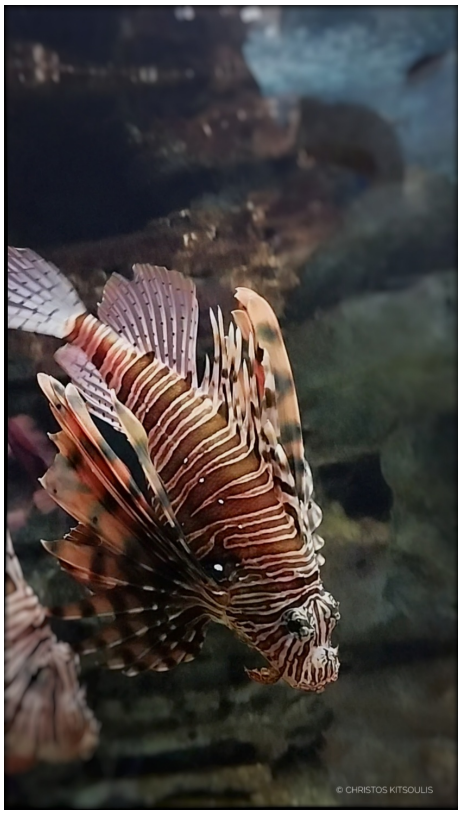
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

hcmr
ΕΛΚΕΘΕ

Master's of Science Program in Bioinformatics
School of Medicine, University of Crete
September 2022
Diploma thesis

Genomic assembly and bioinformatic analyses of devil firefish (*Pterois miles*) genome using Oxford Nanopore, Pacific Biosciences and Illumina sequencing technologies



Student:	**Christos Kitsoulis** [1] [2]
Supervisor:	Assistant Researcher, Dr. Tereza Manousaki [1]
Co-supervisor:	Assistant Professor, Dr. Ioannis Iliopoulos [2],
Committee:	Dr. Tereza Manousaki [1], Dr. Ioannis Iliopoulos [2], Dr. Pavlos Pavlidis [3]

[1]Institute of Marine Biology, Biotechnology & Aquaculture, HCMR, Gournes, Crete
[2]Faculty of Medicine, University of Crete, Voutes, Crete
[3]Foundation for Research & Technology - Hellas, Voutes, Crete

Affirmation

I hereby affirm that the present Master diploma thesis was submitted to the School of Medicine, University of Crete, for the Master's of Science degree (M.Sc.) in "Bioinformatics".

September 2022, Heraklion

Christos Kitsoulis

Abstract

Devil firefish (*Pterois miles*), member of Scorpaenidae, is one of the most successful marine invaders, dominating around the world, that was rapidly spread into the Mediterranean Sea, through the Suez Canal, originating from the Red Sea. In this study we built and analyzed the first reference genome assembly of *P. miles* and explored its evolutionary background. Through genome analysis, we revealed the big amount of transposable elements present on the genome, while with phylogenomic analyses we constructed the first fish phylogeny which includes a member of genus Pterois, based on whole genome sequencing data. Furthermore, we identified six individual genes that encode the two subunits (three genes per subunit) of the toxins secreted from devil firefish's spines and are responsible for the harmful touch, adding a major baseline for additional studies on the lionfish toxins origin and evolution. The high-quality and contiguity genome assembly built here provides a valuable resource for future studies about the species biology, the influence of transposable elements on the evolution of vertebrate genomes and fish toxins evolution.

Keywords:

devil firefish, genome analysis, transposable elements, phylogenomic analyses, toxins, gene duplications

Περίληψη

Το λεοντόψαρο (*Pterois miles*), μέλος των Scorpaenidae, είναι ένας από τους πιο επιτυχημένους θαλάσσιους εισβολείς, κυριαρχώντας σε όλο τον κόσμο, το οποίο εξαπλώθηκε γρήγορα και στη Μεσόγειο Θάλασσα, μέσω της Διώρυγας του Σουέζ, έχοντας προέλευση από την Ερυθρά Θάλασσα. Σε αυτή τη μελέτη κατασκευάσαμε και αναλύσαμε το πρώτο γονιδιώμα αναφοράς του *P. miles* και εξερευνήσαμε το εξελικτικό του υπόβαθρο. Μέσω της ανάλυσης του γονιδιώματος, αποκαλύφθηκε η μεγάλη ποσότητα μεταθετών στοιχείων που υπάρχουν στο γονιδίωμα, ενώ με φυλογενετικές αναλύσεις κατασκευάσαμε την πρώτη φυλογένεση τελεόστεων που να περιλαμβάνει ένα μέλος του γένους *Pterois*, βασισμένο σε δεδομένα ολόκληρων γονιδιωμάτων. Επιπλέον, εντοπίσαμε έξι γονίδια που κωδικοποιούν τις δύο υπομονάδες (τρία γονίδια ανά υπομονάδα) των τοξινών που εκκρίνονται από τις άκανθες του λεοντόψαρου και είναι υπεύθυνα για το επίπονο άγγιγμα, προσθέτοντας μια σημαντική βάση για πρόσθετες μελέτες σχετικά με την προέλευση και την εξέλιξη των τοξινών στα λεοντόψαρα. Το υψηλής ποιότητας και συνέχειας γονιδίωμα που κατασκευάστηκε εδώ αποτελεί μια πολύτιμη πηγή για μελλοντικές μελέτες σχετικά με τη βιολογία του είδους, την επίδραση των μεταθετών στοιχείων στην εξέλιξη των γονιδιωμάτων των σπονδυλωτών και την εξέλιξη των τοξινών στα ψάρια.

Acknowledgements

First of all, I would like to thank my supervisor, Dr. Tereza Manousaki, who has given me the opportunity not only to conduct my thesis to her group but also the ability to scratch a little the area of evolutionary genomics, get passionate about bioinformatics and explore my limits. Her support during my thesis was more than significant.

I also truly appreciate both my co-supervisor Dr. Ioannis Iliopoulos, for his constant helpful attitude, and the Genomics & Bioinformatics group principal investigator Dr. Costas Tsigenopoulos, whose calmness and companionable attitude were really important.

Moreover, I couldn't be more than grateful to meet all the members of the bioinformatics group along this experience. Dr. Vasilis Papadogiannis, our postdoc, for his constant scientific integrity, important advice and valuable assistance during my thesis. Machi Papadopoulou, for our endless conversations, coffee company and her elegant ideas and suggestions about parts of my thesis. Katerina Katirtzoglou and Bastien Mace, for their friendly attitude and pleasant conversations during meal breaks and hangouts. Elsa Iliopoulou, for her tech-assistance and advice at the start of my thesis, despite her personal stress and effort during the ending of hers. Haris Zafeiropoulos for our tech-talks, exchange of ideas and his valuable advice, and Savvas Paragkamian, for his friendly attitude when we were sharing the lab.

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). 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) RI. And so, I would like to thank the administrators of the HPC cluster "Zorbas", Stelios Ninidakis and Antonis Potirakis for their tech-support and troubleshooting, but also the people who built, maintain and support it, Dimitris Sidirokastritis and Vaggelis Pafilis.

Furthermore, I would like to thank the Norwegian Sequencing Centre (`www.sequencing.uio.no`), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities, for their provided sequencing services.

Last but not least, a huge thank you to my parents, Vasilis and Foteini, my sister, Penelope and all my friends for their love, inspiration, understanding and encouragement all these years in everything I have tried, wouldn't be enough.

Contents

List of Abbreviations

ONT - Oxford Nanopore Techologies
WGS - Whole Genome Sequencing
TE - Transposable Element
EDTA - Extensive *de novo* TE Annotator
HPC - High performace computing
PacBio - Pacific Biosciences
LQ - Low Quality
HQ - High Quality
SMRT - Single-Molecule Real Time
CCS - circular consensus sequencing
FL - Full Length
FLNC - Full Length Non-Contatamer
DB - Database
HOGs - Hierarchical Orthogroups
ORF - Open Reading Frame

List of Figures

List of Tables

1 INTRODUCTION

The devil firefish, *Pterois miles* (Bennett, 1828), is a venomous species native to the Indo-Pacific region, from South Africa to Red Sea and East to Sumatra (Schultz, 1986) which belongs to the Scorpaenidae family. The first occurrence of *P. miles*, as a single specimen, in the Mediterranean Sea has been recorded by the Levantine coast in 1991 (Golani and Sonin, 1992), while the second, of two individuals, was almost twenty years later (Bariche et al., 2013). Soon after, the frequency of appearances, along the eastern Mediterranean, 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). While the origin of species colonization in the Mediterranean Sea followed the invasion pattern of other Lessepsian immigrants 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 some 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 of *P. miles* and a con-generic species *P. volitans*, together referred as the invasive lionfish complex (Lyons et al., 2019), in the western Atlantic is one of the fastest and most dominant marine fish introductions to date (Kletou et al., 2016, and references therein). For the Mediterranean, 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 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's dynamics and adaptations to new niche (Barrett, 2015). Data derived from Whole Genome Sequencing (WGS) could provide promising opportunities in the exploration of potential adaptations that shape fitness of invaders, as well as the dynamics of colonization.

Factors associated with *P. miles* biology such as rapid somatic growth, signature anti-predatory defenses (Côté and Smith, 2018), reproductive success, discernible predatory behaviour, low parasitism and ecological flexibility are potential features which explain its rapid distribution in the Mediterranean Sea (Dailianis et al., 2016), whereas the dynamics of species' populations indicate a rapidly progressing increase along the coastlines (Kletou et al., 2016). Yet, only a few genomic references of the species are available, just 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 thesis is to construct and analyze the first high-quality genome assembly of *P. miles*, taking advantage of the constant progress of sequencing technologies, efficiency of computational resources and available bioinformatic methodologies. For that reason, 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 representative genome within the family of Scorpaenidae, this valuable resource could provide a critical conveyance to unveil and highlight the insights of species' biology, fish toxins evolution, species ecology and phylogeny in further investigation of invasive traits across the Mediterranean Sea.

2 MATERIALS AND METHODS

2.1 Sample collection, libraries construction and sequencing

The processes of sample collection, libraries construction and sequencing for genomic data were conducted by group of collaborators, as described by Kitsoulis et al. (2022). Genomic sequencing results and statistics are presented in Table 3.1. For transcriptomic data, the library preparation was conducted using Pacific Biosciences protocol for Iso-Seq™ Express Template Preparation for Sequel and Sequel II Systems. The samples of seven tissues (brain, gonad, gills, heart, liver, muscle and spleen) were sequenced on a PacBio sequel II instrument, on one SMRT cell. Both library preparation and sequencing were carried out by the Norwegian Sequencing Centre (`www.sequencing.uio.no`), hosted by the University of Oslo.

2.2 Genomic data pre-processing

2.2.1 Long reads

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).

2.2.2 Short reads

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 with the following parameters: (i) 4-base sliding window with a cutting-off threshold score lower than 15 Phred (*SLIDINGWINDOW: 4:15*), (ii) leading and trailing bases with score less than 10 Phred are trimmed out (*LEADING: 10, TRAILING: 10*), (iii) reads shorter than 75 bp and average score lower than 30 Phred are removed (*MINLEN: 75, AVGQUAL: 30*).

2.3 *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. The ONT reads were used

for the construction of the initial *de novo* assembly, at first, and the first rounds of polishing, while the Illumina reads were used for the later rounds of polishing, afterwards. 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. 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 the draft assembly by Minimap v2.22 (Li, 2018). Further polishing was performed by Medaka v1.4.4 (`https://github.com/nanoporetech/medaka`) and consequently with Pilon v1.23 (Walker et al., 2014) for which the preprocessed Illumina reads were used, after being mapped against the resulting assembly from Medaka, using Minimap v2.22 (Li, 2018).
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).

2.4 Quality assessment of draft assemblies

The resultant assemblies (e.g. draft, intermediate and final) all through the above procedure were evaluated by two commonly used criteria: (i) the N50 statistic from contigs' size, 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 and visualized (Table 3.2, Figure 3.1) with custom python tool, ELDAR (`https://github.com/ckitsoulis/ELDAR`).

2.5 Genome annotation

2.5.1 Transposable elements annotation

A *de novo* Transposable elements (TEs) library was generated from the previously 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. Sub-libraries (ii)-(iii) were classified again using DeepTE (Yan et al., 2020), a transposon classification tool which depends on convolutional neural network (CNN). The annotation probability threshold was strictly set to 0.8 ("*-prop_thr 0.8*"). A step of headers' correction and reformation, via 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 (Table 3.3), based on an up-to-date TE classification system (Makalowski et al., 2019), a python-based parser "RM_parser.py" was developed for the output files of RepeatMasker. The designed workflow is presented schematically in Figure 2.1.

