UNIVERSITY OF CRETE
SCHOOL OF MEDICINE

Genome analysis of *Lagocephalus sceleratus*: unraveling the genomic aspects of a successful invader

https://upload.wikimedia.org/wikipedia/commons/4/41/Silver_puffer_%28Lagocephalus_sceleratus%29_%2848272090586%29.jpg

March, 2021

by

Theodoros Danis

School of Medicine

University of Crete

A thesis submitted for the degree of

Master in Bioinformatics

Supervisor:

Assistant Researcher, Dr. ***Tereza Manousaki***[1]

Committee members:

Assistant Professor, Dr. ***Ioannis Iliopoulos***[2]

Principal Researcher, Dr. ***George Potamias***[3]

Researcher, Dr. ***Alexandros Kanterakis***[3]

[1] Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Greece

[2] School of Medicine, University of Crete

[3] Computational BioMedicine Laboratory (CBML), Institute of Computer Science, FORTH

I dedicate this work to my family,

and every single person who stood

by my side during the years

Well begun is half done

- Aristotle

Life is simply the reification of

the process living

- Ernst Mayr

ACKNOWLEDGMENTS

Always, this section will be incomplete. Inspiration, motivation and ideas come from a multitude of sources, and it is unrealistic to cover them all properly. I'll do my best of my abilities.

My main supervisor has been Tereza Manousaki. Even though she sometimes misspells my last name, she is a well of knowledge in all kinds of areas and supportive.

The second supervisor has been Costas Tsiggenopoulos, who is always available to discuss all things in bioinformatics and life in sciences.

Additional support was provided by the project "MOdern UNifying Trends in marine biology (MOUNT)" (5002470) funded in the context of the call "Operational Program - Competitiveness, Entrepreneurship and Innovation, 2014-2020

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. Furthermore, I would like to thank the sequencing service was provided by 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

Also, I would like to thank the high performing computing cluster admins, Antonis Potirakis and Stelios Ninidakis and the people who built, maintain and support it Dimitrios Sidirokastritis and Vaggelis Pafilis, for all their help, support and computing power.

I especially have to thank Maria Daskalaki for the precious physiological support, Sandra Ramos for her endless positive energy, Nellina Angelova for extracurricular activities in swimming, beer drinking, commuting to our lab, Klara Eleftheriadi for her aesthetic

corrections (not only in our plots!) and for music entertainment, Vasileios Papadogiannis for his scientific integrity and others that have joined from time to time.

My parents and my brother who have always been supportive, for which I thank them, even though they might not properly understand what I am doing.

Η παρούσα πτυχιακή εργασία πραγματοποιήθηκε για τις απαιτήσεις του μεταπτυχιακού προγράμματος σπουδών Βιοπληροφορικής

στο

Τμήμα Ιατρικής, Πανεπιστήμιο Κρήτης, Ελλάδα

Μάρτιος 2021

CONTENTS

ΠΕΡΙΛΗΨΗ

Η ταξινομική οικογένεια Tetraodontidae περιλαμβάνει διάφορα είδη που ελκύουν ερευνητικό ενδιαφέρον όσον αναφορά την οικολογία και την εξέλιξη τους. Ωστόσο, τα γονιδιώματα αναφοράς και κάθε είδους γενωμικών πηγών, σπανίζουν για τα μέλη της οικογένειας αυτής. Στην παρούσα εργασία, επικεντρώσαμε την προσοχή μας στο λαγοκέφαλο (*Lagocephalus sceleratus*), γνωστό και ως 'invasive sprinter', διότι εισέβαλε από την Ερυθρά Θάλασσα στην Ανατολική και μέρος της Δυτικής Μεσογείου, μέσω της διώρυγας του Suez, μόλις μέσα σε μία δεκαετία. Το γονιδίωμα που κατασκευάστηκε αποτελείται από 235 contigs (N50 = 11,3 Mb) συνολικού μεγέθους 360 Mb με πληρότητα σε 98% σύμφωνα με τα αποτελέσματα του BUSCO. Το γονιδίωμα περιέχει 21,251 γονίδια εκ των οποίων τα 20,578 έχουν ταυτοποιηθεί λειτουργικά. Η φυλογενετική θέση του λαγοκέφαλου είναι ως το γειτονικό είδος του *T. nigroviridis* και μέσω της ανάλυσης των γονιδιακών οικογενειών βρέθηκε ότι 28 οικογένειες έχουν υποστεί διεύρυνση και 13 συρρίκνωση. Επιπλέον, ένας σημαντικός αριθμός γονιδίων του ανοσοποιητικού συστήματος έχει διευρυνθεί υποδεικνύοντας την σημαντική συνεισφορά τους στην επιτυχημένη εισβολή του λαγοκέφαλου. Το υψηλής ποιότητας γονιδίωμα που κατασκευάστηκε στην παρούσα εργασία αναμένεται να αποτελέσει την βάση για περαιτέρω μελέτες της βιολογίας του λαγοκέφαλου.

ABSTRACT

The Tetraodontidae family encompasses several species which attract scientific interest in terms of their ecology and evolution. However, the genomic resources and especially reference assemblies are sparse for the members of this family. The silver-cheeked toadfish (*Lagocephalus sceleratus*) is a well-known 'invasive sprinter' that has invaded and spread, less than a decade, throughout the Eastern and part of the Western Mediterranean Sea from the Red Sea through the Suez Canal. In this study, we focus on the construction of the first high-quality genome assembly of *L. sceleratus* and the exploration of its evolutionary adaptations. The resulted assembly consisted of 235 contigs (N50 = 11,3 Mb) with a total size of 360 Mb and yielded 98% BUSCO completeness. The genome possesses 21,251 predicted encoding genes with annotation of 20,578. The phylogenomic analysis positioned *L. sceleratus* as sister species to *T. nigroviridis* and gene family evolution analysis revealed rapid expansion and contraction of 28 and 13 gene families, respectively. Several genes of immune response have experienced rapid expansion, suggesting their important role in *L. sceleratus*' successive colonisation. The high-quality genome assembly built here is expected to set the ground for future studies on this focal species' invasive biology.

BACKGROUND

The Suez Canal's opening in 1869 initiated a process of invasion from the Red Sea into the Mediterranean, an event commonly known as Lessepsian migration (Por, 1971; Golani 2010). This influx of marine organisms has greatly impacted the local communities in ecological, evolutionary (Sax et al. 2007), and economical terms (Arim et al. 2006). Lessepsian fish comprise nowadays a significant percentage of all recorded invasive species in the Mediterranean Sea (Zenetos et al. 2012) and may be causing several indigenous species displacements (Golani 2010). Lessepsian migration, having both direct and indirect human-driven origins, direct and indirect origins, is a phenomenon suitable for studying fast evolutionary change (Palumbi 2001).

Genome-wide data exploration is a major process to investigate potential adaptive changes that affect invasion success. Comparative genomics is a powerful tool for shaping the species' evolutionary history and unveiling each organism's genomic features. These may include changes in sequence, genes, synteny (i.e., the order of genetic loci within chromosomes), regulatory regions, and many other genomic structural landmarks (Xia, 2013), contributing to the identification of conserved regions, genes, or gene blocks among species. Conserved regions could be useful for the exploration of chromosomal rearrangements or retained lineage specific gene families that are linked with species adaptation and functional innovation. Another source of adaptations is the gene family size (i.e., the number of gene members within each family) which varies across different lineages, with potential impact on species' adaptation (Nei and Rooney, 2005). Differences in the rates of gene gains and losses among species and lineages are proxied by differences in the copy numbers of homologous gene families. These deviations could induce phenotypic novelties in species (Meng and Yang, 2019).

The Silver-cheeked toadfish, *Lagocephalus sceleratus* (Gmelin 1789), is a member of the Tetraodontidae family (called puffers), widely distributed throughout the Indian and Pacific Oceans (Akyol et al. 2005). The first record of *L. sceleratus* invasion in the Mediterranean Sea, was reported in the Gökova Bay, in the south-eastern Aegean Sea coast of Turkey (Filiz and Er, 2004), and two years later in the North Cretan Sea (Kasapidis et al. 2007). Fatal toxicity, capability of fast spreading throughout the entire Levant, Aegean and Ionian Seas (Akyol & Ünal, 2017; Kalogirou, 2013), reduction of important commercial cephalopod species stocks and damaging of fishing gears (Bakiu and Durmishaj 2019) render *L. sceleratus* one of the most significant alien fish (Streftaris and Zenetos 2006). However, the lack of a

high-quality reference genome assembly hampers genome-wide exploration of potential adaptive changes affecting its invasion success.

Following the advances of sequencing technology and bioinformatic methodologies, the aim of this thesis is to provide and analyse the first high-quality genome assembly of *L. sceleratus*. To that end, we combined short but accurate Illumina reads with long but error-prone Oxford Nanopore Technology (ONT) reads and explored the genomic landscape of this successful invader. This valuable and robust genome source of *L. sceleratus*, will enable future studies on ecological, evolutionary and other aspects of the species' biology.

METHODS

Genome & Transcriptome raw data

The processes of sample collection, libraries construction & sequencing were carried out by collaborators and described in this study (Danis et al., 2020). Resulted statistics from raw reads sequencing are summarized in Table 1.

Genomic data pre-processing and Genome size estimation

Quality assessment of the raw DNA Illumina sequence data was performed with FastQC v0.11.8 (Andrews et al. 2010). Low quality reads and adapters were removed using Trimmomatic v0.39 (Bolger et al. 2014). The reads were scanned by a 4-based sliding window with average cutting threshold lower than 15 Phred score. Leading and trailing bases were also filtered out with quality score less than 10. Reads with total length shorter than 75 bp and average score below 30 have been omitted. The RNA filtering protocol was similar to DNA one, differentiating at the total length average score threshold, which was set to "25".

Adapter trimming and length filtering of basecalled ONT data was done using Porechop v0.2.4 (https://github.com/rrwick/Porechop) with default parameters and the extra option --discard_middle to discard reads with internal adapters.

The genome size was estimated using the k-mer histogram method with Kmergenie v1.7051 (Chikhi and Medvedev 2014) from Illumina data.

De novo genome assembly

The whole genome assembly pipeline conducted is described schematically in Figure 1 and containerised by Nelina Angelova et al (2021) (https://github.com/genomenerds/SnakeCube, zenobo reference https://doi.org/10.5281/zenodo.4663113).

