



Genome Resources

A high-quality genome assembly of the Laotian shad (*Tenualosa thibaudeaui*), an endemic species of the Mekong River Basin

Li Wang^{1,2,*}, Liang Lu^{1,2,*}, Kishor Kumar Sarker^{1,2}, Chenhong Li^{1,2}

¹Engineering Research Center of Environmental DNA and Ecological Water Health Assessment, Shanghai Ocean University, Shanghai, China,

²Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai Ocean University, Shanghai, China

*These authors contributed equally to this work.

Address correspondence to C. Li at the address above, or e-mail: chli@shou.edu.cn.

Corresponding Editor: Klaus-Peter Koepfli

Abstract

The Laotian shad (*Tenualosa thibaudeaui*) belongs to the family Clupeidae and is mainly distributed across Lao PRD, Cambodia, and northern Thailand. Due to overfishing and dam reconstruction, the Laotian shad is on the verge of extinction and currently listed as vulnerable by the IUCN. Nanopore and Illumina sequencing data were integrated to generate the first high-quality genome assembly for *T. thibaudeaui*. The assembled genome was 638 Mb in size, including 228 scaffolds with a N50 value of 16.6 Mb. BUSCO analysis revealed the completeness of the assembly to be more than 96%. A total of 24,810 protein-coding genes were predicted. According to the pairwise sequentially Markovian coalescent analysis, the effective population size of the Laotian shad sharply declined from 3 Mya to 20 Kya. We found a significant ratio in contraction of gene families that may reflect secondary gene loss. Our high-quality genome assembly of the Laotian shad will provide a valuable resource for future research in conservation genetics, as well as for investigating the phylogenetics and comparative genomics of shads.

Key words: effective population size, gene families, Laotian shad, Nanopore sequencing

Introduction

The Laotian shad (*Tenualosa thibaudeaui*) (Durand 1940), also known as the Mekong herring, is an endemic species of the Mekong River Basin distributed from northern Thailand, Cambodia, Lao PDR, and possibly into Myanmar and Southern China (Blaber et al. 2003; Poulsen et al. 2004). Although it has become increasingly rare since the 1970s, the species was last recorded in large numbers around 1984 (Roberts 1993; Stone 2016). Because of overfishing and hydropower projects, fishery resources in the Mekong River are declining with an irreversible impact. Assessment based on availability and habitat management, the IUCN declared *T. thibaudeaui* as a “Vulnerable” species in 2011 (Vidthayanon 2011). In the following years, Vu et al. (2015) warned about the risk of extinction for *T. thibaudeaui* along with many other species found in the Mekong Delta. A multidimensional conservation action plan is therefore required for this species but until now, the research related to the Laotian shad has been limited mostly to resource surveys. Meanwhile, recent conservation strategies are becoming more improved by utilizing species-specific genetic data like adaptive alleles to track evolutionary consequences of demographic changes. Irrespective to their importance, little

genetic data have been reported so far for *T. thibaudeaui*. A mitochondrial genome sequenced for phylogenetic analysis by Lavoué et al. (2013) and protein sequences by Egan et al. (2018) to determine trophic niches of Clupeiformes are now the only accessible data through public repositories. In these circumstances, sequencing a whole genome resource for the only potamodromous taxon from the *Tenualosa* clade can aid molecular and conservation efforts. In addition, a series of bioinformatics tools have emerged along with the development of genomics, and now uncovering the genome of *T. thibaudeaui* can make it possible to assess the diversified molecular mechanism of migration throughout the clade. Considering these aspects, firstly, we assembled a high-quality genome for *T. thibaudeaui* using Nanopore and Illumina reads. Secondly, the pairwise sequentially Markovian coalescent (PSMC) model was employed to infer the demographic history from the diploid genome. Finally, we compared the whole genome with all the available genomes of closely related species and outgroups and modeled the gene gain and loss over a time calibrated tree. We are optimistic that our current genome assembly will be considered as a valuable resource for the conservation of this highly threatened species.

