




OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Acrossocheilus fasciatus* using PacBio sequencing and Hi-C technology

Jianbo Zheng¹, Jianhu Jiang¹, Qianlong Rui^{1,2}, Fei Li¹✉, Shili Liu¹ , Shun Cheng¹, Meili Chi¹ & Wenping Jiang¹

Acrossocheilus fasciatus (Cypriniformes, Cyprinidae) is emerged as a newly commercial stream fish in the south of China with high economic and ornamental value. In this study, a chromosome-level reference genome of *A. fasciatus* was assembled using PacBio, Illumina and Hi-C sequencing technologies. As a result, a high-quality genome was generated with a size of 879.52 Mb (accession number: JAVLVS000000000), scaffold N50 of 32.7 Mb, and contig N50 of 32.7 Mb. The largest and smallest scaffold was 60.57 Mb and 16 kb, respectively. BUSCO analysis showed a completeness score of 98.3%. Meanwhile, the assembled sequences were anchored to 25 pseudo-chromosomes with an integration efficiency of 96.95%. Additionally, we found approximately 390.91 Mb of repetitive sequences that accounting for 44.45% of the assembled genome, and predicted 24,900 protein-coding genes. The available genome reported in the present study provided a crucial resource to further investigate the regulation mechanism of genetic diversity, sexual dimorphism and evolutionary histories.

Background & Summary

The genus *Acrossocheilus* belongs to Barbinae, Cyprinidae, and is composed of approximately 26 species, which are mainly native in Laos, Vietnam, and China¹. Meanwhile, these groups exhibit diversiform morphological characteristics and ecological habits, providing a great model for investigating species origin and geographical distribution of freshwater fish². In addition, its flesh is tender, delicious and contains highly polyunsaturated fatty acids (PUFA), possessing a considerable market value. Recently, the freshwater grouper *A. fasciatus* has become a commercially emerging aquaculture fish due to its nutritive and ornamental value³. Moreover, as an omnivorous fish, the growth of *A. fasciatus* requires to feed with moss and other algae plants, which can inhibit the rankness of these aquatic plants, thus playing a role in ecological balance. Previous studies of *A. fasciatus* have primarily focused on its embryos and larval development, gonad histological characteristics, phylogenetic relationships, population structure, and artificial breeding^{4–6}. On the other hand, *A. fasciatus* represents significant difference in growth rate and body size between males and females, with females growing faster than males (Fig. 1a), indicating all-female breeding is of high commercial value in aquaculture⁷. However, our knowledge of *A. fasciatus* on genetic and evolutionary mechanisms have been limited due to lack of genetic resources and genomic information. In this study, we employed an integrated strategy of PacBio, Illumina and Hi-C sequencing technologies to assemble a high-quality genome in a size of 879.52 Mb with scaffold N50 of 32.7 Mb (Fig. 1b,c). We believe that this high-quality of chromosomal-level genome data will provide valuable resources for breeding programs and evolutionary investigation.

Methods

Sample collection and nucleic acid extraction. Mature and healthy *A. fasciatus* were obtained from Zhejiang institute of freshwater fisheries in Huzhou, Zhejiang province, China. Muscle tissues from adult female *A. fasciatus* was prepared for DNA extraction with SDS lysis method, while ovary, kidney, brain, testis, skin, and gill were collected for total RNA extraction using a TRIzol kit following the manufacturer's protocol. Herein,

¹Key Laboratory of Genetics and Breeding, Zhejiang Institute of Freshwater Fisheries, Huzhou, China. ²College of Biological and Environmental Sciences, Zhejiang Wanli University, Ningbo, China. ✉e-mail: lifeibest1022@163.com