In detail, the long ONT reads were used for the construction of the initial *de novo* assembly, and then the Illumina reads were used for the polishing stages. To build the initial assembly, we compared the results from five different softwares, i. SMARTdenovo (https://github.com/ruanjue/smartdenovo) which produces an assembly from all-vs-all raw read alignments without an error correction stage, ii. Canu v1.8 (Pinto 2014) which relies on the *overlap-layout-consensus* (OLC) method and incorporates an error correction step, iii. Raven v1.3.0 (Vaser and Šikić, 2020) based on the *overlap-layout-consensus* (OLC) and novel graph drawings, iv. MECAT pipeline (Xiao et al., 2017) which relies on distance difference

factors (DDFs) to score matched k-mer pairs and v. Flye v2.6 (Kolmogorov et al. 2019) algorithm, a repeat graph assembler.

First, we corrected the ONT dataset with Canu, using default parameters except for *corMinCoverage* which was set to *0*, allowing read correction regardless of the coverage and *corMhapSensitivity* was set to *high*, due to the estimated low coverage of our dataset (~20X). Next, we performed two rounds of assembly, one with SMARTdenovo and one with Canu, with default parameters in both cases. Finally, three different assemblies were constructed totally using Flye with default settings and an approximate genome size of 500 Mb, Raven assembler and MECAT pipeline following default parameters.

The assemblies produced by the different strategies were evaluated with: (1) the N50 sizes of contigs, using QUAST v5.0.2 (Gurevich et al. 2013), and (2) using BUSCO v3.1.0 (Simão et al. 2015) either standalone or through gVolante (Nishimura et al. 2017) against the Actinopterygii ortholog dataset v9, with default parameters.

Based on the quality assessment results, we selected the initial assembly produced using Flye. We polished the selected assembly with two rounds of Racon v1.4.3 (Vaser et al. 2017), using only prepossessed long reads mapped against the assembly with Minimap2 v2.17 (Li 2018). Further polishing was performed with Medaka v0.9.2 (https://github.com/nanoporetech/medaka) and the final polishing completed using Pilon v1.23 (Walker et al. 2014) after mapping the Illumina reads against the partially polished assembly with Minimap2 v2.17.

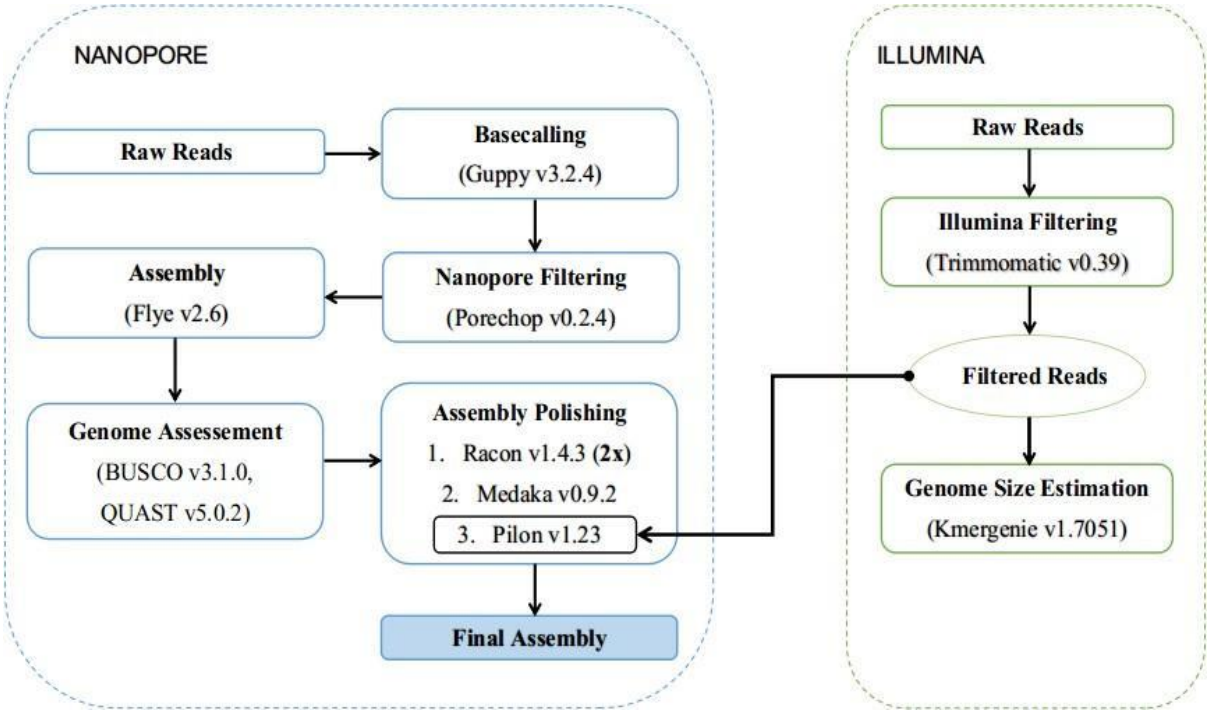

Figure 1. *L. sceleratus* genome assembly pipeline.

Genome annotation

Repeat Annotation

Repeat elements sequences were identified and annotated, in the *L. sceleratus*, prior to gene prediction analysis. A *de-novo* repeat library was constructed using RepeatModeler2 (Flynn et al. 2020), including RECON v1.08 (Bao & Eddy, 2002), RepeatScout v1.0.6 (Price et al. 2005), LtrHarvest (Ellinghaus et al., 2008), which is incorporated in GenomeTools v1.5.9, Ltr_retriever v2.7 (Ou and Jiang 2018), assuming default parameters and the extra LTRStruct pipeline which includes Mafft v7.453 (Katoh and Standley 2013), CD-HIT v4.8.1 (Li & Godzik, 2006) and Ninja v0.95 (Wheeler 2009). Thereafter, sequences that obtained by RepeatModeler, were combined with Repbase v17.01 and a custom database which was constructed with the entries of *Takifugu rubripes*, *Takifugu flavidus* and *Tetraodon nigroviridis* of the FishTEDB (Shao et al. 2018). Finally, RepeatMasker v4.1.0 (Tarailo-Graovac and Chen 2009) was used to annotate repeat elements based on the previous database.

Gene prediction & Functional annotation

After repeat masking, gene prediction was conducted using MAKER2 pipeline v2.31.10 (Holt and Yandell 2011) with two iterative rounds. We used a combined strategy of ab-initio,

homology-based and transcriptome-based methods. In the first round, for homology annotation MAKER2 was initially run in protein2genome mode, while SWISS-PROT (www.uniprot.org) was used for protein sequences extraction of three closely related species, *Mola mola*, *T. nigroviridis* and *T. rubripes*. For RNA-seq annotation, est2genome mode was enabled, which is based on transcriptome evidence. Transcriptome reads from the sequenced (female) *L. sceleratus* mixed tissues (brain, gonad, skin, liver, spleen and muscle) were mapped and assembled through the genome-guide approach, using HISAT2 v2.2.0 (Kim et al. 2015) and StringTie v2.1.1 (Pertea et al. 2015). *Ab initio* prediction was performed with SNAP (Korf 2004) (http://korflab.ucdavis.edu), which was independently trained on *L. sceleratus* genome with default parameters and AUGUSTUS v3.3.3 (Stanke et al. 2006) previously trained through BUSCO v3.1.0 (Simão et al. 2015) with the extra parameter "-long". The second round of MAKER2 was run using the previously trained models with the same settings as round one, except est2genome and protein2genome modes. The previous custom repeat library and MAKER2 repeat library that used for genome masking, remained for both rounds. The completeness of putative genes was assessed using BUSCO v4.0.5 (Simão et al. 2015) against the Actinopterygii odb10 database.

The functional annotation of the predicted genes of *L. sceleratus* was performed by alignment to the UniprotKB/Swissprot database (release-2020_03) with BLASTP v.2.9.0+ (e-value 1e-6, -max_target_seqs=10) (Altschul, S.F et al., 1990). InterProScan v5 (Jones et al. 2014) was used to search motifs and domains against all default databases and the extra of SignalP_GRAM_POSITIVE, SignalP_GRAM_NEGATIVE, SignalP_EUK and TMHMM. Functional annotation results were also retrieved using eggNOG-mapper (Huerta-Cepas et al. 2017) based on fast orthology assignments using precomputed eggNOG v5.0 (Huerta-Cepas et al. 2019) clusters and phylogenies.

Gene Ontology mapping

Gene ontology analysis was carried out using a custom python script (gene_ontology_mapping.py). Gene ontology terms were retrieved through the Uniprot API service (https://www.uniprot.org/help/programmatic_access) and as queries we chose the best blast hits that we extracted after the functional annotation step against UniProtKB/Swiss-Prot.

Phylogenomic analysis

Orthology assignment

To identify paralogous and orthologous genes, 27 whole-genome protein-coding gene sets from teleost fish (Table 8) (Natsidis et al., 2019), adding three new species (*Takifugu bimaculatus. T. flavidus* and *L. sceleratus*) were compared with Orthofinder v2.3.12 (Emms and Kelly 2015) with default parameters. Firstly, the longest isoform of each gene was kept using the *primary_transcript.py* script provided by Orthofinder suite. Then, the *L. sceleratus* longest isoforms over 30 amino acids were extracted with a custom script (longestIsoforms.py).

Species tree reconstruction

To construct a reliable dataset for the phylogenomic analysis, the orthogroups were filtered, keeping those containing a single gene per species to avoid, inclusion of paralogs. Then, we kept those with representation from at least 26 out of the 30 taxa analysed in total, using a custom python script (filtered_orthogroups.py) The genes included within each orthogroup were aligned using MAFFT v7.453 (Katoh and Standley 2013), with the -auto mode. The aligned orthogroups were concatenated using a python script by P. Natsidis (https://github.com/pnatsi/Sparidae_2019/blob/master/concatenate.py). The resulted alignments were filtered with Gblocks v0.91b (Castrecana 2000) to exclude poorly aligned regions with the following parameters: Allowed Gap Positions was set to half, Minimum Length of A Block was set to 8, Minimum Number Of Sequences For A Flanking Position was set to 20 and Minimum Number Of Sequences For A Conserved Position was set to 18.

Then, we ran RAxML-NG v0.9.0 (Kozlov et al. 2019) for phylogenetic tree reconstruction. To select the best model, we used ModelTest-NG v0.1.6 (Darriba et al. 2019) specifying the --topology type parameter to maximum likelihood (ml) mode. The phylogenomic inference was run using the selected model, JTT+I+G4+F*. To assess the branch confidence, we ran 100 bootstrap replicates. The final tree was visualized using R/RStudio (RStudio Team (2021) with a custom script using *Lepisosteus oculatus* as outgroup (phylo_tree_plot.r).

