BMC Genomic Data

DATA NOTE

Open Access

First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima*

Kishor Kumar Sarker[1,2], Liang Lu[1,2], Junman Huang[1,2], Tao Zhou[1,2], Li Wang[1,2], Yun Hu[1,2], Lei Jiang[1,2], Habibon Naher[3], Mohammad Abdul Baki[3], Anirban Sarker[3] and Chenhong Li[1,2*]

Abstract

Objectives: American shad (*Alosa sapidissima*) is an important migratory fish under Alosinae and has long been valued for its economic, nutritional and cultural attributes. Overfishing and barriers across the passage made it vulnerable to sustain. To protect this valuable species, aquaculture action plans have been taken though there are no published genetic resources prevailing yet. Here, we reported the first de novo assembled and annotated transcriptome of *A. sapidissima* using blood and brain tissues.

Data description: We generated 160,481 and 129,040 non-redundant transcripts from brain and blood tissues. The entire work strategy involved RNA extraction, library preparation, sequencing, de novo assembly, filtering, annotation and validation. Both coding and non-coding transcripts were annotated against Swissprot and Pfam datasets. Nearly, 83% coding transcripts were functionally assigned. Protein clustering with clupeiform and non-clupeiform taxa revealed ~82% coding transcripts retained the orthologue relationship which improved confidence over annotation procedure. This study will serve as a useful resource in future for the research community to elucidate molecular mechanisms for several key traits like migration which is fascinating in clupeiform shads.

Keywords: *Alosa sapidissima*, De novo transcriptome, Brain & Blood, Annotation

Objective

Alosa sapidissima is well discussed among the alosines for its biological, nutritional, and commercial calibre [1–4]. Their native range from the North Atlantic coast extends to several freshwater tributaries where come to reproduce by migrating, sometimes up to 1800 km upstream [5–7]. For high fecundity, marketable weight, and sport fishing, this anadromous fish receives an overwhelming demand, which drives up the exploitation. Numerous obstructions on their passage are limiting free movement and segregating the populations into patches [8–12]. Being sensitive to environmental changes, several reports have anticipated the extinction of shad species namely *Tenualosa. reevesii*, *T. thibaudeaui*, and *Alosa killarnensis* [13, 14]. Considering this risk, American shad restoration project and captive rearing has been undertaken in the USA and China, respectively. Despite these efforts, there is no large scale molecular information published to explain key traits that can strengthen a recovery program. Moreover, advanced omics technologies are producing vast amount of genomic data with precision. Therefore, we are reporting annotated transcriptomic resources from *A. sapidissima* for the first time.

*Correspondence: chli@shou.edu.cn
[1] Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai Ocean University, Shanghai 201306, China
Full list of author information is available at the end of the article

For a migratory species, it's a challenge to maintain the ionic-balance in body fluid at a steady-state as it requires a rhythmic alteration between solvent and solutes contents. Moreover, a well-developed signaling system is also required to switch from salt to fresh water and vice versa, and to feed live prey [15–18]. So, the current transcriptomic resource from blood and brain will aim to understand key biological features from molecular level for this precious species. Nevertheless, the resource was initially produced to compare with other shads, but the effort was halted due to biological material transfer incompatibilities during COVID-19 pandemic. Besides, WGS study of *A. sapidissimsa* is under consideration by the G10K consortium [19]. Thereafter, it would be useful to share the data with scientific community to make better use of it.

Data description

A mature individual of 42 cm in SL was euthanized with MS222(1gL^{-1}) prior to extract brain and blood tissues, which were immediately placed in ALLProtect buffer and EDTA-stabilized anticoagulant tubes, respectively and later preserved in −20 °C refrigerator [20]. Total RNA from each sample was extracted with TRIzol and 1 g was used to prepare cDNA libraries (~ 400 bp) for bridge

