Liu et al. *Zool. Res.* 2022, 43(2): 217−220
https://doi.org/10.24272/j.issn.2095-8137.2021.351

ZR | Zoological Research

Letter to the editor Open Access

Chromosome-level genome assembly of the dotted gizzard shad (*Konosirus punctatus*) provides insights into its adaptive evolution

Konosirus punctatus is an economically important marine fishery resource and is widely distributed from the Indian to Pacific oceans. It is a good non-model species for genetic studies on salinity and temperature adaptation. However, a high-quality reference genome has not yet been reported. Here, an 800.00 Mb high-quality chromosome-level genome with a contig N50 length of 2.14 Mb was assembled using Illumina, Pacific Biosciences, and Hi-C sequencing technology. The assembled sequences were anchored to 24 pseudochromosomes by the Hi-C data. In total, 24 298 protein-coding genes were predicted, 91.08% of which were successfully annotated with putative functions. Furthermore, 587 putative genes were identified as being under positive selection. This new high-quality *K. punctatus* reference genome provides a fundamental resource for a deeper understanding of temperature and salinity adaptation and species conservation.

The dotted gizzard shad (*K. punctatus*) (Clupeiformes: Clupeidae) is widely distributed along the coastlines of the Indian and Pacific oceans (Song et al., 2017). As a euryhaline fish, *K. punctatus* can survive in both fresh and seawater (Kuroda et al., 2002) and their spawning grounds and timing are highly related to water temperature and salinity (Kong et al., 2004). The spawning period generally peaks at water temperatures of 17–19 °C (Shan et al., 2020) and the species is known to migrate into brackish water for breeding (Gwak et al., 2015). Thus, specific salinity and water temperature ranges appear to be basic biological factors required for spawning in *K. punctatus* (Kong et al., 2004). These biological properties make *K. punctatus* a valuable model for studying the molecular mechanisms underlying the evolution of salinity and temperature adaptation. Furthermore, the primary food sources of *K. punctatus* are phytoplankton, zooplankton, and algae (Gao et al., 2016), and thus *K. punctatus* plays an important role in material circulation and energy flow in marine ecosystems. However, with the continuous development and utilization of the ocean, the destruction of marine ecosystems and biological resources has intensified, including enormous damage to *K. punctatus* habitat (Li et al., 2017; McCay et al., 2006). *Konosirus punctatus* has a strong regenerative ability and abundant resources, but the resources have sharply declined in recent years (Liu et al., 2020). In recent years, advances in genomic technology, especially third-generation sequencing, has presented a novel opportunity to explore the genetic basis of environmental adaptations. Therefore, high-quality genomes and population resources are essential to understand the critical biological processes related to these adaptations. Thus, high-quality genome assembly will not only benefit the above research areas but also improve our understanding of the adaptive evolution of *K. punctatus*.

We collected a single fish from Zhoushan, Zhejiang Province, China (N29°32′42.60″, E122°26′54.97″) in October 2019. Muscle, eye, gonad, gill, liver, and spleen tissues were collected and preserved in liquid nitrogen before DNA and RNA extraction. Muscle tissue was used for DNA sequencing and all tissues were used for transcriptome sequencing. DNA was extracted from muscle tissue using the phenol/chloroform DNA extraction method. The quantity and quality of DNA were determined using a Qubit fluorometer (Thermo Fisher Scientific, USA) and Agilent 2100 Bioanalyzer (Agilent

Received: 20 December 2021; Accepted: 24 January 2022; Online: 24 January 2022

Foundation items: This study was supported by the National Natural Science Foundation of China (41806156); Zhejiang Provincial Natural Science Foundation of China (LY20C190008, LY22D060001, Y22D064798); Science and Technology Project of Zhoushan (2020C21016); Fund of Guangdong Provincial Key Laboratory of Fishery Ecology and Environment (FEEL-2021-8); Open Foundation from Key Laboratory of Tropical Marine Bio-Resources and Ecology, Chinese Academy of Sciences (LMB20201005); and Open Foundation from Marine Sciences in the First-Class Subjects of Zhejiang (20200201, 20200202).

