

Near chromosome-level and highly repetitive genome assembly of the snake pipefish *Entelurus aequoreus* (Syngnathiformes: Syngnathidae)

Magnus Wolf^{1,2,3,*}, Bruno Lopes da Silva Ferrette¹, Raphael T. F. Coimbra^{1,2}, Menno de Jong¹, Marcel Nebenführ^{1,2}, David Prochotta^{1,2}, Yannis Schöneberg^{1,2}, Konstantin Zapf^{1,2}, Jessica Rosenbaum², Hannah A. Mc Intyre², Julia Maier², Clara C. S. de Souza², Lucas M. Gehlhaar², Melina J. Werner², Henrik Oechler², Marie Wittekind², Moritz Sonnewald⁴, Maria A. Nilsson^{1,5}, Axel Janke^{1,2,5} and Sven Winter^{1,2,6,*}

- 1 Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt am Main, Germany
- 2 Institute for Ecology, Evolution, and Diversity, Goethe University, Frankfurt am Main, Germany
- 3 Institute for Evolution and Biodiversity, University of Münster, Münster, Germany
- 4 Senckenberg Research Institute, Department of Marine Zoology, Section Ichthyology, Frankfurt am Main, Germany
- 5 LOEWE-Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany
- 6 Research Institute of Wildlife Ecology, University of Veterinary Medicine, Vienna, Austria

ABSTRACT

The snake pipefish, *Entelurus aequoreus* (Linnaeus, 1758), is a northern Atlantic fish inhabiting open seagrass environments that recently expanded its distribution range. Here, we present a highly contiguous, near chromosome-scale genome of *E. aequoreus*. The final assembly spans 1.6 Gbp in 7,391 scaffolds, with a scaffold N50 of 62.3 Mbp and L50 of 12. The 28 largest scaffolds (>21 Mbp) span 89.7% of the assembly length. A BUSCO completeness score of 94.1% and a mapping rate above 98% suggest a high assembly completeness. Repetitive elements cover 74.93% of the genome, one of the highest proportions identified in vertebrates. Our demographic modeling identified a peak in population size during the last interglacial period, suggesting the species might benefit from warmer water conditions. Our updated snake pipefish assembly is essential for future analyses of the morphological and molecular changes unique to the Syngnathidae.

Submitted: 20 October 2023
Accepted: 08 January 2024
Published: 11 January 2024

* Corresponding authors. E-mail: Magnus.Wolf@senckenberg.de; Sven.Winter@senckenberg.de

Published by GigaScience Press.

Preprint submitted at <https://doi.org/10.1101/2023.12.12.571260>

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Gigabyte, 2024, 1–13

Subjects Genetics and Genomics, Evolutionary Biology, Marine Biology

INTRODUCTION

The snake pipefish *Entelurus aequoreus* (Linnaeus 1758) is a member of the family Syngnathidae, which currently includes over 300 species of seahorses and pipefishes [1]. *E. aequoreus* shares typical features with other pipefishes, such as the unique, elongated body plan and fused jaws [2]. However, unlike most pipefishes, which are found in benthic habitats, the snake pipefish inhabits more open and deeper seagrass environments and occurs even in pelagic waters [2]. They are ambush predators on small crustaceans and

other invertebrates, thereby indirectly contributing to the overall biodiversity and stability of these fragile habitats [3]. Adult snake pipefishes are poor swimmers equipped with small fins. They rely on their elongated, thin bodies for crypsis in eelgrass habitats [4–6].

The snake pipefish historically ranged from the waters of the Azores northwards to the waters of Norway and Iceland and eastward to the Baltic Sea [7]. However, since 2003, the species has expanded its distribution [8] into the Arctic waters of Spitsbergen [9], the Barents Sea, and the Greenland Sea [10]. Simultaneously, population sizes seem to increase within its former range, as indicated by substantially increased catch rates [11, 12]. Several factors have been proposed to cause this expansion and population growth, including rising sea temperatures, an increased potential for long-distance dispersal of juveniles via ocean currents [4, 7], and an increased reproductive success facilitated by the dispersal of invasive seaweeds [6, 8–10, 13]. The latter explanation has been confirmed by local field experiments in the northern Wadden Sea, suggesting a mutual co-occurrence of the invasive Japanese seaweed (*Sargassum muticum*) and the snake pipefish [5]. Studies based on mtDNA marker regions did not discern any population structure thus far and suggest a previous population expansion in the Pleistocene, around 50–100 thousand years ago (kya) [6]. Yet, a comprehensive analysis of demographic events is better conducted using genomic data, thus requiring a high-quality reference genome, ideally of the same species or at least a closely related one.

Previously, genomes of Syngnathidae have been used to study the evolution of highly specialized morphologies and life-history traits unique to pipefishes and seahorses [14–16]. For instance, the transition to male pregnancy was associated with major genomic restructuring events and parallel modifications of the adaptive immune system. There is a remarkable variability in genome sizes within the family, with estimates ranging from 350 Mbp to 1.8 Gbp [14]. The major shifts in body shape are assumed to be related to gene-family loss and expansion events, along with higher rates of protein and nucleotide evolution [16]. Genomic data obtained using a direct sequencing approach of ultra-conserved elements improved the understanding of the phylogeny of pipefishes [15] and identified a likely radiation of the group in the waters of the modern Indo-Pacific Ocean. Nevertheless, high-quality genomes of Syngnathidae are only available for a few species. According to the NCBI genome database, only 7% of the known species diversity has genome sequences available.