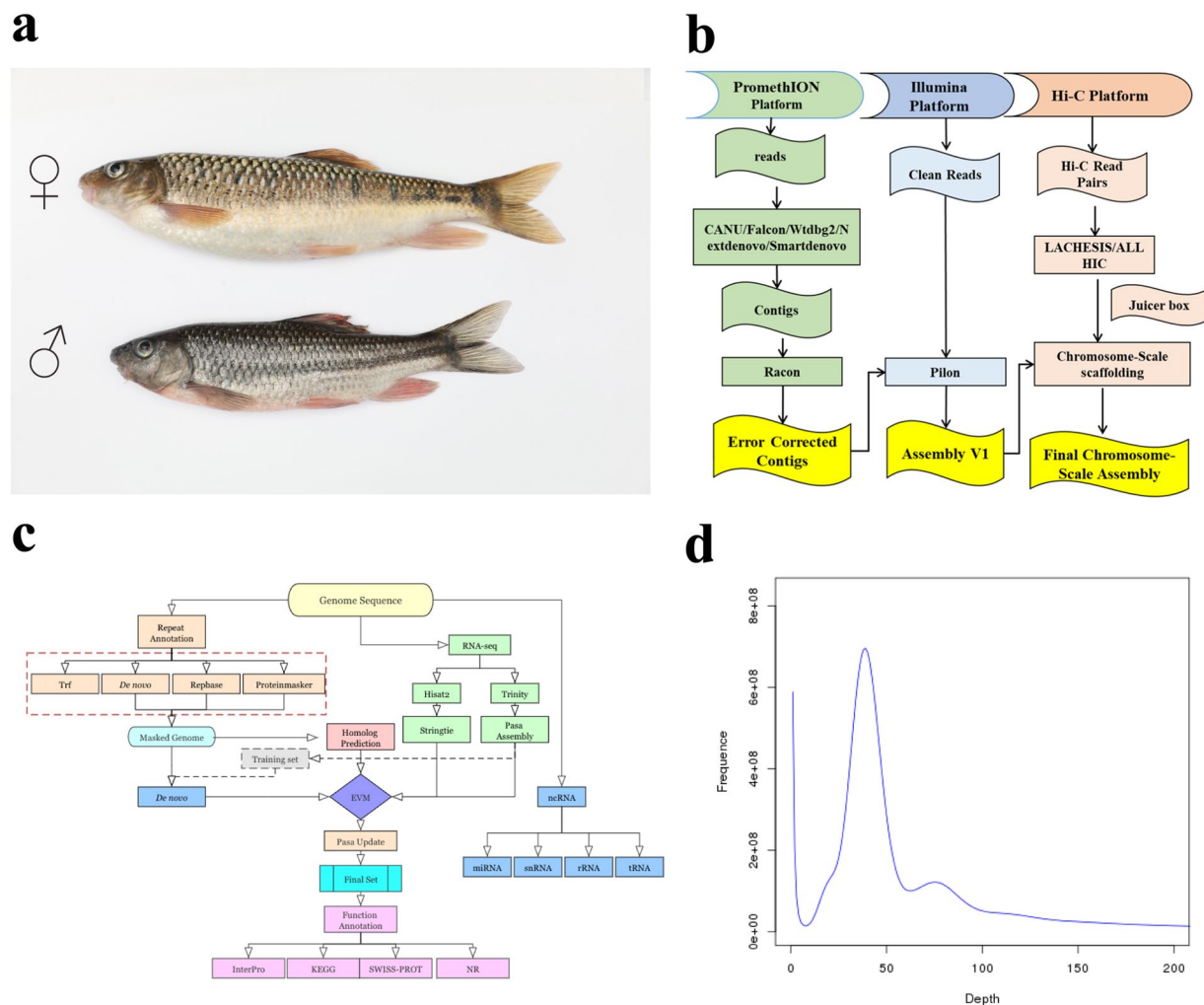


Fig. 1 Workflow of the genome assembly and survey analysis in *A. fasciatus*. **(a)** A picture of female and male *A. fasciatus*. ♂ indicates male individual, and ♀ indicates female individual. **(b)** The work flow used for genome sequencing. **(c)** Flow chart of the genome annotation. **(d)** The 17-mer distribution for the genome size estimation.

Library types	Insert size (bp)	Raw data (Gb)	Clean data (Gb)	Read length (bp)	Sequence coverage (X)
Illumina reads	350	41.00	36.72	150	47.56
PacBio reads	20,000	32.66	22.83	14,447	37.12
Hi-C reads	350	86.32	76.58	150	—
RNA reads	350	41.18	37.51	150	—
Total	—	201.16	173.64	—	—

Table 1. Statistics of the sequencing data for the *A. fasciatus* genome assembly.

the high-quality gDNA was used for genome sequencing, and total RNA isolated from all tissues were used for transcriptome sequencing.

Library construction and genome sequencing. For the Illumina platform (NEB, USA), a paired-end library with an insertion size of 350 bp was generated using NEB Next[®] Ultra[™] DNA Library Prep Kit following manufacturer's recommendations. As a result, a total of 41 Gb Illumina short-reads (coverage of 47.56X, Table 1) with paired-end 150 bp were generated. Simultaneously, HiFi SMRTbell Libraries was prepared using SMRTbell Express Template Prep Kit 2.0 for long-read sequencing with insert size of 20 kb on Pacbio platform. In briefly, gDNA was sheared to 6–20 kb fragments using the g-TUBE, and the ssDNA overhangs were removed with *Exo VII*. Then DNA damage was repaired for Blunt-End ligation, and large insert SMRTbell libraries were constructed after size selection to prepare for sequencing use DNA Sequencing Reagent Kit. For the PacBio platform,

	PacBio		Hi-C	
	Scaffold	Contig	Scaffold	Contig
Total number	134	134	194	115
Total length (bp)	879,520,627	879,520,627	879,520,627	879,528,527
Average_length (bp)	6,563,586	6,563,586	7,648,074	4,533,611
Max length (bp)	60,574,424	60,574,424	36,374,165	54,140,365
Min length (bp)	16,004	16,004	18,174	16,004
N50 length (bp)	32,702,747	32,702,747	22,576,242	33,132,389
N50 number	11	11	16	12
N90 length (bp)	11,417,557	11,417,557	5,356,806	26,858,976
N90 number	28	28	45	23

Table 2. Summary of the assembled genome for *A. fasciatus* genome.

approximately 32 Gb PacBio reads (37.12X coverage, Table 1) were obtained with the longest read of 47.52 kb and the N50 length of 14.56 kb.

