scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the northern snakehead (*Channa argus*) using PacBio and Hi-C technologies

Donglei Sun[1], Haishen Wen[1], Xin Qi[1], Chao Li[2], Lingyu Wang[1], Jianlong Li[1], Mingxin Zhu[1], Xiaoyan Zhang[2] & Yun Li[1] ✉

The evolutionary origins of specialized organs pose significant challenges for empirical studies, as most such organs evolved millions of years ago. The Northern snakehead (*Channa argus*), an air-breathing fish, possesses a suprabranchial organ, a common feature of the Anabantoidei, offering a unique opportunity to investigate the function and evolutionary origins of specialized organs. In this study, a high-quality chromosome-level reference genome of *C. argus* was constructed using PacBio HiFi sequencing and Hi-C technology. The final genome assembly size is 712.14 Mb, with a scaffold N50 of 28.08 Mb. The assembled sequences were anchored to 24 pseudo-chromosomes and predicted 21,643 protein-coding genes. The genome comprises 27.70% repetitive elements and includes 3,588 (98.6%) complete BUSCOs, demonstrating superior contiguity and functional completeness compared to other published *C. argus* assemblies. This genome provides valuable genetic resources for exploring the evolution of the aquatic-aerial bimodal breathing system, including clarifying the evolutionary histories and adaptive strategies.

Background & Summary

The northern snakehead, *Channa argus*, belonging to the Anabantoidei, is an economically important freshwater species that is extensively cultivated in Asia and Africa[1]. Because of its strong growth capacity, high nutritive value, and significant hypoxia tolerance[2], northern snakehead has become extremely popular in the Chinese aquaculture industry, with annual production exceeding 500,000 tons[3]. In recent years, increasing market demands have promoted the development of genetic improvements in several economically important traits of the northern snakehead, such as body color[3], growth[4], and sex-related traits[5]. Additionally, the northern snakehead possesses a suprabranchial organ (SBO) for aerial respiration[6,7], making it an excellent model for investigating the evolution and functional mechanism of the air-breathing organ (ABO) (Fig. 1).

The transformation from aquatic to aerial gas exchange in vertebrates has long been a hot topic of study for evolutionary biologists. Fish conduct aerial respiration, providing critical evidence for the evolution of life from the ocean to land[8]. Notably, some fish have evolved ABOs to adapt to anoxic environments[9,10]. More than 450 fish species across 50 families have been reported to possess ABOs[11], but these organs vary significantly among different fishes. Several types of fish ABOs have been reported, including SBO, modified swim bladder, skin, stomach, oropharyngeal cavity, and intestine[6,10,12]. These organs share features similar to those of higher vertebrate lungs, such as being well-vascularized[13] and having a short blood-gas diffusion distance[14,15]. Fish with bimodal respiration can survive for a certain time out of water and exhibit stronger hypoxia tolerance compared to water breathers[7,16,17]. All species in the Anabantoidei are aquatic air-breathing fish, including snakeheads (*C. argus* and *C. maculata*), Siamese fighting fish (*Betta splendens*), and climbing perch (*Anabas testudineus*)[18]. The natural habitats of air-breathing fish are found in tropical and temperate waters across various aquatic ecosystems[19]. These species are characterized by high phenotypic and morphological plasticity, allowing them to adapt to changing environmental conditions[20]. Prior research on bimodal respiration in fish has primarily

[1]Key Laboratory of Mariculture (Ocean University of China), Ministry of Education (KLMME), Fisheries College, Ocean University of China, Qingdao, 266003, China. [2]School of Marine Science and Engineering, Qingdao Agricultural University, Qingdao, 266109, China. ✉e-mail: yunli0116@ouc.edu.cn

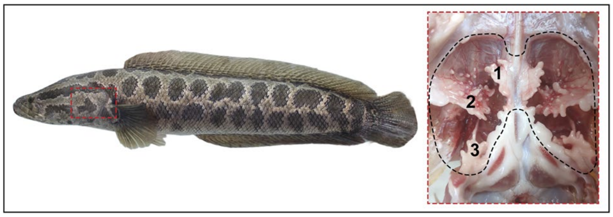

Fig. 1 Morphological photograph of *C. argus* and suprabranchial organ. The suprabranchial organ are made up of areas 1, 2, and 3.