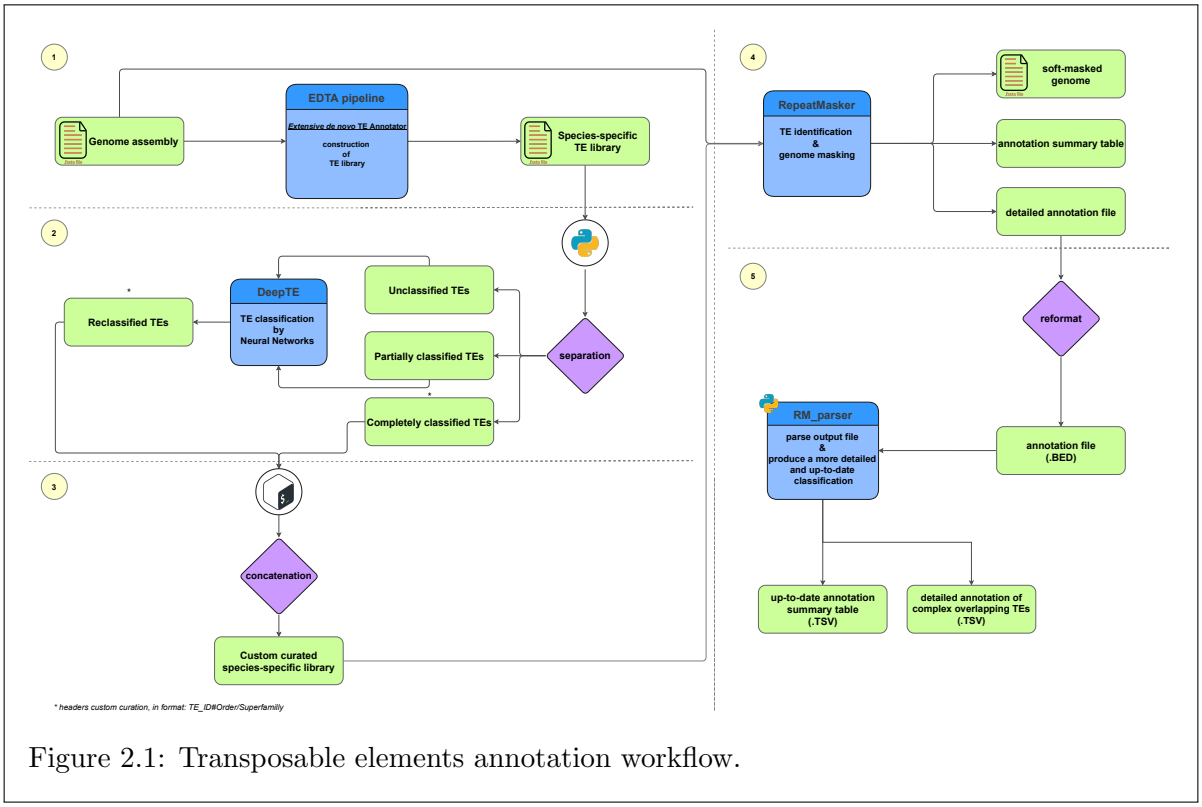


Figure 2.1: Transposable elements annotation workflow.

2.5.2 Long-read transcriptome analysis

After finished sequencing, CCS reads were generated using CCS application on SMRT Link v10.2 and Iso-Seq analysis was performed on them, with default parameters, using Iso-Seq pipeline IsoSeq v3.4 (`https://github.com/PacificBiosciences/IsoSeq`), until the production of high quality 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 IsoSeq pipeline are presented in Table 3.4. In order to have a general view of the shared representation of HQ transcripts between tissues and potential tissue-specific isoforms, the count table of HQ transcripts per tissue, provided by IsoSeq pipeline, was modified and used to visualize their intersections (Supplementary Figure 2) using UpSetR package (`https://github.com/hms-dbmi/UpSetR`), in a custom R script "upset_plot.R". 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 by samtools v1.15.1 (Danecek et al., 2021), and the redundant transcripts were 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.

2.5.3 Structural annotation - gene prediction

After repeat masking, gene prediction was conducted based on a hybrid strategy of transcriptome-based (non-redundant HQ transcripts), homology-based (curated protein sets) and *ab initio* methods, using a semi-automated workflow consisted of 12 tools and intermediate custom python and bash scripts (Figure 2.2). In the first step, HQ isoforms and curated proteomes of 20 actinopterygian species were aligned (splice-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 2.1), 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 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 over iii) a blast hit score of 0.5 and with iv) 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* (by Papadogiannis et al. (2022)) were aligned to 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 *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 resulted annotation of the first round. In the last step, genes were filtered to remove predictions with in-frame STOP codons and those that overlap 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 3.6).

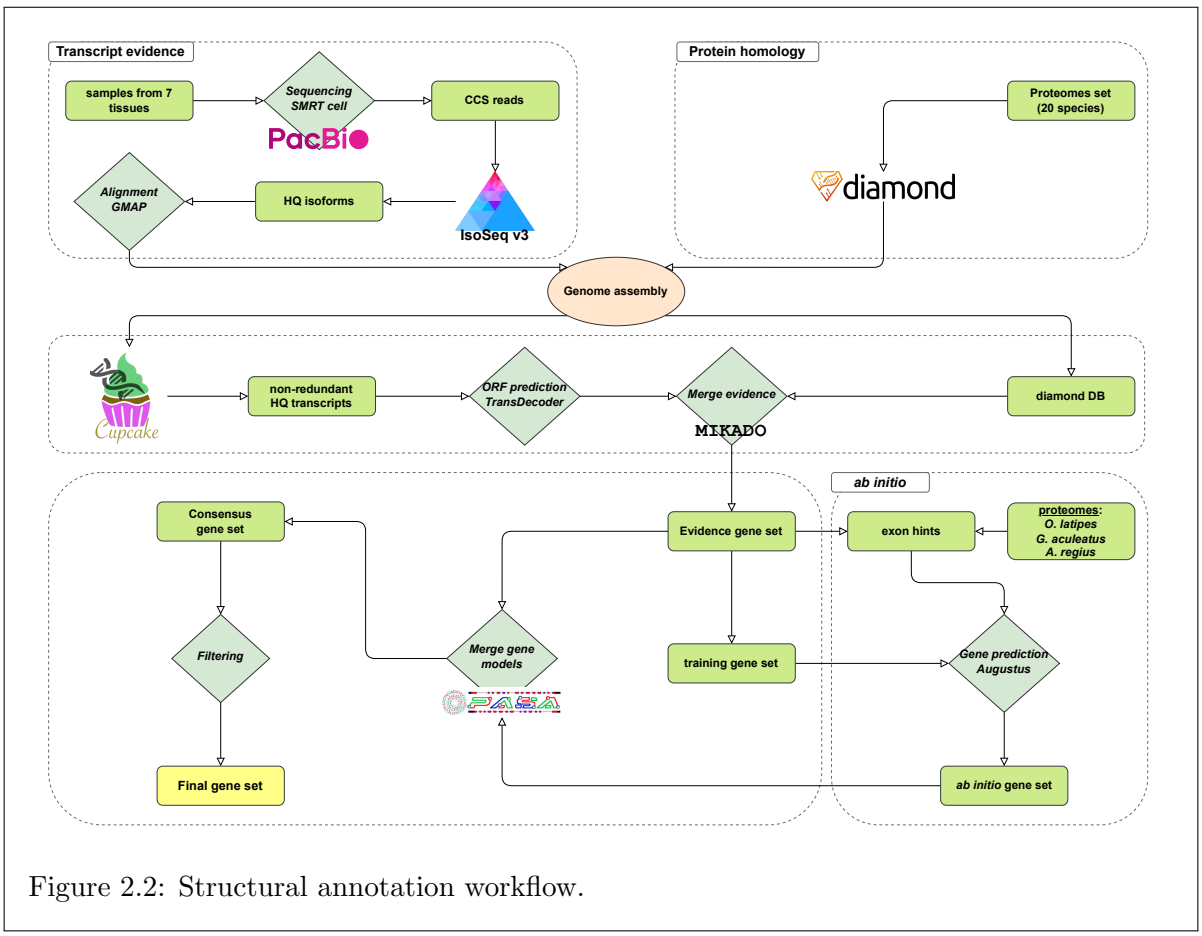


Figure 2.2: Structural annotation workflow.

Table 2.1: Species included in protein homology BLAST database.

Scientific name	Common name	UniProt ID	Number of proteins	Reference
Amphilophus citrinellus	Midas cichlid	UP000261340	31,742	Yunyun 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	Pérez-Sánchez 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)

2.5.4 Functional annotation

The functional annotation of *P. miles* predicted gene set was performed using three different strategies and tools, respectively. 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 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 derived 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.

2.6 Phylogenomic analysis

2.6.1 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 *P.miles*, were compared using OrthoFinder v2.5.2 (Emms and Kelly, 2019), with default parameters. The initial dataset was collected from Genomes - NCBI Datasets (`https://www.ncbi.nlm.nih.gov/datasets/genomes/`) and Ensembl DB (`https://www.ensembl.org/`) based on the following criteria: i) Scaffold level and greater, 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.

2.6.2 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, the ones containing only a single gene copy per species, to exclude potential paralogs. Afterwards, only 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). The aligned orthogroup sequences were corrected, in the case of missing taxa, using a custom python script (gimli_clean.py) and then concatenated to 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.

2.7 Comparative genomic analysis

2.7.1 Synteny analysis

Synteny analysis was performed on gene level between *P. miles* and *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 it) were selected for visualization against the 21 chromosomes of *G. aculeatus*, using Circos (Krzywinski et al., 2009).

2.7.2 Gene families expansion and contraction

Changes in gene families 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, gene families with more than 10 species out of 47 absent and families with a difference of sequences greater than 70 between the species with the maximum number of sequences and the species with the minimum, 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 TIMETREE (`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 (Figure 3.5), as a subset of the phylogenomic tree (Figure 3.3).

2.7.3 Duplication events 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-Gutiérrez et al., 2009), in strict mode. Families with less than three sequences were excluded from the following procedure. From each gene family was estimated the best-fit model 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 substitution model correction in some cases, gene trees and their models along with the phylogenetic species tree were used to estimate duplication events with GeneRax.

2.7.4 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, 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.

2.7.5 Toxin genes evolution in lionfishes

To identify genes responsible for the secreted toxin of devil firefish, toxins (proteins) from 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 2.2) and aligned to the constructed 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 selection 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 managed and visualized with FigTree v.1.4.4 (`http://tree.bio.ed.ac.uk/software/figtree/`).

Table 2.2: Scorpiofish toxins downloaded from NCBI.

Name	**Accession number**	**Length (aa)**	**Species origin**
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 and Shiomi (2011), 3.Ghadessy et al. (1996), 4.Ueda et al. (2006), 5.Chuang and Shiao (2014)

3 RESULTS

3.1 Genomic sequencing results

Sequencing yielded a total of 38.66 Gb of raw genomic ONT reads, from which 36.16 Gb had a quality 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 (Table 3.1).

Table 3.1: Summary of genomic sequencing throughput.

Sequencing technology	**Raw reads**	**Quality-controlled reads**	**Coverage**
Illumina	24,836,074	20,980,358	3.5 x
Oxford Nanopore Technologies	2,245,868	2,224,738	39.5 x

3.2 Genome size and assembly completeness

The final genome assembly contained 660 contigs with a total length of about 902.5 Mb. The longest contig was sized at 36.5 Mb and the N50 statistic value at 14.5 Mb (Table 3.2). At least, 90% of the genome size was represented in the 83 longest contigs of the produced assembly (Figure 3.1). The GC content of the genome was calculated at 40.78% (GC-rich regions at the 42 longest contigs are presented in Supplementary Figure 1). About genome completeness assessment, 3566 out of 3640 BUSCO genes have were present (98%), against the Actinopterigyian ortholog dataset (v.10). By those, 3551 genes (97.6%) were complete, while only 74 (2.0%) were missing (Table 3.2), suggesting a high level of contiguity and completeness of the *de novo* genome assembly.

3.3 Genome annotation

3.3.1 Transposable elements annotation

About 46.5% of the genome assembly ($\sim$ 416.7 Mb) in *P. miles*, consisted of transposable elements (Figure 3.2). 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%, while 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

Table 3.2: Polished genome assembly statistics and completeness.

Number of contigs	660
Total length	902,353,306 bp
GC content (%)	40.78
Longest contig	36,477,432 bp
N50	14,490,642 bp
N75	5,565,202 bp
L90	83
BUSCO completeness score	
Complete	3551 (97.6%)
Single	3515 (96.6%)
Duplicated	36 (1.0%)
Fragmented	15 (0.4%)
Missing	74 (2.0%)
Total number of Actinopterygii orthologs	3,640 (98%)

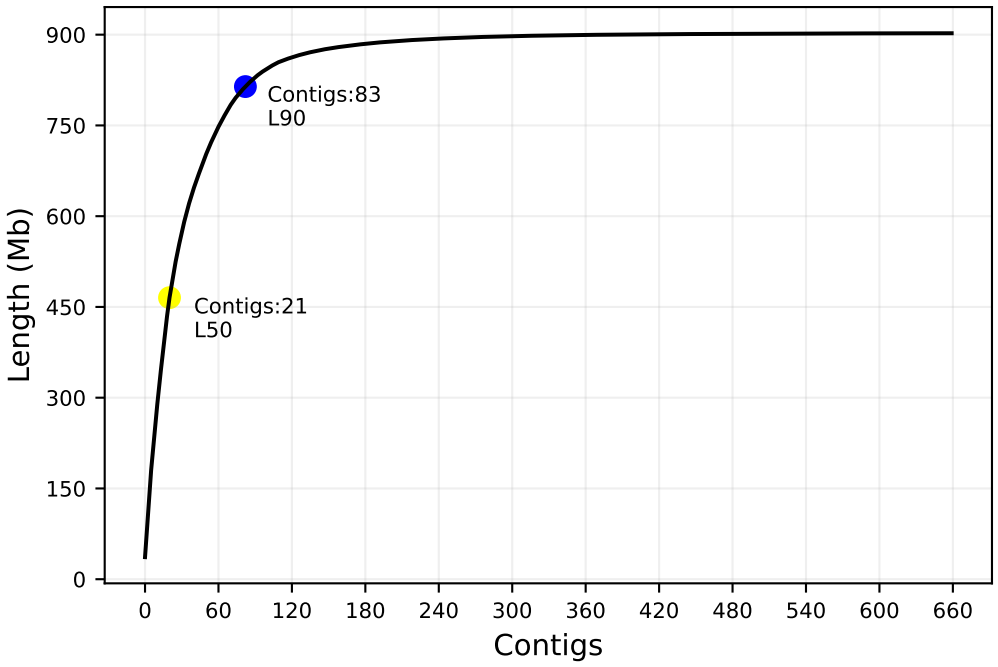


Figure 3.1: Cumulative sum of contigs' lengths of *P. miles* genome assembly. The yellow dot shows the number of contigs that represent at least 50% (L50) and the blue one at least 90% (L90) of the genome size, respectively.

high-confident 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 3.3). The distribution of TE content in the 42 longest contigs is presented in Supplementary Figure 1.