Methods

Sample collection, DNA library preparation, and sequencing

Two individuals of *T. thibaudeaui* were collected from Paksan, Bolikhamsai Province, a central city of Laos (18°21'5.328"N, 103°40'33.239"E) in December 2019 (Fig. 1), and the larger male specimen was used for genome sequencing. Muscle tissues were taken and preserved in All Protect Buffer solution (Beyotime, cat#R0121-100) at -80 °C for further usage. The Ezup Column Animal Genomic DNA Purification Kit (Sangon Biotech) was used to extract HMW genomic DNA and raw DNA was sent to GrandOmics (Wuhan, Hubei, China) for long-read sequencing using the Oxford Nanopore PromethION platform. For short reads, DNA was sheared to ~250 bp using a Covaris M220 focused ultrasonicator (Woburn, MA, USA). Two separate libraries were constructed using a modified method (Meyer and Kircher 2010; Li et al. 2013), and sent to GENEWIZ (Suzhou, Jiangsu, China) and Novogene (Tianjin, China) for sequencing on the Illumina NovaSeq 6000 platform.

Data filtering and genome size estimation

Trim Galore with default settings (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to trim Illumina reads. The parameter “-m 21-s 1G” was used in Jellyfish (Marcais and Kingsford 2011) to count the k-mer frequency. Based on the k-mer distribution, genome size and heterozygosity were estimated using GenomeScope (Vurture et al. 2017). For Nanopore reads, we only kept data with an average quality score larger than 7. The other programs with version details used throughout the analysis procedure are available in Table 1.

Genome assembly

NECAT (Chen et al. 2021) was used to assemble the Nanopore reads. Since the assembly from long reads contained a considerable number of errors, Illumina reads were used to polish the draft assembly employing NextPolish (Hu et al. 2020). BUSCO (Waterhouse et al. 2018) with a database containing 3,354 core genes from the vertebrate lineage was used to assess genome completeness. Purge Haplotigs (https://bitbucket.org/mroachawri/purge_haplotigs/commits/) was used to resolve and treat heterozygous regions of the assembly. Being a non-model species, a de novo library of repeats was built for *T. thibaudeaui* using RepeatModeler (<http://www.repeatmasker.org/RepeatModeler/>) with default parameters prior to masking of the repetitive regions. The library was used as queries for RepeatMasker (<http://www.repeatmasker.org/>) to quantify the repeat content. Before masking, we blasted the library using BLAST (Altschul et al. 1990) against all the available proteins of Clupeiformes from NCBI since transposable elements often can code for certain proteins. We wrote and implemented a python script to remove the repeat content having ≥70% identity threshold with any known protein domain. Finally, only the repeats that did not code for any proteins were retained as true hits.

Gene prediction and annotation

For annotation, we utilized raw RNA-seq data of *Tenualosa ilisha* (Bioproject: PRJNA378699), a sister species to *T. thibaudeaui*. After quality checking, raw reads were assembled with Trinity (Haas et al. 2013) using the reference-free method. Protein-coding genes were annotated using the MAKER pipeline (Holt and Yandell 2011) which is composed of ab initio and evidence-based gene prediction. The HMM within SNAP and Augustus (Stanke and Waack



Fig. 1. Specimens of *Tenualosa thibaudeaui*. The bottom specimen (male) was used for genome sequencing.

Table 1. Programs used in genome assembly, annotation, and analysis of *Tenualosa thibaudeaui*.