Library type	Clean Reads (M)	Clean data (Gb)	N50 (bp)	Max Length (bp)	Sequencing coverage (×)
WGS	1,091.01	109.10	—	—	153.20
PacBio (HiFi)	3.84	27.72	19,051	50,113	38.92
Hi-C	1,278.41	127.84	—	—	179.52

Table 1. Summary of obtained sequencing data generated from multiple sequencing technologies for *C. argus* genome assembly.

focused on their histomorphological and respiratory adaptations[21–24]. Understanding the evolutionary and molecular mechanisms of air-breathing is fundamentally important for theoretical knowledge and aquaculture applications of air-breathing fish. However, few studies have been conducted on this topic.

A high-quality reference genome resource is increasingly important for facilitating genomic breeding programs, investigating biological phenomena, and conserving germplasm[25,26]. Investigating the genomic evolution of *C. argus* may elucidate the underlying molecular mechanisms involved in air-breathing in air-breathing fishes. Although three versions of the *C. argus* genome have been published[1,27,28], the contiguity and completeness of these genome assemblies still require significant improvement. In the present study, PacBio HiFi long-read sequencing and Hi-C technology were integrated to generate a high-quality chromosome-level reference genome for *C. argus*. The assembled genome was approximately 712.14 Mb with a contig N50 of 11.61 Mb and a scaffold N50 of 28.08 Mb. A total of 652.14 Mb of the assembled sequences were anchored to 24 pseudo-chromosomes. We predicted 21,643 protein-coding genes, of which 19,847 (91.70%) were functionally annotated. BUSCO alignment revealed that our final assembly contained 3,588 (98.5%) complete BUSCOs, demonstrating superior contiguity and functional completeness compared to other published *C. argus* assemblies. The successful assembly of a high-quality genome provides valuable genetic resources for elucidating the evolution of the aquatic-aerial bimodal breathing system, including clarifying the evolutionary histories and adaptation strategies.

Methods

Sample collection and genomic DNA sequencing. A two-year-old healthy female *C. argus* was collected from Weishan Lake in Jining, Shandong Province, China. We collected the muscle and blood of this individual for genome and Hi-C sequencing. For whole genome sequencing, high-quality genomic DNA was extracted from the muscle using the QIAamp DNA purification kit (Qiagen). A 100 bp paired-end short reads (PE100) sequencing library was constructed and sequenced by the BGISEQ-500 platform. For the long-read sequencing, a high-fidelity (HiFi) SMRTbell library was prepared using the SMRTbell Express Template Prep Kit 2.0, and then the Circular Consensus Sequencing (CCS) mode was performed on the PacBio Sequel II platform for sequencing. The CCS raw data was processed by CCS version 4.2.0 algorithm (https://github.com/PacificBiosciences/ccs) to obtain the HiFi reads used for genome assembly. Finally, a total of 109.10 Gb short reads and 27.72 Gb PacBio HiFi long reads (N50 length of 19.05 kb) were produced for constructing a high-quality reference genome of *C. argus* (Table 1).

Hi-C library preparation and sequencing. Hi-C technology was applied to construct the chromosome-level genome of *C. argus*. Genomic DNA was extracted from blood samples that had been fixed with formaldehyde at a concentration of 1% and digested by the restriction enzyme Mbo I, followed by repairing 5′ overhangs using biotinylated nucleotides. A 100 bp paired-end Hi-C library was constructed following the Hi-C library preparation protocol (https://www.protocols.io) and then sequenced on the BGISEQ-500 platform. Thereafter, quality control of Hi-C raw reads was performed using HiC-Pro (v 2.8.0)[29]. Finally, the Hi-C library generated a total of 127.84 Gb (179.52 × coverage) of clean data after filtering criteria with short reads (Table 1).

***De novo* genome assembly and chromosome construction.** The genome size of *C. argus* was estimated based on the *k*-mer frequency distribution analysis using the clean data of short-read sequencing. The *k*-mer count frequencies were computed by GenomeScope (v 2.0)[30] with a *k*-mer size of 27. The estimated genome size was 681.47 Mb, and the heterozygosity rate was 0.20% based on the *k*-mer frequency analysis (Table 2, Fig. 2a).

Content	Min	Max
k-mer	27 bp	
Heterozygosity	0.20%	0.20%
Genome Haploid Length	681,316,446 bp	681,473,534 bp
Genome Repeat Length	122,730,973 bp	122,759,271 bp
Genome Unique Length	558,585,473 bp	558,714,263 bp
Model Fit	92.68%	93.37%
Read Error Rate	0.26%	0.26%

Table 2. Statistical results of *k*-mer analysis.