Gene family expansion and contraction

The expansion and contraction of gene families were analysed using CAFE v4.2.1 (De Bie et al. 2006). The sequences were first clustered using MCL (Enright et al. 2002) following the CAFE developers instructions (https://iu.app.box.com/v/cafetutorial-files) and filtering the gene families using a custom python script (cafe_filterin_blast_dump.py) to exclude gene

families that contain at least one or more species with ≥ 100 genes. An ultrametric tree was produced with r8s v1.81 (Sanderson 2003) using the phylogenetic tree produced in our phylogenomic analysis and the divergence time for *Thunnus thynnus* and *Oreochromis niloticus* taken from TIMETREE (http://www.timetree.org/). Finally, CAFE was run with conditional P-values, for each gene family below 0.01.

Synteny analysis

Synteny analyses were performed on two tiers, on whole-genome sequence comparison and at the gene level.

For the whole-genome comparison, the *L. sceleratus* genome assembly was compared with all available Tetraodontiformes genomes to date (*T. nigroviridis*, *T. rubripes*, *T. flavidus*, *T. bimaculatus*). Furthermore, we performed comparisons among *T. nigroviridis,* against *T. rubripes, T. bimaculatus* and *T. flavidus*. Finally, we also aligned the genome of *T. rubripes* against *T. nigroviridis, T. flavidus* and *T. bimaculatus*. For the alignment, we used LAST v1145 (Kielbasa et al., 2011), implementing the sensitive alignment protocol as described for Human-mouse whole-genome project comparison (https://github.com/mcfrith/last-genome-alignments) with e-value cutoff 0.001. On the gene level, we used the one-to-one orthologues outputs of Orthofinder v2.3.12 analysis for the same species comparisons.

Representing ~91% of the genome (Figure 2), the 41 largest contigs of the *L. sceleratus* assembly were chosen for visualisation, both for whole-genome and gene-based synteny results. In each set of fish, the whole-genome pairwise alignments were plotted by custom python scripts (synteny_plot.py) (Figures 3-8, Supplementary Information), while the one-to-one orthologs relationships of all the above-mentioned comparisons, were visualised by Circos (circos_plot.py) (Krzywinski et al., 2009) (Figures 9-13, Supplementary Information).

RESULTS

Genome size and assembly completeness

Sequencing yielded 57.30 Gb of raw Illumina reads and 9.68 Gb, above Q7, of long ONT reads, with N50 of 48.85 Kb. The estimated genome size was ~360 Mb and best predicted k = 81. After quality trimming and filtering, we retained 44.45 Gb Illumina data for genome polishing and 9.67 Gb ONT data (Table 1) employed for the genome assembly. The final assembled and polished genome contained 235 contigs with total length of ~373 Mb, with 41

contigs representing ~91% of the genome (Figure 2) and the largest contig sizing 17 Mb and N50 of 11 Mb (Table 2). Regarding genome completeness, we found 98% (4,513 out of 4,584) of the genes included in the BUSCO Actinopterygian geneset. Of those, 96.20% (4,410) were complete (Table 2), suggesting a high level of completeness and contiguity in the built assembly.

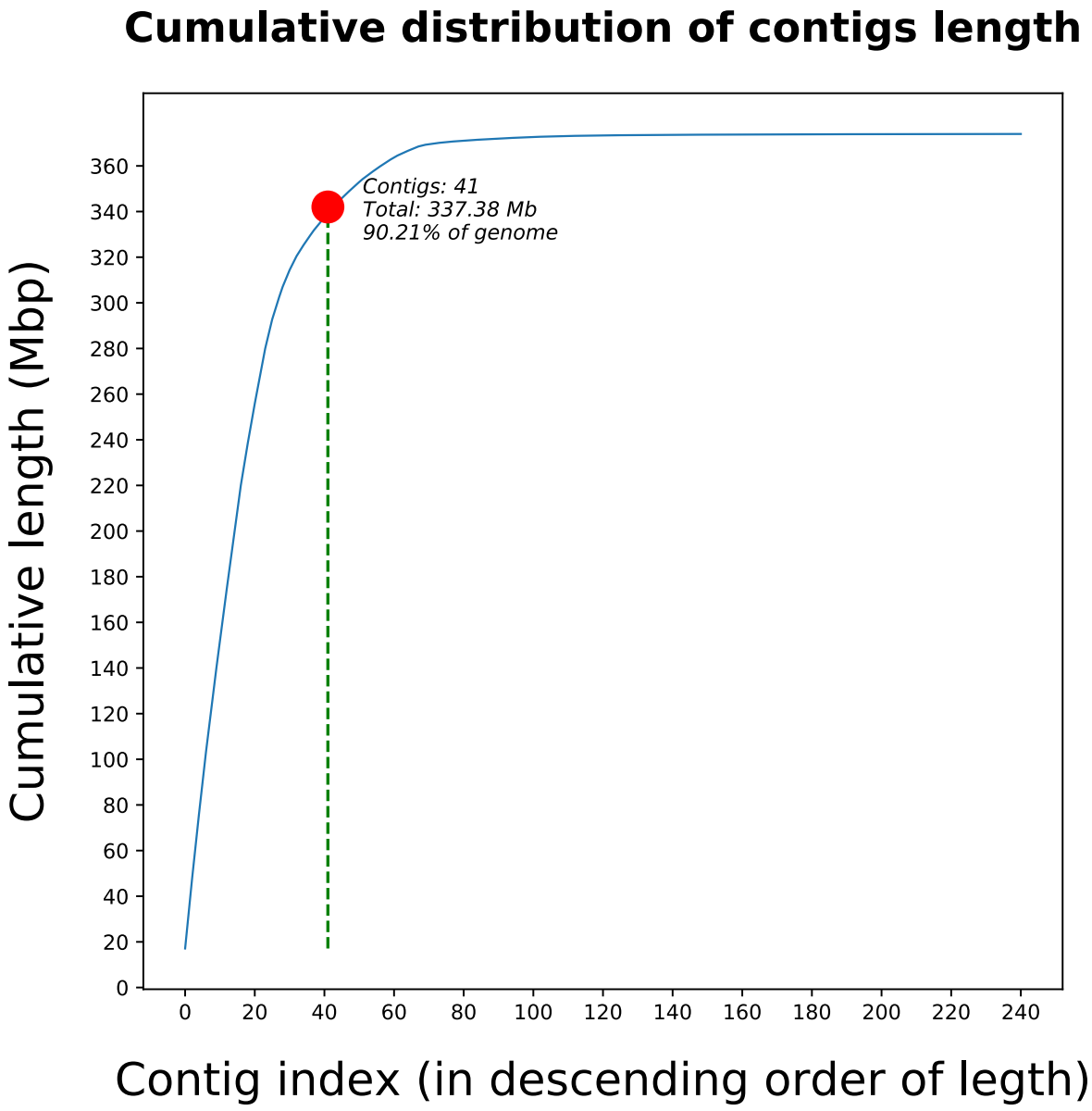

Figure 2. Length distribution plot of the *L. sceleratus* genome assembly.

Repeat annotation, gene prediction and functional annotation

A total of 61.9 Mb of repeat sequences that accounted for 16.55 % of the genome assembly

were masked in *L. sceleratus*. The class of Retroelements makes up 7.81 % of the total assembly and LINEs are the most abundant of this class, with 5.54 %. LTR elements sequences (2.07%) is the second most abundant group in the Retroelements class, and the results also indicated that 2.30% of genome assembly consists of sequences of the class DNA transposons (Table 3).

Gene prediction and functional annotation

The combination of ab initio-based, homologue-based and RNA-seq-based methods resulted in 32,451 putative protein-coding genes. After removing putative genes with Annotation Edit Distance (AED) (Eilbeck et al., 2009) score below one (AED<1), with custom script (longestIsoforms.py) we ended up with 21,251 genes (average gene length and exon size of 587,43 bp and 249,62 bp respectively). A total of 20,578 genes were successfully annotated, accounting for 97% of the predicted gene set (Table 4). The BLAST top hits indicated that *L. sceleratus* genes exhibited among others, 6,704 sequence similarities to *T. rubripes*, and 4,057 to *T. nigroviridis.* The gene number is comparable to those of 10 species (Table 5).

The completeness of the gene set was assessed using BUSCO v4.0.5 (Simão et al. 2015). From a core set of 3,640 single-copy ortholog genes from the *Actinopterygii* (odb 10) lineage, 92.2% were complete (70.5% as single-copy, 21.7% as duplicates), 1.7% were fragmented and 6.1% were not found.

Orthology assignment and Phylogenomic analysis

The total number of genes from all 30 fish proteomes (Natsidis et al., 2019) (Table 8) analysed by Orthofinder was 731,383 while 21,897 orthogroups were identified. After filtering, we chose 731 one-to-one orthogroups to construct the super-alignment. The initial matrix consisted of 494,732 amino acid positions. The Gblocks-filtered matrix contained 252,477 positions (51% of the original), which were used for the phylogenomic analysis.

We identified JTT+I+G4+F as the best model was used for the estimation of the phylogenetic tree (Figure 3). Almost all branches were supported with 100 bootstraps. The recovered phylogenetic position of *L. sceleratus* is within Tetraodontidae and is placed as the sister species to *T. nigroviridis.*The recovered Tetraodontiformes group was the closest group to *Sparus aurata*, a well-known representative species of the Sparidae family: *L. sceleratus* (family:Tetraodontidae) and *T. nigroviridis* (family:Tetraodontidae) exhibited longer branches than the sister group of *T. flavidus* (family:Tetraodontidae), *T. bimaculatus* (family:Tetraodontidae) and *T. rubripes* (family:Tetraodontidae). Instead, almost the same

branch length observed among the branches of *L. sceleratus* and *T. nigroviridis* group and *M. mola*, a Tetraodontiformes class species of Molidae family.

