes species [1]. Mitochondrial genomes of fish are well studied at the molecular level and consist of small circular DNA molecules of about 16 to 17 kb in size containing 37 genes encoding 13 proteins, two ribosomal RNAs, and 22 tRNAs [5]. In addition to the conventional mitochondrial genes, long non-coding RNA genes [4], and size heteroplasmy due to tandem repeat features have been reported in the mitochondrial genomes [3,6-8]. Recent research has shown that normal animal cells appear to harbor several low-level mtDNA haplotypes in addition to the main haplotype [9-12]. Low-level heteroplasmic sites are usually represented below 2-3% compared to the corresponding consensus sequence, and challenge the single haplotype hypothesis of mtDNA [9]. Low-level heteroplasmy has been noted in Atlantic cod (*Gadus morhua*), and includes 17 sites per mitochondrial genome (our unpublished results). Low-level heteroplasmy can be readily detected by high-resolution NGS. Here, SOLiD sequencing has a great advantage in that it makes use of an internal proofreading system that calls bases in pairs resulting in a very high accuracy per base [13,14]. As part of an effort to sequence and analyze the complete *L. piscatorius* genome, we sequenced and assembled the mitochondrial genome by using the Ion Torrent PGM (Ion PGM) and SOLiD5500 next generation sequencing (NGS) platforms. Low-level substitution heteroplasmy in the *L. piscatorius* mitochondrial genome was assessed from reads generated by SOLiD sequencing, and represents the first reported study in a fish species.

MATERIALS AND METHODS

L. piscatorius nucleic acid isolation and next generation sequencing

A skeletal muscle sample of a specimen captured off the coast of Nordl and, Northern Norway, was subjected to mtDNA analysis. Total DNA isolation was performed as previously described [11,15], and all NGS sequencing was run at our in-house facility, Genomic group, Nord University - Norway. Ion PGM sequencing was performed on purified total DNA sheared to approximately 500 bp using a Covaris® S2 ultrasonicator. Subsequently, the libraries for sequencing were constructed using NEBNext® Fast DNA Library Prep Set for Ion Torrent™ according to manufacturer's protocols. Prepared libraries were quality assessed and quantified using a High Sensitivity D1000 Screen Tape on an Agilent Tape Station 2200. The libraries were diluted to approximately 50 pM. Template preparation and enrichment were performed using the Ion PGM™ Template IA 500 kit and One Touch™ ES instrument. Sequencing was performed using the Ion PGM™ Hi-Q™ View Sequencing kit on an Ion 318™ chip v2. Ion PGM total DNA sequencing generated 13,248 high quality mitochondrial reads (average length 380 nt), corresponding to 305 times mtDNA coverage. SOLiD 5500 single read sequencing was performed as previously described [11,15] using the standard protocols given by the manufacturer (Applied Biosystems). The total DNA library was deep sequenced at the SOLiD5500 platform. The sequencing generated 582,165 high quality mitochondrial reads (average length 63 nt), corresponding to 2227 times mtDNA coverage.

Phylogenetic analysis

All available complete mitogenome sequences of the Lophiidae family were included in this study. For a comparative purpose we included representative mitogenome sequences from all five Lophiiformes suborders. *Linophryne bicornis* and *Capros aper* were used as out groups. Sequences were aligned with T-COFFEE Version 11.00.8cbe486 [16], using the M-Coffee mode with clustalw, mafft, muscle and t-coffee as the alignment methods. The control region was trimmed off prior to the alignment leaving tRNAs, rRNAs and protein coding genes untouched. Maximum Likelihood (ML) phylogenetic analysis was performed using RAxML version 8.2.9 [17]. A General Time Reversible model with a discrete Gamma distribution was used. The resulting tree topology was evaluated by a rapid bootstrap analysis with 1000 replications.

RESULTS AND DISCUSSION

The circular *L. piscatorius* mitochondrial genome was found to be 16,472 bp long, and has the typical vertebrate gene content and organization (Figure 1A). The overall sequence identities to the related *L. americanus* (Western Atlantic waters) and *L. litulon* (Northwest Pacific waters) were 94% and 92%, respectively. Interestingly, we noted a 40-bp indel within the control region (CR) that discriminates *L. piscatorius* and *L. americanus* from *L. litulon* (Figure 1B and Figure 2). The close relationship between *L. piscatorius*, *L. americanus*, and *L. litulon* were further supported by phylogenetic analysis (Figure 3). In agreement with previous observations [1]. *Sladenia* was placed as the most basal of the four Lophioidei genera, followed by *Lophiodes*, *Lophiomus* and *Lophius*. Within the *Lophius* genus, *L. piscatorius* and *L. americanus* share more similarity with each other. *L. litulon* was found to represent the most divergent of the three *Lophius* species included in the analysis. The additional Lophiiformes suborders (Ceratioidei, Chaunacoidei, Antennarioidei, and Ogcocephaloidei) were all clustered outside Lophioidei (Figure 3), a feature corroborating previous observations [1]. We then used our SOLiD dataset to investigate low-level substitution heteroplasmy in *L. piscatorius*. The analyses were performed at stringent conditions allowing a

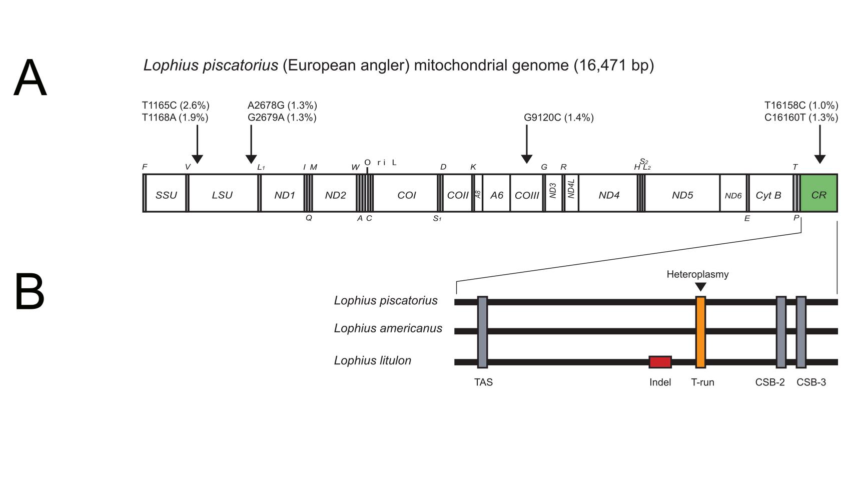
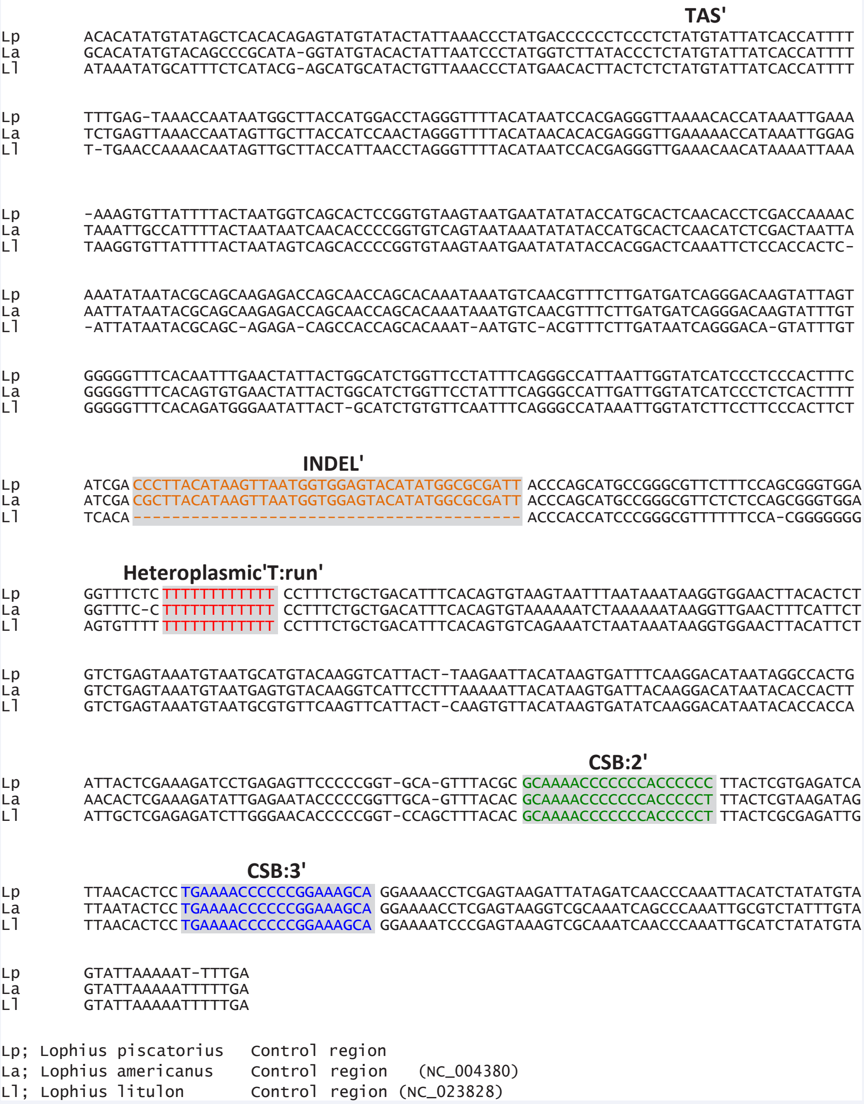


Figure 1 (A) Linear presentation of the circular European anglerfish mitochondrial genome. Gene organization is similar to that of most other vertebrates. SSU and LSU, small and large ribosomal RNA genes; ND1-6, NADH dehydrogenase subunits 1-6 genes; COI-III, Cytochrome c oxidase subunits I to III genes; A6 and A8, ATPase subunit 6 and 8 genes; Cyt B, Cytochrome B subunit gene. All genes, except ND6 and tRNA genes, are encoded by the H-strand (H-strand genes, indicated above the diagram; L-strand genes, indicated below the diagram). tRNA genes are indicated by the standard one-letter symbols for amino acids. OriL, origin of L-strand replication. CR, control region containing the D-loop, H- and L-strand promoters, and origin of H-strand replication. Low-level heteroplasmic substitution sites and their frequencies are indicated above the diagram. (B) Features of the control regions (CR) in the three Lophius species. TAS, termination associated sequence; Indel, 40 bp size variation; T-run, heteroplasmic thymidine homopolymer; CSB-2/3, conserved sequence boxes.



```
                                                                                    TAS'
Lp      ACACATATGTATAGCTCACACAGAGTATGTATACTATTAAACCCTATGACCCCCCTCCCTCTATGTATTATCACCATTTT
La      GCACATATGTACAGCCCGCATA-GGTATGTACACTATTAATCCCTATGGTCTTATACCCTCTATGTATTATCACCATTTT
Ll      ATAAATATGCATTTCTCATACG-AGCATGCATACTGTTAAACCCTATGAACACTTACTCTCTATGTATTATCACCATTTT

Lp      TTTGAG-TAAACCAATAATGGCTTACCATGGACCTAGGGTTTTACATAATCCACGAGGGTTAAAACACCATAAATTGAAA
La      TCTGAGTTAAACCAATAGTTGCTTACCATCCAACTAGGGTTTTACATAACACACGAGGGTTGAAAAACCATAAATTGGAG
Ll      T-TGAACCAAAACAATAGTTGCTTACCATTAACCTAGGGTTTTACATAATCCACGAGGGTTGAAACAACATAAAATTAAA

Lp      -AAAGTGTTATTTTACTAATGGTCAGCACTCCGGTGTAAGTAATGAATATATACCATGCACTCAACACCTCGACCAAAAC
La      TAAATTGCCATTTTACTAATAATCAACACCCCGGTGTCAGTAATAAATATATACCATGCACTCAACATCTCGACTAATTA
Ll      TAAGGTGTTATTTTACTAATAGTCAGCACCCCGGTGTAAGTAATGAATATATACCACGGACTCAAATTCTCCACCACTC-

Lp      AAATATAATACGCAGCAAGAGACCAGCAACCAGCACAAATAAATGTCAACGTTTCTTGATGATCAGGGACAAGTATTAGT
La      AATTATAATACGCAGCAAGAGACCAGCAACCAGCACAAATAAATGTCAACGTTTCTTGATGATCAGGGACAAGTATTTGT
Ll      -ATTATAATACGCAGC-AGAGA-CAGCCACCAGCACAAAT-AATGTC-ACGTTTCTTGATAATCAGGGACA-GTATTTGT

Lp      GGGGGTTTCACAATTTGAACTATTACTGGCATCTGGTTCCTATTTCAGGGCCATTAATTGGTATCATCCCTCCCACTTTC
La      GGGGGTTTCACAGTGTGAACTATTACTGGCATCTGGTTCCTATTTCAGGGCCATTGATTGGTATCATCCCTCTCACTTTT
Ll      GGGGGTTTCACAGATGGGAATATTACT-GCATCTGTGTTCAATTTCAGGGCCATAAATTGGTATCTTCCTTCCCACTTCT

                          INDEL'
Lp      ATCGA CCCTTACATAAGTTAATGGTGGAGTACATATGGCGCGATT ACCCAGCATGCCGGGCGTTCTTTCCAGCGGGTGGA
La      ATCGA CGCTTACATAAGTTAATGGTGGAGTACATATGGCGCGATT ACCCAGCATGCCGGGCGTTCTCTCCAGCGGGTGGA
Ll      TCACA ---------------------------------------- ACCCACCATCCCGGGCGTTTTTTCCA-CGGGGGGG

            Heteroplasmic'T:run'
Lp      GGTTTCTC TTTTTTTTTTTT CCTTTCTGCTGACATTTCACAGTGTAAGTAATTTAATAAATAAGGTGGAACTTACACTCT
La      GGTTTC-C TTTTTTTTTTTT CCTTTCTGCTGACATTTCACAGTGTAAAAAATCTAAAAAATAAGGTTGAACTTTCATTCT
Ll      AGTGTTTT TTTTTTTTTTTT CCTTTCTGCTGACATTTCACAGTGTCAGAAATCTAATAAATAAGGTGGAACTTACATTCT

Lp      GTCTGAGTAAATGTAATGCATGTACAAGGTCATTACT-TAAGAATTACATAAGTGATTTCAAGGACATAATAGGCCACTG
La      GTCTGAGTAAATGTAATGAGTGTACAAGGTCATTCCTTTAAAAATTACATAAGTGATTACAAGGACATAATACACCACTT
Ll      GTCTGAGTAAATGTAATGCGTGTTCAAGTTCATTACT-CAAGTGTTACATAAGTGATATCAAGGACATAATACACCACCA

                                                                CSB:2'
Lp      ATTACTCGAAAGATCCTGAGAGTTCCCCCGGT-GCA-GTTTACGC GCAAAACCCCCCACCCCCC TTACTCGTGAGATCA
La      AACACTCGAAAGATATTGAGAATACCCCCGGTTGCA-GTTTACAC GCAAAACCCCCCACCCCCT TTACTCGTAAGATAG
Ll      ATTGCTCGAGAGATCTTGGGAACACCCCCGGT-CCAGCTTTACAC GCAAAACCCCCCACCCCCT TTACTCGCGAGATTG

                      CSB:3'
Lp      TTAACACTCC TGAAAACCCCCCGGAAAGCA GGAAAACCTCGAGTAAGATTATAGATCAACCCAAATTACATCTATATGTA
La      TTAATACTCC TGAAAACCCCCCGGAAAGCA GGAAAACCTCGAGTAAGGTCGCAAATCAGCCCAAATTGCGTCTATTTGTA
Ll      TTAACACTCC TGAAAACCCCCCGGAAAGCA GGAAAATCCCGAGTAAAGTCGCAAATCAACCCAAATTGCATCTATATGTA

Lp      GTATTAAAAAT-TTTGA
La      GTATTAAAAATTTTTGA
Ll      GTATTAAAAATTTTTGA

Lp; Lophius piscatorius  Control region
La; Lophius americanus   Control region   (NC_004380)
Ll; Lophius litulon      Control region (NC_023828)
```

Figure 2 Nucleotide sequence alignment of the mitochondrial control region (CR) between *L. piscatorius* (Lp), *L. americanus* (La), and *L. litulon* (Ll). TAS, termination associated sequence; Indel, 40 bp size variation; T-run, heteroplasmic thymidine homopolymer; CSB-2/3, conserved sequence boxes.

polymorphic site variant to be recognized only if it accounted for at least 1% of the total reads. Based on these criteria we identified seven heteroplasmic sites (Figure 1A). Four sites were located within the large subunit ribosomal RNA gene, and one with in *COIII*. The latter (G9120C) resulted in a non-synonymous amino acid shift (Ala to Pro) at amino acid position 119. The last two positions were found within CR, at a heteroplasmic thymidine motif (T-run). In addition to substitution heteroplasmy we also noted length variation at the T-run motif (Figure 1A). Interestingly, none of the seven low-level heteroplasmic sites identified in *L. piscatorius* corresponded to the 17 sites recognized in *G. morhua* (our unpublished results). The observation of an apparent non-conserved low-level site heteroplasmy was supported by studies performed in human mitochondrial genomes [6,18]. Humans appear to have individual low-level heteroplasmic haplotypes, but similar variants were recognized among different tissues

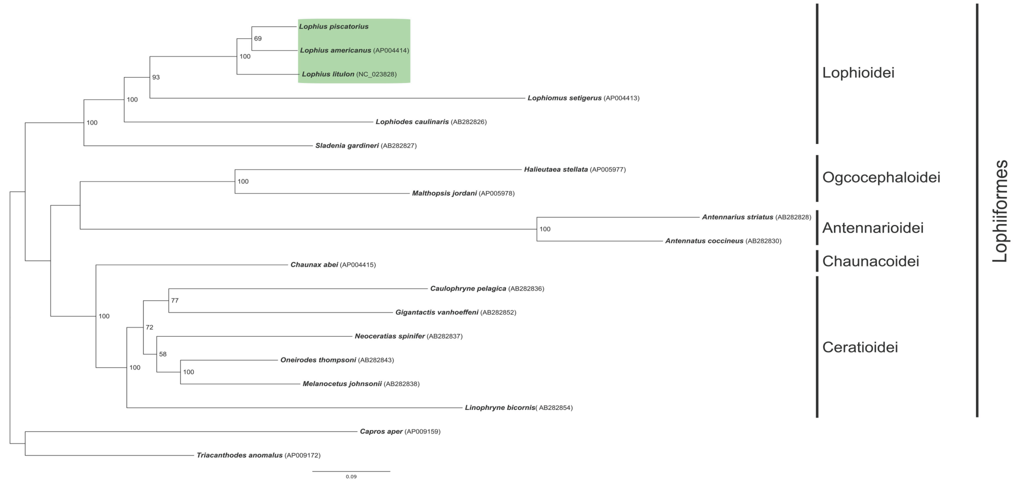

Lophius piscatorius
69
Lophius americanus (AP004414)
100
Lophius litulon (NC_023828)
93
Lophiomus setigerus (AP004413)
100
Lophiodes caulinaris (AB282826)
100
Sladenia gardineri (AB282827)
Halieutaea stellata (AP005977)
100
Malthopsis jordani (AP005978)
Antennarius striatus (AB282828)
100
Antennatus coccineus (AB282830)
Chaunax abei (AP004415)
Caulophryne pelagica (AB282836)
77
Gigantactis vanhoeffeni (AB282852)
100
72
Neoceratias spinifer (AB282837)
58
Oneirodes thompsoni (AB282843)
100
100
Melanocetus johnsonii (AB282838)
Linophryne bicornis(AB282854)
Capros aper (AP009159)
Triacanthodes anomalus (AP009172)
0.09
Lophioidei
Ogcocephaloidei
Antennarioidei
Chaunacoidei
Ceratioidei
Lophiiformes

Figure 3 A phylogenetic maximum-likelihood tree based on 18 complete mitochondrial genome sequences (except for CR). Bootstrap values (%) of 1000 replications are shown at the internal nodes. Only bootstrap values above 50% are shown.

of the same individual [6], and between mothers to child [18]. Low-level mitochondrial heteroplasmic sites are associated with human tumors, and some apparently become selected as somatic mutations during tumor development [19,20].

CONCLUSION

This study reports the complete mitochondrial genome sequence of the European anglerfish *L. piscatorius* and detected seven low-level heteroplasmic substitution sites at frequencies between 1% and 2.6%. *L. piscatorius* was found to be most closely related to *L. americanus*. A 40-bp indel within the control region discriminated the former *Lophius* species and *L. litulon*.

REFERENCES

1. Miya M, Pietsch TW, Orr JW, Arnold RJ, Satoh TP, Shedlock AM, et al. Evolutionary history of anglerfishes (Teleostei: Lophiiformes): a mitogenome perspective. BMC Evol Biol. 2010; 10: 58.
2. Farina AC, Azevedo M, Landa J, Duarte R, Sampedro P, Gostas G, et al. Lophius in the world: a synthesis on the common features and life strategies. ICES J Mar Sci. 2008; 65: 1272-1280.
3. Breines R, Ursvik A, Nymark M, Johansen SD, Coucheron DH. Complete mitochondrial genome sequences of the Arctic Ocean codfishes Arctogadus glacialis and Boreogadus saida reveal oriL and tRNA gene duplications. Polar Biol. 2008; 31: 1245-1252.
4. Jørgensen TE, Bakke I, Ursvik A, Andreassen M, Moum T, Johansen SD1. An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA. BMC Evol Biol. 2014; 14: 182.
5. Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999; 27: 1767-1780.
6. Arnason E, Rand DM. Heteroplasmy of short tandem repeats in mitochondrial DNA of Atlantic cod, Gadus morhua. Genetics. 1992; 132: 211-220.
7. Chen CA, Anonuevo Ablan MC, McManus JW, Bell JD, Tuan VS, Cabanban AS, et al. Variable number of tandem repeats (VNTRs), heteroplasmy, and sequence variation of the mitochondrial control region in the threespot Dascyllus, Dascyllus trimaculatus (Perciformes: Pomacentridae). Zool Stud. 2004; 43: 803-812.
8. Mjelle KA, Karlsen BO, Jrgensen TE, Moum T, Johansen SD. Halibut mitochondrial genomes contain extensive heteroplasmic tandem repeat arrays involved in DNA recombination. BMC Genomics. 2008; 9: 10.
9. He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, et al. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature. 2010; 464: 610-614.
10. Ameur A, Stewart JB, Freyer C, Hagström E, Ingman M, Larsson NG , et al. Ultra-deep sequencing of mouse mitochondrial DNA: mutational patterns and their origins. PLoS Genet. 2011; 7: 1002028.
11. Emblem, Karlsen BO, Evertsen J, Miller DJ, Johansen SD. Mitogenome polymorphism in a single branch sample revealed by SOLiD deep sequencing of the Lophelia pertusa coral genome. Gene. 2012; 506: 344-349.
12. Emblem Å, Okkenhaug S, Weiss ES, Denver DR, Karlsen BO, Moum T, et al. Sea anemones possess dynamic mitogenome structures. Mol Phylogenet Evol. 2014; 75: 184-193.
13. Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T. Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS One. 2011; 6: 19534.
14. Gardner K, Payne BAI, Horvath R, Chinnery PF. Use of stereotypical mutational motifs to define resolution limits for the ultra-deep resequencing mitochondrial DNA. Eur J Hum Genetics. 2014; 23: 413-415.
15. Karlsen BO, Emblem Å, Jørgensen TE, Klingan KA, Nordeide JT, Moum T, et al. Mitogenome sequence variation in migratory and stationary

ecotypes of North-east Atlantic cod. Mar Genomics. 2014; 15: 103-108.

16. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302: 205-217.

17. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post- 199 analysis of large phylogenies. Bioinformatics. 2014; 30: 9.

18. Guo Y, Li CI, Sheng Q, Winther JF, Cai Q, Boice JD, et al. Very low-level heteroplasmy mtDNA variations are inherited in humans. J Genet Genomics. 2013; 40: 607-615.

19. Kloss-Brandstatter A, Weissensteinar H, Erhart G, Schafer G, Forer L Schonherr S, et al. Validation of next-generation sequencing of entire mitochondrial genomes and the diversity of mitochondrial DNA mutations in oral squamous cell carcinoma. PLoS One. 2015; 10: 0135643.

20. Chattopadhyay E, De Sarkar N, Singh R, Ray A, Roy R, Paul M, et al. Genome- wide mitochondrial DNA sequence variations and lower expression of OXPHOS genes predict mitochondrial dysfunction in oral cancer tissue. Tumour Biol. 2016; 37: 11861-11871.

Cite this article

Dubin A, Jørgensen TE, Jakt LM, Moum T, Johansen SD (2017) The Mitochondrial Genome of the European Anglerfish Lophius piscatorius Express Low-Level Substitution Heteroplasmy. Ann Mar Biol Res 4(1): 1019.

