microPublication BIOLOGY

10/18/2022 - *Open Access*

First Ever Whole Genome Sequencing and De Novo Assembly of the Freshwater Angelfish, *Pterophyllum scalare*

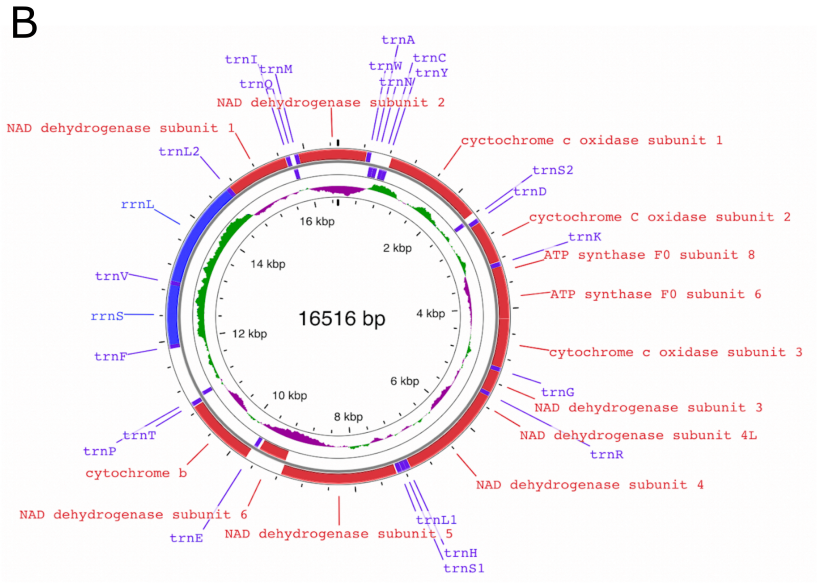
Indeever Madireddy
BioCurious, Santa Clara, CA, USA
BASIS Independent Silicon Valley, San Jose, CA, USA
To whom correspondence should be addressed: indeever@gmail.com

Abstract

This research work is the first to ever sequence and perform a de novo assembly of the genome of the freshwater angelfish, *Pterophyllum scalare*. The final genome assembly consisted of 15,486 contigs and was 734.79 Mb in size with an 86.5% BUSCO score. Functional annotation of the genome revealed 24,247 protein-coding sequences related to other fish species. 14,329 (59%) of the identified genes were orthologous to *Archocentrus centrarchus*, a closely related South American cichlid.

A

Number of Reads	6,937,772
Number of Bases	10,095,840,481
N50 Read Length	3,237
Longest Read	228,316
Shortest Read	17
Mean Read Length	1,455
Median Read Length	628
Mean Read Quality	15.06
Median Read Quality	14.58



