

RESEARCH

Open Access



Long-read de novo genome assembly of Gulf toadfish (*Opsanus beta*)

Nicholas S. Kron^{1*}, Benjamin D. Young^{1,2}, Melissa K. Drown¹ and M. Danielle McDonald¹

Abstract

Background The family Batrachoididae are a group of ecologically important teleost fishes with unique life histories, behavior, and physiology that has made them popular model organisms. Batrachoididae remain understudied in the realm of genomics, with only four reference genome assemblies available for the family, with three being highly fragmented and not up to current assembly standards. Among these is the Gulf toadfish, *Opsanus beta*, a model organism for serotonin physiology which has recently been bred in captivity.

Results Here we present a new, de novo genome and transcriptome assemblies for the Gulf toadfish using PacBio long read technology. The genome size of the final assembly is 2.1 gigabases, which is among the largest teleost genomes. This new assembly improves significantly upon the currently available reference for *Opsanus beta* with a final scaffold count of 62, of which 23 are chromosome scale, an N50 of 98,402,768, and a BUSCO completeness score of 97.3%. Annotation with ab initio and transcriptome-based methods generated 41,076 gene models. The genome is highly repetitive, with ~70% of the genome composed of simple repeats and transposable elements. Satellite DNA analysis identified potential telomeric and centromeric regions.

Conclusions This improved assembly represents a valuable resource for future research using this important model organism and to teleost genomics more broadly.

Keywords Teleost, Model organism, PacBio, HiFi, Toadfish, Genome assembly

Background

Toadfishes are bony fish of the family Batrachoididae, which consists of 78 species including the Amazon toadfish (*Thalassophryne amazonica*), speckled midshipman (*Porichthys myriaster*), plainfin midshipman (*Porichthys notatus*), oyster toadfish (*Opsanus tau*), Lusi-tanian toadfish (*Halobatrachus didactylus*) among others [1]. In general, toadfish are small, demersal ambush

predators with wide mouths (oftentimes with barbels or fleshy projections around them) and eyes set on top of their broadheads [1]. The family exhibits a suite of distinguishing behavioral and physiological characteristics such as paternal care of the nest, complex acoustic communication, the lack of a pelagic larval phase, and, in some lineages, venom and spines [1]. Furthermore, they are tolerant to a range of environmental conditions that occur naturally or a result of anthropogenic impact [2, 3]. These traits make these fish not only interesting on an ecological basis but have made them popular study subjects both as a comparison species to other fish, but also as a potential model for human health and disease.

The genomics of toadfishes are relatively understudied. The bulk of genomic research in batrachoids has focused on karyotype and cytogenetics [4–8]. Particularly, the sequencing, localization, and phylogeny of repetitive

*Correspondence:

Nicholas S. Kron
nicholas.kron@gmail.com

¹ Department of Marine Biology and Ecology, Rosenstiel School of Marine, Atmospheric, and Earth Science, University of Miami, 4600 Rickenbacker Cswy, Miami, FL 33149, USA

² Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80302, USA



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

elements such as rDNA and GATA repeats has been well described in several genera [9–12]. For the species with available genomic resources, little has been reported in the literature on the unique characteristics of this order. *T. amazonica* has been noted to be unique among surveyed teleosts for having a large genome with a unique positive association between chromosome size and GC percentage [13]. Interestingly, two batrachoids, *Chatrabus melanurus*, and *Opsanus beta*, were described as having the highest genomic percentage of transposable elements among 100 surveyed teleost genomes [14]. These few data points suggest batrachoids exhibit unique genomic characteristics among teleosts that warrant further study.

The Gulf toadfish, *Opsanus beta* (Fig. 1), is found inshore within the western Atlantic, from southeastern Florida, USA, through the Bahamas and the Gulf of Mexico. Like other batrachoids, *O. beta* are resilient to various environmental stressors including hypoxia [15–17], ammonia [18–20], and various types of waterborne pollution [21–25] making them an intriguing subject for study. Recent work on *O. beta* has focused on describing the monoaminergic system, with a particular emphasis on serotonin and the role it plays controlling vascular resistance and blood flow [26, 27]. Multiple labs have successfully bred *O. beta* in a laboratory setting [28]. This provides the opportunity for siblings to be used in physiological studies reducing inter-individual variation, and for families to be used for examination of trait heritability, ontogenetic adaptations, and in trans-generational studies, expanding their potential as model organisms.

Despite well described biology and role as model organisms, few genomic resources are available for *O. beta* or other batrachoid fishes. To date, only a single



Fig. 1 Photo of genetic neotype for Gulf toadfish, *Opsanus beta*. The adult male, named Bic, was selected from the toadfish stock at the Toadfish Lab at University of Miami Rosenstiel School for Marine, Atmospheric, and Earth Sciences as DNA and RNA sample donor for genomic and transcriptomic assembly

representative genome assembly generated with modern long read technology exists for the family Batrachoididae; that of *T. amazonica* (GCF_902500255.1). Older genome assemblies generated with short read technology exist for *C. melanurus* and *O. beta*, but these assemblies are highly fragmentary and do not align to modern assembly standards [29]. To remedy this gap in available resources, we present a de novo long-read genome, mitochondrial genome, and transcriptome assemblies and annotations for *O. beta* using modern long read technology.

Methods

A schematic of the computational workflow can be found in Supplemental Fig. 2.

Sample collection, nucleic acid extractions, and sequencing

One adult male *Opsanus beta* (0.068 kg) was selected from the toadfish stock at the University of Miami Rosenstiel School Toadfish Lab. Toadfish are sourced from shrimper roller trawl bycatch in Biscayne Bay, Florida. For a full description of fish housing and care please see [17]. The specimen was sacrificed with an overdose of pharmaceutical grade buffered tricaine methanesulfonate (MS-222) anesthetic at a dose of 3 g.L⁻¹ (pH=8.0), as is considered acceptable by the American Veterinary Medical Association Guidelines on Euthanasia [30].

Five hundred microliters of blood was drawn from the caudal vessel via caudal puncture using a 23 gauge needle attached to a 1 ml disposable syringe that was primed with 500 ul Acetate-Citrate-Dextrose (ACD) anticoagulant buffer (480 mg citric acid, 1.32 mg sodium citrate, 1.47 mg glucose (dextrose), QS to 100 ml with distilled water) and added to another 500 ul of ACD. The blood sample was then shipped overnight at 4 °C to the University of California Davis Genome Center DNA Technologies and Expression Analysis Core Laboratory (UC Davis) for High Molecular Weight (HMW) DNA extraction, PacBio library prep, and HiFi long reads sequencing.

At UC Davis, 10 uL of settled cells were lysed until homogenous at room temperature in 2 mL of lysis buffer (100 mM NaCl, 10 mM Tris-HCl pH 8.0, 25 mM EDTA, 0.5% (w/v) SDS, 100 µg/ml Proteinase K). RNA was removed by treating lysate with 20 µg/ml RNase A for 30 min at 37°C. HMW DNA was then extracted using equal volumes of phenol/chloroform and phase lock gels (Quantabio Cat # 2,302,830). Extracted DNA was precipitated with 0.4X volume of 5 M ammonium acetate and 3X volume of ice cold ethanol, washed twice with 70% ethanol, and finally resuspended in 10 mM Tris, pH 8.0. Purity, yield, and integrity of HMW DNA was assessed with a NanoDrop ND-1000 spectrophotometer, Qubit

2.0 Fluorometer (Thermo Fisher Scientific, MA), and Femto pulse system (Agilent Technologies, CA).

The HiFi SMRTbell library was prepared and sequenced at the UC Davis DNA Technologies Core following standard recommendations from Pacific Biosciences. Briefly, the library was prepared using the SMRTbell prep kit 3.0 (Pacific Biosciences, Menlo Park, CA; Cat. #102–182-700) according to the manufacturer's instructions using sheared (15–18 kb) high molecular weight gDNA. The library was size-selected to remove sequences < 5 kb with the final library having an average size of 15–18 kb. Sequencing used three 8 m SMRT cells (Pacific Biosciences, Menlo Park, CA; Cat #101–389-001) with Sequel II sequencing chemistry 2.0, and 30-h movies on a PacBio Sequel II sequencer.

To obtain as complete as possible transcriptomic snapshot of *O. beta*, the following tissue samples were stored in RNA later (ThermoFisher Scientific, CAT# AM7020) at -80 °C: liver, kidney, brain, heart, gill, esophagus, swimbladder, muscle, skin, gonad, gastrointestinal tract, gallbladder, urinary bladder, and spleen. A whole juvenile *O. beta* was also stored in RNA later. Total RNA was extracted using the Quick-RNA Miniprep Kit (Zymo Research, CAT# R1054) following manufacturer's protocol, before being cleaned and concentrated using the RNA Clean & Concentrator-5 Kit (Zymo Research, CAT# R1013). Total RNA with 260/230 absorbance ratio greater ≥ 1.89 and a concentration > 45 ng/ μ l (RNA HS qubit) were considered of sufficient quality, pooled, and sent to UC Davis for IsoSeq library preparation and sequencing.

At UC Davis, cDNA was constructed using a NEB-Next® Single Cell/Low Input kit (New England Biolabs, Ipswich, MA; Cat. #E6421L) with 500 ng of total RNA as input. Resulting cDNA was amplified for 15 cycles using the cDNA Synthesis Amplification Module. Amplified cDNA was purified using 0.86X SMRTbell cleanup beads. SMRTbell library was constructed from 480 ng of purified cDNA with the SMRTbell prep kit 3.0 (Pacific Biosciences, Menlo Park, CA; Cat. #102–182-700). Resulting Iso-Seq library was sequenced on a single 8 M SMRT cell (Pacific Biosciences, Menlo Park, CA; Cat #101–389-001) on a PacBio Sequel II sequencer using Sequel II 2.0 chemistry.

