scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level assemblies of the White bream *Parabramis pekinensis*

Hailong Gu[1], Xinhui Zhang[2], Wentao Xu[1], Zhijing Yang[1], Ye Xu[1], Xiaoping Miao[3] & Yaming Feng[1]✉

White bream *Parabramis pekinensis* is an omnivorous fish belong to *Cyprinidae* that is widespread in Asia. In this study, we presented chromosome-level genome assemblies of *P. pekinensis* by using PacBio HiFi long reads and Hi-C technology. We assembled high-quality genome of 1.03 Gb with scaffold N50 length of 40.04 Mb, and a total of 98.31% of the assembled sequences were anchored to 24 chromosomes. BUSCO analysis revealed that the genome assembly has a high-level completeness of 98.35% gene coverage. A total of 26,542 protein-coding genes were predicted, of which 92.61% were functionally annotated. The phylogenetic analysis indicated that the lineage leading to *P. pekinensis* was diverged from the lineage to *Megalobrama amblycephala* approximately 7.6 million years ago. The high-quality genome assembly provide valuable resources for evolutionary study and genetic breeding of genus *Parabramis*.

Background & Summary

White bream *Parabramis pekinensis* belongs to family *Cyprinidae* and is widespread in all major freshwaters of Asia, mainly in China[1]. It is also an important freshwater aquaculture economic fish in China. However, bream was always regarded as a generic term such as *Megalobrama terminalis*, *Megalobrama amblycephala* and *P. pekinensis*[2]. They have no significant morphological differences and their chromosome number are all 2n = 48. Above breeds could be crossed with each other and maintain the progeny of both sexes. Thus, they are good materials for exploring the fertility of hybrids[3].

The amino acid content in *P. pekinensis* was about 167.26 ± 24.07 mg/g, which was significantly higher than *Ctenopharyngodon idella* (134.11 ± 18.98 mg/g) and *M. amblycephala* (149.62 ± 13.6 mg/g)[4]. As a herbivorous fish, *P. pekinensis* showed more delicious meat and nutritive value. People living in China were favourite of them and the selling price in market was several times higher than other freshwater herbivorous fishes. The domestic bream market capacity had expanded with a growth rate of 10% since 2008. At present, the demand of domestic bream consumption market is more than 350,000 tons on average[5]. Our previous studies mainly focused on the developmental and reproductive physiology of *P. pekinensis*. *E.g.* we investigated the growth rules of *P. pekinensis* and its potential regulatory mechanisms[6], clarified the characteristics of gonadal development[7,8], and developed artificial breeding techniques for *P. pekinensis*[9]. However, research on this fish was lack of genomics and genetic breeding. Bream were valuable aquatic species in biological research, but only the genome of *M. amblycephala* has been published so far.

Recently, High-through chromosome conformation capture (Hi-C)[10] technology has been used in whole genome sequencing (WGS), which is based on Chromosome Conformation Capture (3 C Technology)[11] and high-throughput sequencing. Hi-C could identify chromatin interactions in the whole genome exactly and obtain chromosome level genomes data completely, which also could be used to discover the potential genetic structure of species, explore species diversity, construct haplotype maps and conduct genome-wide association analysis. At present, WGS, transcriptome sequencing, whole-genome re-sequencing, simplified genome has been widely used in nonpattern organism to candidate sex determination genes or interlocking molecular marker by combining genome-wide association analysis (GWAS), quantitative trait locus (QTL), expression

[1]Taizhou Institute of Agricultural Science, Jiangsu Academy of Agricultural Sciences, Taizhou, 225300, China. [2]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI Marine, Shenzhen, 518081, China. [3]Jiangzhiyuan Fishery Technology Co. Jingjiang, Taizhou, China. ✉e-mail: fym771118@163.com

Stat Type	Contig Length (bp)	Contig Number	Scaffold Length (bp)	Scaffold Number (bp)	Gap Length (bp)	Gap Number
N50	29,690,824	13	40045890	11	100	20
N60	25,824,068	17	38517899	13	100	24
N70	24,171,014	21	37780210	16	100	28
N80	17,584,455	26	35707912	19	100	32
N90	11,160,449	33	32866210	22	100	36
Longest	57,131,658	1	66143525	1	100	39
Total	1,031,041,505	123	1,031,045,405	84	3900	39

Table 1. Statistics of Hi-C assisted assembly.

analysis and functional verification, such as in *Channa argus*[12], *Ctenopharyngodon idellus*[13], *Larimichthys crocea*[14] and *Salmo salar*[15].