Figure S1

```
Lophius piscatorius
16471 bp

GCTAGCGTAGCTTATTTAAAGCATAACACTGAAGATGTTAAGACTGAGTCCTAAAAAACT
CCGTAAGTACTAAAAGTTTGGTCCTGACTTTATTATCAACTATAACTAAACTTACACATG
CAAGTCTCCACACCCCTGTGAAGTACGCCCTATGTATCTCCCCCCAGAGAACAAGGAGCA
GGCATCAGGCACAAGCACACTTAGCCCATAACGCCTTGCTTAGCCACACCCCCACGGGAA
CTCAGCAGTGATAAACATTAAGCCATAAGCGAAAGCTTGACTTAGTTAAAGTTAAGAGGG
CCGGTAAAACTCGTGCCAGCCACCGCGGTTATACGAGAGGCCCAAGTTGACAACAGTCGG
CGTAAAGCGTGGTTAGGCCATCAACCCCCACTAAAGTCGAATGCCCTCAAAGCTGTTATA
CGCACCCGAGGGTTAGAAGTTCAAATACGAAAGTAACTTTATAAGTCTGAACCCACGAAA
GCTACGGCACAAACTGGGATTAGAAACCCCACTATGCCTAGCCCTAAACATTGGCAACAC
AAAACACCCGTTGCCCGCCCGGGCACTACGAGCATTAGCTTAAAACCCAAAGGACTTGGC
GGTGCTTTAGACCCACCTAGAGGAGCCTGTTCTAGAACCGATACCCCCCGTTAAACCTCA
CCCCTTCTTGTCATTACCGCCTATATACCGCCGTCGTCAGCTTACCCTGTGAAGGATTAG
TAGTAAGCAAAGTTGTCACAACCCAAAACGTCAGGTCGAGGTGTAGCCTATGAAGGGGGA
AGAAATGGGCTACATTCACTAATTAAGAGAATACGAACGATGTATTGAAACACACGTCCA
AAGGAGGATTTAGCAGTAAGCAAAAAATAGAGCGTTTTGCTGAAAATGGCCCTGAAGCGC
GCACACACCGCCCGTCACTCTCCCCAAGCTAACGACTAAATATAACTAAATCATAATAAC
TGCAAAGGGGAGGCAAGTCGTAACATGGTAAGTGTACCGGAAAGTGCACTTGGAAAAACC
AGAGCATAGCTAAACAACCAAAGCATCTCCCTTACACCGAGAAGGTGTCCGTGCAAATCG
GACTGCCCTGATACCTAATAGCTAGCCACACCAACTAAACACAACAAAACTACATAAATA
CCCCCTAACCACCCCCCCCCCCCCACCTCAAACAAACCATTTTTCCACCCTAGTACGGGC
GACAGAAAAGGACCTTTGTGAGCAATAGAAAAAGTACCGCAAGGGAAAGCTGAAAGAGAT
ATGAAAAAGCCCAGTAAAGTTTAGAAAAGCAGAGATTAAACTCGTACCTTTTGCATCATG
TTTTAGCCAGTAACACCCAAGCAAAGAGCCCTTTAGTTTGGACCCCCGAAACTAAGCGAG
CTACTCCAAGACAGCCTATTTAAAGGGCACACCCGTCTCTGTGGCAAAAGAGTGGGAAGA
GCTTTGAGTAGAGGTGACAGACCTACCGAGCTTAGTTATAGCTGGTTCCCTGGGAAATGG
ATAGGAGTTCAGCCTTCTAAGTTCTTCACTCCAACCCTTCAACAAGGTGAAAAGAAATCA
GAAGAGTTAATCAAGGGGGGTACAGCCCCCTTGAAAAAAGATACAACTTTAGCAGAAGGA
TAAAGATCATACGAAATTTAAAGGAGAGTCCTCTTGGTGGGCCTAAAAGCAGCCACCCCA
ACAGAAAGCGTTAAAGCTCAAATACGGAACACCCTATATATTCTGACAACCTAATCTTAA
TCCTCTAGTACTACCAGGAAGCCCTATGCGAGCATAGGAATTATACTGCTAATATGAGTA
ATAAGAGAACATAAGATTCTCTCCTAGCACAAGTGTAAATCGGAACGAACCCCCACCGAA
AATTGACGGCCCCAAGAAAAGAGGGAACTGGATGAAAAATAAAAAACTAGAAGAACACCC
AACAATTAACCGTTAACCTCACACTAGAGTGCTGCCAAGGAAAGACTAAAAGGGGAAGAA
GGAACTCGGCAAACACACCCAAGCCTCGCCTGTTTACCAAAAACATCGCCTCTTGTACCT
AATAAATAAGAGGTCCTGCCTGCCCTGTGACAATTAGTTCAACGGCCGCGGTATTTTGAC
CGTGCGAAGGTAGCGCAATCACTTGTCTTTTAAATGGAGACCTGTATGAATGGCAAAACG
AGGGCTTAGCTGTCTCCTCCCCCCAGTCAATGAAATTGATCTCCCCGTGAAGAAGCGGGG
ATACTCACATAAGACGAGAAGACCCTATGGAGCTTTAGACACCAAGGCAGATCACGTTAA
AACCCCTGAATAAAGGAATAAACAAGATGAAGACTACCCTATGTCTTTGGTTGGGGCGAC
CGCGGGGCAAAAAATACCCCCCATGTGGAAAGGGAACACCCTTCCTATCACCCAGAGCTA
CCGCTCTAATTATCAGAATATCTGACCAAAAGATCCGGCTCAAGCCGATCAACGGACCGA
GTTACCCTAGGGATAACAGCGCAATCCCCTTCTAGAGACCCTATCGACAAGGGAGTTTAC
GACCTCGATGTTGGATCAGGACATCCTAATGGTGCAGCTGCTATTAAGGGTTCGTTTGTT
CAACGATTAAAGTCCTACGTGATCTGAGTTCAGACCGGAGTAATCCAGGTCAGTTTCTAT
CTATGATATGATCTTTTCTAGTACGCAAGGACCGAAAAGAGGGGGCCCCTGTAATATACA
CACCCCCTCCTCACCTAATGAAATCAACTAAATTAGGCAAGAGGACATACCTACCTAGCC
AAAGAAAATGGCATGTTAAGGTGGCAGAGCCCGGCCAATGCAAAAGACCTAAGCCCTTTC
CACAGAGGTTCAACTCCTCTCCTTAACTATGATCTCAACCACTGTAATTTATATTATTAA
CCCCCTCGCCTTTATCGTACCCGTCCTCCTAGCTGTCGCATTTCTAACTTTGCTAGAGCG
AAAGGTCTTGGGCTACATACAACTTCGAAAAGGTCCCAACATCGTTGGCCCCTACGGCCT
CCTCCAACCAATCGCAGACGGGGTAAAACTATTCATCAAAGAGCCCATTCGACCTTCTGC
CTCTTCCCCTATTCTATTCTTAGCCGCCCCTATTCTAGCCCTCACCCTTGCCCTAACCCT
TTGGGCCCCAATACCGCTCCCCTACCCAGTCCTTGACCTCAACCTTGGGATTTTATTTAT
CCTAGCATTATCAAGCCTAGCAGTATACTCAATCCTCGGGTCAGGCTGAGCCTCCAACTC
CAAATACGCACTCATCGGTGCCCTCCGGGCTGTCGCACAAACAATCTCATATGAGGTAAG
CCTTGGGCTAATCCTACTCAGCACTATTATCTTCACCGGGAGCTTCACACTTCAAACCTT
TAACACAGCCCAAGAAACCATCTGATTAATCTTACCCGCCTGACCACTAGCCGCCATATG
ATACATCTCCACCCTCGCAGAGACAAACCGAGCCCCATTTGACCTCACCGAAGGAGAATC
TGAACTTGTATCCGGCTTTAACGTCGAATACGCAGGAGGCCCCTTCGCCCTCTTCTTCCT
AGCAGAATATGCAAACATCTTACTAATAAATACGCTCTCAGCCACCCTGTTCTTAGGAGC
AGCCCACATCCCCTCCTTCCCAGAACTAACAACGGCCAACCTCATAACGAAAGCCGCCCT
CCTCTCGGTCGTCTTTCTCTGAGTGCGAGCCTCCTACCCCCGATTTCGATACGACCAACT
AATACACCTCATCTGAAAAAACTTCCTACCCCTCACCCTTGCATTAATTATTTGACACCT
GTCCCTCCCAATTGCGCTCGCAGGACTCCCCCCTCAAATATAACAGGAGCCGTGCCTGAA
AAAAGGACTACTTTGATGGAGTAGATTACGGGGGTTAAAGTCCCCCCGACTCCTTAGAAA
GAAGGGGTTTGAACCCTCCCCGGAGAGATCAAAACTCTCAGTGCTTCCACTACACCACTT
CCTAGTAAAGTCAGCTAACAAAGCTTTTGGGCCCATACCCCAAACATGTCGGTTAAAATC
CCTCCTTCACTAATGAACCCCTTTATCTTATTTATCCTCTTGTCAGGACTTGGCCTAGGC
ACCACCATCACCTTTGCCAGCTCCCACTGGCTTGCAGCCTGGATAGGACTTGAGATCAGC
ACGCTAGCCATCCTACCACTTATAGCACAGCACCACCACCCCCGAGCAGTCGAAGCAACT
ACTAAGTATTTCCTTGTACAGGCAACAGCAGCAGCCATACTTCTATTTGCAAGCACCACC
AACGCCTGACTCACGGGCCAATGGGATATTCAACAAGCCCTGCATCCCCTCCCCTGCACA
ATAATTACACTTGCACTAGCCCTGAAAGTCGGACTTGCCCCCCTACATACCTGACTCCCG
GAAGTACTTCAAGGCCTAGATCTCACCACAGGAGTGATCCTCTCCACCTGACAGAAACTT
GCCCCCTTTGCCCTACTTTTACAGCTCCAACAAAACAACCCCCACCTCCTCATAGCCCTC
GGACTAGTTTCTACCCTTGTGGGCGGATGAGGGGGCTTAAACCAAACACAACTACGAAAA
ATCCTAGCCTACTCATCCATCGCCCACCTTGGCTGAATAATCCTCATCATCCAATTCTCT
CCCTCCCTTACTTTACTCACCCTGGCTATGTACTTTATCATAACTTTCTCAACTTTTACC
CTATTTAAACTACATAACTCAACCAATATTAACTCACTAGCCACCTCCTGAGCAAAAGCC
CCCGCACTTACAGCCCTCACCCCCCTCCTTATATTATCCCTCGGCGGCCTCCCACCACTC
TCAGGCTTCATGCCCAAGTGACTGATCCTTCTAGAACTCACTAAACAAGACTTAGCCCCA
```

```
GCCGCCACCTTTGCTGCACTCTCTGCCCTCCTGAGCTTATACTTCTACTTGCGACTCTCA
TACGCAATGGCCTTAACACTATCCCCCTCCAGTCTCACGGCATCCACCCCTTGACGCCTC
TTTAATACACAACTCACCCTGCCCTTAGCCATCACTACCATACTTGCCCTACTGCTTCTA
CCCTTAACACCTGGTACAACAGCTTTACTGGCCCTTTAGGGGCTTAGGTTAGTACAAGAC
CAAGGGCCTTCAAAGCCCTAAGCGAGAGTGAAACCCTCCCAGCCCCTATAAAACTTGCGG
GACATTACCCCACATCTCCTGCATGCAAAACAGACACTTTAATTAAGCTAAAGCTTTTCT
AGATGAGAAGGCCTCGATCCTACAGTTTCTTAGTTAACAGCTAAGCGCTCAAACCAGCGA
GCATTCATCTACCCTTCCCCCGCCTAATAGGCGGCTAAAGGCGGGGGAAGCCCCGGAGGG
TGACTAACCCTCTTCTTAAGATTTGCAATCTTATATGTAAAAACACCTCAAGGCTTGGTA
AGAAGAGGGCTCGAACCTCTGTTTATGGGGCTACAACCCACCACTTACTCAGCCATCTTA
CCTGTGGCAATCACACGTTGATTTTTCTCGACTAATCACAAAGATATCGGCACCCTTTAT
TTAATCTTTGGAGCCTGAGCCGGAATAGTAGGCACCGCCCTAAGCTTACTTATTCGGGCT
GAACTAAGCCAGCCTGGCGCCCTCTTAGGGGATGACCAAATCTATAACGTTATTGTTACC
GCACACGCCTTTGTAATAATTTTCTTTATGGTTATACCAATCATGATTGGGGGGTTCGGC
AACTGACTTATCCCCTTAATGATCGGGGCCCCCGACATAGCCTTCCCTCGAATGAACAAC
ATGAGCTTCTGACTCCTTCCCCCCTCTTTCCTCCTGCTACTTGCCTCTTCCGGAGTTGAA
GCCGGAGCAGGCACTGGATGAACCGTTTACCCCCCACTAGCAGGGAACCTTGCACACGCA
GGAGCCTCTGTCGACCTAACTATTTTCTCCCTTCACCTGGCCGGGATCTCTTCAATCCTA
GGGGCAATCAACTTTATCACAACAATTATCAACATAAAACCCCCCACGATCTCCCAATAT
CAAACGCCCTTATTCGTATGAGCTGTTTTAATCACAGCAGTCCTTTTACTCCTATCCCTG
CCTGTGCTTGCCGCAGGCATTACCATGCTCTTAACAGATCGGAACTTAAACACCACTTTC
TTTGACCCCACCGGGGGAGGGGACCCTATCCTGTACCAACACTTATTCTGATTCTTCGGC
CACCCTGAAGTTTACATTTTAATTCTACCAGGCTTTGGTATAGTATCCCACATTGTTGCC
TACTATGCAGGTAAAAAAGAACCCTTCGGCTACCTCGGCATGGTCTGAGCTATGATAGCT
ATCGGCCTACTAGGGTTTATCGTCTGAGCCCACCACATATTCACGGTCGGAATAGATGTC
GATACACGAGCTTATTTTACGTCCGCCACAATAATTATTGCCATCCCCACGGGAGTAAAA
GTATTTAGCTGACTAGCGACACTTCACGGAGGTGTAATTAAATGAGAAACCCCTCTCCTG
TGAGCACTAGGGTTTATCTTCCTCTTTACGGTGGGCGGCCTCACCGGCATCGTACTTGCC
AACTCCTCCCTAGATATTATCTTGCACGACACTTACTACGTGGTCGCCCACTTCCACTAT
GTCCTCTCCATGGGGGCTGTTTTCGCCATCATAGGAGCCTTTGTACACTGATTCCCCCTG
TTCTCAGGCTACACACTCCACAGCACATGAACAAAAATCCACTTTGGTATTATATTTGCA
GGTGTTAATCTCACATTCTTCCCCCAACACTTCTTAGGACTGGCAGGCATGCCCCGCCGC
TACTCAGACTACCCGGACGCCTACACCCTGTGAAATACAGTCTCCTCTATTGGCTCACTA
ATTTCACTAATTGCAGTCATCATATTCCTGTTTATCCTTTGAGAAGCATTTACTGCCAAG
CGAGAAGTCCTATCAGTGGAACTTGCTGCAACAAACGTAGAGTGACTACACGGCTGCCCT
CCACCCTATCACACCTTTGAAGAACCTGCATTCGTCCAAGTCCAAACAAACTAAACAAGA
AAGGAGGGAATTGAACCCCCGTAACCCGGTTTCAAGCCGACCACATCACCGCTCTGTCAC
TTTCTTTATAAGACACTAGTAAAGCAGCTATTACACTGCCTTGTCAAGGCAGTACCGCGG
GTTGGAGCCCCGCGTGTCTTACCCAATGGCACATCCCTCCCAACTAGGCTTTCAAGACGC
AGCTTCACCTGTAATAGAAGAACTCCTTCACTTTCACGACCATGTTCTAATAATTGTATT
CCTCATCAGCACACTAGTACTTTACATTATTGTCGCCATGGTTTCAACCAAGCTTACAAG
CAAATACCTGTTAGATTCTCAGGAAGTCGAGATTATTTGAACCATTCTCCCCGCTATCAT
CTTAATTCTAATTGCCCTCCCTTCCTTACGAATCCTCTACCTTATGGATGAAGTAGACGA
CCCCCACCTCACCATTAAAGCAATGGGACACCAATGATACTGAAGTTATGAGTACGCAGA
TTACGTAGACCTTGAGTTTGACTCCTACATAACACCCACCCAAGACCTATTGCCCGGACA
GTTCCGACTTCTCGAAGCAGATCATCGAATGGTAATTCCCGTTGCATCCCCCATTCGAGT
GTTAGTATCCGCAGAAGACGTATTACACTCATGGGCCGTTCCGGCCCTAGGTGTAAAAAT
AGATGCCGTCCCAGGACGACTAAACCAAACAGCCTTTATTACCTTACGACCAGGTGTGTA
TTATGGACAATGCTCAGAAATTTGTGGAGCCAACCATAGCTTTATACCCATTGTAGTGGA
AGCCGTTCCATTAAAACACTTTGAGAACTGATCGGCATTAATAGTAGAAGACGCTTCACT
AAGAAGCTAAACAGGGCCATAGCGTTAGCCTTTTAAGCTAAAGACTGGTGACTCCCAACC
ACCCTTAGTGATATGCCCCAACTCAACCCCACACCTTGATTTGCCATTCTAATCTTCTCC
TGATTAGTATTCTTAACAATTCTGCCCTCAAAAGTGATAGCCCATACTTTCCCAAACGAG
CCAACCCCACAAAGCACAGAGAAATCAGAAACAAACCCCTGAACCTGACCATGACACTAA
GCCTCTTTGACCAATTTATGAGCCCCTGACTTCTTGGAATCCCCCTCATTGCGCTAGCCC
TAGCACTCCCCTGGACACTTTTCCCTACCCCTTCTGCACGATGATTAAATAACCGCCTAC
TCACCCTCCAAGGATGATTCATTAACCGATTCACCCAGCAACTTCTGCAACCCATGAGCC
TCGGAGGACATAAGTGGGCCCTACTACTCACCTCCTTAATACTTTACCTTATTACACTAA
ATATGTTAGGCCTACTTCCCTACACCTTCACTCCGACCACGCAACTATCCCTTAATATTG
GCTTTGCCGTCCCCCTCTGACTAGCAACTGTAATTATTGGCATGCGAAACCAGCCAACTG
TCGCCCTAGGCCACCTCCTCCCAGAAGGAACCCCAACCCCCCTCATCCCAATCCTTATTA
TTATCGAAACAATTAGCCTCTTCATTCGACCCTTAGCCCTAGGAGTCCGACTCACTGCTA
ACCTCACAGCAGGCCACCTCCTCATCCAACTCATTGCCACAGCTGCTTTCGTCCTACTGC
CCCTTATGCCTACAGTTGCCCTTCTTACGGCCACCCTTCTATTCCTACTTACCCTCCTAG
AGGTGGCCGTGGCTATAATTCAAGCCTACGTCTTTGTCCTACTACTAAGCCTCTACCTAC
AAGAAAACGTATAATGACCCATCAAGCACACGCATACCACATAGTAGACCCGAGCCCCTG
GCCCCTCACAGGAGCAGTAGCTGCCTTATTAATAACATCCGGCCTTGCAATTTGATTCCA
TTTCAACTCCATAGTCCTTATCCCCGTCGGGACCACCCTCCTTCTCCTCACCATGCTCCA
ATGATGGCGAGATATCGTACGAGAGGGGACATTTCAGGGACACCACACACCCCCAGTCCA
AAAGGGGCTACGATATGGTATAATCCTATTTATTACTTCAGAAGTTTTCTTCTTCTTGGG
ATTCTTCTGGGCTTTTTACCACGCCAGCCTTGCCCCCACCCCTGAACTGGGGGGCTTCTG
GCCCCCCGCAGGAATCACCCCCCTGGACCCCTTTGAAGTCCCCCTACTTAACACAGCCGT
TCTTCTTGCTTCCGGCGTGACAGTCACCTGAGCCCACCACAGCATCATAGAAGGCGAACG
AAAACAAGCCCTCCAATCCCTCTTCCTAACTATCCTATTAGGATTTTATTTCACCTTCCT
CCAAGGCCTTGAGTACTATGAAGCCCCTTTCACCCTCGCGGACGGCGCCTACGGCTCGAC
CTTCTTTGTAGCAACCGGTTTCCACGGCCTACATGTTATCATCGGCTCCATCTTCCTCAC
TGTATGCCTAATCCGACAAATCCGCCACCACTTCACATCAGAACACCACTTTGGGTTTGA
AGCAGCTGCCTGATACTGACACTTCGTGGACGTTGTCTGACTCTTCCTGTATATCTCAAT
CTACTGATGAGGCTCATAATCTTTCTAGTACTAACGTTAGTATAAGTGACTTCCAATCAC
CCGGTCTTGGTTAAAACCCAAGGAAAGATAATGAACTTAGTCACAACTATCGTCTCCATT
ACCGGCGCCCTCTCCCTTATCCTAGCCACCGTAGCTTTCTGACTCCCACAAATAACCCCA
GACCACGAAAAGCTATCACCCTACGAATGCGGCTTCGACCCCCTTGGATCAGCCCGACTG
CCTTTCTCCCTTCGGTTCTTTTTGGTGGCCATTCTATTCTTATTATTTGACCTAGAAATC
GCCCTGCTATTACCCCTACCTTGAGGAGACCAACTAGCCTCCCCACTAATAACATTCCTA
TGAGCCGCTGCAGTTCTTATTCTTCTCACACTAGGGCTTATCTATGAATGACTCCAAGGG
GGACTAGAATGAGCCGAATAGGTAATTAGTTTAAACAAAACATTTGATTTCGGCTTAAAA
ACTTGTGGTTAAAGTCCACAATTAACCTAATGACCCCCACAAATTTTACATTCTCCGCAG
```

```
CCTTTATATTAGGATTGGCAGGCCTAACATTCCACCGAACGCACCTTCTCTCCGCCCTGC
TATGCCTAGAAGGAATAATACTAGCTCTATTCCTCGCCCTCTCACTGTGATCCCTTCAAC
TAGGAGCAACCAGCTTTTCAGCCACCCCCCTCCTCCTTCTGGCTTTCTCCGCCTGCGAAG
CCAGTGCAGGCTTGGCCCTTTTAGTAGCAACTACGCGGACCCACGGATCCGATCGACTCC
AAACACTCAACCTGCTGCAATGTTAAAAGTCCTCATCCCCACCCTTATGCTAATACCAAC
AATTATGGTCACTAAAGCCAAATGACTATGGCCAACCACCCTCCTCCATAGCCTCCTAAT
CGCACTTATCAGCCTGACTTGACTAAAAAACCTGGGGGAAACGGGGTGGTCCCACCTCAG
CCCCTACATGGCAACAGACCCCCTATCCACACCCCTTCTGGTATTGACTTGCTGGCTGCT
CCCGCTTATAATCCTAGCAAGCCAAGCACACACAGCCTCAGAACCCGTCGGCCGTCAACG
ATTATACATTATCCTCCTCACTTCTCTCCAACTCTTCCTTACTATAGCTTTTAGCGCAAC
TGAAATGATCCTATTTTACATTATATTTGAAGCCACCCTGATCCCCACACTCATTCTAAT
TACCCGGTGGGGCAACCAAACAGAACGTCTTAACGCAGGCACCTACTTCCTTTTCTACAC
CCTAGCAGGCTCACTGCCCCTACTTATTGCCCTACTCTGGCTCCAAAACAATGCGGGCAC
CCTCTCACTTCTCACCCTTCTTTACTCAAACCCCCTACAGCTGGGCGTATGTGCCCACAA
GCTCTGATGAGCAGGCTGCGTACTAGCATTCCTTGTAAAAATGCCTCTGTACGGCATACA
CCTGTGACTACCTAAAGCCCATGTGGAAGCCCCAGTTGCCGGATCAATAATCCTTGCAGC
CGTCCTACTGAAGCTCGGAGGGTATGGCATGATGCGCATACTAGTAATGCTAGAACCCCT
TACCAAGGAATTAAGCTATCCTTTCATTATCCTCGCACTATGGGGTGTAATTATAACAGG
ATCTATTTGCCTCCGCCAAACAGACCTGAAATCCTTAATCGCCTACTCATCAGTCAGCCA
TATGGGTTTGGTGGTAGGAGGTATCCTCATCCAAACCCCTTGAGGCTTCACAGGAGCCCT
AATCTTAATAATCGCCCACGGCCTGACCTCCTCCGCCCTATTCTGCTTAGCCAATACTAA
CTATGAACGCACACACAGCCGAACAATAGTTCTCGCCCGAGGACTTCAAATGGCCCTCCC
CTTGATAACTGCTTGGTGATTTATTGCAAGCCTGGCCAACCTCGCCCTTCCCCCCCTACC
CAACTTAATGGGAGAATTAATGATTATCACCTCCCTCTTCAGCTGATCATGATGAACACT
AGCCTTAACGGGAGCGGGAACCCTTATCACGGCAGGCTACTCCCTGTACATGTTCCTCAT
AACACAACGAGGGCCCCTCCCCACCCACATCCTAGCCCTAGAACCATCACACTCACGCGA
ACATCTGCTAATGGCCCTGCACCTTCTCCCTCTCTTATTACTAACTCTTAAGCCCGAGTT
AATCTGGGGCTGAAGCATATGTAGACATAATTTAATGAAAATATTAGATTGTGATTCTAA
AAATAGGGGTTAAACCCCCCTTGTCCACCGGGAGAGGCTGGCTAGCAACGAAAACTGCTA
ATTCTCGCTACTTTGGTTGAACTCCAAAGCTCACTCGAACACTGCTTCTAAAGGATAATA
GCTCATCCGTTGGTCTTAGGAACCAAAAACTCTTGGTGCAAATCCAAGTGGAAGCTATGC
ACCCTACTTCAACCATAATAGCCTCTAGCTTGCTTATCATTTTTGCCCTACTGGCATATC
CAGTATTTACAACCCTAACCCCTCACCCCACCCACCGAACCTGAGCTTTATCACAAGTGA
AAACAGCTGTTAAACTAGCATTCTTCGCCAGCCTATTACCGCTCTTCCTATTCGTCAACG
AGGGAGCAGAAGCCATTATCACTAGCTGAACCTGGACCAACACACACACCTTTGACATTA
ATATCAGCCTTAAATTTGACATTTACTCGATTATTTTCACACCCGTTGCCCTATACGTCA
CCTGGTCGATTCTAGAATTTGCCTCCTGATACATGCACTCCGACCCTTATATGAACCGAT
TTTTCAAATACCTATTAATCTTCCTTATTACCATAATTATCCTAGTCACAGCTAACAATA
TATTTCAACTCTTTATTGGCTGGGAAGGCGTTGGTATTATATCCTTTTTACTTATCGGGT
GATGATACGGGCGTGCTGACGCTAACACTGCTGCCCTCCAAGCAGTACTGTACAACCGGG
TCGGGGATGTCGGCCTAATTCTCGCAATAGCCTGAATGGCAACCAACCTAAACTCATGAG
AACTACAACAAATGTTCTCATGCACTAAAAATGTTGACCTAACCCTCCCCCTATTAGGCC
TAATCTTAGCCGCCACAGGAAAATCTGCCCAATTTGGCCTGCACCCATGGCTGCCAGCTG
CCATGGAGGGTCCTACGCCGGTATCCGCCCTCCTGCATTCCAGCACCATGGTTGTTGCGG
GAATCTTTCTACTTATTCGTATAAGCCCACTTCTGGAGAATAACCAGACAGCCCTAACAA
CCTGCCTATGCTTAGGAGCACTAACAACCCTCTTCACAGCCACATGCGCCCTCACCCAGA
ATGATATTAAAAAAATCGTTGCTTTCTCTACATCAAGCCAACTTGGCCTAATAATAGTCA
CCATCGGCCTTAATCAACCACAGCTAGCCTTCCTTCACATTTGCACCCATGCCTTCTTTA
AAGCTATGTTATTCCTATGCTCCGGCTCCATCATCCATAGCCTTAATGATGAGCAAGACA
TTCGGAAAATGGGAGGAATACACCACCTCGCCCCTTTCACCTCTTCCTCACTAACCTTGG
GCAGTCTAGCCCTCACAGGGACCCCTTTCCTAGCAGGGTTCTTCTCAAAAGACGCCATTA
TTGAAGCACTCAACACCTCCCACCTAAACGCCTGAGCCCTTGCCCTCACCCTTCTAGCCA
CCTCTTTTACAGCCATCTACAGTCTTCGCATTGTATATTTTGTATCTATGGGTAACCCCC
GATTCAACTCACTATCCCCCATCAATGAAAATAACCCCGCGGTCATCAACCCCATTAAAC
GACTAGCTTGAGGCAGCATCCTAGCGGGGCTTTTAATCACCTCTCACATTACACCTTTAA
AAACACCCGTAATATCCATACCCCTCACCCTAAAAATCGCCGCCCTCGCCGTAACTATCA
CTGGTCTCCTCATCGCCCTAGAACTAGCCCATTTAACTAACAAACACCTCAAACCCTCCC
CCAAAATGAAACCTCACCATTTCTCAAATATGCTTGGATTCTTCCCAGCAATTGTTCACC
ACCACGCCCCTAAAATCAACCTGACACTAGGCCAAACAATTGCAAGTCAAATAGTTGACC
AAACATGATTAGAAAAATCAGGACCCAAGGCAGCTGCCTCACTTAATATGCCCCTAATCA
CTTCCACAAGCAACACCCAACAAGGGATAATTAAAACCTATTTAACGCTATTCCTTCTCA
CCCTAACCTTAGCGGTGGTAGTACTTATAATCTAAACCGCCCGCAAAGCACCCCGGCCCA
GCCCCCGGGTCAACTCCAACACCACAAATAAAGTGAGAAGGAGAACCCAGGCACTAATTA
AGAGCATTCCCCCCCCGAAAGAGTACATCAGTGCAACCCCCCCAATATCCCCCCGAGACA
CAAGAAGCTCCCCAAACTCTTCAACAGATTGCCAGGAAGATTCATACCACCCCCCTAAAA
ATAAGCTTGATGCCAGACCCACCCCACCTAAGTAAAAAACAACAGACAATGCAATCGGAC
GACTACCAAGCCCCTCAGGAAAAGGTTCAGCAGCCAAAGCTGCCGAATATGCAAACACAA
CTAATATTCCCCCGAGATAAATTAGAAACAACACCAGGGATAAAAAAGACCCCCCATGCC
CAACCAAAATCCCACACCCCAAGCCTGCTACAACAACCAACCCTAGAGCAGCAAAATAAG
GAGAGGGATTAGAAGCCACGACCACCATTCCCACCACCAGCCCCACCAAAAATAAATACA
CAACATAAGTCATAATTCCTGCCAGGACTTTAACCAGGACGAATGGCTTGAAAAACCACC
GTTGTTATTCAACTACAAGAACCCTAATGACAAGCCTCCGTAAAACACACCCCCTATTAA
AAATTGCTAATGACGCTTTAGTAGACCTACCCGCCCCCTCCAATATCTCTGCATGATGGA
ACTTTGGCTCCTTACTAGCACTCTGCTTAATTGCCCAAATCTTAACAGGACTATTTCTAG
CCATACACTACACCTCTGATATCGCCACAGCTTTCTCCTCAGTAGCACACATCTGTCGAG
ACGTAAACTACGGATGACTAATCCGCAACCTTCATGCCAATGGGGCCTCCTTCTTCTTTA
TTTGCATCTATATGCATATTGGACGAGGCTTGTATTACGGCTCTTACCTCTACAAAGAAA
CATGAAACGTTGGAGTAGTCCTTCTCCTCCTAGTTATAATGACAGCATTCGTAGGCTACG
TTCTACCGTGGGGCCAAATATCCTTCTGAGGCGCCACCGTCATTACCAACCTTCTATCTG
CCGTCCCCTATGTTGGTAATATATTAGTTCAATGAATCTGAGGGGGCTTCTCAGTAGATA
ACGCCACCCTTACCCGATTCTTTGCCTTCCACTTCCTATTCCCCTTCGTCATCCTAGCAA
TGACCGTCATCCACCTCCTCTTCCTCCATGAGACAGGCTCAAACAACCCCTTAGGAATTA
ACTCAGATGCCGATAAAATCTCCTTCCACCCCTACTACTCCTACAAAGACCTTGTAGGGT
TTGCAATCGTCCTAATTACACTCACGGCCCTAGCCCTCTTTGCCCCAAACCTCTTAGGAG
ACCCAGACAACTTCACCCCAGCAAACCCCTTAGTCACTCCTCCACACATCAAGCCAGAGT
GATACTTCTTATTCGCCTACGCAATCTTGCGCTCCATTCCCAACAAACTAGGGGGCGTTT
TGGCTTTACTTGCCTCCATCCTAATCCTAATAGTCGTACCCATCCTTCACACGTCAAAAC
```

```
AACGAAGCCTCACCTTCCGCCCTGTTACCCAGTTCTTATTCTGAACATTAGTTGCAAACG
TTGCCATTCTTACTTGAATCGGGGGTATGCCCGTAGAACACCCATTTGTAATCATTGGAC
AAGTAGCATCCCTACTCTACTTCGCCCTATTCCTGATCGCCATACCTCTCACTGGTTGAG
TTGAAAATAAATTTCTTGACTGGACCCCCAAATTATAGAGCACTAGTAGCTCAGACCCAG
AGCATCGGTCTTGTAAGCCGAATGTCGGGAGTTAGATTCTCCCCTACTGCTCAAGGGAAA
GGGATTTTAACCCTTACCACTAGCTCCCAAAGCTAGCGCTCTAAACTAAACTACCCCTTG
ACACATATGTATAGCTCACACAGAGTATGTATACTATTAAACCCTATGACCCCCCTCCCT
CTATGTATTATCACCATTTTTTTGAGTAAACCAATAATGGCTTACCATGGACCTAGGGTT
TTACATAATCCACGAGGGTTAAAACACCATAAATTGAAAAAAGTGTTATTTTACTAATGG
TCAGCACTCCGGTGTAAGTAATGAATATATACCATGCACTCAACACCTCGACCAAAACAA
ATATAATACGCAGCAAGAGACCAGCAACCAGCACAAATAAATGTCAACGTTTCTTGATGA
TCAGGGACAAGTATTAGTGGGGGTTTCACAATTTGAACTATTACTGGCATCTGGTTCCTA
TTTCAGGGCCATTAATTGGTATCATCCCTCCCACTTTCATCGACCCTTACATAAGTTAAT
GGTGGAGTACATATGGCGCGATTACCCAGCATGCCGGGCGTTCTTTCCAGCGGGTGGAGG
TTTCTCTTTTTTTTTTTTCCTTTCTGCTGACATTTCACAGTGTAAGTAATTTAATAAATA
AGGTGGAACTTACACTCTGTCTGAGTAAATGTAATGCATGTACAAGGTCATTACTTAAGA
ATTACATAAGTGATTTCAAGGACATAATAGGCCACTGATTACTCGAAAGATCCTGAGAGT
TCCCCCGGTGCAGTTTACGCGCAAAACCCCCCCACCCCCCTTACTCGTGAGATCATTAAC
ACTCCTGAAAACCCCCCGGAAAGCAGGAAAACCTCGAGTAAGATTATAGATCAACCCAAA
TTACATCTATATGTAGTATTAAAAATTTTGA
```

Paper II

The mitochondrial transcriptome of the anglerfish *Lophius piscatorius*

Arseny Dubin, Tor Erik Jørgensen, Lars Martin Jakt and Steinar Daae Johansen*

Genomics group, Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway

Email addresses:

arseny.dubin@nord.no (Arseny Dubin)

tor.e.jorgensen@nord.no (Tor Erik Jørgensen)

lars.m.jakt@nord.no (Lars Martin Jakt)

steinar.d.johansen@nord.no (Steinar Daae Johansen)

* To whom correspondence should be addressed. Genomics group, Faculty of Biosciences and Aquaculture, Nord University, N-8049 Bodø, Norway.

E-mail: steinar.d.johansen@nord.no

Abstract

Objective: Analyze key features of the anglerfish *Lophius piscatorius* mitochondrial transcriptome based on high-throughput total RNA sequencing.

Results: We determined the complete mitochondrial DNA and corresponding transcriptome sequences of *L. piscatorius*. Key features include highly abundant mitochondrial ribosomal RNAs (10-100 times that of mRNAs), and that cytochrome oxidase mRNAs appeared > 5 times more abundant than both NADH dehydrogenase and ATPase mRNAs. Unusual for a vertebrate mitochondrial mRNA, the polyadenylated COI mRNA was found to harbour a 75 nucleotide 3’ untranslated region. The mitochondrial genome expressed several noncanonical genes, including the long noncoding RNAs lncCR-H, lncCR-L and lncCOI. Whereas lncCR- H and lncCR-L mapped to opposite strands in a non-overlapping organization within the control region, lncCOI appeared novel among vertebrates. We found lncCOI to be a highly abundant mitochondrial RNA in antisense to the COI mRNA. Finally, we present the coding potential of a humanin-like peptide within the large subunit ribosomal RNA.

Keywords: Anglerfish; antisense RNA; humanin; mitogenome; long noncoding RNA; lncCOI; mtDNA

Introduction

The mitochondrial genome (mtDNA) gene content and organization is highly conserved among vertebrates [1]. All species investigated to date encode the same 37 canonical gene products of 13 hydrophobic membrane proteins, 2 ribosomal RNAs (mt-rRNAs), and 22 transfer RNAs (tRNAs), as well as several noncanonical peptides and long noncoding RNAs (lncRNAs) [2]. The corresponding mitochondrial transcriptomes are less studied and have mainly been investigated in a small number of vertebrates including some mammalian cells and tissues [3,4] and in gadiform fishes [5,6]. Only minor differences were noted between the mammals and fish. In general, three polycistronic transcripts initiated from two H-strand promoters (HSP_1 and HSP_2) and one L-strand promoter (LSP) are involved in mitochondrial gene expression. Whereas the highly abundant HSP_1 transcript mainly generates mt-rRNAs, the HSP_2 transcript is responsible for most messenger RNAs (mRNAs) and tRNAs. The LSP transcript generates one mRNA and eight tRNAs.

Atlantic cod mt-rRNAs are oligo-adenylated [5], and fold into similar secondary structures as in other fish species [7,8]. Interestingly, several mitochondrial-derived peptides (MDP) have been proposed to be encoded on both strands of the mt-rRNA gene locus [9], and two MDPs (MOTS-c and Humanin) have coding potential in Atlantic cod [2]. Mature tRNAs carry the non-template CCA at their 3' ends and fold into the common tRNA patterns [7,10]. Eleven mature mRNAs were found expressed in the Atlantic cod mitochondria, 10 from the HSP_2 transcript and one from LSP, and two of the HSP_2-specific mRNAs were bicistronic (ND4/4L and ATPase8/6) [6]. All mRNAs, except the LSP-specific ND6 mRNA, were found polyadenylated.

Mitochondrial lncRNAs have been identified and investigated in Atlantic cod [2]. Here, lncCR-H and lncCR-L correspond to different strands of the mitochondrial control region (CR). Both lncRNAs are clearly expressed and appear to generate small stable mitochondrial RNA (mitosRNA) [2,6,11,12]. We recently reported low-level substitution heteroplasmy of the anglerfish *Lophius piscatorius* based on SOLiD deep sequencing [13]. As part of a study to generate a full reference genome and transcriptome for *L. piscatorius*, we here present the complete mitochondrial genome and key features of the corresponding mitochondrial transcriptome.

Main text

Methods

Nucleic acid extraction and high-throughput sequencing

L. piscatorius tissue samples were collected from two specimens obtained by commercial fishery off the coast of Nordland County, Northern Norway, in 2015 (BF1) and 2017 (BF2). Total DNA from BF1 was extracted from muscle tissue and sequenced by the SOLiD5500 and Ion PGM platforms as described previously [13]. Total DNA sequencing (head kidney) of BF2 using the Illumina HiSeqX platform was performed by Dovetail Genomics (Chicago, US) as a service. Total RNA from heart muscle tissue of specimen BF2 was isolated using QIAzol Lysis Reagent (QIAGEN, Hilden - Germany) according to the manufacturers protocol. Cellular rRNA was depleted from 1 ug of total RNA using the RiboMinus Eukaryote System v2 (Thermo Fisher Scientific, Waltham, MA - USA), and whole transcriptome library was constructed using the Ion Total RNA-seq kit v2 (Thermo Fisher Scientific) according to the manufacturers protocols. Manual template preparation on an Ion OneTouch 2 System (Thermo Fisher Scientific) and sequencing of two Ion 540 chips on the Ion GeneStudio S5 System (Thermo Fisher Scientific) were carried out at our Genomics Platform (Nord University) according to the manufacturers protocols. The sequencing resulted in a total of 154,741,088 reads with a mean read length of 169 nt, corresponding to 26 billion nt.

Data analysis

RNA reads were quality trimmed with Cutadapt [14] using q20 as a threshold. The minimum read length was set to 50 nt. Trimmed RNA reads were then mapped to the BF2 mitochondrial genome with CLC Genomics Workbench v12 (QIAGEN). The "Length fraction" parameter was set to 0.9 and "Similarity fraction" to 0.96, requiring at least 90% of the read length to map with 96% similarity. Other parameters were set to their defaults. The resulting BAM file was coordinate sorted with SAMtools [15] and then processed with BEDTools [16] (genomecov command) to obtain a base level coverage of the mitogenome. Mean coverage for each gene and non-coding region was calculated from bed file. Alignments were visually examined to identify non-coding RNAs and polyA tails.

Results

Canonical mitochondrial genes in *L. piscatorius*

Complete mitochondrial genome sequences of two *L. piscatorius* specimens were determined using the Ion PGM and SOLiD5500 technologies (BF1; 2532 times mean coverage; MF994812; [13]) and the Illumina HiSeqX pair-end reads (BF2; 7643 times mean coverage; MN240767). The circular mtDNA possesses the conventional gene content and organization typical in vertebrates (Fig. 1A). Among the nine polymorphic sites between BF1 and BF2, seven were located in protein coding genes, representing both synonymous and non- synonymous amino acid substitutions (Additional file 1: Table S1).

Mitochondrial transcripts from *L. piscatorius* BF2 were generated by Ion S5 sequencing. About 145.2 million quality-filtered total RNA reads were obtained, including 510,484 reads (0.35%) unambiguously identified as mitochondrial transcripts when mapped to the BF2 mitochondrial genome. Several features were noted when inspecting the mitochondrial transcripts and correlating the expression values to specific mitochondrial gene regions (Fig. 1B): (1) reads from mt-rRNA gene transcripts were 10-100 times more abundant than protein coding transcripts. This observation is likely underestimated due to rRNA depletion of input RNA. (2) Of coding transcripts, cytochrome oxidase subunits were the most abundant, with NADH dehydrogenase subunits and ATPase subunits transcripts being much less abundant. (3) Highly abundant lncRNAs mapping to opposite strands within the mitochondrial CR and cytochrome oxidase I gene (COI) were noted. (4) Most mRNAs were polyadenylated and lacked 5' and 3' untranslated regions (UTRs) (Additional file 2: Table S2). A notable exception was the 75 nt 3'UTR of the COI mRNA (see below). Secondary structure predictions of *L. piscatorius* mt-SSU rRNA (Additional file 3: Figure S1) and mt-LSU rRNA (Additional file 4: Figure S2) showed typical fish mitochondrial features [7,8]. Secondary structure predictions of all 22 tRNAs (Additional file 5: Figure S3) followed the general pattern of fish mitochondrial tRNAs [7].

Non-canonical mitochondrial genes in L. piscatorius

The two CR specific lncRNAs (lncCR-H and lncCR-L), transcribed from opposite strands in a non-overlapping organization (Fig. 2A), have previously been reported in Atlantic cod [11,12] and human [17]. The L-strand specific lncCR-L was found to be 30 times more abundant than the L-strand specific ND6 mRNA (Fig. 1B). The vertebrate mitochondrial COI mRNA is unusual due to the presence of a structured 3' UTR. We identified a polyadenylated COI mRNA containing a 75-nt 3'UTR in *L. piscatorius* (Fig. 2B). RNA-Seq data revealed a highly abundant 178 nt antisense RNA to the 5' end of COI mRNA (Figs. 1B and 2B), which appeared novel among vertebrate mitochondrial lncRNAs and named lncCOI.

MDPs have been reported in vertebrates, and the best characterized is the humanin peptide [18]. The humanin gene is located within the mt-LSU rDNA locus. *L. piscatorius* contains a humanin-like open reading frame (ORF) in the mt-LSU rRNA Domain IV, at the exact same location as in Atlantic cod and human (Fig. 2C, left panel). Sequence analysis revealed the derived peptide sequence to be invariant within the *Lophius* genus, highly conserved among fishes, and well conserved between fish and mammals (Fig. 2C, right panel).

Discussion

Here we provide the complete mitochondrial genome sequence and key features of the corresponding transcriptome of the anglerfish *L. piscatorius*. We found all canonical mitochondrial genes to be expressed. Mt-rRNAs were clearly more abundant than mRNAs. Two lncRNAs (lncCR-L and lncCR-H) mapped to the mitochondrial CR, a finding that corroborates recent reports of Atlantic cod and human cells [2,17]. Interestingly, we identified a novel and highly abundant antisense RNA (lncCOI). Finally, we present feature support for the encoding of a humanin-like peptide within the mt-LSU rRNA.

Teleost fish mitochondria generate 10 mature mRNAs from a single primary transcript (HSP_2) that subsequently are translated into 12 mitochondrial proteins in OxPhos complexes I, III, IV and V [2,6]. Thus, the observed differences in transcript abundance may be explained by differential stability of individual mRNAs, and not by transcription initiation. Fish mitochondrial mRNAs contain no, or very short UTRs. A notable exception is the approximately 75-nt 3'UTR of the COI mRNA, which is conserved between fish species [2,6] and mammals [19]. A study in rat showed that the nuclear miR-181c was regulating COI mRNA stability in heart tissue by 3'UTR binding [20]. A similar 75-nt 3'UTR was detected in the polyadenylated *L. piscatorius* COI mRNA. It is plausible, that the 3'UTR structure in *L. piscatorius* contributes to the COI mRNA stability.

A number of mitochondrial lncRNAs have been noted and characterized in vertebrates [reviewed in 2,21,22], but no lncRNA has so far been linked to COI gene sequences. Our observation of lncCOI appears novel among vertebrates. If the highly abundant lncCOI contributes to mRNA stability, translational regulation, or other mitochondrial roles is currently not known. We also detected two CR-specific lncRNAs (lncCR-L and lncCR-H) in *L. piscatorius*. lncCR-L corresponds to the 5' end region of the LSP primary transcript and has been detected in Atlantic cod [6]. lncCR-L appears homologous to the 7S RNA reported inhuman mitochondria more than three decades ago [23], that was recently shown to be aberrantly expressed in human cancer cells [17]. Interestingly, lncCR-L was the most abundant non-ribosomal mitochondrial transcripts in *L. piscatorius*. lncCR-H, on the other hand, corresponds to the 3' end region of the HSP_2 primary transcript. It has been reported in Atlantic cod to be polyadenylated, to harbor a mirror tRNA, a noncoding intergenic spacer, and heteroplasmic tandem repeats [11,12]. Similar to that of Atlantic cod, the *L. piscatorius* lncCR-H contains a mirror tRNA and a polyA tail. lncCR-L and lncCR-H may function as precursors for mitosRNAs [2], but their biological role has not been elucidated.

