Data in Brief 54 (2024) 110400

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

First transcriptome sequencing, assembly, and annotation dataset for the freshwater angelfish, pterophyllum scalare

Indeever Madireddy[a,b,*]

[a] BioCurious, 3108 Patrick Henry Dr, Santa Clara, CA 95054, United States
[b] California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, United States

ARTICLE INFO

Article history:
Received 30 September 2023
Revised 29 February 2024
Accepted 2 April 2024
Available online 5 April 2024

Dataset link: Pterophyllum scalare Transcriptome Assembly (Original data)

Keywords:
MinION
Nanopore
RNA
Cichlid
Craniofacial

ABSTRACT

Cichlids are relevant to biological research for their craniofacial variations that are analogous to human structure and associated congenital anomalies. However, only a limited number of cichlids have genetic information available. Investigating cichlids and adding to the body of knowledge about them may provide better insights into studying developmental biology and craniofacial structure. The angelfish, *Pterophyllum scalare*, is one cichlid for which we lack genetic information including a draft transcriptome assembly.

This work is the first to provide a draft transcriptome and annotation using long-read Nanopore sequencing for *P. scalare*. Total RNA was extracted from angelfish tissue, and a cDNA-PCR library was prepared. Sequencing was performed on a singular R.9.4.1 MinION flow cell for 84 h. Various bioinformatic tools were then employed to assemble the sequencing reads into a transcriptome. The transcriptome was then annotated against various databases.

23 million sequencing reads were collected totalling 21.9 Gb. The N50 sequencing read length was 1255 bp and the mean read length was 938. The data had an initial mean Phred score of 10.04. After assembly, the final transcriptome consists of 98,125 transcripts with a mean length of 1552 and N50 length of 2277. The transcriptome has a completeness of 80.5% as assessed by BUSCO. Functional annotation re-

* Correspondence to: BioCurious, 3108 Patrick Henry Dr, Santa Clara, CA 95054, United States.
E-mail address: indeever@caltech.edu
Social media: @indeever_m

https://doi.org/10.1016/j.dib.2024.110400

vealed pathways related to signal transduction, carbohydrate metabolism, and transcription are the most annotated in the transcriptome.

Specifications Table

Subject	*Omics: Transcriptomics*
Specific subject area	Freshwater Cichlid Transcriptomics
Data format	*Raw, Filtered, Assembly*
Type of data	*Tables, figures, fasta assembly files, fastq read files*
Data collection	*Angelfish total RNA was extracted using the Zymo Quick-RNA kit. The RNA library was then prepped using the cDNA-PCR kit. Sequencing was then performed on a Nanopore MinION device on an R9.4.1 flow cell.*
Data source location	*Angelfish tail fin and bone tissue*
Data accessibility	Repository name: NCBI Data identification number: PRJNA979816 Direct URL to data: www.ncbi.nlm.nih.gov/bioproject/PRJNA979816 Instructions for accessing these data: The raw sequencing reads, and the transcriptome assembly can be accessed and downloaded by visiting the direct URL. The filtered sequencing reads are also available through the URL.

1. Value of the Data

- This data includes sequencing reads and the first draft transcriptome assembly for the freshwater angelfish, a South American cichlid.
- This work also includes a thorough annotation of the draft transcriptome, assessing the transcriptome assembly against various databases.
- Cichlids are model organisms for developmental biology and craniofacial structure. Neural crest development pathways are highly conserved across vertebrates, and the natural variations in facial structure in cichlids are in turn analogous to variations in human facial structure. Adding to the body of knowledge of cichlids will allow for more insights into the causes of human craniofacial anomalies.
- This transcriptomic data may also provide behavioral insights as the angelfish is known for its serial monogamous relationships and thorough parental care of offspring.
- Geneticists interested in functional enrichment analyses or comparative cichlid transcriptomics can use this work as a starting point for future research.