In the present study, we constructed a high-quality chromosome-level reference genome using long-reads generated by a PacBio platform, short reads generated by an Illumina platform, and the Hi-C analysis technology. In addition, phylogenetic analyses based on single-copy genes were performed to understand the relationship between *P. pekinensis* and other species. It is the first genome assembly of *P. pekinensis*, which will be helpful to understand the gene structure, function and arrangement of this species, providing a basis for subsequent studies on genetic breeding, evolutionary analysis and germplasm resource conservation.

Methods

Sample collection, sequencing. We extracted genomic DNA from the muscle samples of a *P. pekinensis* (~1.5 kg, estimated age 2 years, female), provided by Jingjiang Binjiang aquaculture seed farm, Taizhou City, Jiangsu Province, China. Genomic DNA of each sample was extracted from muscle tissue using the QIAamp DNA Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's instructions. The quality and quantity of the extracted DNA were assessed via agarose gel electrophoresis and an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). An Illumina DNA library of *P. pekinensis* with an insert size of 350 bp was constructed using the Genomic DNA Sample Prep kit (Illumina, San Diego, CA, USA) following the manufacturer's instructions, and sequenced on an Illumina HiSeq X Ten platform using a 150-bp paired-end strategy. A total of 58.28 Gb short reads was generated for *P. pekinensis* and used for genome size estimation.

For PacBio HiFi long-read sequencing, about 10 μg gDNA was used to construct long-read libraries by using a SMRTbell Express Template Prep Kit 2.0 based on PacBio's standard protocol (Pacific Biosciences, USA), which were sequenced through a PacBio Sequel II System. A total of 77.96 Gb HiFi reads with N50 sizes of 22,713 bp was obtained using the CCS v6.0.0[16] (Circular Consensus Sequencing) software with the parameter -min-passes 1, which covered approximately 88 × genome (Table 1).

For the high-throughput chromosome conformation capture (Hi-C) sequencing, muscle tissue (about 1 g) from a female individual was collected, and Hi-C library was constructed by using GrandOmics Hi-C kit (the applied restriction enzyme is DpnII, GrandOmics, China) according to the manufacturer's protocol. The Hi-C libraries were then sequenced using DNBSEQ T7 platform (MGI, China) with the paired-end module. In total, 103.47 Gb of raw reads were generated for *P. pekinensis*. Subsequently, fastp v0.19.5[17] was applied to filter the adaptors and low-quality reads. Finally, a total of 101.42 Gb high-quality reads were retained for construction of pseudo-chromosomes.

RNA extraction and transcriptome sequencing. The total RNA samples were extracted from muscle, liver, ovary, kidney and gill tissues using a standard Trizol protocol (Invitrogen, Carlsbad, California, USA), and purified using a Qiagen RNeasy mini kit (Qiagen, USA). RNA with equal amounts from each tissue was mixed for creating Illumina cDNA library followed the manufacture's guideline, which was then sequenced on a HiSeq X Ten platform (Illumina, USA) with paired end mode and PacBio platform (Pacific Biosciences, USA). Around 36 Gb (HiSeq: 34.04 Gb and HiFi: 1.96 Gb) transcriptome data were generated for assistance to gene annotations.