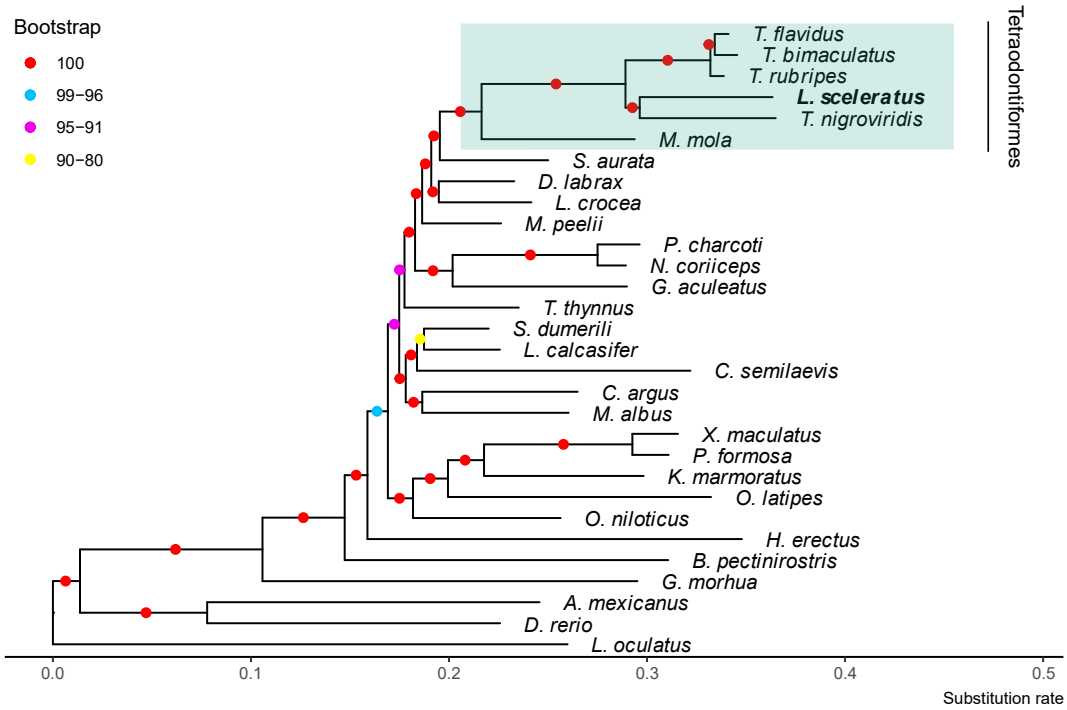

Figure 3. Maximum likelihood tree using JTT +I+ G4+F model and 100 bootstrap replicates. As an outgroup was used the spotted gar (*L. oculatus*).

Gene family evolution

The resulted rapidly expanded and contracted gene families of *L. sceleratus* are shown in Table 9 and Table 10, respectively. The subset of the total gene families with at least one gene family corresponding to a minimum one of the 5 pufferfishes is considered as puffer-specific. Among these, 4 gene families were extracted as *L. sceleratus* specific, 2 as rapidly expanding (Table 6) and 2 gene families as rapidly contracting. Finally, 43 and 45 of *T. bimaculatus* and *T. rubripes* gene families rapidly contracted, respectively (Figure 1-2, Supplementary Information). The gene family analysis also revealed no shared core set of gene families among the five puffer species (Figure 1-2, Supplementary Information).

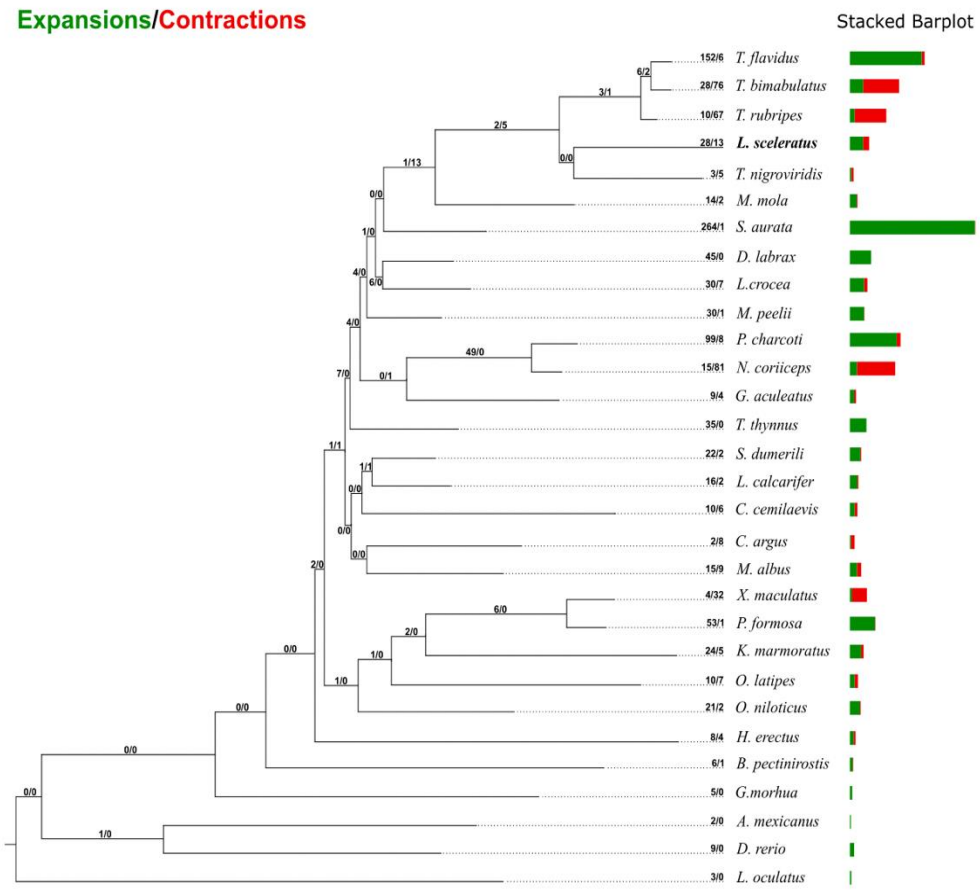

Figure 4. Gene family gain-and-loss analysis, including the number of gained gene families (green) and lost gene families (red). The stacked barplot on the right indicates the gains and losses per species.

Synteny analysis

Whole genome-based synteny

The results that we obtained from whole-genome pairwise alignment of *L. sceleratus* against the four puffers (*T. nigroviridis, T. rubripes, T. flavidus and T. bimaculatus*), summarised in Table 7. Comparisons of *L. sceleratus* against *T. nigroviridis* (Figure 5), *T. rubripes*, *T. flavidus* and *T. bimaculatus* (Figure 9-11, Supplementary information) indicated high colinearity of the focal species genome against the rest. In particular, we aligned ~231.3 Mb of *L. sceleratus* 41 high quality contigs against *T. flavidus* chromosomes, ~224 Mb against *T. bimaculatus*, ~227 Mb against *T. rubripes* and ~154.7 Mb against *T. nigroviridis*.

In addition, whole genome alignmets results obtained from *T. nigroviridis* against *T. flavidus*,

T. bimaculatus and *T. rubripes* (Figure 3-5, Supplementary information) revealed highly contiguous matches along all the species but quite similar matching patterns were observed between chromosomes of *T. nigroviridis* and *T. bimaculatus* and less between *T. nigroviridis* and *T. rubripes*.

Gene-based synteny analysis

Information obtained from gene-based analysis revealed highly conserved synteny between *L. sceleratus* and *T. nigroviridis* (Figure 6) and less conserved between *L. sceleratus* against *T. rubripes* and *T. bimaculatus* (Figure 14-15, Supplementary information). The contigs that do not present any link against *T. nigroviridis* have aligned with regions of the Unplaced chromosome of *T. nigroviridis* (Figure 15 Supplementary information).

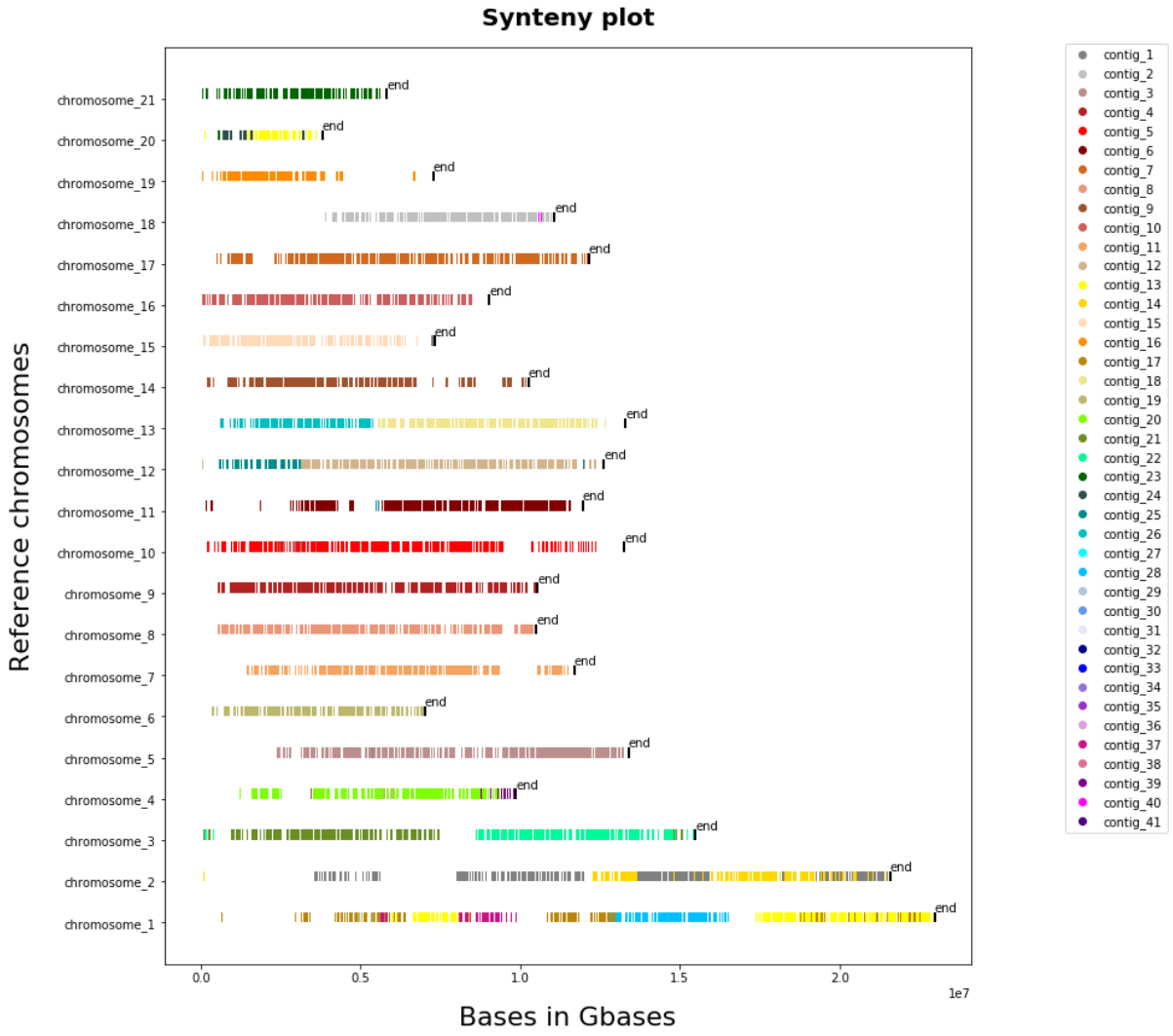

Figure 5. Synteny plot of pairwise whole genome alignment of *L. sceleratus* against *T.*

nigroviridis. **The plot illustrates the contigs of the *L. sceleratus* aligned to *T. nigroviridis* chromosomes (y-axis), grouped by a specific color which are represented on the legend right to the plot.**

Figure 6. Circos plot illustrating syntenic relationships between *L. sceleratus* contigs (right) against *T. nigroviridis* chromosomes (left), based on one-to-one orthologous genes. Ribbons link orthologous genes between the two species and colors represent the different contigs of *L. sceleratus*.