2. Data Description

Cichlid diversity makes cichlids ideal non-model organisms to study craniofacial variation and evolution [1]. Neural crest developmental pathways are highly conserved across vertebrates, and the natural variations in facial structure in cichlids are analogous to variations in human facial structure [1,2]. Thus, the diverse species of cichlids (Fig. 1) can be effective models to study human developmental biology and sequencing the transcriptome of the freshwater angelfish, a poorly understood cichlid, can provide additional insights about these organisms.

2.1. Long read nanopore sequencing data

23.38 million sequencing reads were collected totaling 21.93 Gb (Table 1). The raw N50 read length was 1255 bases and the mean read length was 938 bases. After quality filtering and

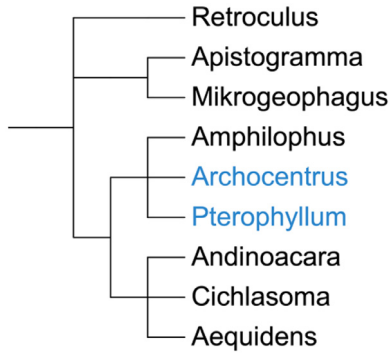

Fig 1. Cladogram of select Cichlid fishes.

Table 1
Pre- and post-filtered sequencing read statistics.

Metric	Pre-filtering	Post-filtering
Reads	23,380,883	14,499,056
Bases	21,933,899,434	11,884,084,049
N50 read length (bases)	1255	1178
Longest read (bases)	884,060	10,179
Shortest read (bases)	6	1
Mean read length (bases)	938	819
Median read length (bases)	762	622
Mean read quality	10.04	15.90
Median read quality	10.51	13.35

Table 2
BUSCO scores for various transcriptome assembly strategies.

Assembly	BUSCO (%)	Single (%)	Duplicated (%)	Fragmented (%)
A. centrarchus stringtie assembly	79.5	40.2	39.9	2.0
P. scalare stringtie assembly	67.8	35.8	32.0	7.6
De-novo assembly	49.5	36.3	13.2	4.6
Combined	81.4	10.2	71.2	1.5
CD-Hit	78.5	21.1	57.4	3.0
Evidential Gene	79.5	47.2	32.3	2.2
Lace super transcripts	68.4	62.3	6.1	5.8
Mapping strategy	**80.5**	**37.1**	**43.4**	**2.0**

adapter trimming, 14.50 million reads remained totaling 11.88 Gb. The N50 read length decreased to 1178 base pairs and the mean read quality significantly increased from 10.04 to 15.90. 84.26% of the reads mapped back to the previously assembled angelfish genome [3].

2.2. Assembly and redundancy reduction

Three transcriptome assemblies were created, and completeness was computed with BUSCO against the Actinopterygii lineage [4]. The first three rows in Table 2 depict the statistics for the initial independent assemblies that were generated. The first two assemblies are genome guided (against the *Archocentrus centrarchus* and *Pterophyllum scalare* genomes respectively), while the

Table 3
Final transcriptome assembly metrics.

Metrics	Values
Transcripts	98,125
Total Length (bases)	152,428,570
Minimum Length (bases)	51
Mean Length (bases)	1552.4
Maximum Length (bases)	15,358
N50 Transcript Length (bases)	2277
GC Content (%)	44.59