Genome size assessment and preliminary assembly. To evaluate the genome size of *P. pekinensis*, we analyzed the K-mer (K = 17) frequency distribution of the filtered 54.13 Gb Illumina reads and obtained the K-mer depth distribution curve (Fig. 1b). The depth of K-mer peak was 34, and the number of k-mer was 31,941,945,635. Therefore, the genome size of *P. pekinensis* estimated about 939.47 Mb Fig. 2. The genome was assembled using the hifiasm[18] software with default parameters, and obtained the genome of preliminary assembly. These contigs were then polished with Nextpolish (v1.10)[19] using the Illumina short reads to fix possible base errors. The primary genome assembly was 1.03 Gb in length and N50 size was 31.22 Mb, consistent with the estimated genome size. Statistics were showed in Table 2.

Chromosome construction and genome assembly. Firstly, the clean paired-end reads were mapped to the draft assembled sequence using bowtie2 (v2.3.2) (-end-to-end --very-sensitive -L 30)[20] to get the unique mapped paired-end reads. Subsequently, HiC-Pro (v2.8.1)[21] pipeline was applied to detect valid ligation products and only valid contact paired reads were retained for further analysis. Finally, the scaffolds were further clustered, ordered, and oriented scaffolds onto chromosomes using LACHESIS[22] software with optimized parameters (CLUSTER_MIN_RE_SITES = 100, CLUSTER NONINFORMATIVE RATIO = 1.4, CLUSTER_MAX_LINK_DENSITY = 2.5, ORDER MIN N RES IN SHREDS = 60, ORDER MIN N RES IN TRUNK = 60). After this,