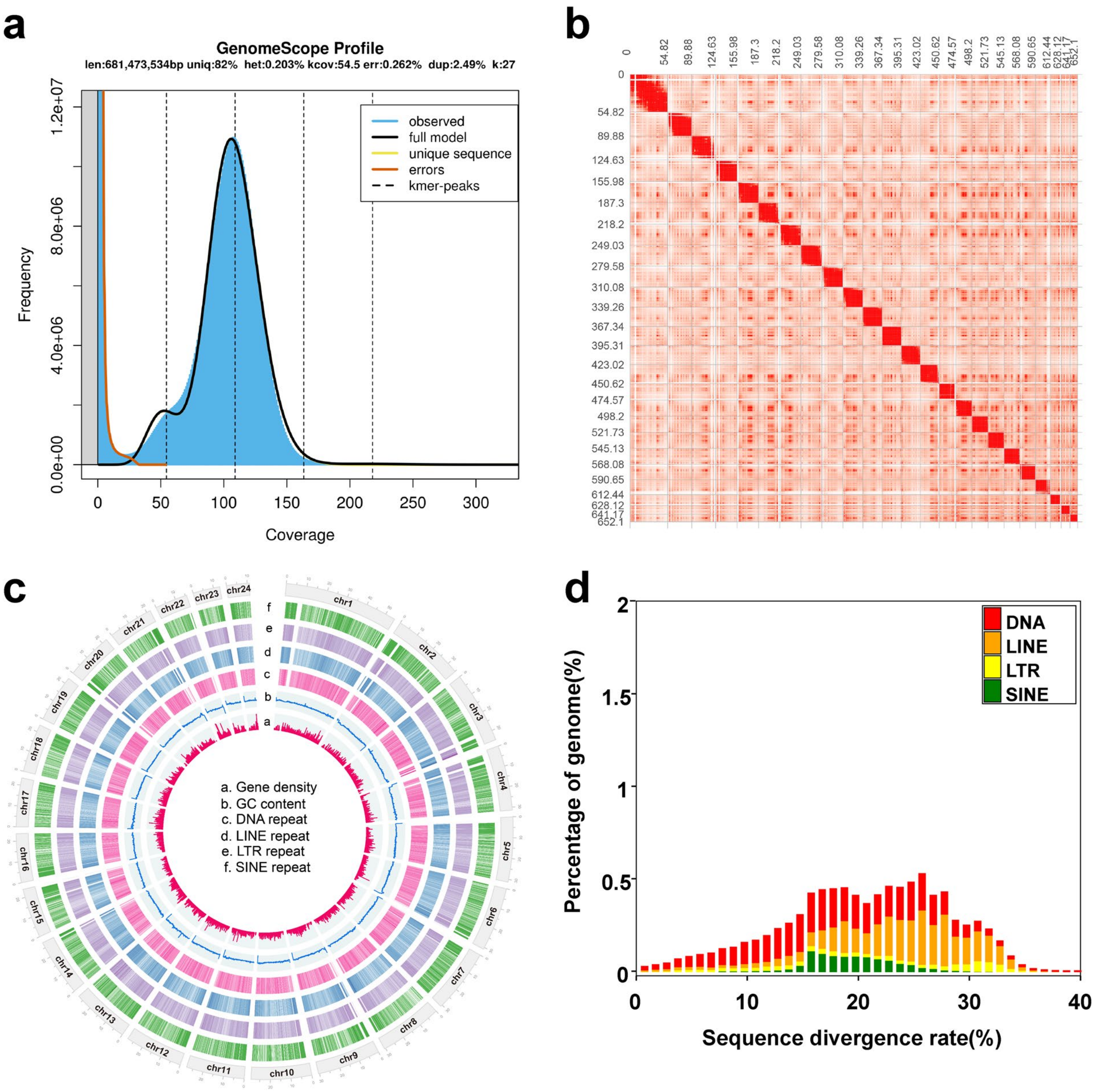

Fig. 2 Chromosome-level genome assembly and annotation of the northern snakehead. (**a**) *k*-mer frequency distribution in the *C. argus* genome. The *k*-mer distributions showed that the genome size was calculated to be 681.47 Mb with a heterozygous rate of 0.20%. **(b)** Hi-C interaction heatmap for the northern snakehead genome. The map shows scaffolded and independently assembled chromosomes at high resolution. **(c)** Characterization of the assembled genome of *C. argus*. The tracks indicate a) gene density, b) GC density, c) DNA repeat, d) LINE repeat, e) LTR repeat, and f) SINE repeat. The densities of genes, GC, and TEs were calculated in 100 kb windows. **(d)** Distribution of divergence rate for TEs in the *C. argus* genome.