Table 1 Overview of all data files/data sets

Label	Name of data file/data set	File types (file extensions)	Data repository and identifier (DOI or accession number)
Data file 1	Method and Code availability	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.17056328 [24]
Data file 2	RNAseq-Brain	SRA file (.sra)	NCBI *Sequence Read Archive* https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR16474177 [25]
Data file 3	RNAseq-Blood	SRA file (.sra)	NCBI *Sequence Read Archive* https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR16474180 [26]
Data file 4	FigS1 Complete work flow	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.17054852 [27]
Data file 5	FigS2 Post trimming quality assessment	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.17054852 [27]
Data file 6	FigS3 Transcript length distribution	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.17054852 [27]
Data file 7	FigS4 BUSCO assessment	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.17054852 [27]
Data file 8	FigS5 Phylogenetic relationship	Image file (.jpg)	*Figshare* https://doi.org/10.6084/m9.figshare.17054852 [27]
Data file 9	Table S1 Preliminary assembly statistics	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.17054948 [28]
Data file 10	Table S2 Final non-redundant assembly statistics	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.17054948 [28]
Data file 11	Table S3 Annotation summery	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.17054948 [28]
Data file 12	Table S4 Species description	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.17054948 [28]
Data file 13	Table S5 Homologue information	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.17054948 [28]
Data file 14	brain.Trinotate.filtered.xls	Spreadsheet (.xls)	*Figshare* https://doi.org/10.6084/m9.figshare.16834564.v2 [29]
Data file 15	brain.Trinity.RSEM.retained.clustered.fasta	Fasta file(.fasta)	*Figshare* https://doi.org/10.6084/m9.figshare.16834564.v2 [29]
Data file 16	brain.Trinity.RSEM.retained.clustered.fasta.transdecoder.pep	Fasta file(.pep)	*Figshare* https://doi.org/10.6084/m9.figshare.16834564.v2 [29]
Data file 17	blood.Trinotate.filtered.xls	Spreadsheet (.xls)	*Figshare* https://doi.org/10.6084/m9.figshare.16834546.v2 [30]
Data file 18	blood.Trinity.RSEM.retained.clustered.fasta	Fasta file(.fasta)	*Figshare* https://doi.org/10.6084/m9.figshare.16834546.v2 [30]
Data file 19	blood.Trinity.RSEM.retained.clustered.fasta.transdecoder.pep	Fasta file(.pep)	*Figshare* https://doi.org/10.6084/m9.figshare.16834546.v2 [30]
Data file 20	Annotation from combined reads	Document file (.docx)	*Figshare* https://doi.org/10.6084/m9.figshare.19308326 [31]

amplification following the manufacturer's instructions. Finally, the purified libraries were loaded into Illumina Novaseq with 2*150 bp paired-end configuration. Raw sequencing reads were trimmed where the base accuracy was strictly confined to 99.99% (Data file 5). To perform assembly, the processed reads were passed through Trinity-v2.11.0 [21, 22] assembler that constructed 195,742 and 158,817 transcripts from brain and blood samples, respectively (Data file 9). The primary number of transcripts was reduced to 160,481 and 129,040 after filtering and clustering non-redundant transcripts at 98% threshold. Quantitative analysis identified 41,572 bp and 17,242 bp from the brain and blood transcriptomes as the longest transcripts with N50 values of 2039 bp and 2096 bp (Data file 10). In both instances, the assembly length distribution remained uniform and comparable to one another (Data file 6). In addition, BUSCO searches against 3354 species from vertebrate lineages found 82.3% and 71.5% of complete universal single-copy genes from brain and blood transcriptomes (Data file 7).

Implication of TransDecoder-v5.5.0 [22] predicted around 80% of assembled transcripts had an ORF, of which 48,579 and 40,948 transcripts were capable of producing functional proteins (Data file 11). Using Blastx, Blastp as well as a series of tools based on HMM, we annotated coding and non-coding transcripts with an e value cut-off at 10^-5. GO analysis ascertained 39,015 and 33,475 proteins had at least one relevant term with molecular function, cellular component or biological process. In both instances, search against Pfam database revealed 70% of proteins with a functional domain. According to the loaded Sqlite database from Trinotate [23], 83% of predicted proteins were functionally annotated. Moreover, we made an assembly and subsequent annotation combining the reads from both tissues. The entire effort and representative datasets can be found in Table 1 (Data file 1, Data file 4 and Data file 14-20). To draw the homologous relationship, we retrieved Refseq proteins of seven other species, including clupeiform and non-clupeiform species from NCBI repository (Data file 12). For brain and blood, we found that 40,304 and 34,301 proteins had orthologue relationships with other species accounting for >82% of total proteins (Data file 13). Finally, to evaluate the phylogenetic relationships, one-to-one orthologue proteins were retrieved. As the datasets from brain tissue extracted more groups of homologue proteins, we used 204 one-to-one orthologue proteins from brain to reconstruct a phylogenetic tree. We have found that *A. sapidissima* was clustered well with the clupeiform clade that was supported with maximum bootstrap value (Data file 8). The constructed phylogeny supports several other previous phylogenetic studies regarding their position [32–34]. However, this present resource will leverage the whole genome study of *A. sapidissima* as well as provide a solid foundation to compare their impressive physiological and behavioral competence with other allies.