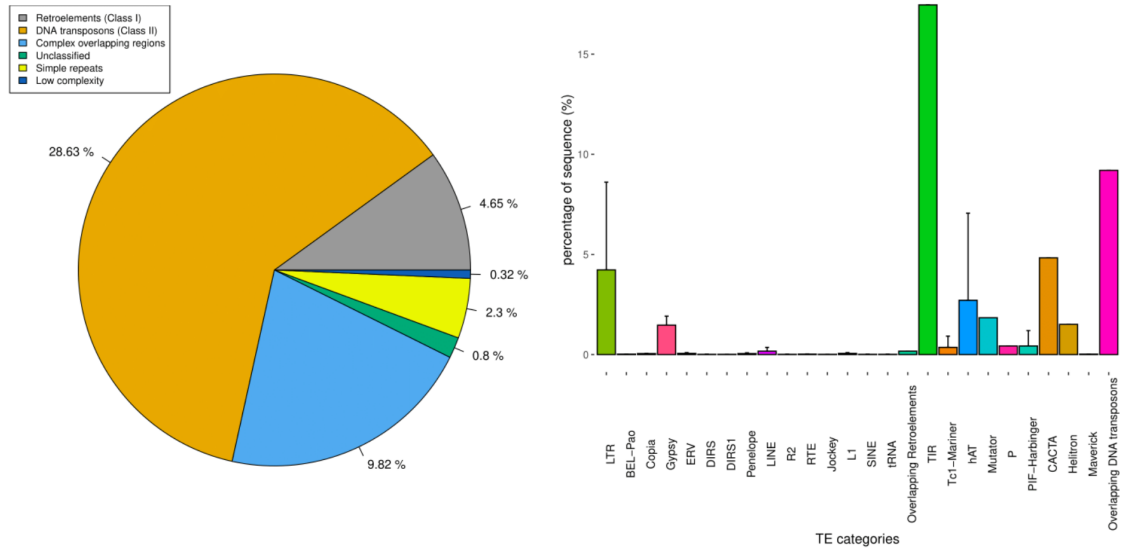


Figure 3.2: Percentage of TE categories representation in the genome of *P. miles*.

3.3.2 Transcriptome analysis

From 6,245,243 CCS reads, Iso-seq analysis yielded a total of 124,307 HQ consensus isoforms. The samples with the most transcripts were the heart, spleen and liver (Table 3.4), and they shared almost 6,500 (unique) of them (Supplementary Figure 2). However, the total amount of unique HQ transcripts shared between all sampled tissues was notably low (∼1,000). From the total number of HQ transcripts, 91,666 (73.75%) were aligned properly to the assembled genome (representing 73.74%) and used as evidence for the gene prediction.

3.3.3 Structural and Functional annotation

The hybrid approach of HQ full-length transcripts-based, homology-based and *ab initio*-based methods resulted in 25,410 candidate protein-coding gene models at total. 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 representing, in size, 382.3 Mb of the genome assembly (Table 3.5). A total 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.
From a core set of 3,640 single-copy ortholog genes (Actinopterigyii lineage, odb 10), 3414 (93.8%) were found to be present in the predicted gene set (Table 3.6), with 3233

Table 3.3: Transposable elements annotation statistics.

Transposable elements	number of elements	length occupied (bp)	% representation
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	2,3489	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 Gupsy/DIRS1 by RepeatMasker
in parenthesis are presented the percentages calculated by RepeatMasker

(88.8%) identified as complete (3082 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).

3.4 Orthology assignment and Phylogenomic analysis

The total number of genes analysed by OrthoFinder, included in the proteomes of all 47 teleost fish species (Supplementary Table 1), was 1,108,753 and 97.8% of them assigned

Table 3.4: From CCS reads 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

Table 3.5: Basic statistics of predicted gene models.

Type	Number	Mean size (bp)	Longest (bp)	Length (Mb)	Genome (%)
gene	24,639	15,519	445,933	382.39	42.37
transcript	25,034	15,442	445,933	386.57	42.84
5' UTR	13,119	228	10,216	2.21	0.24
exon	231,315	207	14,732	48.09	5.33
CDS	227,240	159	6,853	36.16	4.00
intron	206,281	1,609	188,322	338.47	37.5
3' UTR	11,114	914	14,608	9.72	1.07

to 28,397 phylogenetic hierarchical orthogroups (HOGS). After the filtering step, 1,193 HOGs were selected to construct the superalignment matrix. Before trimming, matrix was 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 3.3). At the resulting maximum-likelihood phylogenetic tree, almost all branches were supported with 100 bootstraps. Based on the constructed phylogeny, *P. miles* is placed within the Perciformes clade.

3.5 Comparative genomic analysis

3.5.1 Synteny analysis

Synteny analysis on gene level unveiled high conserved syntenic coding regions between the 42 longest contigs of *P. miles* and chromosomes of *G. aculeatus*, sharing at total 8,035 one-to-one ortholog genes (Figure 3.4).

3.5.2 Gene family size evolution

Gene family evolution analysis estimated 228 rapidly evolving gene families (out of 15,405 included families) of *P. miles*, at a significance level of 0.01 (p-value). From these rapidly evolving families, 136 were identified as expanding and 92 as contracting, respectively. The total number of genes included in these rapidly expanding families was 373. The

Table 3.6: Completeness assessment in each step of structural annotation workflow.

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	69.0	5.7	1.6	23.7	76.7
Genes (Augustus)	82.0	80.7	1.3	5.4	12.6	87.4
Filtered genes (PASA)	88.8	84.7	4.1	5.0	6.2	93.8

corresponding state of the number of estimated gene families' gains-losses inside the Perciformes species is presented in Figure 3.5, as a subset of Figure 3.3.

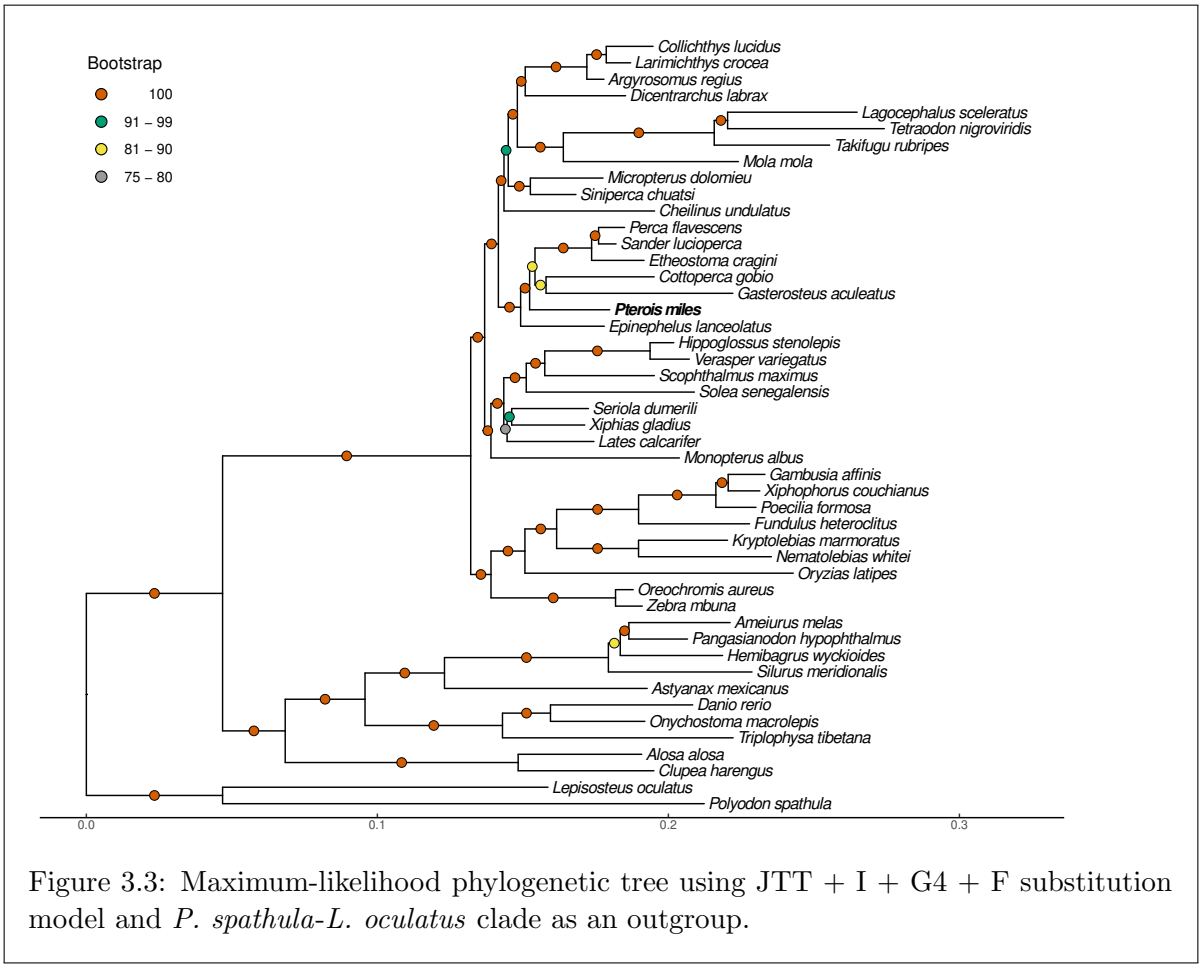


Figure 3.3: Maximum-likelihood phylogenetic tree using JTT + I + G4 + F substitution model and *P. spathula*-*L. oculatus* clade as an outgroup.

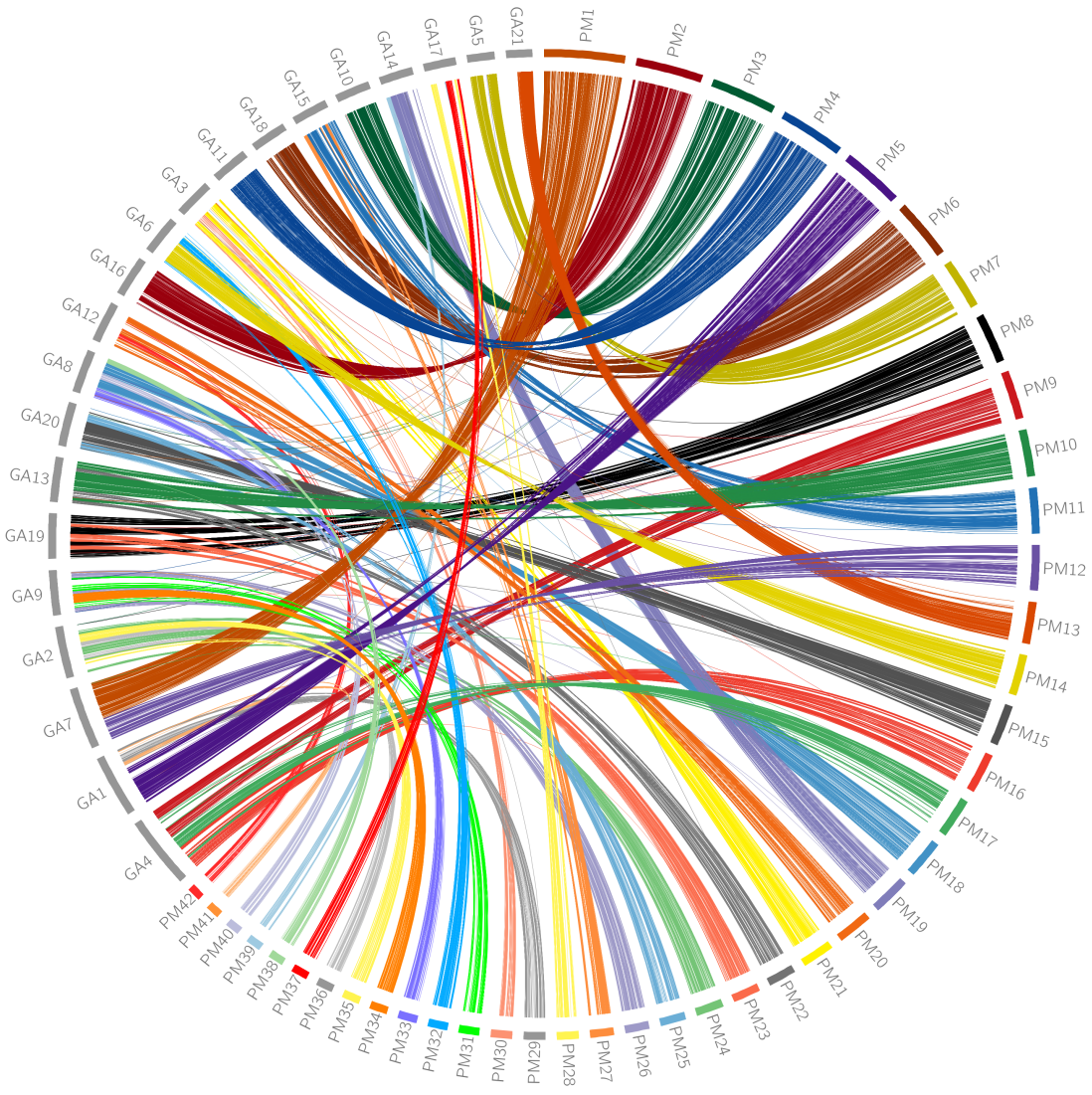


Figure 3.4: Circos plot which presents the syntenic locations of orthologous genes between the 42 longest contigs of *P. miles* (right) and the 21 chromosomes of *G. aculeatus* (left). Ribbons link orthologous genes between the two species, and colors represent the different contigs of *P. miles*.

