Peer Community Journal

Section: Genomics

RESEARCH ARTICLE

Published
2023-07-25

Cite as
Christos V. Kitsoulis, Vasileios Papadogiannis, Jon B. Kristoffersen, Elisavet Kaitetzidou, Aspasia Sterioti, Costas S. Tsigenopoulos and Tereza Manousaki (2023) *Near-chromosome level genome assembly of devil firefish*, Pterois miles , Peer Community Journal, 3: e64.

Correspondence
tereza@hcmr.gr

Peer-review
Peer reviewed and recommended by PCI Genomics,
https://doi.org/10.24072/pci.genomics.100241

Near-chromosome level genome assembly of devil firefish, *Pterois miles*

Christos V. Kitsoulis[1,2], Vasileios Papadogiannis[2], Jon B. Kristoffersen[2], Elisavet Kaitetzidou[2], Aspasia Sterioti[2,3], Costas S. Tsigenopoulos[2], and Tereza Manousaki[2]

Volume **3** (2023), article e64

https://doi.org/10.24072/pcjournal.295

Abstract

Devil firefish (*Pterois miles*), a member of Scorpaenidae family, is one of the most successful marine non-native species, dominating around the world, that was rapidly spread into the Mediterranean Sea, through the Suez Canal, originating from the Indian Ocean. Even though lionfishes (Scorpaenidae) are identified among the most prosperous marine invaders, within this taxonomic group, the genomic resources are scant, while reference genome assemblies are totally absent. Here, we built and analyzed the first reference genome assembly of *P. miles* and explored its evolutionary background. The resulting genome assembly consisted of 660 contigs and scaffolds (N50 = 14,5 Mb) with a total size of about 902 Mb, while delivering 98% BUSCO completeness. We identified and described the large amount of transposable elements present in the genome and based on genomic data we constructed the first teleost phylogeny which includes a member of genus *Pterois*. The high-quality and contiguity *de novo* genome assembly built herein provides a valuable resource for future studies in species' biology and ecology, lionfish phylogeny, the influence of transposable elements on the evolution of vertebrate genomes and fish toxins evolution.

[1]Faculty of Medicine, University of Crete, Heraklion, Greece, [2]Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Greece, [3]Cretaquarium, Hellenic Centre for Marine Research, Heraklion, Greece

CENTRE MERSENNE

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

Introduction

The devil firefish, *Pterois miles* (Bennett, 1828), is a venomous species of the Scorpaenidae family native to the Western Indo-Pacific region, from South Africa to Red Sea and East to Sumatra (Schultz, 1986). The first occurrence of *P. miles*, as a single specimen, in the Mediterranean Sea was recorded off the Levantine coast in 1991 (Golani and Sonin, 1992), while the second, of two individuals, was almost twenty years later in the same area (Bariche et al., 2013). Soon after, the frequency of appearances, along the eastern Mediterranean Sea, rapidly increased (Crocetta et al., 2015; Kletou et al., 2016; Bilge et al., 2017; Mabruk and Rizgalla, 2019; Katsanevakis et al., 2020; Vavasis et al., 2020). Although the origin of species colonization in the Mediterranean Sea followed the invasion pattern of other Lessepsian migrants introduced from the Red Sea, through Suez Canal (Bariche et al., 2013; Dailianis et al., 2016; Kletou et al., 2016; Bariche et al., 2017; Chiesa et al., 2019; Dimitriou et al., 2019), the contribution of long-distance dispersal via aquarium trading remains a possibility (Bariche et al., 2017; Dimitriou et al., 2019). Lionfishes (genus: *Pterois*) are considered among the most thriving invaders in the history of marine invasions (Albins and Hixon, 2008) because of their rapid expansion worldwide (Azzurro et al., 2017). Indeed, the introduction in the western Atlantic of *P. miles* and of its con-generic *P. volitans*, together referred as the invasive lionfish complex (Lyons et al., 2019), is one of the fastest and most dominant marine fish introductions to date (Kletou et al., 2016, and references therein). For the Mediterranean Sea, the Suez Canal is the major pathway responsible for the spread of most of the non-indigenous species that constantly reshape its biodiversity and fishery resources (Kleitou et al., 2022). Invasive non-indigenous marine species, in general, are considered to have a major impact on local biodiversity while threatening marine industries and frequently human health (Bax et al., 2003; Arim et al., 2005; Blakeslee et al., 2019). Furthermore, they are commonly studied in evolutionary biology as models or “natural experiments” in order to explore invasion dynamics and adaptations to new niches (Barrett, 2015). Data derived from Whole Genome Sequencing (WGS) could provide promising opportunities in the exploration of potential adaptations that shape the fitness of invaders, as well as the dynamics of colonization. Further, it will provide a basis for understanding the hybrid origin of the invasive lionfishes (*P. miles* and *P. volitans*) in the Western Atlantic (Wilcox et al., 2018).

Biological characteristics of *P. miles* such as rapid somatic growth, signature anti-predatory defenses (Côté and Smith, 2018), reproductive success, discernible predatory behavior, low parasitism and ecological flexibility might explain its rapid distribution in the Mediterranean Sea (Dailianis et al., 2016), whereas the species population is rising rapidly along the coastlines (Kletou et al., 2016). Yet, few genomic resources of the species are available, including the mitochondrial genome (Dray et al., 2016) and DNA barcoding data (Chiesa et al., 2019, and references therein).

Devil firefish belongs to Scorpaenidae, a large family of venomous marine species including lionfishes, scorpionfishes and stonefishes (Diaz, 2015). Their venom (toxins) is mainly secreted from spines that are present in dorsal, lateral, pelvic and anal fins. These toxins are composed by two subunits α and β to form their active dimeric structure (Kiriake and Shiomi, 2011; Kiriake et al., 2013; Chuang and Shiao, 2014; Campos et al., 2021). The excreted venom is used for defensive purposes alongside other strategies (Campos et al., 2021, and references therein) that lead to successful anti-predatory adaptations. These scorpionfish toxins have multiple biological activities and their range differs between different species, despite their high similarity and conservation in specific domains (Chuang and Shiao, 2014; Campos et al., 2021). So far, toxins from stonefishes (stonustoxin, verrucotoxin and neoverrucotoxin) have been mainly identified and characterized (Ghadessy et al., 1996; Ueda et al., 2006; Kiriake and Shiomi, 2011; Kiriake et al., 2013), while toxins from other genera (*Scorpaena*, *Scorpaenopsis*, *Inimus* and *Pterois*) were recognized by similarity and cloning using the previous ones (Kiriake and Shiomi, 2011; Kiriake et al., 2013; Chuang and Shiao, 2014; Xie et al., 2019; Campos et al., 2021). However, the scorpionfish toxins are relatively understudied, even though there is a high diversity between them (Xie et al., 2019). Due to the absence of genomic data inside this family, the origin and evolution of toxins within scorpaenid species, and specifically in genus *Pterois*, still remains ambiguous.

The aim of this study was to construct, annotate and analyze the first high-quality genome assembly of *P. miles*. Through a combination of Oxford Nanopore Technologies (ONT), Pacific Biosciences (PacBio) and Illumina reads, we explored the genomic background of such a successful and unique invader, the devil firefish. Being the first top-quality genome sequenced within the Scorpaenidae family, this valuable resource could provide a critical conveyance to unveil and highlight the insights of species' biology, ecology and phylogeny in further investigations of invasive traits across the Mediterranean Sea.

Materials and methods

Sample collection, DNA and RNA library construction and sequencing

Animal care and handling were carried out following well established guidelines (Degrazia and Beauchamp, 2019).

A specimen was caught alive in Bali (North coast), Crete, Greece and was transferred in the indoor facilities of Cretaquarium (part of the Hellenic Centre for Marine Research, HCMR). The specimen was kept alive for three weeks in a 150 L tank with semi-closed circuit with airlift, mechanical and biological filtration, water recirculation rates of 30 and 100%/h. Temperature was adjusted at 19 oC +/- 1,0 oC and the photoperiod was set to 12hL:12hD. The specimen was anesthetized using clove oil, and blood was collected with a sterilized and heparinized syringe from the caudal vein of the fish. Blood was kept in sterilized and heparinized 1.5 ml tubes on ice, until DNA extraction.

DNA was extracted, for the purpose of ONT sequencing, from 4ul of blood with the Qiagen Genomic Tip 100/G kit. DNA integrity was assessed by electrophoresis in 0.4 % w/v megabase agarose gel. Three ligation libraries (SQK-LSK109) were constructed using 2.3 microgram DNA as input for each library. The libraries were sequenced on two R9.4.1 flow cells, on the in-house MinION Mk1B and MinION Mk1C sequencers, respectively. The resulting fast5-files were basecalled with Guppy v5.0.11, using the "sup" (super-accuracy) configuration and a minimum quality score of 7.

For the Illumina sequencing, the same procedure and protocol were used for DNA extraction. The template DNA for Illumina sequencing was sheared by ultrasonication in a Covaris instrument. A PCR-free library was prepared with the Kapa Hyper Prep DNA kit with TruSeq Unique Dual Indexing.

For transcriptomic data, total RNA was extracted from seven tissues (brain, gonad, gills, heart, liver, muscle and spleen). Tissue grinding in liquid nitrogen with pestle and mortar was followed by the sample homogenization and lysis in TRIzol buffer (Invitrogen) according to the manufacturer's guidelines. The quantity and the quality of the extracted RNA was estimated with NanoDrop ND-1000 spectrophotometer and with agarose gel (1.5%) electrophoresis, as well as with Agilent 2100 Bioanalyzer using RNA 6000 pico kit. The library preparation was conducted using Pacific Biosciences protocol for Iso-Seq™ Express Template Preparation for Sequel and Sequel II Systems. The seven libraries (one for each tissue) were sequenced on a PacBio sequel II instrument, on one SMRT cell.

Illumina and Pacific Biosciences library preparation and sequencing were carried out by the Norwegian Sequencing Centre (www.sequencing.uio.no), hosted in the University of Oslo.

Genomic data pre-processing

The length filtering and adapter trimming of basecalled ONT reads were carried out with Porechop v0.2.4 (https://github.com/rrwick/Porechop) using default parameters, adding the extra parameter "–discard_middle" to prune reads with potential inner adapters. The quality control was performed using Nanoplot v.1.20 (De Coster et al., 2018). The quality assessment of Illumina reads was performed with FastQC v0.11.9 (Andrews, 2010) while both filtering of low quality reads and adapter trimming, using Trimmomatic v0.39 (Bolger et al., 2014). The reads were processed by Trimmomatic on a 4-base sliding window with an average cutting-off threshold score lower than 15 Phred score. Leading and trailing bases with a quality score less than 10 Phred were trimmed out, while reads shorter than 75 bp and average score lower than 30 Phred were removed.

De novo genome assembly

For the *de novo* genome assembly, using a hybrid approach, the long reads from ONT were combined with short and highly accurate Illumina reads. At first, the ONT reads were used for the construction of the initial draft *de novo* assembly and the first rounds of polishing, while the Illumina reads were used for the later rounds of polishing. The draft assembly was built from ONT reads using the *de novo* assembler Flye v2.9. (Kolmogorov et al., 2019), which uses a repeat graph as core data structure, with default parameters and a genome size estimation of 900Mb. The genome size estimation was based on the corresponding entry of *Pterois volitans* on www.genomesize.com. Then, the draft assembly was polished in two rounds with RACON v1.4.3 (Vaser et al., 2017) using the filtered long reads, mapped against it by Minimap v2.22 (Li, 2018) which resulted in the intermediate assembly. Further polishing was performed on the intermediate genome assembly by Medaka v1.4.4 (https://github.com/nanoporetech/medaka) and then with Pilon v1.23 (Walker et al., 2014) after mapping with Minimap v2.22 (Li, 2018) the preprocessed Illumina reads against the assembly taken from Medaka, resulting in the final reference genome assembly used for all downstream analyses.

The draft, intermediate and final assemblies were evaluated by two commonly used criteria: (i) the N50 statistic from contig sizes, using QUAST v.5.0.2 (Gurevich et al., 2013), and (ii) the completeness score based on the presence of universal single copy ortholog genes, using BUSCO v.5.3 (Manni et al., 2021) against Actinopterygii ortholog dataset 10 (actinopterygii_odb10). BUSCO was run with default parameters adding the extra parameter "—augustus" to enable species-specific training for gene prediction by AUGUSTUS v.3.4 (Stanke et al., 2008). Alternative values (e.g. L90) were calculated (Table 1) with a custom python tool, ELDAR (https://github.com/ckitsoulis/ELDAR; Kitsoulis, 2023b).