Reports in mammals conclude that the humanin peptide has important roles in cellular signaling [18,24-26]. Previously we presented evidence supporting the encoding of humanin- like peptides in Domain IV of the mt-LSU rRNA in gadiform fishes [2], and similar features have recently been reported in avians [27]. Here we show that several anglerfishes, including all *Lophius* species where mtDNA sequences are available, possess humanin-like ORFs. How vertebrate humanin is translated is under debate, but different scenarios may be considered; 1) The humanin ORF is recognized in mt-rRNA by mitochondrial ribosomes and translated in mitochondria. This scenario is supported by a recent study in rat [25]. 2) Translation may also occur in cytosolic ribosomes, which would require mitochondrial export. Interestingly, a chimeric mt-LSU rRNA (lncRNA SncmtRNA) was reported to be expressed in human proliferating cells and localized in the cytoplasm and the nucleus [28,29]. 3) Humanin may also be expressed from a nuclear copy of mt-LSU rRNA (Numt sequence). Studies from human cells provide support for the expression of nuclear-encoded humanin isoforms [30]. The latter scenario may explain why most, but not all, fish species have intact humanin-like ORFs in Domain IV.

Conclusion

Our study provides a mitochondrial transcriptome resource from *L. piscatorius* heart muscle tissue. All mitochondrial genes were expressed, and different mRNAs had different abundances. Two lncRNAs mapped to the control region, we identified one novel lncRNA antisense to the COI mRNA, and the mt-LSU rRNA has the potential of coding a humanin- like peptide.

Limitations

Mitochondrial RNA sequencing was performed in one tissue type in one individual and has to be considered as a *snapshot* of the mitochondrial transcriptome of *L. piscatorius*.

Abbreviations

CR: control region

lncRNA: long noncoding RNA

LSU: large subunit

MDP: mitochondrial-derived peptide

mitosRNA: mitochondrial small RNA

mtDNA: mitochondrial DNA

OxPhos: oxidative phosphorylation

SSU: small subunit

UTR: untranslated region

Declarations

Authors' contributions

AD, TEJ, LMJ and SDJ organized the sequencing of the mitochondrial genomes. AD and SDJ contributed to mtDNA sequence analyses. SDJ directed the research in collaboration with all authors. AD and SDJ wrote the paper in collaboration with all authors. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank Dovetail Genomics for Illumina sequencing of total DNA, and Harald Lorentzen, Bofisk – Bodø, for providing *L. piscatorius* specimen samples.

Competing interests

The authors declare that they have no competing interests.

Availability of data

Accession numbers of mitogenomes are available from GenBank under the accession number MF994812 (BF1) and MN240767 (BF2). The RNA-seq raw sequencing data accession number at NCBI's Sequence Read Archived (SRA) is PRJNA557585.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Fish tissue samples were obtained at site of fisheries, and do not involve research on animals.

Funding

This work was supported by grants from Nord University.

Additional files

Additional file 1: Table S1. Polymorphic sites in the mitochondrial genome of *L. piscatorius* specimens BF1 and BF2.

Additional file 2: Table S2. 5' and 3' sequence features of *L. piscatorius* mitochondrial mRNAs derived from RNA-seq reads.

Additional file 3: Figure S1. Complete secondary structure diagram of *L. piscatorius* mitochondrial small subunit rRNA.

Additional file 4: Figure S2. Complete secondary structure diagram of *L. piscatorius* mitochondrial large subunit rRNA. Polymorphic site between BF1 and BF2 is indicated in Domain I. Low-level heteroplasmic sites in BF1 are indicated in Domains I and VI.

Additional file 5: Figure S3. Secondary structure diagram of *L. piscatorius* mitochondrial tRNAs. Anti-codon triplets and the non-template CCA are indicated.

References

1. Boore JL: Animal mitochondrial genomes. Nucleic Acids Res. 1999;27:1767-1780.
2. Jørgensen TE, Johansen SD: Expanding the coding potential of vertebrate mitochondrial genomes: Lesson learned from the Atlantic cod. In: Mitochondrial DNA – New Insights (M. Mattila, ed). InTech Open. 2018. pp 95-111.
3. Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood A-MJ, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, Filipovska A, Mattick JS: The human mitochondrial transcriptome. Cell. 2011;146:645–658.
4. Gustafsson CM, Falkenberg M, Larsson N-G: Maintenance and expression of mammalian mitochondrial DNA. Ann Rev Biochem. 2016;85:133-160.
5. Bakke I, Johansen S: Characterization of mitochondrial ribosomal RNA genes in gadiformes: sequence variation, secondary structure features, and phylogenetic implications. Mol Phylogen Evol. 2002;25:87-100.
6. Coucheron DH, Nymark M, Breines R, Karlsen BO, Andreassen M, Jørgensen TE, Moum T, Johansen SD: Characterization of mitochondrial mRNAs in codfish reveals unique features compared to mammals. Curr Genet. 2011;57:213-222.
7. Satoh TP, Miya M, Mabuchi K, Nishida M: Structure and variation of the mitochondrial genome of fishes. BMC Genomics. 2016;17:719.
8. Jørgensen TE, Karlsen BO, Emblem Å, Breines R, Andreassen M, Rounge TB, Nederbragt AJ, Jakobsen KS, Nymark M, Ursvik A, Coucheron DH, Jakt LM, Nordeide JT, Moum T, Johansen SD: Mitochondrial genome variation of Atlantic cod. BMC Res Notes. 2018;11:397.
9. Cobb LJ, Lee C, Xiao J, Yen K, Wong RG, Nakamura HK, Mehta HH, Gao Q, Ashur C, Huffman DM, Wan J, Muzumdar R, Barzilai N, Cohen P: Naturally occurring mitochondrial-derived peptides are age-dependent regulators of apoptosis, insulin sensitivity, and inflammatory markers. Aging. 2016;8:796–809.
10. Johansen S, Guddal PH, Johansen T: Organization of the mitochondrial genome of Atlantic cod, *Gadus morhua*. Nucleic Acids Res. 1990;18:411–419.
11. Jørgensen TE, Bakke I, Ursvik A, Andreassen M, Moum T, Johansen SD: An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA. BMC Evol Biol. 2014;14:182.
12. Jørgensen TE, Karlsen BO, Emblem Å, Jaky LM, Nordeide JT, Moum T, Johansen SD: A mitochondrial long noncoding RNA in Atlantic cod harbors complex heteroplasmic tandem repeat motifs. Mitochondrial DNA Part A. 2019;30:307-311.
13. Dubin A, Jørgensen TE, Jakt LM, Moum T, Johansen SD: The mitochondrial genome of the European Anglerfish *Lophius piscatorius* express low-level substitution heteroplasmy. Ann Mar Biol Res. 2017;4:1019.

14. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10-12. doi.org/10.14806/ej.17.1.200.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, et al: The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078-2079.
16. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinform. 2010;26:841-842.
17. Hedberg A, Knutsen E, Løvhaugen AS, Jørgensen TE, Perander M, Johansen SD: Cancer-specific SNPs originate from low-level heteroplasmic variants in human mitochondrial genomes of a matched cell line pair. Mitochondrial DNA Part A. 2019;30:82-91.
18. Guo B, Zhai D, Cabezas E, Welsh K, Nouraini S, Satterthwait AC, Reed JC: Humanin peptide suppresses apoptosis by interfering with Bax activation. Nature. 2003;423:456-461.
19. Slomovic S, Laufer D, Geiger D, Schuster G: Polyadenylation and degradation of human mitochondrial RNA: the prokaryotic past leaves its mark. Mol Cell Biol. 2005;25:6427-6435.
20. Das SD, Ferlito M, Kent OA, Fox-Talbot K, Wang R, Liu D, Raghavachari N, Yang Y, Wheelan SJ, Murphy E, Steenbergen C: Nuclear miRNA regulates the mitochondrial genome in the heart. Cir Res. 2012;110:1596-1603.
21. Dietrich A, Wallet C, Iqbal RK, Gualberto JM, Lotfi F: Organellar non-coding RNAs: emerging regulation mechanisms. Biochimie. 2015;117:48-62.
22. Zhao Y, Sun L, Wang RR, Hu J-F, Cui J: The effects of mitochondria-associated long noncoding RNAs in cancer mitochondria: new players in an old arena. Crit Rev Oncol Hematol. 2018;131:76-82.
23. Ojala D, Crews S, Montoya J, Gelfand R, Attardi G: A small polyadenylated RNA (7S RNA), containing a putative ribosome attachment site, maps near the origin of human mitochondrial DNA replication. J Mol Biol. 1981;150:303-314.
24. Lee C, Yen K, Cohen P: Humanin: a harbinger of mitochondrial-derived peptides? Trends Endocrin Metabol. 2013;24:222–228.
25. Paharkova V, Alvarez G, Nakamura H, Cohen P, Lee KW: Rat humanin is encoded and translated in mitochondria and is located to the mitochondrial compartment where it regulates ROS production. Mol Cell Endocrinol. 2015;413:96-100.
26. Zarate SC, Traetta ME, Codagnone MG, Seilicovich A, Reines AG: Humanin, a mitochondrial-derived peptide released by astrocytes, prevents synapse loss in hippocampal neurons. Front Aging Neurosci. 2019;11:123.
27. Moritz M, Degletagne C, Romestaing C, Duchamp C: Comparative genomic analysis identifies small open reading frames (sORFs) with peptide-encoding

features in avian 16S rDNA. Genomics. 2019; https://doi.org/10.1016/j.ygeno.2019.06.026.

28. Burzio V, Villota C, Villegas J, Landerer E, Boccardo E, Villa LL, Martinez R, Lopez C, Gaete F, Toro V, Rodrigues X, Burzio LO: Expression of a family of noncoding mitochondrial RNAs distinguishes normal from cancer cells. Proc Natl Acad Sci USA. 2009;106:9430-9434.
29. Fitzpatrick C, Bendek MF, Briones M, Farfan N, Silva VA, Nardocci G, Montecino M, Boland A, Deleuze JF, Villegas J, Villota C, Silva V, Lobos-Gonzalez L, Borgna V, Barrey E, Burzio LO, Burzio VA: Mitochondrial ncRNA targeting induces cell cycle arrest and tumor growth inhibition of MDA-MB-231 breast cancer cells through reduction of key cell cycle progression factors. Cell Death Dis. 2019;10:423.
30. Bodzioch M, Lapicka-Bodzioch K, Zapala B, Kamysz W, Kiec-Wilk B, Dembinska-Kiec A: Evidence for potential functionality of nuclearly-encoded humanin isoforms. Genomics. 2009;94:247-256.

Figure legends

Figure 1

Mitochondrial genome organization and transcripts of *L. piscatorius*. (**A**) Mitochondrial genome presented as a linear map of the circular mtDNA. Single nucleotide polymorphisms in BF2 compared to BF1 are indicated above the gene map. Gene abbreviations: mtSSU and mtLSU, mitochondrial small- and large-subunit ribosomal RNA; ND1-6, NADH dehydrogenase subunit 1 to 6; COI-III, cytochrome oxidase subunit I to III; A6 and A8, ATPase subunit 6 and 8; Cyt B, cytochrome b; lncCR-H and lncCR-L, long non-coding RNAs coded by the control region (CR); lncCOI, long noncoding antisense RNA. tRNA genes are indicated by the standard one-letter symbols for amino acids. All genes are H-strand specific, except Q, A, N, C, Y, S_1, E, P, ND6, lncCOI and lncCR-L (L-strand). (**B**) Histogram presentation of mean coverage expression values of mt-rRNAs, mRNAs, and lncRNAs based on Ion Torrent S5 total RNA sequencing

Figure 2

Non-canonical mitochondrial gene products in *L. piscatorius*. (**A**) Schematic view of CR and the long noncoding RNAs lncCR-L (approx. 620 nt) and lncCR-H (approx. 140 nt). Abbreviations: P and F, $tRNA^{Pro}$ and $tRNA^{Phe}$ genes; TAS, termination associated sequence; CSB2 and 3, conserved sequence box 2 and 3. (**B**) Schematic view of the COI mRNA structure and lncCOI (178 nt). The translation initiation codon (GUG) and termination codon (UAA) are indicated. The 3'UTR contains a 75 nt mirror $tRNA^{Ser}$ motif. (**C**) Left panel: Secondary structure diagram of the mt-LSU rRNA Domain IV of *L. piscatorius* with coding potential of a humanin-like peptide. See Additional file 4: Figure S2 for complete secondary structure diagram of mt-LSU rRNA. Right panel: Amino acid alignment of humanin-like peptides in anglerfish, zebrafish (ZF), codfish and mammals. Indicated 'stars' below the alignment represent conserved residues.

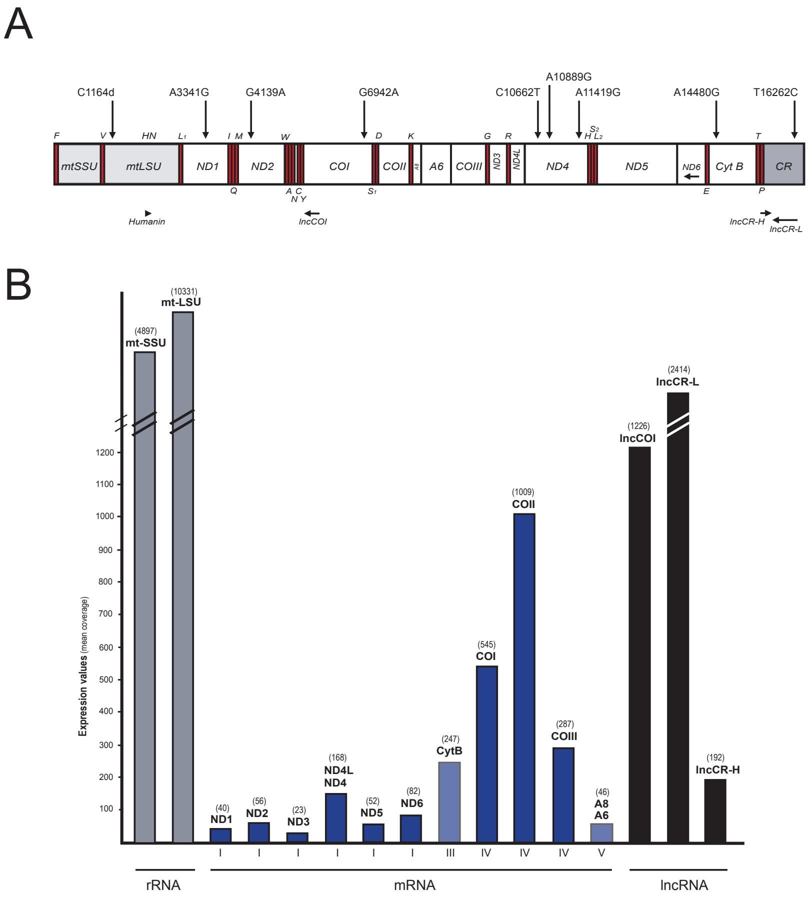
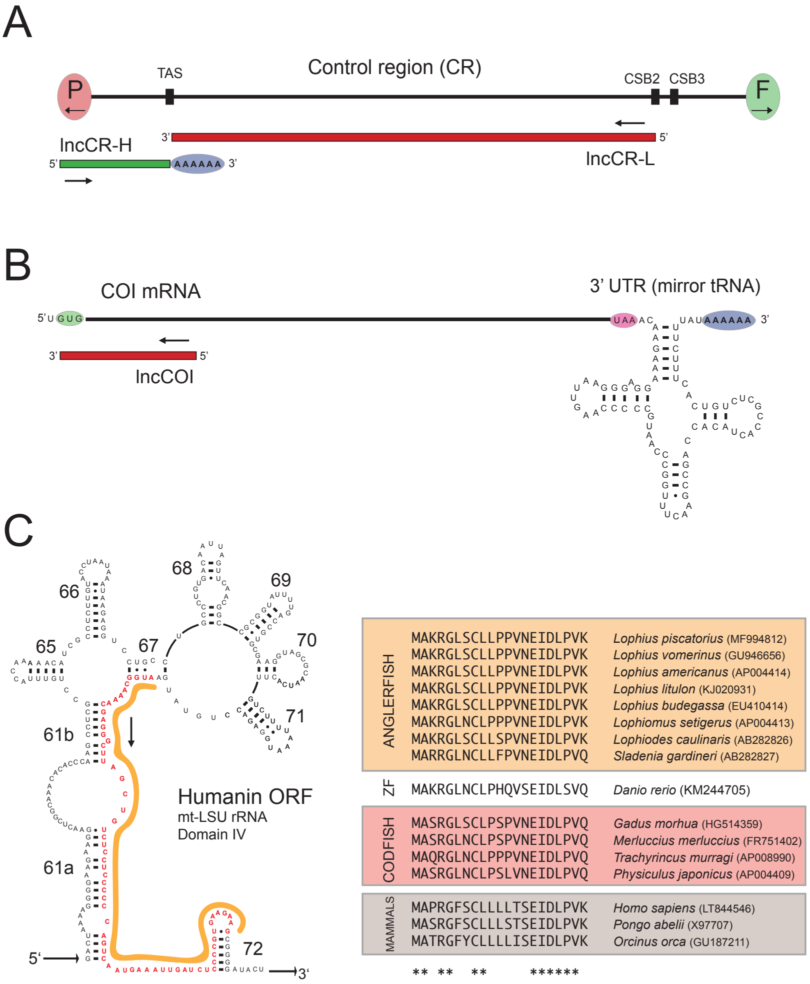

A
C1164d
A3341G
G4139A
G6942A
A10889G
C10662T
A11419G
A14480G
T16262C
F
V
HN
L1
I
M
W
D
K
G
R
S2
H L2
T
mtSSU
mtLSU
ND1
ND2
COI
COII
A8
A6
COIII
ND3
ND4L
ND4
ND5
ND6
Cyt B
CR
Q
A C
N Y
S1
E
P
Humanin
lncCOI
lncCR-H
lncCR-L
B
Expression values (mean coverage)
100
200
300
400
500
600
700
800
900
1000
1100
1200
(4897)
mt-SSU
(10331)
mt-LSU
(40)
ND1
(56)
ND2
(23)
ND3
(168)
ND4L
ND4
(52)
ND5
(82)
ND6
(247)
CytB
(545)
COI
(1009)
COII
(287)
COIII
(46)
A8
A6
(1226)
lncCOI
(2414)
lncCR-L
(192)
lncCR-H
I
I
I
I
I
I
III
IV
IV
IV
V
rRNA
mRNA
lncRNA

Figure 1



ANGLERFISH	MAKRGLSCLLPPVNEIDLPVK	*Lophius piscatorius* (MF994812)
	MAKRGLSCLLPPVNEIDLPVK	*Lophius vomerinus* (GU946656)
	MAKRGLSCLLPPVNEIDLPVK	*Lophius americanus* (AP004414)
	MAKRGLSCLLPPVNEIDLPVK	*Lophius litulon* (KJ020931)
	MAKRGLSCLLPPVNEIDLPVK	*Lophius budegassa* (EU410414)
	MAKRGLNCLPPPVNEIDLPVK	*Lophiomus setigerus* (AP004413)
	MAKRGLSCLLSPVNEIDLPVK	*Lophiodes caulinaris* (AB282826)
	MARRGLNCLLFPVNEIDLPVQ	*Sladenia gardineri* (AB282827)
ZF	MAKRGLNCLPHQVSEIDLSVQ	*Danio rerio* (KM244705)
CODFISH	MASRGLSCLPSPVNEIDLPVQ	*Gadus morhua* (HG514359)
	MASRGLNCLPSPVNEIDLPVQ	*Merluccius merluccius* (FR751402)
	MAQRGLNCLPPPVNEIDLPVQ	*Trachyrincus murragi* (AP008990)
	MASRGLNCLPSLVNEIDLPVQ	*Physiculus japonicus* (AP004409)
MAMMALS	MAPRGFSCLLLLTSEIDLPVK	*Homo sapiens* (LT844546)
	MASRGFSCLLLSTSEIDLPVK	*Pongo abelii* (X97707)
	MATRGFYCLLLLISEIDLPVK	*Orcinus orca* (GU187211)
	** ** ** ******	

Figure 2

Additional file 1: Table S1. Polymorphic sites in the mitochondrial genome of *L. piscatorius* specimens BF1 and BF2

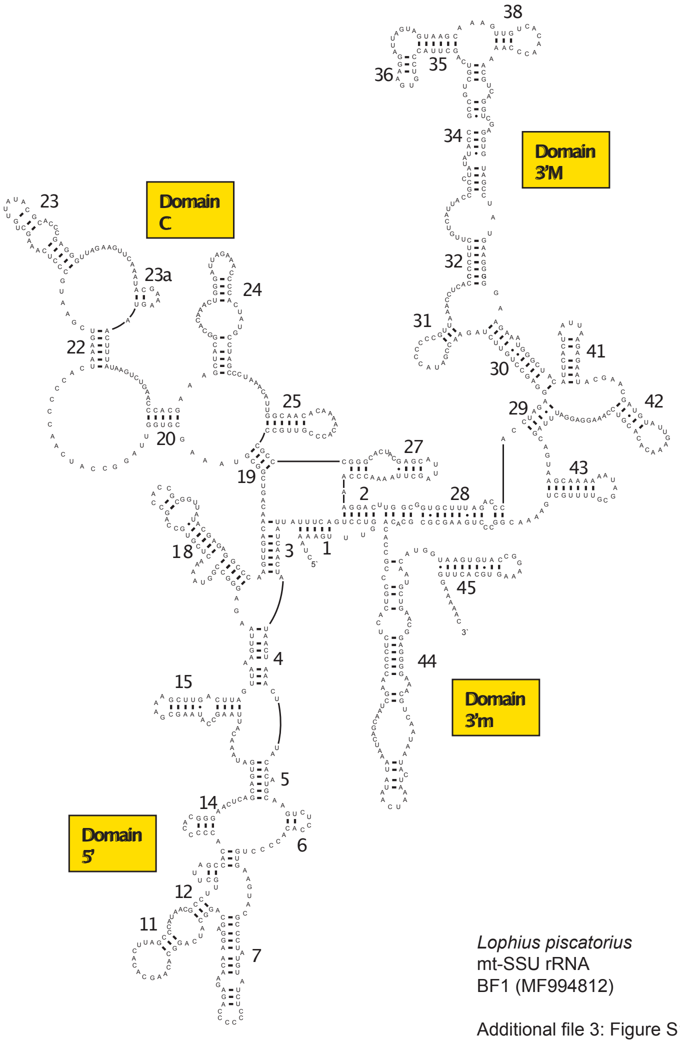
Polymorphic site [1]	Gene region [2]	Codon	Amino acid
C1164d	mtLSU	-	-
A1165C (2.6%) [3]	mtLSU	-	-
T1168A (1.9%) [3]	mtLSU	-	-
A2678G (1.3%) [3]	mtLSU	-	-
G2679A (1.3%) [3]	mtLSU	-	-
A3341G	ND1	AGC to GGC	S to G
G4139A	ND2	AGC to AAC	S to N
G6942A	COI	GAG to GAA	E (synonymous)
G9120C (1.4%) [3]	COIII	TGG to TCG	W to S
C10662T	ND4	ACT to ATT	T to I
A10889G	ND4	AAC to GAC	N to D
A1141G	ND4	TGA to TGG	W (synonymous)
A14480G	CytB	GCA to GCG	A (synonymous)
T16158C (1.0%) [3]	CR	-	-
C16160T (1.3%) [3]	CR	-	-
T16262C	CR	-	-

Notes: [1] Positions are according to the reference sequence in specimen BF1 (MF994812). [2] mtLSU, mitochondrial large subunit ribosomal RNA; ND1, NADH dehydrogenase subunit 1; COI, cytochrome c oxidase subunit I; COIII, cytochrome c oxidase subunit III; ND4, NADH dehydrogenase subunit 4; CytB, cytochrome B. [3] Low-level substitution heteroplasmy detected in the BF1 specimen by SOLiD ligation sequencing at 2227 times mtDNA coverage [Dubin et al. 2017]. Reference: Dubin A, Jørgensen TE, Jakt LM, Moum T, Johansen SD: The mitochondrial genome of the European Anglerfish Lophius piscatorius express low-level substitution heteroplasmy. Ann Mar Biol Res. 2017;4:1019.

Additional file 2: Table S2. 5' and 3' sequence features of *L. piscatorius* mitochondrial mRNAs derived from RNA-seq reads

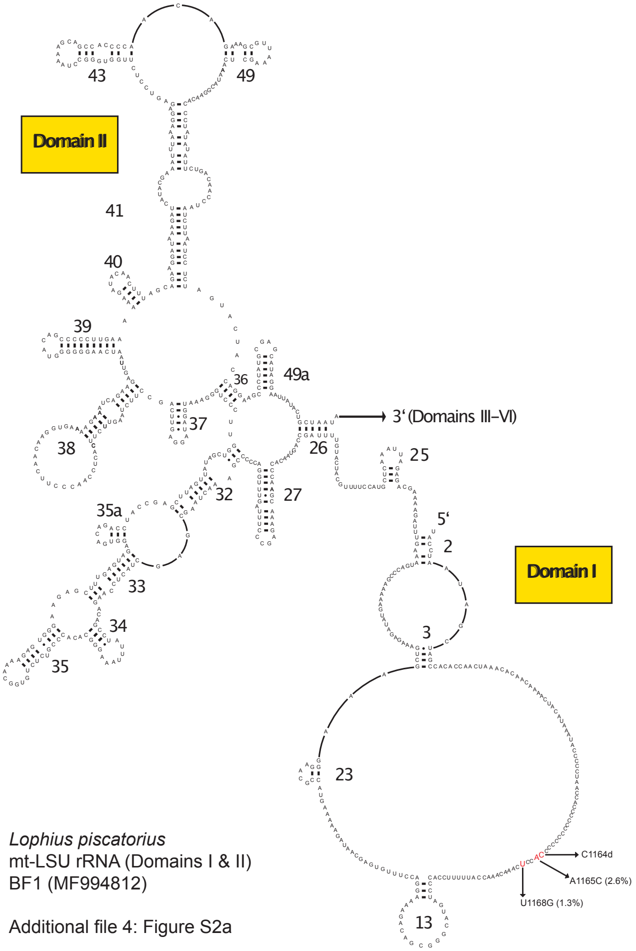
mRNA [1]	5' end [2]	3' end [3]
ND1	**AUG** AUC UCA	CAA AUA **UAA** aaaaaaaaaa
ND2	**AUG** AAC CCC	GCC CUU **Uaa** aaaaaaaa
ND3	**AUG** AAC UUA	GCC GAA **Uaa** aaaaaaaa
ND4L/ND4	**AUG** ACC CCC	AGC AUA **Uaa** aaaaaaaa
ND5	**AUG** CAC CCU	n.d.
ND6	**AUG** ACU UAU	n.d.
CytB	n.d.	AAA UUA **UAG** aaaaaaaaaa
COI	U **GUG** GCA AUC	ACC CGU **UAA** ACAAGAAAGGAGGGAAUUGAACCCCCGUAACCCGGUUUCAAGC CGACCACAUCACCGCUCUGUCACUUUCUUUAUaaaaaaaaaa
COII	n.d.	GAC GCU **Uaa** aaaaaaaa
COIII	n.d.	GGC UCA **UAa** aaaaaaaaaa
A8/A6	U **AUG** CCC CAA	AAC GUA **UAA** aaaaaaaaaa

Notes: [1] Eleven mitochondrial mRNAs: ND1-6, NADH dehydrogenase subunits 1-6 mRNAs; CytB, cytochrome B subunit mRNA; COI-III, cytochrome c oxidase subunits I-III; A8/A6, ATPase 8/6 mRNA. [2] The 5' sequence contains the three first codons. Start codon (AUG/GUG) is under-lined and in bold. Defined 5' end not detected (n.d.) due to low coverage. [3] The 3' sequence contains the three last codons. Termination codon (UAG/UAA) is under-lined and in bold. Non-template adenosines are indicated by lower case 'a'. Defined polyA-tail not detected (n.d.).



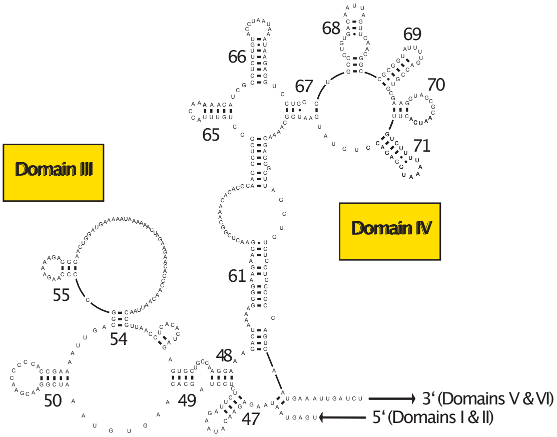
Lophius piscatorius
mt-SSU rRNA
BF1 (MF994812)

Additional file 3: Figure S1



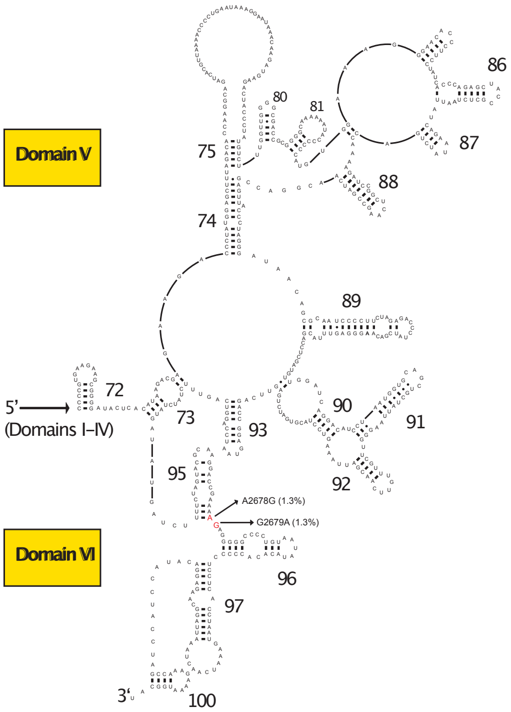
Lophius piscatorius
mt-LSU rRNA (Domains I & II)
BF1 (MF994812)

Additional file 4: Figure S2a



Lophius piscatorius
mt-LSU rRNA (Domains III & IV)
BF1 (MF994812)

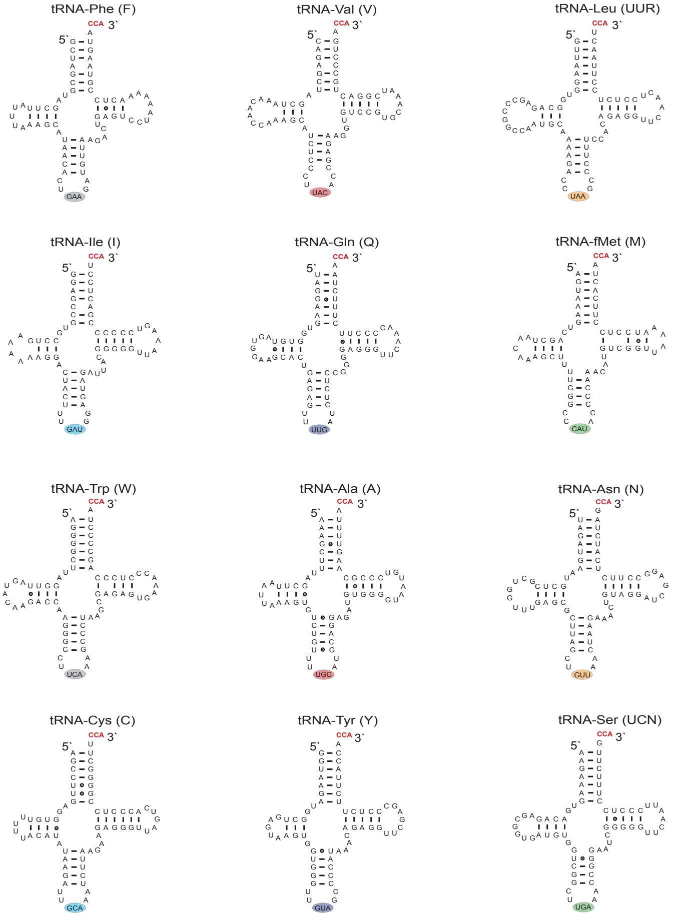
Additional file 4: Figure S2b



Lophius piscatorius
mt-LSU rRNA (Domains V & VI)
BF1 (MF994812)

Additional file 4: Figure S2c

Additional file 5: Figure S3a



tRNA-Phe (F)

tRNA-Val (V)

tRNA-Leu (UUR)

tRNA-Ile (I)

tRNA-Gln (Q)

tRNA-fMet (M)

tRNA-Trp (W)

tRNA-Ala (A)

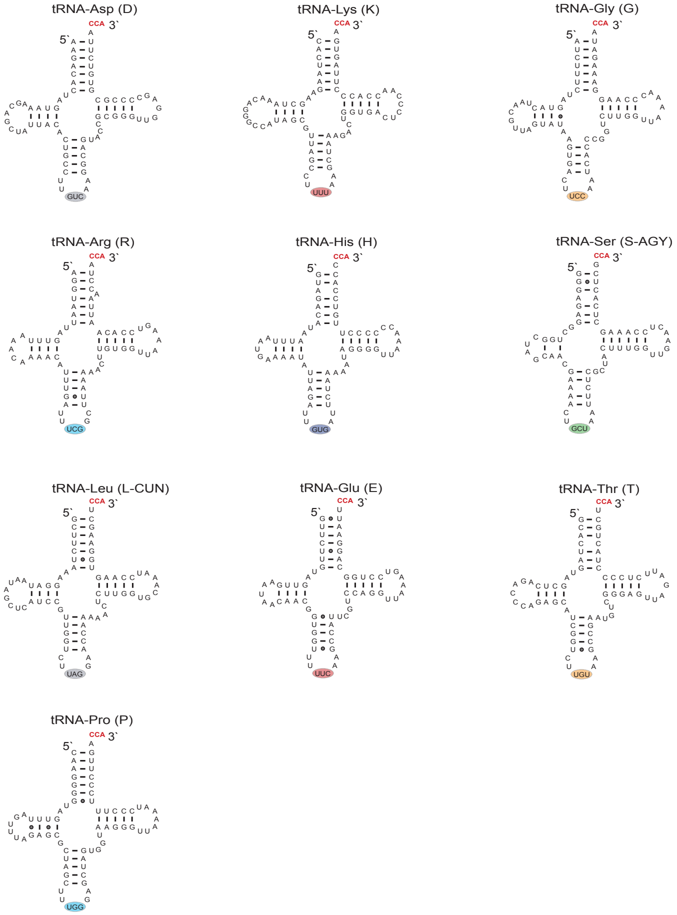
tRNA-Asn (N)

tRNA-Cys (C)

tRNA-Tyr (Y)

tRNA-Ser (UCN)

Additional file 5: Figure S3b



tRNA-Asp (D)

tRNA-Lys (K)

tRNA-Gly (G)

tRNA-Arg (R)

tRNA-His (H)

tRNA-Ser (S-AGY)

tRNA-Leu (L-CUN)

tRNA-Glu (E)

tRNA-Thr (T)

tRNA-Pro (P)

Paper III

BIOLOGY LETTERS

royalsocietypublishing.org/journal/rsbl

Research

Cite this article: Dubin A, Jørgensen TE, Moum T, Johansen SD, Jakt LM. 2019 Complete loss of the MHC II pathway in an anglerfish, *Lophius piscatorius*. *Biol. Lett.* 20190594.
http://dx.doi.org/10.1098/rsbl.2019.0594

Received: 9 August 2019
Accepted: 13 September 2019

Subject Areas:
evolution

Keywords:
major compatibility complex, teleost, anglerfish, *Lophius*, gene loss, genome

Author for correspondence:
Lars Martin Jakt
e-mail: lars.m.jakt@nord.no

Electronic supplementary material is available online at rs.figshare.com.