A draft genome of the snake pipefish was previously assembled using a combination of paired-end and mate-pair sequencing techniques, yielding an assembly with low continuity (N50 3.5 kbp, BUSCO C: 21%) and a large difference between the estimated and assembled genome sizes (1.8 Gbp vs. 557 Mbp) [14]. To obtain a higher quality, near chromosome-scale genome assembly for the snake pipefish, essential for future population, conservation, and evolutionary genomics studies of fish, we used long-read sequencing technologies. This allowed us to gain insights into the genetic properties of the species and to perform demographic analyses based on the Pairwise Sequentially Markovian Coalescent (PSMC) framework [17]. The data generation and analyses presented here were conducted during a six-week master course in 2021 at the Goethe University, Frankfurt am Main, Germany. The concept of high-quality genome sequencing in a course setting has so far yielded three reference-quality fish genomes and has proven to be a successful approach to introducing the technology to a new generation of scientists [18–21].

Table 1. Summary statistics of the snake pipefish reference genome. The table includes information for (A) the raw read sequencing, (B) the scaffold- and contig-level *de novo* assembly, and (C) the BUSCO completeness statistics.

(A) Raw read statistics		
No. short reads	264,111,731	
Mapped short reads (%)	99.53	
Mean short read coverage (x)	23	
No. long reads	130,590,372	
Mapped long reads (%)	98.61	
Mean long-read coverage (x)	205.2	
(B) Assembly statistics (scaffold/contig)		
No. scaffolds/contigs	7,387	7,473
No. scaffolds/contigs (>50 kbp)	466	526
Scaffold/contig L50	12	14
Scaffold/contig N50 (bp)	62,341,166	45,010,074
Total length (bp)	1,662,053,046	1,662,035,846
GC (%)	38.87	38.87
No. of N's per 100 kb	1.03	0.0
Heterozygosity (%)	0.387	
Total interspersed repeats (bp)	1,237,929,559 (74.93%)	
(C) BUSCO completeness		
Clade: <i>Actinopterygii</i>	C: 94.1% [S: 92.6%, D: 1.5%]	
	F: 2.0%, M: 3.9%	
	n: 3,640	
BUSCO: Benchmarking Universal Single-Copy Orthologs (65); C, complete; S, single copy; D, duplicated; F, fragmented; M, missing.		

RESULTS AND DISCUSSION

Genome sequencing and assembly

PacBio’s continuous long read (CLR) technology generated 401 Gbp of long-read data in ~60 million reads with an N50 of 7.9 kb (Table 1). Illumina sequencing yielded 38 Gbp of standard short-read data in approximately 257 million reads with a mean length of 148 bp after filtering. Sequencing of the Omni-C library generated 54.7 Gbp of raw short-read data.

The snake pipefish’s genome was assembled *de novo* to a total size of 1.7 Gbp. It consisted of 2,204 scaffolds, with a scaffold N50 of 62 Mbp and an L50 of 11 (Table 1, Figure 1A). The finalized assembly has 1.0 Ns per 100 kbp and a GC content of 38.84%. Our BUSCO (RRID:SCR_015008) completeness assessment resulted in 94.1% complete core genes, based on the *actinopterygii_obd10* set, showing the high completeness of the assembly. Both long- and short-read data mapped onto the assembly with high mapping rates of 98.6% and 99.5%, respectively. HI-C mapping resulted in 28 larger scaffolds (Figure 1B), indicating the near-chromosome level of the *de novo* assembly. This result aligns with past karyotype estimations of other pipefish and seahorses, predicting 22 and 22-24 chromosomes, respectively [22–24]. The rest of the genome comprises only smaller scaffolds and contigs, which may result from the high amounts of repetitive regions, as detailed in the following section. Our Blobtools (RRID:SCR_017618) analysis of both long- and short-read data (Figure 1C and D) found no apparent signs of contamination. However, we detected and removed background noise of unknown origin in both datasets.

Variant calling identified ~301 million sites (including monomorphic sites), with ~1.3 million found to be biallelic. Genome-wide heterozygosity was determined to be 0.387%, which is in line with other fish species [25, 26]. The GenomeScope (RRID:SCR_017014) results based on short reads suggested a haploid genome size of 1.15 Gbp and an expected