The whole genome assembly pipeline which was used in the present study was previously designed by Danis et al. (2021), containerized by Angelova et al. (2022) (https://github.com/genomenerds/SnakeCube) and ran in the IMBBC High performance computing (HPC) facility "Zorbas" (Zafeiropoulos et al., 2021).

Genome annotation

Transposable elements annotation

A *de novo* transposable elements (TEs) library was generated from the constructed genome assembly of *P. miles*, using the Extensive *de novo* TE Annotator (EDTA) package (Ou et al., 2019), an automated whole-genome TE annotation pipeline, with default parameters. In our case, the RepeatModeler2 (Flynn et al., 2020) was utilized to additionally support the identification of TE families inside the EDTA algorithm, using the extra parameter "—sensitive 1". The non-redundant TE library was then separated into three sub-libraries based on its TEs classification, so far, using a custom python script "library_split.py": i) Classified TE sequences, ii) Unclassified TE sequences in the level of superfamily (partially classified), and iii) Unclassified - Unknown TE sequences. The two latter sub-libraries were classified again using DeepTE (Yan et al., 2020), a transposon classification tool which depends on convolutional neural networks (CNN). The annotation probability threshold was strictly set to 0.8 ("-prop_thr 0.8"). A step of header correction and reformation, using bash commands, in every sub-library occurred before their concatenation to the final TE annotated library, in order to achieve a compatible format for the next steps. Finally, RepeatMasker v4.1.2 (Tarailo-Graovac and Chen, 2009) performed the initial TE annotation and genome soft-masking, utilizing the NCBI/RMblast search engine, based on the previously-described library. To achieve a more accurate and detailed annotation/categorisation of TE (Table 2), based on an up-to-date TE classification system (Makalowski et al., 2019), a python-based parser "RM_parser.py" was developed and used for the output files of RepeatMasker. The designed workflow is presented schematically in Supplementary Figure 1.

Structural annotation - gene prediction

For transcriptome data, circular consensus sequencing (CCS) reads were generated using the CCS application on SMRT Link v10.2 and then Iso-Seq analysis was performed on them, with default parameters, using the Iso-Seq pipeline, IsoSeq v3.4 (https://github.com/PacificBiosciences/IsoSeq), until the production of high quality (HQ) consensus full-length transcripts. The IsoSeq pipeline included three basic steps: i) generation of CCS reads, ii) classification of full-length (FL) reads, and iii) clustering of full-length non-contatamer (FLNC) reads to obtain high-quality consensus transcripts. The number of resulting intermediate reads to final isoforms in the IsoSeq pipeline are presented in Table 3. The HQ transcripts were aligned (spliced-wise) against the soft-masked genome assembly of *P. miles*, using GMAP v2021.08.25 (Wu and Watanabe, 2005). SAM files were sorted and converted to BAM using samtools v1.15.1 (Danecek et al., 2021), and the redundant transcripts were finally collapsed to generate a non-redundant HQ full-length transcripts set, using cDNA Cupcake v28.0 (https://github.com/Magdoll/cDNA_Cupcake) with a minimum alignment coverage equal to 0.99 and a minimum alignment identity of 0.95.

Gene prediction was conducted based on a hybrid strategy of transcriptome-based (non-redundant high quality transcripts), homology-based (curated protein sets) and ab initio methods, using a semi-automated workflow consisting of 12 individual tools and intermediate custom python and bash scripts (Supplementary Figure 2). In the first step, HQ isoforms and curated proteomes of 20 actinopterygian species were aligned (spliced-wise) to the soft-masked genome assembly. The non-redundant HQ trascriptome set was previously aligned using GMAP v2021.08.25 (Wu and Watanabe, 2005). For protein homology evidence, a BLAST database was generated from protein sequences of 20 species (Table 4), being downloaded from UniProtKB/Swiss-Prot (https://www.uniprot.org/), using DIAMOND v2.0.14 (Buchfink et al., 2015). In the second step, Mikado v2.3.3 (Venturini et al., 2018) was used, a python-based pipeline which identifies the "best" set of transcripts from multiple sources, in order to return potential gene models from the transcriptome and protein homology evidence. Homology evidence for each of the predicted transcripts provided to Mikado were generated based on the BLAST DB, using DIAMOND v2.0.14 (Buchfink et al., 2015) while open reading frame (ORF) predictions of Mikado-selected transcripts were produced by Transdecoder v5.5 (https://github.com/TransDecoder/TransDecoder). All information and evidence were merged afterwards to generate the most accurate evidence-based gene models, using Mikado steps "serialise" and "pick ". These gene models had been used in later steps of gene prediction and annotation update. In the third step, Augustus v3.4 (Stanke et al., 2008) was trained with two optimization rounds on a subset of gene models (generated in step 2) that fulfilled specific criteria: i) full length, ii) non-redundant, iii) over a blast hit score of 0.5, and iv) with at least 2 exons. The training set was selected using a custom python script "select_training.py". To take advantage of Augustus ability to incorporate hints (gene, protein, intron etc) for generating high confident gene models, species-specific exon hints and spliced protein alignments were generated and merged, secondarily. For exon hints, the exons coordinates were extracted from the previously produced annotation file using python scripts. For the spliced protein alignments, three well annotated protein sets of species *Oryzias latipes* (downloaded from UniProtKB), *Gasterosteus aculeatus* (downloaded from Ensembl) and *Argyrosomous regius* (Papadogiannis et al., 2023) were aligned to the genome assembly, using Exonerate v2.4 (https://github.com/nathanweeks/exonerate). The annotation files were merged, sorted and then filtered for exonic evidence extraction using python. *Ab initio* prediction on the *P. miles* genome assembly, alongside the generated hints, was performed by Augustus v3.4 (Stanke et al., 2008) with extra parameters "—allow_hinted_splicesites=atac" and "—alternatives-from-evidence=false". In the fourth step, gene models, generated in the previous steps (from Mikado and Augustus), were merged into a consensus gene set, after two updating rounds, using PASA v2.4.1 (Haas et al., 2003), an eukaryotic genome annotation pipeline. For this reason, Mikado-predicted protein coding gene models were loaded into PASA to create the initial MySQL DB of transcripts, the Augustus predictions were loaded to the DB and it was updated later on with the Mikado-predicted genes. The same procedure was followed, as a second updating round, starting this time from the resulting annotation of the first round. In the last step, genes were filtered to remove predictions with in-frame STOP codons and those that overlapped with TEs. For the first case, the gene models were cleaned for potential identical isoforms using Agat (https://github.com/NBISweden/AGAT), the artifacts were recognised using gffread (https://github.com/gpertea/gffread), while they were removed with bash commands. For the second case, candidate models were found using bedtools v2.30 (Quinlan and Hall, 2010) "intersection" command, with a minimum overlapping score of 0.5 "-f 0.50 " and filtered out with bash commands as well. The completeness evaluation of transcripts and genes, in each step, was performed using BUSCO v5.3 (Manni et al., 2021) against the Actinopterygii ortholog dataset 10 (Table 5).

Functional annotation

The functional annotation of *P. miles* predicted gene set was performed using three different strategies and tools. The first approach was based on similarity search (reciprocal hits) against the annotated genes of zebrafish (*D. rerio*) using BLASTp v2.12+ (Altschul et al., 1990) with parameters: "-evalue 1e-6 ", "-max_target_seqs 1" and "-hdps 1", in order to reduce the number of queries for the later annotation update. In the second one, results were fetched with EggNOG-mapper v2.1.7 (Cantalapiedra et al., 2021) based on fast orthology assignments using pre-computed clusters and phylogenies from eggNOG v5.0 database (Huerta-Cepas et al., 2019). For the last approach, annotations were retrieved using PANNZER2 (Törönen and Holm, 2022), a weighted k-nearest neighbour classifier which is based on SANSparallel

(Koskinen and Holm, 2012) for homology similarity against UniProt and enrichment statistics, using various user-defined scoring functions. Prediction of gene names, Gene Ontology (GO) annotations, KEGG pathway IDs, Pfam domains and descriptions from all aforementioned strategies were filtered and assigned to gene models using a custom python script "FUNfilter.py". Gene names, in each case, were selected based on the most frequent occurrence, while KEGG pathway IDs and Pfam domains were obtained directly from EggNOG-mapper. An additional step was performed for GO terms (biological process) being mapped to gene models, by using the assigned gene names as queries and retrieving terms from UniProtKB (https://www.uniprot.org/) with "Retrieve/ID mapping tool", for human, mouse and zebrafish. Finally, GO terms resulted as a set of terms between those predicted by EggNOG-mapper v2.1.7 and UniProtKB.

Phylogenomic analyses

Orthology assignment

To identify orthologous and paralogous genes, 46 whole-genome protein coding gene sets (longest isoforms) from teleost species (Supplementary Table 1), along with the one of *P. miles*, were compared using OrthoFinder v2.5.2 (Emms and Kelly, 2019), with default parameters. The initial dataset of genomes was collected from Genomes-NCBI Datasets (https://www.ncbi.nlm.nih.gov/datasets/genomes/) and Ensembl (https://www.ensembl.org/) based on the following criteria: i) genomes at the chromosome and/or scaffold level of contiguity and ii) N50 > 10Mb. The longest isoform per gene was selected initially (in GFF format) using Agat (https://github.com/NBISweden/AGAT) and then extracted as FASTA file with gffread (https://github.com/gpertea/gffread). Each proteome set was assessed for completeness using BUSCO v.5.3 (Manni et al., 2021) against Actinopterygii ortholog dataset 10. For the final set, only proteomes which exceeded a predefined completeness threshold (90%) were included and only one species per genus was kept for the final analysis.

Species tree inference

The phylogenetic hierarchical orthogroups (HOGs) produced by OrthoFinder were filtered, at first, to select those with complete representation from *P. miles* and then, only those containing a single gene copy per species, to exclude potential paralogs. Afterwards, orthogroups which had a representation from at least 43 out of 47 species (> 91.4%) were selected, using a custom python script (aragorn_orthoX.py). The protein sequences of each HOG were aligned using MAFFT v7.505 (Katoh and Standley, 2013). In cases of HOGs with no representation from certain species, the corresponding sequences were filled with gaps equal to the total length of the HOG alignment using a custom python script (gimli_clean.py). Then, all aligned HOG sequences were concatenated into a superalignment matrix with bash commands. The initial matrix was trimmed to remove spurious sequences and poorly aligned regions using trimAl v1.4.1 (Capella-Gutiérrez et al., 2009), with default parameters and strict mode. ModelTest-NG v0.1.7 (Darriba et al., 2019) was used for the selection of the best-fit model and IQTREE v2.2.0.3 (Minh et al., 2020) for the maximum-likelihood phylogenetic tree inference. To assess the confidence of branches, IQ-TREE was run for 1000 bootstrap replicates (ultrafast bootstrap mode). The phylogenetic tree was finally visualized using R/RStudio Team (2022) and package "ggtree" (Yu, 2020) within a custom R script (tree.R), selecting *Lepisosteus oculatus* and *Polyodon spathula*, as an outgroup clade.

Comparative genomic analysis

Synteny analysis

Synteny analysis was performed at the gene level between *P. miles* and three-spined stickleback, *G. aculeatus*. For this purpose, one-to-one orthologues were selected from the HOGs produced earlier by OrthoFinder, to compare the physical localization of genetic loci within species. The 42 longest contigs of *P. miles* genome assembly (representing ~73.5% of the genome size) were selected for visualization against the 21 chromosomes of *G. aculeatus*, using Circos (Krzywinski et al., 2009).

Gene families expansion and contraction

Changes in gene family size (expansions and contractions) were estimated using CAFE v5 (Mendes et al., 2020). The HOGs summary table of all species from OrthoFinder was retrieved and modified earlier by a custom python script "aragorn_orthoX.py", resulting in a count matrix of genes per species and family, in order to be used by CAFE. Following CAFE's developers instructions, HOGs absent in more than 10 species out of 47 and the ones in which the difference between the maximum and minimum number of genes was greater than 70 ($max_i(n_{genes})$ - $min_i(n_{genes}) > 70$), were filtered out from the analysis. An ultrametric binary tree was produced with R package "ape" (Paradis and Schliep, 2019) in a custom R script "tree_calibration.R". For that, we used the phylogenetic tree, produced earlier, and the divergence times taken from TIME-TREE (http://www.timetree.org/) between 4 different species' combinations, *Polyodon spathula - Danio rerio*, *Danio rerio - Takifugu rubripes*, *Oryzias latipes - Mola mola* and *Dicentrarchus labrax - Mola mola*. CAFE was finally run using 3 different gamma function categories (-k) to estimate λ parameter (corresponding to the rate of change of families), 400 iterations (-I) and a p-value equal to 0.05 (-pvalue). After the analysis, we selected for visualization only the Perciformes clade, as a subset of the phylogenetic tree.