DISCUSSION

Genome size and assembly completeness

The resulted assembly *L. sceleratus* (360 Mb) is comparable in size with that of other puffers, such as *Fugu rubripes* (~365 Mb; Aparicio et al. 2002), *T. flavidus* (~377 Mb; Zhou et al. 2019a), *T. bimaculatus* (~393.15 Mb; Zhou et al. 2019b), *T. obscurus* (~373 Mb; Kang et al. 2020) and *T. nigroviridis* (340 Mb, Jatllon et al. 2004).

We here managed to reconstruct a high-quality pufferfish genome assembly of great contiguity using a single MinION flow cell and half a lane of Illumina data. In particular, the contig N50 value (~11 Mb) of the *L. sceleratus* assembly is considerably greater than that reported for the genomes of *T. bimaculatus* (1.31 Mb; Zhou et al. 2019b) and *T. flavidus* (4.4 Mb; Zhou et al. 2019a). Similarly, our assembly appears of equivalent levels of completeness that other Tetraodontidae genomes, as shown by BUSCO scores (e.g., *T. obscurus* [Kang et al. (2020)] and *T. flavidus* [Zhou et al. (2019a)]).

To our knowledge apart from the present study, only one more teleost highly contiguous reference genome assembly for *Thamnaconus septentrionalis* has been recovered using a single MinION flow cell along with a moderate amount of short Illumina reads so far (Bian et al. 2019).

Repeat content, gene prediction and functional annotation

The percentage of transponsable elements (TEs) found in *L. sceleratus* genome (16.55 % of the genome assembly), is marginally higher than the one found in *T. septentrionalis* (14.2%) (Bian et al., 2019), another species of Tetraodontiformes, *T. obscurus* (11.05%) (kang et al., 2020), and *M. mola* (11%) (Pan et al. 2016). Moreover, it is almost double than that found in *T. rubripes* (7.53%) and threefold higher than *T. nigroviridis* (5.60%) and *T. flavidus* (6.87%) (Gao et al., 2014). *T. rubripes* contains more copies of transposable elements than *T. nigroviridis* and probably these copies contribute to its marginally larger genome size (365-370 Mb) (Jatllon et al., 2004). In contrast, *L. sceleratus* repeat content shows striking differences from the reported *Takifugu* genomes.

However, the *D. holocanthus* genome of Diodontidae family has 36.35% of repetitive sequences that is almost the double size of *L. sceleratus* repeat content. These findings imply that TEs might follow an independent pathway of accumulation and diversification across Tetraodontiformes species and, especially in *L. sceleratus*, after the divergence from the

common ancestor of the species belonging to Takifugu and Tetraodon genus.

Along with TEs variation in fish, positive correlation of genome size and TE repeat content has been observed (Shao et al., 2019). The relatively smaller genome of *T. nigroviridis* possesses 5.6% TEs (~360 Mb), in contrast to zebrafish genome (~1.4 Gb) which is composed of 55% repetitive sequences, among 39 species genomes (Shao et al., 2019). This positive correlation is also reflected in the *L. sceleratus* genome.

Especially, LINEs elements have ~170,000 copies in the present genome of *L. sceleratus* and are the most abundant, as compared with the ~12,300 copies of Fugu genome. This finding indicates dynamic genome evolution between the two species. Previous studies have shown the correlation between TEs and 'adaptability' to new environment and their role to invasion (Yuan et al. 2018, Stapley et al., 2015). Overall, the repeat content of *L. sceleratus* may play a role in its adaptation to the novel environment.

Species tree reconstruction

Although the teleost fish order Tetraodontiformes is a cosmopolitan taxonomic group, including multiple families, large parts of their phylogenetic relationships remained unexplored. In this study we presented the first phylogenetic tree based on whole genome data including the invasive "sprinter*" L. sceleratus*. The recovered phylogenetic position of *L. sceleratus* is within Tetraodontidae and is placed as the sister species to *T. nigroviridis,* showing a long branch, possibly suggesting a faster evolutionary rate. Regarding within-puffers' relationships (*T. nigroviridis, T. rubripes, T. flavidus, T. bimaculatus* and *L. sceleratus*), the resulting phylogenetic position agrees with previous studies (Hughes et al., 2020, Hughes et al., 2018, Meynard et al., 2012, Yamanoue et al., 2009). Moreover, the Tetraodontidae was recovered confidently as a monophyletic clade, which also agrees with a previous study (Yamanoue et al., 2011). Our results suggest that Tetraodontiformes are the closest group to Sparidae and corroborates the results of Natsidis et al. 2019 and other studies (Kawahara et al., 2008, Meynard et al., 2012), based both on mitochondrial and 6 mitochondrial and nuclear genes, respectively.

Synteny analysis

According to our whole-genome synteny analysis of *L. sceleratus* against the four Tetraodontidae species (Figures 5) (Figure 9-11, Supplementary Information), all pairwise comparisons showed highly conserved synteny. However, the genome that exhibited the highest synteny conservation with *L. sceleratus* genome was that of *T. nigroviridis.* This agrees

with our phylogenetic tree that shows these two species as more closely related compared to the rest.

Regarding macro-synteny, we observed absence of highly conserved genomic regions between *L. sceleratus* and the three Tetraodontidae species (*T. rubripes*, *T. bimaculatus* and *T. flavidus*), and especially between *L. sceleratus* and *T. bimaculatus*. The lack of conserved synteny within puffer fish could be explained by the long branches observed in the phylogenetic tree.

Further, we found high levels of rearrangements especially in chromosome one (1) of *T. rubripes, T. flavidus, T. bimaculatus* and in the second (2) chromosome of *T. nigroviridis*. This finding was also supported by the gene-based synteny analysis.

To sum up, the conserved sequence collinearity and synteny between *L. sceleratus* and *T. nigroviridis* corroborates the phylogenetic position of *L. sceleratus* as closest relative to *T. nigroviridis.*

Gene family evolution and adaptation

The ability of taking hold of a new habitat or niche is a challenging component of a species physiology. To achieve establishment, and invader must face environmental challenges that involve both biotic and abiotic factors (Crowl et. al., 2008). Invasive species are facing novel pathogens during the colonisation of new environments and the ability to deal with these new immune challenges is key to their invasion success (Lee and Klasing, 2004). Interestingly, we found several expanded immune related families, including *immunoglobulins (C-Type and V-Type), Ig heavy chain Mem5-like, B-cell receptors* and the *Fish-specific NACHT associated domain* which is related to the main innate immunity (Stein et al., 2007).

In addition, in the expanded gene families we detected major histocompatibility complex (*MHC) class I* genes. MHC genes are crucial for immune response, as are evolved in pathogen-derived recognition by T cells (Germain, 1994), thus initiating the adaptive immune response. The expanded repertoire of *L. sceleratus* immune response might be related to its survival in novel habitats and the detection of a wide range pathogens. Therefore, in this context, we suggest further systematic research and even more in-depth experimental investigations, in order to explore and unravel the role of the expanded genes related to immune response.

Another interesting, expanded gene family was the fucosyltransferase (FUT) gene family. In particular,we detected 24 FUT9 (alpha (1,3) fucosyltransferase 9) genes. Glycosylation is one of the most frequent post-translational modifications of a protein. Before their implication in

immune response, the proteins are glycosylated extending their diversity and functionality (Bednarska et al., 2017). Fucosylation, a type of glycosylation, plays an essential role in cell proliferation, metastasis and immune escape (Jia et al., 2018). FucTC has been shown in mice to regulate leukocyte trafficking between blood and the lymphatic system after its engagement in selectin ligand biosynthesis (Maly et al., 1996). Within Mediterranean Sea, *L. sceleratus* generally inhabits sandy or muddy substrate bottoms, which are very important for juveniles, while adults are more commonly ground in *Posidonia oceanica* meadows (Kalogirou er al., 2013). Probably, *L. sceleratus* ability to survive and conquer successfully new habitats is facilitated by changes in the immune system. In either case, the expanded innate immune system gene families identified in our study seem to be indispensable components of *L. sceleratus* immune system swift of habitat during its growth/maturation and its rapid spread throughout the Eastern Mediterranean (Kalogirou 2011).

Additionally, we detected nine putative genes with no evidence of rapid expansion but functionally characterised as Voltage-sensitive calcium channels (VSCC), which among other functions are insensitive to omega-conotoxin- GVIA (omega-CTx-GVIA) and omega-agatoxin-IVA (omega-Aga-IVA). Omega-conotoxin acts at presynaptic membranes and binds and blocks voltage-gated calcium channels (Cav) (Lewis et al., 2000). Moreover omega-agatoxin inhibits neuronal voltage-gated calcium channels (Bickmeyer et al., 1994, Nishio et al., 1993).

Finally, our dataset also includes the sequences of voltage-gated sodium channels (SCN4AA, SCN4AB). Tetrodoxin (TTX) is a neurotoxin that blocks voltage sodium channels (Narahashi et al., 1964, Venkatesh et al., 2005) and causes severe symptoms such as skeletal muscle paralysis and death (Simmons, 2007). Pufferfishes accumulate high concentrations of TTX in different tissues but remain resilient to TTX toxicity. TTX has been detected in several lineages of pufferfishes, so it is likely that the genetic basis of TTX resistance, which was unveiled in *T. rubripes* and *T. nigroviridis* (Venkatesh et al., 2005), is common in different Tetraodontidae genera including *L. sceleratus* as well.

CONCLUSION

Invader fishes, such as *L. sceleratus,* is a group of fish that can thrive in novel environments. Our analysis provides the first high-quality genome assembly and a comprehensive evolutionary genomic analysis of the species. We uncovered a close phylogenetic position of

L. sceleratus with *T. nigroviridis* untangling the relationships within the puffers' group, that were not clearly resolved in previous non-whole-genome studies. The study also reveals genomic insights into a variety of genomic signatures that may be associated with *L. sceleratus*' invasion effectiveness and colonisation. *L. sceleratus* genome will be an invaluable resource for additional studies on immune response in novel environments, osmoregulation, reconstruction of ancient chromosome rearrangements and will play important role for the species conservation management.

CODE AVAILABILITY

The custom scripts that have been used in this study are available at this github repository (https://github.com/Tdanis/Lagocephalus_genome_analysis/tree/master)

*__JTT__ (general matrix), **I** (allowing for a proportion of invariable sites), **F** (empirical codon frequencies counted from data) and **G4** (discrete Gamma model (Yang, 1994) with default 4 rate categories. The number of categories can be changed with e.g., G8) (http://www.iqtree.org/doc/Substitution-Models)

Table 1. Summary of sequencing results.