Features	Statistics
Assembled genome size	712.14 Mb
Coverage	158.06 ×
Number of scaffolds	414
N50 contigs	11.61 Mb
N50 scaffolds	28.08 Mb
Largest scaffold	54.82 Mb
Number of predicted protein-coding genes	21,643

Table 3. Assembly statistics of the *C. argus* genome.

Chromosome ID	Length (bp)	Percentage
chr1	54,820,115	7.70%
chr2	35,064,694	4.92%
chr3	34,749,973	4.88%
chr4	31,357,850	4.40%
chr5	31,319,733	4.40%
chr6	30,899,339	4.34%
chr7	30,834,011	4.33%
chr8	30,554,426	4.29%
chr9	30,496,680	4.28%
chr10	29,189,983	4.10%
chr11	28,076,099	3.94%
chr12	27,967,726	3.93%
chr13	27,708,905	3.89%
chr14	27,611,317	3.88%
chr15	23,951,042	3.36%
chr16	23,629,612	3.32%
chr17	23,531,644	3.30%
chr18	23,398,285	3.29%
chr19	22,946,491	3.22%
chr20	22,577,630	3.17%
chr21	21,790,777	3.06%
chr22	15,680,959	2.20%
chr23	13,051,962	1.83%
chr24	10,926,970	1.53%
Unanchored	60,004,617	8.43%
Total	712,140,840	

Table 4. Result of *C. argus* genomic assembly at chromosome-level.

The genome of *C. argus* was initially assembled by HiFiasm (v 0.13) using the HiFi reads from the long-read sequencing[31]. Duplicate contigs and redundant sequences in the primary assembly were removed using the Purge_dups program[32]. After *de novo* assembly and polishing, a 712.14 Mb reference genome of *C. argus* with a scaffold N50 length of 11.61 Mb was generated (Table 3). To further construct the chromosome-level genome, Hi-C clean reads were aligned with the primarily assembled genome by BWA (v 0.7.10)[33] to construct inter-/intrachromosomal contact maps. Open-source tools Juicer (v 1.5)[34] and 3D-DNA pipeline[35] were applied to anchor the initially assembled genome scaffolds to chromosomes. Finally, 652.14 Mb of the genome sequence (~91.57% of the assembly) were anchored and oriented into 24 large scaffolds matching the chromosome number of *C. argus* with a scaffold N50 length of 28.08 Mb (Fig. 2b-c, Tables 3,4).

Genomic repeat annotation. We annotated the repetitive sequences in the *C. argus* genome with a combination of *de novo* prediction and homology-based approaches. For *de novo* prediction, the LTR_FINDER (v 1.05)[36] and RepeatScout[37] tools were used to construct *C. argus* repeat library based on the characteristics of the repeat sequences. For the homology-based method, RepeatProteinMask (v 3.3.0) and RepeatMasker (v 4.0.5)[38] were used for predicting based on homologous sequences in RepBase database (http://www.girinst.org/repbase)[39].

The repetitive sequences were annotated in the assembled genome of *C. argus* using combined homology-based and *de novo* predictions of repeats. In total, 197.24 Mb consisted of repetitive sequences, accounting for 27.70% of the genome assembly. Transposable elements (TEs) accounted for the largest number