THE ROYAL SOCIETY
PUBLISHING

Evolutionary biology

Complete loss of the MHC II pathway in an anglerfish, *Lophius piscatorius*

Arseny Dubin, Tor Erik Jørgensen, Truls Moum, Steinar Daae Johansen and Lars Martin Jakt

Genomics group, Faculty of Biosciences and Aquaculture, Nord University, 8049 Bodø, Norway

AD, 0000-0001-8360-5475

Genome studies in fish provide evidence for the adaptability of the vertebrate immune system, revealing alternative immune strategies. The reported absence of the major compatibility complex (MHC) class II pathway components in certain species of pipefish (genus *Syngnathus*) and cod-like fishes (order Gadiformes) is of particular interest. The MHC II pathway is responsible for immunization and defence against extracellular threats through the presentation of exogenous peptides to T helper cells. Here, we demonstrate the absence of all genes encoding MHC II components (CD4, CD74 A/B, and both classical and non-classical MHC II α/β) in the genome of an anglerfish, *Lophius piscatorius*, indicating loss of the MHC II pathway. By contrast, it has previously been reported that another anglerfish, *Antennarius striatus*, retains all MHC II genes, placing the loss of MHC II in the *Lophius* clade to their most recent common ancestor. In the three taxa where MHC II loss has occurred, the gene loss has been restricted to four or five core MHC II components, suggesting that, in teleosts, only these genes have functions that are restricted to the MHC II pathway.

1. Introduction

The vertebrate adaptive immune system that generates diversity through genetic recombination appears to have evolved in the common ancestor of the jawed vertebrates (gnathostomes) [1]. Although this system increased in complexity with gnathostome evolution, it is thought that the acquisition of all required cellular processes, tissues and genes happened relatively quickly as most components are present across all jawed vertebrates [1]. T-cell receptors (TCR), B-cell receptors (BCR) and the Major Histocompatibility Complex (MHC) classes I and II, are all present throughout the gnathostome lineages, from the Chondrichthyes to terrestrial vertebrates [1]. Although the specific sites of haematopoiesis vary, homologous tissues and organs including the thymus and spleen are also present across the gnathostomes [1–4].

An intact adaptive immune system has been found in almost all vertebrate species that have had their genomes sequenced, but recent work has demonstrated the loss of components of the adaptive immune system in the elephant shark, pipefish, coelacanth and the entire Gadiformes order [5–10]. These observations demonstrate an unexpected plasticity of adaptive immunity.

The teleost order Lophiiformes (Anglerfishes) harbours at least 321 living species, approximately half of which express some degree of sexual parasitism [11]. In these species, males attach to the females either temporarily or permanently. In extreme cases, this leads to fusion of male and female circulatory systems [12]. Why this fusion does not result in tissue rejection is unknown, but suggests a specialized adaptive immune system. Phylogenetic inference based on sequencing data and morphology have concluded that male sexual

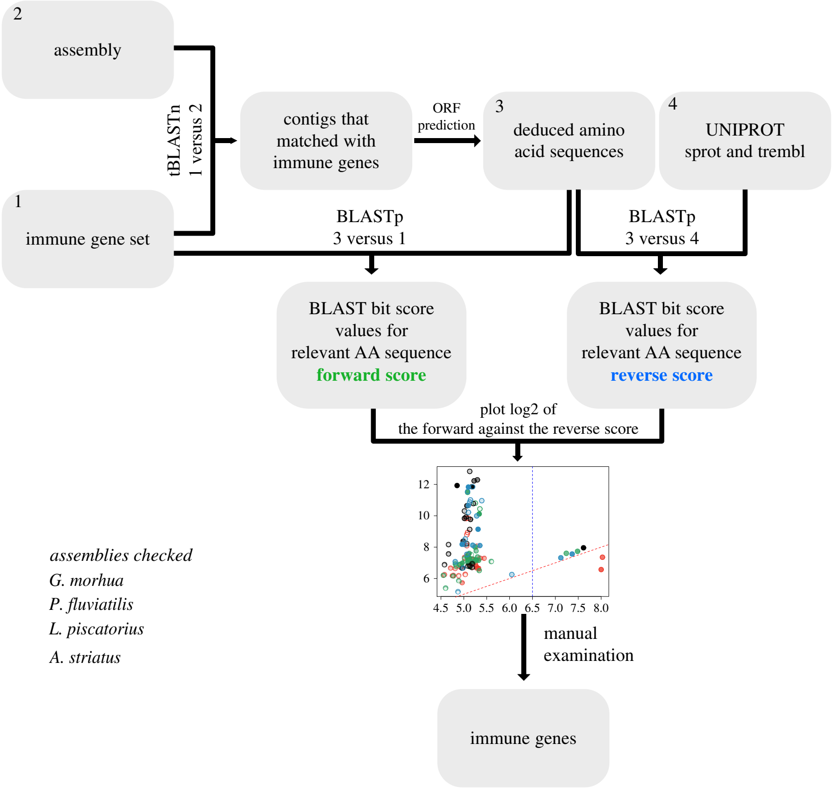


Figure 1. Outline of the gene mining process. Sequence inputs and outputs are shown as boxes with processes indicated by connecting arrows. BLAST inputs are numbered to indicate what was used as a query and subject (denoted as *query(number)* versus *subject(number)*). (Online version in colour.)

parasitism within Lophiiformes must have multiple origins [11,13,14], suggesting a common selective pressure or a shared genetic predisposition.

Here, we present two independently obtained draft genomes of an anglerfish, *Lophius piscatorius*, and show that it has lost all components of the MHC II arm of the adaptive immune system. The MHC II pathway is known to be involved in allogenic rejection [15] and our observations suggest that loss of MHC II may have contributed to the immune tolerance observed in sexually parasitic anglerfish species.

2. Material and methods

(a) Sample collection, DNA isolation and sequencing

Samples from two *L. piscatorius* individuals (referred to as BF1 and BF2) were collected in the Bodø coastal waters, Nordland County in collaboration with local fishermen. BF1 skeletal muscle and BF2 kidney were used for subsequent total DNA isolation, library preparation and sequencing. BF1 total DNA sequencing was performed using Illumina MiSeq and SOLiD 5500 technologies (sequence depth: 24×). BF2 total DNA was sequenced by Dovetail Genomics, USA on an Illumina HiSeq X instrument (sequence depth: 150×) as a service. The Illumina libraries were 300 bp paired-end reads with 600 bp insert size for the MiSeq, and 150 bp paired-end reads with 350 bp insert size for the HiSeq.

(b) Bioinformatic analysis

The raw reads were trimmed from adapters and low-quality bases using Cutadapt [16]. Only Illumina data were used for the assemblies. Prior to assembly, overlapping read pairs were merged using FLASH (v .1.2.11) [17]. Final assemblies were constructed with SPAdes (v. 3.10.0) [18]. Basic assembly statistics were calculated with QUAST (v. 4.4.1) [19] and gene-space completeness assessed using BUSCO (v.2.0) [20] with the actinopterygii dataset (odb9).

MHC genes were identified using methods similar to those used in [10] (figure 1). Briefly, a set of adaptive immune system-related protein sequences (bait-sequences) were used to identify contigs containing potential orthologues. Genes and open reading frames (ORFs) were predicted from these contigs and aligned both to the bait-set and sequences within the UniProt database to separate orthologues from non-orthologous genes containing homologous sequences. The resulting alignment scores were visualized and identities of candidate orthologues manually confirmed by inspection of alignments and annotations.

We performed this analysis on our *L. piscatorius* assemblies as well as on assemblies of *Antennarius striatus*, *Gadus morhua* and *Perca fluviatilis* [10]. If a gene was not identified in *L. piscatorius*, the unassembled reads (SOLiD and Illumina) were searched using tBLASTn. Matching reads were reassembled (CLC GW v. 11, QIAGEN, Aarhus, Denmark), and verified by reciprocal BLASTn against NCBI nr.

More detailed descriptions are provided in the electronic supplementary material.

3. Results

(a) Genome assembly

The resulting *L. piscatorius* assemblies contained 664 (BF1) and 724 (BF2) megabases with N50 values of 6.9 kb and 108 kb, respectively. We used the BUSCO [20] actinopterygii set of 4584 conserved genes to estimate the gene-space completeness of these assemblies. We could detect at least 75% of these genes in both our assemblies (complete and fragmented), with 91.5% of complete genes identified in the BF2 assembly (electronic supplementary material, figure S1).

The gene space completeness of our assemblies is thus similar to that obtained for the *A. striatus* assembly (66.5% complete and 15.8% fragmented, electronic supplementary material, figure S1). Hence, our assemblies are comparable to or better than assemblies in [10] in terms of continuity, coverage and gene-space completeness (electronic supplementary material, figure S1).

(b) Adaptive immune system genes in *L. piscatorius*

We used tBLASTn with a set of adaptive immune system genes to identify orthologous genes in *L. piscatorius* as well as in species previously reported to either have (*A. striatus*, *P. fluviatilis*) or lack (*G. morhua*) genes coding for MHC II components. Candidate orthologues were readily observed for all MHC I genes in all species. By contrast, we were unable to identify genes coding for CD4, CD74 A/B, MHC II α/β in either *L. piscatorius* or *G. morhua* assemblies (figure 2 and table 1). We repeated this analysis using an extended bait set including the non-classical MHC II α/β lineages [21]; this too failed to find any candidate orthologues in *L. piscatorius*. Similarly to the MHC I components we were also able to clearly identify orthologues of 22 out of 23 additional genes that have functions in the adaptive immune system (electronic supplementary material, figure S6, ST2).

To confirm the absence of MHC II orthologues in *L. piscatorius* we also searched for short sequences in the unassembled reads that could be aligned with the missing genes. Using tBLASTn, we identified 18 and 62 reads from BF1 and BF2, respectively, that aligned with an MHC II β subunit. To locate the position of these potential MHC II sequences, the matching reads were assembled into contigs and mapped back to the original assemblies. This identified a region of approximately 300–480 bp in length present in both assemblies. When translated, the predicted reading frame was interrupted by multiple stop codons (electronic supplementary material, figure S2), indicating that the fragment represents a remnant of an MHC II β chain gene. Hence, we conclude that the MHC II pathway is absent in *L. piscatorius*.

To confirm the presence of genes syntenic to CD4 and CD74 in *L. piscatorius*, we identified contigs containing these genes [5] and aligned them to the stickleback (*Gasterosteus aculeatus*) loci (electronic supplementary material, figure S7). All highly conserved genes were identified in either a single (CD74) or three (CD4) contigs and the gene predictions lying within these contigs aligned both in terms of direction and order. CD74 in *L. piscatorius* seems to have been lost through a deletion of a region lying between ndst1a and SCL35A4 that has removed both CD74 and almost all intergenic space. For CD4, we were unable to identify a contig spanning the expected CD4 position; nevertheless, our analysis confirms the presence of the expected syntenic genes within our assembly.

(c) MHC II in *A. striatus*

Antennarius striatus has been reported to contain both MHC I and MHC II pathway genes [10]. Since both *A. striatus* and *L. piscatorius* are members of the Lophiiformes order, we considered the possibility that the identification of *A. striatus* MHC II orthologues could have resulted from cross-contamination of the sample. Although we did observe the presence of cross-contaminating mitochondrial sequences from distantly related teleost taxa, the relative sequencing depth of contaminant and *A. striatus* sequences, combined with the unimodal depth distribution preclude the possibility that the MHC II sequences were derived from contaminant DNA fragments (electronic supplementary material, figure S4). Our observations thus confirm the presence of MHC II, while at the same time highlighting the potential of cross-contamination leading to confounding results.

4. Evolutionary considerations

Of the 81 teleosts examined so far, only members of the Gadiformes order (27 species sequenced), pipefishes (*Syngnathus typhle and S. scovelli*) and now *L. piscatorius* lack a functional MHC II pathway [5,6,9,10]. Notably, all these species lack CD4 and both MHC II α and β chains (classical and non-classical molecules) In addition, CD74 A/B has been completely lost in *L. piscatorius* and the Gadiformes order. Nevertheless, all contain a complete set of MHC I pathway components [6,9,10] (figure 2; electronic supplementary material, figure S6). In contrast to terrestrial vertebrates, teleost MHC II α and β often occupy multiple loci on different chromosomes [22]. This means that loss of the complete MHC II pathway requires multiple independent gene deletions; e.g. the loss of MHC II in *A. striatus* would require around 10 deletions (table 1). This suggests that loss of any critical MHC II pathway component leads to the loss of remaining core parts and argues that genes lost in these species are unlikely to have functions outside of MHC II in teleosts. **Q1**

By contrast, genes such as CIITA, which is conventionally thought to be a specific activator of MHC II gene expression [23] is likely to have roles outside the MHC II system as in the absence of neo-functionalization, it would have no function after MHC II loss.

Hence, our results are consistent with reports indicating an additional role of CIITA in the regulation of MHC I expression [24–26].

Sexual parasitism in anglerfish suggests some form of specialized immune system allowing for allogenic tolerance between fused individuals. Although CD8+ cytotoxic lymphocytes are thought to be the primary effectors of allogenic rejection, it is clear that MHC II components have both enabling and effector functions [15], and a long line of publications show that repression of MHC II components can contribute to immune tolerance or alleviate immune rejection [27–30]. Hence, it is tempting to speculate that the loss of MHC II in the Lophiodei suborder is not restricted to *Lophius* species and has played an enabling role in the development of sexual parasitism.

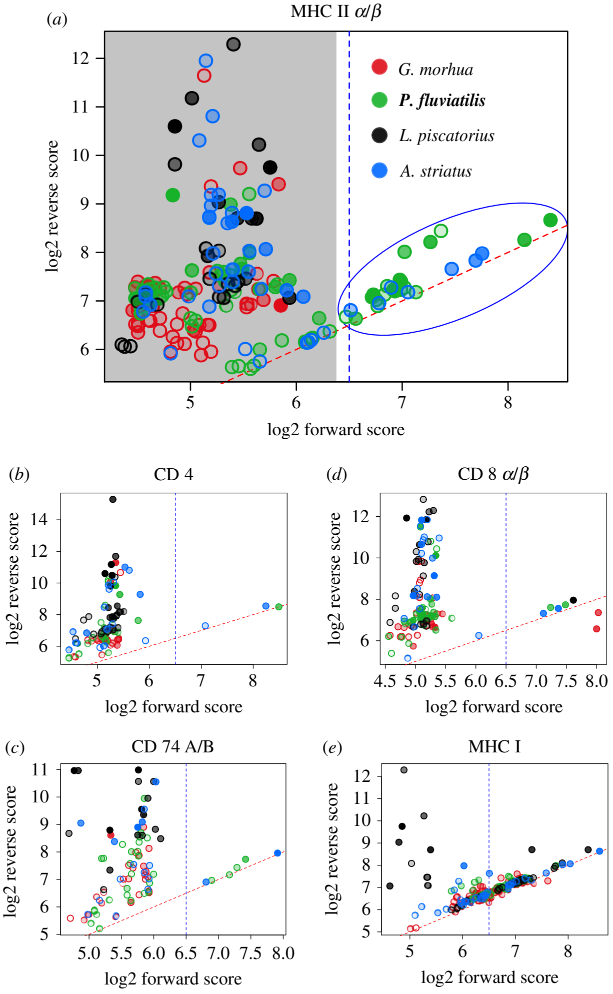


Figure 2. Identification of MHC I and MHC II pathway orthologues. Illustration of identification criteria (*a*). The alignment scores of putative orthologues against the initial bait set (forward score, *X*-axis) plotted versus scores against the UniProt database (reverse score, *Y*-axis). Grey shading indicates hits that were dismissed from further analysis, as they align better to genes not belonging to the MHC pathways. Alignments of short but highly conserved gene fragments can have high alignment scores without indicating functional orthologues; such points are shown with a high transparency by scaling the fill transparency by the ratio of alignment to subject (UniProt) length. Orthologues should have values closer to $Y = X$ indicated by the dashed red line. Candidate orthologues shown within the blue ellipse appear as outliers, and can be identified by a forward score threshold indicated by the dashed blue line. Plots for the MHC II components (*a*–*c*) lack candidate orthologues for *Lophius piscatorius* (black) and *Gadus morhua* (red), whereas candidate orthologues are evident in all species for the MHC I genes (*d*,*e*). Results for the α/β chains of CD8 and MHC II, and the CD 74 A/B genes are shown combined. (Online version in colour.)

Table 1. Number of candidate orthologues identified after forward/reverse screening (Methods) and manual inspection of the plots (figure 2). Numbers in brackets indicate individual hits after the forward score threshold was applied, but before manual examination of UniProt IDs identified unrelated genes.

gene	*Gadus morhua*	*Perca fluviatilis*	*Lophius piscatorius*	*Antennarius striatus*
CD4	**0**	1	**0**	2
CD74 A/B	**0**	4	**0**	2
MHC II α/β	**0**	21 (22)	**0**	6 (7)
CD8 α/β	2	2	1*	2
MHC I	49	34 (35)	18 (19)	12 (13)

*Predicted sequence appears as a fusion protein of α and β chains.

Most Lophiiformes phylogenies place the Lophioidei suborder at the most basal position in the anglerfish taxonomy, followed by Antennarioidei [11,14,31,32]. If Lophioidei is basal, then MHC II loss is likely to be specific to the Lophioidei suborder since *A. striatus* clearly possesses all MHC II components. That would also mean that our observations are unlikely to be relevant to sexual parasitism in other anglerfish clades. However, inferences of higher-order taxonomies are still fraught with difficulty, exemplified by the fact that the phylogenetic topology of the taxa involved based on mitochondrial DNA is subject to change depending on the choice of outgroup [11,13] (electronic supplementary material, figure S5). Exploring the presence and absence of MHC II genes in other anglerfish species can thus provide a test of the conventional phylogeny, as well as the likelihood of MHC II loss being one of the enabling adaptations preventing intra-species tissue rejection.

5. Conclusion

The classical MHC II components are responsible for the presentation of exogenous peptides to T helper cells and constitute an important part of the gnathostome adaptive immune system. Here, we report two draft genome assemblies of *L. piscatorius* and demonstrate a complete loss of the classical MHC II pathway in this species. The finding of a third taxon that lacks MHC II function corroborate the dispensability of MHC II in teleosts, and suggests that the genes lost in all three clades have no function outside of the MHC II.

Data accessibility. Sequences have been deposited to public database Dryad Digital Repository: https://doi.org/10.5061/dryad.4dc007q [33] and GenBank nos. (SRR8809741 and SRR8801493).

Authors' contributions. A.D., L.M.J. and S.D.J. designed the study, collected samples, analysed the data and wrote the first draft of the manuscript. T.M. collected samples. T.E.J. produced sequencing data. All authors contributed significantly to the manuscript, approved the final version and agree to be held accountable for the content therein.

Competing interests. We declare we have no competing interests.

Funding. This study was supported by the Nord University and a matching funds grant by Dovetail Genomics, USA.

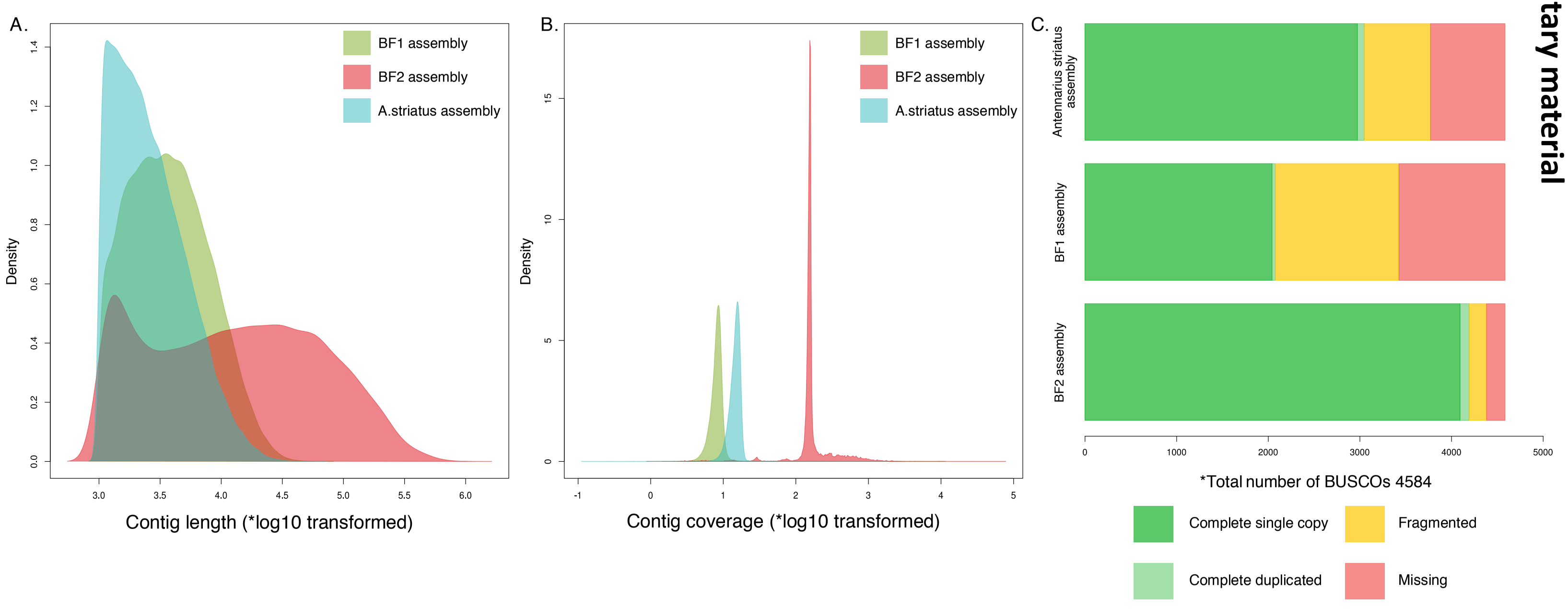
Acknowledgements. We thank local fishermen, in particular, Harald Lorentzen of Bofisk for willingness to collaborate and getting the fish to us fresh. Thanks to prof. Tor Brynjar Stuge, University of Tromsø, for initial discussions.

References

1. Flajnik MF. 2018 A cold-blooded view of adaptive immunity. *Nat. Rev. Immunol.* 18, 438–453. (doi:10.1038/s41577-018-0003-9)
2. Orkin SH, Zon LI. 2008 Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644. (doi:10.1016/j.cell.2008.01.025)
3. Press CM, Evensen Ø. 1999 The morphology of the immune system in teleost fishes. *Fish Shellfish Immun.* **9**, 309–318. (doi:10.1006/fsim.1998.0181)
4. Neely HR, Flajnik MF. 2016 Emergence and evolution of secondary lymphoid organs. *Annu. Rev. Cell Dev. Biol.* **32**, 693–711. (doi:10.1146/annurev-cellbio-111315-125306)
5. Star B *et al*. 2011 The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207–210. (doi:10.1038/nature10342)
6. Haase D, Roth O, Kalbe M, Schmiedeskamp G, Scharsack JP, Rosenstiel P, Reusch TBH. 2013 Absence of major histocompatibility complex class II mediated immunity in pipefish, *Syngnathus typhle*: evidence from deep transcriptome sequencing. *Biol. Lett.* **9**, 20130044. (doi:10.1098/rsbl.2013.0044)
7. Amemiya CT *et al*. 2013 The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316. (doi:10.1038/nature12027)
8. Venkatesh B *et al*. 2014 Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179. (doi:10.1038/nature12826)
9. Small CM, Bassham S, Catchen J, Amores A, Fuiten AM, Brown RS, Jones AG, Cresko WA. 2016 The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biol.* **17**, 258–280. (doi:10.1186/s13059-016-1126-6)
10. Malmstrøm M *et al*. 2016 Evolution of the immune system influences speciation rates in teleost fishes. *Nat. Genet.* **48**, 1204–1210. (doi:10.1038/ng.3645)
11. Miya M *et al*. 2010 Evolutionary history of anglerfishes (Teleostei: Lophiiformes): a mitogenomic perspective. *BMC Evol. Biol.* **10**, 58–84. (doi:10.1186/1471-2148-10-58)
12. Pietsch TW. 2005 Dimorphism, parasitism, and sex revisited: modes of reproduction among deep-sea ceratioid anglerfishes (Teleostei: Lophiiformes). *Ichthyol. Res.* **52**, 207–236. (doi:10.1007/s10228-005-0286-2)
13. Shedlock AM, Pietsch TW, Haygood MG, Bentzen P, Hasegawa M. 2004. Molecular systematics and life history evolution of anglerfishes (Teleostei: Lophiiformes): evidence from mitochondrial DNA. *Steenstrupia* **28**, 129–144.
14. Pietsch TW, Orr JW. 2007 Phylogenetic relationships of deep-sea anglerfishes of the suborder Ceratioidei (Teleostei: Lophiiformes) based on morphology. *Copeia* **2007**, 1–34. (doi:10.1643/0045-8511(2007)7[1:PRODA0]2.0.CO;2)
15. Rosenberg AS, Singer A. 1992 Cellular basis of skin allograft rejection: an *in vivo* model of immune-mediated tissue destruction. *Annu. Rev. Immunol.* **10**, 333–360. (doi:10.1146/annurev.iy.10.040192.002001)
16. Martin M. 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12. (doi:10.14806/ej.17.1.200)
17. Magoc T, Salzberg SL. 2011 FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963. (doi:10.1093/bioinformatics/btr507)
18. Bankevich A *et al*. 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. (doi:10.1089/cmb.2012.0021)
19. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013 QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. (doi:10.1093/bioinformatics/btt086)
20. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018 BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548. (doi:10.1093/molbev/msx319)
21. Dijkstra JM, Grimholt U, Leong J, Koop BF, Hashimoto K. 2013 Comprehensive analysis of MHC class II genes in teleost fish genomes reveals dispensability of the peptide-loading DM system in a large part of vertebrates. *BMC Evol. Biol.* **13**, 260–273. (doi:10.1186/1471-2148-13-260)
22. Grimholt U. 2016 MHC and evolution in teleosts. *Biology* **5**, 6–25. (doi:10.3390/biology5010006)
23. Neefjes J, Jongsma MLM, Paul P, Bakke O. 2011 Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836. (doi:10.1038/nri3084)
24. Martin BK, Chin K-C, Olsen JC, Skinner CA, Dey A, Ozato K, Ting JP-Y. 1997 Induction of MHC Class I expression by the MHC Class II transactivator CIITA. *Immunity* **6**, 591–600. (doi:10.1016/S1074-7613(00)80347-7)
25. Howcroft TK, Raval A, Weissman JD, Gegonne A, Singer DS. 2003 Distinct transcriptional pathways regulate basal and activated major histocompatibility complex class I expression. *Mol. Cell. Biol.* **23**, 3377–3391. (doi:10.1128/MCB.23.10.3377-3391.2003)

26. Gobin SJP, Peijnenburg A, Keijsers V, van den Elsen PJ. 1997 Site α is crucial for two routes of IFNγ-induced MHC Class I transactivation: the ISRE-mediated route and a novel pathway involving CIITA. *Immunity* **6**, 601–611. (doi:10.1016/S1074-7613(00)80348-9)
27. Shizuru J, Gregory A, Chao C, Fathman C. 1987 Islet allograft survival after a single course of treatment of recipient with antibody to L3T4. *Science* **237**, 278–280. (doi:10.1126/science.2955518)
28. Sato Y, Bolzenius JK, Eteleeb AM, Su X, Maher CA, Sehn JK, Arora VK. 2018 CD4$^+$ T cells induce rejection of urothelial tumors after immune checkpoint blockade. *JCI Insight* **3**, e121062. (doi:10.1172/jci.insight.121062)
29. Borges TJ *et al*. 2018 March1-dependent modulation of donor MHC II on CD103$^+$ dendritic cells mitigates alloimmunity. *Nat. Commun.* **9**, 3482–3497. (doi:10.1038/s41467-018-05572-z)
30. Abrahimi P, Qin L, Chang WG, Bothwell ALM, Tellides G, Saltzman WM, Pober JS. 2016 Blocking MHC class II on human endothelium mitigates acute rejection. *JCI Insight* **1**, e85293. (doi:10.1172/jci.insight.85293)
31. Pietsch TW. 1984 Lophiiformes: Development and relationships. In *Ontogeny and systematics of fishes: based on an international symposium dedicated to the memory of elbert halvor ahlstrom* (eds HG Moser, WJ Richards, DM Cohen, MP Fahay, AW Kendall, SL Richardson), pp. 320–325. New York, NY: American Society of Ichthyologists and Herpetologists.
32. Pietsch TW. 1981 The osteology and relationships of the anglerfish genus Tetrabrachium with comments on lophiiform classification. *Fish. Bull.* **79**, 387–419.
33. Dubin A, Jørgensen TE, Moum T, Johansen SD and Jakt LM. 2019 Data from: Complete loss of the MHC II pathway in an anglerfish, Lophius piscatorius. Dryad Digital Repository. (doi:10.5061/dryad.4dc007q)

S1. Assembly statistics



S2. Sequences from the BF1 and BF2 assemblies that could be aligned to MHC II

A.

BF1 MHC class II fragment. Frame -3.

```
ctatcctaaacaaatcaaatgggcaggatgtgacctcagctgtcatttttttctgaagtgttg
   I  L  N  K  S  N  G  Q  D  V  T  S  A  V  I  F  S  E  V  L
cctgatgtggaatagtactgccagattgagtcttacctggagtacatgccaacccctgga
 P  D  V  E  *  Y  C  Q  I  E  S  Y  L  E  Y  M  P  T  P  G
gcataaattacatgcatggataaacacctcagcctaccaaaacaagtgctttgagtctgg
 A  *  I  T  C  M  D  K  H  L  S  L  P  K  Q  V  L  *  V  W
ggcaaatttgcaattgtaatacaggatgagactatggacgatataaaaggagattaattt
 G  K  F  A  I  V  I  Q  D  E  T  M  D  D  I  K  G  D  *  F
ccctcctcatgtctgtagcaacatctcttcctggatcagagaggattataattattggaa
 P  S  S  C  L  *  Q  H  L  F  L  D  Q  R  G  L  *  L  L  E
caatttaactagtatttatttttgtcttagttagcactggcttcatctaatacaagtaaa
 Q  F  N  *  Y  L  F  L  S  *  L  A  L  A  S  S  N  T  S  K
atgcctggcgcagtcatcatataactgcaacttaagaaaaatattgttcttgtgctcatt
 M  P  G  A  V  I  I  *  L  Q  L  K  K  N  I  V  L  V  L  I
tttaacagtataagttgtgattcctgtggaatatgaccttgcagctgatgacaaaattat
 F  N  S  I  S  C  D  S  C  G  I  *  P  C  S  *  *  Q  N  Y
tatca
 Y
```

BF2 MHC class II fragment. Frame +3.

```
atctctttgtgctcctgcttattttcttcatctctcttgctatcccaatacttatttgctaa
   L  F  V  L  L  L  I  F  F  I  S  L  A  I  P  I  L  I  C  *
cttaatccatgcttgtgtgcaaagcctataaatactatcctaaacaaatcaaatgggcag
 L  N  P  C  L  C  A  K  P  I  N  T  I  L  N  K  S  N  G  Q
gatgtgacctcagctgtcatttttttctgaagtgttgcctgatgtggaatagtactgccag
 D  V  T  S  A  V  I  F  S  E  V  L  P  D  V  E  *  Y  C  Q
attgagtcttacctggagtacatgccaacccctggagcataaattacatgcatggataaa
 I  E  S  Y  L  E  Y  M  P  T  P  G  A  *  I  T  C  M  D  K
cacctcagcctaccaaaacaagtgctttgagtctggggcaaatttgcaattgtaatacag
 H  L  S  L  P  K  Q  V  L  *  V  W  G  K  F  A  I  V  I  Q
ga
```

B.

CLUSTAL O(1.2.4) alignment (*stop codons removed)

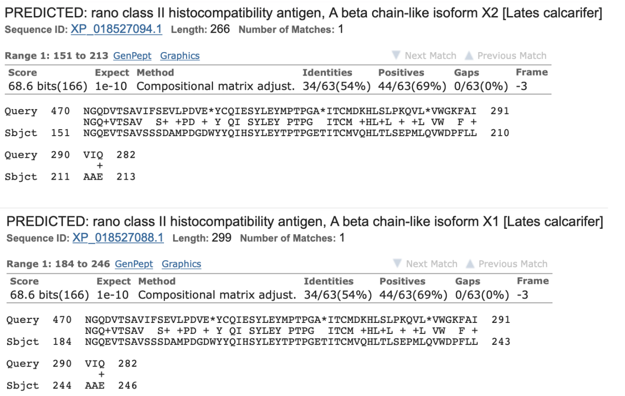
```
bf1      -------------------------------ILNKSNGQDVTSAVIFSEVLPDVEYCQIE 29
bf2      LFVLLLIFFISLAIPILICLNPCLCAKPINTILNKSNGQDVTSAVIFSEVLPDVEYCQIE 60
                                        *****************************


bf1      SYLEYMPTPGAITCMDKHLSLPKQVLVWGKFAIVIQDETMDDIKGDFPSSCLQHLFLDQR 89
bf2      SYLEYMPTPGAITCMDKHLSLPKQVLVWGKFAIVIQ------------------------ 96
         ************************************


bf1      GLLLEQFNYLFLSLALASSNTSKMPGAVIILQLKKNIVLVLIFNSISCDSCGIPCSQNYY 149
bf2      ------------------------------------------------------------ 96
```

C.