Technologies, USA). Total RNA was extracted from all tissues using TRIzol reagent (Invitrogen, USA). The NanoDrop ND-1 000 spectrophotometer (Labtech, USA) and 2 100 Bioanalyzer (Agilent Technologies) were used to check RNA quality. The Illumina NovaSeq 6 000 and PacBio Sequel II platforms were applied for genomic sequencing to generate short and long genomic reads, respectively. Paired-end libraries were constructed with an insert size of 300 bp according to the standard Illumina protocols. A 20 kb DNA SMRTbell sequencing library was sequenced with the PacBio Sequel platform.

In total, 89.92 Gb of clean data were generated by Illumina sequencing (Supplementary Table S1). Jellyfish v2.2.10 was used for K-mer analysis. K-mer analysis showed that the sample genome size was ~797 Mb but was 787 Mb after correction with a heterozygosity rate of 0.96% and repeat sequence ratio of 39.22% (Supplementary Table S2). In total, 84.11 Gb of high-quality data were generated using the PacBio Sequel II platform. The PacBio long reads were used for *de novo* genome assembly with NextDenovo v2.3.1. Arrow in the GenomicConsensus package v2.3.3 was used to polish the genome using the PacBio long reads with MinCoverage. Two rounds of polishing using the Illumina short reads were then applied with Pilon v1.2.3. De-redundancy of the assembled genomes was performed using Purge_haplotigs v1.1.1. Finally, the PacBio sequencing data resulted in an 800 Mb assembly with a contig N50 of 23.07 Mb (Supplementary Table S3). Genome assembly completeness was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.1 (Seppey et al., 2019) to search the genome in the *Actinopterygii* database, which included 4 584 single-copy orthologs. Based on BUSCO analysis of the *K. punctatus* genome, the assembly contained 93.54% complete BUSCOs, 89.44% of which were complete and single copies and 4.10% of which were complete and duplicated (Supplementary Table S4).

An Hi-C sequencing library was constructed to obtain a chromosome-level genome assembly. High-quality Hi-C reads were mapped to the polished *K. punctatus* genome using Bowtie v1.2.22. LACHESIS v1.03with default parameters was applied to perform genome assembly at the chromosome level using corrected contigs and valid Hi-C reads. Juicer v2.0 was used to construct species chromosome and genome-wide interaction maps to appraise the quality of genome assembly at the chromosome level. The Hi-C library generated 72.5 Gb of clean data (Supplementary Table S1). The quality of the sequenced data was evaluated, resulting in 469 685 574 clean reads and 68 764 334 007 bp of clean bases (Supplementary Table S5). Using LACHESIS, the assembled sequences were anchored to 24 pseudochromosomes (Figure 1A). Finally, the *K. punctatus* genome assembly was 0.8 Gb with a contig N50 of 2.02 Mb and scaffold N50 of 32.23 Mb (Supplementary Tables S6, S7). For clarity, Figure 1B shows the distribution of gene density, repeat density, and GC density of the 24 pseudochromosomes of the *K. punctatus* genome. BWA-MEM v0.7.10-r789 and BLASR v5.3.3 were used to evaluate the completeness and accuracy of the genome assembly. Based on evaluation of the genome assembly, we obtained 0.632% heterozygous single nucleotide polymorphisms (SNP) and 0.07% homozygous SNPs (Supplementary Table S8). In addition, the homozygous and heterozygous insertion-deletion (InDel) rates were 0.018% and 0.280%, respectively (Supplementary Table S8). Thus, the assembly showed a high rate of correct single bases.