Limitations

The sample was collected from freshwater captivity located at Songjiang District, Shanghai. Normally, when anadromous fish migrate to freshwater, they need to move against strong water currents and interact with particular abiotic factors. However, in captivity, possible absence of such physical properties might provide less chance to specific gene expression than during migration in the wild.

Abbreviations

SL: Standard Length; BUSCO: Benchmarking Universal Single Copy Orthologs; ORF: Open Reading Frame; HMM: Hidden Markov Model; GO: Gene Ontology; NCBI: National Canter for Biotechnology Information; WGS: Whole Genome Study; G10K: The international Genome 10 K consortium.

Acknowledgements

Our thanks go to the management team at the Lab of Molecular systematics and ecology for maintaining the High Performance Computation Server (HPCS) and supporting our data analysis. We also want to express our gratitude Mr. Roland Nathan Mandal and Miss. Irin Sultana for their technical support.

Authors' contributions

C.L. and K.K.S. designed the project and wrote the primary manuscript. L.J., L.W. and Y.H. collected and prepared the samples. K.K.S., L.L., J.H. and T.Z. performed the data analysis. All authors contributed in manuscript editing and revising the manuscript. The author(s) read and approved the final manuscript.

Funding

This work was supported by "Science and Technology Commission of Shanghai Municipality (19410740500)" and "Shanghai Collaborative Innovation for Aquatic Animal Genetics and Breeding project". Except funding, funder has no role in study design, sample collection, data analysis, and interpretation, or in manuscript writing.

Availability of data and materials

Processed raw data has been deposited in NCBI with open access (https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR16474177 & https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR16474180*)*. Method with its codes and references and all the final product of analysis has been submitted to *figshare* for public usage [24–31]. File type and specific accessible links can be found in Table 1.

Declarations

Ethics approval and consent to participate

All experimental procedures including specimen handling were approved by the Animal Ethics Committee of Shanghai Ocean University, China.

Consent for publication

Not applicable.

Competing interests

Authors are declaring no competing of interests.

Author details

[1]Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai Ocean University, Shanghai 201306, China. [2]Shanghai Collaborative Innovation for Aquatic Animal Genetics and Breeding, Shanghai Ocean

University, Shanghai 201306, China. [3]Department of Zoology, Jagannath University, Dhaka 1100, Bangladesh.

Received: 24 November 2021 Accepted: 18 March 2022
Published online: 28 March 2022