BF1 MHC class II fragment. Top 2 BLASTx hits at NCBI Nr



PREDICTED: rano class II histocompatibility antigen, A beta chain-like isoform X2 [Lates calcarifer]
Sequence ID: XP_018527094.1 Length: 266 Number of Matches: 1

Range 1: 151 to 213 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
68.6 bits(166)	1e-10	Compositional matrix adjust.	34/63(54%)	44/63(69%)	0/63(0%)	-3

```
Query  470  NGQDVTSAVIFSEVLPDVE*YCQIESYLEYMPTPGA*ITCMDKHLSLPKQVL*VWGKFAI  291
            NGQ+VTSAV  S+ +PD + Y QI SYLEY PTPG  ITCM +HL+L + +L VW  F +
Sbjct  151  NGQEVTSAVSSSDAMPDGDWYYQIHSYLEYTPTPGETITCMVQHLTLSEPMLQVWDPFLL  210

Query  290  VIQ  282
             +
Sbjct  211  AAE  213
```

PREDICTED: rano class II histocompatibility antigen, A beta chain-like isoform X1 [Lates calcarifer]
Sequence ID: XP_018527088.1 Length: 299 Number of Matches: 1

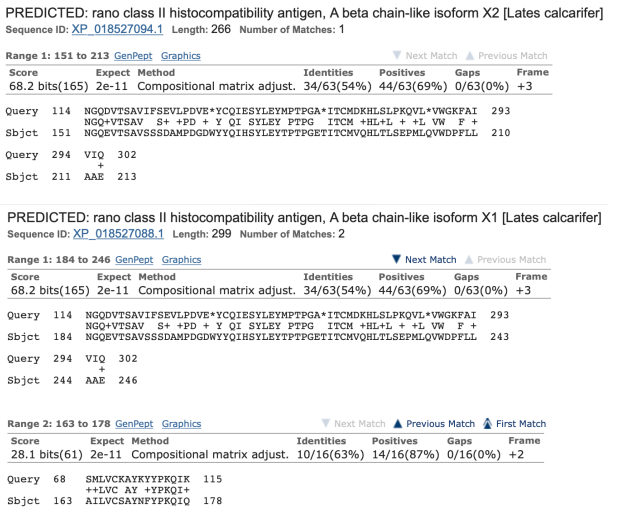
Range 1: 184 to 246 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
68.6 bits(166)	1e-10	Compositional matrix adjust.	34/63(54%)	44/63(69%)	0/63(0%)	-3

```
Query  470  NGQDVTSAVIFSEVLPDVE*YCQIESYLEYMPTPGA*ITCMDKHLSLPKQVL*VWGKFAI  291
            NGQ+VTSAV  S+ +PD + Y QI SYLEY PTPG  ITCM +HL+L + +L VW  F +
Sbjct  184  NGQEVTSAVSSSDAMPDGDWYYQIHSYLEYTPTPGETITCMVQHLTLSEPMLQVWDPFLL  243

Query  290  VIQ  282
             +
Sbjct  244  AAE  246
```

BF2 MHC class II fragment. Top 2 BLASTx hits at NCBI Nr



PREDICTED: rano class II histocompatibility antigen, A beta chain-like isoform X2 [Lates calcarifer]
Sequence ID: XP_018527094.1 Length: 266 Number of Matches: 1

Range 1: 151 to 213 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
68.2 bits(165)	2e-11	Compositional matrix adjust.	34/63(54%)	44/63(69%)	0/63(0%)	+3

```
Query  114  NGQDVTSAVIFSEVLPDVE*YCQIESYLEYMPTPGA*ITCMDKHLSLPKQVL*VWGKFAI  293
            NGQ+VTSAV  S+ +PD + Y QI SYLEY PTPG  ITCM +HL+L + +L VW  F +
Sbjct  151  NGQEVTSAVSSSDAMPDGDWYYQIHSYLEYTPTPGETITCMVQHLTLSEPMLQVWDPFLL  210

Query  294  VIQ  302
             +
Sbjct  211  AAE  213
```

PREDICTED: rano class II histocompatibility antigen, A beta chain-like isoform X1 [Lates calcarifer]
Sequence ID: XP_018527088.1 Length: 299 Number of Matches: 2

Range 1: 184 to 246 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
68.2 bits(165)	2e-11	Compositional matrix adjust.	34/63(54%)	44/63(69%)	0/63(0%)	+3

```
Query  114  NGQDVTSAVIFSEVLPDVE*YCQIESYLEYMPTPGA*ITCMDKHLSLPKQVL*VWGKFAI  293
            NGQ+VTSAV  S+ +PD + Y QI SYLEY PTPG  ITCM +HL+L + +L VW  F +
Sbjct  184  NGQEVTSAVSSSDAMPDGDWYYQIHSYLEYTPTPGETITCMVQHLTLSEPMLQVWDPFLL  243

Query  294  VIQ  302
             +
Sbjct  244  AAE  246
```

Range 2: 163 to 178 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

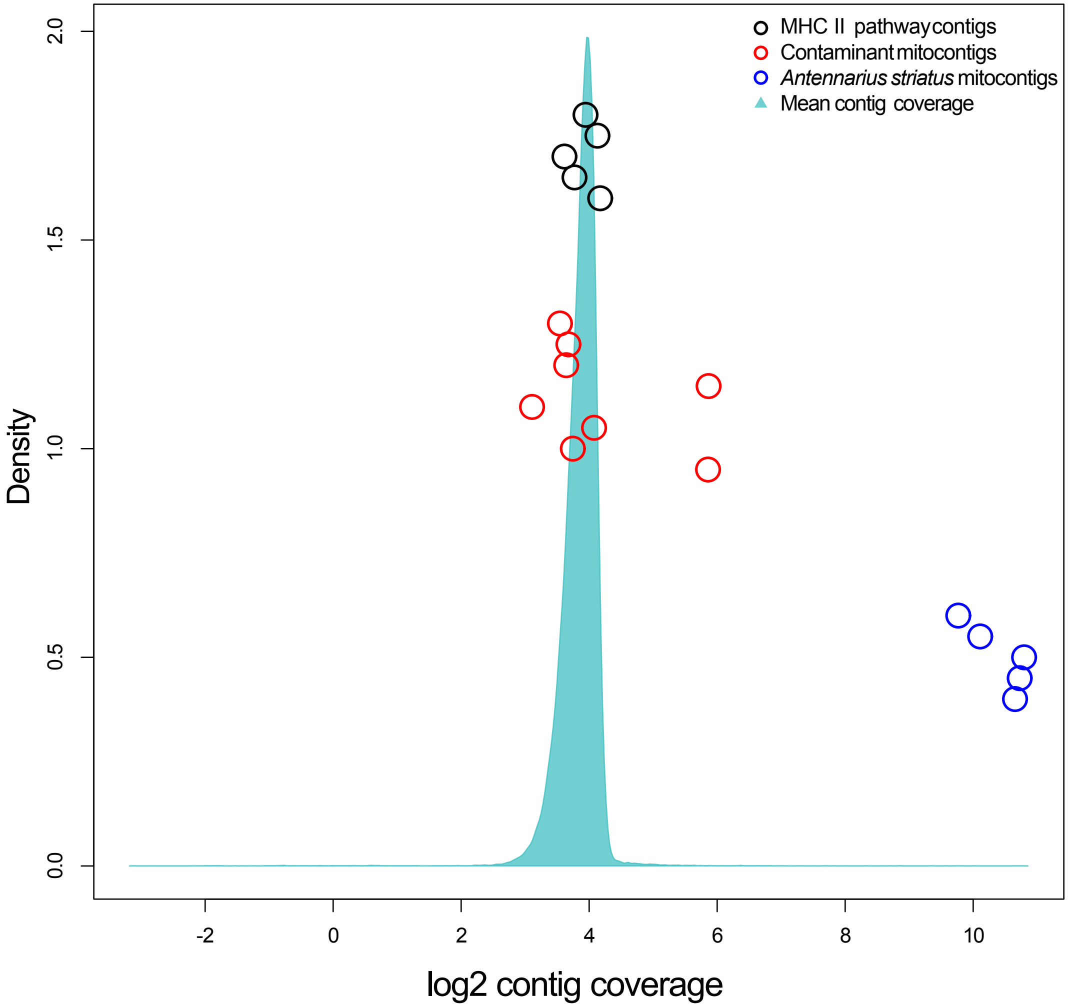
Score	Expect	Method	Identities	Positives	Gaps	Frame
28.1 bits(61)	2e-11	Compositional matrix adjust.	10/16(63%)	14/16(87%)	0/16(0%)	+2

```
Query  68   SMLVCKAYKYYPKQIK  115
            ++LVC AY +YPKQI+
Sbjct  163  AILVCSAYNFYPKQIQ  178
```

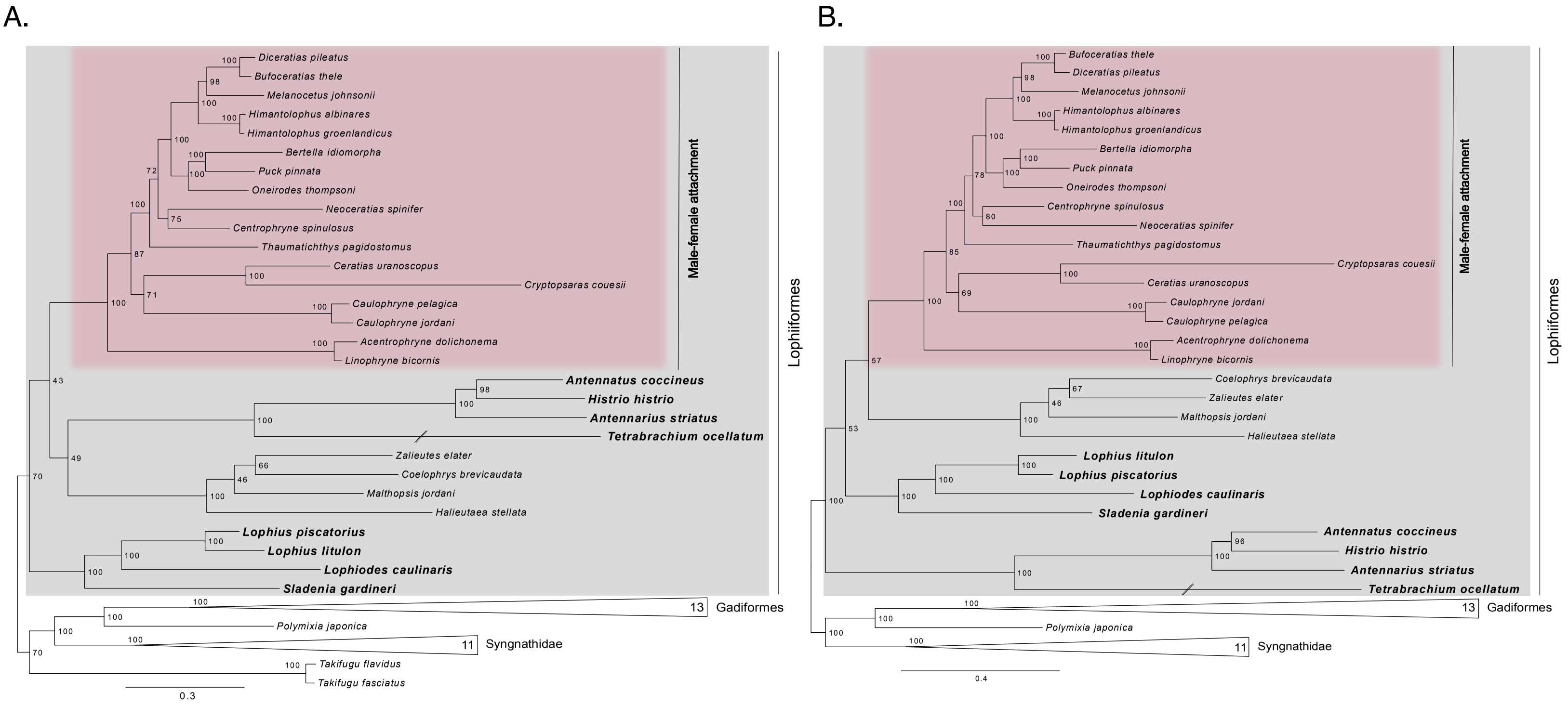
S3. Contaminant mitochondrial sequences in *A. striatus*

Contig identifier	Length (bp)	Match (identity)	Subject/query alignment coordinates
utg7180000251100	1749	Trachurus japonicus (AP003091.1) and Trachurus trachurus (AB108498.1) 98%	7223-5475/1-1749
utg7180000189499	3298	Trachurus japonicus (AP003091.1) 98%	10559-7263/1-3297
utg7180002535071	8275	Trachurus japonicus (AP003091.1 and AP003092.1) 97%	10559-16559/1-6000 1-2276/6001-8275
utg7180002551981	3238	Trachurus japonicus (AP003091.1 and AP003092.1) 94%	5430-2193/1-3238
utg7180000416975	1218	Decapterus maruadsi (KJ004518.1) 87.44% Decapterus macarellus (KM986880.1) 86.86%	9937-11146/9-1218 9938-11147/9-1218
utg7180002761404	2531	Decapterus maruadsi (KJ004518.1) 91%	7861-5334/4-2531
utg7180000037847	2676	Coreoperca loona (KJ644781.1) 86.59% Siniperca scherzeri (AP014527.1) 86.57%	203-2786/1-2609 202-2784/1-2609
utg7180002123755	13110	Emmelichthys struhsakeri (AP004446.1) 79.72% Monodactylus argenteus (AP009169.1) 79.63%	2787-15652/59-12949 2786-15672/59-12972

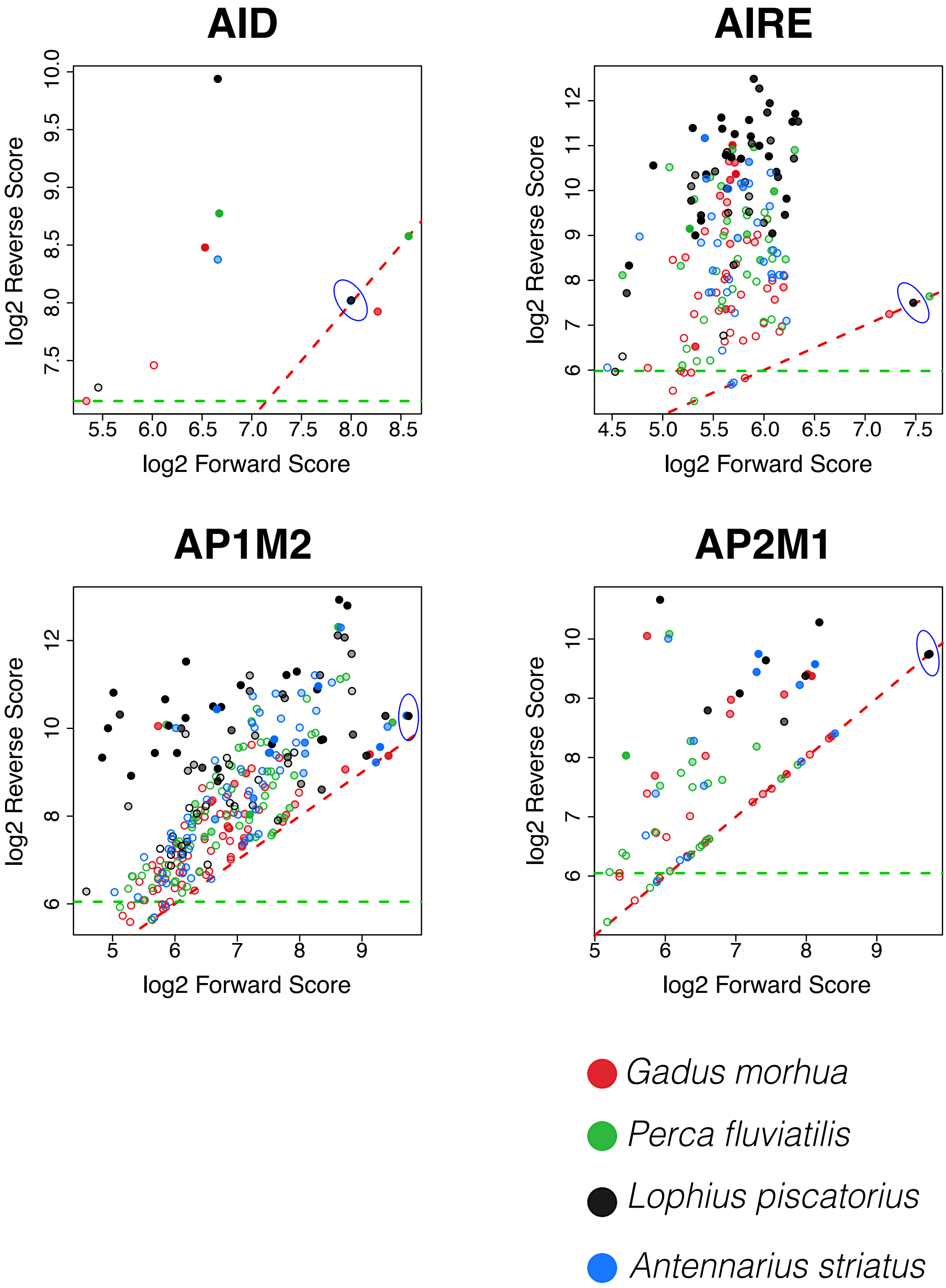
S4. Coverage of MHC II and mitochondrial sequences in *A. striatus*



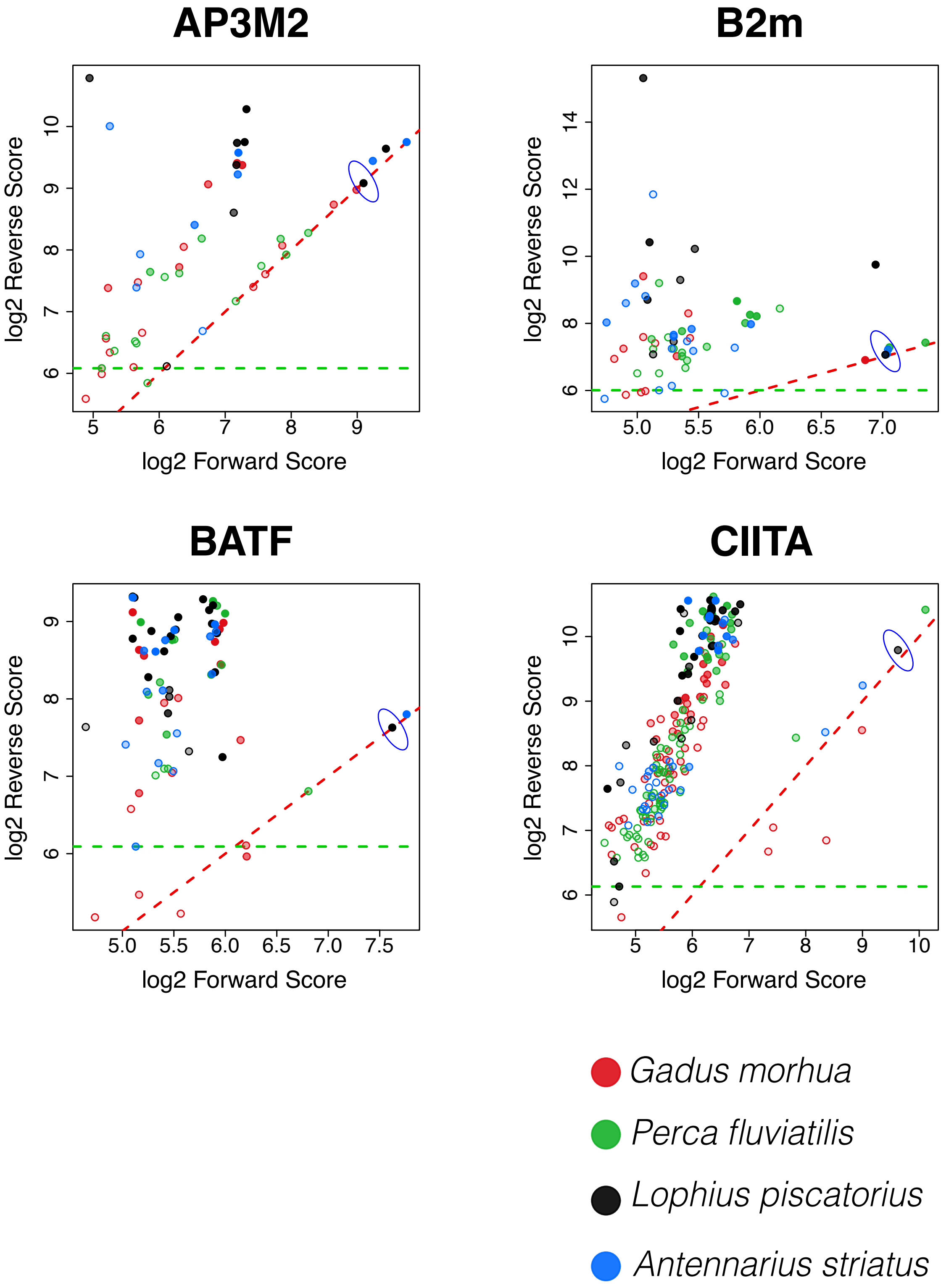
S5. Phylogenetic trees based on complete mitochondrial genome sequences of species in the Lophiiformes order



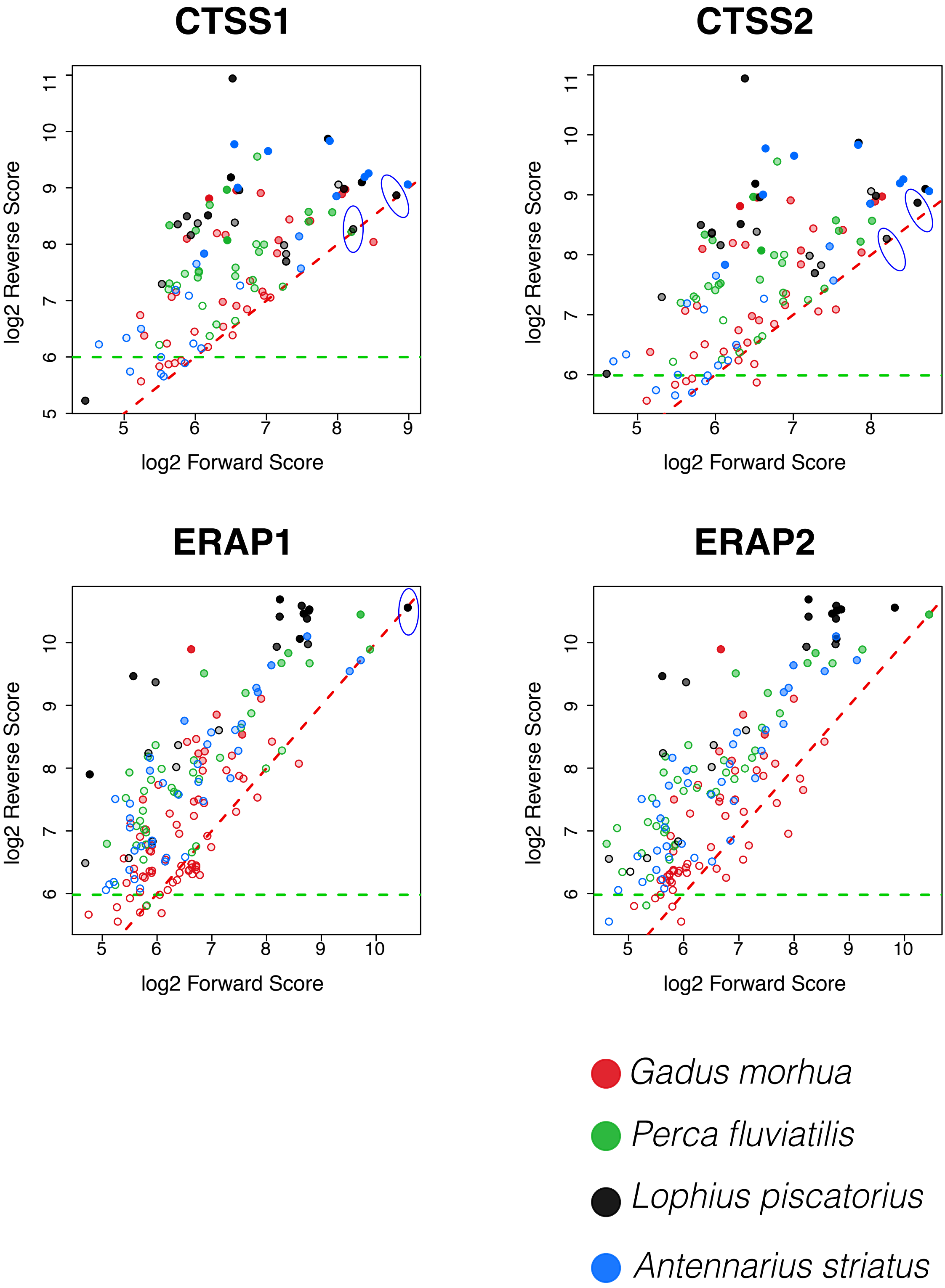
S6. Identification of immune gene orthologues, pages 6-11



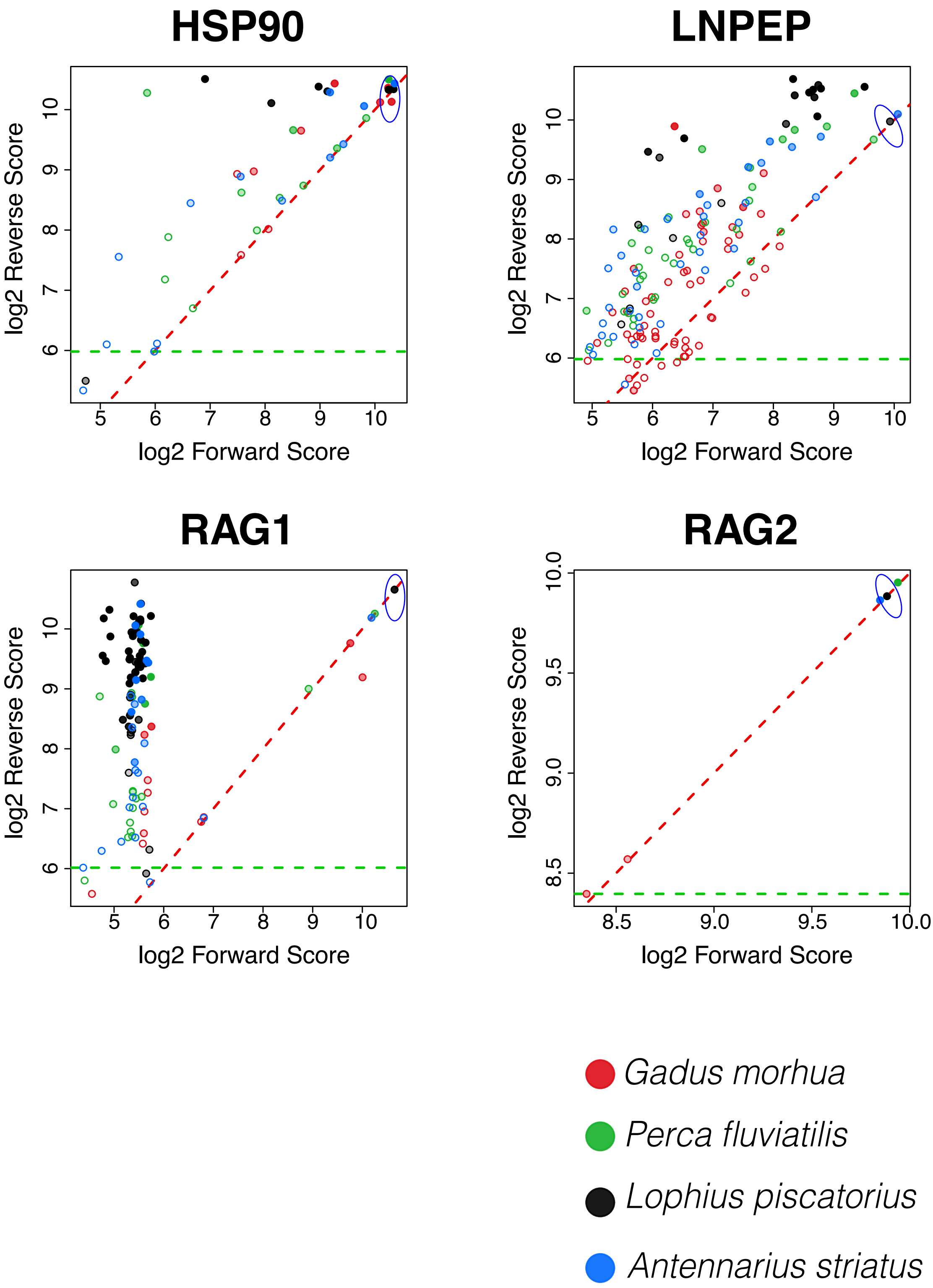
r. Identification of immune gene orthologues (Continued from p.6)



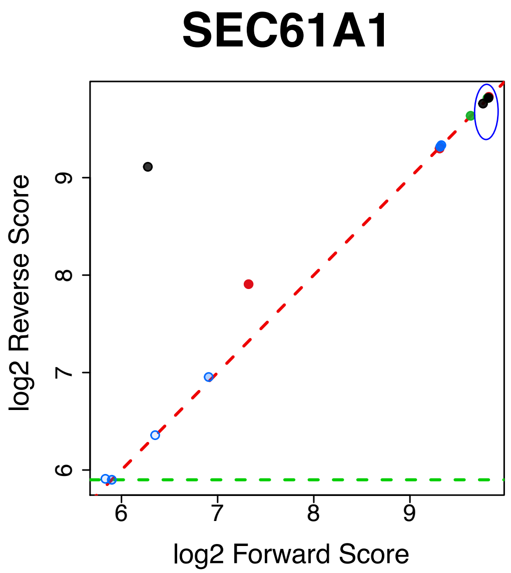
S6. Identification of immune gene orthologues (Continued from p.6)

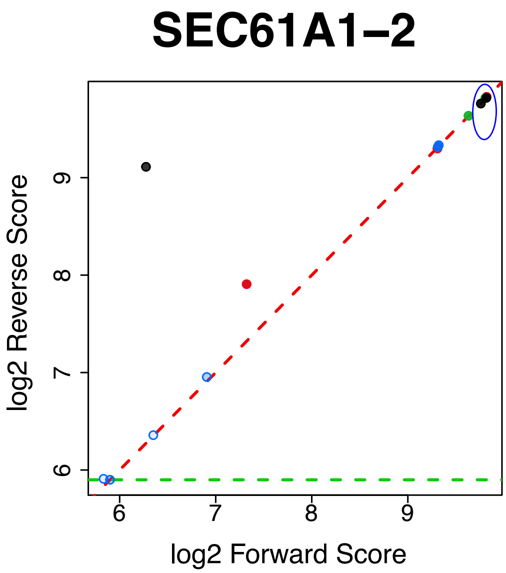


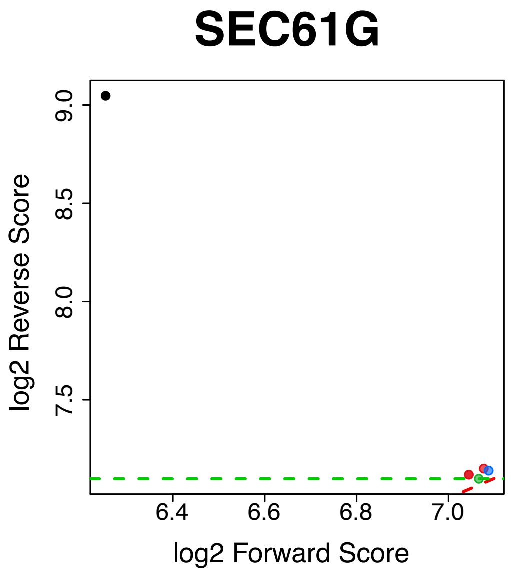
S6. Identification of immune gene orthologues (Continued p.6)

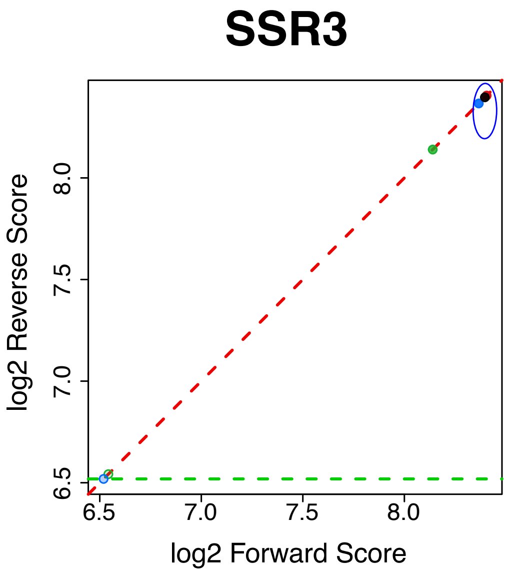


S6. Identification of immune gene orthologues (Continued p.6)









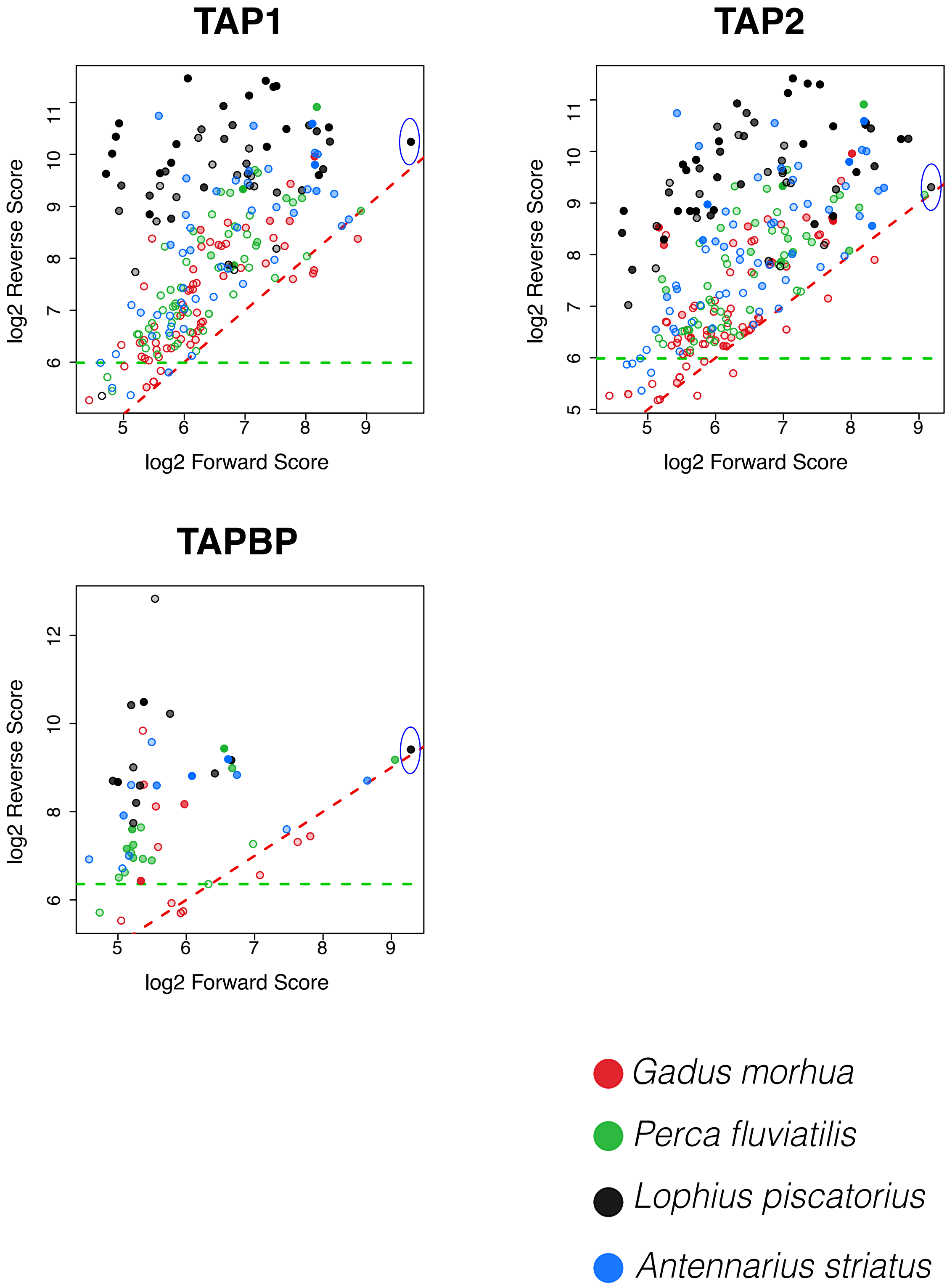
● *Gadus morhua*

● *Perca fluviatilis*

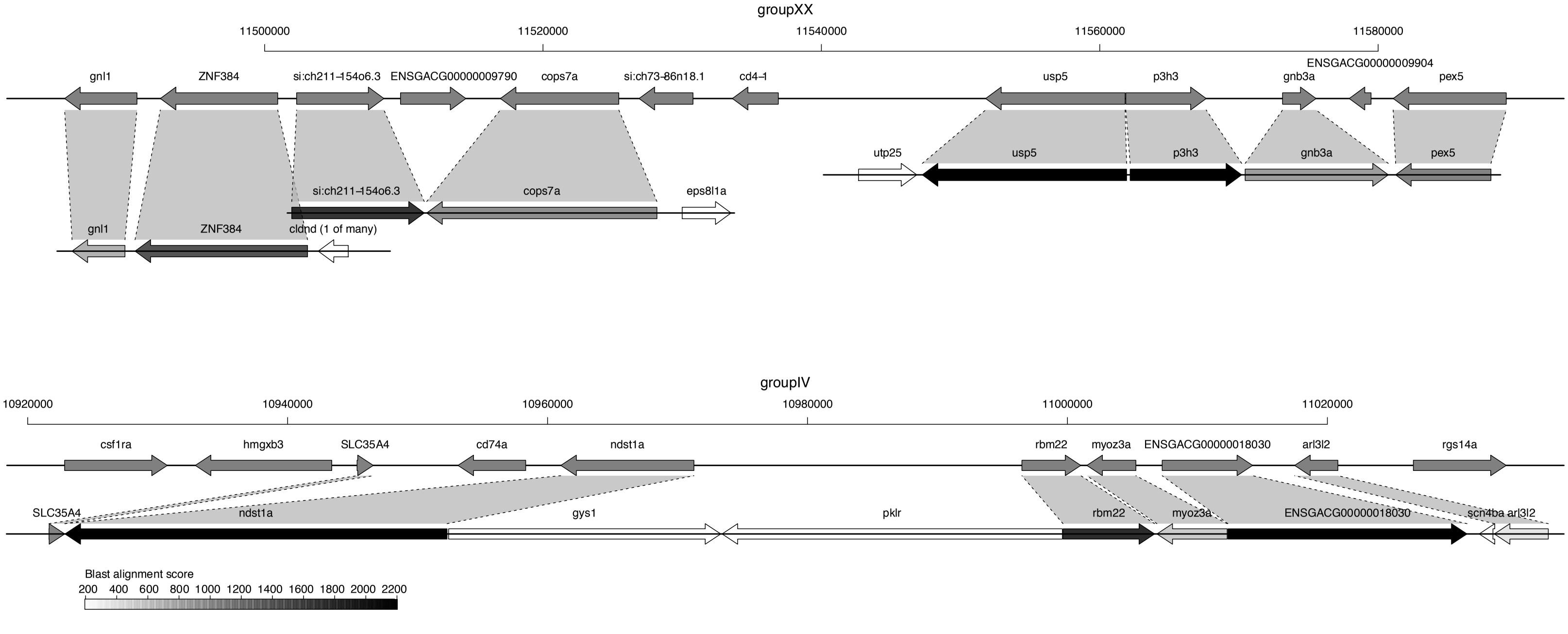
● *Lophius piscatorius*

● *Antennarius striatus*

S6. Identification of immune gene orthologues (Continued p.6)



S7. Gene synteny for the CD74 and CD4 gene regions



Supplementary figure legends

S1. Assembly statistics

Kernel density estimates of log (base 10) transformed contig length (A), coverage (B) and gene completeness (C) for the two *L. piscatorius* and the single *A. striatus* assembly.