Duplication event estimation

To infer gene duplication events in *P. miles* from gene family trees and the estimated phylogenetic tree, we used GeneRax v2.0.4 (Morel et al., 2020). Initially, protein sequences from each HOG, produced by OrthoFinder, were aligned to each other using MAFFT v7.505 (Katoh and Standley, 2013) and trimmed with trimAl v1.4.1 (Capella-Gutierrez et al., 2009), in strict mode. Orthogroups with less than three sequences were excluded from the following procedure. The best-fit model was estimated from each HOG and later was used for the inference of a single maximum-likelihood gene tree, using IQTREE v2.2.0.3 (Minh et al., 2020). After manually inspecting and correcting the estimated substitution models for some cases, gene trees and their models along with the phylogenetic species tree were used to estimate duplication events with GeneRax.

Gene ontology terms descriptive analysis

For GO terms descriptive analysis, firstly we downloaded the core ontology (OBO format) from Gene Ontology DB (http://geneontology.org/docs/download-ontology/), and then the predicted gene set of *P. miles* and their assigned GO terms were mapped with GO biological process descriptions provided by the core ontology, using a custom python script "obo_mapper.py". Then, these GO terms and their descriptions were grouped/mapped into the gene families of their genes (HOGs from OrthoFinder), which were previously identified as rapidly expanding from CAFE and involved in duplication events by GeneRax.

Toxin gene evolution in lionfishes

To identify genes responsible for the secreted toxin of devil firefish, toxins (proteins) identified in other scorpaenid species (Ghadessy et al., 1996; Ueda et al., 2006; Kiriake and Shiomi, 2011; Kiriake et al., 2013; Chuang and Shiao, 2014; Xie et al., 2019) were downloaded from NCBI (Table 6) and aligned to the genome of *P. miles*, using tBLASTx v.2.12 (Altschul et al., 1990). All proteins included were about 700 amino acids long and constituted of three exons. The identified coding regions on the genome of *P. miles* fulfilling specific criteria (a. blast hit against all proteins (toxins), b. non-overlapping to each other, c. with three potential exons) were translated into proteins using similarity results from BLAST and ExPASy Translate tool (Gasteiger, 2003), after the recognition of correct ORFs. The protein sequences were, then, aligned against with MAFFT v7.505 (Katoh and Standley, 2013) and trimmed using trimAl v1.4.1 (Capella-Gutiérrez et al., 2009), in strict mode. Finally, the alignment was manually inspected using Jalview (Waterhouse et al., 2009). ModelTest-NG v.0.1.7 (Darriba et al., 2019) was used for the estimation of the substitution model and IQTREE v2.2.0.3 (Minh et al., 2020) to infer the maximum-likelihood phylogenetic tree. The final unrooted tree was visualized with FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Results

Genomic sequencing results

Sequencing yielded a total of 38.66 Gb of raw genomic ONT reads, from which 36.16 Gb had a Phred quality score above Q7, as well as 3.75 Gb of raw Illumina reads. After the pre-processing steps of both trimming and quality filtering, 35.72 Gb of ONT, for the initial assembly, and 3.16 Gb of Illumina reads, for later polishing, were maintained for the downstream process of *de novo* construction of the genome assembly (Table 7).

Genome size and assembly completeness

The final genome assembly contained 660 contigs with a total length of about 902.3 Mb. The longest contig was sized at 36.5 Mb and the N50 statistic value at 14.5 Mb (Table 1). At least, 90% of the genome size was represented in the 83 longest contigs of the produced assembly (Supplementary Figure 3). The GC content of the genome was calculated at 40.78% (GC-rich regions at the 42 longest contigs are presented in Supplementary Figure 4). For genome completeness assessment, 3,566 out of 3,640 BUSCO genes were present (98%), against the Actinopterigyian ortholog dataset (v.10). Of those, 3,551 genes (97.6%) were complete, while only 74 (2.0%) were missing (Table 1), suggesting a high level of contiguity and completeness of the *de novo* genome assembly.

Table 1 - Polished genome assembly statistics and completeness.

Number of contigs	660
Total length	902,353,306 bp
GC (%)	40.78
Longest contig	36,477,432 bp
N50	14,490,642 bp
L50	21
L75	47
L90	83
BUSCO completeness score	
Complete	3,551 (97.6%)
Single	3,515 (96.6%)
Duplicated	36 (1.0%)
Fragmented	15 (0.4%)
Missing	74 (2.0%)
Total number of orthologs (Actinopterygii)	3,640

Genome annotation

Transposable elements annotation

About 46.5% of the genome assembly (~ 416.7 Mb) in *P. miles*, consisted of transposable elements (Figure 1). Class I Retroelements make up 4.65% of the genome assembly and LTR order is the most dominant with a representation of at least 4.23%, with its superfamily Gypsy of 1.47%. Class II of TEs (DNA transposons) represents a high amount (28.6%) of the whole genome, while elements of the TIR order and its superfamily CACTA were mostly found, with 17.46% and 4.83% respectively, among the high confidence and completely classified DNA TEs. Additionally, 9.8% of the masked genome are regions of complex composition of overlapping TEs, not clearly defined as discrete elements during masking (Table 2). The distribution of TE content in the 42 longest contigs is presented in Supplementary Figure 4.

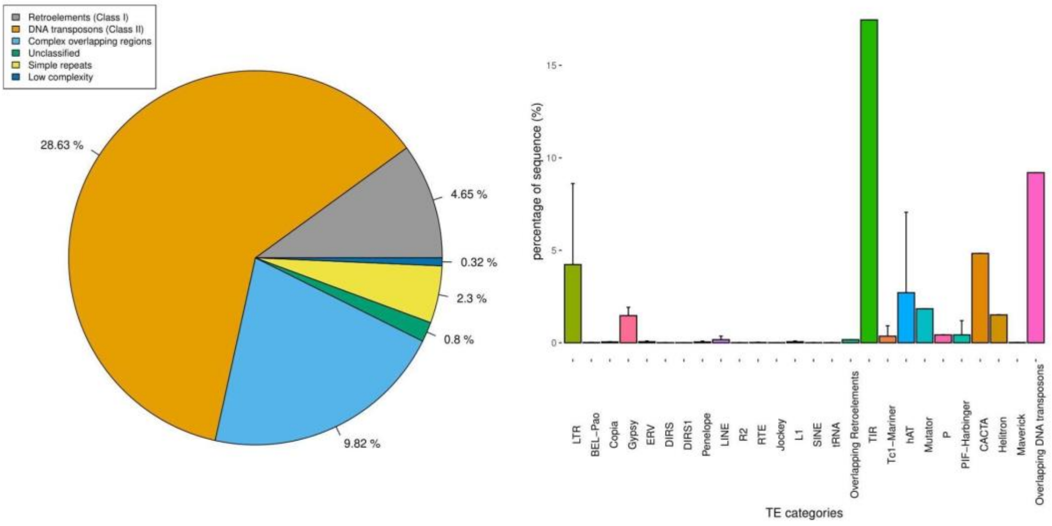

Figure 1 - Percentage of transposable element categories representation in the genome of *P. miles*.

Transcriptome analysis

From 6,245,243 initial CCS reads, Iso-seq analysis yielded a total of 124,307 HQ consensus isoforms. The samples with the higher number of transcripts were the heart, spleen and liver (Table 3), and they shared almost 6,500 of them (Supplementary Figure 5). However, the total amount of HQ transcripts shared between all sampled tissues was notably low (~1,000). From the total number of HQ transcripts, 91,666 were aligned properly to the assembled genome (representing 73.74%) and used as evidence for the gene prediction.

Structural and Functional annotation

The hybrid approach of HQ full-length transcripts-based, homology-based and *ab initio*-based methods resulted in a total of 25,410 candidate protein-coding gene models. We filtered out genes with in-frame stop codons (266 putative genes) and those overlapping with TEs (505 gene models). We ended up with 24,639 potential gene models (Supplementary_file1_annotation.gff), representing in size 382.3 Mb of the genome assembly (Table 5). A subset of 22,473 genes were assigned with gene names and 23,521 matched at least one functional description, accounting 89.7% and 95.4% of the total number of genes, respectively. In terms of GO, KEGG pathway IDs and PFAM domains, 22,115 (88.3%), 15,071 (61.1%) and 20,003 (81.1%) genes were annotated, in each case (Supplementary_file2_functional.tsv).

From a core set of 3,640 single-copy orthologous genes (Actinopterigyii lineage, odb 10), 3,414 (93.8%) were found to be present in the predicted gene set (Table 8), with 3,233 (88.8%) identified as complete (3,082 as single-copy and 151 as duplicated) and 181 (5.0%) as fragmented, while 224 (6.2%) of them were not present, using BUSCO v5.3 (Manni et al., 2021).

Orthology assignment and phylogenomic analysis

The total number of proteins included in the proteomes of all 47 teleost fish species (Supplementary Table 1) and analyzed by OrthoFinder, was 1,108,753 and 97.8% of them were assigned to 28,397 phylogenetic HOGS. After the filtering step, 1,193 HOGs were selected to construct the superalignment matrix. Before trimming, the matrix consisted of 1,018,881 alignment positions, while after filtering it contained 473,254 (46.4%) positions which were used for the phylogenomic analysis. JTT + I + G4 + F was identified as the best-fit model and used for the phylogenetic tree inference (Figure 2). At the resulting maximum-likelihood phylogenetic tree, almost all branches were supported with 100 non-parametric bootstraps. Based on the constructed phylogeny, *P. miles* is placed within the Perciformes clade.

Synteny analysis

Synteny analysis on gene level unveiled high conserved syntenic coding regions between the 42 longest contigs of *P. miles* and the 21 chromosomes of *G. aculeatus*, sharing at total 8,035 one-to-one orthologous genes (Figure 3).

Gene repertoire evolution

Gene family evolution analysis in *P. miles* estimated 228 rapidly evolving gene families (out of 15,405 included families), at a significance level of 0.01 (p-value). From these rapidly evolving families, 136 were identified as expanding and 92 as contracting. The total number of genes included in these rapidly expanding families was 373. The corresponding state of the number of estimated gene families' gains-losses inside the Perciformes species is presented in Figure 4, as a subset of Figure 2.

The number of families included in the duplication events estimation analysis was 23,775 with an average of 43 genes per family. The largest family included 1,638 genes. The total number of genes included in the gene duplication events estimation was 1,036,460. In *P. miles*, 728 gene families were identified with duplication events, including an amount of 2,263 individual genes.

The descriptive analyses for rapidly expanding gene families of CAFE and those involved in duplication events in GeneRax are presented in Figure 5. GO terms were classified into eight categories, associated with metabolism, immune system, development, growth, antimicrobial response, toxin transport, reproduction and locomotion, and the numbers of gene families, involved genes and unique terms were calculated for both results from CAFE and GeneRax. The top terms in both analyses were included in gene families associated with "metabolism", "development", "immune" and "growth".

Table 2 - Transposable element annotation statistics.