Genome size estimation and assembly. Herein, clean data generated from Illumina sequencing were subjected to k-mer analysis to estimate the genome size, heterozygosity, and the proportion of repetitive sequences in *A. fasciatus*. Based on 17-mer frequency distribution using Jellyfish v2.3.0⁸ and GenomeScope v2.0⁹, the genome size was estimated to be 862.9 Mb, with a heterozygosity ratio of 0.56% and repeat sequence ratio of 47.09% (Fig. 1d). The 32.66 Gb raw subreads from the PacBio Sequel platform were filtered out, and the remaining clean subreads were error-corrected by Canu (v1.5)¹⁰ and pre-assembled into contigs using FALCON software¹¹. The assembled scaffolds were polished by Pilon (v1.22)¹² with default parameters. The finally assembled genome was 879.52 Mb in size with 134 contigs and a contig N50 of 32.70 Mb (Table 2).

Hi-C library preparation and sequencing. The Hi-C libraries were constructed following the standard protocol described previously with certain modifications. Firstly, female muscle samples were cross-linked by 4% formaldehyde, and the fixed tissues were homogenised and centrifuged to collect the nuclei, then digested with *Mbo* I enzyme overnight at 37 °C. The proximal chromatin DNA was re-ligated using T4 ligase, and Biotin-labeled Hi-C samples were specifically enriched using magnetic beads. After adding A-tails to the fragment ends, Hi-C sequencing libraries were amplified by PCR and sequenced on Illumina HiSeq-2500 platform (PE 150 bp). For chromosome-level assembly, the raw Hi-C sequencing data were primarily filtered using Hi-C-Pro v2.8.0¹³, and the high-quality clean reads were aligned to the polished *A. fasciatus* genome using BWA (v0.7.10)¹⁴ with default parameters (samtools sort sample.sam-output-fmt BAM-o sample.sort.bam). Finally, 96.95% of the initial assembled sequences were anchored to 25 pseudo-chromosomes that ranged in size from 24.09 to 54.14 Mb (Fig. 2a, Table S1), and the total length of the genome assembly was 879.52 Mb with a contig N50 of 22.57 Mb, and scaffold N50 of 33.13 Mb (Table 2).

Repetitive sequence annotation. Repeat elements in the *A. fasciatus* genome were annotated employing a combined methods of homology alignment and *de novo* searches. The homology-based blast was performed against the RepBase data base (<http://www.girinst.org/replib/>)¹⁵ using Repeatmasker and repeat-proteinmask software for known repeat elements. For *de novo* annotation, we firstly employed LTR_FINDER¹⁶, RepeatModeler¹⁷ and RepeatScout¹⁸ to build a *de novo* repeat library, and then was used to predict repeat elements using Repeatmasker with default parameters. Additionally, Tandem Repeats can be identified using Tandem Repeat Finder (TRF, <http://tandem.bu.edu/trf/trf.html>)¹⁹. In this study, we identified 390.91 Mb of repetitive sequences, accounting for 44.45% of the assembled genome (Table 3).

Gene prediction and functional annotation. Protein-coding genes were annotated through integrating three different strategies of homology, *de novo*, and transcriptome-based prediction methods. For homology-based gene prediction, the published protein sequences of *Sinocyclocheilus grahami*, *Puntius tetrazona* and *Carassius auratus* were aligned to the *A. fasciatus* genome assembly using BLAST²⁰ and Genewise²¹ with default parameters. Five *de novo* programs, including Augustus²², GlimmerHMM²³, SNAP²⁴, GeneID²⁵ and GENSCAN²⁶, were used to predict coding regions in the repeat-masked assembly with default parameters. For the transcriptome-based annotation, the RNA-seq data were *de novo* assembled by Trinity (v2.1.1)²⁷ and splicing variations were identified by PASpipeline (v2.4.1)²⁸. Finally, a non-redundant reference gene set was established by merging the above three methods, resulting in a total of 24,900 protein-coding genes (Fig. 2b, Table 4). Simultaneously, we compared the gene parameters of different elements in *A. fasciatus* and three relative species (*S. grahami*, *C. auratus*, *P. tetrazona*), and the result showed a similar distribution of coding DNA sequence (CDS) length, exon length and number, intron length and mRNA length among the sequenced fish genomes (Fig. 2c).