Purpose	Program	Version
Data filtering and size estimation		
Illumina read trimming	Trim Galore	0.4.1
Nanopore read filtering	Nanofilt	2.7.1
K-mer analysis	Jellyfish	2.2.10
Genome size evaluation	GenomeScope	1.0.0
Genome assembly		
Long-read assembly	NECAT	0.0.1
Error correction	NextPolish	1.4.0
Assembly validation	BUSCO	3.0.1
Redundancy removal	Purge_haplotigs	1.1.2
Gene prediction and annotation		
Repeat sequence identification	RepeatModeler	1.0.8
Masking of repetitive elements	RepeatMasker	4.0.7
Transcriptome assembly	Trinity	2.11.0
Annotation pipeline	Maker	2.31.10
Population history	PSMC	0.6.5
Gene family analysis		
All-by-all BLAST	BLAST	2.10.0
Gene family clustering	Orthofinder	2.5.4
Multiple alignment	Muscle	5.0
Phylogenetic tree	RaxML	8.2
Expansion and contraction	CAFÉ	4.2.1

2003) with the *Danio rerio* gene set were trained to determine the position of genes in the genome. Proteins from the Order Clupeiformes (<https://www.ncbi.nlm.nih.gov/protein/?term=Clupeiformes+>) and assembled transcripts of *T. ilisha* were used for evidence-based gene prediction. The entire procedure was iterated 3 times. Parameter settings are listed in [Supplementary File 2](#). For functional annotation, all products from the coding genes were compared with databases using InterProScan5 (Jones et al. 2014) and KEGG mapper (Ogata et al. 1999).

Analysis of population history of *T. thibaudeaui*

The PSMC method (Li and Durbin 2011) was applied to infer the population history of *T. thibaudeaui*. BWA-MEM algorithm (Li 2013) was used to map 28× pair-end reads against the newly assembled diploid genome of *T. thibaudeaui*. SAMTools (Li et al. 2009) was used to generate the consensus with default settings (<https://github.com/lh3/psmc>). The species base substitution rate (μ) and generation time (g) were set to 3.75×10^{-9} (Li et al. 2021) and 2 yr (Roberts 1993), respectively.

Gene family analysis

To identify the orthologous genes, the protein sequences from 8 species (*Clupea harengus*, *T. ilisha*, *T. thibaudeaui*, *Alosa sapidissima*, *Alosa alosa*, *D. rerio*, *Denticeps clupeoides*, and *Chanos chanos*) were given as input for Orthofinder (Emms and Kelly 2019). Only proteins from the longest isoforms were kept to remove redundancy. After performing an all-by-all BLAST search, a Markov clustering algorithm was applied to cluster the protein

sequences (Fischer et al. 2011). After clustering all the homologous genes, 1,379 single copy genes were retrieved and concatenated into a supermatrix. Multiple alignment was done at the amino acid level using MUSCLE (Edgar 2004) and then translated back to nucleotides for subsequent analyses. An alignment of 4.7 Mb in length was used to reconstruct a maximum likelihood tree under GTR+G model with 1000 bootstraps using RaxML (Stamatakis 2014). To provide calibration points for r8s, the divergence time between *A. sapidissima* and *A. alosa* was selected from the TimeTree.org website (<http://timetree.org/>) (Sanderson 2003). The binary format of the ultrametric tree and cluster information were used as the main parameters to run CAFÉ (Han et al. 2013). Rapidly changing gene families were used to construct the Venn Graph (<http://jvenn.toulouse.inra.fr/app/example.html>).

Results

Genome assembly and evaluation

After filtering low quality and duplicated reads, the Illumina sequencing yielded 32 Gb reads with a depth of 28×. From the Nanopore sequencing, 111 Gb of long reads were generated and only reads >10 kbp were retained. The genome size was estimated approximately as 656 Mb according to the k-mer count. An initial assembly using long reads generated a draft genome with an N50 of 15 Mb. To improve the quality, the assembly was polished with paired-end reads. After polishing, the genome size of *T. thibaudeaui* was 638 Mb consisting of 228 scaffolds of which 189 of them were greater than 5 Mb, accounting for 82.9% of the genome. Consequently, the N50 of the genome was increased to 16.6 Mb and the longest scaffold size was 28.76 Mb (Table 2). BUSCO analysis found that approximately 96.6% out of 3,354 core genes remained fully intact (Supplementary File 1, Table S2).