Transposable elements	Number of elements	Length occupied (bp)	% of genome size
Retroelements (Class I)	173,376	41,925,013	4.65 (8.99)
LTR	165,075	38,160,626	4.23 (8.61)
BEL-Pao	635	141,992	0.02 (0)
Copia	2,380	432,776	0.05 (0.06)
Gypsy	31,595	13,241,124	1.47 (1.92*)
ERV	2,667	529,295	0.06 (0.1)
DIRS	239	123,175	0.01
DIRS1	239	123,175	0.01
Ngaro	0	0	0
Penelope	1,268	407,091	0.05 (0.09)
LINE	3,602	1,523,385	0.17 (0.36)
R2	100	88,857	0.01 (0.02)
RTE	438	169,256	0.02 (0.03)
Jockey	77	60,944	0.01 (0.01)
L1	982	524,323	0.06 (0.1)
SINE	817	132,603	0.01 (0.02)
tRNA	817	132,603	0.01 (0.02)
7L	0	0	0
5S	0	0	0
Overlapping Retroelements	2,375	1,578,133	0.17
DNA transposons (Class II)	1,036,247	258,303,777	28.63 (33.8)
TIR	759,014	157,567,545	17.46
Tc1-Mariner	20,095	3,271,776	0.36 (0.92)
hAT	116,270	24,446,997	2.71 (7.06)
Mutator	109,966	16,578,613	1.84
Merlin	0	0	0
Transib	0	0	0
P	23,489	3,918,909	0.43 (0.01)
PiggyBac	176	17,979	0 (0)
PIF-Harbinger	23,232	3,871,691	0.43 (1.2)
CACTA	311,697	43,611,897	4.83
Helitron	82,240	13,621,796	1.51
Maverick	146	206,087	0.02
Overlapping DNA transposons	177,832	83,009,131	9.2
Unclassified	32,965	7,246,688	0.8 (1.1)
Simple repeats	359,730	20,717,298	2.3 (2.3)
Low complexity	37,555	2,860,435	0.32 (0.32)
Complex overlapping regions	149,699	88,617,633	9.82

*grouped as Gypsy/DIRS1 by RepeatMasker
in parenthesis are presented the percentages calculated by RepeatMasker

Table 3 - Number of reads from CCS to high-quality isoforms

Tissue	CCS	HiFi	FLNC (polyA)	HQ isoforms
muscle	682,522	651,026	679,844	34,862
liver	863,535	814,531	861,396	50,651
heart	747,994	704,226	743,228	73,273
brain 1	16,449	15,739	15,986	2,467
brain 2	33,032	32,232	32,810	7,042
gonad	396,510	377,750	395,080	17,804
spleen	596,858	568,143	594,852	66,103
gills	193,697	184,914	192,470	25,831
consensus isoforms				124,307

Lionfish toxins evolution

The alignment of scorpaenid toxins protein set (blast reciprocal hits) against the genome of *P. miles* revealed a total number of six complete toxin genes, with three exons and two introns each (intron size in bp for 1: $\bar{x}_1$=819, σ_{x1}=489 and 2: $\bar{x}_2$=598 and σ_{x2}=367 respectively), on the 7th longest contig and in a distance between of 50.3 kb (Supplementary_file3_pmiles_toxins.tsv). The phylogeny of scorpaenid toxins showed the separation (with high support) between the two subunits, α and β (Figure 6), which form the functional heterodimer. Also, it confirmed that three genes in devil firefish's genome correspond to α subunit and three to β, respectively.

Table 4 - Species included in protein homology BLAST database

Scientific name	Common name	UniProt ID	Number of proteins	Reference
Amphilophus citrinellus	Midas cichlid	UP000261340	31,742	Lv et al. (2022)
Amphiprion ocellaris	Clown anemonefish	UP000257160	31,745	Ryu et al. (2022)
Astyanax mexicanus	Blind cave fish	UP000018467	39,383	McGaugh et al. (2014)
Betta splendens	Siamese fighting fish	UP000515150	41,617	Fan et al. (2018)
Carassius auratus	Goldfish	UP000515129	82,968	Chen et al. (2019)
Clupea harengus	Atlantic herring	UP000515152	37,255	Kongsstovu et al. (2019)
Danio rerio	Zebrafish	UP000000437	46,841	Howe et al. (2013)
Esox lucius	Northern pike	UP000265140	71,519	Rondeau et al. (2014)
Gymnodraco acuticeps	Ploughfish	UP000515161	39,915	Bista et al. (2022)
Haplochromis burtoni	Burton’s mouthbrooder	UP000264840	34,332	Brawand et al. (2014)
Hippocampus comes	Tiger tail seahorse	UP000264820	27,735	Lin et al. (2016)
Ictalurus punctatus	Channel catfish	UP000221080	40,203	Wang et al. (2022)
Lepisosteus oculatus	Spotted gar	UP000018468	22,463	Braasch et al. (2016)
Oreochromis niloticus	Nile tilapia	UP000005207	74,622	Conte et al. (2017)
Oryzias latipes	Japanese rice fish	UP000001038	36,128	Kasahara et al. (2007)
Perca flavescens	American yellow perch	UP000295070	21,644	Feron et al. (2020)
Salmo salar	Atlantic salmon	UP000087266	82,390	Lien et al. (2016)
Sparus aurata	Gilthead sea bream	UP000472265	69,200	Perez-Sanchez et al. (2019)
Takifugu rubripes	Japanese pufferfish	UP000005226	51,078	Kai et al. (2011)
Xiphophorus maculatus	Southern platyfish	UP000002852	35,279	Schartl et al. (2013)

Table 5 - Basic statistics of predicted gene models

Type	Number	Mean size (bp)	Longest (bp)	Shortest (bp)	Genome size (%)
gene	24,645	15,518	445,933	201	42.37
transcript	25,040	15,440	445,933	201	42.84
5’ UTR	13,119	228	10,216	1	0.24
exon	231,331	207	14,732	3	5.33
CDS	227,256	159	6,853	16	4
intron	206,291	1,609	188,322	21	36.05
3’ UTR	11,114	914	14,608	5	1.07

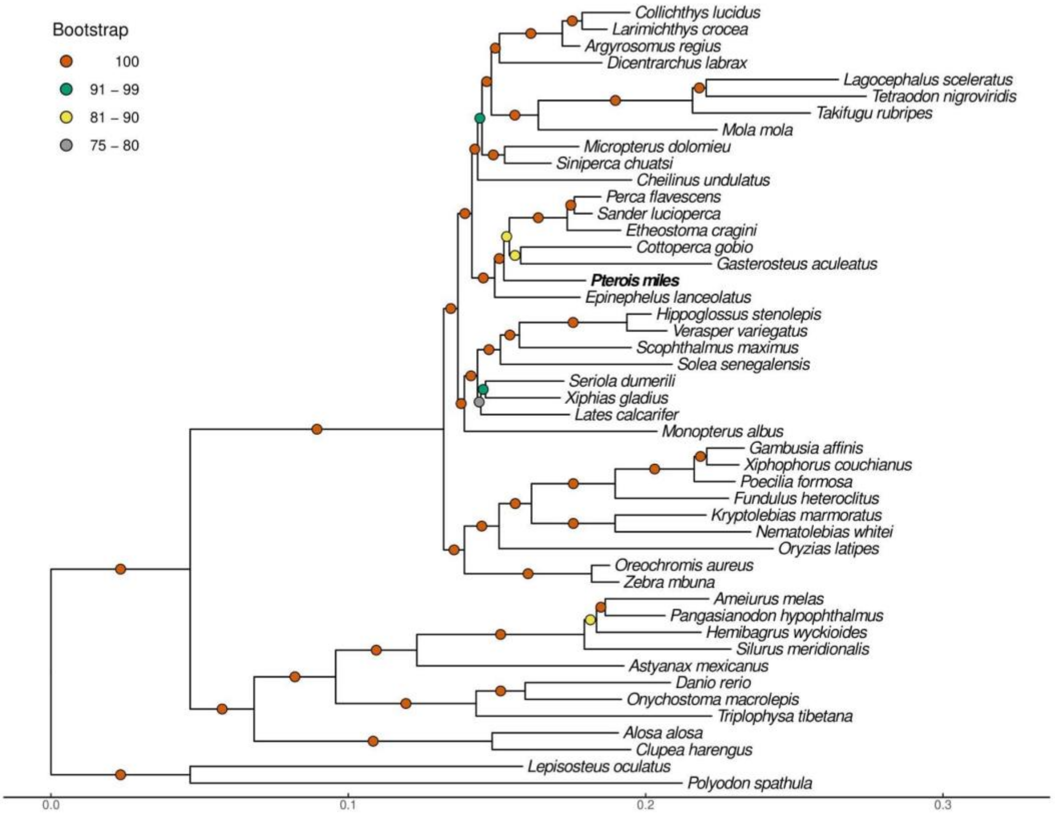

Figure 2 - Maximum-likelihood phylogenetic tree using JTT + I + G4 + F substitution model and *P. spathula* - *L. oculatus* clade as an outgroup.

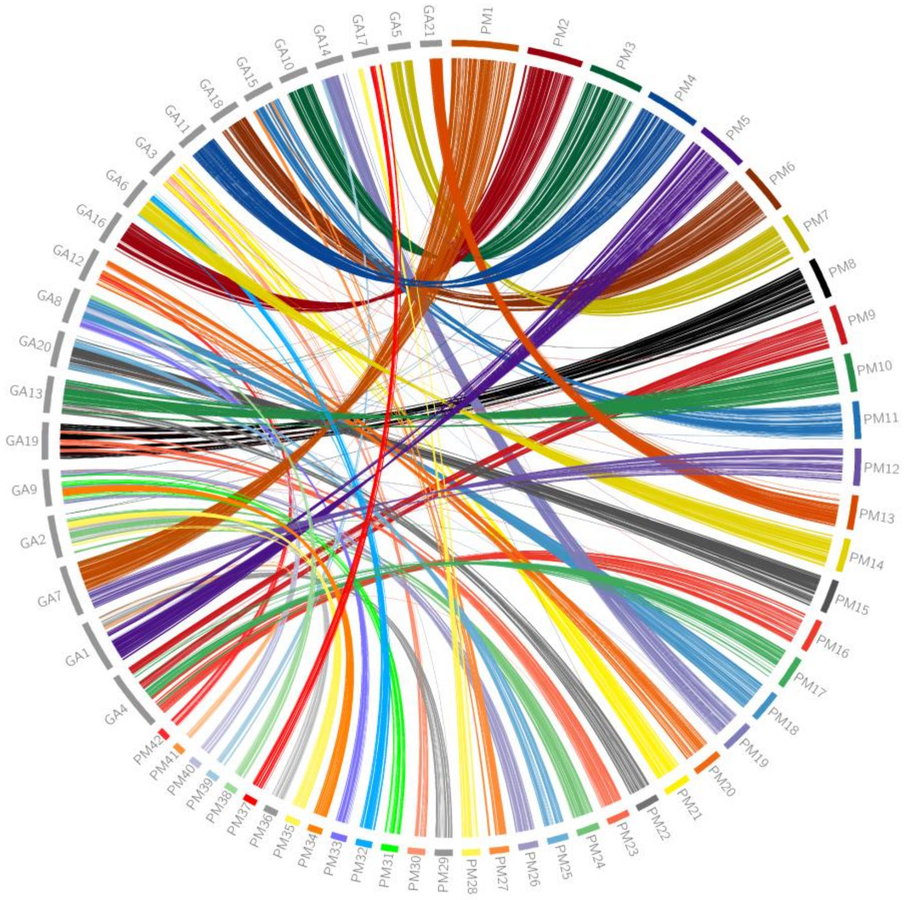

Figure 3 - Circos plot which presents the syntenic locations of orthologous genes between the 42 longest contigs of *P. miles* (right, PM) and the 21 chromosomes of *G. aculeatus* (left, GA). Strips link orthologous genes between the two species, and colors represent the different contigs of *P. miles*.

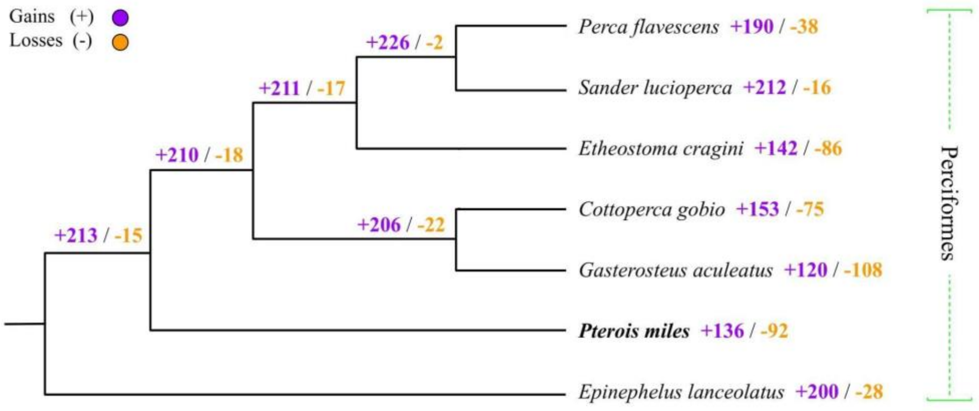
Gains (+)
Losses (-)
+226 / -2
+211 / -17
+210 / -18
+213 / -15
+206 / -22
Perca flavescens +190 / -38
Sander lucioperca +212 / -16
Etheostoma cragini +142 / -86
Cottoperca gobio +153 / -75
Gasterosteus aculeatus +120 / -108
Pterois miles +136 / -92
Epinephelus lanceolatus +200 / -28
Perciformes

Figure 4 - Gene family evolution analysis, including the number of gained (purple) and lost gene families (orange) for the Perciformes clade.