Furthermore, all predicted genes were functionally annotated using public biological function databases of SwissPro²⁹, Nr (<http://www.ncbi.nlm.nih.gov/protein/>), KEGG³⁰ and InterPro³¹ and Pfam (<http://pfam.xfam.org/>). Overall, a total of 24,000 genes (96.40%) were successfully annotated with an average transcript length of 15,927.24 bp and an average CDS length of 1,627.71 bp (Table 5). In addition, non-coding RNAs (ncRNAs) were also annotated, and tRNAscan-SE (v2.0)³² was used to predict tRNAs, and Infernal (1.1)³³ was used to identify

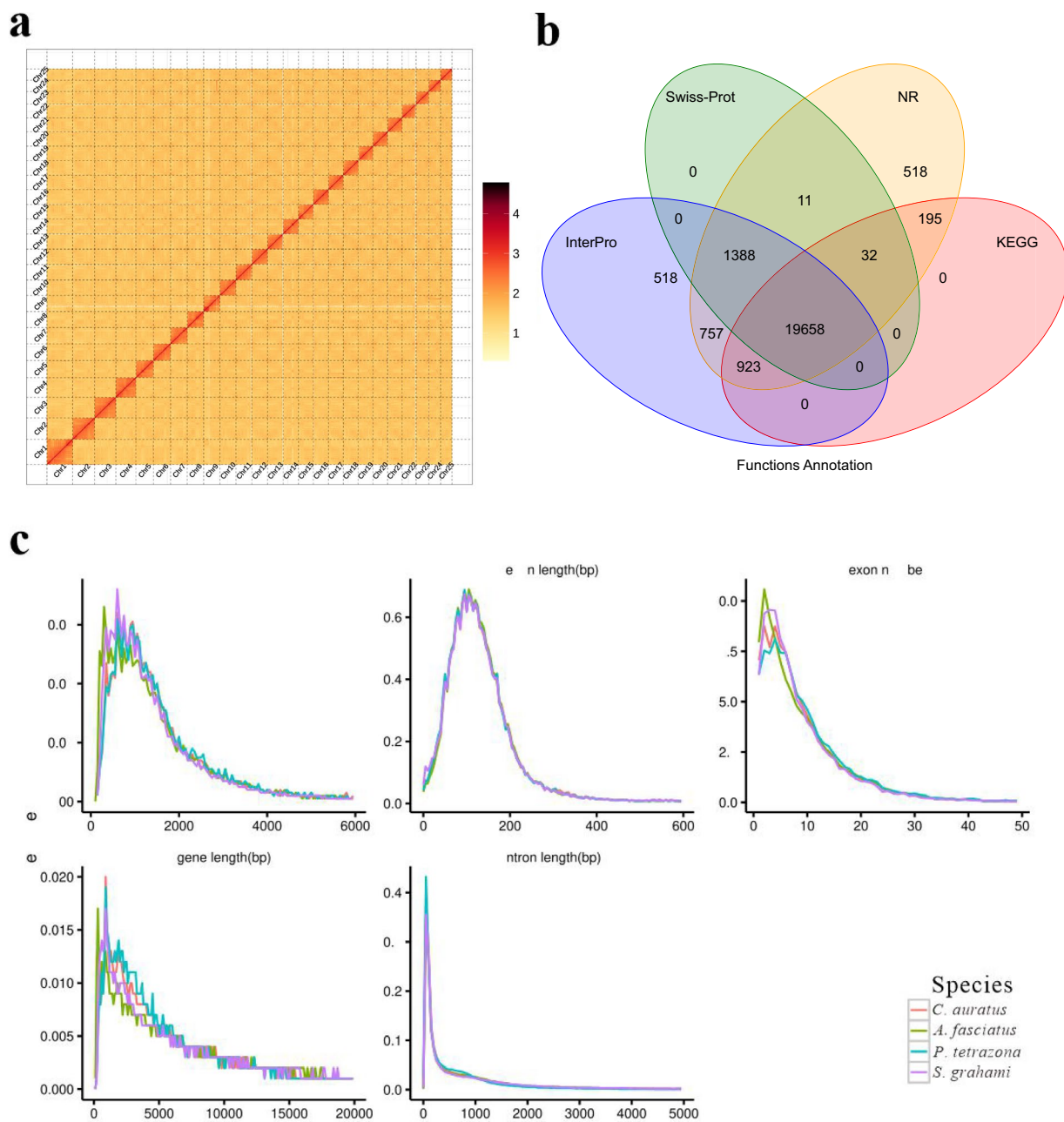


Fig. 2 Chromosomal level assembly of *A. fasciatus* genome and functional annotation. **(a)** Heat maps of Hi-C assembly of *A. fasciatus*. The color bar indicates the logarithm of the strength of the contact density. **(b)** The Venn graph of the numbers of annotated genes with different databases. **(c)** The comparisons of different gene elements in *A. fasciatus* genome with three other fish species.