Repetitive sequences, gene prediction, and annotation

Prior to structural annotation, the position and quantity of repetitive elements were identified, which contributes to 24.64% of the whole genome. Altogether, 164.8 Mb of simple sequence repeats were identified including interspersed repeats (17.44%), simple repeats (6.21%), and low complexity regions (0.99%) (Supplementary File 1, Table S3). A total of 24,810 protein-coding genes from the *T. thibaudeaui* genome were structurally annotated. Predicted genes were distributed on 228 scaffolds with an average of 109 genes per scaffold.

Table 2. Assembly statistics for the genome of *Tenualosa thibaudeaui*.

Metric	Values
Assembly size (nt)	668,877,428
Number of scaffolds	228
Number of long scaffolds (≥5 Mb)	189
Mean sequence length (nt)	2,933,673
Scaffold N50 (nt)	17,052,863
Longest scaffold (nt)	28,763,990
GC content %	42.94

For functional annotation, the predicted genes were searched against known databases. Among those, 23,879 and 15,624 genes were annotated by InterProScan and KEGG Mapper, respectively, which represents 96.25% of the total number of genes (Supplementary File, Table S4).

Population dynamics of *T. thibaudeaui*

The effective population size (N_e) of *T. thibaudeaui* was estimated to have changed during the period from 2 Mya to 10 Kya using PSMC. The effective population size experienced a significant decline from 2 Mya to 200 Kya, although the demographic curve was nearly stable between 200 and 50 Kya and the number remained around 4.0×10^4 . Nonetheless, a slight rebound was observed from 50 to 10 Kya. The demographic trajectory within the given time scale is presented in Fig. 2.

Evolution of gene families

The expansion and contraction analysis was performed using inferred gene families and the estimated phylogenetic tree, and we particularly focused on the comparison between the potamodromous *T. thibaudeaui* and anadromous *T. ilisha*. The time tree showed that *T. thibaudeaui* diverged from its sister species about 12.72 Mya ago (Fig. 3). A total of 203 gene families of *T. thibaudeaui* have been expanded whereas 2,130 gene families have been contracted. Surprisingly, the scenario was quite opposite for the *T. ilisha* where 2,542 and 415 gene families have been expanded and contracted, respectively. We noted that 47 gene families had experienced distinct changes likely according to adaptive preferences, contracting in *T. thibaudeaui* and expanding in *T. ilisha*. However, a Venn Graph using 4 species from 3 representative habitats showed that only 1 gene family (PF07686) containing the Immunoglobulin V-set domain had undergone rapid contraction in *T. thibaudeaui* and *C. harengus*, species that live in only 1 type of habitat i.e. freshwater or saltwater, but showed rapid expansion in the migratory species *T. ilisha* and *A. sapidissima*. (<https://pfam.xfam.org/family/PF07686>).

Discussion

Currently, genomic resources are important in conservation strategies since reference genomes can inform adaptive features from genic and intergenic regions (Formenti et al. 2022). From a conservation perspective, the natural

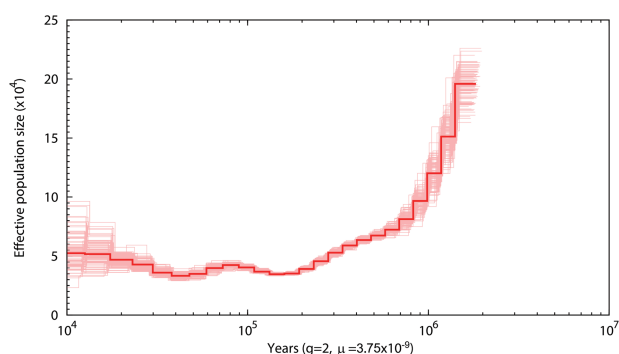


Fig. 2. Historical population size of *Tenulosa thibaudeaui* from 2 Ma to 10 Ka estimated using the PSMC method with 100 bootstraps.