Nuclear genome assembly

PacBio HiFi reads were assembled using methods based upon the vertebrate genome project pipeline [29]. A kmer profile of HiFi reads was generated using meryl v1.3.0 [31] which was then fed into GenomeScope v2.0.0 [32] to estimate genome parameters including genome size as well as bounds for detecting haploid and diploid kmers. HiFi reads were then assembled

into a primary and alternate assemblies using HiFi-asm v0.16.1 [33] with the -l parameter set to 1 for gentle purging and the -purge-max parameter set to the upper bound calculated from GenomeScope2 [32] estimates. The primary and alternate assemblies were then purged of duplicate kmers using purge_dups v1.2.6 [34] with the purging parameter -a set to 80. The primary purged assembly was then corrected with inspector v1.0.1 [35] for three rounds using HiFi reads as input. In the absence of 10 \times linked reads, BioNano optical maps, or Hi-C contact maps, the purged primary assembly was scaffolded using in silico methods. First, the primary purged assembly was scaffolded and gap filled with ntLink v1.3.9 [36] using HiFi reads as input to generate synthetic linked reads for 5 rounds. The scaffolded assembly was then super-scaffolded to a pseudo-chromosome level with RagTag scaffold v2.1.0 [37] using the RefSeq representative genome of *Thalassophryne amazonica* (GCF_902500255.1), the closest available relative within the family Batrachoididae with a chromosome level assembly. The divergence time between *T. amazonica* and *O. beta* is estimated to be 38 million years ago (CI 32.8–39.8 MYA) [38]. RagTag scaffold arranges contigs according to their primary mapping to reference chromosomes without altering contig sequence, and then stitches contigs together with gaps of arbitrary length (100 Ns) to represent gaps of unknown length.

Primary and alternate assemblies were assessed for kmer completeness, QV, and haplotype purging using Merqury v1.3.0 [31]. Genome length and contiguity metrics were calculated using QUAST [39, 40], Genometools [41], and Gfastats [42]. Genome completeness was measured by calculating the number of Actinopterygii single copy orthologs retained in the assembly via BUSCO [43] with the actinopterygii_odb10 database. Completeness was further assessed with read mapping rate by mapping publicly available paired-end Illumina short read RNA [44] and DNA (PRJNA196921) libraries, as well as input HiFi reads to the super-scaffolded assembly.

The primary assembly was screened for microbial contamination via BLAST+v2.13.0 [45] with megablast against an NCBI database of common contaminants in eukaryotic genomes (ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz) parameterized as described in [29]. Primary assembly was also blasted against publicly available databases of representative genome sets for prokaryotes and viruses downloaded with the update_blastdb.pl script (ref_prok_rep_genomes and ref_viruses_rep_genomes). Blastn was parameterized to only report hits with an E-value less than 10E⁻²⁰ and minimum bit score of 1000 as described in [46]. Finally, the super-scaffolded assembly was screened for off-target contaminants

using Kraken2 v2.1.3 [47] and for adaptor contamination with the GenBank Foreign Contamination Screen (FCS) tool [48].

Mitogenome assembly

The mitochondrial genome was assembled from all HiFi reads using the MitoHiFi version 2.14.2 [49, 50]. In addition to the annotations of the mitogenome generated by MitoHiFi [49, 50], the primary mitogenome assembly was also annotated with the Mitos2 webtool [51]. All HiFi reads were then mapped back to the mitogenome assembly using minimap2 v2.25 [52] to assess depth of coverage. The mitogenome assembly and annotations were visualized with the Proksee webtool (<https://proksee.ca/>) [53]. MitoHiFi [49, 50] was run two additional times in contig mode using the initial nuclear genome assembly to compare results and validate the assembly built from raw reads.

Transcriptome assembly

IsoSeq high quality (HQ) transcripts generated from HiFi reads at UC Davis with the IsoSeq3 pipeline. HQ transcripts were cleaned with seqClean (<https://sourceforge.net/projects/seqclean/files/>) using the UniVec vector database as reference (<https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/>). Cleaned HQ transcripts were then aligned to the super-scaffolded assembly using minimap2 v2.25 [52] and BLAT v35.1 [54] and then assembled into gene models using PASA v2.5.2 [55].

Genome Annotation

Transposable elements and dispersed repeats

A custom library of transposable elements (TE) and dispersed repeats was generated from the super-scaffolded assembly using RepeatModeler v2.0.3 [56] with RMBlast (<https://www.repeatmasker.org/rmbblast/>) as the default search engine. In order to eliminate false positive repetitive elements that do in fact originate from coding regions, the SwissProt database [57] was screened for transposable elements with blastp v2.13.0 [45] against a collection of repetitive elements (RepeatPeps.lib) included with RepeatModeler [56]. The resultant version of the Swissport database was subsequently used to screen the de novo repetitive elements identified in the *O. beta* assembly for false positives. High confidence de novo repetitive elements from the *O. beta* assembly were further processed and categorized using repclassifier v1.1 (<https://github.com/darencard/GenomeAnnotation/blob/master/repclassifier>) and RepeatMasker v4.1.2.pl [58] as described in (<https://darencard.net/blog/2022-07-09-genome-repeat-annotation/>) using Actinopterigii repetitive elements from Dfam [59]. Classified repeats were used to generate hard and soft masked versions of

the super-scaffolded assembly with BEDtools v4.1.2 [60]. Hard and soft masked versions of the *T. amazonica* reference were also generated with RepeatMasker [58], using the *T. amazonica* repeat database from FishTEDB [61] and subsequent repeat landscape analysis. Repeat landscapes for both species were generated using the createRepeatLandscape.pl utility from RepeatMasker [58].

Satellite DNA

Satellite DNA sequences were inferred from the super-scaffolded assembly using TRASH [62]. The centromere-like region of scaffold 1 (position 82,000,000–83,000,000) was analyzed for higher order repeats (HORs) with HiCAT v1.1.0 [63] and visualized using StainedGlass v0.5.0 [64].

Telomeres

HiFi reads, initial primary assembly, purged primary assembly, and final super-scaffolded assembly were screened for candidate telomeric repeat monomers with the telomere identification toolkit (tidk) v0.2.31 [65]. Top repeat sequences and canonical telomere marker monomers (TTAGGG and GATA) were then used as input for tidk::search to count putative telomeric repeats at scaffold ends in the super-scaffolded assembly.

Gene Models

Open reading frames (ORFs), coding regions, and protein sequences were then predicted from the PASA transcriptome using transDecoder v5.5.0 (<https://github.com/TransDecoder/TransDecoder>) and used to build training sets for downstream ab initio gene predictors. Gene models were generated from the repeat masked super-scaffolded assembly and transcript based gene models from PASA [55] with the funannotate pipeline v1.8.15 [66]. Briefly, models trained on output from PASA and transdecoder were used as input into ab initio prediction software AUGUSTUS v3.5.0 [67], snap v2013_11_29 [68], glimmerHMM v3.0.4 [69], and GeneMark-ES v3.68.0 [70]. Resulting gene predictions along with PASA [55] predicted transcripts were then passed into EvidenceModeler v1.1.1 [71] to generate a consensus set of high quality gene models. Transfer RNAs (tRNAs) were predicted using trnscan v1.4.0 [72]. Gene models were then refined and UTRs added with funannotate::update [66] using IsoSeq high quality transcripts and *O. beta* illumina short reads (PRJNA313355; [44]).

Functional Annotation

High quality gene models were then annotated using funannotate::annotate [66]. Briefly, funannotate::update screened proteins models for protein domains (Pfam v35.0 [73]), CAZymes (dbCAN v11.0 [74]), biosynthetic

classes (MiBIG v1.4 [75]), peptidases (MEROPS v 12.0 [76]), and homologs (UniprotKB/SwissProt v2023_03 [57]) using HMMER v3.3.2 (<http://hmmer.org/>) and DIAMOND v2.1.7.161 [77]. Funannotate [66] was run with optional eggNOG emapper v2.1.10 [78, 79] using eggNOG database v 5.0.2 and external annotation via InterProScan v5.52–86.0 [80]. Predicted protein models were further annotated for KEGG KO identifiers using ghostKOALA [81].

Comparison with other organisms

To identify regions of collinearity, the super-scaffolded assembly was aligned to the *T. amazonica* reference using NUCmer v3.1 from MUMmer v3.23 [82] with a minimum alignment length (-l) of 500 bases and visualized using Dot (<https://dot.sandbox.bio/>).

Results

Assemblies

Nuclear genome

Initial genome estimates from HiFi reads with GenomeScope2 [32] suggested a 2.09 gigabase genome (nearly twice the size from the current *O. beta* reference),

with ~ 53.5% of the genome composed of repetitive elements, and 0.9% heterozygosity (Fig. 2).

Assembly with HiFiasm [33] generated an initial 2.4 gigabase primary assembly comprising 977 contigs, which was refined to 2.15 gigabase assembly comprising 490 contigs after duplicate purging with purge_dups. Comparison of kmer profiles of initial and purged assemblies suggested successful deduplication of pseudo-haplotype assemblies (Supplemental Fig. 1). Kmer profiling suggested a high-quality assembly with a combined kmer completeness of 98.8% (primary 90.6% and alternate 83.0%) and a combined QV of 60.9 (primary 61.6, alternate 60.3). Merqury [31] screen of purged assembly revealed minor decrement in kmer completeness (primary 89.5%, alternate 87.1%, combined 98.3%) and minor improvement in QV (primary 61.6, alternate 61.3, combined 61.4).