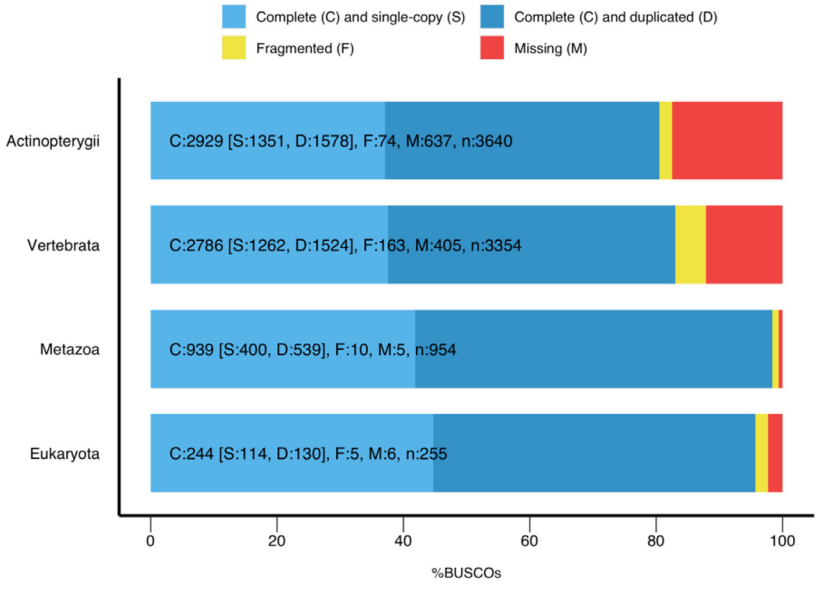

Fig. 2. Transcriptome assembly completeness against various BUSCO lineages.

third is de-novo. When the three assemblies were combined, the BUSCO score increased to 81.4% but duplication in the transcriptome also increased significantly to 71.2%. CD-Hit, Evidential Gene, and Lace were then used to reduce the redundancy on combined assembly. However, none of these tools lead to optimal results. A distinct mapping strategy was tested which maintained the BUSCO score, and limited fragmentation. This mapping strategy was chosen as the final assembly.

2.3. Final transcriptome assembly

Kraken was used to reduce contaminant bacterial and viral transcripts. The BUSCO completeness score remained unchanged after contaminant removal. The final transcriptome consisted of 98,125 transcripts totalling 152.4 Mb. The N50 transcript length was 2277 bases (Table 3).

The completeness of this assembly was then computed against various lineages with BUSCO. Fig. 2 shows these results in reverse taxonomic rank. The transcriptome had a completeness of 80.46% against Actinopterygii, 83.06% completeness against Vertebrata, 98.4% against Metazoa, and 95.7% against Eukaryota lineages.

2.4. Open reading frames

A total of 55,728 open reading frames (ORF) consisting of in-frame start and stop codons were identified with the Transdecoder tool (Table 4). The ORFs had a mean length of 330.9 amino acids and N50 length of 417 amino acids.

2.5. Transcript mapping

Fig. 3 depicts the mapping of the assembled transcripts to the Nr, EggNOG, KEGG, and Pfam databases. Of the 98,125 transcripts, 70,800 mapped to Nr, 22,906 transcripts mapped to EggNOG, 42,474 mapped to KEGG, and 28,063 mapped to Pfam. 73,788 (75.2%) transcripts mapped to at least one database with 13,516 transcripts mapping to all four.

2.6. Functional annotation against Nr database

Over 70% of the 70,800 transcripts aligned to Nr found a hit in *A. centrarchus*, another South American cichlid (Fig. 4).

Table 4
Open-reading frame metrics.

Metrics	Values
Open Reading Frames	55,728
Total Length (amino acids)	18,442,120
Minimum Length amino acids	87
Mean Length (amino acids)	330.9
Maximum Length (amino acids)	4283
N50 Reading Frame Length (amino acids)	417
BUSCO score (%)	77.0