Discussion

Here, we presented and analyzed the first high-quality genome assembly for the lessepsian migrant species *P. miles*, which is, also, the first assembly for the whole Scorpaenidae family. We positioned the species in the teleost tree for the first time and studied its complex repeat and gene content. In this study, a lionfish near-chromosome genome assembly of high-quality and contiguity was constructed from genomic data derived from three MinION flow cells and an Illumina Hiseq4000 platform, finally ending in a total size of 902.3 Mb and distributed in 660 contigs. To our knowledge, this is the first reference genome in Pteroinae and so far the only available genome of the Scorpaenidae family.

Repeat content, gene prediction & functional annotation

The representation of TEs in the *P. miles* genome (46.5% of genome assembly) is notably higher than in other species inside Perciformes (Table 9), such as *G. aculeatus*, 13.02% (Shao et al., 2019), *S. lucioperca*, 39.0% (Nguinkal et al., 2019) and *E. lanceolatus*, 45.1% (Wang et al., 2019). Furthermore, its repetitive content is higher compared to species of similar genome size (0.9-1 Gb) such as *O. niloticus* (21.34%), *A. mexicanus* (25.21%), *O. latipes* (26.74%), *C. idella* (40.08%) and *L. oculatus* (16.06%), as presented by (Shao et al., 2019, and references therein). All aforementioned TE analyses in the different genomes, including ours presented herein, have been implemented using a different strategy for identifying the TEs of each species, a fact that could bias the results not allowing us to do a direct comparison. However, *P. miles* TE content exceeds remarkably most other fish genomes, with potential biological role in the species evolution. Elevated genome-wide repeat content has been previously linked to adaptation (Yuan et al., 2018) and invasiveness (Stapley et al., 2015; Danis et al., 2021) Thus, this higher proportion of TEs in the genome of *P. miles* could potentially play a key factor in adaptive evolution of species and consequently success in thriving in new environments. On the composition of TEs, the percentage of DNA transposons (Class II) in the assembled genome (28.63-33.8%) is only comparable to the corresponding ones in *D. rerio* (46.27% with ∼1.3 Gb genome size, Shao et al. (2019)) and *C. idella* (25.57% with ∼900 Mb genome size, Shao et al., 2019). Despite the positive correlation between genome size and the abundance of TEs in fish genomes, also confirmed here, it would be extremely interesting to investigate further the relationship between the TE heterogeneity, in terms of copy number and composition, and genomes' evolution (Sotero-Caio et al., 2017; Shao et al., 2019). For example, 20 distinct TE superfamilies were recognised in the genome of *P. miles*, from a minimum of 77 Jockey elements to about 311,700 CACTA (Table 2), taking advantage both of the thorough classification resulting from the designed pipeline and the detailed annotation from "RM parser.py" (Supplementary Figure 1).

Additionally, studies have revealed the multi-functional role of TEs on the evolution of vertebrate genomes, from genomic architecture (Sotero-Caio et al., 2017) to their relationship with non-coding RNAs (Bourque et al., 2018) and confluence to transcription regulation (Drongitis et al., 2019; Fueyo et al., 2022). Albeit, it would be noteworthy, in the superfamily level, to explore the patterns in the accumulation of TEs, their roles and consequently their contribution to gene duplication events, and genome dynamics in general.

Phylogenomic analysis, synteny and gene repertoire

Scorpaenidae (order: Perciformes) is a taxonomically widespread family which includes by now 370 marine species (Smith et al., 2018, and references therein), known to be venomous. Despite their worldwide distribution and diversity, this group's biology is clearly understudied, as well as their unexplored phylogeny. Here, we presented the first phylogenetic tree that includes a representative of this family, devil firefish *P. miles*, based on whole genome data (Figure 2). This effort could be an origin for further genomic and evolutionary studies inside this family.

One-to-one orthologous genes between *P. miles* and *G. aculeatus*, exhibited high conserved synteny (Figure 3), which confirms the high quality and completeness of constructed genome assembly. Indeed, between contigs 2, 3, 4, 6 and 7 of devil firefish and chromosomes 16, 10, 11, 18 and 5 of three-spined stickleback, there was high pairwise conservation, respectively. Taking into consideration that the haploid number of *P. miles* should be the same as its con-generic species, *P. volitans*, n=24 (Nirchio et al., 2014), an interesting fact arose. This revealed the fusion of coding regions from more than one contigs of *P. miles* to single chromosomes of *G. aculeatus* (e.g. contigs 1 and 12 to chromosome 7, contigs 5, 29, 41 to chromosome 1, contigs 9, 16, 17 to chromosome 4) and their later rearrangements (e.g. contigs 8, 23 to chromosome 19), an additional support of the high accuracy constructed assembly.

Duplication events estimation and descriptive functional analysis unveiled the extended presence of gene families, being involved in major biological processes, such as metabolism, somatic growth, immunity and reproduction (Figure 5). These families may potentially contribute to species morphology, anti-predatory tactics, rapid spread and adaptation in new marine habitats. Noteworthy, a sufficient number of immune-related gene families were identified, including immunoglobulins (Ig heavy-chain variable, light-chain variable genes), interleukins (interleukin 10 receptor), lysozymes (antimicrobial response), genes contributing to the regulation of antiviral innate immunity (e.g. TRIM35) and transcription factors that regulate the expression of MHC class II genes.

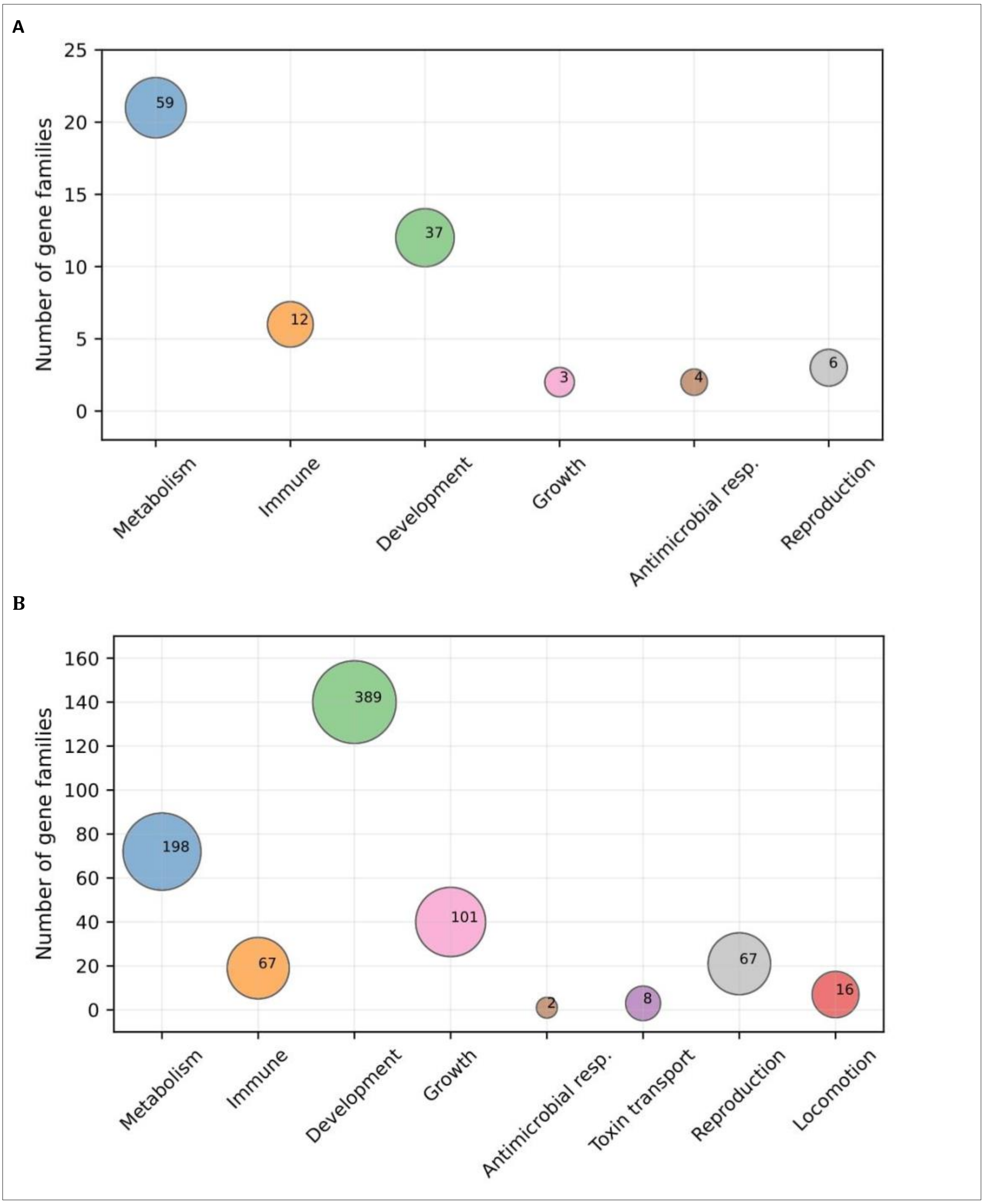

Figure 5 - Number of orthogroups associated with specific biological processes for (A) rapidly expanding from CAFE, and (B) with duplications from GeneRax. The size of each circle is defined by the binary logarithm ($\log_2$) of the number of genes multiplied by the number of unique terms and then adding a scalar factor.The visualization of figures was performed with a custom python script "GO_plots.py"

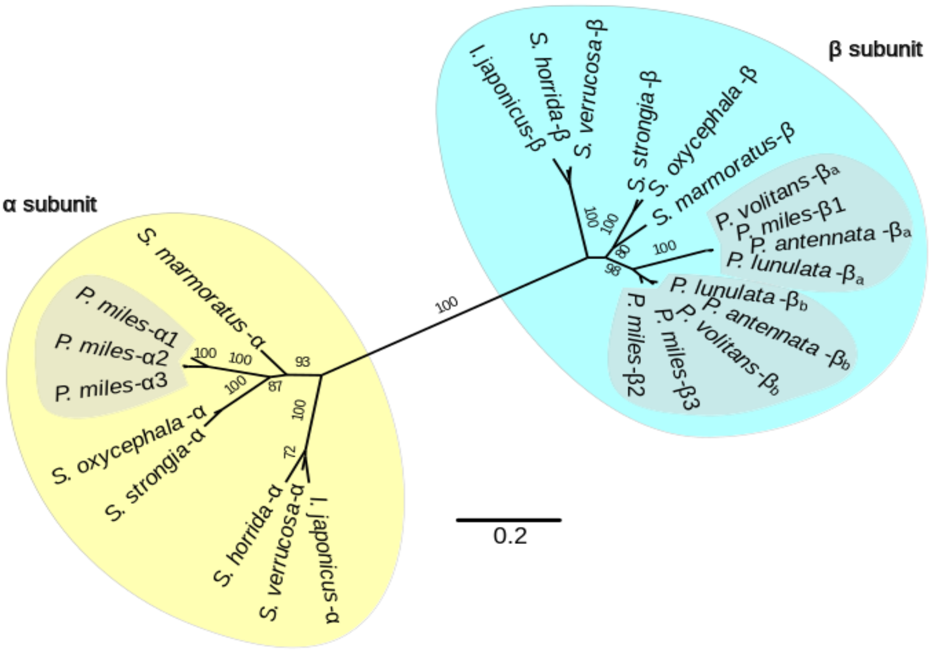

Figure 6 - Maximum-likelihood unrooted phylogenetic tree of the two subunits of scorpaenid toxins. α subunits are presented inside the yellow and β in light blue bubble. For the phylogeny we used the JTT-DCMUT + I + G4 substitution model and conducted 100 non-parametric bootstrap replicates.