	Denovo + Repbase		TE Proteins		Combined TEs	
	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	62,039,748	7.05	3,890,780	0.44	63,377,571	7.21
LINE	12,947,711	1.47	17,737,914	2.02	24,472,470	2.78
SINE	402,411	0.05	0	0	402,411	0.05
LTR	295,839,499	33.64	14,401,516	1.64	296,730,746	33.74
Unknown	18,113,674	2.06	0	0	18,113,674	2.06
Total	380,053,264	43.21	36,025,623	4.10	382,870,399	43.53

Table 3. Classification of the predicted repeat sequences in the genome of *A. fasciatus*. Note: TE, transposable element; LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeats.

	Gene set	Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	Augustus	39,649	8,793.02	1,128.93	6.39	176.78	1,422.96
	GlimmerHMM	82,692	9,420.66	651.96	4.38	148.76	2,592.23
	SNAP	55,978	15,341.35	816.46	5.86	139.32	2,988.45
	Geneid	30,650	17,697.02	1,364.34	6.40	213.17	3,024.50
	Genscan	30,914	19,582.74	1,547.51	8.21	188.43	2,500.51
Homolog	Ptet	21,870	14,149.81	1,651.62	9.29	177.73	1,507.06
	Sgra	22,672	13,058.43	1,556.54	8.61	180.68	1,510.44
	Caur	23,079	13,629.20	1,623.40	9.01	180.28	1,499.77
RNAseq	PASA	27,840	13,376.06	1,383.89	8.23	168.18	1,658.94
	Transcripts	50,168	23,864.39	3,076.21	10.43	294.92	2,204.35
EVM		35,375	12,103.93	1,281.60	7.43	172.60	1,684.29
Pasa-update*		35,122	12,467.91	1,299.95	7.50	173.23	1,717.05
Final set*		24,900	15,927.24	1,627.71	9.56	170.21	1,669.96

Table 4. Statistical analysis of predicted protein-coding genes in *A. fasciatus* genome. Note: EVM, EVidenceModeler.

	Number	Percent(%)
Total	24,900	—
Swissprot	21,089	84.70
Nr	23,482	94.30
KEGG	20,808	83.60
InterPro	23,244	93.30
GO	16,364	65.70
Pfam	19,986	80.30
Annotated	24,000	96.40
Unannotated	900	3.60

Table 5. Summary of functional annotation in *A. fasciatus* genome.

Type	Copy number	Average length (bp)	Total length (bp)	% of genome	
miRNA	2,606	118.27	308,203	0.035042	
tRNA	17,604	75.85	1,335,287	0.15	
rRNA	rRNA	9,157	135.62	1,241,850	0.14
	18S	153	744.68	113,936	0.012954
	28S	442	422.19	186,608	0.021217
	5.8S	62	156	9,672	0.001100
	5S	8,500	109.60	931,634	0.11
snRNA	snRNA	2,548	146.53	373,349	0.042449
	CD-box	303	149.79	45,385	0.005160
	HACA-box	85	150.09	12,758	0.001451
	splicing	2,101	145.30	305,275	0.034709
	scaRNA	52	183.58	9,546	0.001085
	Unknown	7	55	385	0.000044

Table 6. Statistics of annotated non-coding RNAs in the *A. fasciatus* genome assembly.

rRNAs, snRNAs, and miRNAs. In total, 43,620 non-coding RNAs were predicted, including 17,604 tRNAs, 9,157 rRNAs, 2,606 miRNAs and 2,548 snRNAs (Table 6).

Gene family construction. Firstly, the protein sequences of other 13 fish species, including *P. tetrazona*, *S. grahami*, *C. auratus*, *Opsariichthys bidenswere*, *Cyprinus carpio*, *Danio rerio*, *Ictalurus punctatus*, *Megalobrama amblycephala*, *Ctenopharyngodon idellus*, *Micropterus salmoides*, *Oreochromis niloticus*, *Cynoglossus semilaevis*, *Larimichthys crocea*, were downloaded from the public database. The low quality of sequences with less than 50 amino acids were then filtered out and only retained the longest predicted transcript per locus. Next, similarities between the protein sequences of all species were identified employing an all-to-all BLAST search with