Homology comparison and *de novo* prediction were used to annotate the repetitive sequences of the *K. punctatus* genome. RepeatMasker v4.0.7 and RepeatProteinMask v4.1.0 were used to search the genome sequences for known repeat elements based on the RepBase database. LTR_FINDER v1.0.2 and RepeatModeler v2.0 were used to establish the *de novo* repeat sequence library, and RepeatMasker v4.0.7 was used to predict genes. A total of 327.23 Mb of repeat sequences were detected, accounting for 40.88% of the assembled genome (Supplementary Table S9). In total, 19.91% (159.37 Mb) of the repeat sequences were annotated using the *de novo* method (Supplementary Table S9). Repetitive sequences primarily consisted of DNA transposable elements (151.38 MB; 18.91% assembly), long terminal repeat elements (72.75 Mb; 9.09%), and long interspersed elements (39.37 Mb, 4.92%) (Supplementary Table S10). Three strategies based on ab initio, homology, and RNA sequencing (RNA-seq) were applied to predict the protein-coding genes. AUGUSTUS v2.7 and GENSCAN v1.0 were used for *ab initio* gene prediction. For homology-based prediction, protein sequences of *Anabas testudineus*, *Clupea harengus*, *Amphiprion ocellaris*, *Denticeps clupeoides*, and *Acanthochromis polyacanthus* were downloaded from the NCBI database and aligned to the *K. punctatus* genome using tBLASTn (e-value=1e-5). GeneWise v2.4.0 was used to predict the exact gene structure of the corresponding genomic region in each blast. For transcriptome-based prediction, RNA-seq reads were directly mapped to the genome using TopHat v2.1.1. The mapped reads were subsequently assembled into gene models (Cufflinks-set) using CUFFLINKS v2.02. EvidenceModeler (EVM) v.1.1.1 was used to integrate the above three predicted gene sets into a non-redundant and more complete gene set. Finally, PASA v2.0.2 was applied to combine the transcriptome assembly results, correct the EVM annotation results, and add untranslated region (UTR), variable splicing, and other information to obtain a final gene set. In total, 24 298 protein-coding genes were predicted with an average gene length of 16 809 bp (Supplementary Table S11). Furthermore, 22 131 predicted genes (91.08%) were successfully annotated based on alignment with nucleotide, protein, and annotation databases (i.e., InterPro, NR, SwissProt, TrEMBL, KOG, GO, and KEGG) using BLAST+ v2.2.28 (Supplementary Table S12). The annotations for non-coding RNA (ncRNA) included transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA), and small nuclear RNA (snRNA). tRNAscan-SE v1.3.1 was used to identify the tRNA sequence in the genome. As rRNA is highly conserved, the rRNA sequence of a closely related species was selected as the reference sequence, and rRNA in the genome was found via blast alignment with a threshold e-value<1e-10. The covariance model of the Rfam family and INFERNAL v1.1 were used to predict the miRNA and snRNA sequences in the genome. Finally, 338 miRNAs, 2 752 tRNAs, 371 rRNAs, and 884 snRNAs were identified in the assembled