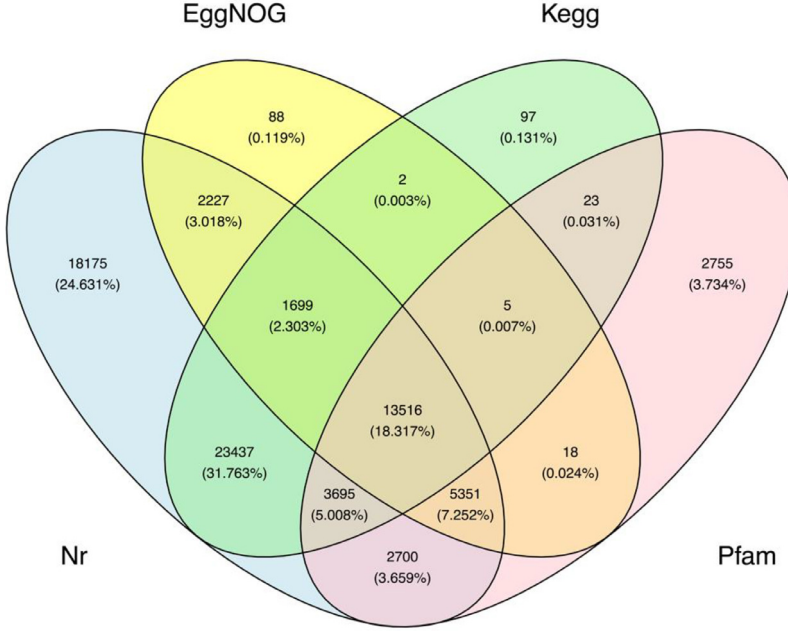

Fig. 3. Assembled transcript mapping against various databases.

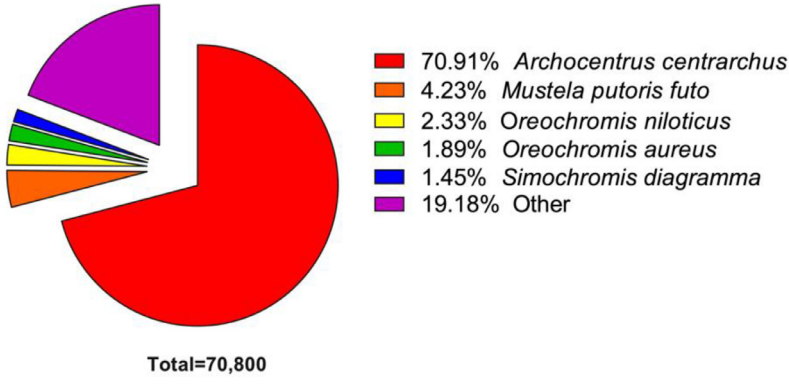

Fig. 4. Top blast hits of transcripts.

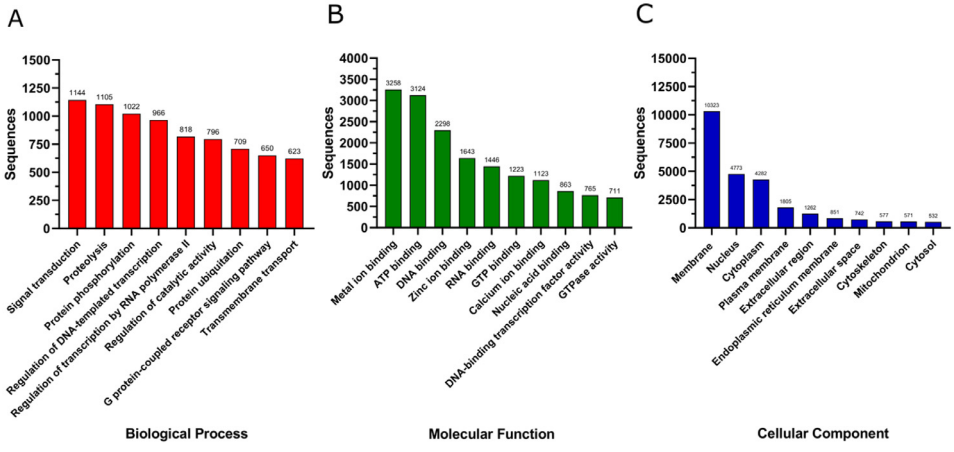

Fig. 5. Gene ontology classification.

Transcripts that mapped were then annotated with Gene Ontology (GO). The GOs were first classified by function under biological processes, molecular function, or cellular component categories and then further classified into sub-categories The top 10 subcategories are listed in Fig. 5. Signal transduction is the most highly expressed biological process, while metal ion binding and membrane transcripts were the most expressed in molecular function and cellular component respectively.

2.7. Clusters of orthologous genes annotation

The EGGNOG mapper tool was used to perform Clusters of Orthologous Genes (COG) analysis on the transcriptome (Fig. 6A). Transcription had the greatest number of transcripts assigned to it out of all the Information Storage and Processing functions. Signal Transduction had the greatest number of transcripts classified out of all the cellular processes and signal functions. Signal Transduction was also the most classified category overall. Carbohydrate transport and metabolism had the greatest number of transcripts out of all the metabolism functions.