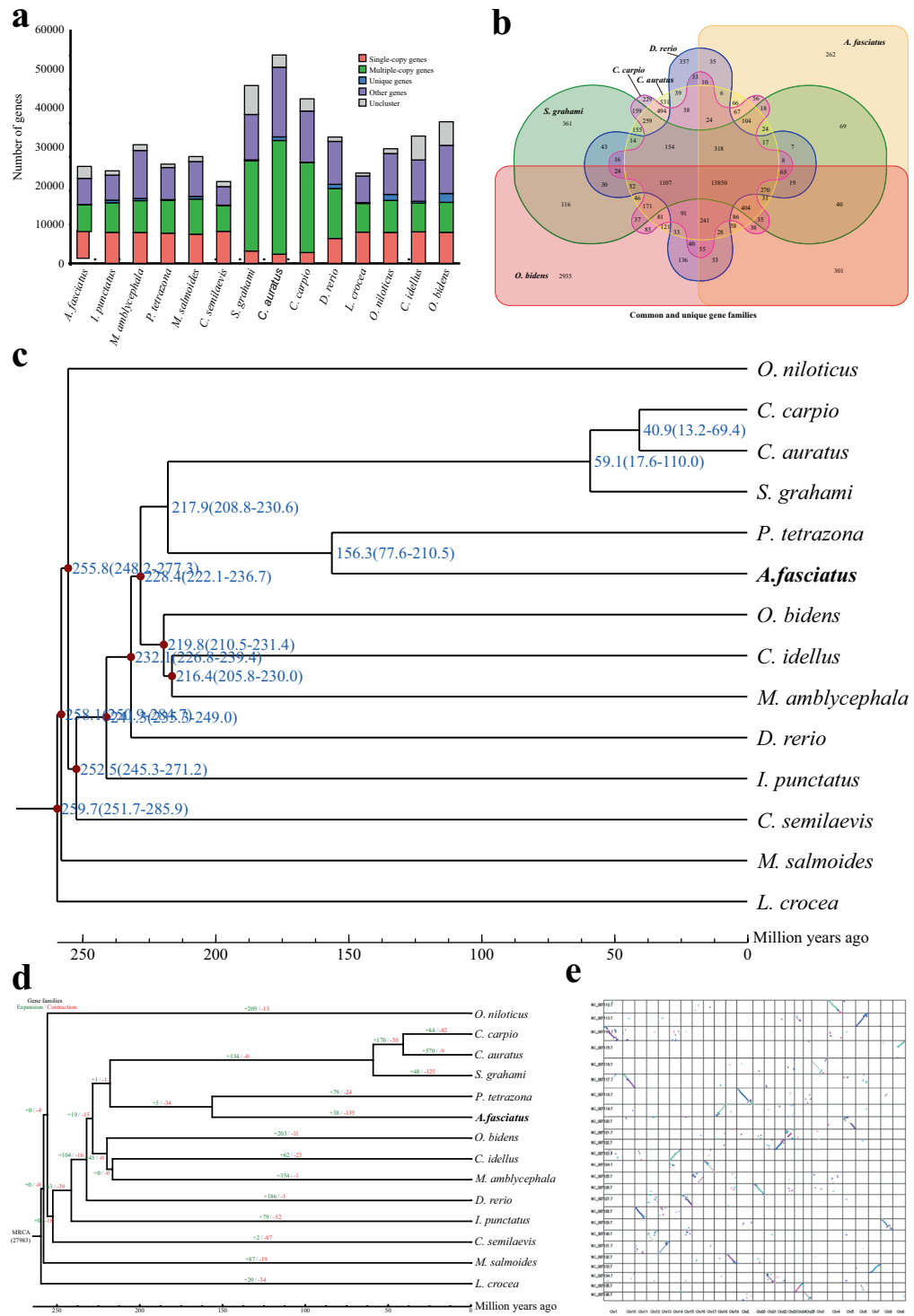


Fig. 3 Comparative genomic analysis reveals phylogenetic positioning and genome evolution of *A. fasciatus*. **(a)** Statistics of orthologous gene families in 14 representative fish species. **(b)** Venn diagram of shared and unique orthologous gene families in *A. fasciatus* and four other teleosts. **(c)** Phylogenetic analysis and divergence time tree of *A. fasciatus* and other representative species. **(d)** Statistical analysis of contraction and expansion of gene families. **(e)** Comparative synteny analysis between *A. fasciatus* and zebrafish.

an e-values of $1e^{-5}$. Finally, orthologous gene clusters were performed using the the OrthoMCL³⁴. In summary, we identified 27,983 gene families shared by *A. fasciatus* and the additional 13 species, and 10,524 gene families and 604 single-copy gene families were found in all species, respectively (Fig. 3a). Moreover, gene families from *A. fasciatus*, *O. bidens*, *S. grahami*, *D. rerio*, *C. carpio* and *C. auratus*, were further clustered, of which 13,850 gene families were shared by these fish species, and 262 gene families were specific to *A. fasciatus* (Fig. 3b).

In addition, functional annotation was conducted for unique gene families in *A. fasciatus*, and revealed that Phosphatidylinositol signaling system, GABAergic synapse, Vitamin digestion and absorption, Lysine degradation, Synaptic vesicle cycle were enriched.