Initial screen of purged assembly with inspector [35] identified a 100% mapping rate of HiFi reads to the purged assembly and 48× coverage. Misassembly correction with three rounds of inspector [35] was able to reduce the number of small-scale errors per megabase

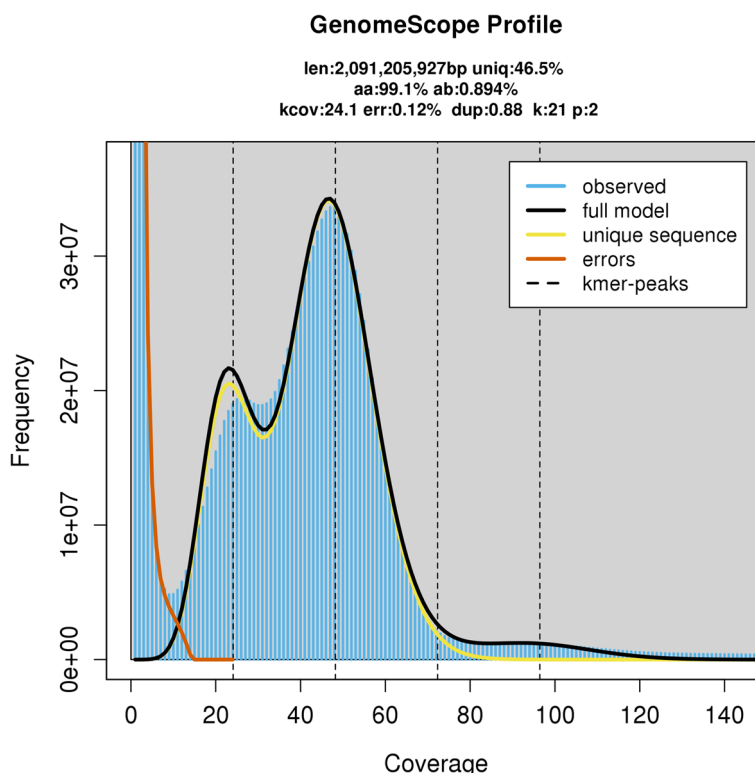


Fig. 2 Genome parameters estimated from kmer profile of HiFi reads with meryl and genomescope v2 using a calculated ideal kmer size of 21. Bimodal distribution is typical of heterozygous genomes. The smaller and larger peaks representing hetero and homozygous kmers respectively, with larger peak at sequencing depth of coverage (48x). Profile analysis estimates a 2.09 gigabase genome, with 46.5% of sequence being non-repetitive and a heterozygosity rate of 0.9%

from 36 to 0.5 and larger structural errors per megabase from 526 to 168 (Supplemental File 1).

Initial scaffolding with ntLink grouped the 490 contigs into 317 scaffolds. Subsequent mounting to the chromosome scale reference assembly of relative *T. amazonica* (GCF_902500255.1) with RagTag reduced the total scaffold count to 62 scaffolds, 31 of which were greater than 1 megabase in length (Fig. 3). Of the 62 final scaffolds, 23 had high sequence similarity and comparable length to chromosome-scale scaffolds of *T. amazonica* (Fig. 4), suggesting a chromosome scale assembly. Pairwise alignment of the final super-scaffolded assembly with that of *T. amazonica* indicated several inverted segments, most notably on scaffolds 4, 9, and 17 (Fig. 5). Screening of the initial and final

assembly for adapters and microbial sequences did not identify any contamination.

The final assembly was highly contiguous, with a total length of 2,151,823,914 bp, a largest contig size of 142,919,290 bp, an N50 of 98,402,768 bp, and an L50 of 10 (Table 1). The final assembly also scored highly in terms of completeness, with 96.1% of Actinopterygii universal single copy orthologs being found as complete and single copy, 1.2% duplicated, 0.9% fragmented, and only 1.8% missing (Fig. 5). This assembly markedly improves upon the current *O. beta* reference assembly (GCA_900660325) in terms of contiguity and completeness, with metrics similar to those of the *T. amazonica* chromosome-scale reference (GCF_902500255.1) (Table 1, Fig. 5). Alignment of the 345,629 contigs in

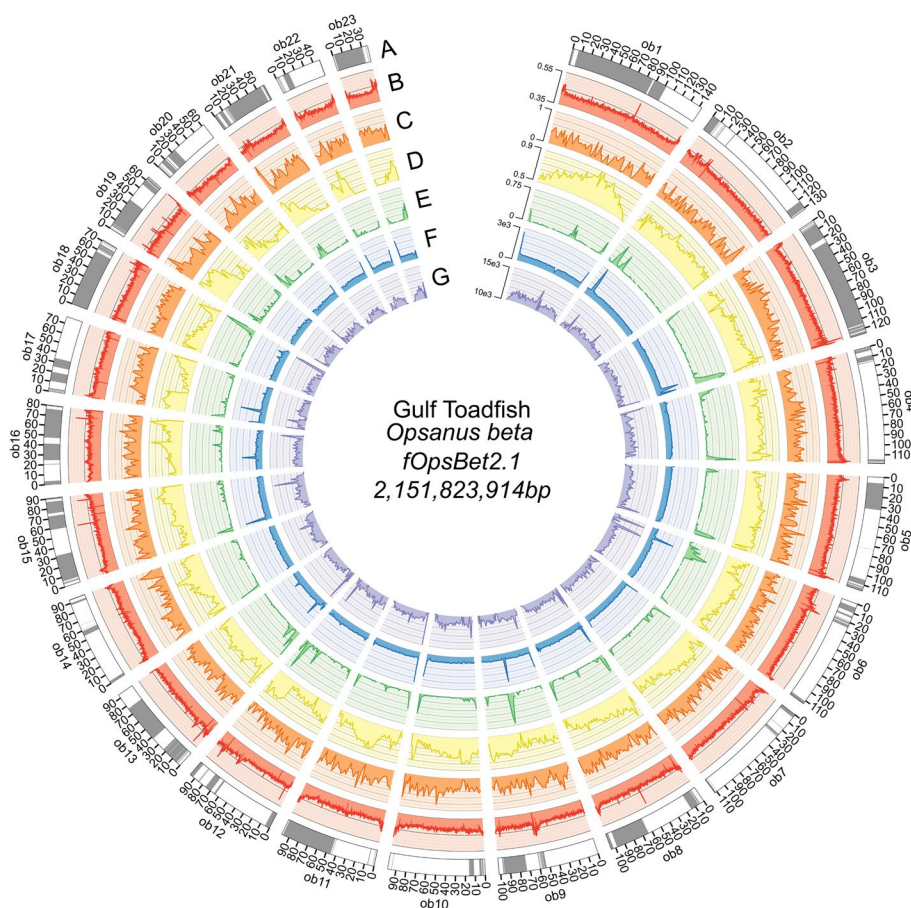


Fig. 3 Circos ideogram of fOpsBet2.1 chromosome-scale genome assembly. **A** 23 chromosome-scale scaffold (ob1-ob23) lengths with contigs represented as alternating grey and white regions. **B** Genome GC content. Y-axis shows 35% (0.35) to 55% (0.55). **C** Gene density calculated using the GFF3 from funannotate::update. Y-axis ranges from 0% (0) to 100% (1) of bases in sliding window. **D** Transposable element (TE) density calculated using the GFF3 generated with RepeatModeler + reclassifier. Y-axis ranges from 50% (0.5) to 90% (0.9) of bases in sliding window. **E** Satellite DNA density calculated using the GFF3 generated by TRASH. Y-axis ranges from 0% (0) to 75% (0.75) of bases in sliding window. **F** Frequency of TTAGGG telomeric satellite calculated with tidk. Y axis shows the counts for each sliding window, with peaks identifying telomeric repeats. **G** Frequency of GATA satellite calculated with tidk. Y-axis shows the counts of GATA satellites for each sliding window. All tracks were generated with a window size of 2Mbp. The 39 unplaced scaffolds are not represented