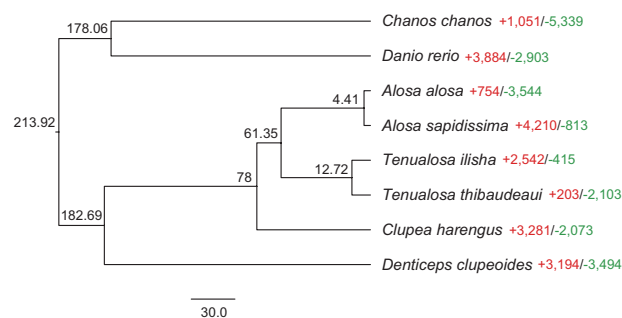


Fig. 3. A time tree of *Tenulosa thibaudeaui*. The “+” sign indicates the number of expanded gene families, whereas “-” sign indicates the number of contracted gene families.

existence of *T. thibaudeaui* is under serious threat but there are very limited genetic resources. Even for the entire clade of *Tenulosa* (5 species), the genomic resources are yet insignificant to understand the dispersal, adaptation, and diversification of these species across their range of habitats. In order to produce a reference genome for the most vulnerable species from this group, we integrated Nanopore and Illumina reads to accomplish genome assembly of *T. thibaudeaui*, resulting in 228 scaffolds with 99.1% of them over 10 kb and 62.7% over 100 kb in length, which reflected a high level of contiguity. In addition, the completeness of the assembly was higher compared with the existing genome assembly of *T. ilisha*, but the genome size of *T. thibaudeaui* was sufficiently shorter (~17%) (Mohindra et al. 2019). The effective population size of *T. thibaudeaui* showed a downward trend with a relatively low recovery curve (Fig. 2). According to the literature, this decline coincided with the Quaternary glaciation about 2 to 3 million years ago (Kerdelhué et al. 2009). Suspecting the degree of environmental sensitivity by *T. thibaudeaui*, it may be inferred that habitat deterioration brought by temperature changes during the global cooling period led to an extensive decrease in population size (Blaber et al. 2003; Riede 2004). However, there are still 2 possible populations of *T. thibaudeaui* which have been divided by Khong Phapheng Falls, the world’s widest waterfall, into an upstream and downstream cluster (Poulsen et al. 2004). The present genome assembly resource can be used as a baseline for investigating the drift or loss of genetic variability within these populations.

Gene families consist of genes sharing similar molecular functions that evolve or disappear as a result of gene duplication or secondary deletion events, respectively. Over time, the increase or decrease in the copy number of genes from certain families can be a result of both random processes and natural selection (Demuth et al. 2006). In the present study, the expansion and contraction analysis of gene families revealed a remarkable ratio for the Laotian shad in terms of gene gain and loss. The ratio of contraction compared with expansion was strikingly higher, ~10 times higher than that observed in the other species. The particular comparison between anadromous and potamodromous shad species revealed a pattern for contraction, where several crucial members including SHC-transforming protein, DNA-binding protein inhibitor, and Bone morphogenetic proteins (BMP) were found to be contracted for *T. thibaudeaui*. Each of these families might have limited role for potamodromous movement but may be highly essential for diadromous movement. For example,

members from the BMP family preliminarily are responsible for bone maintenance but these also have a clinical role in homeostasis (Wang et al. 2014). Furthermore, proteins from the SHC family are involved in enzymatic activity to draw signaling molecules to respective pathways (Ahmed and Prigent 2017). Since anadromous migrants need to tolerate a wide range of salinity compared with freshwater dwellers and require a well developed signaling system, it might be necessary to be equipped with more genes related to homeostasis. Experimental evidence is required to test these ideas for the species under discussion. However, the habitat range of the Laotian shad is sharply limited compared with the anadromous species which might be 1 reason for the significant contraction of copy numbers of certain gene families. Since secondary gene loss is common throughout the tree of life after gene duplication events according to habitat preferences and adaptive requirements, the present results raise 2 key questions: First, did *T. thibaudeaui* evolve from a marine ancestor? Second, is it a basal species to all the remaining members of *Tenualosa*? Previous phylogenetic studies suggested a freshwater origin of the genus *Tenualosa* (Bloom and Lovejoy 2014), but we encourage a phylogenomic study to address these questions.