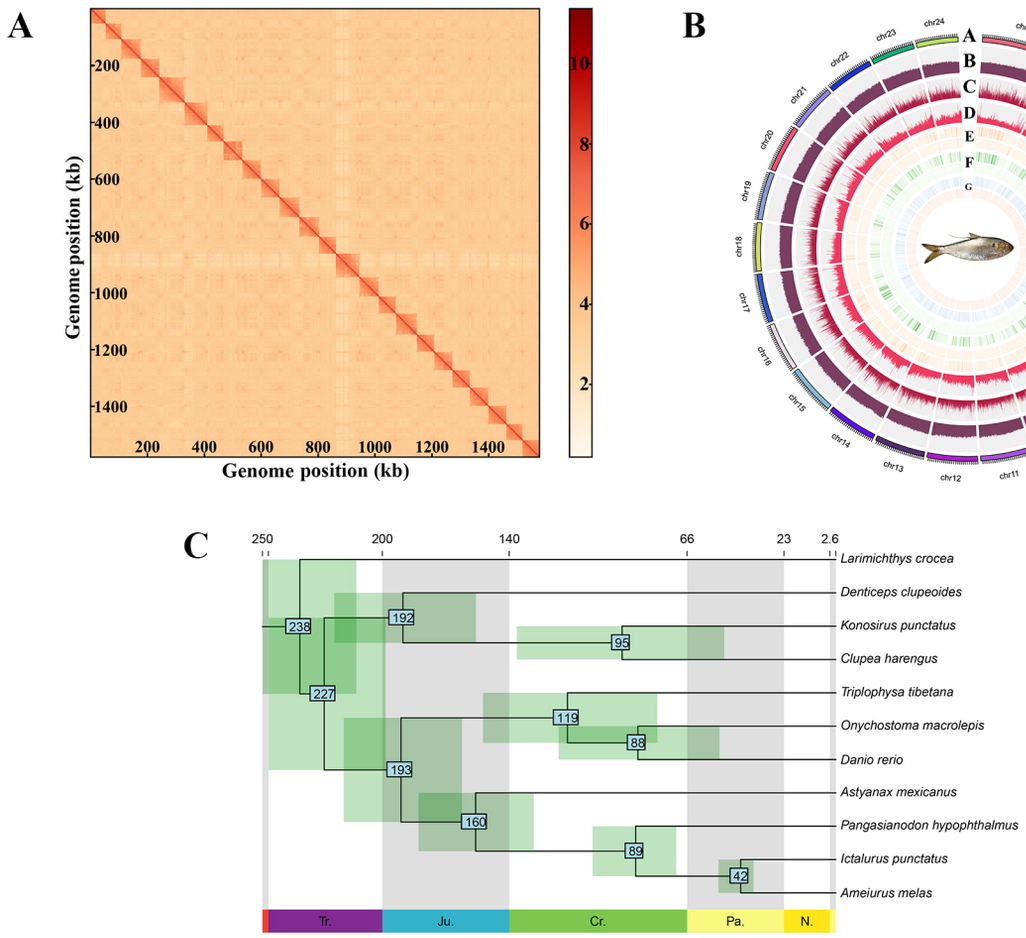

Figure 1 Genomic analyses of *K. punctatus*

A: Genome-wide Hi-C heatmap of *K. punctatus*. B: Genome characteristics of *K. punctatus*: a) genomic information; b) GC content distribution; c) second-generation read depth distribution; d) third-generation read depth distribution; e) outer circle shows homozygous SNP distribution, inner circle shows heterozygous SNP distribution; f) outer circle shows homozygous InDel distribution, inner circle shows heterozygous InDel distribution; g) outer ring is single-copy BUSCO, inner ring shows duplicated BUSCOs. C: Estimated divergence time of 11 species.

genome (Supplementary Table S13).

Orthologous groups were constructed using ORTHOMCL v2.0.9 with default settings based on the filtered BLASTP results. Single-copy orthologous genes shared by all 11 species (i.e., *Larimichthys crocea*, *Clupea harengus*, *Denticeps clupeoides*, *Danio rerio*, *Astyanax mexicanus*, *Ictalurus punctatus*, *Pangasianodon hypophthalmus*, *Onychostoma macrolepis*, *Triplophysa tibetana*, and *Ameiurus melas*) were further aligned using MUSCLE v3.8.31. Based on comparative genomics, 21 276 gene families were identified, including 2 018 single-copy homologous gene families (Supplementary Table S14). In addition, 24 298 genes of *K. punctatus* were clustered into 16 782 gene families, including 1 409 unique gene families (Supplementary Table S14). jModelTest/ProTest was applied to select the optimal sequence substitution model. RAxML v8.2.12 was then applied to construct the phylogenetic tree of the 11 species using the maximum-likelihood (ML) approach. The MCMCTree tool in PAML v4.5 was used to calibrate the divergence dates for other nodes on the phylogenetic tree using single-copy orthologs obtained from the TimeTree database and seven reference divergence times. Results showed that *K. punctatus* and *Clupea harengus* were clustered together, and the divergence time between the two species was 95 million years ago (Ma) (Figure 1C).

Gene family expansion and contraction analyses were performed using statistical tests in CAFÉ v3.1. Based on the gene family groupings of the species, the branch-site model and likelihood ratio test (LRT) in CODEML in PAML v4.5 were used to estimate the non-synonymous to synonymous mutation (dN/dS) ratio. A total of 512 expanded gene families and 2 099 contracted gene families were identified in the *K. punctatus* genome compared to the most recent common ancestor (Supplementary Table S15). A total of 587 positively selected genes (PSGs) were identified in the *K. punctatus* genome (Supplementary Table S16). Several PSGs may play an important role in the adaptive evolution of *K. punctatus*. Thus, further studies are needed to determine the putative roles of gene-related functions in adaptive evolution in these expanded, contracted, and PSG families.

In this study, we assembled a high-quality chromosome-level genome of *K. punctatus*, only the second reference

genome in the family Clupeidae. This study provides valuable genomic data for further research on the molecular mechanisms underlying adaptation in broadly saline fish and the functional validation of candidate genes that contribute to environmental adaptation.

DATA AVAILABILITY

The whole genome project of *Konosirus punctatus* was deposited at NCBI/BioProject (PRJNA664835, PRJNA665107, PRJNA665552, PRJNA666237). The raw sequencing reads of DNA are available at SRA (Illumina raw reads: SRR12702103 and PacBio raw reads: SRR12827990), the raw sequencing reads of RNA are available at SRA (SRR12690112), and the raw sequencing reads of Hi-C are available at SRA (SRR12719226 and SRR12719225). The assembled genome was deposited at the National Genomic Data Center (https://bigd.big.ac.cn/gwh/) under accession No. GWHBFWL00000000. The genome data were deposited in Figshare (https://figshare.com/s/46cf39eaa8bca2f04344).