Sequencing technology	**Raw Reads**	**Quality-controlled Reads**	**Coverage**
Illumina	57,303,140	44,475,382	38 x
MinION	552,476	484,152	20 x

Table 2. Polished genome assembly statistics and completeness.

Total contigs	235
Total contig sequence	373,851,781 bp
GC (%)	46,7
Contig N50	11,297,640 bp
Contig N75	6,386,954 bp
Longest contig	17,085,954 bp
BUSCO completeness score	
Complete	96.20 %
Single	94.40 %

Duplicated	2.20 %
Fragmented	1.40 %
Missing	2.00 %
Total number of Actinopterygii orthologs	4,492 (98%)

Table 3. Repeat annotation statistics.

Repetitive Elements	**Number of elements**	**Length occupied (bp)**	**Percentage of sequence**
Retroelements	204,205	29,198,482	7.81 %
SINEs:	5,447	743,745	0.20 %
Penelope	3,627	2,572,446	0.69 %
LINEs:	171,108	20,706,770	5.54 %
CRE/SLACS	0	0	0.00%
L2/CR1/Rex	45,521	7,208,493	1.93 %
R1/LOA/Jockey	573	179,409	0.05 %
R2/R4/NeSL	48,675	3,708,881	0.99 %
RTE/Bov-B	47,510	3,856,263	1.03 %
L1/CIN4	16,080	2,144,919	0.57 %
LTR elements:	27,650	7,747,967	2.07 %
BEL/Pao	410	259,661	0.07 %
Ty1/Copia	235	96,868	0.03 %
Gypsy/DIRS1	13,595	3,917,10	1.05 %

Retroviral	4,582	1,298,868	0.35 %
DNA transposons	59,679	8,587,997	2.30 %
hobo-Activator	25,182	2,911,788	0.78 %
Tc1-IS630-Pogo	14,227	2,885,165	0.77 %
En-Spm	0	0	0.00 %
MuDR-IS905	0	0	0.00 %
PiggyBac	644	131,052	0.04 %
Tourist/Harbinger	2,399	392,537	0.10 %
Other (Mirage,P-element, Transib)	99	5,804	0.00 %
Unclassified	113,777	1,126,996	5.97%
Small RNA	0	0	0.00%
Satellites	20	85,955	0.02%
Simple repeats	9,305	1,126,996	0.30%

Table 4. Summary statistics of functional annotated protein-coding genes.

Type	**Number**	**Per cent (%)**
Blast	18,805	88%
InterProScan	20,347	96%
EggNog-Mapper	17,849	84%
Predicted genes	20,578	97%
Total Genes	**21,251**	

Table 5. Fish genome gene prediction summary statistics.

Species	Genes	Reference
Lagocephalus sceleratus	21,251	Current study
Danio rerio	25,403	Howe et al., 2013
Amphiprion percula	27,240	Tan et al., 2018
Diodon holocanthus	20,840	Xu et al., 2019 (bioRxiv)
Tetraodon nigroviridis	23,118	llon et al., 2004
Takifugu rubripes	22,760	Aparicio et al., 2002
Takifugu flavidus	29,416	Zhou et al.,2019a
Takifugu bimaculatus	21,117	Zhou et al.,2019b
Takifugu obscurus	22,105	kang et al., 2020
Epinephelus akaara	23,903	Ge et al., 2019
Thamnaconus septentrionalis	22,067	Bian et al., 2019

Table 6. *L. sceleratus* specific genes from the 2 rapidly expanded gene families.

Entry	Protein names	Gene names	Organism
Q4RTI4	Chromosome 1 SCAF14998	GSTENG00029235001	*Tetraodon nigroviridis*
A0A3P9KVW8	Uncharacterized protein	-	*Oryzias latipes*
A0A3B4VFG6	Reverse transcriptase domain-containing protein	-	*Seriola dumerili*

A0A0G2L2B6	Reverse transcriptase domain-containing protein	-	-
A0A3B5QQH6	Uncharacterized protein	-	*Xiphophorus maculatus*

Table 7. Summary of whole genome synteny plots. The last row contains the contigs that are represented within the chromosomes of the puffers.

Species	Chromosomes																					
T.bimaculatus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
T.flavidus	1	3	2	5	4	7	8	6	14	10	12	13	11	9	15	18	17	16	19	-	20	22
T.rubripes	1	2	13	21	19	20	3	17	15	7	11	9	17	14	22	5	10	2	16	-	6	18
T.nigroviridis	2	-	5	12	11	1	9	18	7	8	16	13	4	20	15	-	6	10	12	10	4	19
Contigs	1, 14	6,2	4,38	7	22,33, 34	20, 39	8,35 36	1,2 24	18, 26	19, 31	10, 32	23, 27	3,29	13, 37	11	9	2,40	16	6	17, 18	12, 25	15,30

Table 8. Species included in the phylogenetic tree.

Species	Series (for Percomorphaceae)	Source	Reference	#of proteins
A. mexicanus	(Ostariophysi)	Ensembl database	McGaugh, S. E. et al., 2014	22,998
B. pectinirostris	Gobiaria	NCBI ftp server	You, X. et al., 2014	21,541

C. argus	Anabantaria	GigaDB	Xu, J. et al., 2017	20,541
C. semilaevis	Carangaria	NCBI ftp server	Chen, S. et al., 2014	24,489
D. labrax	Eupercaria	species database	Tine, M. et al., 2014	26,719
D. rerio	(Ostariophysi)	Ensembl database	Howe, K. et al., 2013	25,644
G. aculeatus	Eupercaria	Ensembl database	Jones, F. C. et al., 2012	20,625
G. morhua	(Paracanthopterygii)	Ensembl database	Star, B. et al., 2011	19,978
H. erectus	Syngnatharia	GigaDB	Lin, Q. et al., 2017	20,788
K. marmoratus	Ovalentaria	NCBI ftp server	Kelley, J. L. et al., 2016	25,257
L. calcarifer	Carangaria	NCBI ftp server	Vij, S. et al., 2016	22,221
L. crocea	Eupercaria	NCBI ftp server	Ao, J. et al., 2015	28,009
L. oculatus	(Holostei)	Ensembl database	Braasch, I. et al., 2016	18,304
L. sceleratus	Eupercaria	in-house sequenced	Current study	21,251
M. albus	Anabantaria	NCBI ftp server	Yi, M. et al., 2014	24,943
M. mola	Eupercaria	GigaDB	Pan, H. et al., 2016	19,605
M. peelii	Eupercaria	GigaDB	Austin, C. M. et al., 2017	26,539

N. coriiceps	Eupercaria	NCBI ftp server	Shin, S. C. et al., 2014	25,937
O. latipes	Ovalentaria	Ensembl database	Kasahara, M. et al. 2007	19,603
O. niloticus	Ovalentaria	Ensembl database	Brawand, D. et al. 2014	21,383
P. charcoti	Eupercaria	provided by authors	Ahn, D. H. et al., 2017	32,713
P. formosa	Ovalentaria	Ensembl database	Warren, W. C. et al., 2018	23,315
S. aurata	Eupercaria	in-house sequenced	Pauletto, M. et al., 2018	61,850
S. dumerili	Carangaria	NCBI ftp server	Araki et al., unpublished	24,000
T. bimaculatus	Eupercaria	Uniprot ftp server	Zhou, Z., et al., 2019b	19,334
T. flavidus	Eupercaria	Uniprot ftp server	Gao et al., 2014	29,076
T. nigroviridis	Eupercaria	Ensembl database	Jatllon, O. et al., 2004	19,511
T. rubripes	Eupercaria	Ensembl database	Aparicio, S. et al., 2002	18,433
T. thynnus	Pelagiaria	species database	Nakamura, Y. et al., 2013	26,433
X. maculatus	Ovalentaria	Ensembl database	Schartl, M. et al., 2013	20,343

Table 9. Number of contigs that are involved in rapidly expanded gene families and its functionality.

Function	**# of contigs**
Transposase	24
Fucosyltransferase 9 (alpha (1,3) fucosyltransferase)	23
ENV polyprotein (coat polyprotein)	21
K02A2.6-like	12
SCAN domain	12
B-cell receptor CD22-like	7
Reverse transcriptase (RNA-dependent DNA polymerase)	6
Podospora anserina S mat genomic DNA chromosome	6
Ig heavy chain Mem5-like	6
Si ch211-286b4.4	6
Nuclear migration along micro-filament	5
Si ch211-81n22.1	5
Immunoglobulin C-Type	5
Immunoglobulin V-Type	5
Fish-specific NACHT associated domain	5
DDE superfamily endonuclease	4
Receptor	4
protein dimerization activity	3
Immunoglobulin V-set domain	3

cytoskeletal anchoring at nuclear membrane	3
Early B-cell factor 3	3
Early B-cell factor 1	2
Early B-cell factor 2	2
Sad1 and UNC84 domain containing 1	2
Serpentine type 7TM GPCR chemoreceptor Srx	2
Fucosyltransferase 7 (alpha (1,3) fucosyltransferase)	2
Immunoglobulin C1-set domain	2
Retrotransposable element Tf2 155 kDa protein type 1-like	2
Belongs to the MHC class I family	2
Early B-cell factor	2
Solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3	1
Coreceptor activity involved in Wnt signaling pathway, planar cell polarity pathway	1
Phosphodiesterase 4D	1
B-cell receptor	1
Interleukin 4 induced 1	1
Metal dependent phosphohydrolase with conserved ‘HD’ motif.	1
ST3 beta-galactoside alpha-2,3-sialyltransferase 1	1
N-acetyltransferase	1
Reverse transcriptase	1

Galactosyltransferase	1
Transposition, RNA-mediated	1
DNA-binding transcription factor activity, RNA polymerase II-specific	1
Endonuclease/Exonuclease/phosphatase family	1
Phosphodiesterase 4A, cAMP-specific	1
Receptor accessory protein-like 2	1
Monovalent cation proton antiporter 1 (CPA1) transporter (TC 2.A.36) family	1
Si dkey-24p1.1	1
Calcium binding protein	1
Pao retrotransposon peptidase	1
SH3 domain binding kinase family, member 2	1
Ribonuclease H protein	1
Purinergic receptor P2Y, G-protein coupled, 12	1
Fucosyltransferase 4 (alpha (1,3) fucosyltransferase, myeloid-specific)	1
Calcium ion binding	1

Table 10. Number of contigs that are involved in rapidly contracted gene families and its functionality.