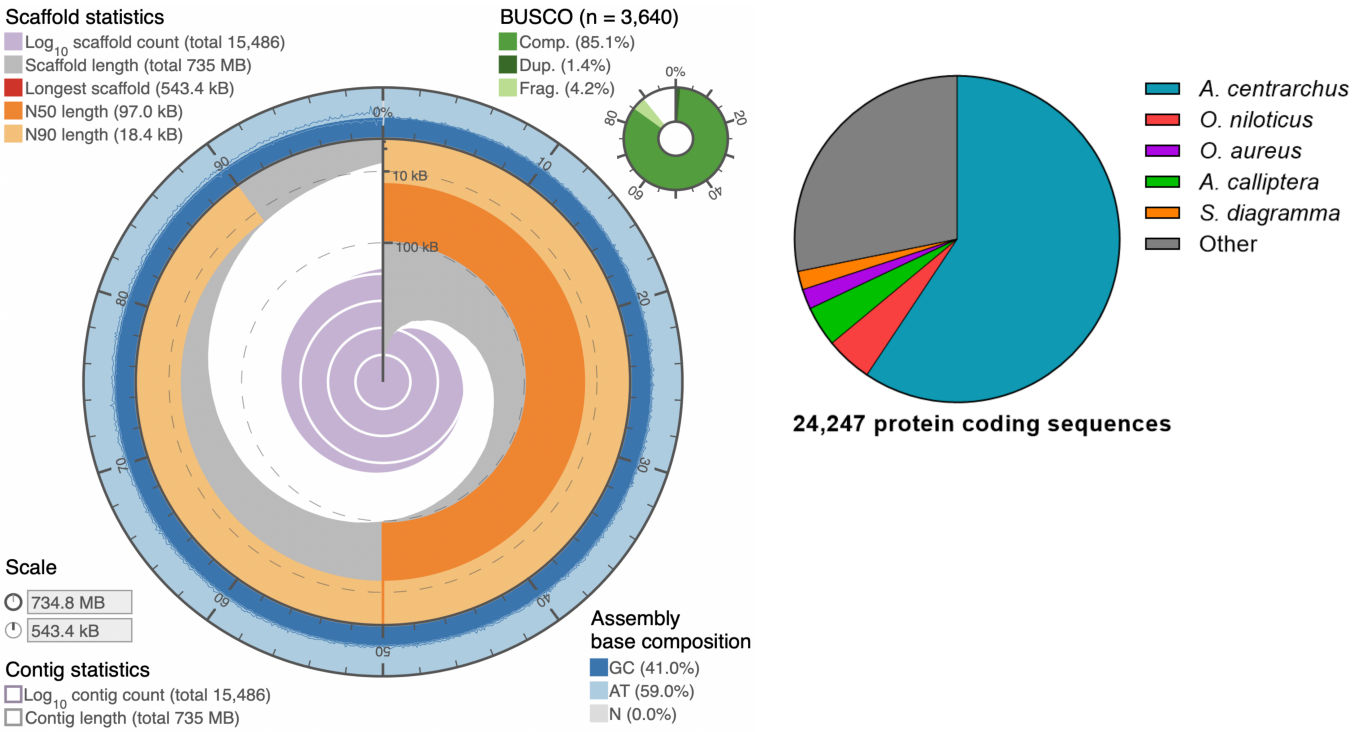


Figure 1. Long-read nanopore sequencing of angelfish DNA leads to a robust genome assembly:

A) Table depicting the read statistics of the nanopore sequencing including the number of collected reads, the mean read length, and mean read quality. **B)** Angelfish mitochondrial genome annotated with the 37 genes. Violet text refers to tRNAs, red text refers to protein-coding genes, and blue text corresponds to rRNAs. The green histogram indicates + GC skew while the purple histogram indicates -GC skew. **C)** Snail plot showing the assembled nuclear genome with its corresponding statistics. **D)** Orthologous genes were most commonly identified in the following organisms: *Archocentrus centrarchus*, *Oreochromis niloticus*, *Oreochromis aureus*, *Astatotilapia calliptera*, and *Simochromis diagramma*.

Description

10/18/2022 - *Open Access*

The freshwater angelfish, *Pterophyllum scalare*, is a popular freshwater cichlid kept by aquarium hobbyists around the world. Originally from South America, these fish are well known for their monogamous breeding patterns and thorough parental care of offspring (Cacho et al., 2007). Although the behaviors of the angelfish have been well studied, very little is known about the nuclear genetics of the angelfish as its genome has never been fully sequenced and assembled (Gómez-Laplaza and Gerlai, 2020). Cichlids are of especial importance to biomedical research, for they have been used as model organisms to study craniofacial variation and neurobiology (Powder and Albertson, 2016; Maruska and Fernald, 2018). Investigating the genome of the angelfish may enable its use as a model organism for further biological research. In this work, I sequenced, assembled, and annotated the complete genome of the freshwater angelfish in addition to the full mitochondrial genome with Oxford Nanopore Technologies (Lu et al., 2016).

With the MinION MK1B device, 6.94 million sequencing reads (figure A) and an estimated 10.1 gigabases at a 3.24 kb N50 read length were collected (Steinig and Coin, 2022). Two flow cells (R10.4) were used to collect this sequencing data, and the flow cells were run for 72 hours each. The reads collected had a mean read quality of 15.06 and a median read quality of 14.58 corresponding to an estimated 97% sequencing accuracy. Reads were collected at an average translocation speed of 220 bases per second.

Collected reads were then screened to identify potential contaminant organisms in the sequencing data. The kraken2 tool (ver. 2.1.2) identified that *Pseudomonas aeruginosa*, a common opportunistic aquatic pathogen, was the largest contaminant of the sequencing reads (Wood and Salzberg, 2014; Souza et al., 2019).

The mitochondrial genome (figure B) of the angelfish was assembled from the sequencing reads. All 37 conserved mitochondrial genes including 2 rRNAs, 13 genes and 22 tRNAs common to eukaryotic organisms were identified, indicating a complete and robust assembly. This new assembly was found to be 25 bp longer than the reference mitochondrial assembly, with a 99.1% similarity.