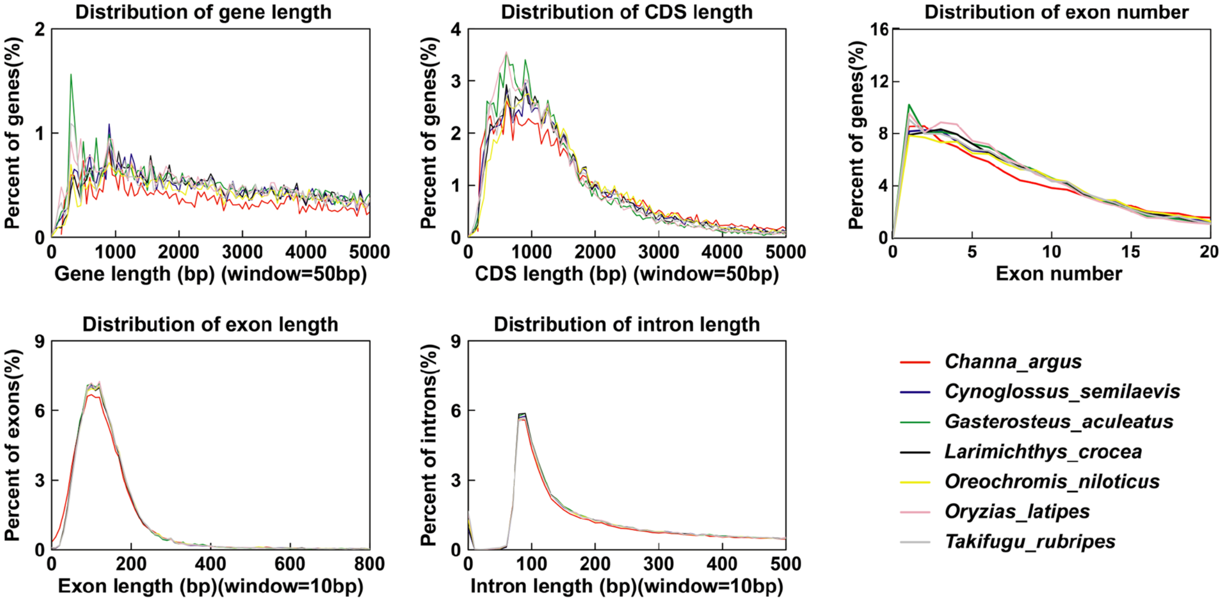

Fig. 3 Comparisons of the genomic elements in *C. argus* and related species.

Database	Numbers	Percent
Total	21,643	100%
Swissprot	18,541	85.67%
KEGG	17,985	83.10%
TrEMBL	19,658	90.83%
Interpro	18,242	84.29%
GO	14,856	68.64%
NR	19,755	91.28%
Overall	19,847	91.70%

Table 5. Statistics for gene function annotation in *C. argus* genome.

of repeats, which occupied 19.49% of the genome assembly (Supplementary Table 1). These TEs included DNA repeat elements (6.18%), long interspersed nuclear elements (LINEs, 10.37%), short interspersed nuclear elements (SINEs, 3.23%), long terminal repeats (LTRs, 4.28%), and others (0.48%) (Fig. 2d, Supplementary Table 1). Their distributions across each chromosome were illustrated in Fig. 2c.

Protein-coding gene prediction and function annotation. The gene structure annotation was conducted by combining *de novo* prediction, homologous prediction, and transcriptome-based prediction. For *de novo* prediction, the gene structures were predicted using AUGUSTUS (v 3.2.2)[40], GlimmerHMM (v 3.0.4)[41] and GENESCAN (v 1.0)[42] with default settings, respectively. Amino acid sequences of six teleost species, including tongue sole (*Cynoglossus semilaevis*), three-spine stickleback (*Gasterosteus aculeatus*), large yellow croaker (*Larimichthys crocea*), Nile tilapia (*Oreochromis niloticus*), Japanese medaka (*Oryzias latipes*), and pufferfish (*Takifugu rubripes*) were downloaded from the National Center for Biotechnology Information (NCBI) database and used for homology-based annotation using Genewise (v 2.4.0)[43]. For the transcriptome-based prediction, RNA-seq data from 3 tissues (skin, gill, and eye) were assembled by TRINITY (v 2.1.1)[44]. Then PASA (v 2.0.1)[45] was used to align the transcripts to the genome, and Transdecoder (http://transdecoder.github.io) was performed to identify Open Reading Frames (ORFs). Finally, the results of gene structure prediction based on the above methods were integrated by EVidenceModeler (v 1.1.1)[46].

For gene functional annotation, the sequences of predicted protein-coding genes were aligned to the public protein databases, including TrEMBL[47], SwissProt[47], InterPro[48], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[49] and NR (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins) using BLASTP with the threshold of E-value of 1e-5. InterProScan (v 4.7)[50] was employed to obtain protein domain annotation and Gene Ontology (GO) annotation[51].