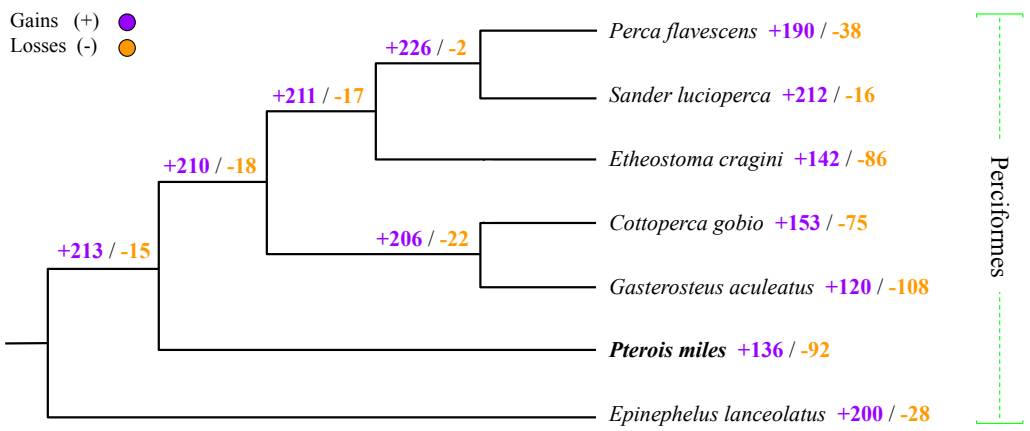


Figure 3.5: Gene family evolution analysis, including the number of gained (purple) and lost gene families (orange) for the Perciformes clade.

3.5.3 Gene duplication events

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 for 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.

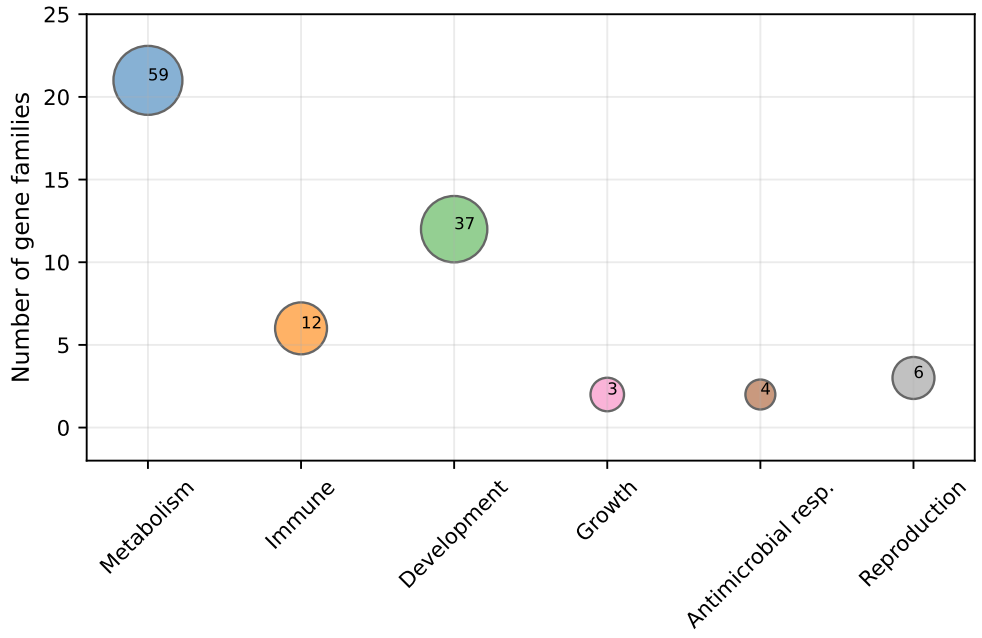
3.5.4 Gene ontology terms descriptive functional analysis

The descriptive analysis for rapidly expanding gene families of CAFE and those involved in duplication events in GeneRax are presented in Figure 3.6. GO terms were classified into eight categories, associated with metabolism, immune system, development, growth, antimicrobial response, toxin transport, reproduction and locomotion, and the number of gene families, involved genes and unique terms was calculated for both results form CAFE and GeneRax. The top terms were included in gene families associated with "metabolism", "development", "immune" and "growth", in both analyses.

3.5.5 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 (mean introns size 1: 819 and 2: 598 bp respectively), on the 7^{th} longest contig and in a distance between of 50.3 kb. The phylogeny of scorpaenid toxins showed the separation (with high support) between the two subunits, α and β (Figure 3.7), which form the functional heterodimer. Also, it confirmed that each triplet of genes in devil firefish's genome correspond to each of the subunits, three for α and three for β.

(a)



(b)

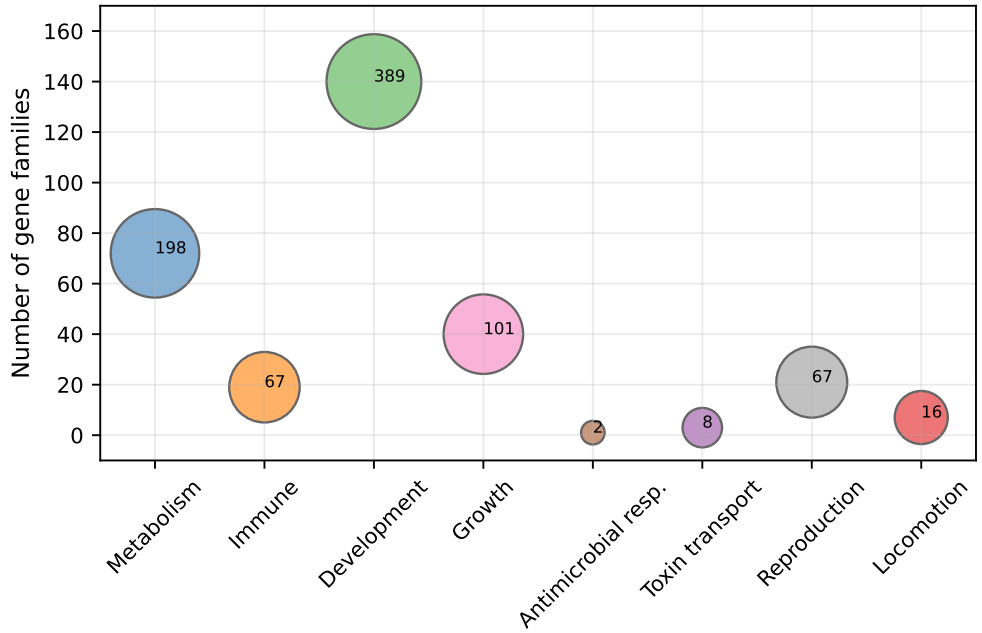


Figure 3.6: Number of gene families associated with specific biological processes for (a) rapidly expanding from CAFE and (b) with duplications from GeneRax. The size of each circle is the binary logarithm (log2) of the number of genes multiplied by the number of unique terms and then adding a scalar factor. The visualization of figures was performed with a custom python script "GO_plots.py".

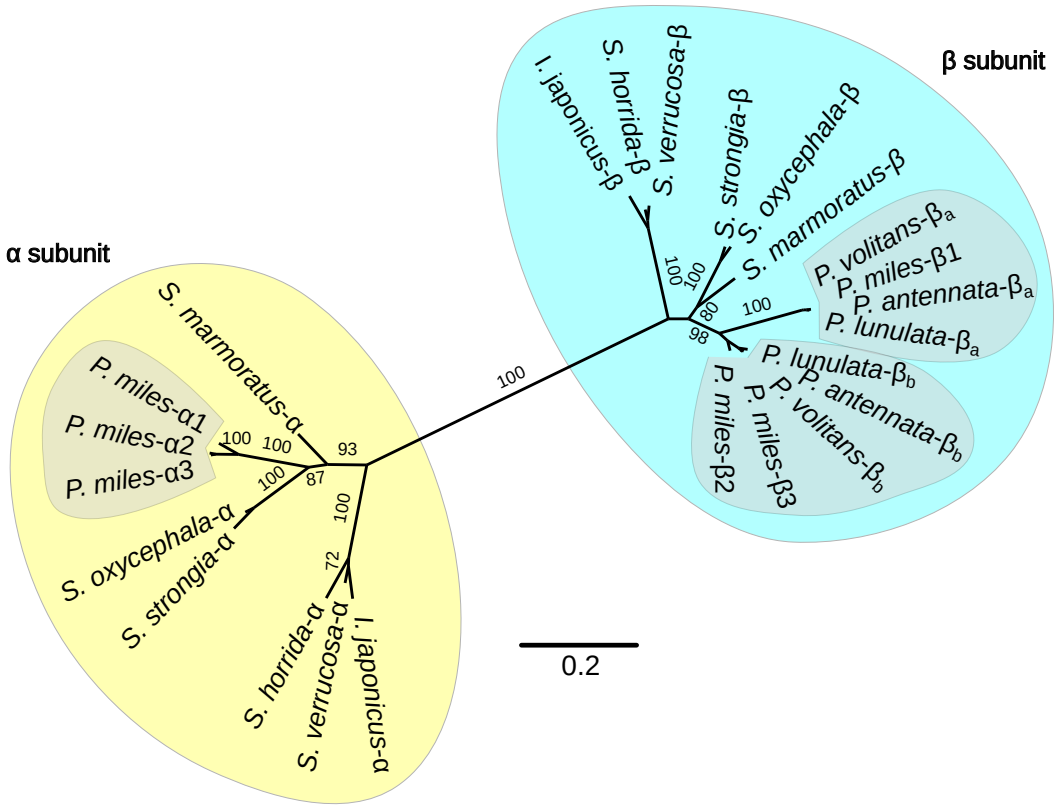


Figure 3.7: Maximum-likelihood unrooted phylogenetic tree of the two subunits of scorpaenid toxins. α subunits are presented inside the yellow and β in light blue bubble. For the phylogeny we used the JTT-DCMUT + I + G4 substitution model and conducted 100 bootstrap replicates.

4 DISCUSSION

Here, we presented and analyzed the first high quality genome assembly for the lessepsian migrant species *P. miles*, the first assembly for the whole family Scorpaenidae. We positioned the species in the teleost tree for the first time and studied its gene content. Our analysis revealed multiple interesting expanded gene families and especially a group of genes producing the toxin of this famous invasive species.

4.1 Genome size and assembly completeness

In this study, a lionfish genome assembly of high-quality and contiguity was constructed from genomic data derived from three MinION flow cells and an Illumina Hiseq4000 platform, with a total size of 902.3 Mb. To our knowledge, this is the first reference genome in Pteroinae and is the only representative in the family of Scorpaenidae, so far.

4.2 Repeat content, gene prediction & functional annotation

The representation of TEs in *P. miles* genome (46.5% of genome assembly) is notably higher than in other species inside Perciformes, 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). 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 superfamilies were recognised in the genome of *P. miles*, from a minimum of 77 Jockey elements to about 311,700 CACTA (Table 3.3), taking advantage both of the thorough classification resulting from the designed pipeline and the detailed annotation from "RM_parser.py" (Figure 2.1). 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, noteworthy would be the exploration of patterns in the

accumulation of TEs, in superfamily level, their roles and consequently their contribution to gene duplication events, and genome dynamics in general.

Table 4.1: 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	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)

4.3 Phylogenomic positioning of *P. miles*

Scorpaenidae (order: Perciformes) is a wide taxonomic marine family which includes by now 370 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 3.3). This effort could be an origin for further genomic and evolutionary studies inside this family.

4.4 Synteny analysis

One-to-one orthologous genes between *P. miles* and *G. aculeatus*, exhibited high conserved synteny (Figure 3.4), 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. Additionally, an interesting fact arose, which 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).

4.5 Gene families evolution and adaptation

Duplication events estimation and descriptive analysis unveiled the extended presence of gene families, being involved in major biological processes, such as metabolism, somatic growth, immunity and reproduction (Figure 3.6). 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. An interesting finding was a detected duplication in gene family of meprins (meprin-F in fish, Marín (2015)), proteins that are involved in toxins transport. Based on the results, it could 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.

4.6 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). 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 fish cytolysins and their evolution. For this reason, taking advantage of the first genome assembly of *P. miles*, we investigated the presence and identification of toxin genes on the genome, and reconstructed the phylogeny of lionfish toxins (Figure 3.7). 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). Our findings confirmed the important contribution of genomic data to the exploration of the evolutionary history of lionfish toxins in the first place, and of fish toxins on a larger scale, consequently.

5 CONCLUSION

Devil firefish is one of the most successful marine invasive species around the world, and a landmark for studies associated with unique phenotypes and fish toxins evolution, as well. In this study, we provide the first high-quality genome assembly of devil firefish, its 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 about the influence of transposable elements on genome evolution, the correlation between gene duplications and adaptation to new niche, lionfish rapid spread worldwide and its dominance, and fish toxins evolution as well as their potential pharmaceutical applications.

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://github.com/ckitsoulis/ELDAR`
- `https://github.com/genomenerds/SnakeCube`

REFERENCES

M. Albins and M. Hixon. Invasive Indo-Pacific lionfish *Pterois volitans* reduce recruitment of Atlantic coral-reef fishes. *Marine Ecology-progress Series - MAR ECOL-PROGR SER*, 367:233–238, 09 2008. doi: 10.3354/meps07620.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: https://doi.org/10.1016/S0022-2836(05)80360-2.

S. Andrews. FASTQC. A quality control tool for high throughput sequence data, 2010.

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

J. Ao, Y. Mu, L.-X. Xiang, D. Fan, M. Feng, S. Zhang, Q. Shi, L.-Y. Zhu, T. Li, Y. Ding, L. Nie, Q. Li, W. ren Dong, L. Jiang, B. Sun, X. Zhang, M. Li, H.-Q. Zhang, S. Xie, Y. Zhu, X. Jiang, X. Wang, P. Mu, W. Chen, Z. Yue, Z. Wang, J. Wang, J.-Z. Shao, and X. Chen. Genome Sequencing of the Perciform Fish *Larimichthys crocea* Provides Insights into Molecular and Genetic Mechanisms of Stress Adaptation. *PLOS Genetics*, 11(4):e1005118, apr 2015. doi: 10.1371/journal.pgen.1005118.