S2. Sequences from the BF1 and BF2 assemblies that could be aligned to MHC II

A. Nucleotide and deduced amino acid sequences of the identified MHC II fragments. Residues shown in red mark the amino acids that were aligned by BLAST to MHC II β sequences.

B. Alignment of amino acid sequences from both assemblies (with stop codons removed) that aligned to MHC II β confirm that the same sequence was identified in both assemblies.

C. Top 2 BLASTx hits for the identified fragments. The figure shows direct screen-grabs from the NCBI BLAST web service.

S3. Contaminant mitochondrial sequences in *A. striatus*

Hits that we consider identifiable to a species level are marked with green. All sequences were identified with BLASTn and e-value threshold of 10. Hits shorter than a 1000 bp were discarded.

S4. Coverage of MHCII and mitochondrial sequences in *A. striatus*

Kernel density estimate of log2 transformed mean coverage for all contigs (cyan), mean coverage of MHC II containing contigs (black points) and mitochondrial contigs (red and blue points). Mitochondrial contigs identified as *A. striatus* sequences (blue points) were sequenced at 16 to 64 times the depth observed for contaminant mitochondrial sequences (red points). Each point represents one contig. Mean coverage for each contig was calculated by mapping quality trimmed reads to the assembly, converting bam to by-base coverage bed files and calculating the mean.

S5. Phylogenetic trees based on complete mitochondrial genome sequences of species in the Lophiiformes order

The internal branch ordering is dependent on the choice of outgroups, with the position of the Antennariodei and Lophiodei clades occupying the most basal position in A and B respectively.

The scale indicates the number of substitutions per site. The *Tetrabrachium ocellatum* branch length has been halved due to its extreme length. Node support values are bootstrap probabilities based on 500 iterations.

Phylogenetic relationships were inferred using a partitioned maximum likelihood analysis (with first, second and third codon positions, rRNA and tRNA as partitions) and a GTR GAMMA model as implemented in RaxML [1].

A. Tree created using *Polymixia japonica*, Gadiformes, Syngnathidae and Tetraodontiformes as the outgroups.

B. Tree created using *Polymixia japonica*, Gadiformes and Syngnathidae as the outgroups.

S6. Identification of immune gene orthologues. Pages 6-11

Illustration of identification criteria. Scores of alignments of putative orthologue sequences to the initial bait set (forward score, X-axis) plotted against scores obtained by alignment to sequences in the UniProt database (reverse score, Y-axis). The point fill transparency indicates the ratio of the alignment length to the length of the UniProt subject. Solid fills (alpha=1) correspond to full length alignments (i.e. the alignment covers the complete UniProt sequence). indicates relationship between the alignment length and length of the UniProt subject. Solid fill colour corresponds to 1/1 relationship. Orthologues should lie close to the Y=X line indicated by the dashed red line. The green dashed line shows the inferred e-10 e-value threshold. Points that we think represent orthologous sequences are marked with a blue ellipse. Peptide IDs corresponding to the selected points are collected in supplementary table 2, along with comments about the selection process., along with some comments about selection process. For additional information see *gene.hits.tsv* and *esm_pisc_pep.fasta*.

S7. Gene synteny for the CD74 and CD4 gene regions in *L.piscatorius* and *G. aculeatus*

The locations of orthologues to genes that are usually found in the CD4 and CD74 regions were identified in the BF2 assembly of *L. piscatorius* in the same manner as described above. To verify the identity of predicted genes we also aligned them to the NCBI nr database and manually inspected the resulting alignments. The synteny of the genes lying in the identified contigs was tested in *Gasterosteus aculeatus*. Genscan predicted peptides were blasted against *G. aculeatus* sequences (Ensembl *gasterosteus aculeatus* core 97.1). The top scoring alignment was taken as the gene identity for the genscan predictions and the matching contigs were aligned to their respective *G. aculeatus* loci using coordinates provided by Ensembl and the genscan predictions using a custom R-script. Top and bottom panel: CD4 and CD74 loci respectively. Genome positions in *G. aculeatus* are indicated by the scale bar; groupXX and groupIV are linkage group identifiers.

Shading of *L. piscatorius* gene predictions indicate the blast alignment score. Plots are to scale.

Supplementary Methods

****To find all scripts referred to in this ESM *see esm_scripts.txt***

Sequencing and genome assembly

The raw reads were trimmed from adapters and low quality bases using Cutadapt [2] with 25 as a quality threshold. Only Illumina data was used for the assemblies. Prior to assembly, overlapping read pairs were merged using FLASH (v1.2.11) [3]. Final assemblies were constructed with SPAdes (v3.10.0) [4] employing 6 kmer lengths (21, 33, 55, 77, 99, 127/103). Basic assembly statistics were calculated with QUAST (v4.4.1) [5] and gene-space completeness assessed using BUSCO (v2.0) [6] with the actinopterygii dataset (odb9). The trimmed reads were used to approximate the genome size with Jellyfish (v2.2.6) [7] and a suite of perl scripts (http://josephryan.github.com/estimate_genome_size.pl/).

Orthologue identification

In order to identify orthologues of adaptive immune system genes in *Lophius piscatorius* genome assemblies without the use of a predetermined e-value and bit score thresholds, we developed the strategy described below. Each step described was implemented in a short python (.py) or shell (.sh) script as indicated.

1. Identification of contigs that contain immune genes

We used a set of full-length amino acid sequences of 29 immune genes [8] and HSP90 from 10 species to search for orthologues in both our assemblies (BF1 and BF2). We also performed the same procedure for previously published draft genome assemblies of *Antennarius striatus*, *Gadus morhua and Perca fluviatilis* [8] as positive and negative controls to validate our strategy.

Sequences for the following species were obtained from Ensembl:

Danio rerio, Gadus morhua, Gasterosteus aculeatus, Tetraodon nigroviridis , Orytzias latipes, Oreochromis niloticus, Takifugu rubripes, Xiphophorus maculatus, Poecilia formosa, Astyanax mexicanus

Genes in the dataset:

1.AID
2.AIRE
3.AP1M2
4.AP2M1
5.AP3M2
6.B2m
7.BATF
8.CD4
9.CD8a
10.CD8b
11.CIITA
12.CTSS1
13.CTSS2
14.ERAP1
15.ERAP2
16.HSP90
17.CD74a
18.CD74b
19.LNPEP
20.MHCI
21.MHCIIa
22.MHCIIb
23.RAG1
24.RAG2
25.SEC61A1
26.SEC61A1-2
27.SEC61G
28.SSR3
29.TAP1
30.TAP2
31.TAPBP

To identify contigs containing candidate orthologues, we aligned the peptide sequences encoded by these genes to assemblies using tBLASTn (*manyfish_blast.py*). To reduce the false negative rate at this step we used a permissive e-value threshold of 1 (compared to the e-10 threshold usually used) but limited the number of target sequences to 50 and relied on post-filtering to remove incorrect matches.

Identifiers of contigs containing alignments to the seed genes were extracted from the BLAST output and split by assembly and gene into separate files (*process_blast.py*) which were used for downstream analyses.

2. Contig extraction and gene prediction

Selected contigs were subjected to gene prediction by Genscan [10], resulting in a set of amino acid sequences for each immune gene and matching contig. These sequence sets included both the amino acid sequences of orthologous immune genes and unrelated sequences located within the same contigs. To identify the orthologues, we used two further BLASTp screens which we refer to as Forward and the Reverse BLAST.

All predicted peptides from the BF2 *L. piscatorius* assembly (including non-immune peptides) sorted by gene can be found in the *esm_pisc_pep.fasta* file.

3. Forward BLAST

In order to provide alignment scores that could be compared to those in an extended blast against UniProt (step 4), we aligned amino acid sequence sets identified by Genscan in step 2 to the initial seed set (*forward_blast.sh*) using BLASTp. Again we used an e-value threshold of 1 and limited the number of target sequences to 50.

Peptide sequences aligning to their respective seed genes from step 1 were selected for further analyses (*filter_forward_blast.sh*). For example, peptides derived from contigs identified by tBLASTn with AIRE as a query were filtered to remove all peptides not aligned to AIRE.

We refer to the BLASTp bit score values obtained in this search as the Forward BLAST score.

4. Reverse BLAST

The majority of alignments obtained in the first rounds of blast with the seed set of immune genes are likely to involve proteins that are not orthologous, but which contain domains with some homology with seed set domains. Such sequences should align with better scores to their true orthologues, at least some of which we would expect to find within the UniProt database.

Hence, we aligned the candidate immune peptide sequences from step 2 to the UniProt KB database (*reverse_blast.sh*). Again, the e-value threshold was 1 and the number of target sequences in the output was limited to 50. This is similar to the rationale for reciprocal blast, and for this reason we refer to this step as reverse blast even though technically both step 3 and 4 are done in the same direction.

We refer to the BLASTp bit score values obtained in this search as the Reverse BLAST score.

5. Comparison of the Forward and Reverse BLAST scores

In theory, immune gene orthologues should align to the initial immune set (from step 1) and to the UniProt with similar bit scores, i.e. have similar Forward and Reverse scores, whereas non-orthologous sequences should be aligned with a higher score to their true orthologues present in Uniprot and hence have higher Reverse scores.

To determine whether it was possible to separate true orthologues by comparing forward and reverse scores we plotted forward against reverse (log)scores (*R script bl_revision.R, functions.R see functions_and_R_scripts.txt*). Since truncated orthologue sequences would still have similar forward and reverse scores (reflecting their identity), we also visualised the ratio of the alignment length to the Uniprot sequence length using alpha transparency values for points such that points reflecting alignments to non-truncated sequences (i.e. similar sequence length) appear as solid points.

6. Visual/manual examination of plots/hits

To verify the identity of candidate immune gene orthologues we used the *identify* function in R to examine the UniProt annotation of selected alignments. For most immune genes, orthologues were easily identifiable as they lied on/or very close to the forward = reverse score line and were aligned to a UniProt protein annotated as the desired immune gene orthologue. However, for some genes we observed multiple points on or close to this line (AP1M2, AP3M2, CTSS1/2), the UniProt annotation did not match with the selected gene (ERAP1, TAP1), or none of the points on the plot fitted our criteria (SEC61G, ERAP2). In this case, we chose several points that might represent an orthologue and examined their top 5 UniProt hits (*gene.hits.tsv*)

Supplementary Table 2

Gene name	Orthologous predicted protein	Comments
AID	NODE_3337_length_66067_cov_44.9838|GENSCAN_predicted_peptide_3|233_aa	Top scoring UniProt hit belongs to correct gene
AP2M1	NODE_326_length_249814_cov_47.6866|GENSCAN_predicted_peptide_14|1074_aa NODE_6641_length_27025_cov_42.6987|GENSCAN_predicted_peptide_1|1026_aa	Top scoring UniProt hit of the first peptide and third of the second peptide belongs to correct gene
AP3M2	NODE_11858_length_6461_cov_171.208|GENSCAN_predicted_peptide_1|405_aa	Top scoring UniProt hit belongs to correct gene
B2m	NODE_26321_length_534_cov_183.369|GENSCAN_predicted_peptide_1|92_aa	Top scoring UniProt hit belongs to correct gene
BATF	NODE_6109_length_31083_cov_52.3689|GENSCAN_predicted_peptide_2|251_aa	Top scoring UniProt hit belongs to correct gene
CIITA	NODE_2303_length_93200_cov_43.342|GENSCAN_predicted_peptide_4|1556_aa	Top scoring UniProt hit belongs to correct gene
CTSS1	NODE_2178_length_97347_cov_44.3687|GENSCAN_predicted_peptide_7|470_aa	Top scoring UniProt hit belongs to correct gene
CTSS2	NODE_969_length_161682_cov_49.7592|GENSCAN_predicted_peptide_5|404_aa	Top scoring UniProt hit belongs to correct gene
HSP90	NODE_170_length_315068_cov_45.2866|GENSCAN_predicted_peptide_20|709_aa	Top scoring UniProt hit belongs to correct gene
LNPEP	NODE_1284_length_138026_cov_45.1581|GENSCAN_predicted_peptide_8|971_aa	Top scoring UniProt hit belongs to correct gene
RAG 1	NODE_1604_length_120776_cov_52.7006|GENSCAN_predicted_peptide_2|1015_aa	Top scoring UniProt hit belongs to correct gene
RAG 2	NODE_1604_length_120776_cov_52.7006|GENSCAN_predicted_peptide_3|533_aa	Top scoring UniProt hit belongs to correct gene
SEC61A1	NODE_148_length_331228_cov_45.7995|GENSCAN_predicted_peptide_18|494_aa	Top scoring UniProt hit belongs to correct gene
SEC61A1-2	NODE_4460_length_48594_cov_76.0295|GENSCAN_predicted_peptide_1|454_aa	Top scoring UniProt hit belongs to correct gene
SSR3	NODE_3618_length_61353_cov_45.4674|GENSCAN_predicted_peptide_2|296_aa	Top scoring UniProt hit belongs to correct gene
TAP2	NODE_4645_length_46251_cov_64.2593|GENSCAN_predicted_peptide_5|875_aa	Top scoring UniProt hit belongs to correct gene
TAPBP	NODE_11231_length_7753_cov_428.536|GENSCAN_predicted_peptide_1|456_aa	Top scoring UniProt hit belongs to correct gene
ERAP1	NODE_1839_length_110210_cov_47.4795|GENSCAN_predicted_peptide_7|886_aa	After examination, top UniProt hit belongs to correct gene
AIRE	NODE_2144_length_98088_cov_45.0458|GENSCAN_predicted_peptide_7|218_aa	Second UniProt hit and two others belong to correct gene
TAP1	NODE_39_length_479181_cov_44.6292|GENSCAN_predicted_peptide_18|1443_aa	Fusion prediction. Selected first part of the sequence
AP1M2	NODE_938_length_164357_cov_43.2764|GENSCAN_predicted_peptide_8|1716_aa	Fusion prediction. Selected last part of the sequence
ERAP2	Unclear orthology due to too many paralogous aminopeptidases.	?
SEC61G	See figure legend for detail.	Special case. See figure legend.

Supplementary Table 2

The table includes identifiers of the predicted peptides (column 2) that we consider to be orthologues to the target set of immune genes (column 1). Column 3 contains short comments on how this gene was identified. For most genes, the top blast hit lying on the X=Y line corresponded clearly to a UniProt protein annotated as the respective target gene. However, for some genes we had to examine the annotation of additional hits, due to non-informative description of the top hit, e.g. in the case of AIRE the top UniProt hit was described as Chromosome_15_SCAF14992. In addition, some predicted peptides combined products from two adjacent genes (Fusion prediction). For these genes the alignment coordinates had to be examined. The predicted protein sequences can be found in *esm_pisc_pep.fasta* and a summary of the blast output for selected points is provided in *gene.hits.tsv.*

SEC61G of *L. piscatorius* was a special case. Although SEC61G is a highly conserved gene, it is short and one exon primarily contains low-complexity sequence. This results in alignments to the second exon having low BLAST scores leading to its exclusion from the gene prediction and resulting in a truncated protein sequence. However, a manual examination of the BLAST output clearly demonstrated that complete sequence was aligned with a high sequence similarity (but low score). Similarly, running BLAST with '-dust no' provided the full alignment with a high alignment score. It is notable that SEC61G is one of the genes that Malmstrøm et al. failed to identify in a number of species.

7. Unassembled reads search

Protein sequences from genes for which we failed to identify *L. piscatorius* orthologues with the Forward/Reverse BLAST strategy were used in a tBLASTn search of the unassembled read pools. In this case, we included both Illumina and SOLiD reads. Reads that were aligned to the missing protein sequences were re-assembled with CLC Genomics Workbench. The resulting contigs were aligned to the NCBI nr database with BLASTn. If this approach failed to identify missing orthologues, we aligned selected unassembled reads to the NCBI nr database. After this, we reported orthologues that we failed to identify as actually missing.

Construction of phylogenetic trees

All sequences were obtained from genbank (see accession numbers in the section below). Then, mitogenomes were split by gene according to their annotations. First, each protein coding gene, each tRNA and rRNA were aligned separately with T-Coffee [9]. Then, alignments were trimmed from the ends, to remove end gaps and sequences were concatenated into new mitogenome sequences for all species. Datasets were partitioned by the first, second and third codon positions for protein coding genes, then rRNA and tRNA were put as separate partitions. To construct the trees we used RaxML [1] using the GTR GAMMA model with 500 rapid bootstrap (-f a option) iterations.

Sequences used to construct the trees:

Polymixiidae

NC_002648	*Polymixia japonica*

Tetraodontiformes

GQ409967	*Takifugu fasciatus*
KJ562276	*Takifugu flavidus*

Syngnathiformes

KJ184525	*Syngnathoides biaculeatus*
KU925872	*Syngnathus typhle*
KJ184524	*Solegnathus hardwickii*
AP012309	*Doryrhamphus japonicus*
AP013027	*Hippocampus histrix*
KJ184528	*Trachyrhamphus serratus*
KP861226	*Syngnathus schlegeli*
JX970973	*Hippocampus comes*
NC_010272	*Hippocampus kuda*
NC_022722	*Hippocampus erectus*
KJ139455	*Corythoichthys flavofasciatus*

Gadiformes

AP018148	*Gadiculus argenteus thori*
X99772	*Gadus morhua*
KC844053	*Lota lota*
NC_008225	*Ventrifossa garmani*
NC_015102	*Micromesistius poutassou*
NC_004377	*Physiculus japonicus*
NC_015094	*Pollachius virens*
NC_010122	*Arctogadus glacialis*
NC_015120	*Merluccius merluccius*
NC_010121	*Boreogadus saida*
NC_008224	*Trachyrincus murrayi*
NC_008222	*Bathygadus antrodes*
NC_008124	*Bregmaceros nectabanus*

Lophiiformes

AB282831	*Tetrabrachium ocellatum*
AB282828	*Antennarius striatus*
AP005977	*Halieutaea stellata*
AB282837	*Neoceratias spinifer*
AB282847	*Thaumatichthys pagidostomus*
AB282836	*Caulophryne pelagica*
AB282830	*Antennatus coccineus*
AB282841	*Bufoceratias thele*
AB282842	*Diceratias pileatus*
AB282827	*Sladenia gardineri*
AB282855	*Acentrophryne dolichonema*
AB282854	*Linophryne bicornis*
AB282840	*Himantolophus groenlandicus*
AB282829	*Histrio histrio*
AB282849	*Centrophryne spinulosa*
AB282839	*Himantolophus albinares*
AB282835	*Zalieutes elater*
AB282826	*Lophiodes caulinaris*
AP005978	*Malthopsis jordani*
AB282833	*Chaunax pictus*
AB282838	*Melanocetus johnsonii*
AB282834	*Coelophrys brevicaudata*
AB282845	*Chaenophryne melanorhabdus*
AB282843	*Oneirodes thompsoni*
AB282846	*Bertella idiomorpha*
AB282844	*Puck pinnata*
AB282851	*Ceratias uranoscopus*
AB282850	*Cryptopsaras couesii*
NC_004383	*Caulophryne jordani*
MF994812	*Lophius piscatorius*
KJ020931	*Lophius litulon*

Sources.

1. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. (doi:10.1093/bioinformatics/btu033)

2. Martin M. 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10-12. (doi:10.14806/ej.17.1.200)

3. Magoc T, Salzberg SL. 2011 FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963. (doi:10.1093/bioinformatics/btr507)

4. Bankevich A et al. 2012 SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19, 455–477. (doi:10.1089/cmb.2012.0021)

5. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013 QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. (doi:10.1093/bioinformatics/btt086)

6. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018 BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35, 543–548. (doi:10.1093/molbev/msx319)

7. Marçais G, Kingsford C. 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770. (doi:10.1093/bioinformatics/btr011)

8. Malmstrøm M et al. 2016 Evolution of the immune system influences speciation rates in teleost fishes. Nat. Genet. 48, 1204–1210. (doi:10.1038/ng.3645)

9. Notredame C, Higgins DG, Heringa J. 2000 T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. J. Mol. Biol. 302, 205–217. (doi:10.1006/jmbi.2000.4042)

10. Burge C, Karlin S. 1997 Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94. (doi:10.1006/jmbi.1997.0951)

Paper IV

Assembly and annotation of an anglerfish genome

Arseny Dubin, Tor Erik Jørgensen, Truls Moum, Steinar Daae Johansen and Lars Martin Jakt

Genomics group, Faculty of Biosciences and Aquaculture, NORD University, 8049 Bodø, Norway

Author for correspondence:

Lars Martin Jakt

lars.m.jakt@nord.no

Abstract

The anglerfishes (Lophiiformes) comprise at least 321 species that have evolved a wide range of behavioural and morphological adaptations. Most notably, members of the Ceratioidei suborder have developed a mode of reproduction where the male attaches to, and essentially becomes a parasite of the female. Nevertheless, anglerfishes remain poorly studied with a small amount of genomic resources available. Here we present an annotated chromosome level genome assembly of the monkfish *Lophius piscatorius*. We validate the genome assembly and annotation by comparison to five other teleost species. The assembly was determined to 715 Mb, which is typical of teleost genomes. About 32% of the genome (232Mb) consists of repeats and low-complexity regions, dominated by Tc1/Mariner Class II transposons. We also observed a teleost-specific bimodal distribution in intron lengths and, as recently reported, a complete loss of MHCII pathway genes. The availability of this genome will greatly facilitate further analyses of anglerfish biology and the evolutionary relationships among teleost species.

Introduction

The anglerfishes

Anglerfishes comprise the teleost order Lophiiformes, which includes at least 321 living species with new members being discovered relatively frequently (Shedlock et al. 2004; Pietsch et al. 2009a, b; Miya et al. 2010; Pietsch and Sutton 2015; Ho 2016; Ho and Ma 2016; Rajeeshkumar et al. 2017; Arnold and Pietsch 2018). The order is further divided into 5 suborders, members of which display extraordinary morphological diversity and occupy a wide range of habitats (Miya et al. 2010; Betancur-R et al. 2017). Four suborders; Lophioidei, Antennarioidei, Ogcocephaloidei and Chaunacoidei, include shallow/deep water benthic ambush predators, while Ceratioidei, the most species rich suborder, is represented by meso/bathypelagic and abyssal-benthic forms. Despite the diverse appearance of its members, both morphology and molecular-based phylogenies support the monophyly of the Lophiiformes order (Pietsch and Orr 2007; Miya et al. 2010). The current phylogenies place the Lophioidei suborder as the most basal clade in the anglerfish tree, followed by Antennarioidei or Ogcocephaloidei, Chaunacoidei and Ceratioidei. Batrachoidiformes (frogfishes) was historically considered to be the clade most closely related to Lophiiformes (Shedlock et al. 2004; Miya et al. 2010); however this hypothesis has recently been challenged as a result of phylogenies based on complete mitochondrial genomes and nuclear ultra conserved elements (Alfaro et al. 2018) and the order is now grouped with Tetraodontiformes (Miya et al. 2010; Betancur-R et al. 2017).

Special adaptations in anglerfishes

Anglerfishes are known for several unique adaptations. The most famous is their luring apparatus (illicium) from which their name is derived. The illicium is formed from a modified dorsal fin spine (Miya et al. 2010) and with a few exceptions is found in all Lophiiformes species. In Lophioidei and Antennarioidei, the illicium is relatively immobile, but in some Chaunacoidei and Ogcocephaloidei members, and in most Ceratioidei females it is retractable.

The tip of the illicium usually ends with a fleshy outgrowth or esca. The appearance of the esca is often species specific and can include complex structures to lure prey. For most species the esca represents an outgrowth of skin, as in *Lophius piscatorius*. The esca of most Ceratioidei members contains a photophore organ filled with bioluminescent bacteria (Hulet and Musil 1968; O'day 1974; Pietsch and Orr 2007). Species that belong to the Linophryne genus also have a luminous barbel that generates light from photogenic granules instead of bacteria (Hansen and Herring 2009). In some batfish species the esca is known to release a chemical compound which attracts marine gastropods (Nagareda and Shenker 2009).

Another notable anglerfish adaptation, common for benthic species, is their ability to "walk" using modified pectoral fins. Members of the Antennarioidei suborder (also known as frogfishes) use their pectoral fins for walking along the bottom rather than for swimming (Dickson and Pierce 2019).

Sexual parasitism in anglerfishes

The deep-water suborder Ceratioidei, or sea devils, represents the most diverse and species dense clade in the order, accounting for half of the living anglerfish species. Members of this suborder display extreme sexual dimorphism and male-to-female attachment. Here, males attach to the females either temporarily or permanently, and in extreme cases fuse with the female body. After fusion males become dependent on the female circulatory system for their nutrient supply (Pietsch 1976, 2005, 2009; Munk 2000; Pietsch and Orr 2007). Why such fusion does not result in tissue rejection is unknown. When overlaid on a phylogenetic tree, male sexual parasitism has a patchy distribution within the suborder, with obligatory male parasitism and temporary attachment having no clear evolutionary pathway. Hence, the general consensus is that male parasitism evolved multiple times within the suborder (Pietsch 1976, 2005, 2009; Pietsch and Orr 2007; Miya et al. 2010). This suggests a common selective pressure, or a shared genetic predisposition which may be present across the order.

The monkfish *Lophius piscatorius*

Lophius piscatorius is commonly known as monkfish in English and belongs to the most apical clade in the Lophioidei suborder (Miya et al. 2010; Betancur-R et al. 2017). It is a dorso-ventrally flattened bathydemersal fish, which can be found along the European continental shelf, from Gibraltar to the Barents Sea and Iceland (Thangstad 2006; Farina et al. 2008). *Lophius* members are opportunistic feeders and display low prey selectivity. Their diet consists mainly of various bony marine fish, cephalopods and crustaceans (Thangstad 2006; Issac et al. 2017).

An immune system without MHCII

The lack of immune rejection of males attached to females suggest the presence of a modified immune system in anglerfishes. The fact that sexual parasitism seems to have arisen multiple times suggests that immune system modifications are shared across the clade and may have been present in their common ancestor. Indeed, we have recently demonstrated that *L. piscatorius*, in common with the Gadiformes order and pipefish, lacks all key components of the MHCII pathway (Dubin et al. In Press; Star et al. 2011; Haase et al. 2013; Malmstrøm et al. 2016). This indicates that the loss of MHCII may have been one of the enabling adaptations allowing sexual parasitism to evolve within anglerfish. However, it has previously been reported that the striated frogfish *Antennarius striatus* contains a complete MHCII system (Malmstrøm et al. 2016). As the most recently published anglerfish phylogenies suggest that A. striatus has a more recent common ancestor with the Ceratioidei order than *L. piscatorius* it is not possible for the loss of MHCII in *L. piscatorius* to be shared across the species which exhibit sexual parasitism if these phylogenies are correct.

Hence, the loss of MHCII pathway in *L. piscatorius*, if shared across the Ceratioidei would both invalidate currently accepted phylogenies and be likely to have contributed to the observed immune tolerance. However, on its own, the loss of MHCII is unlikely to be sufficient (Rosenberg and Singer 1992; Waldmann 2010; Abrahimi et al. 2016) and we would expect other immune system modifications to exist within the anglerfishes. Such modifications may be shared across the clade or be specific to the subclades. Here we extend our previous work to provide an annotated chromosome level anglerfish genome assembly. This can provide a reference point for the investigation of the development of sexual parasitism and the other strange and wonderful anglerfish adaptations.

Methods

Sampling and genome sequencing

Samples from an *L. piscatorius* female (referred to as BF2) were collected in the Bodø coastal waters, Nordland County in collaboration with local fishermen. Total DNA extraction from kidney and sequencing was performed by Dovetail Genomics, USA on an Illumina HiSeq X instrument (sequence depth: 150x) as a paid service. In addition to the regular paired-end library, HiC and Dovetail Chicago libraries were made. The libraries were 150 bp paired-end reads with 350 bp insert size for the HiSeq, and variable length for the HiC and Chicago libraries. Total RNA from heart was isolated using QIAzol (QIAGEN, Hilden - Germany) according to the manufacturers protocol. Cellular rRNA depletion was performed using the riboMinus Eukaryote System v2 (Thermo Fisher Scientific, Waltham, MA - USA). RNA seq libraries were constructed using the Ion Total RNA-seq kit v2 following the manufacturer's standard protocol and sequenced using the Ion Torrent PGM system.

Assemblyand scaffolding

The raw Illumina reads were trimmed from adapters and low-quality bases using Cutadapt with a q25 threshold (Martin 2011). Prior to assembly, overlapping read pairs were merged using FLASH (v1.2.11) (Magoc and Salzberg 2011). The assembly was constructed with SPAdes (v3.10.0) (Bankevich et al. 2012) using 6kmer lengths

(21,33,55,77,99,105). The scaffolding of the SPAdes assembly was performed using HiC and Chicago data with the HiRise scaffolding pipeline by Dovetail Genomics. The gene-space completeness of the final assembly was assessed with BUSCO v3 using the Actinopterygii odb9 database (Waterhouse et al. 2018). The Ion Torrent (RNA) reads were quality trimmed using Cutadapt (Martin 2011) with q20 as a threshold. The minimum read length was set to 50 nt. A *de novo* genome guided transcriptome assembly was performed using Trinity v2.8.5 (Grabherr et al. 2011) with default parameters.

Repeat annotation

Prior to annotation by MAKER we constructed a de-novo repeat library using RepeatModeler version 1.0.11 with default parameters. Repeats that failed to be classified were aligned to the transposase database (Ou and Jiang 2018) with BLASTx, and searched against the Dfam database (Hubley et al. 2016) of repetitive elements with nhmmer (Wheeler and Eddy 2013). The resulting files were processed with scripts from (https://github.com/uio-cels/Repeats/). In order to characterize the repeat content of the genome in more detail and compare it to other fish species we also ran RepeatMasker with the Repbase database + De-novo repeat library for *Takifugu rubripes* ((Kai et al. 2011, GCF_901000725), *Gadus morhua* (Tørresen et al. 2017) and *Lepisosteus oculatus* ((Braasch et al. 2016, GCF_000242695). The RepeatMasker output was processed with a script (https://github.com/4ureliek/Parsing-RepeatMasker- Outputs).

Gene prediction

Genome-wide gene prediction was performed with MAKER (Holt and Yandell 2011), following the approach described in (Varadharajan et al. 2018). First, hidden Markov model profiles were generated de-novo with GeneMark-ES v4 (Lomsadze 2005; Borodovsky and Lomsadze 2011) and SNAP (version 2006-07-28c) (Korf 2004) using the core eukaryotic gene data set (CEGMA). Next, the resulting HMM profiles, the RepeatModeller library, the de-novo transcriptome assembly and the Uniprot KB database were used as an input for MAKER.

Gene annotation

Translations of maker gene predictions were used to search for homologous sequences in 7 well annotated genomes using blastp (Altschul et al. 1990; Camacho et al. 2009). Peptide sequences, and associated gene annotation for *Ciona intestinalis*, *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripens* and *Tetraodon nigroviridis* were obtained from Ensembl (Zerbino et al. 2018) version 97. Blast databases were created for both *L. piscatorius* and the Ensembl set of peptides. Blastp was run in both directions (*L. piscatorius* vs Ensembl, and Ensembl vs *L. piscatorius*) using default parameters but with a tabular output format and limiting the output to 10 matches per query. The resulting data sets were processed using a Perl script to add gene identities, remove alignments with E-values larger than 1e-10 and to infer reciprocal ranks. The resulting output was analysed using R where the data-sets were reduced to include only the highest scoring alignment for each ensembl-maker gene pair. Score-ratios were calculated for all alignments representing the fraction of the alignment score to the maximum alignment score for the given Ensembl gene (score[i,g] / max(score[,g]), where i is the i-th alignment score to gene g).

The genome locations of *L. piscatorius* genes were extracted from gff files produced by the maker script (maker_1.genes.gff). Ensembl gene annotation was extracted from locally installed species specific core Ensembl databases (ciona_intestinalis_core_97_3, danio_rerio_core_97_11, gasterosteus_aculeatus_core_97_1, lepisosteus_oculatus_core_97_1, oryzias_latipes_core_97_1, takifugu_rubripes_core_97_5, tetraodon_nigroviridis_core_97_8). Plots showing chromosomal level synteny were produced using custom R functions. Gene family descriptions and memberships were obtained from the public ensembl_compara_97 SQL portal (Herrero et al. 2016). Gene family membership for *L. piscatorius* genes were assigned based on the plurality of a vote by candidate orthologues from each species. Ensembl databases were accessed using the the RMySQL package (https:// cran.r-project.org/web/packages/RMySQL/index.html).

Results

A chromosome level assembly

We assembled and annotated the genome of *Lophius piscatorius* (monkfish). The final scaffolded assembly length was 715 Mb long, which was consistent with k-mer based genome size estimates. 90% of the complete assembly (including contigs) was contained within a set of chromosome-sized scaffolds (47-19 million base pairs) (Fig. 1A,B). The remaining 10% was found within smaller contigs (~160 kb or less) that could not be successfully scaffolded (Fig. 2).

The contigs that were not included in the final scaffolds can be divided into two parts by their size; the majority of contigs and the majority of the sequence (more than 85%) was found in contigs up to 1000 bp (Fig. 2). However, there were also contigs of lengths up to 160 kb base pairs that were not scaffolded. We suspected these to comprise or contain sequences from *Spraguea lophii*, an intracellular parasite that is often found in members of the *Lophius* genus (Mansour et al. 2013; Campbell et al. 2013). To check this we searched the assembly for matches to the parasite genome sequence (NCBI, BioProject: PRJNA269798) using blastn. About 75% (by cumulated length) of matching sequences were found within the unscaffolded contigs. With the exception of 11 contigs, all contigs without matches (~370,000) to the parasite genome were shorter than 10,000 bp.

We also found around 19,000 loci within the main chromosomes that could be aligned to the parasite genome. However, these alignments represented only five parasite contigs, one of which could be aligned to the chromosomal scaffolds at almost 19,000 different loci. This sequence (GenBank: MQSS01000231.1) could also be aligned to multiple loci within a number of teleost species with up to 75% nucleotide identity. These observations suggest that this sequence is not a parasite sequence, but rather an *L. piscatorius* repeat sequence.

We also assessed the gene-space completeness of the assembly with the BUSCO software using the Actinopterygii-specific database. For comparison we included our initial draft genome assembly (BF1) (Dubin et al. In Press), contigs before scaffolding (BF2.c) and the final chromosome-level assembly (BF2.s). Of 4584 orthologues, 91% and 95% were found to be complete in the contig- and scaffold-level assemblies respectively (Fig. 1C).

These observations argue that our genome assembly is nearly complete with the exception of sequences from very repetitive regions that SPADES was unable to assemble into sufficiently large regions to be successfully scaffolded. It also suggests that although the assembly may contain some sequences derived from parasite DNA, that most exogenous sequences were excluded by the scaffolding process.