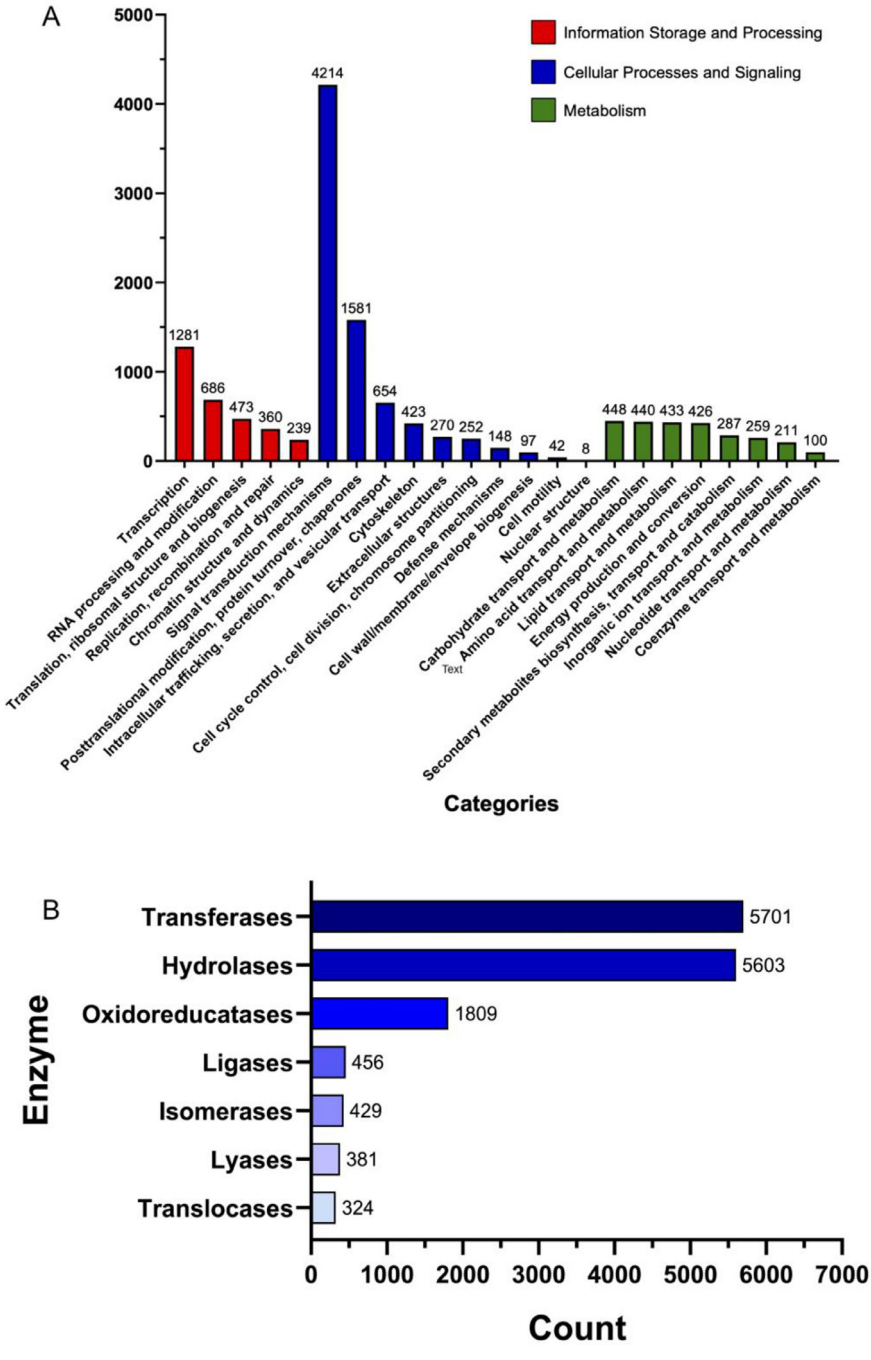

Fig. 6. A) Clusters of orthologous groups annotation of transcriptome. B) Enzyme annotation.