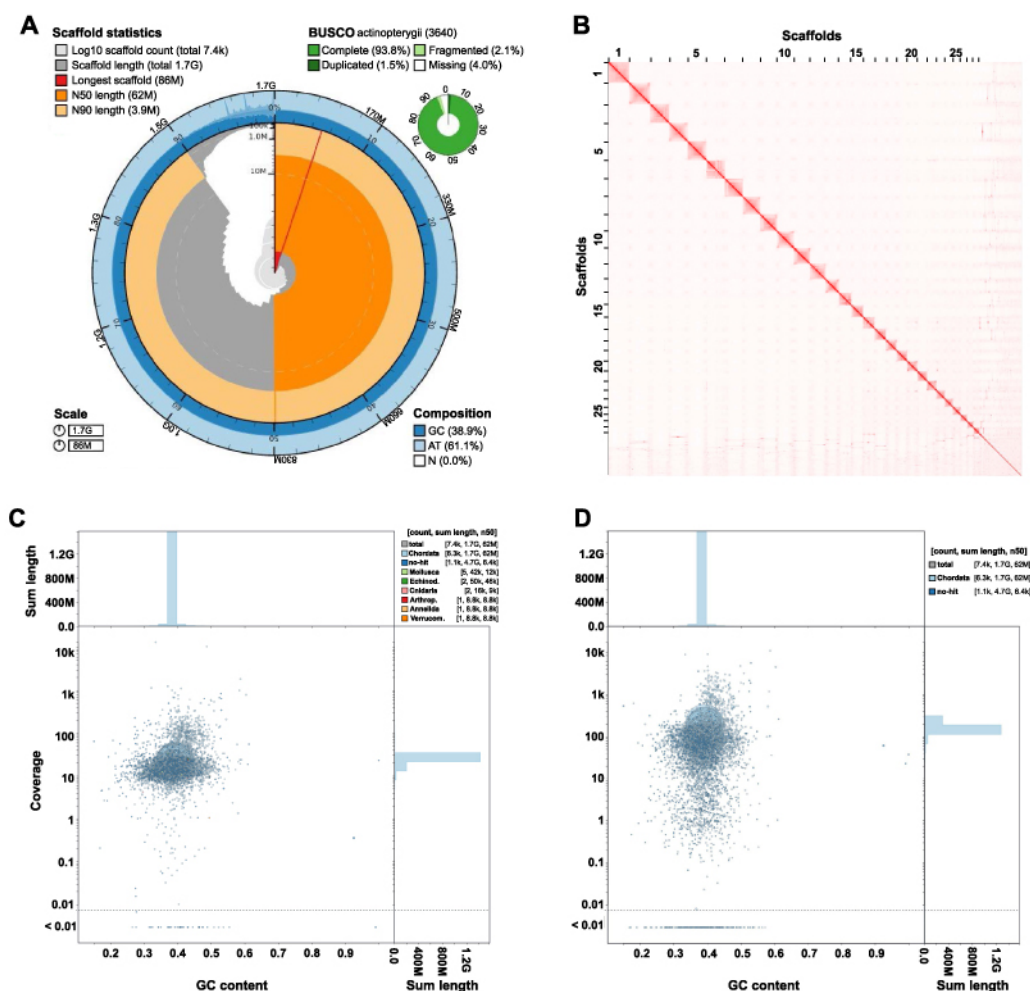


Figure 1. Assembly characteristics and quality assessments of the *de novo* *Entelurus aequoreus* genome. (A) The snail plot summarizes different assembly properties. Scaffold statistics are depicted in the innermost circle, and colors red to orange represent the longest scaffold N50 and N90, respectively. GC composition is shown in the outer blue circle. BUSCO completeness statistics are depicted in the small green circle. (B) Omni-C contact density map indicating 28 larger scaffolds and the near-chromosome level of the assembly. (C,D) The BlobPlot analysis compares GC content (x-axis), assembly coverage (y-axis), and taxonomic BLAST assignments of contigs (color) for both the Omni-C short reads (C) and PacBio long reads (D).

genome-wide heterozygosity of 1%. These estimates were around 362 Mbp shorter and 0.57% more heterozygous than the final assembly. This, again, might be explained by the high repeat content of the genome.

Annotation

In total, 0.9 Gbp, or 74.93%, of the entire assembly, were identified as repetitive during our *de novo* repeat-modeling (using RepeatModeler, RRID:SCR_015027) and repeat-masking (using RepeatMasker, RRID:SCR_012954) as shown in Figure 2. This high repeat content contrasts with other fish genomes [27]. However, it is similar, although at a smaller scale, to the closest relative, *Nerophis ophidion* (65.7%) [14], and other syngnathid fish genomes, such as seadragons [28]. The first draft assembly of the snake pipefish had a repeat content of

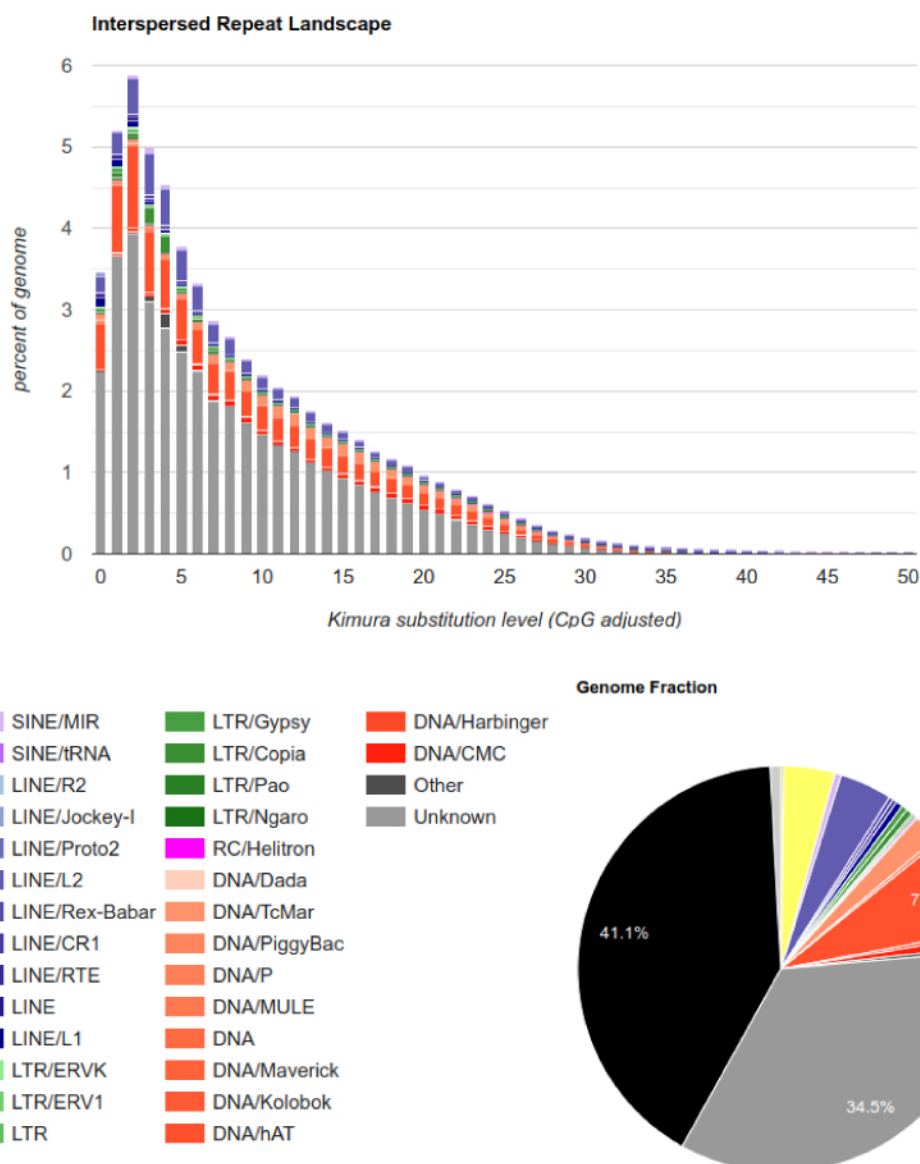


Figure 2. Repeat landscape of the *de novo* *Entelurus aequoreus* genome. Colors represent different types of RE, and gray areas indicate unclassified types of repetitive regions.