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

B.J..L, K.Z., S.F.Z., and Y.F.L. conceived and designed the research. B.J.L., K.Z., S.F.Z., Y.F.L., J.S.L., Y.P., X.J., Y.P.W., S.X.Z., L.G., L.Q.L., and Z.M.L. conducted the experiments, analyzed the data, and wrote the manuscript. All authors read and approved the final version of the manuscript.

Bing-Jian Liu[1,2,3], Kun Zhang[1,3], Shu-Fei Zhang[2], Yi-Fan Liu[1,3], Jia-Sheng Li[1,3], Ying Peng[1,3], Xun Jin[1,3], Yun-Peng Wang[1,3], Si-Xu Zheng[1,3], Li Gong[1,3], Li-Qin Liu[1,3], Zhen-Ming Lü[1,3,*]

[1] *National Engineering Laboratory of Marine Germplasm Resources Exploration and Utilization*, *Zhejiang Ocean University*, *Zhoushan*, *Zhejiang* 316022, *China*

[2] *Guangdong Provincial Key Laboratory of Fishery Ecology and Environment*; *South China Sea Fisheries Research Institute*, *Chinese Academy of Fisheries Sciences*, *Guangzhou*, *Guangdong* 510300, *China*

[3] *National Engineering Research Center for Facilitated Marine Aquaculture*, *Marine Science and Technology College*, *Zhejiang Ocean University*, *Zhoushan*, *Zhejiang* 316022, *China*

*Corresponding author, E-mail: nblzmnb@163.com

REFERENCES

Gao YJ, Lü ZB, Yang YY, Wang YH, Ren ZH, Cong XR. 2016. Structure and species diversity of ichthyoplankton in spring in Laizhou Bay. *Acta Ecologica Sinica*, **36**(20): 6565−6573. (in Chinese)

Gwak WS, Lee YD, Nakayama K. 2015. Population structure and sequence divergence in the mitochondrial DNA control region of gizzard shad *Konosirus punctatus* in Korea and Japan. *Ichthyological Research*, **62**(3): 379−385.

Kong LB, Kawasaki M, Kuroda K, Kohno H, Fujita K. 2004. Spawning characteristics of the konoshiro gizzard shad in Tokyo and Sagami Bays, central Japan. *Fisheries Science*, **70**(1): 116−122.

Kuroda K, Kong L, Kawasaki M, Fujita K. 2002. Long-term fluctuations in the catch data of konoshiro gizzard shad, *Konosirus punctatus*, around Japan. *Bulletin of the Japanese Society of Fisheries Oceanography*, **66**(4): 239−246.

Li M, Xu BD, Ma QY, Zhang CL, Ren YP, Wan R, et al. 2017. Generalized additive model reveals effects of spatiotemporal and environmental factors on the relative abundance distribution of *Konosirus punctatus* in the Yellow River estuary and its adjacent waters. *Journal of Fishery Sciences of China*, **24**(5): 963−969. (in Chinese)

Liu BJ, Zhang K, Zhu KH, Shafi M, Gong L, Jiang LH, et al. 2020. Population genetics of *Konosirus punctatus* in Chinese coastal waters inferred from two mtDNA genes (COI and Cytb). *Frontiers in Marine Science*, **7**: 534.

McCay DPF, Whittier N, Ward M, Santos C. 2006. Spill hazard evaluation for chemicals shipped in bulk using modeling. *Environmental Modelling & Software*, **21**(2): 156−169.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *In*: Kollmar M. Gene Prediction. Methods in Molecular Biology, vol 1962. New York: Humana, 227–245.

Shan LZ, Ma JZ, Shao XB, Zhang LN, Yan MC, Fang J, et al. 2020. Study on the technique of artificial propagation and larva nursery of *Clupanodon punctatus* in Yue Qing Bay. *Fisheries Science & Technology Information*, **47**(3): 130−134. (in Chinese)

Song N, Gao TX, Ying YP, Yanagimoto T, Han ZQ. 2017. Is the Kuroshio Current a strong barrier for the dispersal of the gizzard shad (*Konosirus punctatus*) in the East China Sea?. *Marine and Freshwater Research*, **68**(5): 810−820.