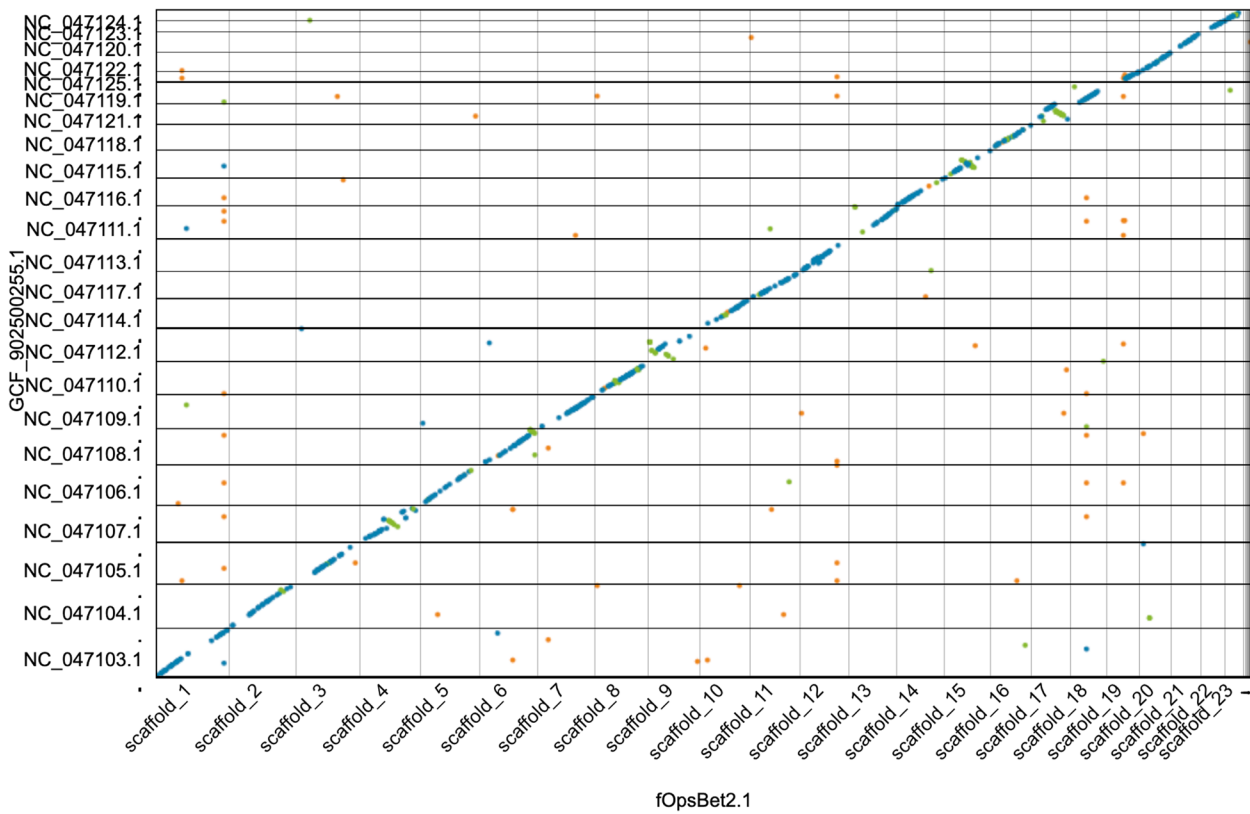


Fig. 4 Dot plot of pairwise alignment of fThaAma1.1 and fOpsBet2.1 chromosome-scale scaffolds. Pairwise alignments are colored blue from forward, green for reverse, and orange for repetitive alignments. Only alignments 4, kilobases and longer are represented. The two assemblies show a high degree of collinearity. In addition, several *O. beta* scaffolds contain large inversions, including scaffolds 4, 9, 15, and 17. Observable gaps in the alignments (such as on scaffold 1) consist primarily of repetitive alignments less than 4 kilobases in length

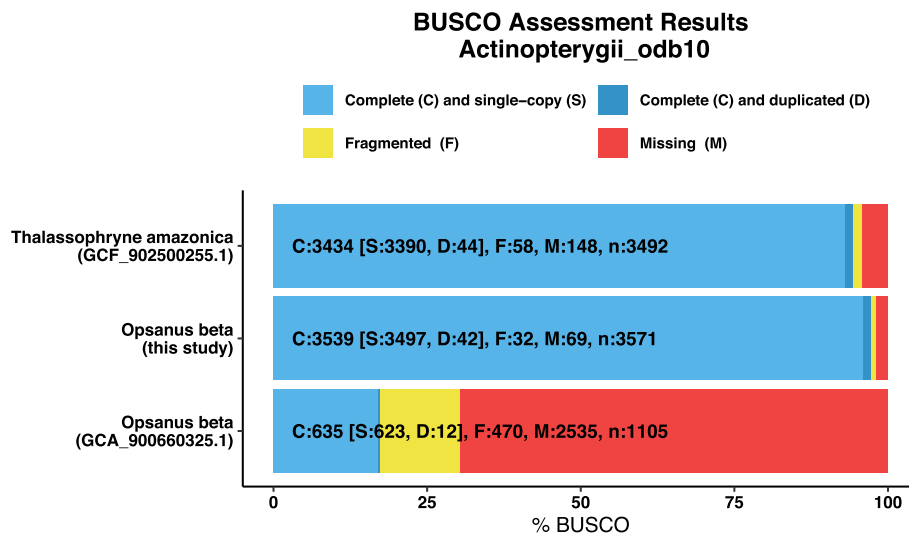


Fig. 5 Number of identified universal single copy orthologs identified by BUSCO analysis in de novo assembly of *Opsanus beta* from this study, current reference assembly for *O. beta*, and chromosome-scale reference assembly of close relative *Thalassophryne amazonica*. The new *O. beta* assembly is the most complete Batrachoididae genome assembly currently available, at 97.3% complete (96.1% single copy, 1.3% duplicated). Light blue = Complete and single copy. Dark Blue = Complete and Duplicated. Yellow = Fragmented. Red = Missing. Analysis run using the Actinopterygii_odb10 database

Table 1 Summary statistics from Quast of the de-novo assembly presented here as compared to current reference assemblies for *O. beta*, and *T. amazonica*. All statistics are generated using only contigs with length greater than 500 bp

	<i>Opsanus beta</i>	<i>Opsanus beta</i>	<i>Thalassophryne amazonica</i>
Assembly	fOpsBet2.1 (this study)	fOpsBet1.1	fThaAma1.1
Accession	JAUHV1000000000	GCA_900660325.1	GCF_902500255.1
# scaffolds	62	126,799	463
Largest scaffold	142,919,290	36,221	175,462,960
Total length	2,151,823,914	1,028,607,202	2,446,591,988
GC (%)	41.99	41.48	41.99
N50	98,402,768	3335	118,421,481
N90	62,083,177	1726	72,818,716
auN	99,667,142.5	3964	114,997,723.2
L50	10	102,666	9
L90	20	272,280	19
# N's per 100 kbp	2.72	52.17	514.77

the current *O. beta* reference to the final assembly with minimap2 resulted in 337,383 (97.6%) primary mapped sequences, with 147,794 being secondary and 20,546 supplemental, for a final mapping rate of 98.6%. Mapping of raw illumina reads used to assemble the current reference (SRR2034069) to the final assembly with bwa-mem2 (v2.2.1) resulted in a mapping rate of 98.4%, whereas mapping of HiFi reads to the final assembly with mimi- nap2 resulted in a mapping rate of 100%.

Mitochondrial genome

MitoHiFi [49, 50] identified the speckled midshipman (*Porichthys myriaster*) as the closest relative with an available mitochondrial genome sequence in GenBank (AP006739.1), which was used to identify candidate mitochondrial reads from HiFi reads.

The primary mitogenome assembly measured 19,381 bp in length and included: two rRNAs (12 s and 16 s), 13 protein coding genes, and 24 tRNAs. In addition to the expected number of mitochondrial genes, the assembly contained two extra phenylalanine tRNAs and D-loop-like control regions than is typical of vertebrate mitogenomes (Fig. 6A). The duplication and atypical arrangement of tRNAs and control region resembled the unique mitogenome organization of other toadfishes, namely that of *P. myriaster* (Fig. 6B) [83].

Running MitoHiFi [49, 50] with the initial primary nuclear genome assembly, as well as the purged alternate assembly as input, generated identical mitogenomes. Comparison with an unverified *O. beta* mitogenome sequence assembled from Illumina short reads in GeneBank (OP056998.1; 19,394 bp) via BLASTN and Clustal Omega showed that the two sequences were 99.85% identical when the unverified sequences was rotated to being at position 13,771 (data not shown).

Transcriptome

Initial IsoSeq processing of raw HiFi RNA reads identified 150,842 high quality transcripts. Cleaning with seqClean trimmed 3262 transcripts and removed 11. Alignment to the super-scaffolded assembly resulted in 143,021 (95%) genome-aligned transcripts, 142,958 of which were longer than 200 bases. PASA [55] assembled transcript alignments into 44,006 gene models. PASA [55] with transdecoder (<https://github.com/TransDecoder/TransDecoder>) identified 48,306 coding domain sequences, 45,602 of which could be propagated to the final genome assembly to be used as input for ab initio gene predictors. Of the predicted coding domain sequences, 40,806 (84%) were marked as complete (containing 5' and 3' UTRs).

Annotation

Transposable Elements

De novo modeling of repetitive elements with Repeat-Modeler identified 4,615 transposable elements, with 1,699 assigned to known families and 2,916 unknown. Curation with repclassifier using the Dfam repeat database for Actinopterygii and known de novo families further improved repeat annotation to 3,026 assigned to known families and 1,589 remaining unknown. Masking of repeats identified 78,515,112 bases (3.6%) as simple repeats, 1,587,765,143 bases (66.5%) as interspersed repeats, and a total repeat content of approximately 70.1%; roughly 20% more of the genome than initial estimates by GenomeScope2 (Fig. 7B) [32]. DNA repeats (22.7%), LINEs (16.8%), and LTRs (13.4%) represented the major components of the repeat landscape, with a further 9.1% of the genome belonging to unclassified repeats. Major contributing classes of repeats included DNA/TcMar (9.9%), LTR/Gypsy (9.3%), LINE/L2 (6.2%), LINE/RTE (6%). Kimura distance-based copy divergence