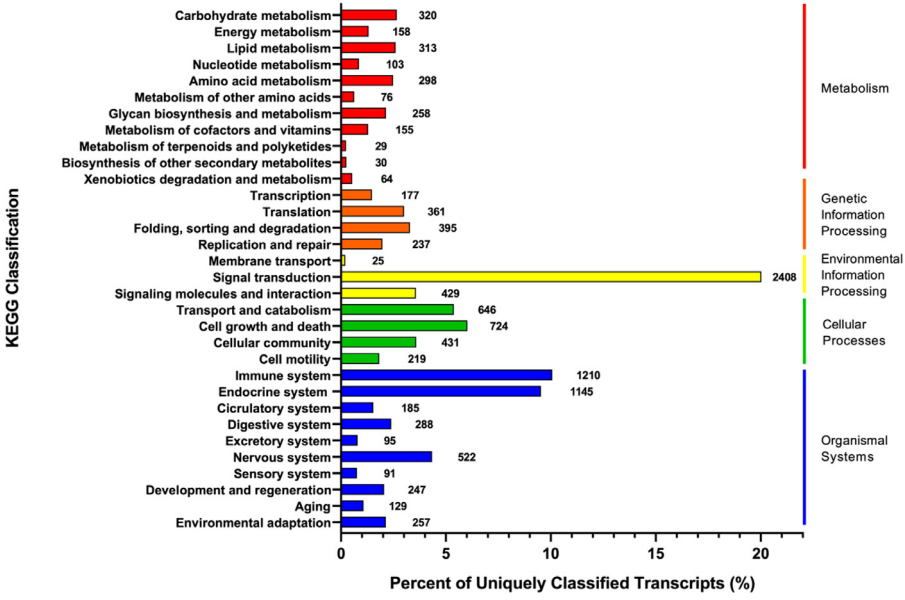

Fig. 7. Functional classification of KO identified in transcriptome.

Enzyme distribution was then assessed against the Nr database. Fig. 6B shows the number of transcripts expected to express the specific class of enzyme. Transferases were the most expressed class of enzyme with hydrolases coming second.

2.8. Functional annotation against KEGG database

From the 42,474 transcripts that aligned to the KEGG database, 8770 unique KEGG orthologs (KO) were identified in the transcriptome (Fig. 7). These orthologs were then classified by function. Note that more than sum of the bars in the figure do not add to 8770 as KEGG orthologs can be involved in multiple pathways and thus have multiple functions. Carbohydrate "metabolism", "folding, sorting, and degradation", "signal transduction", "cell growth and death", and "immune system" were the top functions in each KEGG category.

3. Experimental Design, Materials and Methods

3.1. RNA extraction

Total RNA was extracted from a 25 mg tail fin and bone clipping of a recently deceased aquarium-raised angelfish using the Zymo Quick-RNA Miniprep Plus kit following the standard protocol. Tail fin tissue was initially preserved in Longhorn PrimeStore MTM. Approximately 3.1 µg of total RNA was isolated at a concentration of 57 ng/µL as measured with a Denovix DS-11 spectrophotometer. 260/280 and 260/230 wavelength ratios for the extracted RNA were 1.92 and 2.11 respectively which indicated a high purity extraction. 100 ng of extracted total RNA was run on a 1% TAE agarose gel for 45 min at 120 Vs to crudely validate the quality of the extraction. No smearing was visible on the gel, and two clear bands corresponding to the 28S and 18 s rRNA were identified. Total RNA was stored in nuclease-free water at −80 °C until use.

3.2. Library preparation and sequencing

The standard Nanopore cDNA-PCR (PCS-111) library preparation was performed on 228 ng of total RNA. While selecting for full-length transcripts with PCR, extension time was set to 3 min to limit bias towards shorter transcripts and ensure through amplification. 14 PCR cycles were performed. The final library was 132 ng at a concentration of 11 ng/μL. 44 ng of the cDNA library were loaded into a single R9.4.1 flow cell on a MinION Mk1B device. Unpaired long reads were base called at High Accuracy with the Guppy base caller. Sequencing was performed for 82 h, with a flush buffer "refuel" at 58 h.

