Chromosome-Level Genome Assembly and Annotation of a Sciaenid Fish, *Argyrosomus japonicus*

Linlin Zhao[1], Shengyong Xu[2], Zhiqiang Han[2], Qi Liu[3], Wensi Ke[3], An Liu[3], and Tianxiang Gao[2,*]

[1]First Institute of Oceanography, Ministry of Natural Resources, Qingdao, Shandong, China

[2]Fishery College, Zhejiang Ocean University, Zhoushan, Zhejiang, China

[3]Wuhan Gooalgene Technology Co., Ltd., Wuhan, Hubei, China

*Corresponding author: E-mail: gaotianxiang0611@163.com.

Accepted: 17 November 2020

Abstract

Argyrosomus japonicus is an economically and ecologically important fish species in the family Sciaenidae with a wide distribution in the world's oceans. Here, we report a high-quality, chromosome-level genome assembly of *A. japonicus* based on PacBio and Hi-C sequencing technology. A 673.7-Mb genome containing 282 contigs with an N50 length of 18.4 Mb was obtained based on PacBio long reads. These contigs were further ordered and clustered into 24 chromosome groups based on Hi-C data. In addition, a total of 217.2 Mb (32.24% of the assembled genome) of sequences were identified as repeat elements, and 23,730 protein-coding genes were predicted based on multiple approaches. More than 97% of BUSCO genes were identified in the *A. japonicus* genome. The high-quality genome assembled in this work not only provides a valuable genomic resource for future population genetics, conservation biology and selective breeding studies of *A. japonicus* but also lays a solid foundation for the study of Sciaenidae evolution.

Key words: *Argyrosomus japonicus*, PacBio sequencing, de novo assembly, genome annotation, phylogenetic structure.

Significance

Limited genetic and genomic information for marine species restricts breeding development and resources conservation. In this study, we obtained a chromosome-level genome assembly of *Argyrosomus japonicus*, which will contribute to research on the genomics, evolution, and conservation biology of this species.

Introduction

The family Sciaenidae is one of the largest groups in the order Perciformes, comprising 66 genera with 291 species worldwide and 14 genera with 37 species in China (Zhu et al. 1963; Nelson 2006). *Argyrosomus japonicus* is a large-bodied fish in Sciaenidae that is widely distributed in estuaries and near-shore coastal waters (<100 m depth) of the Pacific and Indian Oceans surrounding Australia, South Africa, India, Pakistan, China, Korea, and Japan (Nakabo 2013). *Argyrosomus japonicus* was usually identified as *Nibea japonica* in previous studies in China, but this has been proven to be an invalid species name (synonym of *A. japonicus*) (Shen and Wu 1993; Xu 2010). The biology of *A. japonicus* has been well studied in Australia and South Africa, but little information is available from other areas of its distributional range (Lo et al. 2015). Due to the long-distance geographical isolation and different life histories of its populations, the population structure of *A. japonicus* might differ among regions.

In recent decades, overfishing and habitat degradation have led to a dramatic decrease in the population resources of *A. japonicus* in the wild ocean (Silberschneider et al. 2009). To cope with the declining wild stocks and growing seafood demand, the aquaculture of *A. japonicus* has been initiated in various areas worldwide (Wei et al. 2012). Due to its characteristics of disease resistance and rapid growth, this species is regarded as an important mariculture species in China, South Africa, and Australia (Bolton et al. 2013). Increased productivity achieved via genetic improvement with selective breeding has been a key factor facilitating the development of major aquaculture industries. To date, genomes from six species of sciaenid fishes have been sequenced including *Larimichthys crocea*, *Larimichthys polyactis*, *Collichthys lucidus*, *Miichthys miiuy*, *Sciaenops ocellatus*, and *Nibea albiflora* (detailed information in supplementary table S1, Supplementary Material online). However, the limited genetic information of *A. japonicus* restricts its further breeding and germplasm conservation. Therefore, there is an urgent need to obtain high-quality chromosome-level genome resources for *A. japonicus* to facilitate its selective breeding and reveal its phylogenetic relationships within its distribution region.