References

1. Limburg KE. American Shad in its native range. Am Fish Soc Symp. 2003;35:125–40.
2. Bi YH, Chen XW. Mitochondrial genome of the American shad Alosa sapidissima. Mitochondrial DNA. 2011;22(1-2):9–11.
3. Wang J, Yu ZS, Wang X, Yang SS, Zhang DG, Zhang Y. The next-generation sequencing reveals the complete mitochondrial genome of Alosa sapidissima (Perciformes: Clupeidae) with phylogenetic consideration. Mitochondrial DNA B. 2017;2(1):304–6.
4. Guo YJ, Xing ZK, Yang G, Liu JL, Chen CX, Xu DW. American shad muscle nutrition composition determination and analysis. China Feed. 2010;8:39–40.
5. Brown BL, Smouse PE, Epifanio JM, Kobak CJ. Mitochondrial DNA mixed-stock analysis of American Shad: coastal harvests are dynamic and variable. Trans Am Fish Soc. 1999;128(6):977–94.
6. Rasmussen JL, Regier HA, Sparks RE, Taylor WW. Dividing the waters: the case for hydrologic separation of the north American Great Lakes and Mississippi River basins. J Great Lakes Res. 2011;37(3):588–92.
7. Pearcy WG, Fisher JP. Ocean distribution of the American shad (Alosa sapidissima) along the Pacific coast of North America. Fish B-Noaa. 2011;109(4):440–53.
8. Harris JE, Hightower JE. Movement patterns of American Shad transported upstream of dams on the Roanoke River, North Carolina and Virginia. North Am J Fish Manage. 2011;31(2):240–56.
9. Haro A, Castro-Santos T. Passage of American Shad: paradigms and realities. Mar Coast Fish. 2012;4(1):252–61.
10. Grote AB, Bailey MM, Zydlewski JD. Movements and demography of spawning American Shad in the Penobscot River, Maine, prior to dam removal. Trans Am Fish Soc. 2014;143(2):552–63.
11. Mulligan KB, Haro A, Noreika J. Effect of backwatering a streamgage weir on the passage performance of adult American Shad (Alosa sapidissima). J Ecohydraulics. 2021:1–13. https://doi.org/10.1080/24705357.2021.1945500.
12. Hasselman DJ, Bentzen P, Narum SR, Quinn TP. Formation of population genetic structure following the introduction and establishment of non-native American shad (Alosa sapidissima) along the Pacific coast of North America. Biol Invasions. 2018;20(11):3123–43.
13. Guo Q, Liu XJ, Ao XF, Qin JJ, Wu XP, Ouyang S. Fish diversity in the middle and lower reaches of the Ganjiang River of China: threats and conservation. PLoS One. 2018;13(11):e0205116. https://doi.org/10.1371/journal.pone.0205116.
14. IUCN. The IUCN Red List of Threatened Species, vol. 2021-2; 2021.
15. Cao QQ, Gu J, Wang D, Liang FF, Zhang HY, Li XR, et al. Physiological mechanism of osmoregulatory adaptation in anguillid eels. Fish Physiol Biochem. 2018;44(2):423–33.
16. Mohindra V, Dangi T, Tripathi RK, Kumar R, Singh RK, Jena JK, et al. Draft genome assembly of Tenualosa ilisha, Hilsa shad, provides resource for osmoregulation studies. Sci Rep. 2019;9(1):16511. https://doi.org/10.1038/s41598-019-52603-w.
17. Xu GC, Bian C, Nie ZJ, Li J, Wang YY, Xu DP, et al. Genome and population sequencing of a chromosome-level genome assembly of the Chinese tapertail anchovy (Coilia nasus) provides novel insights into migratory adaptation. Gigascience. 2020;9(1):giz157. https://doi.org/10.1093/gigascience/giz157.
18. Finlay RW, Poole R, Rogan G, Dillane E, Cotter D, Reed TE. Hyper- and hypo-osmoregulatory performance of Atlantic Salmon (Salmo salar) Smolts infected with Pomphorhynchus tereticollis (Acanthocephala). Front Ecol Evol. 2021;9(529). https://doi.org/10.3389/fevo.2021.689233.
19. 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(7856):737.
20. Beekman JM, Reischl J, Henderson D, Bauer D, Ternes R, Peña C, et al. Recovery of microarray-quality RNA from frozen EDTA blood samples. J Pharmacol Toxicol Methods. 2009;59(1):44–9.
21. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–U130.
22. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512.
23. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A tissue-mapped axolotl De novo transcriptome enables identification of limb regeneration factors. Cell Rep. 2017;18(3):762–76.
24. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima* Figshare; 2021. https://doi.org/10.6084/m9.figshare.17056328.
25. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima*. Sequence Read Archive. 2021. https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR16474177.
26. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima*. Sequence Read Archive. 2021. https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR16474180.
27. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima* Figshare; 2021. https://doi.org/10.6084/m9.figshare.17054852.
28. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima* Figshare; 2021. https://doi.org/10.6084/m9.figshare.17054948.
29. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima*: Figshare; 2021. https://doi.org/10.6084/m9.figshare.16834564.v2.
30. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of an anadromous shad, *Alosa sapidissima* Figshare; 2021. https://doi.org/10.6084/m9.figshare.16834546.v2.
31. Sarker KK, Lu L, Huang J, Zhou T, Wang L, Hu Y, et al. First report of de novo assembly and annotation from brain and blood transcriptome of ananadromous shad, *Alosa sapidissima* Figshare 2022; 2022. https://doi.org/10.6084/m9.figshare.19308326.v1.
32. Bloom DD, Lovejoy NR. The evolutionary origins of diadromy inferred from a time-calibrated phylogeny for Clupeiformes (herring and allies). P Roy Soc B-Biol Sci. 2014;281(1778):20132081. https://doi.org/10.1098/rspb.2013.2081.
33. Bloom DD, Burns MD, Schriever TA. Evolution of body size and trophic position in migratory fishes: a phylogenetic comparative analysis of Clupeiformes (anchovies, herring, shad and allies). Biol J Linn Soc. 2018;125(2):302–14.
34. Hughes LC, Orti G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci U S A. 2018;115(24):6249–54.

Publisher's Note