S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. ming Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. S. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. K. Edwards, N. Doggett, A. Zharkikh, S. V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner. Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. *Science*, 297(5585):1301–1310, aug 2002. doi: 10.1126/science.1072104.

K. Araki, J. Aokic, J. Kawase, K. Hamada, A. Ozaki, H. Fujimoto, I. Yamamoto, and H. Usuki. 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, 2018. doi: 10.1155/2018/7984292.

M. Arim, S. R. Abades, P. E. Neill, M. Lima, and P. A. Marquet. Spread dynamics of invasive species. *PNAS*, 2005.

E. Azzurro, B. Stancanelli, V. D. Martino, and M. Bariche. 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, 2017. doi: 10.3391/bir.2017.6.2.01.

M. Bariche, M. Torres, and E. Azzurro. The presence of the invasive lionfish *Pterois miles* in the Mediterranean Sea. *Mediterranean Marine Science*, 14:292–294, 2013. ISSN 1108393X. doi: 10.12681/mms.428.

M. Bariche, P. Kleitou, S. Kalogirou, and G. Bernardi. Genetics reveal the identity and origin of the lionfish invasion in the Mediterranean Sea. *Scientific Reports*, 7, 12 2017. doi: 10.1038/s41598-017-07326-1.

S. C. H. Barrett. Foundations of invasion genetics: the Baker and Stebbins legacy. *Mol. Ecol.*, 24(9):1927–1941, 2015. doi: 10.1111/mec.13014.

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

C. Bian, J. Li, X. Lin, X. Chen, Y. Yi, X. You, Y. Zhang, Y. Lv, and Q. Shi. Whole genome sequencing of the Blue Tilapia (*Oreochromis aureus*) provides a valuable genetic resource for biomedical research on Tilapias. *Marine Drugs*, 17(7):386, 2019. doi: 10.3390/md17070386.

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

I. Bista, S. A. McCarthy, J. Wood, Z. Ning, H. W. Detrich, Iii, T. Desvignes, J. Postlethwait, W. Chow, K. Howe, J. Torrance, M. Smith, K. Oliver, Vertebrate Genomes Project Consortium, E. A. Miska, and R. Durbin. The genome sequence of the channel bull blenny, *Cottoperca gobio* (Günther, 1861). *Wellcome Open Res.*, 5:148, jun 2020. doi: 10.12688/wellcomeopenres.16012.1.

I. Bista, J. M. D. Wood, T. Desvignes, S. A. McCarthy, M. Matschiner, Z. Ning, A. Tracey, J. Torrance, Y. Sims, W. Chow, M. Smith, K. Oliver, L. Haggerty, W. Salzburger, J. H. Postlethwait, K. Howe, M. S. Clark, W. H. Detrich, C.-H. C. Cheng, E. A. Miska, and R. Durbin. Genomics of cold adaptations in the Antarctic notothenioid fish radiation. *bioRxiv*, 2022. doi: 10.1101/2022.06.08.494096.

A. M. Blakeslee, T. Manousaki, K. Vasileiadou, and C. K. Tepolt. An evolutionary perspective on marine invasions. *Evolutionary Applications*, 2019. doi: 10.1111/eva.12906.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 04 2014. doi: 10.1093/bioinformatics/btu170.

G. Bourque, K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, Z. Izsvák, H. L. Levin, T. S. Macfarlan, D. L. Mager, and C. Feschotte. Ten things you should know about transposable elements. *Genome Biology*, 19(1), Nov. 2018. doi: 10.1186/s13059-018-1577-z.

I. Braasch, A. R. Gehrke, J. J. Smith, K. Kawasaki, T. Manousaki, J. Pasquier, A. Amores, T. Desvignes, P. Batzel, J. Catchen, A. M. Berlin, M. S. Campbell, D. Barrell, K. J. Martin, J. F. Mulley, V. Ravi, A. P. Lee, T. Nakamura, D. Chalopin, S. Fan, D. Wcisel, C. Caestro, J. Sydes, F. E. Beaudry, Y. Sun, J. Hertel, M. J. Beam, M. Fasold, M. Ishiyama, J. Johnson, S. Kehr, M. Lara, J. H. Letaw, G. W. Litman, R. T. Litman, M. Mikami, T. Ota, N. R. Saha, L. Williams, P. F. Stadler, H. Wang, J. S. Taylor, Q. Fontenot, A. Ferrara, S. M. Searle, B. Aken, M. Yandell, I. Schneider, J. A. Yoder, J. N. Volff, A. Meyer, C. T. Amemiya, B. Venkatesh, P. W. Holland, Y. Guiguen, J. Bobe, N. H. Shubin, F. D. Palma, J. Alföldi, K. Lindblad-Toh, and J. H. Postlethwait. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics*, 48:427–437, 3 2016. ISSN 15461718. doi: 10.1038/ng.3526.

D. Brawand, C. E. Wagner, Y. I. Li, M. Malinsky, I. Keller, S. Fan, O. Simakov, A. Y. Ng, Z. W. Lim, E. Bezault, J. Turner-Maier, J. Johnson, R. Alcazar, H. J. Noh, P. Russell, B. Aken, J. Alföldi, C. Amemiya, N. Azzouzi, J.-F. Baroiller, F. Barloy-Hubler, A. Berlin, R. Bloomquist, K. L. Carleton, M. A. Conte, H. D'Cotta, O. Eshel, L. Gaffney, F. Galibert, H. F. Gante, S. Gnerre, L. Greuter, R. Guyon, N. S. Haddad, W. Haerty, R. M. Harris, H. A. Hofmann, T. Hourlier, G. Hulata, D. B. Jaffe, M. Lara, A. P. Lee, I. MacCallum, S. Mwaiko, M. Nikaido, H. Nishihara, C. Ozouf-Costaz, D. J. Penman, D. Przybylski, M. Rakotomanga, S. C. P. Renn, F. J. Ribeiro, M. Ron, W. Salzburger, L. Sanchez-Pulido, M. E. Santos, S. Searle, T. Sharpe, R. Swofford, F. J. Tan, L. Williams, S. Young, S. Yin, N. Okada, T. D. Kocher, E. A. Miska, E. S. Lander, B. Venkatesh, R. D. Fernald, A. Meyer, C. P. Ponting, J. T. Streelman, K. Lindblad-Toh, O. Seehausen, and F. D. Palma. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518):375–381, 2014. doi: 10.1038/nature13726.

B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, Jan 2015. doi: 10.1038/nmeth.3176.

M. Cai, Y. Zou, S. Xiao, W. Li, Z. Han, F. Han, J. Xiao, F. Liu, and Z. Wang. Chromosome assembly of *Collichthys lucidus*, a fish of Sciaenidae with a multiple sex chromosome system. *Scientific Data*, 6, 12 2019. ISSN 20524463. doi: 10.1038/s41597-019-0139-x.

F. V. Campos, H. B. Fiorotti, J. B. Coitinho, and S. G. Figueiredo. Fish cytolysins in all their complexity. *Toxins*, 13, 12 2021. doi: 10.3390/toxins13120877.

C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12):5825–5829, 10 2021. doi: 10.1093/molbev/msab293.

S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25 (15):1972–1973, 06 2009. doi: 10.1093/bioinformatics/btp348.

Z. Chen, Y. Omori, S. Koren, T. Shirokiya, T. Kuroda, A. Miyamoto, H. Wada, A. Fujiyama, A. Toyoda, S. Zhang, T. G. Wolfsberg, K. Kawakami, A. M. Phillippy, J. C. Mullikin, and S. M. B. and. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Science Advances*, 5(6), 2019. doi: 10.1126/sciadv.aav0547.

P. Cheng, Y. Huang, Y. Lv, H. Du, Z. Ruan, C. Li, H. Ye, H. Zhang, J. Wu, C. Wang, R. Ruan, Y. Li, C. Bian, X. You, C. Shi, K. Han, J. Xu, Q. Shi, and Q. Wei. The American Paddlefish Genome Provides Novel Insights into Chromosomal Evolution and Bone Mineralization in Early Vertebrates. *Molecular Biology and Evolution*, 38 (4):1595–1607, 12 2020. doi: 10.1093/molbev/msaa326.

S. Chiesa, E. Azzurro, and G. Bernardi. The genetics and genomics of marine fish invasions: a global review. *Reviews in Fish Biology and Fisheries*, 29:837–859, 12 2019. doi: 10.1007/s11160-019-09586-8.

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

M. A. Conte and T. D. Kocher. An improved genome reference for the African cichlid, *Metriaclima zebra*. *BMC Genomics*, 16(1), sep 2015. doi: 10.1186/s12864-015-1930-5.

M. A. Conte, W. J. Gammerdinger, K. L. Bartie, D. J. Penman, and T. D. Kocher. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics*, 18(1), 2017. doi: 10.1186/s12864-017-3723-5.

I. M. Côté and N. S. Smith. The lionfish *Pterois sp.* invasion: Has the worst-case scenario come to pass? *J. Fish Biol.*, 92(3):660–689, mar 2018. doi: 10.1111/jfb.13544.

F. Crocetta, D. Agius, P. Balistreri, M. Bariche, Y. K. Bayhan, M. Çakir, S. Ciriaco, M. Corsini-Foka, A. Deidun, R. E. Zrelli, D. Ergüden, J. Evans, M. Ghelia, M. Giavasi, P. Kleitou, G. Kondylatos, L. Lipej, C. Mifsud, Y. Özvarol, A. Pagano, P. Portelli, D. Poursanidis, L. Rabaoui, P. J. Schembri, E. Taşkin, F. Tiralongo, and A. Zenetos. New Mediterranean biodiversity records (october 2015). *Mediterranean Marine Science*, 16:682–702, 2015. doi: 10.12681/mms.1477.

T. Dailianis, O. Akyol, N. Babali, M. Bariche, F. Crocetta, V. Gerovasileiou, R. Ghanem, M. Gökoglu, T. Hasiotis, A. Izquierdo-Muñoz, D. Julian, S. Katsanevakis, L. Lipej, E. Mancini, C. Mytilineou, K. O. B. Amor, A. Özgül, M. Ragkousis, E. Rubio-Portillo, G. Servello, M. Sini, C. Stamouli, A. Sterioti, S. Teker, F. Tiralongo, and D. Trkov. New Mediterranean Biodiversity Records (July 2016). *Mediterranean Marine Science*, 17:608–626, 2016. doi: 10.12681/mms.1734.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of SAMtools and

BCFtools. *GigaScience*, 10(2), 02 2021. doi: 10.1093/gigascience/giab008. URL `https://doi.org/10.1093/gigascience/giab008`. giab008.

T. Danis, V. Papadogiannis, A. Tsakogiannis, J. B. Kristoffersen, D. Golani, D. Tsaparis, A. Sterioti, P. Kasapidis, G. Kotoulas, A. Magoulas, C. S. Tsigenopoulos, and T. Manousaki. Genome Analysis of *Lagocephalus sceleratus*: Unraveling the Genomic Landscape of a Successful Invader. *Frontiers in Genetics*, 12, 2021. doi: 10.3389/fgene.2021.790850.

D. Darriba, D. Posada, A. M. Kozlov, A. Stamatakis, B. Morel, and T. Flouri. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, 37(1):291–294, 08 2019. doi: 10.1093/molbev/msz189.

W. De Coster, S. D'Hert, D. T. Schultz, M. Cruts, and C. V. Broeckhoven. Nanopack: Visualizing and processing long-read sequencing data. *Bioinformatics*, 34:2666–2669, 8 2018. doi: 10.1093/bioinformatics/bty149.

J. H. Diaz. Marine Scorpaenidae envenomation in travelers: Epidemiology, management, and prevention. *J. Travel Med.*, 22(4):251–258, 2015. doi: 10.1111/jtm.12206.

A. C. Dimitriou, N. Chartosia, J. M. Hall-Spencer, P. Kleitou, C. Jimenez, C. Antoniou, L. Hadjioannou, D. Kletou, and S. Sfenthourakis. Genetic data suggest multiple introductions of the lionfish (*Pterois miles*) into the Mediterranean Sea. *Diversity*, 11, 9 2019. doi: 10.3390/d11090149.

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

L. Dray, M. Neuhof, A. Diamant, and D. Huchon. The complete mitochondrial genome of the devil firefish *Pterois miles* (Bennett, 1828) (Scorpaenidae). *Mitochondrial DNA*, 27:783–784, 1 2016. doi: 10.3109/19401736.2014.945565.

D. Drongitis, F. Aniello, L. Fucci, and A. Donizetti. Roles of Transposable Elements in the Different Layers of Gene Expression Regulation. *International Journal of Molecular Sciences*, 20(22):5755, Nov. 2019. doi: 10.3390/ijms20225755.

D. M. Emms and S. Kelly. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238, 2019. doi: https://doi.org/10.1186/s13059-019-1832-y.

G. Fan, J. Chan, K. Ma, B. Yang, H. Zhang, X. Yang, C. Shi, H. Chun-Hin Law, Z. Ren, Q. Xu, Q. Liu, J. Wang, W. Chen, L. Shao, D. Gonçalves, A. Ramos, S. D. Cardoso, M. Guo, J. Cai, X. Xu, J. Wang, H. Yang, X. Liu, and Y. Wang. Chromosome-level reference genome of the Siamese fighting fish *Betta splendens*, a model species for the study of aggression. *GigaScience*, 7(11), 07 2018. doi: 10.1093/gigascience/giy087.

R. Feron, M. Zahm, C. Cabau, C. Klopp, C. Roques, O. Bouchez, C. Eché, S. Valière, C. Donnadieu, P. Haffray, A. Bestin, R. Morvezen, H. Acloque, P. T. Euclide, M. Wen,

E. Jouano, M. Schartl, J. H. Postlethwait, C. Schraidt, M. R. Christie, W. A. Larson, A. Herpin, and Y. Guiguen. 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 (2):531–543, jan 2020. doi: 10.1111/1755-0998.13133.