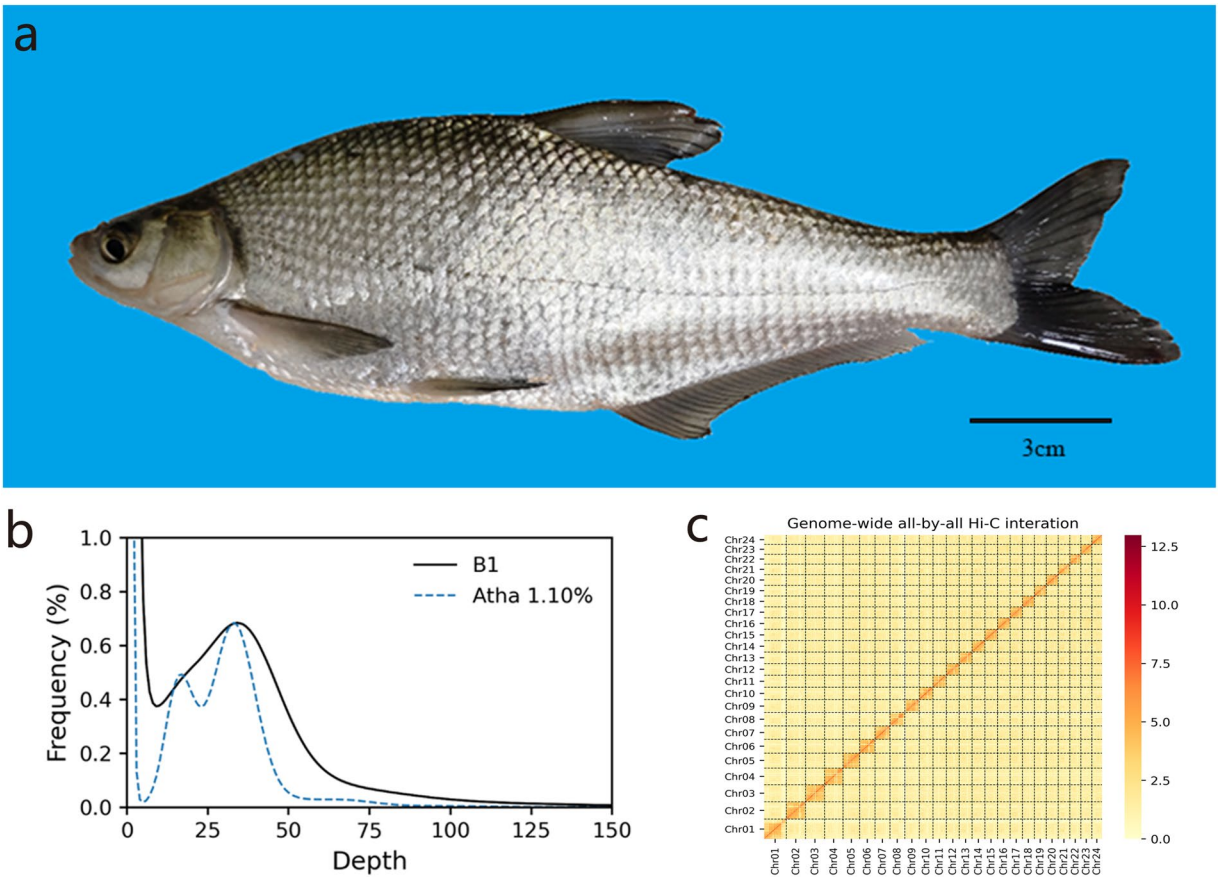

Fig. 1 *Parabramis pekinensis* (**a**). Frequency distribution of 17-mer depth in the second-generation sequencing data of *P. pekinensis* (**b**). Hi-C chromatin interaction map of the *P. pekinensis* assembly (**c**).

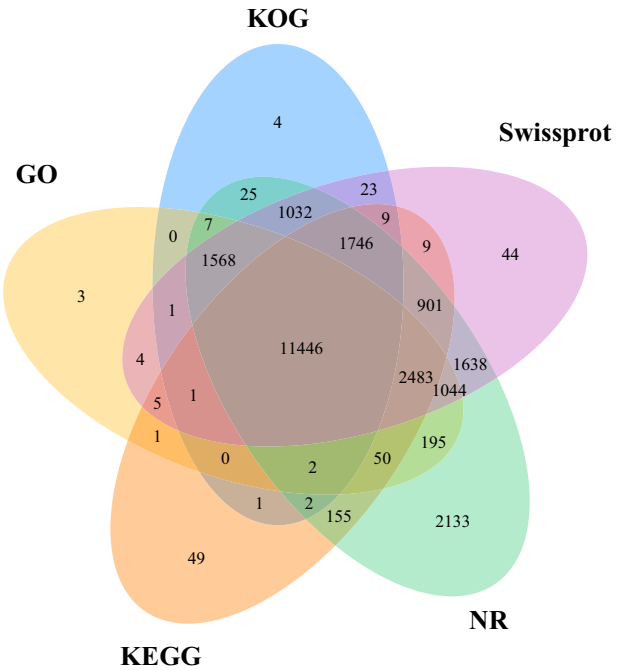

Fig. 2 The Venn figure of function annotations from different databases.

Juicebox v1.11.08[23] was employed to visually inspect and rectify any errors in the order and orientation of the contigs or assembly errors within the contigs. We hence obtained the final genome assembly with a size of 1.03 Gb,

Library type		Raw data (Gb)	Clean data (Gb)	Read length (bp)	Coverage (×)
Short reads		58.28	54.13	150	57.64
HiFi		/	77.96	22,223[a]	88.02
Hi-C		103.47	101.42	150	108
mRNA	NGS	37.75	34.04	150	/
	CCS	/	1.96	2,556[a]	/

Table 2. Statistics of sequencing data of *P. pekinensis*. [a]This number is read N50.

Chr	Length	Contig Num
Chr01	66,143,025	6
Chr02	57,690,969	7
Chr03	57,131,658	1
Chr04	55,387,430	2
Chr05	50,141,793	1
Chr06	45,758,010	3
Chr07	45,037,298	3
Chr08	44,330,729	3
Chr09	43,305,917	7
Chr10	41,085,222	3
Chr11	40,045,590	4
Chr12	39,073,843	3
Chr13	38,517,799	2
Chr14	38,105,042	1
Chr15	37,932,325	2
Chr16	37,780,110	2
Chr17	37,045,656	1
Chr18	35,969,630	1
Chr19	35,707,812	2
Chr20	35,579,175	2
Chr21	35,306,415	1
Chr22	32,866,010	3
Chr23	32,787,021	2
Chr24	30,952,917	1
Total	1,013,681,396	63

Table 3. Scaffold clustering 24 chromosome length statistics.

of which 98.31% are anchored into 24 chromosomes (Table 3). The scaffold and contig N50 values of the overall chromosome-level genome assembly are 40.04 Mb and 29.69 Mb, respectively (Table 1).