Table 6 - Scorpionfish toxins downloaded from NCBI

Name	Accession number	Length (aa)	Species
pltoxin-a 1	BAM74455.1	699	*Pterois lunulata*
pltoxin-b 1	BAM74456.1	698	*Pterois lunulata*
patoxin-a 2	BAK18812.1	699	*Pterois antennata*
patoxin-b 2	BAK18813.1	698	*Pterois antennata*
pvtoxin-a 2	BAK18814.1	699	*Pterois volitans*
pvtoxin-b 2	BAK18815.1	698	*Pterois volitans*
ijtoxin-a 1	BAM74457.1	703	*Inimicus japonicus*
ijtoxin-b 1	BAM74458.1	700	*Inimicus japonicus*
stonustoxin-a 3	AAC60022.1	703	*Synanceia horrida*
stonustoxin-b 3	AAC60021.1	700	*Synanceia horrida*
neoverrucotoxin-a 4	BAF41221.1	703	*Synanceia verrucosa*
neoverrucotoxin-b 4	BAF41222.1	700	*Synanceia verrucosa*
Tx-a 5	AIC84045.1	703	*Sebastapistes strongia*
Tx-b 5	AIC84046.1	700	*Sebastapistes strongia*
Tx-a 5	AIC84047.1	703	*Scorpaenopsis oxycephala*
Tx-b 5	AIC84048.1	700	*Scorpaenopsis oxycephala*
Tx-a 5	AIC84049.1	702	*Sebastiscus marmoratus*
Tx-b 5	AIC84050.1	700	*Sebastiscus marmoratus*

1.Kiriake et al. (2013), 2.Kiriake et al. (2011), 3.Ghadessy et al. (1996), 4.Ueda et al. (2006), 5.Chuang and Shiao (2014)

An interesting finding was a detected duplication in the gene family of meprins (meprin-F in fish, Marın (2015)), proteins that are involved in toxins transport. Based on the results, it could be worth additional studies on genes responsible for the unique morphology (e.g. spines development) of devil firefish and its successful adaptation to new habitats, with the contribution of more genomic data inside the family of Scorpaenidae, that would become available in the future.

Lionfish toxins evolution

Scorpaeniform fish toxins are multifunctional proteins that have, among others, lethal, cytolytic, hemolytic, inflammatory, nociceptive and neuromuscular activities (Campos et al., 2021). Scorpaeniform fish use their venom (toxins) mostly for defense, when the threat touches their spines (Diaz, 2015; Campos et al., 2021). These toxins are formed by two subunits α and β (Kiriake and Shiomi, 2011; Kiriake et al., 2013; Chuang and Shiao, 2014; Campos et al., 2021), being actively organized in either heterodimeric or tetrameric proteins (Campos et al., 2021).

Table 7 - Summary of genomic sequencing throughput

Sequencing technology	Raw reads	Quality-controlled reads	Coverage
Illumina	24,836074	20,980,358	3.5x
MinION	2,245,868	2,224,738	39.5x

Table 8 - BUSCO completeness assessment (%) in each step of structural annotation

Type (Source)	Complete	Single	Duplicate	Fragmented	Missing	Final
HQ isoforms (PacBio)	77.3	19.9	57.4	1.9	20.8	79.2
Consensus transcripts (Mikado)	74.7	69	5.7	1.6	23.7	76.7
Genes (Augustus)	82	80.7	1.3	5.4	12.6	87.4
Consensus genes (PASA-filtered)	88.8	84.7	4.1	5	6.2	93.8

In spite of the identification of toxins in other lionfishes (*P. lunulata*, *P. volitans* and *P. antennata*), through cDNA cloning and immunoblotting (Kiriake and Shiomi, 2011; Kiriake et al., 2013), the lack of genomic data inside this genus makes it difficult to understand their relationship with other scorpaenid fish cytolysins and their evolution. We investigated the presence and identification of toxin genes on the genome, and reconstructed the phylogeny of lionfish toxins (Figure 6). The identification of three toxin genes per subunit in devil firefish and their phylogeny inside toxins from various scorpaenid fishes, rejected a previous hypothesis about the evolution of lionfish toxins. This theory proposed the absence of α subunit gene in species of genus *Pterois* and the origination of toxin genes from the β subunit of scorpaenids and a later duplication event occurred prior to the speciation of Pteroinae (Chuang and Shiao, 2014).

Table 9 - Fish genome size and TE content comparison

Species	Genome size (Mb)	TE content (%)	Reference
Pterois miles	902.5	46.51	present study
Gasterosteus aculeatus	461.5	13.02	Shao et al., 2019
Sander lucioperca	1014	39.0	Nguinkal et al., 2019
Epinephelus lanceolatus	1128	45.1	Wang et al., 2019
Oreochromis niloticus	927.3	21.34	Shao et al., 2019
Astyanax mexicanus	1191.2	25.21	Shao et al., 2019
Oryzias latipes	868.9	26.74	Shao et al., 2019
Ctenopharyngodon idella	900.5	40.08	Shao et al., 2019
Lepisosteus oculatus	945.8	16.06	Shao et al., 2019

Conclusion

In this study, we provide the first near-chromosome and high-quality genome assembly of devil firefish, its complex repeat and gene content, we construct the first phylogeny including a member of genus *Pterois*, based on whole genome sequencing data, and baseline the evolution of lionfish toxins. All the analyses performed here, highlighted the importance of *P. miles* genome as a valuable resource for further studies regarding the influence of transposable elements on genome evolution, the correlation between gene duplications and adaptation to new niches, lionfish rapid spread worldwide and its dominance, and scorpaenid toxins evolution.

Acknowledgements

The authors would like to thank the scuba diver who kindly provided the specimen and the aquarists of the Cretaquarium (C.K. Doxa, T. Doulamis, P. Grigoriou, G. Vardanis, E. Pinakis) for keeping the species alive as well as for their assistance at the sampling. This research was supported through computational resources provided by IMBBC (Institute of Marine Biology, Biotechnology and Aquaculture) of the HCMR (Hellenic Centre for Marine Research)–Zorbas HPC infrastructure. Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, LifeWatchGreece RI, and the CMBR (Centre for the study and sustainable exploitation of Marine Biological Resources). Preprint version 6 of this article has been peer-reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100241; Irisarri, 2023)

Scripts and code availability

All customs scripts, designed workflows and used software commands that have been used during this study are available at the following GitHub repositories:

https://github.com/ckitsoulis/Pterois-miles-Genome (https://doi.org/10.5281/zenodo.8167419; Kitsoulis, 2023a)

https://github.com/ckitsoulis/ELDAR (https://doi.org/10.5281/zenodo.8167397; Kitsoulis, 2023b)

https://github.com/genomenerds/SnakeCube

Data and supplementary information availability

Genomic data from Illumina and Oxford Nanopore Technologies, and transcriptomic data from Pacific Biosciences can be accessed in the European Nucleotide Archive (ENA) under the IDS ERR10286272, ERR10286273 and ERR10286274-81 respectively. The genome assembly along with the raw data have been deposited to ENA under the study with accession PRJEB56286. Supplementary figures/tables, gene and functional annotations can be found in the supplementary data files (https://doi.org/10.1101/2023.01.10.523469).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

References

Al Mabruk SAA, Rizgalla J (2019) First record of lionfish (Scorpaenidae: Pterois) from Libyan waters. Journal of the Black Sea / Mediterranean Environment, 25, 108–114.

Albins M, Hixon M (2008) Invasive Indo-Pacific lionfish Pterois volitans reduce recruitment of Atlantic coral-reef fishes. Marine Ecology Progress Series, 367, 233–238. https://doi.org/10.3354/meps07620

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

Andrews S (2010) FASTQC. A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angelova N, Danis T, Lagnel J, Tsigenopoulos CS, Manousaki T (2022) SnakeCube: containerized and automated pipeline for de novo genome assembly in HPC environments. BMC Research Notes, 15, 98. https://doi.org/10.1186/s13104-022-05978-5

Ao J, Mu Y, Xiang L-X, Fan D, Feng M, Zhang S, Shi Q, Zhu L-Y, Li T, Ding Y et al. (2015) Genome Sequencing of the Perciform Fish Larimichthys crocea Provides Insights into Molecular and Genetic Mechanisms of Stress Adaptation. PLOS Genetics, 11, e1005118. https://doi.org/10.1371/journal.pgen.1005118

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffels A, Rash S, Hoon S, Smit A et al. (2002) Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. Science, 297, 1301–1310. https://doi.org/10.1126/science.1072104

Araki K, Aokic J, Kawase J, Hamada K, Ozaki A, Fujimoto H, Yamamoto I, Usuki H (2018) Whole Genome Sequencing of Greater Amberjack (Seriola dumerili) for SNP Identification on Aligned Scaffolds and Genome Structural Variation Analysis Using Parallel Resequencing. International Journal of Genomics, 2018, 1–12. https://doi.org/10.1155/2018/7984292

Arim M, Abades SR, Neill PE, Lima M, Marquet PA (2006) Spread dynamics of invasive species. Proceedings of the National Academy of Sciences, 103, 374–378. https://doi.org/10.1073/pnas.0504272102

Azzurro E, Stancanelli B, Di Martino V, Bariche M (2017) Range expansion of the common lionfish Pterois miles (Bennett, 1828) in the Mediterranean Sea: an unwanted new guest for Italian waters. BioInvasions Records, 6, 95–98. https://doi.org/10.3391/bir.2017.6.2.01

Bariche M, Kleitou P, Kalogirou S, Bernardi G (2017) Genetics reveal the identity and origin of the lionfish invasion in the Mediterranean Sea. Scientific Reports, 7. https://doi.org/10.1038/s41598-017-07326-1

Bariche M, Torres M, Azzurro E (2013) The presence of the invasive Lionfish Pterois miles in the Mediterranean Sea. Mediterranean Marine Science, 14, 292–294. https://doi.org/10.12681/mms.428

Barrett SCH (2015) Foundations of invasion genetics: the Baker and Stebbins legacy. Molecular Ecology, 24, 1927–1941. https://doi.org/10.1111/mec.13014

Bax N, Williamson A, Aguero M, Gonzalez E, Geeves W (2003) Marine invasive alien species: A threat to global biodiversity. Marine Policy, 27, 313–323. https://doi.org/10.1016/S0308-597X(03)00041-1

Bian C, Li J, Lin X, Chen X, Yi Y, You X, Zhang Y, Lv Y, Shi Q (2019) Whole Genome Sequencing of the Blue Tilapia (Oreochromis aureus) Provides a Valuable Genetic Resource for Biomedical Research on Tilapias. Marine Drugs, 17, 386. https://doi.org/10.3390/md17070386

Bilge G, Filiz H, Yapıcı S (2017) Occurrences of Pterois miles (Bennett, 1828) between 1992 and 2016 from Turkey and the Mediterranean Sea. Journal of the Black Sea / Mediterranean Environment, 23, 201–208.

Bista I, Wood JMD, Desvignes T, McCarthy SA, Matschiner M, Ning Z, Tracey A, Torrance J, Sims Y, Chow W, Smith M, Oliver K, Haggerty L, Salzburger W, Postlethwait JH, Howe K, Clark MS, I WHDII, Cheng C-HC, Miska EA, Durbin R (2022) Genomics of cold adaptations in the Antarctic notothenioid fish radiation. bioRxiv, 2022.06.08.494096. https://doi.org/10.1101/2022.06.08.494096

Blakeslee AMH, Manousaki T, Vasileiadou K, Tepolt CK (2019) An evolutionary perspective on marine invasions. Evolutionary Applications. https://doi.org/10.1111/eva.12906

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C (2018) Ten things you should know about transposable elements. Genome Biology, 19, 199. https://doi.org/10.1186/s13059-018-1577-z

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J et al. (2016) The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics, 48, 427–437. https://doi.org/10.1038/ng.3526

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E et al. (2014) The genomic substrate for adaptive radiation in African cichlid fish. Nature, 513, 375–381. https://doi.org/10.1038/nature13726

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12, 59–60. https://doi.org/10.1038/nmeth.3176

Cai M, Zou Y, Xiao S, Li W, Han Z, Han F, Xiao J, Liu F, Wang Z (2019) Chromosome assembly of Collichthys lucidus, a fish of Sciaenidae with a multiple sex chromosome system. Scientific Data, 6. https://doi.org/10.1038/s41597-019-0139-x

Campos F V., Fiorotti HB, Coitinho JB, Figueiredo SG (2021) Fish Cytolysins in All Their Complexity. Toxins, 13, 877. https://doi.org/10.3390/toxins13120877

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution, 38, 5825–5829. https://doi.org/10.1093/molbev/msab293

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chen Z, Omori Y, Koren S, Shirokiya T, Kuroda T, Miyamoto A, Wada H, Fujiyama A, Toyoda A, Zhang S et al. (2019) De novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication. Science Advances, 5. https://doi.org/10.1126/sciadv.aav0547

Cheng P, Huang Y, Lv Y, Du H, Ruan Z, Li C, Ye H, Zhang H, Wu J, Wang C et al. (2021) The American Paddlefish Genome Provides Novel Insights into Chromosomal Evolution and Bone Mineralization in Early Vertebrates. Molecular Biology and Evolution, 38, 1595–1607. https://doi.org/10.1093/molbev/msaa326

Chiesa S, Azzurro E, Bernardi G (2019) The genetics and genomics of marine fish invasions: a global review. Reviews in Fish Biology and Fisheries, 29, 837–859. https://doi.org/10.1007/s11160-019-09586-8