J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117:9451–9457, 4 2020. doi: 10.1073/PNAS.1921046117.

R. Fueyo, J. Judd, C. Feschotte, and J. Wysocka. Roles of transposable elements in the regulation of mammalian transcription. *Nature Reviews Molecular Cell Biology*, 23(7): 481–497, Feb. 2022. doi: 10.1038/s41580-022-00457-y.

Z. Gao, X. You, X. Zhang, J. Chen, T. Xu, Y. Huang, X. Lin, J. Xu, C. Bian, and Q. Shi. A chromosome-level genome assembly of the striped catfish (*Pangasianodon hypophthalmus*). *Genomics*, 113(5):3349–3356, 2021. doi: https://doi.org/10.1016/j.ygeno.2021.07.026.

E. Gasteiger. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, 31(13):3784–3788, 2003. doi: 10.1093/nar/gkg563.

F. J. Ghadessy, D. Chen, R. M. Kini, M. Chung, K. Jeyaseelan, H. E. Khoo, and R. Yuen. Stonustoxin Is a Novel Lethal Factor from Stonefish (*Synanceja horrida*) Venom. *Journal of Biological Chemistry*, 271(41):25575–25581, 1996. doi: 10.1074/jbc.271.41.25575.

D. Golani and O. Sonin. New records of the red sea fishes, *Pterois miles* (scorpaenidae) and *Pteragogus pelycus* (labridae) from the Eastern Mediterranean Sea. *Japanese Journal of Ichthyology*, 39, 1992.

I. Guerrero-Cózar, J. Gomez-Garrido, C. Berbel, J. F. Martinez-Blanch, T. Alioto, M. G. Claros, P.-A. Gagnaire, and M. Manchado. Chromosome anchoring in Senegalese sole (*Solea senegalensis*) reveals sex-associated markers and genome rearrangements in flatfish. *Scientific Reports*, 11(1), jun 2021. doi: 10.1038/s41598-021-92601-5.

B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith, Jr, L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg, and O. White. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, 31(19):5654–5666, oct 2003. doi: 10.1093/nar/gkg770.

K. Howe, M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J. C. Barrett, R. Koch, G.-J. Rauch, S. White, W. Chow, B. Kilian, L. T. Quintais, J. A. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.-H. Vogel, T. Eyre, R. Banerjee, J. Chi, B. Fu, E. Langley, S. F. Maguire, G. Laird, D. Lloyd, E. Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey, A. Hunt, J. Burton, S. Sims,

K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Elliott, G. Threadgold, G. Harden, D. Ware, S. Begum, B. Mortimore, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N. Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N. Barker, C. Lloyd, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D. Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D. Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S. Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders, J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington, N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. D. Cooper, A. Tromans, D. Grafham, C. Skuce, R. Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird, S. Palmer, I. Gehring, A. Berger, C. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A. Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, C. Lanz, G. Raddatz, K. Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, S. C. Schuster, N. P. Carter, J. Harrow, Z. Ning, J. Herrero, S. M. J. Searle, A. Enright, R. Geisler, R. H. A. Plasterk, C. Lee, M. Westerfield, P. J. de Jong, L. I. Zon, J. H. Postlethwait, C. Nüsslein-Volhard, T. J. P. Hubbard, H. R. Crollius, J. Rogers, and D. L. Stemple. Correction: Corrigendum: The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 505(7482):248–248, dec 2013. doi: 10.1038/nature12813.

J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, and P. Bork. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, 11 2019. doi: 10.1093/nar/gky1085.

O. Jaillon, J.-M. Aury, F. Brunet, J.-L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet, S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biémont, Z. Skalli, L. Cattolico, J. Poulain, V. de Berardinis, C. Cruaud, S. Duprat, P. Brottier, J.-P. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K. J. McKernan, P. McEwan, S. Bosak, M. Kellis, J.-N. Volff, R. Guigó, M. C. Zody, J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quétier, W. Saurin, C. Scarpelli, P. Wincker, E. S. Lander, J. Weissenbach, and H. R. Crollius. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, oct 2004. doi: 10.1038/nature03025.

A. J. Jasonowicz, A. Simeon, M. Zahm, C. Cabau, C. Klopp, C. Roques, C. Iampietro, J. Lluch, C. Donnadieu, H. Parrinello, D. P. Drinan, L. Hauser, Y. Guiguen, and J. V. Planas. Generation of a chromosome-level genome assembly for Pacific halibut (*Hippoglossus stenolepis*) and characterization of its sex-determining genomic region. *Molecular Ecology Resources*, jun 2022. doi: 10.1111/1755-0998.13641.

W. Kai, K. Kikuchi, S. Tohari, A. K. Chew, A. Tay, A. Fujiwara, S. Hosoya, H. Suetake, K. Naruse, S. Brenner, Y. Suzuki, and B. Venkatesh. Integration of the genetic map and genome assembly of fugu facilitates insights into distinct features of genome evolution in teleosts and mammals. *Genome Biol. Evol.*, 3:424–442, 2011.

M. Kasahara, K. Naruse, S. Sasaki, Y. Nakatani, W. Qu, B. Ahsan, T. Yamada, Y. Nagayasu, K. Doi, Y. Kasai, T. Jindo, D. Kobayashi, A. Shimada, A. Toyoda, Y. Kuroki, A. Fujiyama, T. Sasaki, A. Shimizu, S. Asakawa, N. Shimizu, S. ichi Hashimoto, J. Yang, Y. Lee, K. Matsushima, S. Sugano, M. Sakaizumi, T. Narita, K. Ohishi, S. Haga, F. Ohta, H. Nomoto, K. Nogata, T. Morishita, T. Endo, T. Shin-I, H. Takeda, S. Morishita, and Y. Kohara. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145):714–719, jun 2007. doi: 10.1038/nature05846.

K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30 (4):772–780, 01 2013. doi: 10.1093/molbev/mst010.

S. Katsanevakis, D. Poursanidis, R. Hoffman, J. Rizgalla, S. B. S. Rothman, Y. Levitt-Barmats, L. Hadjioannou, D. Trkov, J. M. Garmendia, M. Rizzo, A. G. Bartolo, M. Bariche, F. Tomas, P. Kleitou, P. J. Schembri, D. Kletou, F. Tiralongo, C. Pergent, G. Pergent, E. Azzurro, M. Bilecenoglu, A. Lodola, E. Ballesteros, V. Gerovasileiou, M. Verlaque, A. Occhipinti-Ambrogi, E. Kytinou, T. Dailianis, J. Ferrario, F. Crocetta, C. Jimenez, J. Evans, M. Ragkousis, L. Lipej, J. A. Borg, C. Dimitriadis, G. Chatzigeorgiou, P. G. Albano, S. Kalogirou, H. Bazairi, F. Espinosa, J. B. Souissi, K. Tsiamis, F. Badalamenti, J. Langeneck, P. Noel, A. Deidun, A. Marchini, G. Skouradakis, L. Royo, M. Sini, C. N. Bianchi, Y. R. Sghaier, R. Ghanem, N. Doumpas, J. Zaouali, K. Tsirintanis, O. Papadakis, C. Morri, M. E. Çinar, J. Terrados, G. Insacco, B. Zava, E. Soufi-Kechaou, L. Piazzi, K. O. B. Amor, E. Andriotis, M. C. Gambi, M. M. B. Amor, J. Garrabou, C. Linares, A. Fortič, M. Digenis, E. Cebrian, M. Fourt, M. Zotou, L. Castriota, V. D. Martino, A. Rosso, C. Pipitone, M. Falautano, M. García, R. Zakhama-Sraieb, F. Khamassi, A. M. Mannino, M. H. Ktari, I. Kosma, M. Rifi, P. K. Karachle, S. Yapıcı, A. R. Bos, P. Balistreri, A. A. Esplá, J. Tempesti, O. Inglese, I. Giovos, D. Damalas, S. Benhissoune, M. F. Huseyinoglu, W. Rjiba-Bahri, J. Santamaría, M. Orlando-Bonaca, A. Izquierdo, C. Stamouli, M. Montefalcone, H. Cerim, R. Golo, S. Tsioli, S. Orfanidis, N. Michailidis, M. Gaglioti, E. Taşkın, E. Mancuso, A. Žunec, I. Cvitković, H. Filiz, R. Sanfilippo, A. Siapatis, B. Mavrič, S. Karaa, A. Türker, F. Monniot, J. Verdura, N. E. Ouamari, M. Selfati, and A. Zenetos. Unpublished mediterranean records of marine alien and cryptogenic species. *BioInvasions Records*, 9:165–182, 2020. doi: 10.3391/bir.2020.9.2.01.

J. L. Kelley, M.-C. Yee, A. P. Brown, R. R. Richardson, A. Tatarenkov, C. C. Lee, T. T. Harkins, C. D. Bustamante, and R. L. Earley. 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(7):2145–2154, jun 2016. doi: 10.1093/gbe/evw145.

A. Kiriake and K. Shiomi. Some properties and cDNA cloning of proteinaceous toxins

from two species of lionfish (*Pterois antennata* and *Pterois volitans*). *Toxicon*, 58(6-7): 494–501, 2011. doi: 10.1016/j.toxicon.2011.08.010.

A. Kiriake, Y. Suzuki, Y. Nagashima, and K. Shiomi. 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, 2013. doi: 10.1016/j.toxicon.2013.04.021.

C. V. Kitsoulis, V. Papadogiannis, J. B. Kristoffersen, E. Kaitetzidou, A. Sterioti, C. S. Tsigenopoulos, and T. Manousaki. A high-quality reference genome assembly for the devil firefish, *Pterois miles*, Mar 2022. URL `https://doi.org/10.5281/zenodo.6380502`.

P. Kleitou, D. K. Moutopoulos, I. Giovos, D. Kletou, I. Savva, L. L. Cai, J. M. Hall-Spencer, A. Charitou, M. Elia, G. Katselis, and S. Rees. Conflicting interests and growing importance of non-indigenous species in commercial and recreational fisheries of the Mediterranean Sea. *Fish. Manag. Ecol.*, 29(2):169–182", 2022. doi: 10.1111/fme.12531.

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

M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner. Assembly of long error-prone reads using repeat graphs. *Nature biotechnology*, 37(5):540–546, 2019. doi: 10.1038/s41587-019-0072-8.

S. Kongsstovu, S. O. Mikalsen, E. Homrum, J. A. Jacobsen, P. Flicek, and H. A. Dahl. Using long and linked reads to improve an Atlantic herring (*Clupea harengus*) genome assembly. *Scientific Reports*, 9, 12 2019. doi: 10.1038/s41598-019-54151-9.

J. P. Koskinen and L. Holm. SANS: high-throughput retrieval of protein sequences allowing 50% mismatches. *Bioinformatics*, 28(18):i438–i443, 09 2012. doi: 10.1093/bioinformatics/bts417.

M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res.*, 19(9):1639–1645, 2009.

H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–310, 2018. doi: https://doi.org/10.1093/bioinformatics/bty191.

S. Lien, B. F. Koop, S. R. Sandve, J. R. Miller, M. P. Kent, T. Nome, T. R. Hvidsten, J. S. Leong, D. R. Minkley, A. Zimin, F. Grammes, H. Grove, A. Gjuvsland, B. Walenz, R. A. Hermansen, K. von Schalburg, E. B. Rondeau, A. D. Genova, J. K. A. Samy, J. O. Vik, M. D. Vigeland, L. Caler, U. Grimholt, S. Jentoft, D. I. Våge, P. de Jong, T. Moen, M. Baranski, Y. Palti, D. R. Smith, J. A. Yorke, A. J. Nederbragt, A. Tooming-Klunderud, K. S. Jakobsen, X. Jiang, D. Fan, Y. Hu, D. A. Liberles, R. Vidal, P. Iturra, S. J. M. Jones, I. Jonassen, A. Maass, S. W. Omholt, and W. S. Davidson. The

Atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602): 200–205, 2016. doi: 10.1038/nature17164.

Q. Lin, S. Fan, Y. Zhang, M. Xu, H. Zhang, Y. Yang, A. P. Lee, J. M. Woltering, V. Ravi, H. M. Gunter, W. Luo, Z. Gao, Z. W. Lim, G. Qin, R. F. Schneider, X. Wang, P. Xiong, G. Li, K. Wang, J. Min, C. Zhang, Y. Qiu, J. Bai, W. He, C. Bian, X. Zhang, D. Shan, H. Qu, Y. Sun, Q. Gao, L. Huang, Q. Shi, A. Meyer, and B. Venkatesh. The seahorse genome and the evolution of its specialized morphology. *Nature*, 540(7633):395–399, 2016. doi: 10.1038/nature20595.

D. Liu, X. Wang, H. Guo, X. Zhang, M. Zhang, and W. Tang. 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, 10 2021. doi: 10.1111/1755-0998.13429.

T. J. Lyons, Q. M. Tuckett, and J. E. Hill. Data quality and quantity for invasive species: A case study of the lionfishes. *Fish and Fisheries*, 20:748–759, 7 2019. doi: 10.1111/faf.12374.

S. A. A. A. Mabruk and J. Rizgalla. First record of lionfish (scorpaenidae: *Pterois*) from Libyan waters. *J. Black Sea/Mediterranean Environment*, 25:108–114, 2019.

W. Makalowski, V. Gotea, A. Pande, and I. Makalowska. *Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics*, pages 177–207. Springer New York, 2019. ISBN 978-1-4939-9074-0. doi: 10.1007/978-1-4939-9074-0_6. URL `https://doi.org/10.1007/978-1-4939-9074-0_6`.

M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. 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(10):4647–4654, 07 2021. doi: 10.1093/molbev/msab199.