Phylogenetic and evolutionary analysis. All single-copy gene families were subjected to multiple sequence alignment to generate a super alignment matrix by MUSCLE³⁵, and a phylogenetic tree was constructed using RAxML³⁶. Subsequently, the MCMCTree package in PAML³⁷ was used to estimate divergence times. As expected, evolutionary analysis demonstrated that *A. fasciatus* and *P. tetrazona* were clustered into one clade, and their divergence time was estimated to be 156.3 million years ago (Fig. 3c). Furthermore, gene expansions and contractions were analyzed employing CAFE (v3.1)³⁸ with default parameters based on the divergence times and phylogenetic relationships. A total of 38 and 135 gene families significantly expanded and contracted in *A. fasciatus*, respectively (Fig. 3d). Finally, chromosome synteny between *A. fasciatus* and *D. rerio* were carried out using MCScanX software³⁹, and visual diagram was generated by Circos. Synteny relationships analysis showed that the chromosomes of *A. fasciatus* displayed a high homology with the *D. rerio* chromosomes (Fig. 3e).

Data Records

All sequencing data had been uploaded to NCBI database via the project PRJNA1012810. The genomic Illumina sequencing data were deposited in the Sequence Read Archive at SRR25949940⁴⁰, SRR25949941⁴¹. The genomic PacBio sequencing data were deposited in the SRA at NCBI SRR25933437⁴². The transcriptomic sequencing data were deposited in the SRA at NCBI SRR25949840⁴³, SRR25949841⁴⁴, SRR25949842⁴⁵, SRR25949843⁴⁶, SRR25949844⁴⁷, SRR25949845⁴⁸. The Hi-C sequencing data were deposited in the SRA at NCBI SRR25947115⁴⁹, SRR25947116⁵⁰, SRR25947117⁵¹. The final chromosome assembly was deposited in the GenBank at NCBI with accession number: JAVLVS000000000⁵². The genome annotation file was also available in figshare⁵³. The data for the gene family construction was available in the figshare database⁵⁴.

Technical Validation

DNA quantification and qualification. DNA degradation and contamination was monitored on 1.5% agarose gels. DNA purity was checked using the NanoPhotometer[®] spectrophotometer (IMPLEN, CA, USA). DNA concentration was measured using Qubit[®] DNA Assay Kit in Qubit[®] 2.0 Fluorometer (Life Technologies, CA, USA).

Quality control of raw sequencing data. To make sure reads reliable and without artificial bias (low quality paired reads, which mainly resulted from base-calling duplicates and adapter contamination) in the following analyses, raw data were firstly processed through a series of quality control (QC) procedures in-house C scripts. QC standards as the following: (1) Removing reads with $\geq 10\%$ unidentified nucleotides (N); (2) Removing reads with $>50\%$ bases having phred quality < 5 .

RNA quality evaluation. Before transcriptomes sequencing, the quality of total RNA from six tissues was validated. The concentration was measured by Qubit Fluorometer, and the integrity was detected using Agilent 2100 Bioanalyzer. Overall, RNAs samples with a total RNA amount $\geq 10\mu\text{g}$, RNA integrity ≥ 8 , and rRNA ratio ≥ 1.5 were served as libraries construction.

Evaluation of the assembled genome. The completeness and accuracy of the *A. fasciatus* genome assembly were evaluated by multiple methods. First, Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.4.4)⁵⁵ and Core Eukaryotic Genes Mapping Approach (CEGMA, v2.5)⁵⁶ were used to assess the completeness of the assembled genome. The BUSCO results revealed that 98.3% of the complete BUSCOs and 0.7% of the fragmented BUSCOs were found in 3640 single-copy orthologs of actinopterygii_odb10, and 1.0% of BUSCOs was missing. Moreover, CEGMA evaluation showed that 96.77% (240/248) core eukaryotic genes (CEGs) were obtained. In addition, Merqury (v1.3)⁵⁷ was ran to evaluate the accuracy of genome assembly, and a high quality value (QV) of 44.81 indicated that this assembly was of good quality. Taken together, these results suggested that the assembled *A. fasciatus* genome was of high quality at chromosome level.

Code availability

No special codes or scripts were used in this work, and Data processing was carried out based on the protocols and manuals of the corresponding bioinformatics software.

Received: 27 September 2023; Accepted: 25 January 2024;