3.3. Quality control and adapter processing

Pychopper was used to trim sequencing adapters from the Nanopore reads. Only reads that successfully passed adapter removal and had a quality score above 7 were used in downstream analysis. Read rescue was attempted on unclassified reads that did not initially pass the adapter trimming, and any successfully rescued reads were kept. All other reads were discarded. Read statistics (Table 1) were generated with Nanoq [5].

3.4. Transcriptome assembly

A combined mapping and de-novo assembly strategy was employed to assemble the transcriptome of the angelfish. Minimap2 (-ax splice) was first used to map the cDNA reads to the angelfish genome [6]. Due to the high fragmentation of the angelfish genome, reads were also mapped to the genome of *Archocentrus centrarchus* (GCA 007364275.2), a close relative. StringTie (-L -m 50) was used to assemble transcripts from both mappings [7]. Next, Trinity was used to perform a de-novo assembly of the same reads to better identify transcripts that the genomes may not cover well [8] with (–min contig length 50). Thus, three independent transcriptomes were assembled.

3.5. Redundancy and contamination removal

These three assemblies needed to be combined into a single uniform assembly. Simply concatenating the three assemblies led to high duplication of transcripts as assessed with BUSCO. The CD-Hit tool, the Evidential Gene pipeline, and the assembling of super transcripts with Lace were tested to reduce redundancy [9–11]. However, all of these methods led to a significant decrease in completeness. To address this issue, a new mapping strategy was employed. Since the *A. centrarchus* mapping assembly had the highest completeness in the Actinopterygii lineage, the other two assemblies were mapped against it with Minimap2 (-ax asm5). Transcripts that successfully mapped to the *A. centrarchus* assembly were dropped, and the remaining unmapped transcripts were kept in the final assembly. Approximately 25% of the transcripts assembled against the angelfish genome, and 35% of the de-novo assembled transcripts did not map and were subsequently kept. The employed mapping strategy remained the best at minimizing duplication and maximizing completeness. The Kraken tool was then used to remove contaminant bacterial, archaeal, and viral transcripts from the final assembly [12].

3.6. Functional annotation

Functional annotation of the transcriptome was performed against four databases: NCBI non-redundant (Nr), EggNOG, KEGG, and Pfam [13–17]. To map transcripts against Nr, EggNOG and

Pfam, the OmicsBox bioinformatics software was employed [18]. Gene ontology terms were then assigned with OmicsBox to the successfully mapped Nr transcripts with standard pre-defined parameters. Enzyme classification was also generated with the Omics Box tool. COG annotation was performed with the EggNOG mapper tool inbuilt with OmicsBox. The KEGG Automatic Annotation server (https://www.genome.jp/kegg/kaas/) was used to generate KEGG Orthology assignments with the single directional best hit method. KEGG mapper - reconstruct (www.genome.jp/kegg/mapper/reconstruct.html) was then used to functionally classify the KO assigned transcripts. Transdecoder was used to determine coding potential of the transcripts [19]. The possible protein coding sequences were then blasted against the NR, Ref-Seq, and Swissprot databases with Diamond [20].

Limitations

This dataset is limited to sequencing data from a single individual and a limited number of sampled tissues. The data provided here likely does not represent the full transcriptome space of the angelfish. Future sequencing depth in multiple individuals and tissue types is required to build a more thorough transcriptome. Nanopore technology also has lower accuracy compared to alternative short-read sequencing technologies.

Ethics Statement

The angelfish used in this work was raised by the author from birth and its parents were purchased from local pet stores. The angelfish passed away due to natural causes prior to the start of experimentation and was not harmed for the purpose of this research. Although IACUC approval was not required for an already deceased animal, all US and journal rules and regulations were followed in the handling of the biological material.