I. Marín. Origin and diversification of meprin proteases. *PLoS ONE*, 10, 8 2015. ISSN 19326203. doi: 10.1371/journal.pone.0135924.

S. McGaugh, J. Gross, B. Aken, M. Blin, R. Borowsky, D. Chalopin, H. Hinaux, W. Jeffery, A. Keene, L. Ma, P. Minx, D. Murphy, K. O'Quin, S. Rétaux, N. Rohner, S. Searle, B. Stahl, C. Tabin, J. Volff, M. Yoshizawa, and W. W. The cavefish genome reveals candidate genes for eye loss. *Nature Communications*, 5, 10 2014. doi: 10.1038/ncomms6307.

F. K. Mendes, D. Vanderpool, B. Fulton, and M. W. Hahn. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*, 36(22-23):5516–5518, 12 2020. doi: 10.1093/bioinformatics/btaa1022.

B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, 02 2020. doi: 10.1093/molbev/msaa015.

B. Morel, A. M. Kozlov, A. Stamatakis, and G. J. Szöllősi. 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(9):2763–2774, 06 2020. doi: 10.1093/molbev/msaa141.

S. Nath, D. E. Shaw, and M. A. White. Improved contiguity of the threespine stickleback genome using long-read sequencing. *G3 Genes—Genomes—Genetics*, 11(2), 01 2021. doi: 10.1093/g3journal/jkab007.

J. A. Nguinkal, R. M. Brunner, M. Verleih, A. Rebl, L. de los Ríos-Pérez, N. Schäfer, F. Hadlich, M. Stüeken, D. Wittenburg, and T. Goldammer. The first highly contiguous genome assembly of pikeperch (*Sander lucioperca*), an emerging aquaculture species in Europe. *Genes*, 10, 9 2019. doi: 10.3390/genes10090708.

S. Ou, W. Su, Y. Liao, K. Chougule, J. R. Agda, A. J. Hellinga, C. S. B. Lugo, T. A. Elliott, D. Ware, T. Peterson, N. Jiang, C. N. Hirsch, and M. B. Hufford. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, 20, 12 2019. doi: 10.1186/s13059-019-1905-y.

H. Pan, H. Yu, V. Ravi, C. Li, A. P. Lee, M. M. Lian, B.-H. Tay, S. Brenner, J. Wang, H. Yang, G. Zhang, and B. Venkatesh. The genome of the largest bony fish, ocean sunfish (*Mola mola*), provides insights into its fast growth rate. *GigaScience*, 5(1), sep 2016. doi: 10.1186/s13742-016-0144-3.

V. Papadogiannis, A. Tsakogiannis, J. B. Kristoffersen, O. Nousias, C. Batargias, D. Chatziplis, C. S. Tsigenopoulos, and T. Manousaki. Chromosome assembly for the meagre, *Argyrosomus regius*, reveals species adaptations and sciaenid sex-related locus evolution. *in prep.*, 2022.

E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019.

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

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 01 2010. doi: 10.1093/bioinformatics/btq033.

B. N. Reid, R. L. Moran, C. J. Kopack, and S. W. Fitzpatrick. 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(2):404–420, nov 2020. doi: 10.1111/1755-0998.13275.

N. M. Reid, C. E. Jackson, D. Gilbert, P. Minx, M. J. Montague, T. H. Hampton, L. W. Helfrich, B. L. King, D. E. Nacci, N. Aluru, S. I. Karchner, J. K. Colbourne, M. E. Hahn, J. R. Shaw, M. F. Oleksiak, D. L. Crawford, W. C. Warren, and A. Whitehead.

The Landscape of Extreme Genomic Variation in the Highly Adaptable Atlantic Killifish. *Genome Biology and Evolution*, 9(3):659–676, 03 2017. doi: 10.1093/gbe/evx023.

E. B. Rondeau, D. R. Minkley, J. S. Leong, A. M. Messmer, J. R. Jantzen, K. R. von Schalburg, C. Lemon, N. H. Bird, and B. F. Koop. 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(7):e102089, 2014. doi: 10.1371/journal.pone.0102089.

T. Ryu, M. Herrera, B. Moore, M. Izumiyama, E. Kawai, V. Laudet, and T. Ravasi. A chromosome-scale genome assembly of the false clownfish, *Amphiprion ocellaris*. *G3 Genes—Genomes—Genetics*, 12(5), 03 2022. doi: 10.1093/g3journal/jkac074.

M. Schartl, R. B. Walter, Y. Shen, T. Garcia, J. Catchen, A. Amores, I. Braasch, D. Chalopin, J.-N. Volff, K.-P. Lesch, A. Bisazza, P. Minx, L. Hillier, R. K. Wilson, S. Fuerstenberg, J. Boore, S. Searle, J. H. Postlethwait, and W. C. Warren. The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. *Nat. Genet.*, 45(5):567–572, 2013.

E. T. Schultz. *Pterois volitans* and *Pterois miles*: Two valid species. *Copeia 1986*, pages 686–690, 8 1986. doi: 10.2307/1444950.

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

F. Shao, A. Ludwig, Y. Mao, N. Liu, and Z. Peng. Chromosome-level genome assembly of the female western mosquitofish (*Gambusia affinis*). *GigaScience*, 9(8), 08 2020. doi: 10.1093/gigascience/giaa092.

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

Y. Shen, D. Chalopin, T. Garcia, M. Boswell, W. Boswell, S. A. Shiryev, R. Agarwala, J.-N. Volff, J. H. Postlethwait, M. Schartl, P. Minx, W. C. Warren, and R. B. Walter. *X. couchianus* and *X. hellerii* genome models provide genomic variation insight among Xiphophorus species. *BMC Genomics*, 17(1), jan 2016. doi: 10.1186/s12864-015-2361-z.

W. L. Smith, E. Everman, and C. Richardson. Phylogeny and Taxonomy of Flatheads, Scorpionfishes, Sea Robins, and Stonefishes (Percomorpha: Scorpaeniformes) and the Evolution of the Lachrymal Saber. *Copeia*, 106(1):94 – 119, 2018. doi: 10.1643/CG-17-669.

C. G. Sotero-Caio, I. Platt, Roy N., A. Suh, and D. A. Ray. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biology and Evolution*, 9(1): 161–177, 02 2017. doi: 10.1093/gbe/evw264.

M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5): 637–644, 01 2008. doi: 10.1093/bioinformatics/btn013.

L. Sun, T. Gao, F. Wang, Z. Qin, L. Yan, W. Tao, M. Li, C. Jin, L. Ma, T. D. Kocher, and D. Wang. Chromosome-level genome assembly of a cyprinid fish *Onychostoma macrolepis* by integration of nanopore sequencing, Bionano and Hi-C technology. *Molecular Ecology Resources*, 20(5):1361–1371, jul 2020. doi: 10.1111/1755-0998.13190.

M. Tarailo-Graovac and N. Chen. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, 25(1):4.10.1–4.10.14, 2009. doi: https://doi.org/10.1002/0471250953.bi0410s25.

A. W. Thompson, H. Wojtas, M. Davoll, and I. Braasch. Genome of the Rio Pearlfish (*Nematolebias whitei*), a bi-annual killifish model for Eco-Evo-Devo in extreme environments. *G3 Genes|Genomes|Genetics*, 12(4), feb 2022. doi: 10.1093/g3journal/jkac045.

H.-F. Tian, Q.-M. Hu, and Z. Li. 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(1), 12 2020. doi: 10.1093/g3journal/jkaa032.

M. Tine, H. Kuhl, P.-A. Gagnaire, B. Louro, E. Desmarais, R. S. Martins, J. Hecht, F. Knaust, K. Belkhir, S. Klages, R. Dieterich, K. Stueber, F. Piferrer, B. Guinand, N. Bierne, F. A. M. Volckaert, L. Bargelloni, D. M. Power, F. Bonhomme, A. V. M. Canario, and R. Reinhardt. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, 5(1), dec 2014. doi: 10.1038/ncomms6770.

P. Törönen and L. Holm. PANNZER—A practical tool for protein function prediction. *Protein Science*, 31(1):118–128, 2022. doi: https://doi.org/10.1002/pro.4193.

A. Ueda, M. Suzuki, T. Honma, H. Nagai, Y. Nagashima, and K. Shiomi. 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(11):1713–1722, 2006. doi: 10.1016/j.bbagen.2006.08.017.

R. Vaser, I. Sović, N. Nagarajan, and M. Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737–746, 2017. doi: 10.1101/gr.214270.116.

C. Vavasis, G. Simotas, E. Spinos, E. Konstantinidis, S. Minoudi, A. Triantafyllidis, and C. Perdikaris. Occurrence of *Pterois miles* in the Island of Kefalonia (Greece): the Northernmost Dispersal Record in the Mediterranean Sea. *Thalassas*, 36:171–175, 4 2020. doi: 10.1007/s41208-019-00175-x.

L. Venturini, S. Caim, G. G. Kaithakottil, D. L. Mapleson, and D. Swarbreck. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*, 7(8), 2018. doi: 10.1093/gigascience/giy093.

S. Vij, H. Kuhl, I. S. Kuznetsova, A. Komissarov, A. A. Yurchenko, P. V. Heusden, S. Singh, N. M. Thevasagayam, S. R. S. Prakki, K. Purushothaman, J. M. Saju, J. Jiang, S. K. Mbandi, M. Jonas, A. H. Y. Tong, S. Mwangi, D. Lau, S. Y. Ngoh, W. C.

Liew, X. Shen, L. S. Hon, J. P. Drake, M. Boitano, R. Hall, C.-S. Chin, R. Lachumanan, J. Korlach, V. Trifonov, M. Kabilov, A. Tupikin, D. Green, S. Moxon, T. Garvin, F. J. Sedlazeck, G. W. Vurture, G. Gopalapillai, V. K. Katneni, T. H. Noble, V. Scaria, S. Sivasubbu, D. R. Jerry, S. J. O'Brien, M. C. Schatz, T. Dalmay, S. W. Turner, S. Lok, A. Christoffels, and L. Orbán. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLOS Genetics*, 12(4):e1005954, apr 2016. doi: 10.1371/journal.pgen.1005954.

B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 9, 11 2014. doi: 10.1371/journal.pone.0112963.

D. Wang, X. Chen, X. Zhang, J. Li, Y. Yi, C. Bian, Q. Shi, H. Lin, S. Li, Y. Zhang, and X. You. Whole Genome Sequencing of the Giant Grouper (*Epinephelus lanceolatus*) and High-Throughput Screening of Putative Antimicrobial Peptide Genes. *Marine Drugs*, 17, 8 2019. ISSN 16603397. doi: 10.3390/md17090503.

H. Wang, B. Su, I. A. E. Butts, R. A. Dunham, and X. Wang. Chromosome-level assembly and annotation of the blue catfish *Ictalurus furcatus*, an aquaculture species for hybrid catfish reproduction, epigenetics, and heterosis studies. *GigaScience*, 11, 2022. doi: 10.1093/gigascience/giac070.

W. C. Warren, R. García-Pérez, S. Xu, K. P. Lampert, D. Chalopin, M. Stöck, L. Loewe, Y. Lu, L. Kuderna, P. Minx, M. J. Montague, C. Tomlinson, L. W. Hillier, D. N. Murphy, J. Wang, Z. Wang, C. M. Garcia, G. C. W. Thomas, J.-N. Volff, F. Farias, B. Aken, R. B. Walter, K. D. Pruitt, T. Marques-Bonet, M. W. Hahn, S. Kneitz, M. Lynch, and M. Schartl. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nature Ecology & Evolution*, 2(4):669–679, feb 2018. doi: 10.1038/s41559-018-0473-y.

A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 01 2009. doi: 10.1093/bioinformatics/btp033.

B. Wu, C. Feng, C. Zhu, W. Xu, Y. Yuan, M. Hu, K. Yuan, Y. Li, Y. Ren, Y. Zhou, H. Jiang, Q. Qiu, W. Wang, S. He, and K. Wang. The Genomes of Two Billfishes Provide Insights into the Evolution of Endothermy in Teleosts. *Molecular Biology and Evolution*, 38(6):2413–2427, 02 2021. doi: 10.1093/molbev/msab035.

T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 02 2005. doi: 10.1093/bioinformatics/bti310.

B. Xie, H. Yu, H. Kerkkamp, M. Wang, M. Richardson, and Q. Shi. Comparative transcriptome analyses of venom glands from three scorpionfishes. *Genomics*, 111(3): 231–241, 2019. doi: 10.1016/j.ygeno.2018.11.012.

X. Xu, C. Shao, H. Xu, Q. Zhou, F. You, N. Wang, W. Li, M. Li, and S. Chen. Draft genomes of female and male turbot *Scophthalmus maximus*. *Scientific Data*, 7(1), mar 2020. doi: 10.1038/s41597-020-0426-6.

H. Yan, A. Bombarely, and S. Li. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics*, 36(15): 4269–4275, 05 2020. doi: 10.1093/bioinformatics/btaa519.

X. Yang, H. Liu, Z. Ma, Y. Zou, M. Zou, Y. Mao, X. Li, H. Wang, T. Chen, W. Wang, and R. Yang. Chromosome-level genome assembly of *Triplophysa tibetana*, a fish adapted to the harsh high-altitude environment of the Tibetan Plateau. *Molecular Ecology Resources*, 19(4):1027–1036, may 2019. doi: 10.1111/1755-0998.13021.

G. Yu. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics*, 69(1), mar 2020. doi: 10.1002/cpbi.96.

L. Yunyun, L. Yanping, L. Yi, W. Zhengyong, Y. Yexin, Q. Chuanjie, S. Qiong, and M. Xidong. An Updated Genome Assembly Improves Understanding of the Transcriptional Regulation of Coloration in Midas Cichlid. *Frontiers in Marine Science*, 9, 2022. doi: 10.3389/fmars.2022.950573.