Published online: 03 February 2024

References

- Zhao, D. *et al.* Natural selection drives the evolution of mitogenomes in *Acrossocheilus*. *PLoS one* **17**(10), e0276056 (2022).
- Han, C. *et al.* Characteristics and phylogenetic studies of *Acrossocheilus* parallens (Cypriniformes, Barbinae) complete mitochondrial genome. *Mitochondrial. DNA A DNA Mapp. Seq. Anal.* **27**(6), 4708–4709 (2016).
- Fu, S. *et al.* A histological study of testis development and ultrastructural features of spermatogenesis in cultured *Acrossocheilus fasciatus*. *Tissue Cell.* **48**(1), 49–62 (2016).
- Zhang, L. & Yang, J. Genetic diversity and population demography of the endemic species *Acrossocheilus longipinnis* (Teleostei, Cyprinidae) based on mtDNA *COI* and *cyt b* gene sequences. *Mitochondrial. DNA A DNA Mapp. Seq. Anal.* **29**(3), 403–408 (2018).
- Zhang, L. & Yang, J. Genetic diversity and population structure of *Acrossocheilus yunnanensis* (Teleostei, Cyprinidae) inferred from four mitochondrial gene sequences. *Mitochondrial. DNA A DNA Mapp. Seq. Anal.* **29**(4), 606–614 (2018).
- Wei, W. *et al.* Integrated mRNA and miRNA expression profile analysis of female and male gonads in *Acrossocheilus fasciatus*. *Biology.* **11**(9), 1296–1300 (2022).

7. Yan, Y. *et al.* Age and growth of *Acrossocheilus fasciatus* (Barbinae, Cyprinidae) from the Puxi Stream in the Huangshan Mountain, China. *J. Freshwater Ecol.* **25**(1), 79–83 (2011).
8. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**(6), 764–770 (2011).
9. Ranallo-Benavidez, T. *et al.* GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**(1), 1432 (2020).
10. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**(5), 722–736 (2017).
11. Chen-Shan, C. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**(12), 1050–1054 (2016).
12. Bruce, J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* **9**(11), e112963 (2017).
13. Burton, J. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**(12), 1119–1125 (2013).
14. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**(4), 357–359 (2012).
15. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic Genome Res.* **110**(1–4), 462–467 (2005).
16. Zhao, X. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, 265–268 (2007).
17. Abrusán, G. *et al.* TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics.* **25**(10), 1329–1330 (2009).
18. Price, A. *et al.* De novo identification of repeat families in large genomes. *Bioinformatics.* **21**(1), i351–i358 (2005).
19. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**(2), 573–580 (1999).
20. Mount, D. W. Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc.* **2007**, pdb.top17 (2007).
21. Birney, E. *et al.* GeneWise and Genomewise. *Genome Res.* **14**(5), 988–995 (2004).
22. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–439 (2006).
23. Majoros, W. *et al.* TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics.* **20**(16), 2878–2879 (2004).
24. Stanke, M. *et al.* Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* **7**, 62 (2006).
25. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics.* **5**, 59 (2004).
26. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**(1), 78–94 (1997).
27. Grabherr, M. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**(7), 644–652 (2011).
28. Roberts, A. *et al.* Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics.* **27**(17), 2325–2329 (2011).
29. UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**(5), 2699 (2018).
30. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
31. Finn, R. *et al.* InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **45**(D1), D190–D199 (2017).
32. Chan, P. *et al.* tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**(16), 9077–9096 (2021).
33. Nawrocki, E. & Eddy, S. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29**(22), 2993–2995 (2013).
34. Li, L., Stoeckert, C. & Roos, D. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**(9), 2178–2189 (2003).
35. Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**(5), 1792–1797 (2004).
36. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* **30**(9), 1312–1313 (2014).
37. Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* **23**(1), 212–226 (2006).
38. De, B. *et al.* CAFE: a computational tool for the study of gene family evolution. *Bioinformatics.* **22**(10), 1269–1271 (2006).
39. Wang, Y. *et al.* MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**(7), e99 (2012).
40. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949940> (2023).
41. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949941> (2023).
42. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25933437> (2023).
43. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949840> (2023).
44. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949841> (2023).
45. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949842> (2023).
46. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949843> (2023).
47. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949844> (2023).
48. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25949845> (2023).
49. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25947115> (2023).
50. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25947116> (2023).
51. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR25947117> (2023).
52. Zheng, J. Chromosome-level genome assembly of *Acrossocheilus fasciatus*. *GenBank* <https://identifiers.org/ncbi/insdc:JAVLVS000000000> (2023).
53. Zheng, J. Genome annotation of Function annotation result of *Acrossocheilus fasciatus* genome genes. *figshare.* <https://doi.org/10.6084/m9.figshare.24198819.v1> (2023).
54. Zheng, J. Gene family construction. *figshare.* <https://doi.org/10.6084/m9.figshare.24198885.v1> (2023).
55. Simão, F. *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**(19), 3210–3212 (2015).
56. Parra, G. *et al.* CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* **23**(9), 1061–1067 (2007).
57. Rhie, A. *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**(1), 245 (2020).

Acknowledgements

This work was funded by the Key R&D Program of Zhejiang Province (2023C02050), and the “Agriculture, Farmer and Rural area” cooperation project of Zhejiang Province (2023SNJF070).

Author contributions

J.Z. and F.L. conceived this study and wrote the manuscript; J.J. and Q.R. collected the samples; S.L. and S.C. analyzed the data; M.C. and W.J. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-024-02999-6>.

Correspondence and requests for materials should be addressed to F.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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024