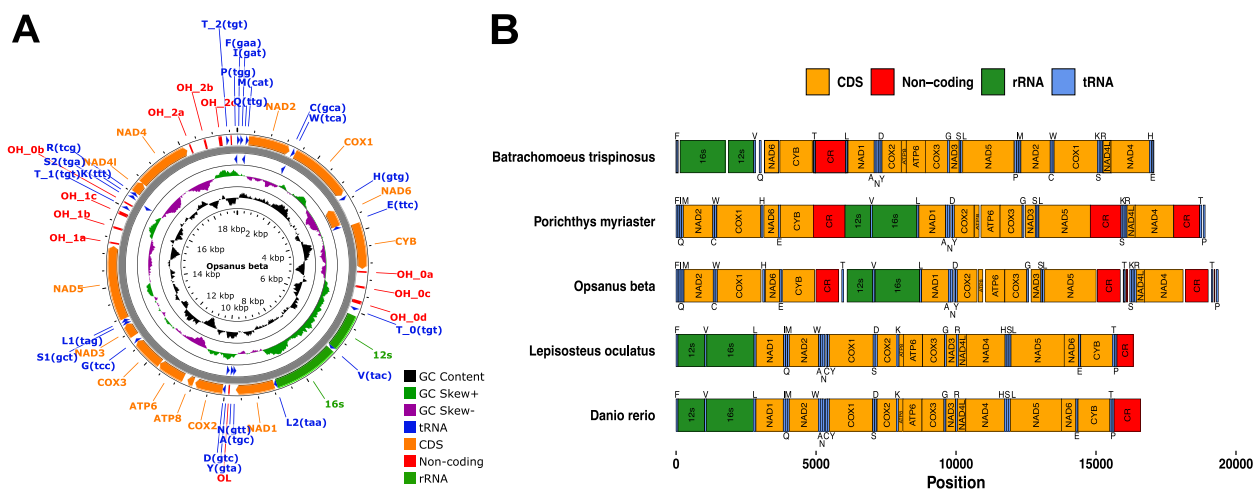


Fig. 6 Mitochondrial genome assembly of *Opsanus beta*. Tracks, moving from outermost to innermost, represent: genomic features of the heavy strand and light strand, GC skew, GC content, and sequence length. Arrows represent genes and their respective orientation on each strand as identified by Mitos2. Labels and features are colored according to their gene type (tRNA (blue), Coding domain sequence (orange), rRNA (green), and non-coding/regulatory features (red). Compared to the typical vertebrate mitochondrial genome, *O. beta* retains three threonine tRNAs (T_0, 1, and 2) and three D-loop like control regions (OH_0, 1, and 2), as opposed to one. Tracks visualize mitochondrial features for *Opsanus beta*, two other Batrachooids (*B. trispinosus* and *P. myriaster*) and two “typical” teleosts (*L. oculatus* as a basal teleost, and *D. rerio* as a model teleost). CDS and rRNA are labeled by their gene symbol within each box, while tRNAs are labeled by the codon product amino acid either above or below for heavy and light strand encoded tRNAs respectively. Grey polygons represent conserved sequence regions as determined by pairwise BLASTn alignments with a minimum eval of 1e-6 and a word size of 7. All mitochondrial sequences were linearized to start at the first base of the Phenylalanine tRNA. Gene order and pairwise alignments demonstrate conserved “typical” vertebrate arrangement in outgroup teleosts and unique Batrachooid arrangement, with highly conserved gene order between *P. myriaster* and *O. beta*. An even more highly derived order was observed in *B. trispinosus* with some conserved gene blocks with the two other Batrachooids as shown previously

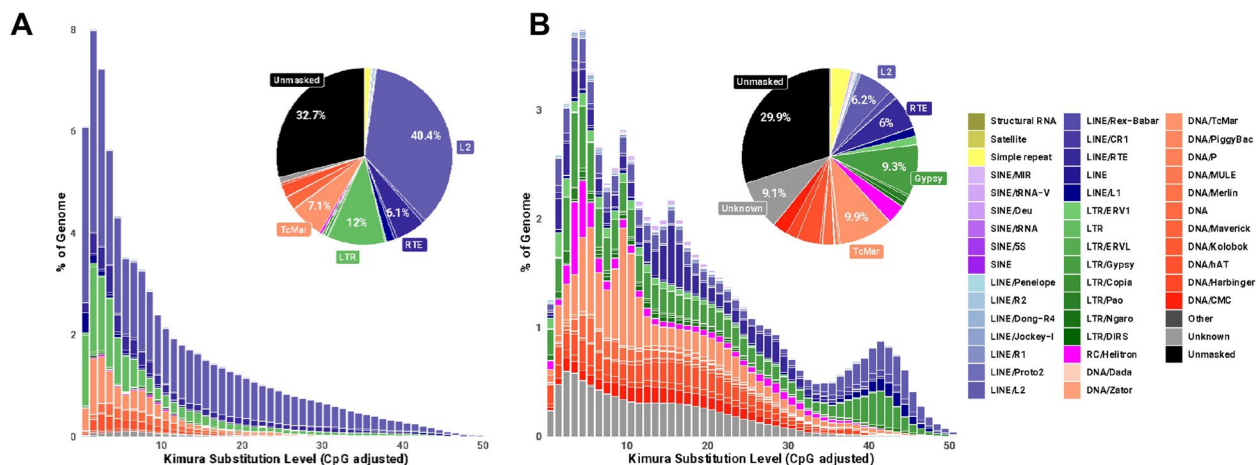


Fig. 7 Repeat landscape of curated de novo repeats in FishTEDB in *Thalassophryne amazonica* (A) and curated repeats in *Opsanus beta* (B). Bar plot: proportion of genome covered by each transposable element (TE) class against Kimura-2 distance parameter, binned by values of 1 from 0 to 50. Smaller/larger Kimura values represent lower/higher divergence from reference, suggesting newer/older repeat divergence, respectively. Inset pie chart: total proportion of genome covered by each repeat class. Repeat classes that cover greater than 5% of the genome are labeled

analysis suggests progressive expansion of TE families in the *O. beta* with recent expansions in DNA/TcMar, LTR/Gypsy, and RC/Helitron classes (Fig. 7B).

To validate the high repeat content of *O. beta* genome, the *T. amazonica* reference assembly was also masked with RepeatMasker (<https://github.com/rmhubble/>

RepeatMasker) using the FishTEDB [61] *T. amazonica* specific repeat library. The *T. amazonica* genome exhibited a similarly high repeat content, with an estimated 72% repeat content (70.5% interspersed). However, the dominant families annotated were distributed differently. While LTRs represented a similar fraction of the genome

at 12.08%, DNA elements represented only 12.17%, and LINEs accounted for 44.48% of the genome (Fig. 7A).

Telomeres

Analysis of scaffolds with the telomere identification toolkit (tidk) using monomers previously used to mark telomeres histologically in other Batrachoids (TTAGGG and GATA, [12]) identified peaks in repeat frequency at the ends of some chromosome-scale scaffolds but not all (Fig. 3). *de novo* search of canonical telomere repeats in raw HiFi reads, primary and alternate assemblies, purged assemblies, and final super-scaffolded assembly identified the canonical 'AACCTT' as common but not the most common among candidate telomere monomers. This suggests sequencing depth was insufficient to adequately penetrate and capture full telomeres.

Satellite DNA

Repeat analysis with TRASH [62] identified 6,722 unique monomer sequences, ranging from 7 to 834 bases long. Peaks of satellite DNA were observed at one or both ends of the 23 chromosome-scale scaffolds, with several scaffolds (*e.g.*, scaffolds 1, 9, and 11) exhibiting satellite peaks in central regions of the

scaffold suggesting possible metacentric centromeres (Fig. 3). For example, the centromere-like array of scaffold 1 spans a region of roughly 5 megabases (80 Mb to 85 Mb) consisting of a core repeat array approximately 570 Kb long (position 82,278,000–82,848,999) composed of ~13,000 copies of a 45 bp monomer, flanked on both sides by interspersed arrays comprised of 128 bp monomers (Fig. 8). While telomere search did not identify consistent telomeric regions in chromosome-scale scaffolds, the presence of satellite peaks at scaffold ends suggests that subtelomeric regions were captured.

Gene Models

Consensus gene prediction using transcript based and *ab initio* predictors via funannotate [66] generated 41,468 gene models, comprising 34,426 mRNAs and 7,042 tRNAs. Updating gene models with funannotate::update [66], using IsoSeq HQ transcripts and publicly available short reads [44], resulted in 41,706 high quality gene models with 38,994 mRNAs and 6,607 tRNAs. Of those, 30,930 CDS were marked as complete, with 17,200 having annotated 5' and 3' UTRs (Supplemental File 1).

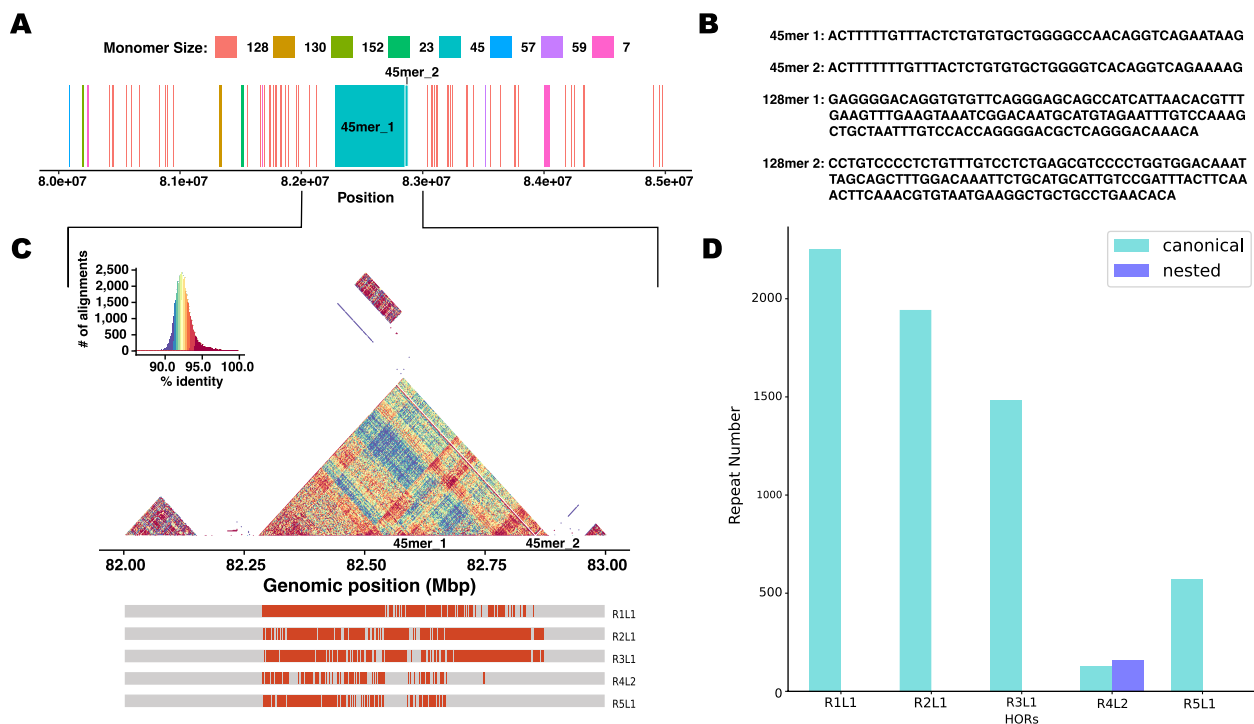


Fig. 8 Putative centromeric region of chromosome-scale super-scaffold Scaffold_1. 5 Mbp region of Scaffold 1 with major clusters of monomers represented as colored blocks, with each color representing a consensus monomer size as identified by TRASH. Consensus sequence from TRASH for most common 45mer and 128mer of putative centromeric and pericentromeric region. Identity heatmap of 82–83Mbp region of Scaffold 1 with occurrence of canonical Higher Order Repeats (HOR) of most common 45mer (45mer 1) as colored strips. Render generated with StainedGlass. Repeat number of canonical HORs of 45mer 1 in putative centromeric region. Analysis regenerated with HiCAT

Functional Annotation

Annotation of protein models with funannotate for GO terms, InterPro [80], eggNOG [78, 79], PFAM, COG, and others resulted in 18,218 (47%) protein models being annotated for all databases and 32,772 (84%) proteins with annotations from at least one (Table 2). Additional annotation of predicted protein models with ghostKO-ALA [81] resulted in 24,914 of 38,994 models (63.9%) assigned a KEGG KO identifier/K number.