57.2% [14], and our improved long-read assembly identified 17.7% additional repeats that were missing from the previous assembly [14]. So far, among vertebrates, only the lungfish *Neoceratodus forsteri* [29] has more transposable elements (TEs) than the snake pipefish.

The annotation of the genome, featuring *de novo* and homology-based identification approaches, resulted in 33,202 genes with an average length of 13,828 bp. Each gene had, on average, 7.32 exons and 6.25 introns with average lengths of 188 bp and 2,240 bp, respectively. In total, we identified 243,038 exons and 207,467 introns within our annotation. The total number of genes is ~30% higher compared to other annotated genomes in the order Syngnathiformes, such as 23,458 for the tiger tail seahorse (*Hippocampus comes*) [16]

or 24,927 for the greater pipefish (*Syngnathus acus*) [30] made by the NCBI Eukaryotic Genome Annotation pipeline. Notably, as these two genomes are considerably smaller (492 Mbp and 324 Mbp, respectively), we can assume that the large-scale genome increase in this species included many coding sequences. The high content of repetitive regions and the lack of transcriptomic data might also have increased the number of false positive gene-calls; however, our BUSCO completeness analysis of the predicted proteins resulted in 82.6% complete sequences, with only 6.8% duplicated ones. Additionally, 5.3% of the coding sequences appeared fragmented, and 12.1% were missing from the *actinopterygii_obd10* OrthoDB set. Our functional annotation resulted in hits for 89% of the predicted proteins.

Demographic inference

The demographic inference analysis of the snake pipefish genome (Figure 3) using the PSMC framework [17] traced population changes over the past 1 million years. Given the chosen substitution rate and generation time, there was a steady increase in the effective population size (N_e), starting at 15 thousand individuals 1 million years ago, which peaked at an N_e of 250 thousand individuals 100 kya. Thereafter, N_e decreased until reaching 30 thousand individuals at 10 kya and stagnated until the end of the model. The previously suggested population expansion during the Pleistocene (50–100 kya) was therefore confirmed by this model. However, the population expansion was followed by another population decline that was not resolved by Braga Goncalves *et al.* [6]. Our result may point to a conclusion different from that drawn by the authors. This is because the snake pipefish might have inhabited a comparably small population during the Holocene and only recently expanded its distribution. This expansion resulted in a large population with a high degree of homogenization, as observed by Braga Goncalves and colleagues [6]. Given that the presented peak in population size parallels with the last interglacial period between the Penultimate Glacial Period (135–192 kya [31]) and the last glacial period (present – 20 kya [32]), we assume that the snake pipefish largely benefitted from the warmer water conditions during the interglacial period, as seen in the present range expansion.

MATERIAL AND METHODS

Sampling, DNA extraction, and sequencing

A single individual of *Entelurus aequoreus* (Linnaeus 1758) (NCBI: [txid42861](#), [marinespecies.org:taxname: 127379](#)) was caught by trawling during an annual monitoring expedition to the Dogger Bank in the North Sea in July 2021 (trawl start coordinates 54.993633, 2.940833; end coordinates 55.0077, 2.929867) with the permission of the Maritime Policy Unit of the UK Foreign and Commonwealth Office. The study complied with the ‘Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization’. The sample was initially frozen at –20 °C and later stored at –80 °C.

High-molecular-weight genomic DNA was extracted from muscle tissue, following the protocol by Mayjonade *et al.* [33], with the addition of Proteinase K. We evaluated the quantity and quality of the DNA with the Genomic DNA ScreenTape on the Agilent 2200 TapeStation system (Agilent Technologies), as well as with the Qubit® dsDNA BR Assay Kit.

For long-read sequencing, a PacBio SMRT Bell CLR library was prepared using the SMRTbell Express Prep kit v3.0 kit (Pacific Biosciences – PacBio, Menlo Park, CA, USA) and sequenced on the PacBio Sequel IIe platform. A proximity-ligation library was compiled

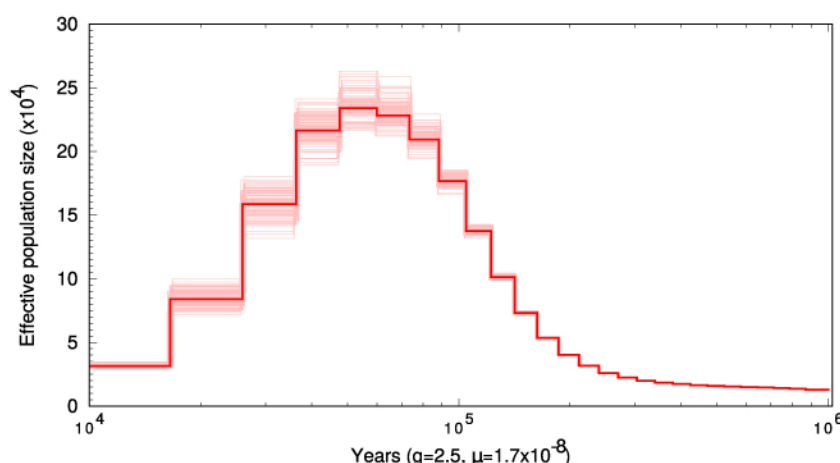


Figure 3. Demographic history of the snake pipefish estimated using the PSMC framework. Using a generation time of 2.5 years [72] and a substitution rate of 1.7×10^{-8} per site per generation [71], a model was created covering the last 10 kya to 1 Mya. The x-axis represents time in number of years ago and the y-axis shows the effective population N_e size in tens of thousands of individuals. The model indicates a peak in N_e of 250 thousand individuals during the Pleistocene around 100 thousand years ago.

with muscle tissue following the Dovetail™ Omni-C protocol (Dovetail Genomics, Santa Cruz, California, USA). In addition, a standard whole-genome 150 bp paired-end Illumina library was prepared using the NEBNext Ultra II library preparation kit (New England Biolabs Inc., Ipswich, USA). Finally, the proximity ligation and the paired-end library were shipped to Novogene (UK) for sequencing on the Illumina NovaSeq 6000 platform (RRID:SCR_016387).