Chuang P-S, Shiao J-C (2014) Toxin gene determination and evolution in scorpaenoid fish. Toxicon, 88, 21–33. https://doi.org/10.1016/j.toxicon.2014.06.013

Conte MA, Gammerdinger WJ, Bartie KL, Penman DJ, Kocher TD (2017) A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. BMC Genomics, 18, 341. https://doi.org/10.1186/s12864-017-3723-5

Conte MA, Kocher TD (2015) An improved genome reference for the African cichlid, Metriaclima zebra. BMC Genomics, 16, 724. https://doi.org/10.1186/s12864-015-1930-5

Côté IM, Smith NS (2018) The lionfish Pterois sp. invasion: Has the worst-case scenario come to pass? Journal of Fish Biology, 92, 660–689. https://doi.org/10.1111/jfb.13544

Crocetta F, Agius D, Balistreri P, Bariche M, Bayhan YK, Çakir M, Ciriaco S, Corsini-Foka M, Deidun A, El Zrelli R et al. (2015) New mediterranean biodiversity records (October 2015). Mediterranean Marine Science, 16, 682–702. https://doi.org/10.12681/mms.1477

Dailianis T, Akyol O, Babali N, Bariche M, Crocetta F, Gerovasileiou V, Ghanem R, Gökoglu M, Hasiotis T, Izquierdo-Muñoz A et al. (2016) New Mediterranean Biodiversity Records (July 2016). Mediterranean Marine Science, 17, 608–626. https://doi.org/10.12681/mms.1734

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. GigaScience, 10. https://doi.org/10.1093/gigascience/giab008

Danis T, Papadogiannis V, Tsakogiannis A, Kristoffersen JB, Golani D, Tsaparis D, Sterioti A, Kasapidis P, Kotoulas G, Magoulas A et al. (2021) Genome Analysis of Lagocephalus sceleratus: Unraveling the Genomic Landscape of a Successful Invader. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.790850

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T (2020) ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Molecular Biology and Evolution, 37, 291–294. https://doi.org/10.1093/molbev/msz189

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018) NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics, 34, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Degrazia D, Beauchamp TL (2019) Beyond the 3 Rs to a More Comprehensive Framework of Principles for Animal Research Ethics. ILAR Journal, 60, 308–317. https://doi.org/10.1093/ilar/ilz011

Diaz JH (2015) Marine Scorpaenidae Envenomation in Travelers: Epidemiology, Management, and Prevention. Journal of Travel Medicine, 22, 251–258. https://doi.org/10.1111/jtm.12206

Dimitriou AC, Chartosia N, Hall-Spencer JM, Kleitou P, Jimenez C, Antoniou C, Hadjioannou L, Kletou D, Sfenthourakis S (2019) Genetic data suggest multiple introductions of the lionfish (Pterois miles) into the Mediterranean Sea. Diversity, 11. https://doi.org/10.3390/d11090149

Ding W, Zhang X, Zhao X, Jing W, Cao Z, Li J, Huang Y, You X, Wang M, Shi Q, Bing X (2021) A Chromosome-Level Genome Assembly of the Mandarin Fish (Siniperca chuatsi). Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.671650

Dray L, Neuhof M, Diamant A, Huchon D (2016) The complete mitochondrial genome of the devil firefish Pterois miles (Bennett, 1828) (Scorpaenidae). Mitochondrial DNA, 27, 783–784. https://doi.org/10.3109/19401736.2014.945565

Drongitis D, Aniello F, Fucci L, Donizetti A (2019) Roles of Transposable Elements in the Different Layers of Gene Expression Regulation. International Journal of Molecular Sciences, 20, 5755. https://doi.org/10.3390/ijms20225755

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology, 20, 238. https://doi.org/10.1186/s13059-019-1832-y

Fan G, Chan J, Ma K, Yang B, Zhang H, Yang X, Shi C, Law H, Ren Z, Xu Q et al. (2018) Chromosome-level reference genome of the Siamese fighting fish Betta splendens, a model species for the study of aggression. GigaScience. https://doi.org/10.1093/gigascience/giy087

Feron R, Zahm M, Cabau C, Klopp C, Roques C, Bouchez O, Eché C, Valière S, Donnadieu C, Haffray P et al. (2020) Characterization of a Y-specific duplication/insertion of the anti-Mullerian hormone type II receptor gene based on a chromosome-scale genome assembly of yellow perch, Perca flavescens. Molecular Ecology Resources, 20, 531–543. https://doi.org/10.1111/1755-0998.13133

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences, 117, 9451–9457. https://doi.org/10.1073/PNAS.1921046117

Fueyo R, Judd J, Feschotte C, Wysocka J (2022) Roles of transposable elements in the regulation of mammalian transcription. Nature Reviews Molecular Cell Biology, 23, 481–497. https://doi.org/10.1038/s41580-022-00457-y

Gao Z, You X, Zhang X, Chen J, Xu T, Huang Y, Lin X, Xu J, Bian C, Shi Q (2021) A chromosome-level genome assembly of the striped catfish (Pangasianodon hypophthalmus). Genomics, 113, 3349–3356. https://doi.org/10.1016/j.ygeno.2021.07.026

Gasteiger E (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research, 31, 3784–3788. https://doi.org/10.1093/nar/gkg563

Ghadessy FJ, Chen D, Kini RM, Chung MaxeyCM, Jeyaseelan K, Khoo HE, Yuen R (1996) Stonustoxin Is a Novel Lethal Factor from Stonefish (Synanceja horrida) Venom. Journal of Biological Chemistry, 271, 25575–25581. https://doi.org/10.1074/jbc.271.41.25575

Golani D, Sonin O (1992) New Records of the Red Sea Fishes, Pterois miles (Scorpaenidae) and Pteragogus pelycus (Labridae) from the Eastern Mediterranean Sea. Japanese Journal of Ichthyology, 39, 167–169. https://doi.org/10.11369/jji1950.39.167

Guerrero-Cózar I, Gomez-Garrido J, Berbel C, Martinez-Blanch JF, Alioto T, Claros MG, Gagnaire P-A, Manchado M (2021) Chromosome anchoring in Senegalese sole (Solea senegalensis) reveals sex-associated markers and genome rearrangements in flatfish. Scientific Reports, 11, 13460. https://doi.org/10.1038/s41598-021-92601-5

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

Haas BJ (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research, 31, 5654–5666. https://doi.org/10.1093/nar/gkg770

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L et al. (2014) Erratum: Corrigendum: The zebrafish reference genome sequence and its relationship to the human genome. Nature, 505, 248–248. https://doi.org/10.1038/nature12813

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ et al. (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research, 47, D309–D314. https://doi.org/10.1093/nar/gky1085

í Kongsstovu S, Mikalsen SO, Homrum E, Jacobsen JA, Flicek P, Dahl HA (2019) Using long and linked reads to improve an Atlantic herring (Clupea harengus) genome assembly. Scientific Reports, 9. https://doi.org/10.1038/s41598-019-54151-9

Irisarri I (2023) The genome of a dangerous invader (fish) beauty. Peer Community in Genomics. http://doi.org/10.24072/pci.genomics.100241

Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature, 431, 946–957. https://doi.org/10.1038/nature03025

Jasonowicz AJ, Simeon A, Zahm M, Cabau C, Klopp C, Roques C, Iampietro C, Lluch J, Donnadieu C, Parrinello H et al. (2022) Generation of a chromosome-level genome assembly for Pacific halibut (Hippoglossus stenolepis) and characterization of its sex-determining genomic region. Molecular Ecology Resources, 22, 2685–2700. https://doi.org/10.1111/1755-0998.13641

Kai W, Kikuchi K, Tohari S, Chew AK, Tay A, Fujiwara A, Hosoya S, Suetake H, Naruse K, Brenner S, Suzuki Y, Venkatesh B (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. https://doi.org/10.1093/gbe/evr041

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature, 447, 714–719. https://doi.org/10.1038/nature05846

Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30, 772–780. https://doi.org/10.1093/molbev/mst010

Katsanevakis S, Poursanidis D, Hoffman R, Rizgalla J, Rothman SBS, Levitt-Barmats Y, Hadjioannou L, Trkov D, Garmendia JM, Rizzo M et al. (2020) Unpublished mediterranean records of marine alien and cryptogenic species. BioInvasions Records, 9, 165–182. https://doi.org/10.3391/bir.2020.9.2.01

Kelley JL, Yee M-C, Brown AP, Richardson RR, Tatarenkov A, Lee CC, Harkins TT, Bustamante CD, Earley RL (2016) The Genome of the Self-Fertilizing Mangrove Rivulus Fish, Kryptolebias marmoratus : A Model for Studying Phenotypic Plasticity and Adaptations to Extreme Environments. Genome Biology and Evolution, 8, 2145–2154. https://doi.org/10.1093/gbe/evw145

Kiriake A, Shiomi K (2011) Some properties and cDNA cloning of proteinaceous toxins from two species of lionfish (Pterois antennata and Pterois volitans). Toxicon, 58, 494–501. https://doi.org/10.1016/j.toxicon.2011.08.010

Kiriake A, Suzuki Y, Nagashima Y, Shiomi K (2013) Proteinaceous toxins from three species of scorpaeniform fish (lionfish Pterois lunulata, devil stinger Inimicus japonicus and waspfish Hypodytes rubripinnis): Close similarity in properties and primary structures to stonefish toxins. Toxicon, 70, 184–193. https://doi.org/10.1016/j.toxicon.2013.04.021

Kitsoulis C (2023a) ckitsoulis/Pterois-miles-Genome: v1.0.0 (v1.0.0). Zenodo https://doi.org/10.5281/zenodo.8167419

Kitsoulis C (2023b) ckitsoulis/ELDAR: v1.0.0 (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.8167397

Kleitou P, Moutopoulos DK, Giovos I, Kletou D, Savva I, Cai LL, Hall-Spencer JM, Charitou A, Elia M, Katselis G, Rees S (2022) Conflicting interests and growing importance of non-indigenous species in commercial and recreational fisheries of the Mediterranean Sea. Fisheries Management and Ecology, 29, 169–182. https://doi.org/10.1111/fme.12531

Kletou D, Hall-Spencer JM, Kleitou P (2016) A lionfish (Pterois miles) invasion has begun in the Mediterranean Sea. Marine Biodiversity Records, 9, 46. https://doi.org/10.1186/s41200-016-0065-y

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology, 37, 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koskinen JP, Holm L (2012) SANS: high-throughput retrieval of protein sequences allowing 50% mismatches. Bioinformatics, 28, i438–i443. https://doi.org/10.1093/bioinformatics/bts417

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: An information aesthetic for comparative genomics. Genome Research, 19, 1639–1645. https://doi.org/10.1101/gr.092759.109

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A et al. (2016) The Atlantic salmon genome provides insights into rediploidization. Nature, 533, 200–205. https://doi.org/10.1038/nature17164

Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, Lee AP, Woltering JM, Ravi V, Gunter HM et al. (2016) The seahorse genome and the evolution of its specialized morphology. Nature, 540, 395–399. https://doi.org/10.1038/nature20595

Liu D, Wang X, Guo H, Zhang X, Zhang M, Tang W (2021) Chromosome-level genome assembly of the endangered humphead wrasse Cheilinus undulatus : Insight into the expansion of opsin genes in fishes. Molecular Ecology Resources, 21, 2388–2406. https://doi.org/10.1111/1755-0998.13429

Lv Y, Li Y, Liu Y, Wen Z, Yang Y, Qin C, Shi Q, Mu X (2022) An Updated Genome Assembly Improves Understanding of the Transcriptional Regulation of Coloration in Midas Cichlid. Frontiers in Marine Science, 9. https://doi.org/10.3389/fmars.2022.950573

Lyons TJ, Tuckett QM, Hill JE (2019) Data quality and quantity for invasive species: A case study of the lionfishes. Fish and Fisheries, faf.12374. https://doi.org/10.1111/faf.12374

Makałowski W, Gotea V, Pande A, Makałowska I (2019) Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. In:, pp. 177–207. https://doi.org/10.1007/978-1-4939-9074-0_6

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution, 38, 4647–4654. https://doi.org/10.1093/molbev/msab199

Marín I (2015) Origin and Diversification of Meprin Proteases. PLOS ONE, 10, e0135924. https://doi.org/10.1371/journal.pone.0135924

McGaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H, Jeffery WR, Keene A, Ma L et al. (2014) The cavefish genome reveals candidate genes for eye loss. Nature Communications, 5, 5307. https://doi.org/10.1038/ncomms6307