Repeats

Repeat masking using a de-novo generated repeat library and the Repbase database identified 32% of the genome (232 Mb) as repeats and low-complexity regions. For comparison, we performed the same analysis on the Atlantic cod (v2) (*Gadus morhua*), *T. rubripes* and *L. oculatus* genomes. 33.8, 17, 20% of these genomes were comprised of repeats respectively. The *L. piscatorius* genome appears to be dominated by the class 2 DNA transposons with the highest counts belonging to Tc1/mariner family (Fig. 4, table S1).

Gene predictions and features

The MAKER annotation pipepeline reported 45,552 candidate genes, which is considerably higher than the ~20,000 genes expected for a typical teleost genome (Table S2). We examined the distributions of exon, intron and gene lengths as well as the distribution of exon number per gene, and compared it to a set of other fish and the *Ciona intestinalis* genome (Fig. 3). The exon length distribution appears conserved for all of the examined species (Fig. 3A). A small peak of short exons that is specific to *L. piscatorius* is likely to represent erroneous ORF predictions.

The intron length distribution follows two distinct patterns (Fig. 3B). In all teleosts, including *L. piscatorius* we observe a bimodal distribution with two distinct peaks for short and long introns. In contrast both non-teleost species have a unimodal distribution with a peak that lies between the two teleost peaks. This suggests a teleost specific intron size reduction that may be related to the small genome sizes of teleosts. Furthermore, the log-normality of the distributions indicate that intron length is to some extent governed by exponential processes (i.e. long introns are more likely to change in length than short introns).

In general, gene length appears to be similar for all species, with most gene lengths lying in a broad peak between 256-32,000 bp long. In addition, stickleback, zebrafish and medaka have an additional peak of short genes at around 128 bp (Fig. 3D). *L. piscatorius* has an overall similar distribution but with higher counts, which probably reflects a high false positive prediction rate. The *L. piscatorius* distribution also has a left shoulder in the 256 to 1000 bp region which is not observed for any of the other species. Exon numbers per gene were highly conserved for all examined species and follow the same distribution pattern, with *L. piscatorius* having the highest counts presumably due to false positive gene predictions.

We also calculated the proportions of genomes comprising genes, exons, introns and coding regions before and after filtering based on reciprocal blast performed for annotation (Table 1). Prior to filtering, more than 80% of the genome was genic; after filtering by removal of genes for which we were unable to find similar proteins this reduced to 36% and resulted in a gene feature composition similar to that of the other teleosts analysed.

Gene annotation

We performed reciprocal blast for protein sequences predicted by the maker pipeline against the proteomes of 7 well annotated genomes. We chose 5 teleost species based both on the completeness of their assemblies (i.e. primarily chromosome level) and on the extent of their annotation (*Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes* and *Tetraodon nigroviridis*). In addition we included spotted gar (*Lepisosteus oculatus*) and *Ciona intestinalis* as outgroup species, which have not undergone either the teleost specific (*L. oculatus*) or any of the vertebrate specific gene duplications (*C. intestinalis*) (Dehal 2002; Braasch et al. 2016). The number and type of genes that have been predicted in these species varies considerably depending both on the species and the extent of annotation (Table S2). For example, more genes have been predicted in *D. rerio* for all gene categories except for lincRNA. This is more likely due to the greater amount of studies that have been performed in this species rather than because of a greater system complexity. In contrast the much smaller gene numbers observed in *C. intestinalis* are likely to reflect a biological difference rather than an experimental artefact.

Of the 45,552 protein sequences predicted by the maker pipeline only 23,864 could be aligned with an E-value of less than 1e-10 to sequences from any of the included species (Table 2, Figure 5). This, and the numbers of annotated coding regions typically found in teleosts, (Table S2) suggests that about half of the maker predictions reflect analytical artefacts. Approximately 18,000 of these alignments have a reciprocal rank of 1 in at least one species; that is, that same alignment has the top score in both the forward and reciprocal blast screens. Such alignments are more likely to represent simple one-to-one orthology and are more likely to represent true orthologues.

To assess the implication of non-reciprocal alignments we calculated a score-ratio (SR) as the alignment score between a given pair of *L. piscatorius* and Ensembl genes divided by the maximal alignment score for the Ensembl gene. A score less than, but close to one, suggests either gene duplication (in *L. piscatorius*) or gene loss (in the cognate species). Low scores, however, suggest the presence of gene fragments, or incomplete gene predictions. With the exception of alignments to *C. intestinalis*, the majority of alignments identified in the forward screen have score ratios of 1, with the largest number of such alignments identified in *O. latipes* (Fig. 6A, Table 2). Score ratios below 1 drop rapidly and at similar rates in all teleosts with the highest number of alignments being observed for *O.* latipes throughout the range. Of the teleost species, the fewest alignments are observed for *T. nigroviridis* at all score ratios. Of the vertebrates, *L. oculatus* (spotted gar) has the fewest reciprocal alignments (SR=1), but the score ratio does not drop as rapidly as for the teleosts and at lower ratio *L. oculatus* has a number of alignments similar to that observed for the teleosts. This is reasonable since *L. oculatus* is not descended from an ancestor that has undergone the teleost specific genome duplication and it is thus likely that there are more many-to-one orthologue relationships between *L. piscatorius* and *L. oculatus* than for the teleost species analysed here.

Curiously we found 93 maker proteins which had one-to-one orthologues only with *C. intestinalis* (Fig. S1). Whether these represent analytical artefacts or genes which have been lost from other teleost species is an interesting question, that needs further analysis to answer. Similarly we observed 126 genes that had a score ratio less than 1, but higher than 0.8 across all 7 species (Fig. S2). These genes may represent loophiformes specific gene duplications and should be inspected more closely.

The biggest difference in the rate of decrease of the score-ratio between *L. oculatus* and the teleosts is observed at score-ratios above 0.8, suggesting alignments more likely to include true orthologous pairings than at lower scores. Indeed, the distributions of alignment scores for alignments with score-ratios between 0.8 and 1 are similar (but shifted to the left) to those with the maximal score-ratio (Fig. 5B,C). However, the order of numbers of observed alignments is mirrored at intermediate alignment scores (2^{7}-2^{9}), with *C. intestinalis* and *L. oculatus* having the highest numbers of alignments with intermediate score-ratios. This pattern is consistent with such alignments representing real orthologies due to the additional genome duplications that have occurred in the teleost lineage.

Gene family composition

To assess to what extent candidate orthologues identified in separate species are consistent with each other we made use of the protein family classification provided by Ensembl compara (Herrero et al. 2016). All top scoring blast matches that had a score ratio of 0.8 or higher were considered as orthologues and used to assign family identifiers to the predicted genes. Gene families were then assigned to *L. piscatorius* genes by a simple plurality vote. Although the number of species from which we could identify orthologues varied, the majority of families was identical throughout (Fig 5.D).

We then asked whether we could observe any obvious increases or decreases in gene copy numbers based on the number of genes belonging to individual families by plotting the log- median ratio of gene family membership numbers for the different species. For the vast majority of families *L. piscatorius* contained the median number of genes (Fig. 7). This supports our gene annotation since the number of outliers is low. We were, however, intrigued to find an apparent loss of a number of genes known to be involved in immune responses (Ig and heavy chain V, an interleukin subunit and an immune receptor), since we have already demonstrated the loss of the MHCII pathway in *L. piscatorius* (Dubin et al. In Press). However, a closer examination found that these were not identified either because they are located to highly repetitive regions (Ig and heavy chain V) that had failed to assemble, or the genes appeared as part of fusion predictions.

Global syntenywith teleosts

In teleosts orthologues are commonly found on equivalent chromosome pairs; i.e. the set of genes found on a chromosome in a given species are predominantly located on a single chromosome in a second teleost species. If our orthology based annotation is generally correct we should be able to observe this in alignments between *L. piscatorius* and other teleosts. Indeed, chord diagrams linking gene positions between *L. piscatorius* and the teleost species (Fig. 8) clearly show, that in general, individual *L. piscatorius* scaffolds map to single chromosomes in the teleost species analysed. This relationship is strongest for *T. rubripes* and weakest for *D. rerio*, which is consistent with the expected phylogenetic relationship (Betancur-R et al. 2017). This strongly argues that the scaffolds in our assembly represent complete chromosomes of *L. piscatorius*.

Although we observe a one-to-one relationship between most *L. piscatorius* and *T. rubripes* chromosome pairs, the two longest *L. piscatorius* scaffolds (seq23 and seq1) map to two chromosomes in each of the teleost species analysed. This suggests either that these long scaffolds arose through chromosome fusion, or that they represent scaffolding artefacts. It may be possible to assess the likelihood of which explanation is true through further analysis of the evidence used for scaffolding, but experimental analysis (eg. chromosome painting (Ried 1998)) would be preferable. In the meanwhile we note that this observation is not specific to our assembly and that in fact some of the longer chromosomes in stickleback (*G. gasterosteus*), fugu (*T. rubripes*) and *T. nigroviridis* map to two chromosomes in our assembly in a reciprocal manner.

We also visualised the relationships between genome locations in orthologues as simple scatter plots. These largely show the same pattern, but in addition show evidence for a conserved order of genes along the chromosomes (Fig. 9). This order is readily observed in all of the teleost species with the exception of *D. rerio,* where the internal gene order appears to have been largely rearranged. These observations further support the quality of both our genome assembly and annotation.

Discussion

Scaffolding andsequence contamination

The *L. piscatorius* assembly described here was created through an initial contig assembly followed by scaffolding using Chicago (Putnam et al. 2016) and HiC libraries (Burton et al. 2013; Kaplan and Dekker 2013). The scaffolding utilised the increased probability of ligation between genome fragments physically close to each other, either in a reconstituted chromatin (Chicago) or within native chromatin (HiC). More than 90% of the completed assembly can be found within a set of chromosome sized scaffolds (47-30 million base pairs), with 10% found within smaller contigs that could not be included into the larger scaffolds [figure: 1]. The combined length of the chromosome sized scaffolds is similar to the estimated genome length and it seems likely that these scaffolds represent at least 90% of the genome.

The remaining 10% of the sequences fall into a set of short (<1000 bp) contigs that comprise 85% of the non-scaffolded sequences and a smaller set of longer contigs (up to 160,000 bp). The longer contigs appear to be comprised primarily of sequences derived from the intracellular parasite *S. lophii*. This parasite is intracellular (Mansour et al. 2013; Campbell et al. 2013) and hence it is difficult (though not impossible) to physically separate from the host DNA. Although we were able to find sequences from the chromosomal scaffolds that could be aligned to *S. lophii* these could only be aligned to 5 of the 439 parasite contigs and included only about 0.05% of the parasite genome. This argues that the scaffolding methods themselves (Chicago and HiC) are to a large extent able to mitigate the problems of contaminating sequences. This is reasonable, as the scaffolding process is able to separate sequences from individual molecules present within the same cell, which is arguably a much more difficult task. One of the sequences from *S. lophii* aligned at close to 19,000 locations within the chromosomal scaffolds and could also be aligned to multiple locations within other teleost species, suggesting that this sequence is derived from repetitive *L. piscatorius* sequences and may represent a contamination of the parasite genome with host DNA.

Assembly validation by annotation and global comparison

We have performed annotation based on de-novo gene prediction and orthology identification in order to facilitate downstream analyses of the genome. This process allowed us to assess the quality of the assembly since chromosome gene content and order is highly conserved among teleosts. The ability to observe a conserved chromosomal gene content relies on both the orthology classification to be generally correct and for the assembly to accurately represent chromosomes. We show that both the chromosomal gene content and order within our assembly is conserved in non-*D. rerio* teleosts and this validates both the assembly itself and the annotation.

We observed a strong conservation of protein family size in the set of annotated genes, with *L. piscatorius* having sizes typical for the set of species analysed. However, a closer inspection of some families that appear to contain missing members in *L. piscatorius* suggested that these were either present in highly repetitive regions that could not be scaffolded, or were missed by the annotation process due to being fused with other proteins. This demonstrates that neither our annotation nor our assembly is yet complete and emphasises the need for directed search strategies in order to show the loss of genes from a species (Dubin et al. In Press).

A typical teleost genome

Our assembly and annotation indicate that the *L. piscatorius* genome is a fairly typical teleost genome both in size, as well as in gene, exon and repeat contents. Although there is some apparent variation in gene and coding richness it should be remembered that even though we have used well characterized teleost genomes, the annotation and assemblies of these genomes should not be considered final. Hence, observed variances, will to some extent represent analytical artefacts in both our and prior annotation processes.

The distributions of gene feature (exons, introns, genes) were also similar to the other teleosts analysed here. Notably, we observed a teleost specific bimodal distribution in intron lengths. Although a bimodality of the intron length has been noted in *D. rerio* previously (Moss et al. 2011), the distributions of 4 other species (*G. aculeatus*, *O. latipes*, *T. rubripes*, and *T. nigroviridis*) were described as having 'monotonically decreasing frequency distributions'. The difference in our observations stem from the fact that we have inspected the distribution of log-transformed intron lengths. For the non-teleost species analysed here, this results in roughly uni-modal normal distributions; for all the teleosts we see two clearly separated peaks. Interestingly, the teleost species with the shortest genome included in this analysis (*G. aculeatus*, around 400 Mb) has the smallest long intron peak whereas the longest genome (*D. rerio*, 1.3Gb) has by far the largest long intron peak (Fig. 3B). This is consistent with a general relationship between intron and genome size observed across the vertebrates (Hara et al. 2018). Among the vertebrates, teleosts in general have small compact genomes, and our observations suggest that a reduction in intron size may be one of the mechanisms underlying a reduction in genome size after genome duplication.

The bimodality in the teleost intron length distributions are apparent only after log-transformation. For the non-teleost species log transformation results in something approaching a normal distribution, indicating that the mechanisms driving intron size distribution are not additive, but rather multiplicative. This implies that the rate of change in intron size is a function of the intron size itself, which is intuitive since large introns have a larger probability of being destinations for transposon translocations or to support internal sequence duplications (either by local transposon hopping or by meiotic slippage) (Rogozin et al. 2012; Huff et al. 2016). Similarly genetic rearrangements that can reduce intron size are more likely to occur in a large intron than a small intron. We thus argue that intron size should be considered in log space and that our observations relate to how teleost genomes have evolved.

Future prospects

The genome assembly and annotation presented here provides a rich resource for researching the genetic mechanisms that have allowed the anglerfishes to evolve into the huge range of weird and fantastic morphologies and behaviours we can observe today. In particular, we have an interest in identifying further modifications to the immune system that may have allowed the development of sexual parasitism. This is not only intrinsically interesting from an evolutionary perspective, but may also lead to insights with potential clinical application in reducing immune rejection after transplantation.

We also note that our annotation process in itself suggests the presence of genes which appear to have been specifically retained in the anglerfish clade; however, this will require further analyses from both this and related genome sequences in order to confirm and we look forward to the publication of additional anglerfish genomes.

The availability of an annotated genome and sequences from different individuals will also allow us to start to consider population genetic questions, including the effective population sizes and potential inbreeding. These are especially interesting for the anglerfishes as the range of reproductive strategies and the habitats which they occupy suggest difficulties in following more normal mating strategies. Understanding how populations change is especially important due to the large changes in marine environments caused by current and historical human activity. For species like the anglerfish that are not easily monitored this is most easily performed by population genetics approaches, and our assembly will greatly facilitate this.

Figure legends

Figure 1. A chromosome level assembly

A. Lengths and sequence depth (coverage) of the assembly scaffolds. The final scaffolds had sizes typical of chromosomes ranging from 47 to 25 Mb (Mega-bases). The sequencing depth was generally even across the genome, though distinct regions of deeper coverage can be observed on all chromosomes; these are likely to contain repetitive sequences. B. Scaffold and contig lengths. Each point represents a single scaffold, with positions along the X-axis giving the cumulative length of the preceding scaffolds. The N-50 and N-90 bands are indicated. The assembly is cleanly divided into a set of chromosome length scaffolds and a set of contigs which are orders of magnitude shorter. C. Gene space completeness. The stacked segments indicate the proportion of Actinopterygii conserved genes that were detected, either completely or fragmented across contigs, or missing. BF1 refers to a draft assembly prepared previously (Dubin et al. In Press) from a separate individual. BF2.c and BF2.s refer to the assembly prior to and after scaffolding respectively. Comp.D refers to genes that appear to have been duplicated in our assembly.

Figure 2. Parasite DNAin unscaffolded contigs

Lengths and parasite content of unscaffolded contigs. Each point represents a single scaffold, with positions along the X-axis giving the cumulative length of the preceding contigs. Red points contain matches to the sequences from the intracellular parasite *Spraguea lophii*. These are present primarily in the longer contigs. About 85% of the total sequences were present in short contigs that cannot be aligned to the parasite genome.

Figure3. Gene feature properties

Distributions of the sizes of exons (A), introns (B) and genes (D) as well as the number of exons per gene (C) for monkfish (L. pis), medaka (O. lat*)*, spotted gar (L. ocu*)*, stickleback (G. acu*)*, zebra fish (D. rer*)* and ciona (C. int*)*. The length of exons and genes, as well as the number of exons per gene have similar distributions across the 6 species, though the gene lengths are generally shorter for *Ciona*. In contrast, intron lengths have a characteristic bimodal distribution only in the teleost species.

Figure4. Repeat abundance by class

Repeat abundances for *L. piscatorius* (L. pis), *T. rubripes* (T. fug), *G. morhua* (G. mor) and *L. oculatus* (L.ocu) based on a de-novo generated repeat library and the Repbase database (GIRI). *T. rubripes is* an outlier having the smallest amount of transposable elements. This is correlated with genome reduction in this species and is documented in literature (Aparicio 2002; Guo et al. 2010). All species have a relatively high abundances of class 2 DNA transposons, dominated by the Tc1/mariner family. *L. oculatus* and *L. piscatorius* have a high counts of LINE type repeats compared to the other species. All species contain a large portion of unclassified repetitive elements.

Figure5. Reciprocal blast alignments

Numbers of candidate genes predicted by the Maker pipeline for which blast alignments with an e-value (random expectation) less than 1e-10 could be found. The total set of genes was further divided by the score-ratio parameter (SR) and the proportion of the subject (SC) or the query (QC) that were included in the alignment. The score ratio refers to the fraction of the alignment score to the maximum alignment score for the given Ensembl gene (score[i,g] / max(score[,g])). The subject in this case is the Ensembl gene (the candidate orthologue) and the query is the Maker predicted peptide (the candidate gene). A total of 45,552 genes were predicted by the maker pipeline. Of these about half could be aligned to known Ensembl proteins. Abbreviations: *Ciona intestinalis* (C. int), *Danio rerio* (D. rer), *Gasterosteus aculeatus* (G. acu), *Lepisosteus oculatus* (L. ocu), *Oryzias latipes* (O. lat), *Takifugu rubripiens* (T. rub), *Tetraodon nigroviridis* (T.nig), score ratio (SR), subject coverage (SC), query coverage (QC).

Figure 6. Orthology assignment

A. Quantiles plot of the score ratios for pairs of *L. piscatorius* gene candidates and Ensembl candidate orthologues. The score-ratio for a given pair of genes is the ratio of the alignment score for a given pair and the maximal alignment score for that Ensembl gene. Only a single score is plotted for each *L. piscatorius* gene candidate per species. For alignments with vertebrates these mostly have score ratios of 1, indicating that most genes have a single orthologue in the cognate species. Fewer alignments with score ratios of 1 were observed for *C. intestinalis* and *L. oculatus* (spotted gar) which is consistent with those genomes not having undergone either the vertebrate genome duplication (*C. intestinalis*), or the teleost specific duplication (*L. oculatus*). B. and C. Quantiles plots of alignment scores for alignments with score ratios of 1 (B) or between 0.8 and 1 (C). D. Candidate genes were assigned to protein families by a plurality vote of the family memberships of their candidate orthologues. Numbers indicate the number of candidate genes for each class of vote (number of orthologues voting) and the maximal number of votes (winning vote). Thus 4197 candidate genes had candidate orthologues in all 7 species, all of which belong to the same protein family, whereas 2183 genes had orthologues in all species, but one of these belonged to a separate family giving a winning score of 6.

Abbreviations: *Ciona intestinalis* (C. int), *Danio rerio* (D. rer), *Gasterosteus aculeatus* (G. acu), *Lepisosteus oculatus* (L. ocu), *Oryzias latipes* (O. lat), *Takifugu rubripiens* (T. rub), *Tetraodon nigroviridis* (T.nig).

Figure 7. Family sizes

Candidate *L. piscatorius* genes were assigned to families based on the family membership of their candidate orthologues. The inferred family sizes (numbers of member genes) were calculated and compared to the median family size across the 7 species analysed. Each point represents a single gene family and its position in Y is the log2 of the ratio of the *L. piscatorius* (red) or Ensembl species (grey) family size to the median family sizes. All family sizes were incremented by one to allow for log-transformation. The family sizes in *L. piscatorius* are highly consistent with the typical family sizes.

Figure 8. Chromosomal orthology

Arc diagrams showing connections between the genome locations of orthologue pairs between *L. piscatorius* and *D. rerio* (A), and *T. rubripes* (B) show that orthologue pairs tend to be found on conserved chromosomes. This relationship appears stronger between *L. piscatorius* and *T. rubripes*. *L. piscatorius* chromosomes are indicated by blue arc segments. Red arcs indicate *D. rerio* and *T. rubripes* chromosomes.

Figure 9. Chromosomal synteny

Plots of positions of genes in *L. piscatorius* and candidate orthologues in *D. rerio* and *T. rubripes*. The colour of the points indicate the strand in the two species (+/+ purple, +/- red, -/ + blue, -/- black for the strands of the *L. piscatorius*/cognate species). The chromosome gene content is clearly conserved in both species, but the internal chromosomal order in *D. rerio* is scrambled. In contrast long regions of chromosomal synteny can be observed between *L. piscatorius* and *T. rubripes* as diagonal lines of points.

Table legends

Table 1. Genome composition

Genome sizes (column 1) and the proportion of the indicated genomes occupied by genes, exons, introns and coding regions. Row names indicate the names of the Ensembl databases.

Table 2. Reciprocal blast alignments

Numbers of alignments visualised in Figure 5.

Abbreviations: *Ciona intestinalis* (C. int), *Danio rerio* (D. rer), *Gasterosteus aculeatus* (G. acu), *Lepisosteus oculatus* (L. ocu), *Oryzias latipes* (O. lat), *Takifugu rubripiens* (T. rub), *Tetraodon nigroviridis* (T.nig), score ratio (SR), subject coverage (SC), query coverage (QC).

Supplementary figures

Figure S1. Counts of reciprocal alignments for species combinations

Number of *L. piscatorius* gene candidates that could be aligned to sequences with a reciprocal rank of 1 in different species. Left panel indicates the species in which alignments were found. The numbers and bars following that indicate the number of alignments for each species combination. The final two columns (numbers & bars) show the cumulative number of candidates aligned.

Abbreviations: *Ciona intestinalis* (C. int), *Danio rerio* (D. rer), *Gasterosteus aculeatus* (G. acu), *Lepisosteus oculatus* (L. ocu), *Oryzias latipes* (O. lat), *Takifugu rubripiens* (T. rub), *Tetraodon nigroviridis* (T.nig).

Figure S2. Counts of non-reciprocal alignments for species combinations

Number of *L. piscatorius* gene candidates that could be aligned to sequences with a reciprocal score ratio between 1 and 0.8 in different species. Data visualised as in Fig. S2.

Supplementary tables

Table S1. Gene numbers in Ensembl species used

Numbers of different categories of genes reported by Ensembl version 97 for species used in these analyses.

References.

Abrahimi P , Qin L, Chang WG, et al (2016) Blocking MHC class II on human endothelium mitigates acute rejection. JCI Insight 1:. doi: 10.1172/jci.insight.85293

Alfaro ME, Faircloth BC, Harrington RC, et al (2018) Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. Nature Ecology & Evolution 2:688–696. doi: 10.1038/s41559-018-0494-6

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Aparicio S (2002) Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. Science 297:1301–1310. doi: 10.1126/science.1072104

Arnold RJ, Pietsch TW (2018) Fantastic Beasts and Where to Find Them: A New Species of the Frogfish Genus *Histiophryne* Gill (Lophiiformes: Antennariidae: *Histiophryninae*) from Western and South Australia, with a Revised Key to Congeners. Copeia 106:622–631. doi: 10.1643/CI-18-112

Bankevich A, Nurk S, Antipov D, et al (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19:455–477. doi: 10.1089/cmb.2012.0021

Betancur-R R, Wiley EO, Arratia G, et al (2017) Phylogenetic classification of bony fishes. BMC Evolutionary Biology 17:. doi: 10.1186/s12862-017-0958-3

Borodovsky M, Lomsadze A (2011) Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES. Current Protocols in Bioinformatics 35:4.6.1-4.6.10. doi: 10.1002/0471250953.bi0406s35

Braasch I, Gehrke AR, Smith JJ, et al (2016) The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics 48:427–437. doi: 10.1038/ng.3526

Burton JN, Adey A, Patwardhan RP, et al (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology 31:1119– 1125. doi: 10.1038/nbt.2727

Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi: 10.1186/1471-2105-10-421

Campbell SE, Williams TA, Yousuf A, et al (2013) The Genome of *Spraguea lophii* and the Basis of Host-Microsporidian Interactions. PLoS Genetics 9:e1003676. doi: 10.1371/journal.pgen.1003676

Dehal P (2002) The Draft Genome of *Ciona intestinalis*: Insights into Chordate and Vertebrate Origins. Science 298:2157–2167. doi: 10.1126/science.1080049

Dickson BV, Pierce SE (2019) How (and why) fins turn into limbs: insights from anglerfish. Earth and Environmental Science Transactions of the Royal Society of Edinburgh 109:87–103. doi: 10.1017/S1755691018000415

Dubin A, Jørgensen TE, Moum T, et al (In Press) Complete loss of the MHC II pathway in an anglerfish, Lophius piscatorius. Biol Lett

Farina AC, Azevedo M, Landa J, et al (2008) Lophius in the world: a synthesis on the common features and life strategies. ICES Journal of Marine Science 65:1272–1280

Grabherr MG, Haas BJ, Yassour M, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644–652. doi: 10.1038/nbt.1883

Guo B, Zou M, Gan X, He S (2010) Genome size evolution in pufferfish: an insight from BAC clone-based Diodon holocanthus genome sequencing. BMC Genomics 11:396. doi: 10.1186/1471-2164-11-396

Haase D, Roth O, Kalbe M, et al (2013) Absence of major histocompatibility complex class II mediated immunity in pipefish, *Syngnathus typhle*: evidence from deep transcriptome sequencing. Biology Letters 9:20130044–20130044. doi: 10.1098/rsbl.2013.0044

Hansen K, Herring PJ (2009) Dual bioluminescent systems in the anglerfish genus *Linophryne* (Pisces: *Ceratioidea*). Journal of Zoology 182:103–124. doi: 10.1111/j.1469-7998.1977.tb04144.x

Hara Y, Yamaguchi K, Onimaru K, et al (2018) Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. Nature Ecology & Evolution 2:1761–1771. doi: 10.1038/s41559-018-0673-5

Herrero J, Muffato M, Beal K, et al (2016) Ensembl comparative genomics resources. Database (Oxford) 2016:. doi: 10.1093/database/bav096 Ho H-C (2016) Records of deep-sea anglerfishes (Lophiiformes: Ceratioidei) from Indonesia, with descriptions of three new species. Zootaxa 4121:267. doi: 10.11646/zootaxa.4121.3.3

Ho H-C, Ma W-C (2016) Revision of southern African species of the anglerfish genus *Chaunax* (Lophiiformes: *Chaunacidae*), with descriptions of three new species. Zootaxa 4144:175. doi: 10.11646/zootaxa.4144.2.2

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:. doi: 10.1186/1471-2105-12-491

Hubley R, Finn RD, Clements J, et al (2016) The Dfam database of repetitive DNA families. Nucleic Acids Research 44:D81–D89. doi: 10.1093/nar/gkv1272 Huff JT, Zilberman D, Roy SW (2016) Mechanism for DNA transposons to generate introns on genomic scales. Nature 538:533–536. doi: 10.1038/nature20110

Hulet WH, Musil G (1968) Intracellular Bacteria in the Light Organ of the Deep Sea Angler Fish, *Melanocetus murrayi*. Copeia 1968:506. doi: 10.2307/1442019

Issac P , Robert M, Le Bris H, et al (2017) Investigating feeding ecology of two anglerfish species, Lophius piscatorius and *Lophius budegassa* in the Celtic Sea using gut content and isotopic analyses. Food Webs 13:33–37. doi: 10.1016/j.fooweb.2017.08.001

Kai W, Kikuchi K, Tohari S, et al (2011) Integration of the Genetic Map and Genome Assembly of Fugu Facilitates Insights into Distinct Features of Genome Evolution in Teleosts and Mammals. Genome Biology and Evolution 3:424–442. doi: 10.1093/gbe/evr041

Kaplan N, Dekker J (2013) High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature Biotechnology 31:1143–1147. doi: 10.1038/nbt.2768

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59. doi: 10.1186/1471- 2105-5-59

Lomsadze A (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research 33:6494–6506. doi: 10.1093/nar/gki937

Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. doi: 10.1093/bioinformatics/btr507

Malmstrøm M, Matschiner M, Tørresen OK, et al (2016) Evolution of the immune system influences speciation rates in teleost fishes. Nature Genetics 48:1204–1210. doi: 10.1038/ng.3645

Mansour L, Ben Hassine OK, Vivares CP, Cornillot E (2013) Spraguea lophii (Microsporidia) parasite of the teleost fish, Lophius piscatorius from Tunisian coasts: Evidence for an extensive chromosome length polymorphism. Parasitology International 62:66–74. doi: 10.1016/j.parint.2012.09.007

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10. doi: 10.14806/ej.17.1.200

Miya M, Pietsch TW, Orr JW, et al (2010) Evolutionary history of anglerfishes (Teleostei: Lophiiformes): a mitogenomic perspective. BMC Evolutionary Biology 10:58

Moss SP, Joyce DA, Humphries S, et al (2011) Comparative Analysis of Teleost Genome Sequences Reveals an Ancient Intron Size Expansion in the Zebrafish Lineage. Genome Biology and Evolution 3:1187–1196. doi: 10.1093/gbe/evr090

Munk O (2000) Histology of the fusion area between the parasitic male and the female in the deep-sea anglerfish *Neoceratias spinifer* Pappenheim, 1914 (Teleostei, Ceratioidei). Acta Zoologica 81:315–324

Nagareda BH, Shenker J (2009) Evidence for chemical luring in the polka-dot batfish *Ogcocephalus cubifrons* (Teleostei: Lophiiformes: *Ogcocephalidae*). Florida Scientist 72:11–17

O’day WT (1974) Bacterial luminescence in the deep-sea anglerfish Oneirodes acanthias (Gilbert, 1915). Contributions in science 255:1–12

Ou S, Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for

Identification of Long Terminal Repeat Retrotransposons. Plant Physiology 176:1410–

1422. doi: 10.1104/pp.17.01310

Pietsch Theodore W, Arnold Rachel J, Hall David J (2009a) A Bizarre New Species of Frogfish of the Genus *Histiophryne* (Lophiiformes: *Antennariidae*) from Ambon and Bali, Indonesia. Copeia 2009:37–45. doi: 10.1643/CI-08-129

Pietsch TW (1976) Dimorphism, Parasitism and Sex: Reproductive Strategies among Deepsea Ceratioid Anglerfishes. Copeia 1976:781. doi: 10.2307/1443462

Pietsch TW (2005) Dimorphism, parasitism, and sex revisited: modes of reproduction among deep-sea ceratioid anglerfishes (Teleostei: Lophiiformes). Ichthyological Research 52:207–236. doi: 10.1007/s10228-005-0286-2

Pietsch TW (2009) Oceanic anglerfishes: extraordinary diversity in the deep sea. University of California Press, Berkeley

Pietsch TW, Johnson JW, Arnold RJ (2009b) A New Genus and Species of the Shallow-Water Anglerfish Family Tetrabrachiidae (Teleostei: Lophiiformes: Antennarioidei) from Australia and Indonesia. Copeia 2009:483–493. doi: 10.1643/CI-08-192

Pietsch TW, Orr JW (2007) Phylogenetic Relationships of Deep-Sea Anglerfishes of the Suborder Ceratioidei (Teleostei: Lophiiformes) Based on Morphology. Copeia 2007:1–34

Pietsch TW, Sutton TT (2015) A New Species of the Ceratioid Anglerfish Genus *Lasiognathus* Regan (Lophiiformes: *Oneirodidae*) from the Northern Gulf of Mexico. Copeia 103:429–432. doi: 10.1643/CI-14-181

Putnam NH, O’Connell BL, Stites JC, et al (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Research 26:342–350. doi: 10.1101/gr.193474.115

Rajeeshkumar MP, Meera KM, Hashim M (2017) A New Species of the Deep-Sea Ceratioid Anglerfish Genus *Oneirodes* (Lophiiformes: *Oneirodidae*) from the Western Indian Ocean. Copeia 105:82–84. doi: 10.1643/CI-16-467

Ried T (1998) Chromosome painting: a useful art. Human Molecular Genetics 7:1619–1626. doi: 10.1093/hmg/7.10.1619

Rogozin IB, Carmel L, Csuros M, Koonin EV (2012) Origin and evolution of spliceosomal introns. Biology Direct 7:11. doi: 10.1186/1745-6150-7-11

Rosenberg AS, Singer A (1992) Cellular Basis of Skin Allograft Rejection: An In Vivo Model of Immune-Mediated Tissue Destruction. Annu Rev Immunol 10:333–360. doi: 10.1146/annurev.iy.10.040192.002001

Shedlock AM, Pietsch TW, Haygood MG, et al (2004) Molecular systematics and life history evolution of anglerfishes (Teleostei: Lophiiformes): Evidence from mitochondrial DNA. Steenstrupia 28:129–144

Star B, Nederbragt AJ, Jentoft S, et al (2011) The genome sequence of Atlantic cod reveals a unique immune system. Nature 477:207–210. doi: 10.1038/nature10342

Thangstad T (2006) Anglerfish (Lophius spp) in Nordic waters. Nordic Council of Ministers

Tørresen OK, Star B, Jentoft S, et al (2017) An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics 18:. doi: 10.1186/s12864-016-3448-x

Varadharajan S, Sandve SR, Gillard GB, et al (2018) The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication. Genome Biology and Evolution 10:2785–2800. doi: 10.1093/gbe/evy201

Waldmann H (2010) Tolerance: an overview and perspectives. Nature Reviews Nephrology 6:569–576. doi: 10.1038/nrneph.2010.108

Waterhouse RM, Seppey M, Sim„o FA, et al (2018) BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Molecular Biology and Evolution 35:543–548. doi: 10.1093/molbev/msx319

Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. Bioinformatics 29:2487–2489. doi: 10.1093/bioinformatics/btt403

Zerbino DR, Achuthan P, Akanni W, et al (2018) Ensembl 2018. Nucleic Acids Res 46:D754–D761. doi: 10.1093/nar/gkx1098