Data Availability

Pterophyllum scalare Transcriptome Assembly (Original data) (NCBI)

CRediT Author Statement

Indeever Madireddy: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgments

I am deeply grateful to Johan Sosa and Kurt Chang from BioCurious for their continual guidance throughout the course of this work. I would also like to thank Yuanyu Lin from the University of North Carolina, Chapel Hill for his advice on the manuscript. I am also thankful to Dr. Karen Allendoerfer for her continual support. The flow cell and library preparation kit used in this work were provided as a courtesy of Oxford Nanopore Technology. No external monetary funding was received.

Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K.E. Powder, R.C. Albertson, Cichlid fishes as a model to understand normal and clinical craniofacial variation, Dev. Biol. 415 (2) (2016) 338–346.
[2] M. Schartl, Beyond the zebrafish: diverse fish species for modeling human dis- ease, Dis. Model. Mech. 7 (2) (2014) 181–192 h t 10.1016/j.ydbio.2015.12.018, doi:10.1242/dmm.012245.
[3] I. Madireddy, First ever whole genome sequencing and de novo assembly of the freshwater angelfish, pterophyllum scalare, MicroPubl. Biol. 2022 (2022), doi:10.17912/micropub.biology.000654.
[4] F.A. Sim ão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single- copy orthologs, Bioinformatics 31 (19) (2015) 3210–3212 Accessed 23 June 2023, doi:10.1093/bioinformatics/btv351.
[5] E. Steinig, L. Coin, Nanoq: ultra-fast quality control for nanopore reads, J. Open Source Softw. 7 (69) (2022) 2991 Accessed 23 June 2023, doi:10.21105/joss.02991.
[6] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinform. (Oxf., Engl.) 34 (18) (2018) 3094–3100, doi:10.1093/bioinformatics/bty191.
[7] M. Pertea, G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, S.L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nat. Biotechnol. 33 (3) (2015) 290–295, doi:10.1038/nbt.3122.
[8] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. LeDuc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat. Protoc. 8 (8) (2013) 1494–1512 Accessed 23 June 2023, doi:10.1038/nprot.2013.084.
[9] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152, doi:10.1093/bioinformatics/bts565.
[10] N.M. Davidson, A.D.K. Hawkins, A. Oshlack, SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes, Genome Biol. 18 (1) (2017) 148 Accessed 23 June 2023, doi:10.1186/s13059-017-1284-1.
[11] Gilbert, D.: Gene-omes built from mRNA seq not genome DNA, Notre Dame (2013). http://arthropods.eugenes.org/EvidentialGene/about/EvigeneRNA2013poster.pdf (Accessed 7 April 2023).
[12] D.E. Wood, S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol. 15 (3) (2014) 46, doi:10.1186/gb-2014-15-3-r46.
[13] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410 Accessed 23 June 2023, doi:10.1016/S0022-2836(05)80360-2.
[14] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hern ández-Plaza, S.K. Forslund, H. Cook, D.R. Mende, I. Letunic, T. Rattei, L. Jensen, C. von Mering, P. Bork, eggNOG 5.0: a hierarchical, functionally and phylogenetically anno- tated orthology resource based on 5090 organisms and 2502 viruses, Nucleic Acids Res. 47 (D1) (2019) 309–314 Accessed 23 June 2023, doi:10.1093/nar/gky1085.
[15] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30, doi:10.1093/nar/28.1.27.
[16] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, M. Ishiguro-Watanabe, KEGG for taxonomy-based analysis of pathways and genomes, Nucleic Acids Res. 51 (D1) (2023) 587–592, doi:10.1093/nar/gkac963.
[17] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. Salazar, E.L.L. Sonnham- mer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman, Pfam: the protein families database in 2021, Nucleic Acids Res. 49 (D1) (2021) 412–419, doi:10.1093/nar/gkaa913.
[18] B. Bioinformatics, OmicsBox – Bioinformatics Made Easy (2019). https://www.biobam.com/omicsbox.
[19] B.J. Haas, TransDecoder (2023). https://github.com/TransDecoder/TransDecoder (Accessed 23 June 2023).
[20] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (1) (2015) 59–60 Accessed 23 June 2023, doi:10.1038/nmeth.3176.