Finally, we used BUSCO (Benchmarking Universal Single-Copy Orthologs)[24] and CEGMA[25] to assess the completeness of the genome and core gene. BUSCO against actinopterygii_odb10 database was employed to assess the completeness of this genome assembly, showing that the assembly contains 98.35% of complete BUSCO genes including 96.73% single-copies and 1.62% duplicates, which indicated major conserved genes were assembled completely and results were highly credible. Moreover, we obtained 243 core genes (97.98%), which including 183 complete core genes (73.79%), and it indicated that assembled genome has complete core genes.

Repeats and gene annotation. A combined strategy based on de novo searches and homologue alignments was used to annotate whole-genome repeat elements. GMATA[26] was used to search for simple repeats (SSRs) with short repeat units, and the genome was subjected to SSR softmask. Tandem Repeats were re-searched with Tandem Repeats Finder (TRF)[27], and softmask was performed on tandem repeats. MITE-hunter was used to construct the TE library, and RepeatModeler[28] was used to search the repeat sequences de novo to construct the RepMod library. The combination of TE, RepMod and Repbase library revealed that 572.45 Mb repeats were predicted which accounting for 55.52%. Tandem repeats were predicted 9.50 Mb, which accounted for 0.92%. Short interspersed nuclear elements (SINEs) and long terminal repeats (LTRs) accounted for 1.17% and 14.21% of the whole genome, respectively, and long interspersed nuclear elements (LINEs) accounted for 8.46%.

For protein-coding gene prediction, a comprehensive strategy combining homology-based prediction, transcript-based prediction and de novo prediction was employed. For de novo predictions, the gene structures were identified with AUGUSTUS v3.2.1[29]. For homology-based predictions, protein sequences from

Gene set	Total number of genes	Average gene length(bp)	Average CDS length(bp)	Average exons number per gene	Average exon length(bp)	Average intron length(bp)
	26,542	19,708.88	1,684.26	10.06	167.4	1,989.22

Table 4. Protein coding gene annotation statistics.

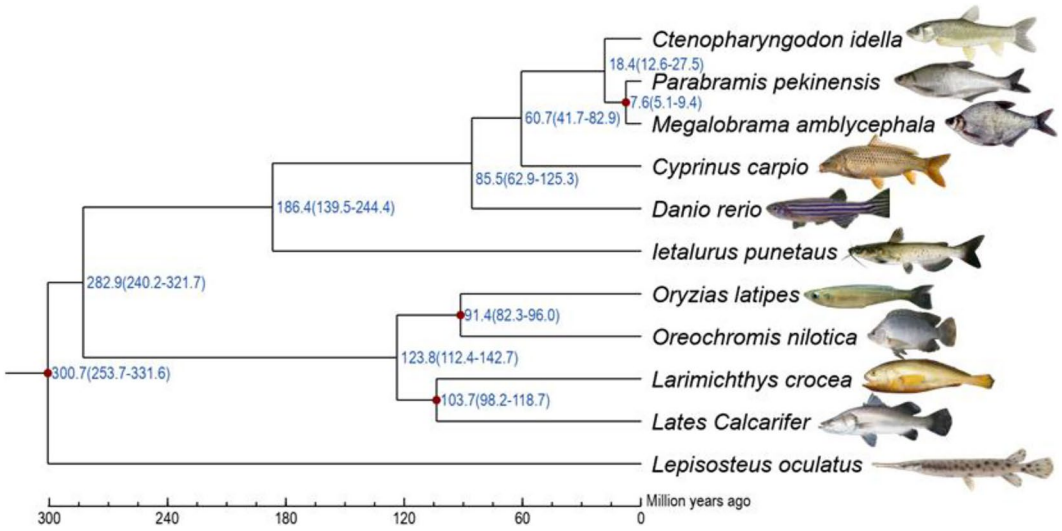

Fig. 3 The phylogenetic analysis of *P. pekinensis* and other 10 species. The numbers in blue near the nodes indicate the divergence time (million years ago, MYA), and the blue numbers inside the brackets are bootstrap values.