Figure 1. A chromosome level assembly

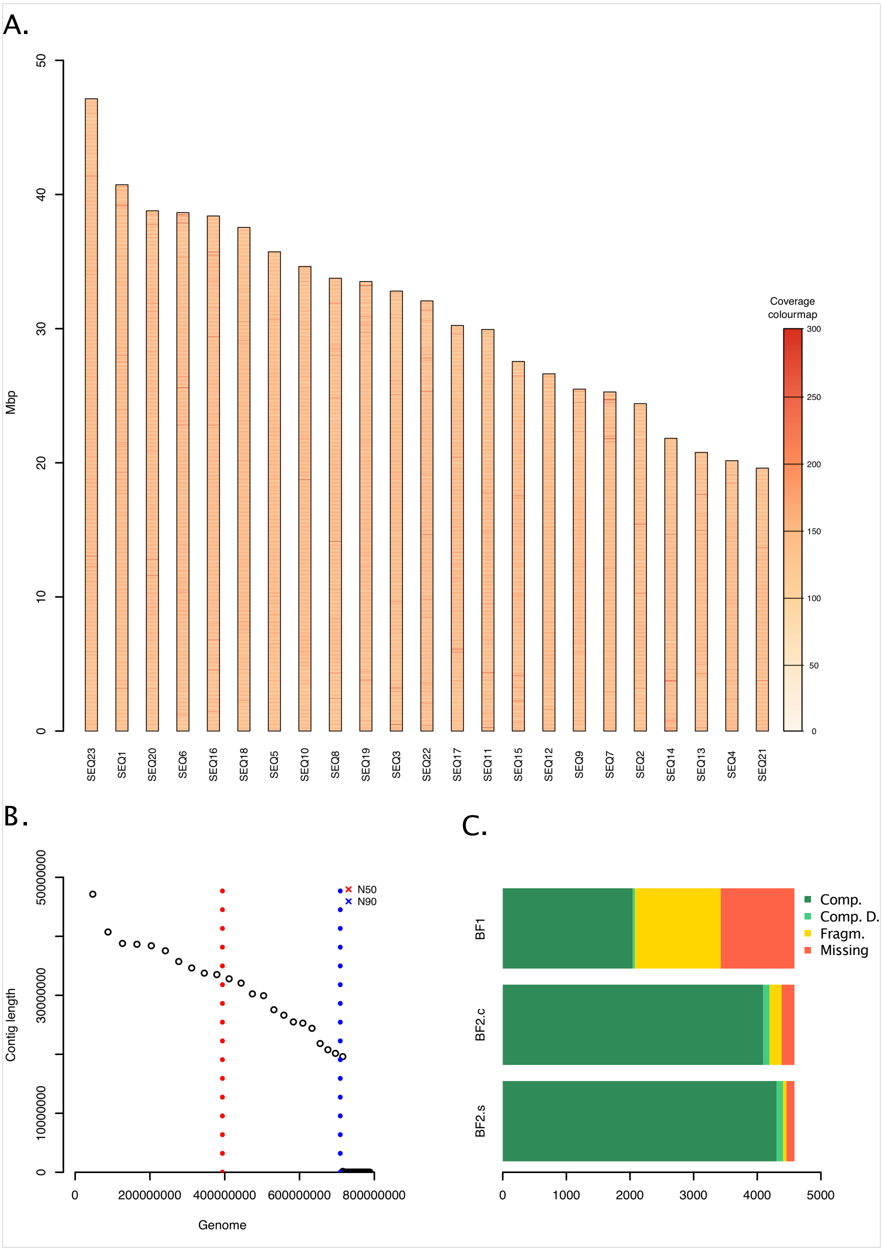


Figure 2. Parasite DNA in unscaffolded contigs

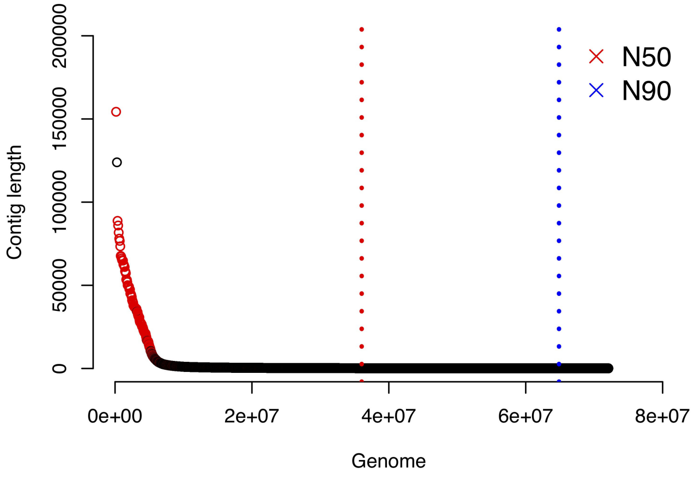
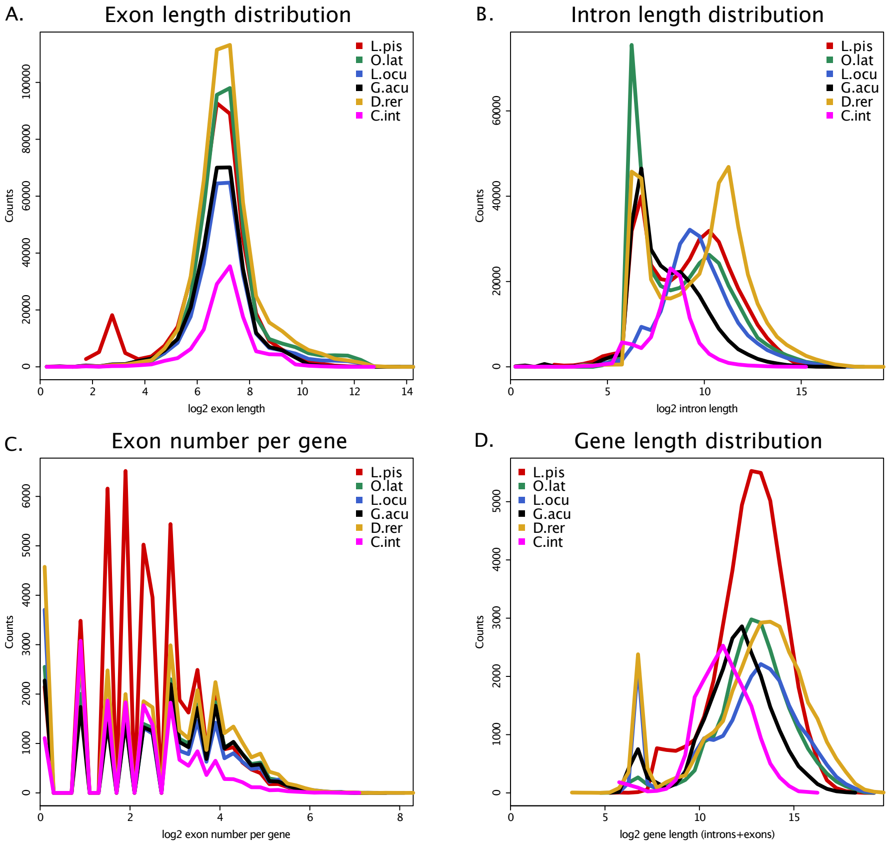


Figure 3. Gene feature properties



A. Exon length distribution

B. Intron length distribution

C. Exon number per gene

D. Gene length distribution

Figure 4. Repat abundance by class

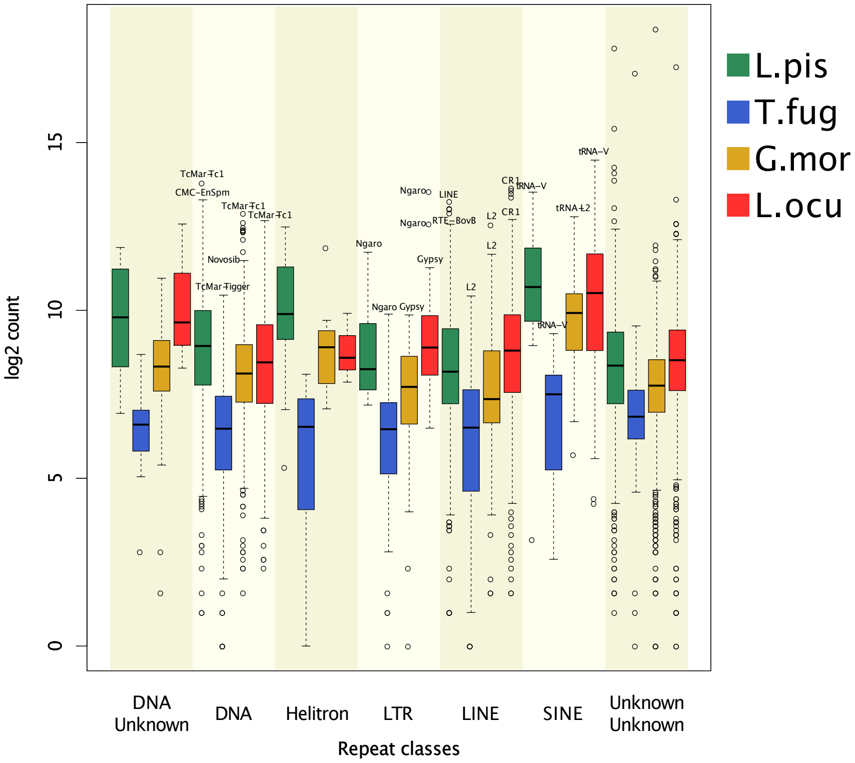


Figure 5. Reciprocal blast alignments

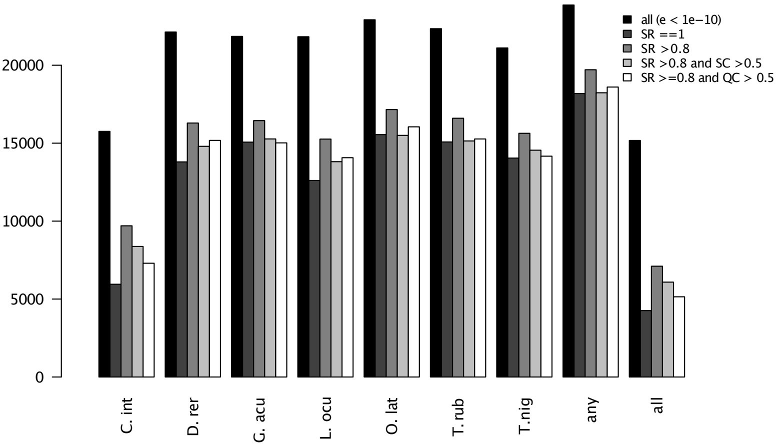
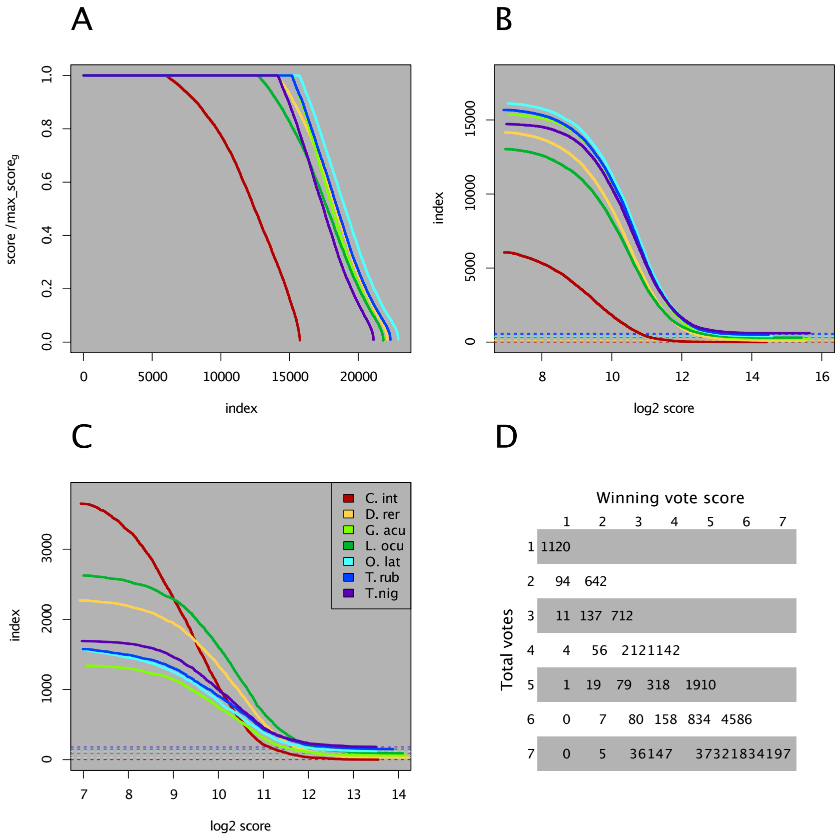


Figure 6. Orthology assignment



	1	2	3	4	5	6	7
1	1120						
2	94	642					
3	11	137	712				
4	4	56	212	1142			
5	1	19	79	318	1910		
6	0	7	80	158	834	4586	
7	0	5	36	147	373	2183	4197

Figure 7. Family sizes

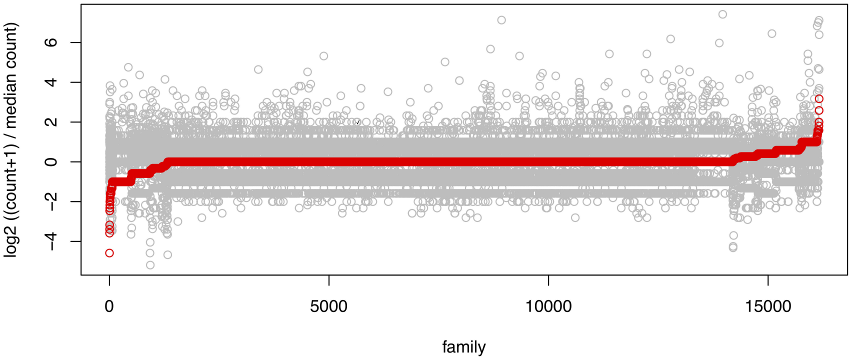


Figure 8. Chromosomal orthology

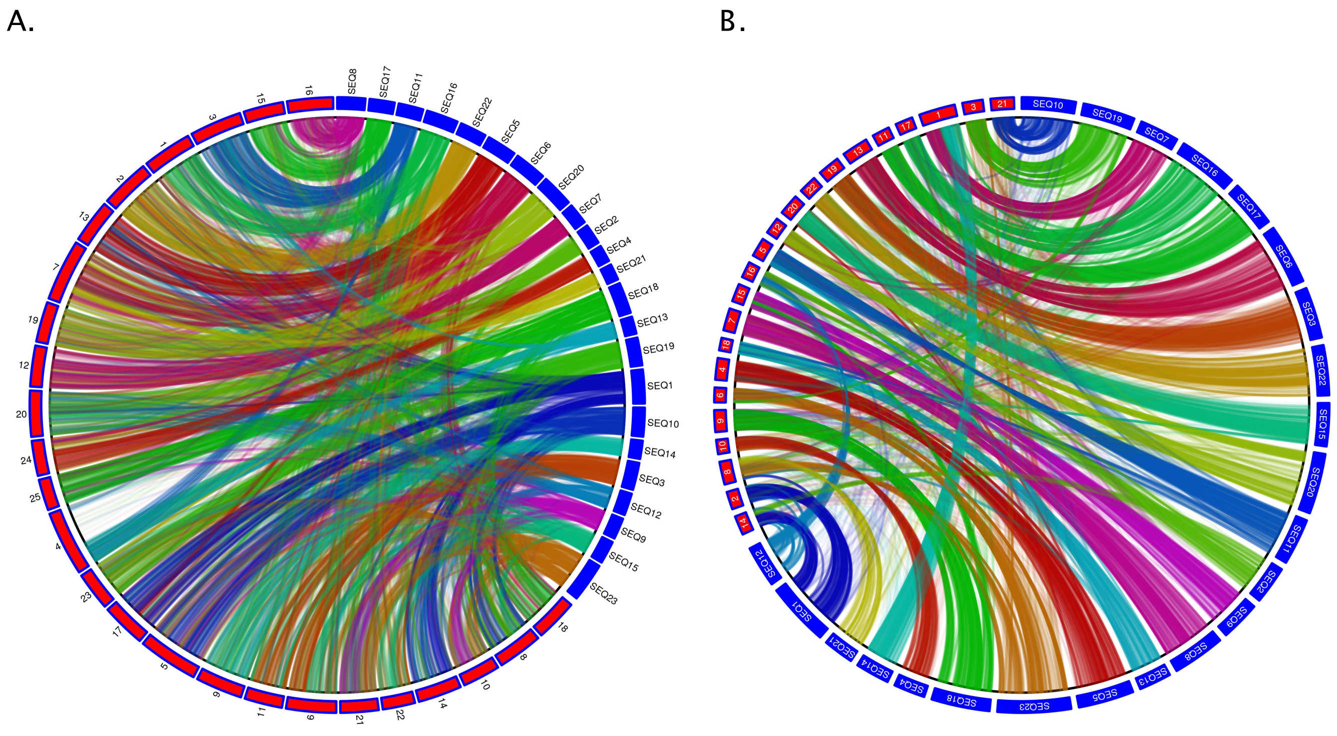


Figure 9. Chromosomal synteny

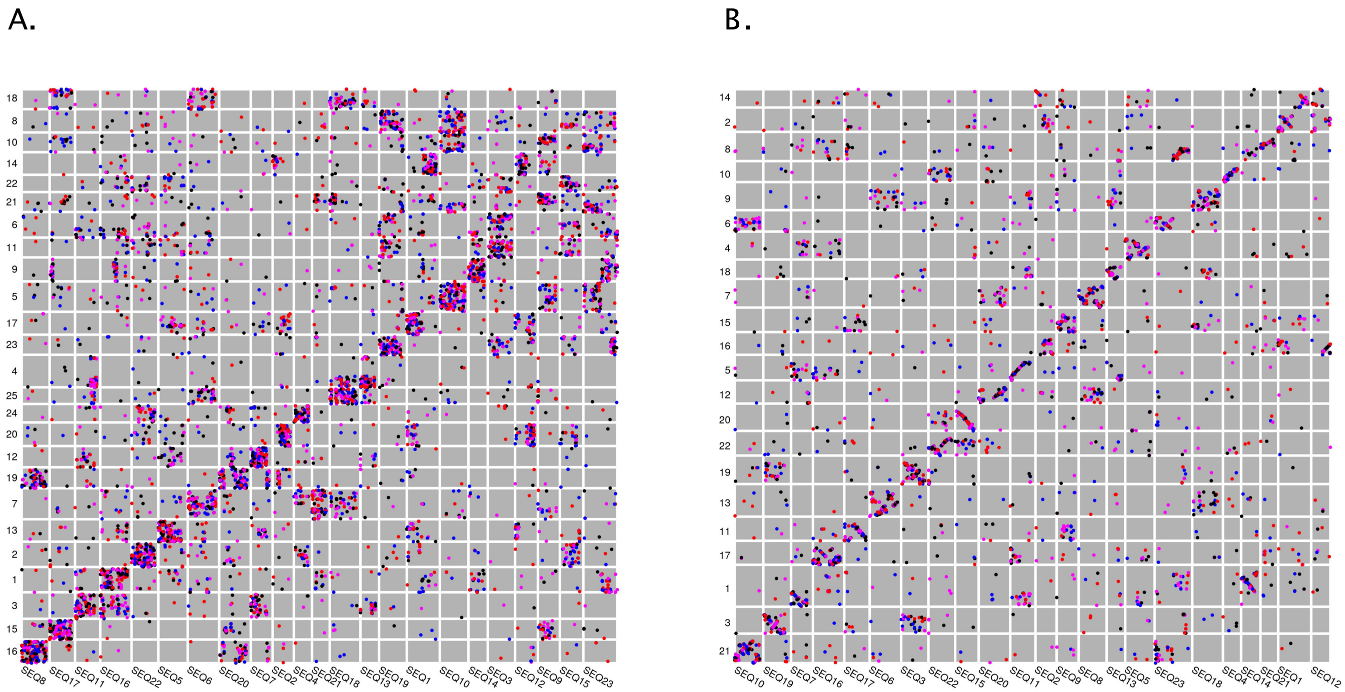


Table 1. Genome composition

genome		genic	exon	intron	coding
Lophius all	715681327	0.818314547139219	0.0706292229418471	0.747685324197371	0.0648653559183891
Lophius (with orthologue)	715681327	0.363496875195124	0.0501039424212894	0.313392932773835	0.0450475648640194
ciona_intestinalis_core_97_3	78310945	0.622732454575794	0.260289375897584	0.36244307867821	0.205589231492482
danio_rerio_core_97_11	1345118429	0.584661264796336	0.0468447711677289	0.537816493628607	0.0356981013453946
gasterosteus_aculeatus_core_97_1	400804237	0.433476333235469	0.0885441837282773	0.344932149507192	0.0802355515019169
lepisosteus_oculatus_core_97_1	891160407	0.502272451159289	0.0552529068989484	0.447019544260341	0.0348137021756174
oryzias_latipes_core_97_1	734057086	0.576981076918587	0.0904012961193593	0.486579780799228	0.0536397723705102
takifugu_rubripes_core_97_5	391484725	0.45922010622509	0.119514877623897	0.339705228601193	0.0791520844140215
tetraodon_nigroviridis_core_97_8	358618246	0.345913852358756	0.0864601183733412	0.259453733985415	0.0827950594571811

Table 2. Reciprocal blast alignments

	C. int	D. rer	G. acu	L. ocu	O. lat	T.rub	T.nig	any	all
all (e < 1e−10)	15756	22132	21850	21829	22917	22344	21110	23864	15174
SR == 1	5950	13799	15067	12607	15550	15079	14044	18184	4260
SR >0.8	9700	16292	16448	15261	17157	16599	15633	19711	7107
SR > 0.8 and SC > 0.5	8375	14798	15272	13814	15500	15143	14549	18234	6086
SR >=0.8 and QC > 0.5	7297	15178	15019	14069	16049	15270	14165	18601	5145

Figure S1. Counts of reciprocal alignments for species combinations

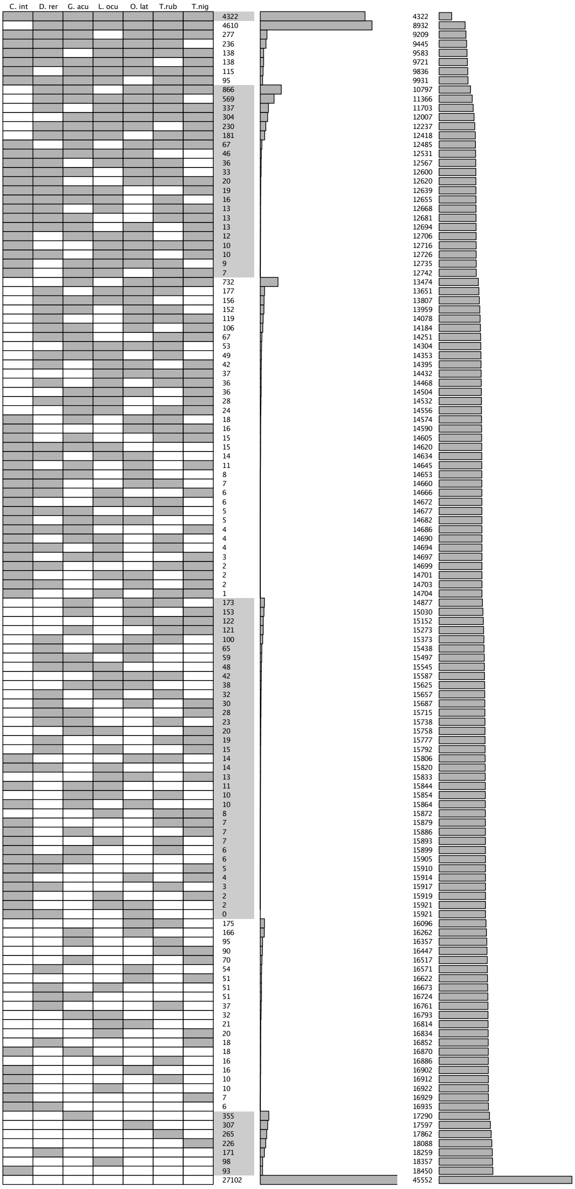


Figure S2. Counts of non-reciprocal alignments for species combinations

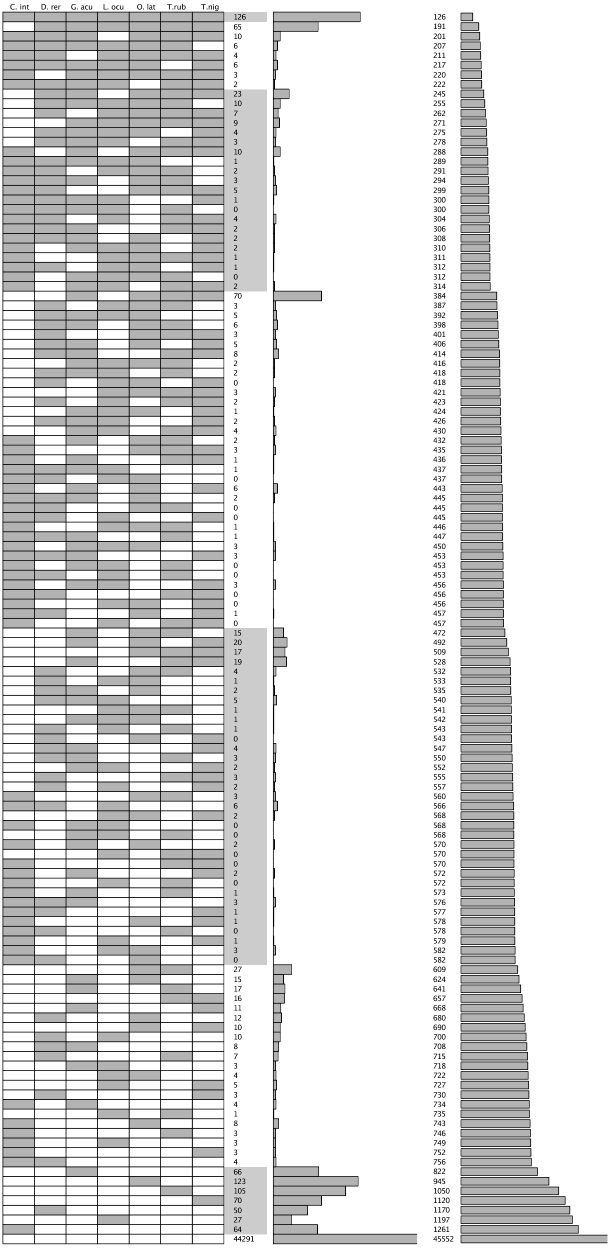


Table S1. Gene numbers in Ensembl species used

	C. int	D. rer	G. acu	L. ocu	O. lat	T.rub	T.nig
protein_coding	11806	25094	18342	17225	23587	20503	19589
rRNA	4	1919	67	1841	63	80	80
lincRNA	0	1466	0	1969	0	0	0
miRNA	161	431	390	243	230	153	397
snoRNA	51	242	227	146	189	170	177
processed_transcript	0	1045	0	0	0	0	0
snRNA	21	487	65	104	99	68	128
antisense	0	672	0	0	0	0	0
pseudogene	17	12	40	31	26	34	141
misc_RNA	10	93	10	25	59	23	7
unprocessed_pseudogene	0	221	0	0	0	0	0
TR_J_gene	0	76	0	0	3	0	0
TR_V_gene	0	70	0	0	0	0	0
sense_intronic	0	57	0	0	0	0	0
sRNA	0	4	0	0	41	1	0
IG_V_gene	0	0	0	0	16	28	0
processed_pseudogene	0	27	0	4	1	1	6
transcribed_unprocessed_pseudogene	0	28	0	0	0	0	0
scaRNA	0	11	0	0	8	8	0
IG_V_pseudogene	0	17	0	0	0	0	0
TEC	0	14	0	0	0	0	0
polymorphic_pseudogene	0	10	0	0	0	0	0
ribozyme	0	4	0	0	3	2	0
sense_overlapping	0	7	0	0	0	0	0
IG_J_gene	0	0	0	0	3	1	0
IG_C_pseudogene	0	3	0	0	0	0	0
IG_J_pseudogene	0	3	0	0	0	0	0
IG_C_gene	0	2	0	0	0	0	0
IG_pseudogene	0	2	0	0	0	0	0
TR_D_gene	0	2	0	0	0	0	0
TR_V_pseudogene	0	1	0	0	0	0	0

List of previously published theses for PhD in Aquaculture / PhD in Aquatic Biosciences, Nord University

No. 1 (2011)
PhD in Aquaculture
Chris André Johnsen
Flesh quality and growth of farmed Atlantic salmon (*Salmo salar* L.) in relation to feed, feeding, smolt type and season
ISBN: 978-82-93165-00-2

No. 2 (2012)
PhD in Aquaculture
Jareeporn Ruangsri
Characterization of antimicrobial peptides in Atlantic cod
ISBN: 978-82-93165-01-9

No. 3 (2012)
PhD in Aquaculture
Muhammad Naveed Yousaf
Characterization of the cardiac pacemaker and pathological responses to cardiac diseases in Atlantic salmon (*Salmo salar* L.)
ISBN: 978-82-93165-02-6

No. 4 (2012)
PhD in Aquaculture
Carlos Frederico Ceccon Lanes
Comparative Studies on the quality of eggs and larvae from broodstocks of farmed and wild Atlantic cod
ISBN: 978-82-93165-03-3

No. 5 (2012)
PhD in Aquaculture
Arvind Sundaram
Understanding the specificity of the innate immune response in teleosts: Characterisation and differential expression of teleost-specific Toll-like receptors and microRNAs
ISBN: 978-82-93165-04-0

No. 6 (2012)
PhD in Aquaculture
Teshome Tilahun Bizuayehu
Characterization of microRNA during early ontogeny and sexual development of Atlantic halibut (*Hippoglossus hippoglossus* L.)
ISBN: 978-82-93165-05-7

No. 7 (2013)
PhD in Aquaculture
Binoy Rajan
Proteomic characterization of Atlantic cod skin mucosa – Emphasis on innate immunity and lectins
ISBN: 978-82-93165-06-04

No. 8 (2013)
PhD in Aquaculture
Anusha Krishanthi Shyamali Dhanasiri
Transport related stress in zebrafish: physiological responses and bioremediation
ISBN: 978-82-93165-07-1

No. 9 (2013)
PhD in Aquaculture
Martin Haugmo Iversen
Stress and its impact on animal welfare during commercial production of Atlantic salmon (*Salmo salar* L.)
ISBN: 978-82-93165-08-8

No. 10 (2013)
PhD in Aquatic Biosciences
Alexander Jüterbock
Climate change impact on the seaweed *Fucus serratus*, a key foundational species on North Atlantic rocky shores
ISBN: 978-82-93165-09-5

No. 11 (2014)
PhD in Aquatic Biosciences
Amod Kulkarni
Responses in the gut of black tiger shrimp *Penaeus monodon* to oral vaccine candidates against white spot disease
ISBN: 978-82-93165-10-1

No. 12 (2014)
PhD in Aquatic Biosciences
Carlo C. Lazado
Molecular basis of daily rhythmicity in fast skeletal muscle of Atlantic cod *(Gadus morhua)*
ISBN: 978-82-93165-11-8

No. 13 (2014)
PhD in Aquaculture
Joanna Babiak
Induced masculinization of Atlantic halibut (*Hippoglossus hippoglossus* L.): towards the goal of all-female production
ISBN: 978-82-93165-12-5

No. 14 (2015)
PhD in Aquaculture
Cecilia Campos Vargas
Production of triploid Atlantic cod: A comparative study of muscle growth dynamics and gut morphology
ISBN: 978-82-93165-13-2

No. 15 (2015)
PhD in Aquatic Biosciences
Irina Smolina
Calanus in the North Atlantic: species identification, stress response, and population genetic structure
ISBN: 978-82-93165-14-9

No. 16 (2016)
PhD in Aquatic Biosciences
Lokesh Jeppinamogeru
Microbiota of Atlantic salmon (*Salmo salar L*.), during their early and adult life
ISBN: 978-82-93165-15-6

No. 17 (2017)
PhD in Aquatic Biosciences
Christopher Edward Presslauer
Comparative and functional analysis of microRNAs during zebrafish gonadal development
ISBN: 978-82-93165-16-3

No. 18 (2017)
PhD in Aquatic Biosciences
Marc Jürgen Silberberger
Spatial scales of benthic ecosystems in the sub-Arctic Lofoten-Vesterålen region
ISBN: 978-82-93165-17-0

No. 19 (2017)
PhD in Aquatic Biosciences
Marvin Choquet
Combining ecological and molecular approaches to redefine the baseline knowledge of the genus Calanus in the North Atlantic and the Arctic Oceans
ISBN: 978-82-93165-18-7

No. 20 (2017)
PhD in Aquatic Biosciences
Torvald B. Egeland
Reproduction in Arctic charr – timing and the need for speed
ISBN: 978-82-93165-19-4

No. 21 (2017)
PhD in Aquatic Biosciences
Marina Espinasse
Interannual variability in key zooplankton species in the North-East Atlantic: an analysis based on abundance and phenology
ISBN: 978-82-93165-20-0

No. 22 (2018)
PhD in Aquatic Biosciences
Kanchana Bandara
Diel and seasonal vertical migrations of high-latitude zooplankton: knowledge gaps and a high-resolution bridge
ISBN: 978-82-93165-21-7

No. 23 (2018)
PhD in Aquatic Biosciences
Deepti Manjari Patel
Characterization of skin immune and stress factors of lumpfish, *Cyclopterus lumpus*
ISBN: 978-82-93165-21-7

No. 24 (2018)
PhD in Aquatic Biosciences
Prabhugouda Siriyappagouder
The intestinal mycobiota of zebrafish – community profiling and exploration of the impact of yeast exposure early in life
ISBN: 978-82-93165-23-1

No. 25 (2018)
PhD in Aquatic Biosciences
Tor Erik Jørgensen
Molecular and evolutionary characterization of the Atlantic cod mitochondrial genome
ISBN: 978-82-93165-24-8

No. 26 (2018)
PhD in Aquatic Biosciences
Yangyang Gong
Microalgae as feed ingredients for Atlantic salmon
ISBN: 978-82-93165-25-5

No. 27 (2018)
PhD in Aquatic Biosciences
Ove Nicolaisen
Approaches to optimize marine larvae production
ISBN: 978-82-93165-26-2

No. 28 (2019)
PhD in Aquatic Biosciences
Qirui Zhang
The effect of embryonic incubation temperature on the immune response of larval and adult zebrafish (*Danio rerio*)
ISBN: 978-82-93165-27-9

No. 29 (2019)
PhD in Aquatic Biosciences
Andrea Bozman
The structuring effects of light on the deep-water scyphozoan *Periphylla periphylla*
ISBN: 978-82-93165-28-6

No. 30 (2019)
PhD in Aquatic Biosciences
Helene Rønquist Knutsen
Growth and development of juvenile spotted wolffish (*Anarhichas minor*) fed microalgae incorporated diets
ISBN: 978-82-93165-29-3

No. 31 (2019)
PhD in Aquatic Biosciences
Shruti Gupta
Feed additives elicit changes in the structure of the intestinal bacterial community of Atlantic salmon
ISBN: 978-82-93165-30-9

No. 32 (2019)
PhD in Aquatic Biosciences
Peter Simon Claus Schulze
Phototrophic microalgal cultivation in cold and light-limited environments
ISBN: 978-82-93165-31-6

No. 33 (2019)
PhD in Aquatic Biosciences
Maja Karoline Viddal Hatlebakk
New insights into Calanus glacialis and C. finmarchicus distribution, life histories and physiology in high-latitude seas
ISBN: 978-82-93165-32-3

Anglerfishes possess a number of extraordinary adaptations. Perhaps the strangest of these is male sexual parasitism, where the male clamps on to the female with his jaws and never lets go. In extreme cases the male's body will degrade and simply become a pair of testicles that are fed by the female's blood supply. In essence, the male's testicles have been transplanted to the female's body. In humans, and most vertebrates, it is not possible to simply transplant an organ from one individual to another as the immune system of the recipient will reject the organ and kill the transplanted cells. Why there is no rejection of sexually parasitic males is a bit of a mystery, but it does suggest that anglerfishes have a specialised immune system and that studying this may teach us how immune rejection can be avoided after transplantation. Though interesting, we currently, know very little about the anglerfishes and their immune system.In this thesis we provide the first genome assembly of a local anglerfish *Lophius piscatorius* (breiflabb). We discovered that *L. piscatorius* lacks an important part of the immune system that is known to be involved in rejection. Surprisingly, we also observed a previously unreported fundamental property of fish genomes that may indicate specific mechanisms of their genome evolution.The work described in this thesis provides a new genome resource which we have used to make discoveries that are both general to genome evolution and specific to anglerfish immunology.

NORD
University

ISBN: 978-82-93165-33-0
Trykk: Trykkeriet, Nord universitet
www.nord.no