Discussion

This new assembly represents a substantial improvement over the currently available *O. beta* NCBI reference assembly in terms of size, contiguity, and completeness. This assembly is the second batrachoid genome assembly generated with modern, long-read technology after *T. amazonica*. Furthermore, our assembly boasts the highest BUSCO completeness score for any batrachoid species, followed by *T. amazonica* (95.6% complete; single: 94.4%, duplicate: 1.2%, fragmented: 1.1%, missing 3.3%) and *C. melanurus* (27.7% complete; single: 27.3%, duplicate: 0.4%, fragmented: 15%, missing 57.3%). By leveraging in silico methods and the high quality *T. amazonica* reference assembly, which was generated with the full gamut of assembly technology (long reads, chromium linked reads, optical maps, and Hi-C chromatin contact maps), we were able to achieve a highly contiguous scaffolded assembly with high completeness scores using only HiFi long-reads at 48X coverage. These marked improvements over the current reference, which was generated using Illumina short read technology, demonstrating the capacity of high fidelity long read technology in penetrating through repetitive and difficult-to-sequence regions to improve assembly efforts. This capacity is critical for highly repetitive genomes such as that of *O. beta*.

Compared to other teleosts, this *O. beta* genome assembly is among the largest assembly available,

numbering 31st out of 1,566 teleost reference assemblies in terms of length on the NCBI at the time of writing. Many of the teleosts with longer genomes than *O. beta* are salmonids. The large size of salmonid genomes is a result of a salmonid-specific whole genome duplication event [84], which is reflected in both salmonid assemblies scoring high in BUSCO duplication rates for teleost single copy orthologs (*O. mykiss*: 38% and *S. salar*: 45%). Conversely, the *T. amazonica* reference and the *O. beta* assembly presented here exhibit low BUSCO duplication scores, suggesting the large size of batrachoid genomes is driven by the observed high proportions of repetitive DNA, consistent with what has been reported from the current, short-read reference assemblies of batrachoids *C. melanurus* and *O. beta* previously [14].

Transposable elements have been demonstrated to play a key role in the evolution of eukaryotic genomes, serving as the raw material for novel protein sequences, regulatory elements, and promoting structural changes that contribute to speciation [85]. In teleosts, increased genome size is associated with increased TE content [86], with expansion of different TE families potentially associated with life history features such as habitat and migratory behavior [86–88]. Transposable elements can become de-repressed as a result of abiotic stress and can drive rapid adaptation towards resilience [89]. In teleosts, such effects have been observed in response to salinity and temperature [90–93]. Mudminnows (Umbridae), which exhibit genome expansion as a result of TE proliferation as we suggest for *O. beta* and other batrachoids, are noted for being very resilient to temperature and oxygen levels [94]. In a survey of 100 teleost genomes, batrachoids exhibited the highest proportion of genomic TE content [14]. Likely as a result of their demersal lifestyle, Toadfishes are also noted for their exceptional resilience to abiotic stressors [95]. *O. beta* is a eurythermal, euryhaline, and hypoxia resistant fish, experiencing swings of salinity (5 ppt to 50 ppt)[3], temperature (18–32C) [96], oxygen concentration [15], and pH (up to pH of 9.4)[96] in their natural habitat. This raises the possibility that the expansion of TE driving *O. beta* genome size may be related to adaptation to abiotic stress. Whether the high TE content in batrachoids is related to their hardy nature remains to be studied, but *O. beta* may serve as a good model to investigate the relationship between TE, abiotic stress, and genome size in teleosts.

Without Hi-C contact maps we could not directly confirm that our super-scaffolds capture full chromosomal sequences. While the karyotype of species within *Opsanus* have not been described, the 2n arrangements of other batrachoids in the genera *Amphychthys*, *Batrachoides*, *Halobatrachus*, *Halophryne*, and *Thalassophryne* is 46, while species in *Porichthys* range between 44 and

Table 2 Protein model annotation statistics

Annotation Database	Protein models annotated
go terms	21,430
interproscan	29,438
eggNOG	32,465
pfam	24,222
cazyme	423
merops	1216
busco	5148
KEGG	24,914
COG	31,703

48 [4–8, 97]. The 2n arrangement within batrachoids is highly constrained, with 2n=46 considered basal and likely retained in *Opsanus* [98]. Given the size and number of the largest 23 scaffolds in our pseudo-haploid super-scaffolded assembly relative to the common karyotype of batrachoids and the size of chromosome sequences in *T. amazonica*, we believe these largest scaffolds represent chromosomal sequences. While the chromosome number is constrained in Batrachoididae, pericentric inversion has been noted to be an important component of karyotypic evolution in batrachoids [98]. Conforming to this trend, several large inversions relative to *T. amazonica* were apparent in *O. beta* scaffolds 4, 9, 15, and 17. Transposable elements are noted to play a role in pericentric inversions in other demersal fish [99], raising the possibility that high TE content in batrachoid genomes may play a role in the previously observed importance of pericentric inversion in batrachoid genome evolution.

Telomeric regions have been identified in batrachoids histochemically using probes matching (TTAGGG)_n telomeric repeats, (GATA)_n satellite DNA, 5 s rDNA, and 18 s rDNA [9, 11, 12]. In silico search for telomeric repeats and (GATA)_n satellites suggests some scaffolds contain at least partial telomeres, but the assembly overall is not at a telomere-to-telomere scale. Concentration of satellite DNA at the terminals of chromosome-scale scaffolds suggests that, while a full telomere-to-telomere scale was not achieved, subtelomeric regions were captured (Fig. 3). Additionally, peaks in satellite DNA concentration at either the center or towards the ends of chromosome-scale scaffolds suggests the capture of metacentric, acrocentric, and telocentric centromere regions.

Mitochondrial gene order in *O. beta* differed significantly from that typical of vertebrates found in other teleosts such as zebrafish or spotted gar [100, 101]. However, mitochondrial genomic rearrangements are not unheard of in other fish lineages: 32 fish families (14% out of 250 examined by Satoh et al.) contain gene order that differs from the typical vertebrate mitochondrial genome [102]. Furthermore, control region duplications similar to those observed in our assembly are known from gulper eels [103], the cyprinid *Rivulus marmoratus* [104], and notothenioid fishes [105] as well as other vertebrate lineages such as snakes [106]. In fact, the organization of the mitochondrial genome in *O. beta* greatly resembles that of closely related Batrachoididae, *P. myriaster* (30). The genomic arrangement is nearly identical save for the two additional duplicated threonine tRNAs that flank the duplicate control regions, as well as small non-coding regions [107]. Given the retention in *O. beta* of toadfish specific

gene blocks first observed in *P. myriaster* and *B. trispinosus* (30), as well as a highly similar but slightly more gene rich arrangement when compared to *P. myriaster*, we propose that the *O. beta* mitochondrial genomic arrangement represents the more ancestral state of Batrachoididae. Thus, the mitochondrial genomic arrangements observed in Porichthynae representative *P. myriaster* arose from gene loss of duplicated tRNAs from the ancestral state, whereas the arrangement observed in Halophryinae representative *B. trispinosus* arose from loss of duplicated control regions and genes as well as further rearrangements.

Conclusion

Here we presented a de novo genome assembly for the Gulf toadfish *Opsanus beta* generated with long read technology that substantially improves over the currently available reference assembly. This assembly represents only the second highly contiguous long read genome in the poorly represented family Batrachoididae. We also highlight the highly repetitive nature and unique mitochondrial arrangement of batrachoids. Given the utility of this model species in diverse fields, it is our hope that this resource aids in the integration of genomics into toadfish research.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-10747-8>.

Supplementary Material 1: Supplemental Figure 1. Merqury kmer profile analysis. Stacked density Kmer profiles of primary and alternate assemblies before (A,B) and after (C,D) purging of duplicate kmers with `purge_dups`. Bimodal distribution with peaks at 25 and 48 represents hetero and homozygous kmers respectively, with larger peak at read coverage depth. Green, purple, and yellow densities in Kmer copy number plots (A,C) represent artificial duplications, which are reduced in large peak after purging (C). Reduction of peak at 48 after purging (D) of primary assembly (red density) indicates improved phasing of pseudo-haplotypes.

Supplementary Material 2: Supplemental Figure 2. Genome assembly and annotation pipeline flowchart. Colored boxes represent data files and white boxes analysis tools. Arrows represent the flow of data. Text in italics represent quality assessment (QA) and quality control (QC) tools used at various stages, with an "*" denoting the suite of tools used for assembly completeness and contiguity assessment. Schematic is divided into major phases of the workflow: kmer profiling of HiFi reads for genome estimates. Genome and mitogenome assembly and QC from HiFi reads, Scaffolding of primary assembly to a super-scaffolded assembly. Modeling and annotation of repetitive elements. Assembly of transcriptome from ISO-seq HiFi reads. Consensus modeling of coding features and annotation using `funannotate` and `interproscan`.