Function	**# of contigs**
Glutamate receptor, ionotropic	7
regulator of G-protein signaling	7

protein heterodimerization activity	6
Belongs to the beta gamma-crystallin family	6
ryanodine receptor	5
Belongs to the G-protein coupled receptor 1 family	3
Beta/gamma crystallins	3
7 transmembrane receptors (rhodopsin family)	2
Opsin 6, group member a	2
fibulin-like extracellular matrix protein	2
Proprotein convertase subtilisin kexin type	2
Belongs to the G-protein coupled receptor 1 family. Opsin subfamily	2
Teleost multiple tissue opsin	2
Glutamyl aminopeptidase	1
Opsin 8, group member c	1
Aminopeptidase puromycin sensitive	1
C-terminus of histone H2A	1
Endoplasmic reticulum aminopeptidase 1	1
Opsin 4xb	1
Regulator of G-protein signaling 3-like	1
Subtilase family	1
Serine-type endopeptidase activity	1

regulator of G-protein	1
Histone H2A	1
aminopeptidase	1
Ryanodine Receptor TM 4-6	1
Opsin 4a (melanopsin)	1
Rhodopsin	1
Fibulin 2	1
Protein-chromophore linkage	1
Green-sensitive opsin-like	1
Opsin 1 (cone pigments), long-wave-sensitive	1
H2A histone family, member	1
H2A histone family, member Y2	1
Incorporated into fibronectin-containing matrix fibers. May play a role in cell adhesion and migration along protein fibers within the extracellular matrix (ECM). Could be important for certain developmental processes and contribute to the supra-molecular organization of ECM architecture, to those of basement membranes	1
Vertebrate ancient long opsin	1
Peptidase family M1 domain	1
Ataxin 10	1
Transcription intermediary factor 1-alpha-like	1
Belongs to the histone H2B family	1

negative regulation of adenylate cyclase-inhibiting adrenergic receptor signaling pathway involved in heart process	1
Belongs to the protein kinase superfamily. Tyr protein kinase family	1
Opsin 4.1	1
H2A histone family, member Y	1
Si ch211-196h16.12	1
Histone H1 like	1
Leucyl cystinyl aminopeptidase	1
Gamma-crystallin M2-like	1
Histone H2A-like	1
Belongs to the histone H2A family	1

REFERENCES

Akyol O, Unal V, Ceyhan T and Bilecenoglu M (2005) First confirmed record of *Lagocephalus sceleratus* (Gmelin, 1789) in the Mediterranean. Journal of Fish Biology 66: 1183-1186

Akyol, O., and Ünal, V. (2017). Long journey of *Lagocephalus sceleratus* (Gmelin, 1789) throughout the Mediterranean Sea. *Natural and Engineering Sciences* 2(3): 41-47. doi: 10.28978/nesciences.369534

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

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

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. ming, Dehal, P., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science.* 297, 1301–1310. doi:10.1126/science.1072104

Arim, M., Abades, S. R., Neill, P. E., Lima, M., and Marquet, P. A. (2006). Spread dynamics of invasive species. *Proc.Natl.Acad.Sci.U.S.A.* 103, 374–378. doi:10.1073/pnas.0504272102

Bakiu, P.A and Durmisshaj (2019). First record of the silver-cheeked toadfish *Lagocephalus sceleratus* (Gmelin, 1789) in Albanian waters. Pp. 237-238. In: Kousteni, V., Bakiu, R. A., Benhmida, A., Crocetta, F., Martino, V. Di, Dogrammatzi, A., et al. New Mediterranean Biodiversity Records (April 2019). Med. Mar. Sci. 20(1), 230–247

Bao, Z. and Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research, 12*, 1269–1276. https://doi.org/10.1101/gr.88502

Bednarska, N. G., Wren, B. W., & Willcocks, S. J. (2017). The importance of the glycosylation of antimicrobial peptides: natural and synthetic approaches. In Drug Discovery Today (Vol. 22, Issue 6, pp. 919–926). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2017.02.001

Bian, L., Li, F., Wang, P., Zhang, S., Liu, K., Liu, X., et al. (2019). Chromosome-level genome assembly of the greenfin horse-faced filefish (*Thamnaconus septentrionalis*) using Oxford Nanopore PromethION sequencing and Hi-C technology. *bioRxiv Genomics*, 1–25. doi:10.1101/798744

Bickmeyer, U., Rössler, W., & Wiegand, H. (1994). Omega AGA toxin IVA blocks high-voltage-activated calcium channel currents in cultured pars intercerebralis neurosecretory cells of adult locusta migratoria. *Neuroscience Letters*, *181*, 113–116. https://doi.org/10.1016/0304-3940(94)90572-X

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

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17, 540-552

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., Song, W., et al. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics*, *46*, 253–260. https://doi.org/10.1038/ng.2890

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30, 31–37. doi:10.1093/bioinformatics/btt310

Crowl, T. A., Crist, T. O., Parmenter, R. R., Belovsky, G., and Lugo, A. E. (2008). The spread of invasive species and infectious disease as drivers of ecosystem change. *Frontiers in Ecology and the Environment*, *6*, 238–246. https://doi.org/10.1890/070151

Darriba, D, Posada, D, Kozlov, MA, Stamatakis, A, Morel, B, Flouri, T, ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models, Molecular Biology and Evolution, Volume 37, Issue 1, January 2020, Pages 291–294, https://doi.org/10.1093/molbev/msz189

De Bie, T., Cristianini, N., Demuth, J. P. and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, *22*, 1269–1271. https://doi.org/10.1093/bioinformatics/btl097

Demuth, J. P., and Hahn, M. W. (2009). The life and death of gene families. *BioEssays*, *31*(1), 29–39. https://doi.org/10.1002/bies.080085

Deniz Magazin (Istanbul) 68, 52-54. [in Turkish]

Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 2009 Feb 23, 10:67. doi: 10.1186/1471-2105-10-67. PMID: 19236712; PMCID: PMC2653490.

Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 18 (2008). https://doi.org/10.1186/1471-2105-9-18

Emms, D.M., Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16, 157 (2015). https://doi.org/10.1186/s13059-015-0721-

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002 Apr 1;30(7):1575-84. doi: 10.1093/nar/30.7.1575. PMID: 11917018; PMCID: PMC101833.

Filiz, H. and Er, M. (2004). Akdenizin yeni misafiri (New guests in the Mediterranean Sea).

Flajnik, M. F. (2018). A cold-blooded view of adaptive immunity. In *Nature Reviews Immunology* (Vol. 18, Issue 7, pp. 438–453). Nature Publishing Group. https://doi.org/10.1038/s41577-018-0003-9

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

Gao, Y., Gao, Q., Zhang, H., Wang, L., Zhang, F., Yang, C., & Song, L. (2014). Draft Sequencing and Analysis of the Genome of Pufferfish Takifugu flavidus. DNA Research, 21(6), 627–637. https://doi.org/10.1093/dnares/dsu025

Ge, H., Lin, K., Shen, M., Wu, S., Wang, Y., Zhang, Z., et al. (2019). De novo assembly of a chromosome-level reference genome of red-spotted grouper (*Epimetheus akaara*) using nanopore sequencing and Hi-C. *Mol. Ecol. Resour.* 19, 1461–1469. doi:10.1111/1755-0998.13064

Germain, R. N. (1994). MHC-dependent antigen processing and peptide presentation: Providing ligands for T lymphocyte activation. In *Cell* (Vol. 76, Issue 2, pp. 287–299). Elsevier. https://doi.org/10.1016/0092-8674(94)90336-0

Golani, D and Appelbaum-Golani, B. (2010). *FISH INVASIONS of the MEDITERRANEAN SEA*: Change and Renewal, PENSOFT Publishers

Guidelines for the treatment of animals in behavioral research and teaching. (1977). *Anim. Behav.* 53, 229–234

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

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

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*, 498–503. https://doi.org/10.1038/nature12111

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C. and Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*, *34*, 2115–2122. https://doi.org/10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., et al. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*, D309–D314. https://doi.org/10.1093/nar/gky1085

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., et al. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 6249–6254. https://doi.org/10.1073/pnas.1719358115

Hughes, L. C., Ortí, G., Saad, H., Li, C., White, W. T., Baldwin, C. C., Crandall, K. A., et al. (2021). Exon probe sets and bioinformatics pipelines for all levels of fish phylogenomics. *Molecular Ecology Resources*, *21*, 816–833. https://doi.org/10.1111/1755-0998.13287

Jatllon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Maucell, E., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto karyotype. *Nature* 431, 946–957. doi:10.1038/nature03025

Jia, L., Zhang, J., Ma, T., Guo, Y., Yu, Y., & Cui, J. (2018). The Function of Fucosylation in Progression of Lung Cancer. Frontiers in oncology,8, 565. https://doi.org/10.3389/fonc.2018.00565

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kalogirou, S. (2013). Ecological characteristics of the invasive pufferfish *Lagocephalus sceleratus* (Gmelin, 1789) in Rhodes, Eastern Mediterranean Sea. A case study. *Med. Mar. Sci.* 14, 251–260. doi:10.12681/mms.364

Kalogirou, S. (2013). Ecological characteristics of the invasive pufferfish *Lagocephalus sceleratus* (Gmelin, 1789) in the eastern Mediterranean Sea – a case study from Rhodes. Mediterranean Marine Science, 14(2), 251-260. doi:https://doi.org/10.12681/mms.364

Kang, S., Kim, J., Jo, E., Lee, S. J., Jung, J., Kim, B., et al. (2020). Chromosomal-level assembly of *Takifugu obscurus* (Abe, 1949) genome using third-generation DNA sequencing and Hi-C analysis. *Mol. Ecol. Resour.* 00, 1–11. doi:10.1111/1755- 0998.13132

Kang, S., Kim, J., Jo, E., Lee, S. J., Jung, J., Kim, B., Lee, J. H., Oh, T., Yum, S., Rhee, J., & Park, H. (2020). Chromosomal-level assembly of Takifugu obscurus (Abe, 1949) genome using third-generation DNA sequencing and Hi-C analysis. Molecular Ecology Resources, 1755-0998.13132. https://doi.org/10.1111/1755-0998.13132

Kasapidis, P., Peristeraki, P., Tserpes, G. and Magoulas, A. (2007). First record of the Lessepsian migrant *Lagocephalus sceleratus* (Gmelin 1789) (Osteichthyes: Tetraodontidae) in the Cretan Sea (Aegean, Greece). *Aquat. Invasions* 2, 71–73. doi:10.3391/ai.2007.2.1.9

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software

Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010

Kawahara, R., Miya, M., Mabuchi, K., Lavoué, S., Inoue, J. G., Satoh, T. P., Kawaguchi, A. and Nishida, M. (2008). Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives): A new perspective based on whole mitogenome sequences from 75 higher teleosts. *Molecular Phylogenetics and Evolution*, *46*, 224–236. https://doi.org/10.1016/j.ympev.2007.07.009

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011 Mar;21(3):487-93. doi: 10.1101/gr.113985.110. Epub 2011 Jan 5. PMID: 21209072; PMCID: PMC3044862.