The final nuclear genome assembly (figure C) consisted of 15,486 contigs totaling 734.79 Mb with a final BUSCO score of 86.5% and a 41% GC content (Simão et al., 2019). The genome size and GC content is similar to other fish species such as the Asian seabass (*Lates calcarifer*) and the Nile tilapia (*Oreochromis niloticus*) (Lu and Luo, 2020). The N50 contig length of the assembled genome was 96,962 bp, and the longest contig was 543,394 bp. Repeatmasker (ver. 4.1.1) masked 12.47% of the genome containing simple repeat sequences (Chen, 2004). An interactive version of the whole genome can be found here: https://indeeverm.github.io/assembly-stats/

NCBI blastp (ver. 2.12.0) performed functional annotation of the genome through the GenSAS platform (States and Gish, 1994; Humann et al, 2019). 24,247 unique protein-coding sequences orthologous to other species were identified in the angelfish genome against the refseq vertebrate-other database. A majority of genes, 59%, were orthologous to *Archocentrus centrarchus*, a closely related South American cichlid (figure D). Timetree suggests that *A. centrarchus* and *P. scalare* diverged between 28.7 to 72.4 million years ago (Kumar et al., 2017).

Future work would involve RNA sequencing of the angelfish to build an appropriate transcriptome of the organism. Illumina sequencing could also be performed to improve the current assembly.

Methods

Angelfish

The angelfish used in this work died of natural causes prior to the start of experimentation and was neither euthanized nor harmed in any way for the purpose of this research. Although IACUC approval was not required, all US regulations were followed in the collection and handling of the biological material. Angelfish tissue was obtained from an angelfish raised by the author from birth. Angelfish muscle and skin tissue were collected post-mortem with a punch biopsy. Fish tissue was stored in a DNA shield at -80°C until use.

Angelfish Genomic DNA Extraction

Angelfish genomic DNA was extracted with the NEB Monarch genomic DNA purification kit. The standard protocol for tissue samples was followed with the following specifications. 30 mg of angelfish muscle tissue was lysed with the provided tissue lysis buffer and proteinase K for 1.5 hours at 56°C and shaking at 2000 rpm. Samples were vortexed every 5 minutes for 5 seconds during lysis. Cell debris was pelleted at 12000 x g after lysis and the supernatant was transferred to a new microcentrifuge tube. RNAse A addition was not skipped to minimize RNA carryover. Genomic DNA was washed three times with wash buffer 1 and eluted in 60µL of TE buffer heated to 70°C. DNA concentration was measured with a Denovix DS-11 spectrophotometer.

Extracted genomic DNA was then run on a 1 percent TBE agarose gel in pulsed-field gel electrophoresis in order to verify DNA length and quality. Samples were run in the CHEF-DR II system for 20 hours at 3.5 V/cm with a pulse shift every 25 seconds. Genomic DNA fragments were approximately 20-40kb in length.

Library prep

The standard ligation library prep (SQK-LSK112) was performed on the angelfish genomic DNA following the manufacturer's protocol.

Base-calling

Base-calling was performed at high accuracy by the Guppy base-caller (ver. 6.1.5).

Nuclear Genome Assembly

I performed whole-genome assembly with the flye de novo assembler (ver. 2.9) built into the Project Galaxy bioinformatics tool (ver. 22.01) with one round of polishing (Kolmogorov et al., 2019; Afgan et al., 2019) . Shorter reads (< 1000 bp) were eliminated as they did not improve the assembly and increased computation time. Only reads with a quality score above 9 were used in the assembly. A BUSCO (Benchmarking Universal Single-Copy Orthologs) score, a basis for the completeness of a genome, was calculated for the assembly with the reference being the *Actinopterygii* lineage and was found to be 88% for the initial Flye assembly. Neither polishing the genome with Racon nor including lower quality reads in the assembly improved this score (Vaser et al., 2017). Smaller-sized contigs (< 5000 bp) were then removed from the final assembly as they ware found to not contribute to the BUSCO score. Kraken2 was used again to identify and remove contaminating contigs. Mitochondrial contigs were also removed from the nuclear genome assembly. The genome snail plot was generated using https://github.com/rjchallis/assembly-stats (ver. 17.02).

Mitochondrial Genome Assembly

I assembled the mitochondrial genome of the angelfish against one already assembled by Hao et al (Hao et al., 2016). The Epi2Me fastq custom alignment program (ver. 3.4.2) was used to align the 6.94 million collected reads against the mitochondrial reference genome to filter out unnecessary genomic reads. The Shasta de novo assembler (ver. 0.10.0) then assembled the remaining mitochondrial reads. This assembly resulted in a 16,516 base pair mitochondrial genome (Shafin et al., 2020). The MITOS web server (http://mitos.bioinf.uni-leipzig.de/index.py) was used to annotate the genome (Bernt et al., 2013). This plot was generated with https://proksee.ca (ver. 1.0.0).