Oreochromis niloticus, *Megalobrama amblycephala*, *Larimichthys crocea*, *Ictalurus punctatus*, *Cyprinus carpio*, *Oryzias latipes*, *Lates calcarifer*, *Danio rerio*, and *Ctenopharyngodon idell* were downloaded from NCBI and GeMoMa v1.6.4[30] was adopted with default parameters. For transcript-based prediction, the short reads generated from the Illumina platform were assembled using Trinity v2.5.2[31], the gene structures were identified using PASA v2.3.3[32]. Finally, gene sets were integrated by the Evidence Modeler (EVM) pipeline v1.0, which based on PASA prediction > = GeMoMa prediction > = ab initio prediction to count results. A total of 26,542 protein-coding genes were predicted and average gene length was 19,708.88 bp (Table 4). The BUSCO completeness values were 96.04% of the predicted proteins. Next, we compared the number of genes, gene length, coding DNA sequence (CDS) length, exon length, and intron length distributions with those of other fish species.

We used Blastp and InterproScan to compare *P. pekinensis* protein genes codes with public protein databases, which involved NCBI non-redundan (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of orthologous groups for eukaryotic complete genomes (KOG), Gene Ontology (GO), Swissport. The results of these comparisons were merged and integrated to predict the functional information of the obtained genes. In this study, 26,542 proteins were predicted and 24,581 genes encoding proteins were compared to functional databases, accounted for 92.61% of total genes.

Compared to the KEGG database, the results of functional pathway clustering were shown that 16,860 protein-coding genes were enriched in 35 biochemical pathways, forming five major categories: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Metabolism and Organismal Systems. Compared to the GO database, the results of functional clustering showed that 16,810 genes were annotated into 60 secondary functions from three aspects, including Cellular component, biological process and Molecular function.

Phylogenetic tree. A species tree was constructed using those single-copy orthologs from the whole genomes of *P. pekinensis* and other 10 representative ray-finned species, including *Lepisosteus oculatus*, *O. nilotica*, *D. rerio*, *O. latipes*, *L. calcarifer*, *I. punetaus*, *C. carpio*, *C. idella*, *M. amblycephala* and *L. crocea* were downloaded from the NCBI. The protein-coding genes of 11 representative species were selected for gene family analysis using OrthoMCL v1.1[33] with default parameters (Li *et al.*[33]). A total of 1,224 homologous single-copy genes were obtained. These homologous single-copy genes were compared using the "-align" parameter of Muscle v3.7[34] (Edgar[34]). Gblock v0.19b[35] was employed to extract conserved sequences in comparison results with the parameter "-b4 = 5 -t = d -e = 0.2". The phylogenetic tree was subsequently constructed by PhyM v3.0[36] (Guindon *et al.*[36]). Based on the phylogenetic topology, MCMCTREE in the PAML v4.9e[37] package was employed to estimate divergence times among *P. pekinensis* and other examined species, with assistance of fossil records from the TimeTree v5.0[38]. Here, our phylogenetic tree showed that *P. pekinensis* was evolutionarily closer to *M. amblycephala*, which also belongs to *Culterinae*, with a divergence time of 7.6 million years ago (Mya). In addition, the two species have