H. Zafeiropoulos, A. Gioti, S. Ninidakis, A. Potirakis, S. Paragkamian, N. Angelova, A. Antoniou, T. Danis, E. Kaitetzidou, P. Kasapidis, J. B. Kristoffersen, V. Papadogiannis, C. Pavloudi, Q. V. Ha, J. Lagnel, N. Pattakos, G. Perantinos, D. Sidirokastritis, P. Vavilis, G. Kotoulas, T. Manousaki, E. Sarropoulou, C. S. Tsigenopoulos, C. Arvanitidis, A. Magoulas, and E. Pafilis. 0s and 1s in marine molecular research: a regional HPC perspective. *GigaScience*, 10(8), 08 2021. doi: 10.1093/gigascience/giab053.

N. Zhao, H. Guo, L. Jia, B. Guo, D. Zheng, S. Liu, and B. Zhang. Genome assembly and annotation at the chromosomal level of first Pleuronectidae: *Verasper variegatus* provides a basis for phylogenetic study of Pleuronectiformes. *Genomics*, 113(2):717–726, mar 2021. doi: 10.1016/j.ygeno.2021.01.024.

S. Zheng, F. Shao, W. Tao, Z. Liu, J. Long, X. Wang, S. Zhang, Q. Zhao, K. L. Carleton, T. D. Kocher, L. Jin, Z. Wang, Z. Peng, D. Wang, and Y. Zhang. Chromosome-level assembly of southern catfish (*Silurus meridionalis*) provides insights into visual adaptation to nocturnal and benthic lifestyles. *Molecular Ecology Resources*, 21(5): 1575–1592, may 2021. doi: 10.1111/1755-0998.13338.

SUPPLEMENTARY

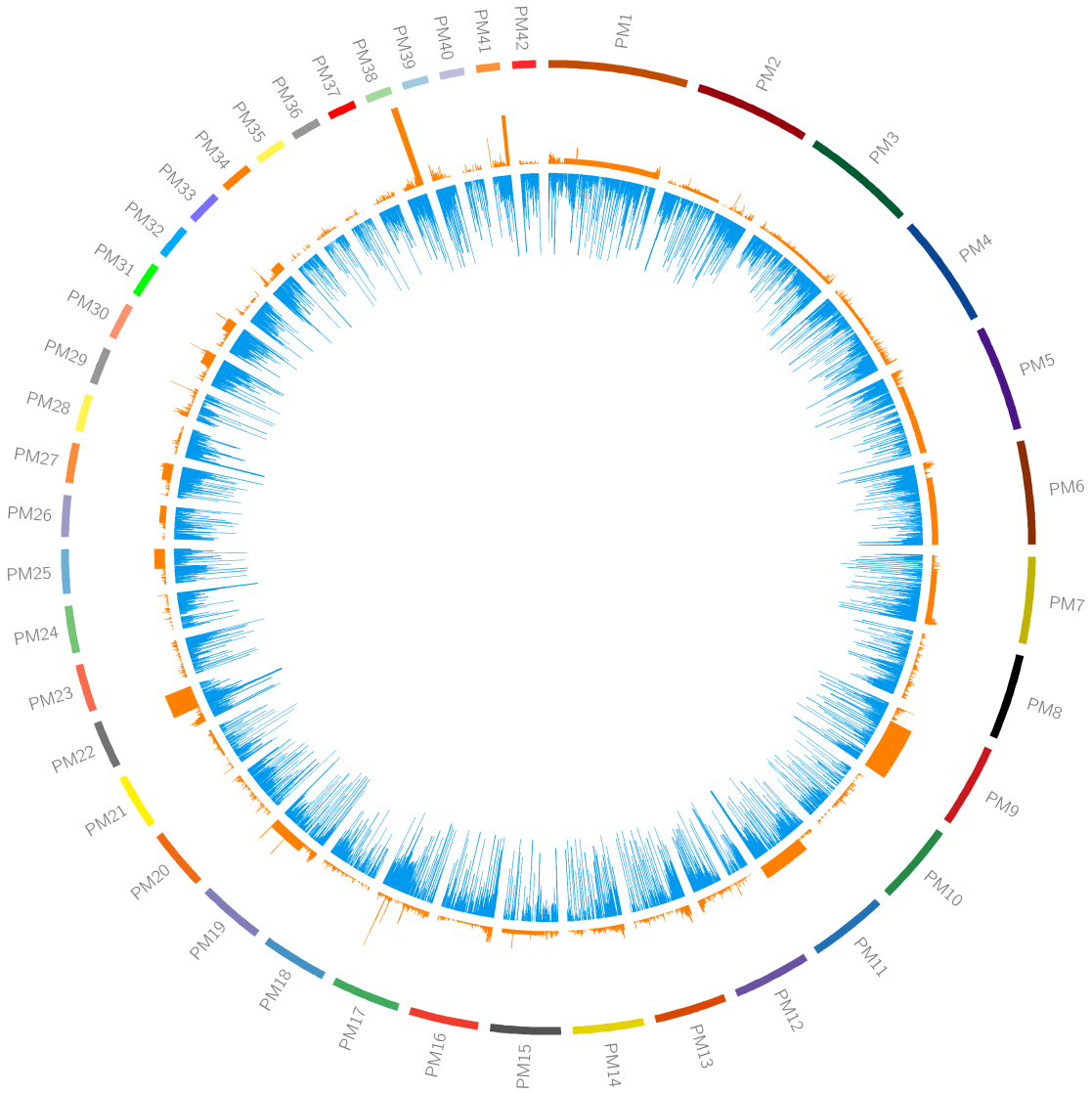


Figure 1: Circos plot which presents the GC (orange - outer circle) and repetitive elements (blue - inner circle) content as histograms in the 42 longest contigs of *P. miles*.

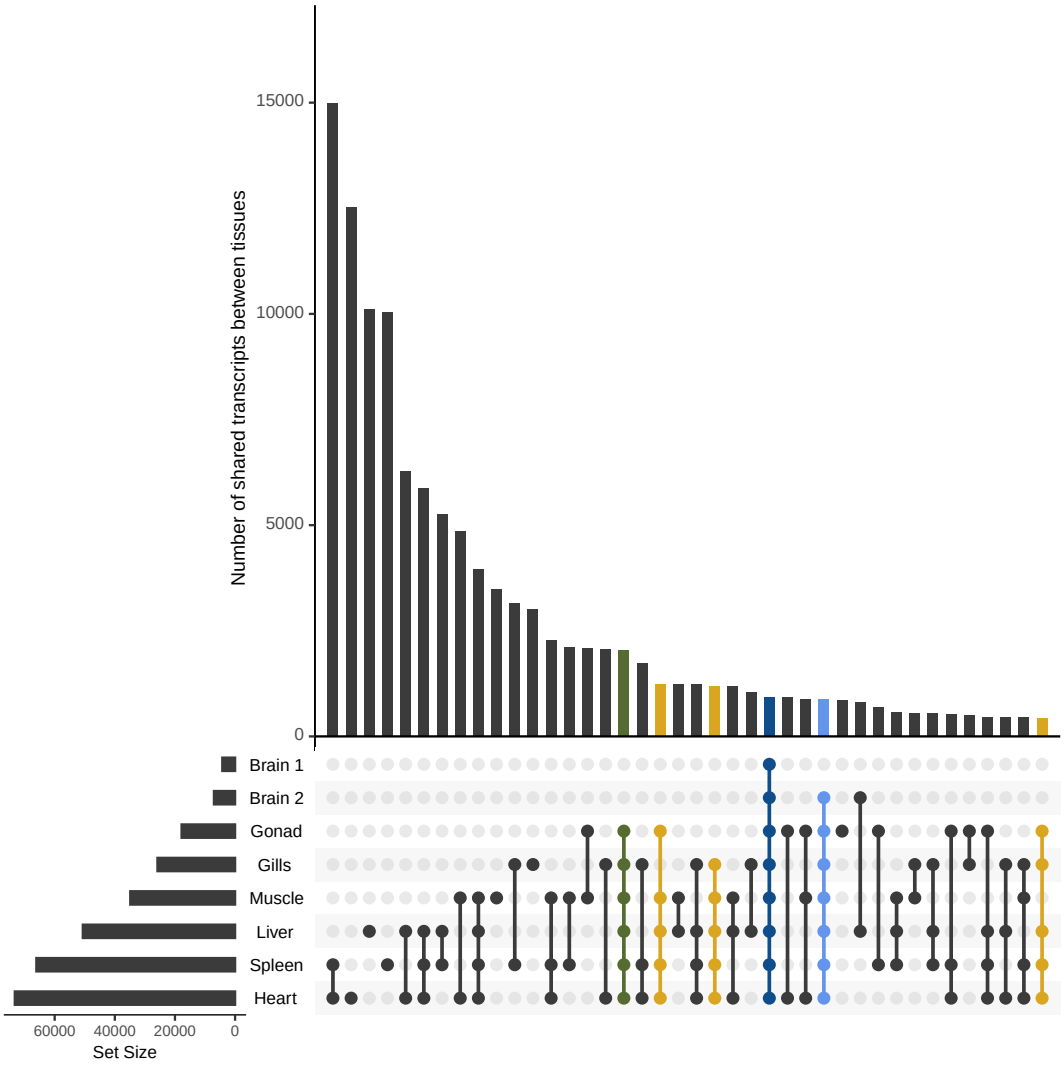


Figure 2: Number of detected transcripts in each tissue and their intersections with the others using UpSetR. Blue color shows the intersection between all tissues, light blue for 7, green for 6 and yellow-gold for 5.

Table 1: Species included in the phylogenomic analysis.

Scientific name	**Source**	**Reference**	**no. of proteins**	**BUSCO (%)**
Alosa alosa	NCBI ftp	unpublished yet	26,440	90.9
Ameiurus melas	NCBI ftp	unpublished yet	24,354	94.1
Argyrosomous regius	in-house	Papadogiannis et al. (2022)	24,443	95.3
Astyanax mexicanus	Ensembl DB	McGaugh et al. (2014)	22,998	97.3
Cheilinus undulatus	NCBI ftp	Liu et al. (2021)	23,369	99.2
Clupea harengus	NCBI ftp	Kongsstovu et al. (2019)	26,846	98.6
Collichthys lucidus	NCBI ftp	Cai et al. (2019)	28,508	91.7
Cottoperca gobio	NCBI ftp	Bista et al. (2020)	21,322	96.2
Danio rerio	Ensembl DB	Howe et al. (2013)	25,644	96.7
Dicentrarchus labrax	NCBI ftp	Tine et al. (2014)	23,380	95.9
Epinephelus lanceolatus	NCBI ftp	Wang et al. (2019)	24,223	99.9
Etheostoma cragini	NCBI ftp	Reid et al. (2020)	21,874	97.6
Fundulus heteroclitus	NCBI ftp	Reid et al. (2017)	27,033	99.4
Gambusia affinis	NCBI ftp	Shao et al. (2020)	23,272	99.3
Gasterosteus aculeatus	NCBI ftp	Nath et al. (2021)	20,779	98.6
Hemibagrus wyckioides	NCBI ftp	Shao et al. (2021)	22,794	95.6
Hippoglossus stenolepis	NCBI ftp	Jasonowicz et al. (2022)	2,1840	98.8
Kryptolebias marmoratus	NCBI ftp	Kelley et al. (2016)	22,228	99.5
Lagocephalus sceleratus	in-house	Danis et al. (2021)	21,333	91.7
Larimichthys crocea	NCBI ftp	Ao et al. (2015)	28,009	97.7
Lates calcarifer	NCBI ftp	Vij et al. (2016)	22,221	96.2
Lepisosteus oculatus	Ensembl DB	Braasch et al. (2016)	18,339	95.8
Micropterus dolomieu	NCBI ftp	unpublished yet	24,828	99.0
Mola mola	NCBI ftp	Pan et al. (2016)	21,404	94.1
Monopterus albus	NCBI ftp	Tian et al. (2020)	22,143	97.0
Nematolebias whitei	NCBI ftp	Thompson et al. (2022)	21,342	95.8
Onychostoma macrolepis	NCBI ftp	Sun et al. (2020)	24,754	93.0
Oreochromis aureus	NCBI ftp	Bian et al. (2019)	27,995	99.7
Oryzias latipes	NCBI ftp	Kasahara et al. (2007)	23,620	96.4
Pangasianodon hypophthalmus	NCBI ftp	Gao et al. (2021)	21,245	93.5
Perca flavescens	NCBI ftp	Feron et al. (2020)	23,990	99.6
Poecilia formosa	NCBI ftp	Warren et al. (2018)	23,165	98.6
Polyodon spathula	NCBI ftp	Cheng et al. (2020)	30,763	97.2
Pterois miles	in-house	current study	24,639	93.8
Sander lucioperca	NCBI ftp	Nguinkal et al. (2019)	25,044	99.7
Scophthalmus maximus	NCBI ftp	Xu et al. (2020)	21,737	99.5
Seriola dumerili	NCBI ftp	Araki et al. (2018)	23,276	98.0
Silurus meridionalis	NCBI ftp	Zheng et al. (2021)	22,769	95.2
Siniperca chuatsi	NCBI ftp	Ding et al. (2021)	22,756	99.0
Solea senegalensis	NCBI ftp	Guerrero-Cózar et al. (2021)	23,462	99.3
Takifugu rubripes	Ensembl DB	Aparicio et al. (2002)	21,411	93.8
Tetraodon nigroviridis	Ensembl DB	Jaillon et al. (2004)	19,600	88.6
Triplophysa tibetana	NCBI ftp	Yang et al. (2019)	24,310	93.1
Verasper variegatus	Ensembl DB	Zhao et al. (2021)	21,273	97.8
Xiphias gladius	NCBI ftp	Wu et al. (2021)	21,527	99.7
Xiphophorus couchianus	NCBI ftp	Shen et al. (2016)	22,879	99.7
Zebra mbuna	NCBI ftp	Conte and Kocher (2015)	26,063	99.6