Kim, D., Langmead, B. and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*, 357–360. https://doi.org/10.1038/nmeth.3317

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

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

Kozlov, MA, Darriba, D, Flouri, T, Morel, B, Stamatakis, A, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference, Bioinformatics, Volume 35, Issue 21, 1 November 2019, Pages 4453–4455, https://doi.org/10.1093/bioinformatics/btz305

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639-45. doi: 10.1101/gr.092759.109. Epub 2009 Jun 18. PMID: 19541911; PMCID: PMC275213

Lee, K. A., and Klasing, K. C. (2004). A role for immunology in invasion biology. *Trends in Ecology and Evolution*, *19*, 523–529. https://doi.org/10.1016/j.tree.2004.07.012

Lewis RJ, Nielsen KJ, Craik DJ, Loughnan ML, Adams DA, Sharpe IA, Luchian T, Adams DJ, Bond T, Thomas L, Jones A, Matheson JL, Drinkwater R, Andrews PR, Alewood PF. Novel omega-conotoxins from Conus catus discriminate among neuronal calcium channel subtypes. J Biol Chem. 2000 Nov 10;275(45):35335-44. doi: 10.1074/jbc.M002252200.

PMID: 10938268.

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

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Maly, P. et al. The alpha (1,3) fucosyltransferase Fuc-TVII controls leukocyte trafficking through an essential role in L-, E-, and P-selectin ligand biosynthesis. Cell 86, 643–653 (1996).

McGaugh, S. E., Gross, J. B., Aken, B., Blin, M., Borowsky, R., Chalopin, D., Hinaux, H., Jeffery, et al. (2014). The cavefish genome reveals candidate genes for eye loss. *Nature Communications*, *5*, 1–10. https://doi.org/10.1038/ncomms6307

Meng, Y., & Yang, R. L. (2019). [Comparative analysis of gene family size provides insight into the adaptive evolution of vertebrates]. *Yi Chuan = Hereditas*, *41*, 158–174. https://doi.org/10.16288/j.yczz.18-225

Meynard CN, Mouillot D, Mouquet N, Douzery EJP (2012) A Phylogenetic Perspective on the Evolution of Mediterranean Teleost Fishes. PLoS ONE 7: e36443. https://doi.org/10.1371/journal.pone.0036443

Michael J. Sanderson, r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock, Bioinformatics, Volume 19, Issue 2, 22 January 2003, Pages 301–302, https://doi.org/10.1093/bioinformatics/19.2.301

Narahashi, t., Moore, j. W., & Scott, w. R. (1964). Tetrodotoxin Blockage of Sodium Conductance Increase in Lobster Giant Axons. *The Journal of General Physiology*, *47*, 965–974. https://doi.org/10.1085/jgp.47.5.965

Natsidis, P., Tsakogiannis, A., Pavlidis, P., Tsigenopoulos, C. S. and Manousaki, T. (2019). Phylogenomics investigation of sparids (Teleostei: Spariformes) using high-quality proteomes highlights the importance of taxon sampling. *Communications Biology*, *2*, 1–10. https://doi.org/10.1038/s42003-019-0654-5

Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., Wainwright, P. C, et al. (2012). Resolution of ray-finned fish phylogeny and timing of diversification.

Proceedings of the National Academy of Sciences of the United States of America, *109*, 13698–13703. https://doi.org/10.1073/pnas.1206625109

Nei, M., & Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. In *Annual Review of Genetics* 39, pp. 121–152). NIH Public Access. https://doi.org/10.1146/annurev.genet.39.073003.112240

Nishimura, O., Hara, Y., and Kuraku, S. (2017). GVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* 33, 3635–3637. doi:10.1093/bioinformatics/btx445

Nishio, H., Kumagaye, K. Y., Kubo, S., Chen, Y. N., Momiyama, A., Takahashi, T., Kimura, T., & Sakakibara, S. (1993). Synthesis of ω-agatoxin IVA and its related peptides. *Biochemical and Biophysical Research Communications*, *196*, 1447–1453. https://doi.org/10.1006/bbrc.1993. 2354

Olyarnik, S. V., Bracken, M. E. S., Byrnes, J. E., Hughes, A. R., Hultgren, K. M., and Stachowicz, J. J. (2009). *Ecological Factors Affecting Community Invasibility* (pp. 215–238). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79236-9_12

Ou, S. and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, *176*, 1410–1422. https://doi.org/10.1104/pp.17.01310

Palumbi, S. R. (2001). *The Evolution Explosion: how Humans* C*ause Rapid Evolutionary Change*. New York, W.W. Norton & Co

Pan, H., Yu, H., Ravi, V. et al. The genome of the largest bony fish, ocean sunfish (Mola mola), provides insights into its fast growth rate. GigaSci 5, 36 (2016). https://doi.org/10.1186/s13742-016-0144-3

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T. and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*, 290–295. https://doi.org/10.1038/nbt.3122

Pinto, A. (2014). Secure because math: A deep-dive on machine learning-based monitoring. *Black Hat Briefings* 25, 1–11. doi: 10.1101/gr.215087.116.Freely

Por, F.D. 1971. One hundred years of Suez Canal – A century of Lessepsian migration:

retrospect and viewpoints. Systematic Zoology 20: 138-159

Price, A. L., Jones, N. C. and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*, i351–i358. https://doi.org/10.1093/bioinformatics/bti1018

Renwick JH (1971) The mapping of human chromosomes. Annual Reviews in Genetics 5: 81–120

Sathyajith, C., Yamanoue, Y., Yokobori, S. I., Thampy, S., & Vattiringal Jayadradhan, R. K. (2019). Mitogenome analysis of dwarf pufferfish (Carinotetraodon travancoricus) endemic to southwest India and its implications in the phylogeny of Tetraodontidae. *Journal of Genetics*, *98*(5), 1–11. https://doi.org/10.1007/s12041-019-1151-9

Sax, D. F., Stachowicz, J. J., Brown, J. H., Bruno, J. F., Dawson, M. N., Gaines, S. D., et al. (2007). Ecological and evolutionary insights from species invasions. *Trends Ecol. Evol.* 22, 465–471. doi: 10.1016/j.tree.2007.06.009

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

Shao, F., Wang, J., Xu, H. and Peng, Z. (2018). FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. *Database*, *2018*(2018), 106. https://doi.org/10.1093/database/bax106

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351.

Simmons, M. A. (2007). Tetrodotoxin. In *xPharm: The Comprehensive Pharmacology Reference* (pp. 1–4). Elsevier Inc. https://doi.org/10.1016/B978-008055232-3.62740-0

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W435-9. doi: 10.1093/nar/gkl200. PMID: 16845043; PMCID: PMC1538822.4

Stapley, J., Santure, A. W., & Dennis, S. R. (2015). Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Molecular Ecology*, *24*, 2 235–2252. https://doi.org/10.1111/mec.13089

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Rounge, T. B., et al. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, *477*, 207–210. https://doi.org/10.1038/nature10342

Stein, C., Caccamo, M., Laird, G. et al. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. Genome Biol 8, R251 (2007). https://doi.org/10.1186/gb-2007-8-11-r251

Streftaris, N. and Zenetos, A. (2006). Alien marine species in the Mediterranean - the 100 "worst invasives" and their impact. *Med. Mar. Sci.* 7, 87–118. doi:10.12681/mms.180

Tarailo-Graovac, M. and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. In *Current Protocols in Bioinformatics: Vol. Chapter 4* (Issue SUPPL. 25). Curr Protoc Bioinformatics. https://doi.org/10.1002/0471250953.bi0410s25

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

Vaser, R. and Šikić, M. (2020). Raven: A de novo genome assembler for long reads. In *bioRxiv* (p.2020.08.07.242461). bioRxiv. https://doi.org/10.1101/2020.08.07.242461

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

Venkatesh, B., Lu, S. Q., Dandona, N., See, S. L., Brenner, S., & Soong, T. W. (2005). Genetic basis of tetrodotoxin resistance in pufferfishes. *Current Biology*, *15*, 2069–2072. https://doi.org/10.1016/j.cub.2005.10.068

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9. doi: 10.1371/journal.pone.0112963

Wheeler, T. J. (2009). Large-scale neighbor-joining with NINJA. *Lecture Notes in Computer*

Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *5724 LNBI*, 375–389. https://doi.org/10.1007/978-3-642-04 235-6_31

Xia, X. (2013). What is comparative Genomics? Comparative Genomics (Springer), pp. 1-20.

Xiao, C. Le, Chen, Y., Xie, S. Q., Chen, K. N., Wang, Y., Han, Y., Luo, F. and Xie, Z. (2017). MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature Methods*, *14*, 1072–1074. https://doi.org/10.1038/nmeth.4432

Xu, J., Bian, C., Chen, K., Liu, G., Jiang, Y., Luo, Q., You, X., et al. (2017). Draft genome of the Northern snakehead, Channa argus. In *GigaScience* (Vol. 6, Issue 4, pp. 1–5). Oxford University Press. https://doi.org/10.1093/gigascience/gix011

Yamanoue Y, Miya M, Doi H, Mabuchi K, Sakai H, Nishida M (2011) Multiple Invasions into Freshwater by Pufferfishes (Teleostei: Tetraodontidae): A Mitogenomic Perspective. PLoS ONE 6: e17410. https://doi.org/10.1371/journal.pone.0017410

Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* **39,** 306–314 (1994). https://doi.org/10.1007/BF00160154

You, X., Bian, C., Zan, Q., Xu, X., Liu, X., Chen, J., Wang, J., et al. (2014). Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nature Communications*, *5*(1), 1–8. https://doi.org/10.1038/ncomms6594

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

Zenetos, A., Gofas, S., Morri, C., Rosso, A., Violanti, D., Garcia Raso, J. E., et al. (2012). Alien species in the Mediterranean Sea by 2012. A contribution to the application of European Union's Marine Strategy Framework Directive (MSFD). Part 2. Introduction trends and pathways. *Mediterr. Mar. Sci.* 13, 328. doi:10.12681/mms.327

Zhao, T., and Eric Schranz, M. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 2165–2174.

https://doi.org/10.1073/pnas.1801757116

Zhou, Y., Xiao, S., Lin, G., Chen, D., Cen, W., Xue, T., et al. (2019a). Chromosome genome assembly and annotation of the yellowbelly pufferfish with PacBio and Hi-C sequencing data. *Sci. data* 6, 267. doi:10.1038/s41597-019-0279-z

Zhou, Z., Liu, B., Chen, B., Shi, Y., Pu, F., Bai, H., et al. (2019b). The sequence and de novo assembly of *Takifugu bimaculatus* genome using PacBio and Hi-C technologies. *Sci. data* 6, 187. doi:10.1038/s41597-019-0195-2