More importantly, reference genomes are generally used to guide the sequencing and assembly of new genomes from closely related species and for population genomic studies (Brandies et al. 2019). For the genus *Tenualosa*, there are 5 species with oceanodromous, anadromous, and potamodromous life history. With the exception *T. ilisha* and now *T. thibaudeaui*, there are no other genomic resources available for this economically important group of fish. Thus, our genome resource will facilitate the genome assembly of other members of *Tenualosa* and enable comparative genome studies of shads from various habitats.

Conclusion

The combined assembly from Nanopore and Illumina reads generated a high-quality reference genome for *T. thibaudeaui* with good contiguity (228 scaffolds) and completeness (96.6%). The annotated genome can provide information about genomic regions having an influential role in local adaptations within certain ecotypes. This may help to execute viable conservation action plans like translocation. Moreover, this will facilitate better comparative genomic studies among *Tenualosa* shads when more genome resources of the other species become available.

Supplementary material

Supplementary material is available at *Journal of Heredity* online.

Funding

This work was supported by “Science and Technology Commission of Shanghai Municipality (19410740500)” and Asian Cooperation Fund Project (125A0607).

Acknowledgments

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. The authors would like to thank Shanghai Oceanus Supercomputing Center (SOSC) for providing their computational resources. Thanks to A. Sarker, Q. Xue, and Souliphet Laolouteng for their contributions to sample collection.

Data availability

The raw reads have been uploaded, Illumina reads are under number of SRR15422198 (<https://dataview.ncbi.nlm.nih.gov/object/SRR15422198>), Nanopore reads are under number of SRR15422199 (<https://dataview.ncbi.nlm.nih.gov/object/SRR15422199>). The genome assembly file is under accession number of PRJNA753982 (<https://dataview.ncbi.nlm.nih.gov/object/PRJNA753982>).

References

- Ahmed SBM, Prigent SA. Insights into the Shc family of adaptor proteins. *J Mol Signal*. 2017;12(2):1–17.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–410.
- Blaber SJM, Milton DA, Brewer DT, Salini JP. Biology, fisheries, and status of tropical shads *Tenualosa* spp. in south and southeast Asia. In: American Fisheries Society Symposium. Cleveland, Queensland 4163, Australia. 2003. p. 49–58. <https://www.researchgate.net/publication/289263909>
- Bloom DD, Lovejoy NR. The evolutionary origins of diadromy inferred from a time-calibrated phylogeny for Clupeiformes (herring and allies). *Proc Biol Sci*. 2014;281(1778):20132081.
- Brandies P, Peel E, Hogg CJ, Belov K. The value of reference genomes in the conservation of threatened species. *Genes (Basel)*. 2019;10(11):846–862.
- Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, Wang YX, Xing JF, Huang ZJ, Wang DP, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun*. 2021;12(1):1–10.
- Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. The evolution of mammalian gene families. *PLoS One*. 2006;1(1):e85.
- Durand J. *Notes sur quelques poissons d'espèces nouvelles ou peu connues des eaux douces Cambodgiennes*. Institut Océanographique de l'Indochine. Note 36; 1940. p. 1–40. <https://www.semanticscholar.org/paper/Notes-sur-quelques-poissons-d'esp%C3%A8ces-nouvelles-ou-Durand/d9812e5ec288e0d7d0875fc692dd4877b3e33593>
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–1797.
- Egan JP, Bloom DD, Kuo CH, Hammer MP, Tongnunui P, Iglesias SP, Sheaves M, Grudpan C, Simons AM. Phylogenetic analysis of trophic niche evolution reveals a latitudinal herbivory gradient in Clupeoidei (herrings, anchovies, and allies). *Mol Phylogenet Evol*. 2018;124:151–161.
- Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20(1):1–14.
- Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics*. 2011;35(1):6–12.
- Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crotini A, Godoy JA, Höglund J, et al.; European Reference Genome Atlas (ERGA) Consortium. The era of reference genomes in conservation genomics. *Trends Ecol Evol*. 2022;37(3):197–202.
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. De novo transcript

- sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 2013;8(8):1494–1512.
- Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol.* 2013;30(8):1987–1997.
- Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.* 2011;12(1):1–14. doi:10.1186/1471-2105-12-491
- Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics.* 2020;36(7):2253–2255.
- Jones P, Binns D, Chang HY, Fraser M, Li WZ, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30(9):1236–1240.
- Kerdelhué C, Zane L, Simonato M, Salvato P, Rousselet J, Roques A, Battisti A. Quaternary history and contemporary patterns in a currently expanding species. *BMC Evol Biol.* 2009;9:220.
- Lavoué S, Miya M, Musikasinthorn P, Chen WJ, Nishida M. Mitogenomic evidence for an Indo-West Pacific origin of the Clupeoidei (Teleostei: Clupeiformes). *PLoS One.* 2013;8(2):e56485.
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.1303.3997.* 2013. doi:10.6084/M9.FIGSHARE.963153.V1
- Li J, Bian C, Yi Y, Yu H, You X, Shi Q. Temporal dynamics of teleost populations during the Pleistocene: a report from publicly available genome data. *BMC Genomics.* 2021;22(1):1–11.
- Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature.* 2011;475(7357):493–496.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–2079.
- Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ. Capturing protein-coding genes across highly divergent species. *Biotechniques.* 2013;54(6):321–326.
- Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27(6):764–770.
- Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc.* 2010;2010(6):pdb prot5448.
- Mohindra V, Dangi T, Tripathi RK, Kumar R, Singh RK, Jena JK, Mohapatra T. Draft genome assembly of *Tenualosa ilisha*, Hilsa shad, provides resource for osmoregulation studies. *Sci Rep.* 2019;9(1):16511.
- Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1999;27(1):29–34.
- Poulsen AF, Hortle KG, Valbo-Jorgensen J, Chan S, Chhohon CK, Viravong S, Bouakhamvongsa K, Suntornratana U, Yoorong N, Nguyen TT, et al. *Distribution and ecology of some important riverine fish species of the Mekong River Basin.* Phnom Penh, Cambodia: Mekong River Commission; 2004. MRC Technical Paper No. 10.
- Riede K. *Global register of migratory species—from global to regional scales.* Federal Agency for Nature Conservation; 2004. p. 329. Final Report of the R&D-Projekt 808 05 081. <https://www.researchgate.net/publication/235225727>
- Roberts TR. Artisanal fisheries and fish ecology below the great waterfalls of the Mekong River in southern Laos. *Nat Hist Bull Siam Soc.* 1993;41:31–62. <https://vdocuments.net/artisanal-fisheries-and-fish-ecology-below-the-great.html?page=1>
- Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics.* 2003;19(2):301–302.
- Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post analysis of large phylogenies. *Bioinformatics.* 2014;30(9):1312–1313.
- Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics.* 2003;19(Suppl 2):ii215–ii225.
- Stone R. Dam-building threatens Mekong fisheries. *Science.* 2016;354(6316):1084–1085.
- Vidthayanon C. *Tenualosa thibaudeaui.* The IUCN Red List of Threatened Species 2011. e.T21627A9303248. Version 3.1.
- Vu NU, Au VH, Nguyen BL. Diversity of fish in Hau River, Mekong Delta, Vietnam. *J Fish Sci Technol.* 2015. Special issue:167–174. <https://slideshare.vn/ngu-nghiep/diversity-of-fish-in-hau-river-mekong-delta-vietnam-waycuq.html>
- Vurtture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* 2017;33(14):2202–2204.
- Wang RN, Green J, Wang Z, Deng Y, Qiao M, Peabody M, Zhang Q, Ye J, Yan Z, Denduluri S, et al. Bone Morphogenetic Protein (BMP) signaling in development and human diseases. *Genes Dis.* 2014;1(1):87–105.
- Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol.* 2018;35(3):543–548.