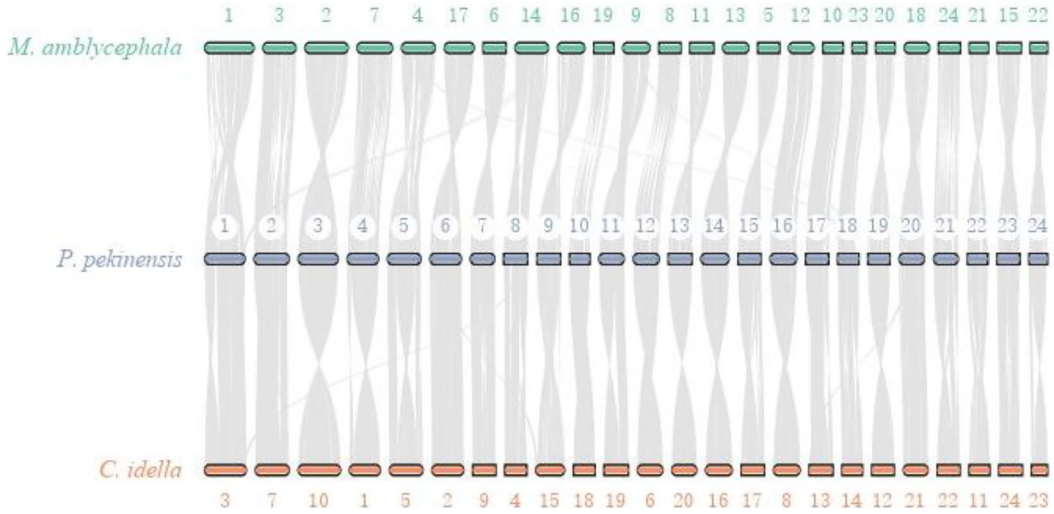

Fig. 4 A synteny analysis of the chromosomes among *P. pekinensis*, *M. amblycephala* and *C. idella* Chromosomes are ordered from the longest to the shortest.

a common ancestor with *C. idella*, which belong to the same family *Cyprinidae*, and the divergence time of the two clades was about18.4 Mya (Fig. 3).

Based on the protein coding sequences and gene structure, JCVI v190213[39] was used to perform chromosomal collinearity analysis among *P. pekinensis*, *M. amblycephala* and *C. idella*. Through collinearity analysis, it can be seen that these four species have a good collinearity relationship, and the chromosomes show a one-to-one correspondence relationship, which also indicates that the assembled genome is complete and high-quality (Fig. 4).

Data Records

The raw reads of the MGI, HiFi, Hi-C and transcriptome sequencing for *P. pekinensis* were deposited at NCBI under the accession number PRJNA1082057. Raw reads are available in the Sequence Reads Archive (SRA) with the accession number SRP509380[40]. The genome assemblies were deposited at GenBank with the accession number GCA_041684115.1[41]. The annotation files have been submitted to appropriate public Figshare database[42].

Technical Validation

The final genome assembly of *P. pekinensis* was 1.03 Gb in size with an N50 of 31.22 Mb, consisting of 24 chromosomes. Also, the Hi-C chromosome interaction intensity signal demonstrated the high quality of genome assembly. We also performed BUSCO analysis with the actinopterygii_odb10 database to assess the completeness of the genome assembly and protein-coding gene for *P. pekinensis*. More than 98.35% and 96.04% of complete BUSCOs were identified in genome assembly and protein-coding genes, respectively, showing the genome assembly and annotation of *P. pekinensis* are complete and high-quality. Furthermore, we compared the conservation synteny among *P. pekinensis*, *M. amblycephala* and *C. idella* to validate the chromosome assembly. We observed highly conserved synteny and strict correspondence of chromosome assignment (Fig. 4).

Code availability

All software used in this work is publicly available, and the software versions and parameters used are described in the Methods section. The parameters not mentioned in the analysis were used as default parameters suggested by the developer. No custom script was used in this study.

Received: 11 November 2024; Accepted: 13 March 2025;
Published online: 27 May 2025

References

1. Zhao, C. *et al.* Growth Characteristics of *Parabramis pekinensis* in Xiaolangdi Reservoir of the Yellow River. *Anh Agr Sci Bul.* **25**(01), 64–67 (2019).
2. Hu, X. *et al.* Characterization of the mitochondrial genome of *Megalobrama terminalis* in the Heilong River and a clearer phylogeny of the genus *Megalobrama*. *Sci Rep.* **9**, 8509 (2019).
3. Lin, Y. A Comparative of the Karyotypes in Chinense Bream, Herbivorous Bream and Their Hybrid. *Zoological Research* **5**(zk), 65–66 (1984).
4. Ling, M. Nutritive Quality and Nutritional Component in the Muscle of *Megalobrama tarminalis* and *Parabramis pekinensis*. *Chinese Journal of Zoology* (2013).
5. Liu, X. *et al. China Fisheries Statistics Yearbook* (China Agricultural Press, 2023).
6. Xu, W. *et al.* Effects of morphological traits on body weight and analysis of growth-related genes of *Parabramis pekinensis* at different ages. *BMC Zool.* **8**, 13 (2023).