Pre-processing and genome size estimation

The PacBio subreads were converted from BAM into FASTQ format using the PacBio Secondary Analysis Tool BAM2fastx v.1.3.0 [34]. Quality control, trimming, and filtering of the Illumina reads were performed using fastp v0.23.1 (RRID:SCR_016962) [35] with the settings “-g -3 -l 40 -y -Y 30 -q 15 -u 40 -c -p -j -h -R -w N”. To estimate the genome size of the snake pipefish, we performed *k*-mer profiling using the standard short-read Illumina data. We first ran Jellyfish v2.3.0 (RRID:SCR_005491) [36] to generate a histogram of *k*-mers with a length of 21 bp. Subsequently, we used this data to obtain a genome profile using GenomeScope v2.0 (RRID:SCR_017014) [37]. We further tested alternative *k*-mer lengths between 17- and 25-mers. No significant differences in the estimated genome size were detected except for the 17-mer, which resulted in a smaller genome size estimation of ~500 Mbp.

Genome assembly and polishing

We assembled the genome from the PacBio long-read data using WTDBG v.2.5 (RRID:SCR_017225) [38]. The resulting assembly was first polished using the PacBio data with Flye v.2.9 (RRID:SCR_017016) [39], using Minimap v.2.17 [40] for mapping. Afterwards, we conducted two rounds of short-read polishing by mapping reads onto the assembly with BWA-MEM v.0.7.17 (RRID:SCR_010910) [41], followed by error correction with Pilon v1.23 (RRID:SCR_014731) [42].

Assembly quality control and scaffolding

The polished assembly contigs were anchored into chromosome-scale scaffolds utilizing the generated proximity-ligation Omni-C data. First, the data were mapped and filtered to the assembly following the Arima Hi-C mapping pipeline used by the Vertebrate Genome Project [43]. In brief, reads were mapped using BWA-MEM v.0.7.17 [41], the mapped reads were filtered with samtools v.1.14 (RRID:SCR_002105) [44], and the duplicated reads were removed with “MarkDuplicates” in Picard v.2.26.10 (RRID:SCR_006525) [45]. The filtered mapped reads were then used for proximity-ligation scaffolding in YaHs v.1.1 [46]. Gaps in the scaffolded assembly were closed with TGS-GapCloser v.1.1.1 (RRID:SCR_017633) [47] using a subset (25%) of the PacBio subreads due to computational constraints. To further improve the assembly’s contiguity, scaffolding and gap-closing were performed a second time using a different subset of PacBio reads for gap-closing. The PacBio read subsets were generated with Seqtk v.1.3 (RRID:SCR_018927) [48] using the random number generator seeds 11 and 18. Gene set completeness was analyzed with BUSCO v.5.4.7 [49] using the Actinopterygii set of core genes (*actinopterygii_odb10*). Assembly continuity was evaluated using QUAST v5.0.2 (RRID:SCR_001228) [50], and mapping rates were assessed by QualiMap v2.2.1 (RRID:SCR_001209) [51]. Finally, BlobToolsKit v.4.0.6 [52] performed contamination screening.

Repeat landscape analysis and genome annotation

The TE annotation was done in three steps. First, we used RepeatMasker v4.1.5 [53] to annotate and hard-mask known Actinopterygii repeats from Repbase (RRID:SCR_021169), which comprises a database of eukaryotic repetitive DNA element sequences [54]. Secondly, a *de novo* library of TE was created from the hard-masked genome assembly using RepeatModeler v2.0.4 [55], which includes RECON v1.08 (RRID:SCR_021170) [56], RepeatScout v1.0.6 (RRID:SCR_014653) [57], as well as LTRharvest and LTR_retriever (RRID:SCR_018970 and RRID:SCR_017623, respectively) [58, 59]. Finally, predicted repeats were annotated with a second run of RepeatMasker on the hard-masked assembly obtained in the first run. The results of both RepeatMasker runs were then combined. A summary of TEs and the relative abundance of repeat classes in the genome are shown in Table 2 and Figure 2.

The genome was annotated using the BRAKER3 pipeline (RRID:SCR_018964) [60–65], combining a *de novo* gene calling and a homology-based gene annotation. For protein references, we combined the vertebrate-specific protein collection from OrthoDB (RRID:SCR_011980) and the protein collection of the greater pipefish (*Syngnathus acus*) genome [30] made by the NCBI (see: GCF_901709675.1, last accessed 12th October 2023). To further filter genes based on the support of introns and using extrinsic homology evidence, we used TSEBRA [66] with an “intron_support=0.1”. The resulting set of proteins was tested for completeness using BUSCO v.5.4.7 [49] in “protein mode” and run against the Actinopterygii-specific set of core genes. Finally, functional annotation was done using InterProScan v5 (RRID:SCR_005829) [67].

Variant calling and demographic inference

The preprocessed short reads were mapped to the final assembly using BWA-MEM v.0.7.17 [41], followed by the removal of duplicate reads with “MarkDuplicates” in Picard v.2.26.10 [45] and the evaluation of the mapping quality using Qualimap v2.2.1 [51]. Indels

Table 2. Repeat content of the genome assembly. Class: class of the repetitive regions. Count: number of occurrences of the repetitive region. bpMasked: number of base pairs masked; %Masked: percentage of base pairs masked. LINE: Long Interspersed Nuclear Elements (include retroposons); LTR: Long Terminal Repeat elements (including retroposons); SINE: Short Interspersed Nuclear Elements; RC: Rolling Circle.