Supplementary Material 3: Supplemental File 1. This file contains the results of various QA/QC tools used throughout the assembly and scaffolding process of fOpsBet2.1 genome assembly. Each tab of the file is titled according to the tool used. The tools include QUAST, gfastats, and `genometools` for assembly size and contiguity metrics, `inspector` for assembly error evaluation, Merqury for QV and kmer completeness, `funannotate` for `genemodel` annotation, and BUSCO for completeness.

Acknowledgements

We would like to thank the DNA Technologies and Expression Analysis Core Laboratory (UC Davis, California), specifically Noravit Chumchim and Dr. Ruta Sahasrabudhe, for their help in preparing and sequencing the DNA and RNA samples for this project. We would also like to thank IDSC at the University of Miami for computational resources used on Pegasus HPC. We would like to thank the editor and reviewers for their time and effort in editing and reviewing the manuscript. Finally, we would like to thank Bic the toadfish for the ultimate sacrifice in the name of science.

Authors' contributions

Study conception: BDY, MDM, and NSK. Wet lab work: MDM. Genomic lab work: BDY. Data analysis: NSK and BDY. Initial manuscript: NSK, BDY, MD, MDM. Manuscript edits: NSK, BDY, MD, MDM. Manuscript was approved by all authors before submission.

Funding

This research was supported by the National Science Foundation grant number IOS-1754550.

Availability of data and materials

All analyses were performed on the University of Miami high performance computing (HPC) cluster, PEGASUS. This HPC operates with linux operating system x86_64 with CentOS v7. All analysis processing scripts are available at [https://github.com/Nicholas-Kron/fOpsBet2.1_assembly]. All data is housed under the NCBI at project number PRJNA944659. Raw PacBio Sequel II subreads were deposited in the NCBI SRA under the accessions SRR24134099, SRR24134098, SRR24065177, SRR24065176, SRR23992258, and SRR23992257. IsoSeq subreads were deposited in the NCBI SRA under accession numbers SRR24283848 and SRR24283847. The *O. beta* genome draft assembly has been deposited at DDBJ/ENA/GenBank under the accession JAUHVI000000000. The *O. beta* mitogenome assembly was deposited in GenBank under accession number JAUHVI000000000. The *O. beta* transcriptome assembly was deposited in TSA under accession number GKQW000000000. The version described in this paper is version JAUHVI010000000. This assembly (fOpsBet2.1_genomic.fa.gz) and additional results for analysis steps can be found in a public repository at Zenodo under the DOI: <https://doi.org/10.5281/zenodo.10246873>.

Declarations

Ethics approval and consent to participate

All protocols were carried out with the approval of the University of Miami Institutional Animal Care and Use Committee (IACUC).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 31 March 2024 Accepted: 29 August 2024

Published online: 18 September 2024

References

- Greenfield D, Winterbottom R, Collette B. Review of the Toadfish Genera (Teleostei: Batrachoididae). 2008.
- Ultsch GR, Jackson DC, Moalli R. Metabolic oxygen conformity among lower vertebrates: The toadfish revisited. *J Comp Physiol.* 1981;142:439–43.
- McDonald MD, Grosell M. Maintaining osmotic balance with an aglomerular kidney. *Comp Biochem Physiol A Mol Integr Physiol.* 2006;143:447–58.
- Nirchio M, Fenocchio A, Swarca A, Dias A, Giuliano—Caetano L, Ron E, Cytogenetic Characterization of *Thalassophryne maculosa* Gunther, et al. (Pisces: Batrachoididae) from Margarita Island. *Venezuela Caribb J Sci.* 1861;2004(40):218–22.
- Palazón Fernández JL, Nirchio M, Sarasquete C. Conventional karyotype and nucleolar organizer regions of the toadfish *Halobatrachus didactylus* (Schneider, 1801) (Pisces: Batrachoididae). *Cariotipo y regiones organizadoras del nucleolo del pez sapo marino Halobatrachus didactylus* (Schneider, 1801) (Pisces: Batrachoididae). 2003.
- de Sousa RPC, Oliveira-Filho AB, Vallinoto M, de Cioffi M B, Molina W, de Oliveira EH, et al. Cytogenetics description in *Batrachoides surinamensis*, (Batrachoididae: Batrachoidiformes): What does the estuary have to say? *Estuar Coast Shelf Sci.* 2018;213:253–9.
- Nirchio M, Turner BJ, Pérez JE, Gaviria JI, Cequea H. Karyotypes of three species of toadfish (Batrachoididae: Teleostei) from Venezuela. *Sci Mar.* 2002;66:1–4.
- Nirchio M, Fenocchio AS, Swarca AC, Pérez JE. Karyology of the toadfish *Porichthys plectrodon* (Jordan and Gilbert, 1882) (Batrachoididae) from Margarita Island. *Venezuela Mar Biol.* 2004;146:161–5.
- Úbeda-Manzanaro M, Merlo MA, Palazón JL, Sarasquete C, Rebordinos L. Sequence characterization and phylogenetic analysis of the 5S ribosomal DNA in species of the family Batrachoididae. *Genome.* 2010;53:723–30.
- Merlo MA, Cross I, Palazón JL, Úbeda-Manzanaro M, Sarasquete C, Rebordinos L. Evidence for 5S rDNA Horizontal Transfer in the toadfish *Halobatrachus didactylus* (Schneider, 1801) based on the analysis of three multigene families. *BMC Evol Biol.* 2012;12:201.
- Úbeda-Manzanaro M, Merlo MA, Palazón JL, Cross I, Sarasquete C, Rebordinos L. Chromosomal mapping of the major and minor ribosomal genes, (GATA)n and U2 snRNA gene by double-colour FISH in species of the Batrachoididae family. *Genetica.* 2010;138:787–94.
- Merlo A, Cross I, Palazón JL, Sarasquete C, Rebordinos L. Chromosomal mapping of the major and minor ribosomal genes, (GATA)n and (TTAGGG)n by one-color and double-color FISH in the toadfish *Halobatrachus didactylus* (Teleostei: Batrachoididae). *Genetica.* 2007;131:195–200.
- Borůvková V, Howell WM, Matoulek D, Symonová R. Quantitative Approach to Fish Cytogenetics in the Context of Vertebrate Genome Evolution. *Genes.* 2021;12:312.
- Reinar WB, Tørresen OK, Nederbragt AJ, Matschiner M, Jentoft S, Jakobsen KS. Teleost genomic repeat landscapes in light of diversification rates and ecology. *Mob DNA.* 2023;14:14.
- McDonald MD, Gilmour KM, Walsh PJ, Perry SF. Cardiovascular and respiratory reflexes of the gulf toadfish (*Opsanus beta*) during acute hypoxia. *Respir Physiol Neurobiol.* 2010;170:59–66.
- Panlilio JM, Marin S, Lobl MB, McDonald MD. Treatment with the selective serotonin reuptake inhibitor, fluoxetine, attenuates the fish hypoxia response. *Sci Rep.* 2016;6:31148.
- Sebastiani J, Sabatelli A, McDonald MD. Mild hypoxia exposure impacts peripheral serotonin uptake and degradation in Gulf toadfish (*Opsanus beta*). *J Exp Biol.* 2022;225:jeb244064.
- Wang Y, Walsh PJ. High ammonia tolerance in fishes of the family Batrachoididae (Toadfish and Midshipmen). *Aquat Toxicol Amst Neth.* 2000;50:205–19.
- Veauvy CM, McDonald MD, Audekerke JV, Vanhouette G, Camp NV, der Linden AV, et al. Ammonia affects brain nitrogen metabolism but not hydration status in the Gulf toadfish (*Opsanus beta*). *Aquat Toxicol.* 2005;74:32–46.
- Barimo JF, Walsh PJ. The effects of acute and chronic ammonia exposure during early life stages of the gulf toadfish. *Opsanus beta Aquat Toxicol.* 2005;75:225–37.
- Grosell N, McDonald MD, Wood CM, Walsh PJ. Effects of prolonged copper exposure in the marine gulf toadfish (*Opsanus beta*): I. Hydromineral balance and plasma nitrogenous waste products. *Aquat Toxicol.* 2004;68:249–62.
- Wood CM, Grosell M, Danielle McDonald M, Playle RC, Walsh PJ. Effects of waterborne silver in a marine teleost, the gulf toadfish (*Opsanus beta*): Effects of feeding and chronic exposure on bioaccumulation and physiological responses. *Aquat Toxicol.* 2010;99:138–48.
- Reddam A, Mager EM, Grosell M, McDonald MD. The impact of acute PAH exposure on the toadfish glucocorticoid stress response. *Aquat Toxicol.* 2017;192:89–96.
- Amador MHB, Schauer KL, McDonald MD. Does fluoxetine exposure affect hypoxia tolerance in the Gulf toadfish, *Opsanus beta*? *Aquat Toxicol Amst Neth.* 2018;199:55–64.