Additional Information

A pipeline for this work is available on GitHub: https://github.com/IndeeverM/Angelfish-Genome-Assembly.

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAMQGT000000000. The version described in this paper is version JAMQGT020000000.

Reagents

Kit	Use
Nanopore Ligation Sequencing Kit (SQK-LSK112)	Library Prep
Flow Cell Wash Kit (EXP-WSH004)	Washing of Flow Cell
NEB Monarch Genomic DNA Purification Kit	Extracting Angelfish Genomic DNA

Acknowledgements: I would like to acknowledge Mr. Johan Sosa and Mr. Kurt Chang from BioCurious for their support on this project. I would also like to thank Mr. Yuanyu Lin from the University of North Carolina at Chapel Hill for his advice.

References

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al., Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46: W537-W544. PubMed

10/18/2022 - *Open Access*

ID: 29790989

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al., Stadler PF. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol 69: 313-9. PubMed ID: 22982435

Cacho MS, Yamamoto ME, Chellappa S. 2007. Mating system of the Amazonian cichlid angel fish, Pterophyllum scalare. Braz J Biol 67: 161-5. PubMed ID: 17505764

Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4: Unit 4.10. PubMed ID: 18428725

Gómez-Laplaza LM, Gerlai R. 2020. Food Quantity Discrimination in Angelfish (*Pterophyllum scalare*): The Role of Number, Density, Size and Area Occupied by the Food Items. Front Behav Neurosci 14: 106. PubMed ID: 32655384

Hao G, Wu Q, Zhong H, Zhou Y. 2016. Complete mitochondrial genome of Pterophyllum scalare (Perciformes, Cichlidae). Mitochondrial DNA A DNA Mapp Seq Anal 27: 4215-4216. PubMed ID: 26000948

Humann JL, Lee T, Ficklin S, Main D. 2019. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. Methods Mol Biol 1962: 29-51. PubMed ID: 31020553

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34: 1812-1819. PubMed ID: 28387841

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37: 540-546. PubMed ID: 30936562

Lu, G., & Luo, M. 2020. Genomes of major fishes in world fisheries and aquaculture: Status, application and perspective. *Aquaculture and Fisheries*, *5*(4), 163-173. DOI: 10.1016/j.aaf.2020.05.004

Lu H., Giordano F., Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 14: 265-279. PubMed ID: 27646134

Maruska KP, Fernald RD. 2018. Astatotilapia burtoni: A Model System for Analyzing the Neurobiology of Behavior. ACS Chem Neurosci 9: 1951-1962. PubMed ID: 29522313

Powder KE, Albertson RC. 2016. Cichlid fishes as a model to understand normal and clinical craniofacial variation. Dev Biol 415: 338-346. PubMed ID: 26719128

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al., Paten B. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol 38: 1044-1053. PubMed ID: 32686750

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-2. PubMed ID: 26059717

Souza CF, Baldissera MD, Descovi SN, Zeppenfeld CC, Verdi CM, Santos RCV, da Silva AS, Baldisserotto B. 2019. Grape pomace flour alleviates Pseudomonas aeruginosa-induced hepatic oxidative stress in grass carp by improving antioxidant defense. Microb Pathog 129: 271-276. PubMed ID: 30802491

States DJ, Gish W. 1994. Combined use of sequence similarity and codon bias for coding region identification. J Comput Biol 1: 39-50. PubMed ID: 8790452

Steinig, E., & Coin, L. 2022. Nanoq: ultra-fast quality control for nanopore reads. *Journal of Open Source Software*, *7*(69), 2991. DOI: 10.21105/joss.02991

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27: 737-746. PubMed ID: 28100585

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15: R46. PubMed ID: 24580807

Funding: Funding was raised through experiment.com. www.doi.org/10.18258/27105

Author Contributions: Indeever Madireddy: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project, resources, software, supervision, validation, visualization, writing - original draft, writing - review editing.

Reviewed By: Anonymous

microPublication BIOLOGY

10/18/2022 - *Open Access*

History: Received September 22, 2022 **Revision Received** October 14, 2022 **Accepted** October 18, 2022 **Published Online** October 18, 2022 **Indexed** November 1, 2022

Citation: Madireddy, I (2022). First Ever Whole Genome Sequencing and De Novo Assembly of the Freshwater Angelfish, *Pterophyllum scalare*. microPublication Biology. 10.17912/micropub.biology.000654