Class	Count	bpMasked	%masked
ARTEFACT	4	84	0.00%
DNA	2,765,297	372,407,739	22.40%
LINE	850,222	167,337,419	10.06%
LTR	177,214	55,439,687	3.33%
PLE	1	0	0.00%
RC	32,348	3,385,084	0.20%
SINE	435,464	32,709,572	1.95%
Unknown	3,628,328	534,216,084	32.14%
Low complexity	127,733	3,095,322	0.19%
Satellite	21,221	7,145,469	0.43%
Simple repeat	1,437,090	61,077,339	3.67%
rRNA	4,394	534,599	0.03%
scRNA	5	504	0.00%
snRNA	695	46,845	0.00%
tRNA	6,029	533,812	0.03%
Total	9,486,045	1,237,929,559	74.93%

in the BAM files were first identified and then realigned with “RealignerTargetCreator” and “IndelRealigner” as part of the Genome Analysis Toolkit v3.8-1 [68]. Subsequently, samtools v.1.14 [44] was used to check and remove unmapped, secondary, QC-failed, duplicated, and supplementary reads, keeping only reads mapped in proper pairs in non-repetitive regions of the 28 chromosome-scale scaffolds.

Sambamba v 1.0.0 (RRID:SCR_024328) [69] was used to estimate site depth statistics. Minimum and maximum thresholds for the global site depth were set to $d \pm (5 \times \text{MAD})$, where d is the global site depth distribution median and MAD is the median absolute deviation. Variant calling was performed using the bcftools v1.17 (RRID:SCR_005227) [70] commands “mpileup” and “call” [-m]. Variants were then filtered with bcftools “filter” [-e “ $\text{DP} < d - (5 \times \text{MAD}) \parallel \text{DP} > d + (5 \times \text{MAD}) \parallel \text{QUAL} < 30$ ”], thus removing sites with low quality and out of range depth. Finally, bcftools was used to estimate the genome-wide heterozygosity as the proportion of heterozygous sites using the “stats” command.

Long-term changes in the effective population size (N_e) over time were estimated using the PSMC model [17]. This analysis used the diploid consensus genome sequences generated by bcftools v1.17 [70] with the script “vcfutils.pl” from the processed BAM files, as described above. Sites with read-depth up to a third of the average depth or above twice each sample’s median depth and with a consensus base quality < 30 were removed. PSMC was executed using 25 iterations, employing a maximum $2N_0$ -scaled coalescent time of 15, an initial θ/ρ ratio of 5, and 64 atomic time intervals ($4 + 25 \times 2 + 4 + 6$) to infer the scaled mutation rate, the recombination rate, and the free population size parameters, respectively. We performed 100 bootstrap replicates by randomly sampling with replacement 1 Mb blocks from the consensus sequence for all individuals. A mutation rate μ of 1.7×10^{-9} per site per generation [71] and a generation length of 2.5 years [72] were employed for plotting.

DATA AVAILABILITY

The *de novo* genome and all underlying raw data were uploaded to NCBI under the BioProject PRJNA1005573, BioSample SAMN36988691, genome assembly GCA_034508595.

All other data, including the repeat and gene annotation, is available in the GigaDB repository [73].

ABBREVIATIONS

CLR, continuous long reads; kya, thousand years ago; MAD, median absolute deviation; PSMC, Pairwise Sequentially Markovian Coalescent; TE, transposable element.

DECLARATIONS

Ethics approval and consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author contributions

MW, BF, MS, AJ, and SW designed the study. SW, JR, HMI, JM, CDS, LG, MJW, HO, and MWI performed laboratory procedures and sequencing. MW, BF, RC, MDJ, MN, DP, YS, KZ, JR, HMI, JM, CDS, LG, MJW, HO, MWI, MAN, and SW conducted bioinformatic processing and analyses. All authors contributed to writing this manuscript.

Acknowledgements

The present study is a result of the Centre for Translational Biodiversity Genomics (LOEWE-TBG) and was supported through the program 'LOEWE-Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz' of Hesse's Ministry of Higher Education, Research, and the Arts.

REFERENCES

- 1 Froese R, Pauly D. FishBase. 2023; www.fishbase.org. Accessed 9 August 2023.
- 2 Dawson C. Syngnathidae. In: Smith M, Heemstra P (eds), *Smiths' Sea Fishes*. Berlin: Springer-Verlag, 1986; pp. 445–458.
- 3 O'Gorman EJ. Multitrophic diversity sustains ecological complexity by dampening top-down control of a shallow marine benthic food web. *Ecology*, 2021; **102**: e03274. doi:10.1002/ecy.3274.
- 4 Vincent ACJ, Berglund A, Ahnesj I. Reproductive ecology of five pipefish species in one eelgrass meadow. *Environ. Biol. Fish*, 1995; **44**: 347–361. doi:10.1007/BF00008250.
- 5 Polte P, Buschbaum C. Native pipefish *Entelurus aequoreus* are promoted by the introduced seaweed *Sargassum muticum* in the northern Wadden Sea. *North Sea Aquat. Biol.*, 2008; **3**: 11–18. doi:10.3354/ab00071.
- 6 Braga Goncalves I, Cornetti L, Couperus AS et al. Phylogeography of the snake pipefish, *Entelurus aequoreus* (Family: Syngnathidae) in the northeastern Atlantic Ocean. *Biol. J. Linn. Soc.*, 2017; **122**: 787–800. doi:10.1093/biolinnean/blx112.
- 7 Wheeler A. Key to the Fishes of Northern Europe: A Guide to the Identification of More than 350 Species. London: Frederick Warne & Co. Ltd, 1978. ISBN-10:0723220646.
- 8 Harris MP, Beare D, Toresen R et al. A major increase in snake pipefish (*Entelurus aequoreus*) in northern European seas since 2003: potential implications for seabird breeding success. *Mar. Biol.*, 2007; **151**: 973–983. doi:10.1007/s00227-006-0534-7.
- 9 Fleischer D, Schaber M, Piepenburg D. Atlantic snake pipefish (*Entelurus aequoreus*) extends its northward distribution range to Svalbard (Arctic Ocean). *Polar Biol.*, 2007; **30**: 1359–1362. doi:10.1007/s00300-007-0322-y.