A total of 21,643 protein-coding genes were predicted in the *C. argus* genome assembly through integrated transcriptome sequences, *de novo* prediction, and homology-based strategies (Supplementary Table 2). The predicted gene length and CDS length averaged 17,549 bp and 1,966 bp, respectively. The average number of exons per gene was 11.26, and the average length of exons and introns were 175 bp and 1,519 bp, respectively (Supplementary Table 2). The gene structures of *C. argus* were relatively conserved compared to other teleosts

Statistics	Result
Reads mapping rate (%)	99.70
Genome average sequencing depth (×)	158.06
Coverage of genome (%)	99.91
Coverage of genome > 4× (%)	99.75
Coverage of genome > 10× (%)	99.52
Coverage of genome > 20× (%)	99.12

Table 6. Coverage assessment of the *C. argus* genome using WGS data.

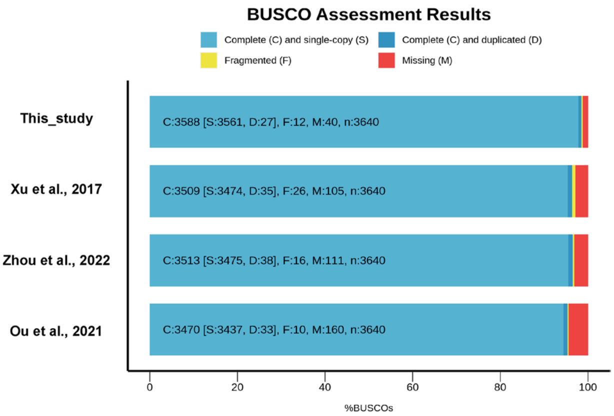

Fig. 4 Quality assessment of assembled genome of *C. argus*. Comparison of BUSCO scores (Actinopterygii_odb10) for the assemblies of *C. argus* with three other assemblies.

Type	Gene Number	Percent (%)
Complete BUSCOs (C)	3,588	98.6
Complete and single-copy BUSCOs (S)	3,561	97.8
Complete and duplicated BUSCOs (D)	27	0.7
Fragmented BUSCOs (F)	12	0.3
Missing BUSCOs (M)	40	1.1
Total BUSCO groups searched	3,640	100

Table 7. Assessment of *C. argus* genome completeness by BUSCO.

with available annotation data (Fig. 3). For gene function prediction, 19,847 genes were successfully annotated, representing 91.70% of all predicted protein-coding genes of *C. argus* (Table 5). Consequently, the high integration efficiency, mapping ratio, recognition rate of single-copy orthologues, and gene number indicate that the genome assembly of *C. argus* is of high quality.

Data Records

The assembled genome has been deposited in the NCBI Assembly database with GenBank accession JAJQTP000000000[52]. All the raw sequencing data utilized in this study, including WGS, HiFi, and Hi-C have been deposited in the NCBI SRA database under accession numbers SRP375296[53]. The genome annotation files were also available in figshare[54].

Technical Validation

Evaluating the completeness of the genome assembly and annotation. To evaluate the completeness and consistence of the genome assembly, clean short reads were mapped onto the assembled genome using BWA software with default parameters[55]. The mapping ratios of the assembly was 99.70%, with a genome coverage of 99.91% (Table 6). Subsequently, Benchmarking Universal Single-Copy Orthologs (BUSCO) (v 3.0)[56] analysis was conducted using the Actinopterygii_odb10 database to assess the completeness and quality of the genome assembly. Of the 3,640 single-copy orthologues, 98.6% (3,588) were fully identified in the genome assembly (Fig. 4, Table 7). This study significantly improves the assembly contiguity and functional completeness compared to previously published *C. argus* assemblies[1,27,28], leading to its selection as the reference genome for *C. argus* in the NCBI database.

Code availability

All software and pipelines used for data analyses were executed according to the manual and protocols of the published bioinformatic tools. The version and code/parameters of software have been described in Methods.