Mendes FK, Vanderpool D, Fulton B, Hahn MW (2021) CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics, 36, 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ (2020) GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. Molecular Biology and Evolution, 37, 2763–2774. https://doi.org/10.1093/molbev/msaa141

Nath S, Shaw DE, White MA (2021) Improved contiguity of the threespine stickleback genome using long-read sequencing. G3 Genes|Genomes|Genetics, 11. https://doi.org/10.1093/g3journal/jkab007

Nguinkal JA, Brunner RM, Verleih M, Rebl A, de los Ríos-Pérez L, Schäfer N, Hadlich F, Stüeken M, Wittenburg D, Goldammer T (2019) The First Highly Contiguous Genome Assembly of Pikeperch (Sander lucioperca), an Emerging Aquaculture Species in Europe. Genes, 10, 708. https://doi.org/10.3390/genes10090708

Nirchio M, Eheman N, Siccha-Ramirez R, Pérez JE, Rossi AR, Oliveira C (2014) Karyotype of the invasive species Pterois volitans (Scorpaeniformes: Scorpaenidae) from Margarita Island, Venezuela. Revista de Biología Tropical, 62, 1365. https://doi.org/10.15517/rbt.v62i4.13029

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology, 20, 275. https://doi.org/10.1186/s13059-019-1905-y

Pan H, Yu H, Ravi V, Li C, Lee AP, Lian MM, Tay B-H, Brenner S, Wang J, Yang H, Zhang G, Venkatesh B (2016) The genome of the largest bony fish, ocean sunfish (Mola mola), provides insights into its fast growth rate. GigaScience, 5, 36. https://doi.org/10.1186/s13742-016-0144-3

Papadogiannis V, Manousaki T, Nousias O, Tsakogiannis A, Kristoffersen JB, Mylonas CC, Batargias C, Chatziplis D, Tsigenopoulos CS (2023) Chromosome genome assembly for the meagre, Argyrosomus regius, reveals species adaptations and sciaenid sex-related locus evolution. Frontiers in Genetics, 13. https://doi.org/10.3389/fgene.2022.1081760

Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35, 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pérez-Sánchez J, Naya-Català F, Soriano B, Piazzon MC, Hafez A, Gabaldón T, Llorens C, Sitjà-Bobadilla A, Calduch-Giner JA (2019) Genome Sequencing and Transcriptome Analysis Reveal Recent Species-Specific Gene Duplications in the Plastic Gilthead Sea Bream (Sparus aurata). Frontiers in Marine Science, 6. https://doi.org/10.3389/fmars.2019.00760

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Reid NM, Jackson CE, Gilbert D, Minx P, Montague MJ, Hampton TH, Helfrich LW, King BL, Nacci DE, Aluru N et al. (2017) The Landscape of Extreme Genomic Variation in the Highly Adaptable Atlantic Killifish. Genome Biology and Evolution, 9, 659–676. https://doi.org/10.1093/gbe/evx023

Reid BN, Moran RL, Kopack CJ, Fitzpatrick SW (2021) Rapture-ready darters: Choice of reference genome and genotyping method (whole-genome or sequence capture) influence population genomic inference in Etheostoma. Molecular Ecology Resources, 21, 404–420. https://doi.org/10.1111/1755-0998.13275

Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, von Schalburg KR, Lemon C, Bird NH, Koop BF (2014) The Genome and Linkage Map of the Northern Pike (Esox lucius): Conserved Synteny Revealed between the Salmonid Sister Group and the Neoteleostei. PLoS ONE, 9, e102089. https://doi.org/10.1371/journal.pone.0102089

Ryu T, Herrera M, Moore B, Izumiyama M, Kawai E, Laudet V, Ravasi T (2022) A chromosome-scale genome assembly of the false clownfish, Amphiprion ocellaris. G3 Genes|Genomes|Genetics, 12. https://doi.org/10.1093/g3journal/jkac074

Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff J-N, Lesch K-P et al. (2013) The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. Nature Genetics, 45, 567–572. https://doi.org/10.1038/ng.2604

Schultz ET (1986) Pterois volitans and Pterois miles: Two Valid Species. Copeia, 1986, 686. https://doi.org/10.2307/1444950

Shao F, Han M, Peng Z (2019) Evolution and diversity of transposable elements in fish genomes. Scientific Reports, 9, 15399. https://doi.org/10.1038/s41598-019-51888-1

Shao F, Ludwig A, Mao Y, Liu N, Peng Z (2020) Chromosome-level genome assembly of the female western mosquitofish (Gambusia affinis). GigaScience, 9. https://doi.org/10.1093/gigascience/giaa092

Shao F, Pan H, Li P, Ni L, Xu Y, Peng Z (2021) Chromosome-Level Genome Assembly of the Asian Red-Tail Catfish (Hemibagrus wyckioides). Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.747684

Shen Y, Chalopin D, Garcia T, Boswell M, Boswell W, Shiryev SA, Agarwala R, Volff J-N, Postlethwait JH, Schartl M et al. (2016) X. couchianus and X. hellerii genome models provide genomic variation insight among Xiphophorus species. BMC Genomics, 17, 37. https://doi.org/10.1186/s12864-015-2361-z

Smith WL, Everman E, Richardson C (2018) Phylogeny and Taxonomy of Flatheads, Scorpionfishes, Sea Robins, and Stonefishes (Percomorpha: Scorpaeniformes) and the Evolution of the Lachrymal Saber. Copeia, 106, 94–119. https://doi.org/10.1643/CG-17-669

Sotero-Caio CG, Platt RN, Suh A, Ray DA (2017) Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biology and Evolution, 9, 161–177. https://doi.org/10.1093/gbe/evw264

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24, 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stapley J, Santure AW, Dennis SR (2015) Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Molecular Ecology, 24, 2241–2252. https://doi.org/10.1111/mec.13089

Sun L, Gao T, Wang F, Qin Z, Yan L, Tao W, Li M, Jin C, Ma L, Kocher TD, Wang D (2020) Chromosome-level genome assembly of a cyprinid fish Onychostoma macrolepis by integration of nanopore sequencing, Bionano and Hi-C technology. Molecular Ecology Resources, 20, 1361–1371. https://doi.org/10.1111/1755-0998.13190

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics, 25. https://doi.org/10.1002/0471250953.bi0410s25

Thompson AW, Wojtas H, Davoll M, Braasch I (2022) Genome of the Rio Pearlfish (Nematolebias whitei), a bi-annual killifish model for Eco-Evo-Devo in extreme environments. G3 Genes|Genomes|Genetics, 12. https://doi.org/10.1093/g3journal/jkac045

Tian H-F, Hu Q-M, Li Z (2021) A high-quality de novo genome assembly of one swamp eel (Monopterus albus) strain with PacBio and Hi-C sequencing data. G3 Genes|Genomes|Genetics, 11. https://doi.org/10.1093/g3journal/jkaa032

Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RST, Hecht J, Knaust F, Belkhir K, Klages S, Dieterich R et al. (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications, 5, 5770. https://doi.org/10.1038/ncomms6770

Törönen P, Holm L (2022) PANNZER —A practical tool for protein function prediction. Protein Science, 31, 118–128. https://doi.org/10.1002/pro.4193

Ueda A, Suzuki M, Honma T, Nagai H, Nagashima Y, Shiomi K (2006) Purification, properties and cDNA cloning of neoverrucotoxin (neoVTX), a hemolytic lethal factor from the stonefish Synanceia verrucosa venom. Biochimica et Biophysica Acta (BBA) - General Subjects, 1760, 1713–1722. https://doi.org/10.1016/j.bbagen.2006.08.017

Vaser R, Sović I, Nagarajan N, Šikić M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research, 27, 737–746. https://doi.org/10.1101/gr.214270.116

Vavasis C, Simotas G, Spinos E, Konstantinidis E, Minoudi S, Triantafyllidis A, Perdikaris C (2020) Occurrence of Pterois miles in the Island of Kefalonia (Greece): the Northernmost Dispersal Record in the Mediterranean Sea. Thalassas: An International Journal of Marine Sciences, 36, 171–175. https://doi.org/10.1007/s41208-019-00175-x

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience, 7. https://doi.org/10.1093/gigascience/giy093

Vij S, Kuhl H, Kuznetsova IS, Komissarov A, Yurchenko AA, Van Heusden P, Singh S, Thevasagayam NM, Prakki SRS, Purushothaman K et al. (2016) Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. PLOS Genetics, 12, e1005954. https://doi.org/10.1371/journal.pgen.1005954

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE, 9, e112963. https://doi.org/10.1371/journal.pone.0112963

Wang D, Chen X, Zhang X, Li J, Yi Y, Bian C, Shi Q, Lin H, Li S, Zhang Y, You X (2019) Whole Genome Sequencing of the Giant Grouper (Epinephelus lanceolatus) and High-Throughput Screening of Putative Antimicrobial Peptide Genes. Marine Drugs, 17, 503. https://doi.org/10.3390/md17090503

Wang H, Su B, Butts IAE, Dunham RA, Wang X (2022) Chromosome-level assembly and annotation of the blue catfish Ictalurus furcatus , an aquaculture species for hybrid catfish reproduction, epigenetics, and heterosis studies. GigaScience, 11. https://doi.org/10.1093/gigascience/giac070

Warren WC, García-Pérez R, Xu S, Lampert KP, Chalopin D, Stöck M, Loewe L, Lu Y, Kuderna L, Minx P et al. (2018) Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. Nature Ecology & Evolution, 2, 669–679. https://doi.org/10.1038/s41559-018-0473-y

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Wilcox CL, Motomura H, Matsunuma M, Bowen BW (2018) Phylogeography of Lionfishes (Pterois) Indicate Taxonomic Over Splitting and Hybrid Origin of the Invasive Pterois volitans. Journal of Heredity, 109, 162–175. https://doi.org/10.1093/jhered/esx056

Wu B, Feng C, Zhu C, Xu W, Yuan Y, Hu M, Yuan K, Li Y, Ren Y, Zhou Y, Jiang H, Qiu Q, Wang W, He S, Wang K (2021) The Genomes of Two Billfishes Provide Insights into the Evolution of Endothermy in Teleosts. Molecular Biology and Evolution, 38, 2413–2427. https://doi.org/10.1093/molbev/msab035

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics, 21, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Xie B, Yu H, Kerkkamp H, Wang M, Richardson M, Shi Q (2019) Comparative transcriptome analyses of venom glands from three scorpionfishes. Genomics, 111, 231–241. https://doi.org/10.1016/j.ygeno.2018.11.012

Xu X, Shao C, Xu H, Zhou Q, You F, Wang N, Li W, Li M, Chen S (2020) Draft genomes of female and male turbot Scophthalmus maximus. Scientific Data, 7, 90. https://doi.org/10.1038/s41597-020-0426-6

Yan H, Bombarely A, Li S (2020) DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics, 36, 4269–4275. https://doi.org/10.1093/bioinformatics/btaa519

Yang X, Liu H, Ma Z, Zou Y, Zou M, Mao Y, Li X, Wang H, Chen T, Wang W, Yang R (2019) Chromosome-level genome assembly of Triplophysa tibetana , a fish adapted to the harsh high-altitude environment of the Tibetan Plateau. Molecular Ecology Resources, 19, 1027–1036. https://doi.org/10.1111/1755-0998.13021

Yu G (2020) Using ggtree to Visualize Data on Tree-Like Structures. Current Protocols in Bioinformatics, 69. https://doi.org/10.1002/cpbi.96

Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z (2018) Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics, 19, 141. https://doi.org/10.1186/s12864-018-4516-1

Zafeiropoulos H, Gioti A, Ninidakis S, Potirakis A, Paragkamian S, Angelova N, Antoniou A, Danis T, Kaitetzidou E, Kasapidis P et al. (2021) 0s and 1s in marine molecular research: a regional HPC perspective. GigaScience, 10. https://doi.org/10.1093/gigascience/giab053

Zhao N, Guo H, Jia L, Guo B, Zheng D, Liu S, Zhang B (2021) Genome assembly and annotation at the chromosomal level of first Pleuronectidae: Verasper variegatus provides a basis for phylogenetic study of Pleuronectiformes. Genomics, 113, 717–726. https://doi.org/10.1016/j.ygeno.2021.01.024

Zheng S, Shao F, Tao W, Liu Z, Long J, Wang X, Zhang S, Zhao Q, Carleton KL, Kocher TD et al. (2021a) Chromosome-level assembly of southern catfish (Silurus meridionalis) provides insights into visual adaptation to nocturnal and benthic lifestyles. Molecular Ecology Resources, 21, 1575–1592. https://doi.org/10.1111/1755-0998.13338