- 10 Rusyaev SM, Dolgov AV, Karamushko OV. Captures of snake pipefish *Entelurus aequoreus* in the Barents and Greenland Seas. *J. Ichthyol.*, 2007; **47**: 544–546. doi:10.1134/S0032945207070090.
- 11 Kloppmann MHF, Ulleweit J. Off-shelf distribution of pelagic snake pipefish, *Entelurus aequoreus* (Linnaeus, 1758), west of the British Isles. *Mar. Biol.*, 2007; **151**: 271–275. doi:10.1007/s00227-006-0480-4.
- 12 van Damme CJ, Couperus AS. Mass occurrence of snake pipefish in the Northeast Atlantic: result of a change in climate? *J. Sea Res.*, 2008; **60**: 117–125. doi:10.1016/j.seares.2008.02.009.
- 13 Lindley J, Kirby R, Johns D et al. Exceptional abundance of the snake pipefish (*Entelurus aequoreus*) in the north-eastern North Atlantic Ocean. ICES Document. 2006.
- 14 Roth O, Solbakken MH, Tørresen OK et al. Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. *Proc. Natl. Acad. Sci. USA*, 2020; **117**: 9431–9439. doi:10.1073/pnas.1916251117.
- 15 Stiller J, Short G, Hamilton H et al. Phylogenomic analysis of Syngnathidae reveals novel relationships, origins of endemic diversity and variable diversification rates. *BMC Biol.*, 2022; **20**: 75. doi:10.1186/s12915-022-01271-w.
- 16 Lin Q, Fan S, Zhang Y et al. The seahorse genome and the evolution of its specialized morphology. *Nature*, 2016; **540**: 395–399. doi:10.1038/nature20595.
- 17 Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*, 2011; **475**: 493–496. doi:10.1038/nature10231.
- 18 Prost S, Winter S, de Raad J et al. Education in the genomics era: generating high-quality genome assemblies in university courses. *GigaScience*, 2020; **9**(6): g1aa058. doi:10.1093/gigascience/g1aa058.
- 19 Prost S, Petersen M, Grethlein M et al. Improving the chromosome-level genome assembly of the siamese fighting fish (*Betta splendens*) in a University Master's course. *G3 (Bethesda)*, 2020; **10**: 2179–2183. doi:10.1534/g3.120.401205.
- 20 Winter S, Prost S, de Raad J et al. Chromosome-level genome assembly of a benthic associated Syngnathiformes species: the common dragonet, *Callionymus lyra*. *GigaByte*, 2020; **2020**: gigabyte6. doi:10.46471/gigabyte.6.
- 21 Winter S, de Raad J, Wolf M et al. A chromosome-scale reference genome assembly of the great sand eel, *Hyperoplus lanceolatus*. *J. Hered.*, 2023; **114**: 189–194. doi:10.1093/jhered/esad003.
- 22 Vitturi R, Catalano E. Karyotypes in two species of the genus *Hippocampus* (Pisces: Syngnathiformes). *Mar. Biol.*, 1988; **99**: 119–121. doi:10.1007/BF00644985.
- 23 Vitturi R, Libertini A, Campolmi M et al. Conventional karyotype, nucleolar organizer regions and genome size in five Mediterranean species of Syngnathidae (Pisces, Syngnathiformes). *J. Fish Biol.*, 1998; **52**: 677–687. doi:10.1111/j.1095-8649.1998.tb00812.x.
- 24 Small CM, Bassham S, Catchen J et al. The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biol.*, 2016; **17**: 258. doi:10.1186/s13059-016-1126-6.
- 25 Tigano A, Jacobs A, Wilder AP et al. Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation. *Genome Biol. Evol.*, 2021; **13**(6): evab098. doi:10.1093/gbe/evab098.
- 26 Barry P, Broquet T, Gagnaire P-A. Age-specific survivorship and fecundity shape genetic diversity in marine fishes. *Evol. Lett.*, 2022; **6**: 46–62. doi:10.1002/evl3.265.
- 27 Shao F, Han M, Peng Z. Evolution and diversity of transposable elements in fish genomes. *Sci. Rep.*, 2019; **9**: 15399. doi:10.1038/s41598-019-51888-1.
- 28 Small CM, Healey HM, Currey MC et al. Leafy and weedy seadragon genomes connect genic and repetitive DNA features to the extravagant biology of syngnathid fishes. *Proc. Natl. Acad. Sci. USA*, 2022; **119**: e2119602119. doi:10.1073/pnas.2119602119.
- 29 Meyer A, Schloissnig S, Franchini P et al. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature*, 2021; **590**: 284–289. doi:10.1038/s41586-021-03198-8.
- 30 Scott-Somme K, McTierney S, Brittain R et al. The genome sequence of the greater pipefish, *Syngnathus acus* (Linnaeus, 1758). *Wellcome Open Res.*, 2023; **8**: 274. doi:10.12688/wellcomeopenres.19528.1.
- 31 Obrochta SP, Crowley TJ, Channell JE et al. Climate variability and ice-sheet dynamics during the last three glaciations. *Earth Planet. Sci. Lett.*, 2014; **406**: 198–212. doi:10.1016/j.epsl.2014.09.004.
- 32 Armstrong E, Hopcroft PO, Valdes PJ. A simulated Northern Hemisphere terrestrial climate dataset for the past 60,000 years. *Sci. Data*, 2019; **6**: 265. doi:10.1038/s41597-019-0277-1.