7. Gu, H., Feng, Y., Yang, Z. & Wang, J. Histological Observation of Gonadal Development of White Bream *Parabramis pekinensi*. *Chinese Journal of Fisheries* **35**, 03 (2022).
8. Gu, H. *et al*. Differential study of the *Parabramis Pekinensis* intestinal microbiota according to different habitats and different parts of the intestine. *Ann Microbiol* **71**, 5 (2021).
9. Gu, H., Cai, Y., Yang, Z., Tang, W. & Feng, Y. Study on the stress response of *Parabramis pekinensis* to different oxytocin. *Freshwater Fisheries* **52**(01), 21–29 (2022).
10. Lieberman-Aiden, E. *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950), 289–93 (2009).
11. Dekker, J. *et al*. Capturing chromosome conformation. *Science* **295**(5558), 1306–11 (2002).
12. Liu, H. *et al*. High-density genetic linkage map and QTL fine mapping of growth and sex in snakehead (*Channa argus*). *Aquaculture* **519**, 0044–8486 (2020).
13. Wang, Y. *et al*. The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. *Nat Genet.* **47**, 625–631 (2015).
14. Lin, A. *et al*. Study of sex-specific molecular markers and some sex-related genes in Large Yellow Croaker (*Larimichthys crocea*). Jimei Univerdity (2017).
15. Carmichael, S. N. *et al*. Identification of a Sex-Linked SNP Marker in the Salmon Louse (*Lepeophtheirus salmonis*) Using RAD Sequencing. *Plos One* **8**(10), 271–272 (2013).
16. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
17. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
18. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**(2), 170–175 (2021).
19. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
20. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
21. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16**(1), 1–11 (2015).
22. Joshua, N. Burton & Andrew, Adey. Chromosome-scale Contiging of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**(12), 1119–1125 (2013).
23. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**(1), 99–101 (2016).
24. Simao, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**(19), 3210–2 (2015).
25. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* **23**(9), 1061–7 (2007).
26. Wang, X. & Wang, L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. *Front Plant Sci.* **7**, 1350 (2016).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**(2), 573–80 (1999).
28. Bedell, J. A., Korf, I. & Gish, W. Masker Aid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**(11), 1040–1 (2000).
29. Stanke, M. *et al*. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**(5), 637–44 (2008).
30. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**(9), e89 (2016).
31. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology* **29**(7), 644 (2011).
32. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1–22 (2008).
33. Li, L., Stoeckert, C. & Roos, D. Identification of ortholog groups for eukaryotic genomes. *Genome research* **13**(9), 2178–2189 (2003).
34. Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5), 1792–1797 (2004).
35. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**(4), 540–552 (2000).
36. Guindon, S. *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**(3), 307–321 (2010).
37. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**(8), 1586–1591 (2007).
38. Kumar, S. *et al*. TimeTree 5: an expanded resource for species divergence times. *Molecular biology and evolution* **39**(8), msac174 (2022).
39. Tang, H. *et al*. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP509380 (2024).
41. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_041684115.1 (2024).
42. https://doi.org/10.6084/m9.figshare.25868809 (2024).

Acknowledgements

This work was supported by the Project of Seed Industry revitalization of Jiangsu Province (JBGS [2021] 138).

Author contributions

Y. Feng conceived the study. H. Gu and W. Xu collected samples. Bioinformatics analysis was performed by X. Zhang, Z. Yang and Y. Xu. H. Gu and X. Zhang wrote and revised the original manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

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.