25. Cartolano MC, Berenshtein I, Heuer RM, Pasparakis C, Rider M, Hamerschlag N, et al. Impacts of a local music festival on fish stress hormone levels and the adjacent underwater soundscape. *Environ Pollut*. 2020;265: 114925.
26. Amador MHB, McDonald MD. Molecular and functional characterization of the Gulf toadfish serotonin transporter SLC6A4. *J Exp Biol*. 2018;221 Pt 7:jeb170928.
27. Sebastiani J, McDonald MD. The role of uptake and degradation in the regulation of peripheral serotonin dynamics in Gulf toadfish, *Opsanus beta*. *Comp Biochem Physiol A Mol Integr Physiol*. 2021;258: 110980.
28. Frank L, Serafy J, Grosell M. A large aerobic scope and complex regulatory abilities confer hypoxia tolerance in larval toadfish, *Opsanus beta*, across a wide thermal range. *Sci Total Environ*. 2023;899:165491.
29. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021;592:737–46.
30. Leary S, Underwood W, Anthony R, Cartner S, Grandin C, Gwaltney-Brant S, et al. AVMA Guidelines for the Euthanasia of Animals: 2020 Edition. 2020.
31. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:245.
32. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11:1432.
33. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–5.
34. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36:2896–8.
35. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol*. 2021;22:312.
36. Coombe L, Warren RL, Wong J, Nikolic V, Birol I. ntLink: A Toolkit for De Novo Genome Assembly Scaffolding and Mapping Using Long Reads. *Curr Protoc*. 2023;3:e733.
37. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. 2022;23:258.
38. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol*. 2017;34:1812–9.
39. Mikheenko A, Pribelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QAST-LG. *Bioinformatics*. 2018;34:i142–50.
40. Mikheenko A, Valin G, Pribelski A, Saveliev V, Gurevich A. Icarus: visualizer for de novo assembly evaluation. *Bioinformatics*. 2016;32:3321–3.
41. Gremme G, Steinbiss S, Kurtz S. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Trans Comput Biol Bioinform*. 2013;10:645–56.
42. Formenti G, Abueg L, Brajuka A, Brajuka N, Gallardo-Alba C, Giani A, et al. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics*. 2022;38:4214–6.
43. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol*. 2021;38:4647–54.
44. Schauer KL, LeMoine CMR, Pelin A, Corradi N, McDonald MD, Warren WC, et al. A proteinaceous organic matrix regulates carbonate mineral production in the marine teleost intestine. *Sci Rep*. 2016;6:34494.
45. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
46. Chen Y, González-Pech RA, Stephens TG, Bhattacharya D, Chan CX. Evidence That Inconsistent Gene Prediction Can Mislead Analysis of Dinoflagellate Genomes. *J Phycol*. 2020;56:6–10.
47. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20:257.
48. Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol*. 2024;25:60.
49. Uliano-Silva M, Ferreira JGRN, Krashenninnikova K, Consortium DT of L, Formenti G, Abueg L, et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads. *BMC Bioinformatics*. 2023. <https://doi.org/10.1186/s12859-023-05385-y>.
50. Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Resour*. 2020;20:892–905.
51. Donath A, Jühling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, et al. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Res*. 2019;47:10543–52.
52. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.
53. Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C, et al. Proksee: in-depth characterization and visualization of bacterial genomes. *Nucleic Acids Res*. 2023;51:W484–92.
54. Kent WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Res*. 2002;12:656–64.
55. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 2003;31:5654–66.
56. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci*. 2020;117:9451–7.
57. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2023;51:D523–31.
58. Hubley R. rmhubley/RepeatMasker. 2024.
59. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA*. 2021;12:2.
60. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
61. Shao F, Wang J, Xu H, Peng Z. FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. *Database*. 2018;2018:bax106.
62. Wlodzimierz P, Hong M, Henderson IR. TRASH: Tandem Repeat Annotation and Structural Hierarchy. *Bioinformatics*. 2023;39:btad308.
63. Gao S, Yang X, Guo H, Zhao X, Wang B, Ye K. HiCAT: a tool for automatic annotation of centromere structure. *Genome Biol*. 2023;24:58.
64. Vollger MR, Kerpedjiev P, Phillippy AM, Eichler EE. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics*. 2022;38:2049–51.
65. Brown M, González De la Rosa PM, Mark B. 'A Telomere Identification Toolkit'. Zenodo. 2023. <https://doi.org/10.5281/zenodo.10091385>.
66. Palmer JM, Stajich J. Funannotate v1.8.1: Eukaryotic genome annotation. 2020.
67. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 2008;24:637–44.
68. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5:59.
69. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. 2004;20:2878–9.
70. Lomsadze A, Ter-Hovhannissyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 2005;33:6494–506.
71. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 2008;9:R7.
72. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res*. 2021;49:9077–96.
73. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49:D412–9.
74. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2012;40:W445–51.
75. Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, et al. MIBiG 3.0: a community-driven effort to annotate

- experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* 2023;51:D603–10.
76. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* 2018;46:D624–32.
 77. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods.* 2021;18:366–8.
 78. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47 Database issue:D309–14.
 79. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol.* 2021;38:5825–9.
 80. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40.
 81. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol.* 2016;428:726–31.
 82. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol.* 2004;5:R12.
 83. Miya M, Satoh TP, Nishida M. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. *Biol J Linn Soc.* 2005;85:289–306.
 84. Dimos B, Phelps M. A homology guide for Pacific salmon genus *Oncorhynchus* resolves patterns of ohnolog retention, resolution and local adaptation following the salmonid-specific whole-genome duplication event. *Ecol Evol.* 2023;13: e9994.
 85. Almojil D, Bourgeois Y, Falis M, Hariyani I, Wilcox J, Boissinot S. The Structural, Functional and Evolutionary Impact of Transposable Elements in Eukaryotes. *Genes.* 2021;12:918.
 86. Shao F, Han M, Peng Z. Evolution and diversity of transposable elements in fish genomes. *Sci Rep.* 2019;9:15399.
 87. Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, et al. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics.* 2018;19:141.
 88. Carotti E, Carducci F, Canapa A, Barucca M, Greco S, Gerdol M, et al. Transposable Elements and Teleost Migratory Behaviour. *Int J Mol Sci.* 2021;22:602.
 89. Colonna Romano N, Fanti L. Transposable Elements: Major Players in Shaping Genomic and Evolutionary Patterns. *Cells.* 2022;11:1048.
 90. Carotti E, Carducci F, Canapa A, Barucca M, Biscotti MA. Transposable Element Tissue-Specific Response to Temperature Stress in the Stenothermal Fish *Puntius tetrazona*. *Animals.* 2023;13:1.
 91. Carotti E, Carducci F, Greco S, Gerdol M, Di Marino D, Perta N, et al. Transcriptional Contribution of Transposable Elements in Relation to Salinity Conditions in Teleosts and Silencing Mechanisms Involved. *Int J Mol Sci.* 2022;23:5215.
 92. Carducci F, Biscotti MA, Forconi M, Barucca M, Canapa A. An intriguing relationship between teleost Rex3 retroelement and environmental temperature. *Biol Lett.* 2019;15:20190279.
 93. Carducci F, Barucca M, Canapa A, Carotti E, Biscotti MA. Mobile Elements in Ray-Finned Fish Genomes. *Life.* 2020;10:221.
 94. Lehmann R, Kovařík A, Ocalewicz K, Kirtiklis L, Zuccolo A, Tegner JN, et al. DNA Transposon Expansion is Associated with Genome Size Increase in Mudminnows. *Genome Biol Evol.* 2021;13:evab228.
 95. Molina JM, Kunzmann A, Reis JP, Guerreiro PM. Metabolic Responses and Resilience to Environmental Challenges in the Sedentary Batrachoid *Halobatrachus didactylus* (Bloch & Schneider, 1801). *Animals.* 2023;13:632.
 96. Barimo JF, Serafy JE, Frezza PE, Walsh PJ. Habitat use, urea production and spawning in the gulf toadfish *Opsanus beta*. *Mar Biol.* 2007;150:497–508.
 97. Brum MJJ, Affonso PR A M, Mota LCG. Cytogenetic characterization of *Porichthys porosissimus* (Valenciennes, 1857) (Batrachoididae, Batrachoidiformes) from the Rio de Janeiro Coast, Brazil. *Chromosome Sci.* 2001;5:15–8.
 98. Costa GWWF, Molina WF. Karyoevolution of the toadfish *Thalassophryne nattereri* (Batrachoidiformes: Batrachoididae). *Genet Mol Res GMR.* 2009;8:1099–106.
 99. Rodríguez ME, Cross I, Arias-Pérez A, Portela-Bens S, Merlo MA, Liehr T, et al. Cytogenomics Unveil Possible Transposable Elements Driving Rearrangements in Chromosomes 2 and 4 of *Solea senegalensis*. *Int J Mol Sci.* 2021;22:1614.
 100. The Complete Sequence of the Zebrafish (*Danio rerio*) Mitochondrial Genome and Evolutionary Patterns in Vertebrate Mitochondrial DNA. <https://genome.cshlp.org/content/11/11/1958.short>. Accessed 30 Jan 2024.
 101. Animal mitochondrial genomes - PubMed. <https://pubmed.ncbi.nlm.nih.gov/10101183/>. Accessed 30 Jan 2024.
 102. Satoh TP, Miya M, Mabuchi K, Nishida M. Structure and variation of the mitochondrial genome of fishes. *BMC Genomics.* 2016;17:719.
 103. Inoue JG, Miya M, Tsukamoto K, Nishida M. Evolution of the Deep-Sea Gulper Eel Mitochondrial Genomes: Large-Scale Gene Rearrangements Originated Within the Eels. *Mol Biol Evol.* 2003;20:1917–24.
 104. Lee J-S, Miya M, Lee Y-S, Kim CG, Park E-H, Aoki Y, et al. The complete DNA sequence of the mitochondrial genome of the self-fertilizing fish *Rivulus marmoratus* (Cyprinodontiformes, Rivulidae) and the first description of duplication of a control region in fish. *Gene.* 2001;280:1–7.
 105. Minhas BF, Beck EA, Cheng C-HC, Catchen J. Novel mitochondrial genome rearrangements including duplications and extensive heteroplasmy could underlie temperature adaptations in Antarctic notothenioid fishes. *Sci Rep.* 2023;13:6939.
 106. Kumazawa Y, Ota H, Nishida M, Ozawa T. Gene rearrangements in snake mitochondrial genomes: highly concerted evolution of control-region-like sequences duplicated and inserted into a tRNA gene cluster. *Mol Biol Evol.* 1996;13:1242–54.
 107. Miya M, Satoh TP, Nishida M. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences: PHYLOGENETIC POSITION OF TOADFISHES. *Biol J Linn Soc.* 2005;85:289–306.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.