- 33 Mayjonade B, Gouzy J, Donnadieu C et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques*, 2016; **61**: 203–205. doi:10.2144/000114460.
- 34 PacBio secondary analysis tool BAM2fastx v.1.3.0, GitHub. 2023; <https://github.com/PacificBiosciences/pbbioconda>.
- 35 Chen S, Zhou Y, Chen Y et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 2018; **34**: i884–i890. doi:10.1093/bioinformatics/bty560.
- 36 Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 2011; **27**: 764–770. doi:10.1093/bioinformatics/btr011.
- 37 Vurture GW, Sedlazeck FJ, Nattestad M et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 2017; **33**: 2202–2204. doi:10.1093/bioinformatics/btx153.
- 38 Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods*, 2020; **17**: 155–158. doi:10.1038/s41592-019-0669-3.
- 39 Kolmogorov M, Yuan J, Lin Y et al. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, 2019; **37**: 540–546. doi:10.1038/s41587-019-0072-8.
- 40 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 2018; **34**: 3094–3100. doi:10.1093/bioinformatics/bty191.
- 41 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; <https://doi.org/10.48550/arXiv.1303.3997>.
- 42 Walker BJ, Abeel T, Shea T et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 2014; **9**: e112963. doi:10.1371/journal.pone.0112963.
- 43 Vertebrate Genome Project. Arima Genomics Mapping Pipeline 042817. https://github.com/VGP/vgp-assembly/blob/master/pipeline/salsa/arima_mapping_pipeline.sh.
- 44 Danecek P, Bonfield JK, Liddle J et al. Twelve years of SAMtools and BCFtools. *GigaScience*, 2021; **10**(2): giab008. doi:10.1093/gigascience/giab008.
- 45 Picard toolkit. 2019; Broad Institute, GitHub Repository. <https://broadinstitute.github.io/picard/>.
- 46 Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*, 2023; **39**(1): btac808. doi:10.1093/bioinformatics/btac808.
- 47 Xu M, Guo L, Gu S et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*, 2020; **9**(9): giaa094. doi:10.1093/gigascience/giaa094.
- 48 Seqtk v.1.3. <https://github.com/lh3/seqtk>.
- 49 Manni M, Berkeley MR, Seppey M et al. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.*, 2021; **1**: e323. doi:10.1002/cpz1.323.
- 50 Mikheenko A, Prjibelski A, Saveliev V et al. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, 2018; **34**: i142–i150. doi:10.1093/bioinformatics/bty266.
- 51 Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 2016; **32**: 292–294. doi:10.1093/bioinformatics/btv566.
- 52 Challis R, Richards E, Rajan J et al. BlobToolKit – Interactive quality assessment of genome assemblies. *G3 (Bethesda)*, 2020; **10**(4): 1361–1374. doi:10.1534/g3.119.400908.
- 53 Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013; <http://www.repeatmasker.org>.
- 54 Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, 2015; **6**: 11. doi:10.1186/s13100-015-0041-9.
- 55 Flynn JM, Hubley R, Goubert C et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*, 2020; **117**: 9451–9457. doi:10.1073/pnas.1921046117.
- 56 Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.*, 2002; **12**: 1269–1276. doi:10.1101/gr.88502.
- 57 Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics*, 2005; **21**(Suppl 1): i351–i358. doi:10.1093/bioinformatics/bti1018.
- 58 Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.*, 2018; **176**: 1410–1422. doi:10.1104/pp.17.01310.

- 59 Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.*, 2008; 9: 18. doi:10.1186/1471-2105-9-18.
- 60 Bruna T, Lomsadze A, Borodovsky M. GeneMark-ETP: automatic gene finding in eukaryotic genomes in consistency with extrinsic data. bioRxiv. 2023; <https://doi.org/10.1101/2023.01.13.524024>.
- 61 Bruna T, Hoff KJ, Lomsadze A et al. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.*, 2021; 3: lqaa108. doi:10.1093/nargab/lqaa108.
- 62 Kovaka S, Zimin AV, Pertea GM et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.*, 2019; 20: 278. doi:10.1186/s13059-019-1910-1.
- 63 Hoff KJ, Lange S, Lomsadze A et al. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 2016; 32: 767–769. doi:10.1093/bioinformatics/btv661.
- 64 Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 2015; 12: 59–60. doi:10.1038/nmeth.3176.
- 65 Gabriel L, Bruna T, Hoff KJ et al. BRAKER3: fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv. 2023; <https://doi.org/10.1101/2023.06.10.544449>.
- 66 Gabriel L, Hoff KJ, Bruna T et al. TSEBRA: transcript selector for BRAKER. *BMC Bioinform.*, 2021; 22: 566. doi:10.1186/s12859-021-04482-0.
- 67 Jones P, Binns D, Chang H-Y et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 2014; 30: 1236–1240. doi:10.1093/bioinformatics/btu031.
- 68 Genome analysis toolkit v3.8-1. Broad Institute. 2022; <https://gatk.broadinstitute.org/>.
- 69 Tarasov A, Vilella AJ, Cuppen E et al. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 2015; 31: 2032–2034. doi:10.1093/bioinformatics/btv098.
- 70 Danecek P, Bonfield JK, Liddle J et al. Twelve years of SAMtools and BCFtools. *GigaScience*, 2021; 10(2): giab008. doi:10.1093/gigascience/giab008.
- 71 He L, Long X, Qi J et al. Genome and gene evolution of seahorse species revealed by the chromosome-level genome of *Hippocampus abdominalis*. *Mol. Ecol. Resour.*, 2022; 22: 1465–1477. doi:10.1111/1755-0998.13541.
- 72 Schultz J. *Entelurus aequoreus*: IUCN red list of threatened species. e.T18258072A44775951. 2014.
- 73 Wolf M, Ferrette BLS, Coimbra RTF et al. Supporting data for “Near chromosome-level and highly-repetitive genome assembly of the snake pipefish *Entelurus aequoreus* (Syngnathiformes: Syngnathidae)”. *GigaScience Database*, 2023; <http://dx.doi.org/10.5524/102493>.