Received: 16 August 2024; Accepted: 17 December 2024;
Published online: 27 December 2024

References

1. Xu, J. *et al*. Draft genome of the Northern snakehead, *Channa argus*. *Gigascience* **6**, gix011, https://doi.org/10.1093/gigascience/gix011 (2017).
2. Liu, J., Cui, Y. & Liu, J. Resting metabolism and heat increment of feeding in mandarin fish (*Siniperca chuatsi*) and Chinese snakehead (*Channa argus*). *Comp. Biochem. Phys. A* **127**, 131–138, https://doi.org/10.1016/s1095-6433(00)00246-4 (2000).
3. Sun, D. *et al*. The genetic basis and potential molecular mechanism of yellow-albino northern snakehead (*Channa argus*). *Open Biol.* **13**, 220235, https://doi.org/10.1098/rsob.220235 (2023).
4. Liu, H. *et al*. High-density genetic linkage map and QTL fine mapping of growth and sex in snakehead (*Channa argus*). *Aquaculture* **519**, 734760, https://doi.org/10.1016/j.aquaculture.2019.734760 (2020).
5. Sun, D. *et al*. Comparative study of candidate sex determination regions in snakeheads (*Channa argus* and *C. maculata*) and development of novel sex markers. *Aquaculture* **575**, 739771, https://doi.org/10.1016/j.aquaculture.2023.739771 (2023).
6. Jiang, Y. *et al*. Comparative transcriptome analysis between aquatic and aerial breathing organs of *Channa argus* to reveal the genetic basis underlying bimodal respiration. *Mar. Genom.* **29**, 89–96, https://doi.org/10.1016/j.margen.2016.06.002 (2016).
7. Lefevre, S. *et al*. Air-breathing fishes in aquaculture. What can we learn from physiology? *J. Fish Biol.* **84**, 705–731, https://doi.org/10.1111/jfb.12302 (2014).
8. Graham, J. B. & Lee, H. J. Breathing air in air: in what ways might extant amphibious fish biology relate to prevailing concepts about early tetrapods, the evolution of vertebrate air breathing, and the vertebrate land transition? *Physiol. Biochem. Zool.* **77**, 720–731, https://doi.org/10.1086/425184 (2004).
9. Li, N. *et al*. Genome sequence of walking catfish (*Clarias batrachus*) provides insights into terrestrial adaptation. *BMC Genom.* **19**, 1–16, https://doi.org/10.1186/s12864-018-5355-9 (2018).
10. Huang, S., Cao, X. & Tian, X. Transcriptomic analysis of compromise between air-breathing and nutrient uptake of posterior intestine in loach (*Misgurnus anguillicaudatus*), an air-breathing fish. *Mar. Biotechnol.* **18**, 521–533, https://doi.org/10.1007/s10126-016-9713-9 (2016).
11. Martin, K. L. Theme and variations: amphibious air-breathing intertidal fishes. *J Fish Biol* **84**, 577–602, https://doi.org/10.1111/jfb.12270 (2014).
12. Wang, K. *et al*. African lungfish genome sheds light on the vertebrate water-to-land transition. *Cell* **184**, 1362–1376, https://doi.org/10.1016/j.cell.2021.01.047 (2021).
13. Munshi, J. S. D., Olson, K. R., Ojha, J. & Ghosh, T. K. Morphology and vascular anatomy of the accessory respiratory organs of the air-breathing climbing perch, *Anabas testudineus* (Bloch). *Am. J. Anat.* **176**, 321–331, https://doi.org/10.1002/aja.1001760306 (1986).
14. Frick, N. T., Scott Bystriansky, J. & Stuart Ballantyne, J. The metabolic organization of a primitive air-breathing fish, the Florida gar (*lepisosteus platyrhincus*). *J. Exp. Zool. Part A* **307**, 7–17, https://doi.org/10.1002/jez.a.338 (2007).
15. Icardo, J. M. Lungs and gas bladders: morphological insights. *Acta Histochem.* **120**, 605–612, https://doi.org/10.1016/j.acthis.2018.08.006 (2018).
16. Sayer, M. D. Adaptations of amphibious fish for surviving life out of water. *Fish Fish.* **6**, 186–211, https://doi.org/10.1111/jfb.12270 (2005).
17. Ip, Y. K. & Chew, S. F. Air-breathing and excretory nitrogen metabolism in fishes. *Acta Histochem.* **120**, 680–690, https://doi.org/10.1016/j.acthis.2018.08.013 (2018).
18. Rüber, L., Britz, R. & Zardoya, R. Molecular phylogenetics and evolutionary diversification of labyrinth fishes (Perciformes: Anabantoidei). *Syst. Biol.* **55**, 374–397, https://doi.org/10.1080/10635150500541664 (2006).
19. Berra, T. M. Freshwater fish distribution. Academic press (2001).
20. Huang, C. Y., Lin, C. P. & Lin, H. C. Morphological and biochemical variations in the gills of 12 aquatic air-breathing anabantoid fish. *Physiol. Biochem. Zool.* **84**, 125–134, https://doi.org/10.1086/658996 (2011).
21. Adamek-Urbańska, D., Błażewicz, E., Sobień, M., Kasprzak, R. & Kamaszewski, M. Histological study of suprabranchial chamber membranes in Anabantoidei and Clariidae fishes. *Animals* **11**, 1158, https://doi.org/10.3390/ani11041158 (2021).
22. Ishimatsu, A. Evolution of the cardiorespiratory system in air-breathing fishes. *Aqua-BioScience Monographs* **5**, 1–28 (2012).
23. Milsom, W. K. New insights into gill chemoreception: receptor distribution and roles in water and air breathing fish. *Resp. Physiol. Neurobi.* **184**, 326–339, https://doi.org/10.1016/j.resp.2012.07.013 (2012).
24. Damsgaard, C. *et al*. Evolutionary and cardio-respiratory physiology of air-breathing and amphibious fishes. *Acta Physiol.* **228**, e13406, https://doi.org/10.1111/apha.13406 (2020).
25. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nat. Rev. Genet.* **11**, 697–709, https://doi.org/10.1038/nrg2844 (2010).
26. Mohanty, B. P. *et al*. Omics technology in fisheries and aquaculture. *Advances in Fish Research* **7**, 1–30 (2019).
27. Ou, M. *et al*. Chromosome-level genome assemblies of *Channa argus* and *Channa maculata* and comparative analysis of their temperature adaptability. *Gigascience* **10**, giab070, https://doi.org/10.1093/gigascience/giab070 (2021).
28. Zhou, C. *et al*. Chromosome-Scale Assembly and Characterization of the Albino Northern Snakehead, *Channa argus* var. (Teleostei: Channidae) Genome. *Front. Mar. Sci.* **9**, 839225, https://doi.org/10.3389/fmars.2022.839225 (2022).
29. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 1–11, https://doi.org/10.1186/s13059-015-0831-x (2015).
30. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
31. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
32. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
33. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595, https://doi.org/10.1093/bioinformatics/btp698 (2010).
34. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
35. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).

36. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
37. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
38. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4–10, https://doi.org/10.1002/0471250953.bi0410s05 (2009).
39. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* **6**, 1–6, https://doi.org/10.1186/s13100-015-0041-9 (2015).
40. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439, https://doi.org/10.1093/nar/gkl200 (2006).
41. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
42. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* **268**, 78–94, https://doi.org/10.1006/jmbi.1997.0951 (1997).
43. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
44. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
45. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
46. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
47. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48, https://doi.org/10.1093/nar/28.1.45 (2000).
48. Mitchell, A. L. *et al*. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360, https://doi.org/10.1093/nar/gky1100 (2019).
49. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361, https://doi.org/10.1093/nar/gkw1092 (2017).
50. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
51. Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338, https://doi.org/10.1093/nar/gky1055 (2019).
52. Li, Y., Wen, H. & Sun, D. *Channa argus* breed panmixia, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JAJQTP000000000 (2024).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP375296 (2022).
54. Sun, D. Genome annotation of function annotation result of northern snakehead (*Channa argus*). *figshare* https://doi.org/10.6084/m9.figshare.26582638.v1 (2024).
55. Li, H. & Durbin, R. Fast and accurate short read align ment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1186/s44342-024-00012-5 (2009).
56. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 227-245 (2019).

Acknowledgements

This work was supported by Shandong Technical System of Fish Industry (SDAIT-12-03). We acknowledge the support of the High-Performance Biological Supercomputing Center at the Ocean University of China for this research.

Author contributions

Y.L. and H.W. conceived and supervised the study. Y.L. and H.W. coordinated and supervised the whole study. D.S. collected the sample and conducted the genome assembly and analysis. D.S. drafted the manuscript. Y.L. revised the manuscript. X.Q., C.L., L.W., J.L., M.Z. and X.Z. participated in discussions and provided suggestions for manuscript improvement. All authors have read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Ethics statement

This work was approved by the Animal Research and Ethics Committees of Ocean University of China (Permit Number: 20141201). All the methods used in this study were carried out following approved guidelines.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-04314-9.

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.