In the present study, we constructed a chromosome-level genome assembly of *A. japonicus* by combining Illumina short reads, PacBio long reads, and Hi-C sequencing data. We expected that this chromosome-level genome will promote studies on the selective breeding and population genetics of *A. japonicas*.

Materials and Methods

Sample Collection and Sequencing

A male individual of *A. japonicus* was sampled from a breeding farm in Zhoushan City, Zhejiang Province, for genome sequencing. Fresh muscle, eye, skin, gonad, gut, kidney, liver, brain, and blood samples were collected and quickly frozen in liquid nitrogen for 1 h before storage at −80 °C. Muscle tissue was used for DNA sequencing, whereas all tissues were used for transcriptome sequencing.

Total genomic DNA was extracted from fresh muscle using the standard phenol/chloroform method and sequenced using the PacBio Sequel II platform (for genome assembly) and the Illumina NovaSeq platform (for genome surveying and base correction after assembly). To obtain a chromosome-level genome, a 0.2-ml blood sample from the same individual was used for Hi-C library construction and sequencing with the same method used in a previous study (Gong et al. 2018). RNA was extracted from different tissues using the TRIzol Reagent (Invitrogen), then mixed in equal amounts and subjected to RNA-seq using the Illumina NovaSeq platform.

Genome Assembly and Assessment

Before genome assembly, the size, heterozygosity, and repeat content of the *A. japonicus* genome were estimated through k-mer analysis with jellyfish (Marcais and Kingsford 2011).

To assemble the genome of *A. japonicus*, we applied Canu to analyze PacBio long reads (Koren et al. 2017). To correct random sequencing errors in the assembled genome, two steps of genome sequence polishing were applied: We first used the Quiver algorithm to polish the genome using PacBio long reads (Chin et al. 2013), and another round of genome-wide base-level correction was performed using Pilon with the Illumina clean reads (Walker et al. 2014). For chromosome-level scaffolding, the cleaned Hi-C reads were mapped to the assembled genome using BWA, and only uniquely mapped read pairs were considered for subsequent analysis (Li and Durbin 2009). We then used LACHESIS to cluster, order, and orient the assembled contigs (Burton et al. 2013). To evaluate the quality of the assembled genome, its completeness and accuracy were assessed via sequenced read mapping and benchmarking universal single-copy ortholog (BUSCO) analysis (Seppey et al. 2019).

The Annotation of Repetitive Elements

Repeat sequences were identified in the *A. japonicus* genome via a combination of homology-based and de novo approaches. First, we used Tandem Repeats Finder (TRF) to detect tandem repeats and RepeatModeler to detect repeat sequences in the assembled genome (Benson 1999; Tarailo and Chen 2009). Based on the Repbase library, we used RepeatMasker and RepeatProteinMasker to annotate repeat elements and TE proteins, respectively (Bao et al. 2015).

Gene Prediction and Functional Annotation

For gene structure prediction, we used a combination of de novo, homology-based and transcriptome-based strategies to predict genes in the *A. japonicus* genome. The de novo approach was implemented using Augustus (Stanke et al. 2008). For homology-based prediction, TBlastN was used to align protein sequences from big head croker (*C. lucidus* from Sciaenidae), larger yellow croaker (*L. crocea* from Sciaenidae), ocellaris Clownfish (*Amphiprion ocellaris* from Pomacentridae), spiny chromis (*Acanthochromis polyacanthus* from Pomacentridae), eastern happy (*Astatotilapia calliptera* from Cichlidae), and Climbing perch (*Anabas testudineus* from Anabantidae) to the assembled genome of *A. japonicus* (Gertz et al. 2006). Then, the transcriptome sequence reads were aligned to the genome using the TopHat package, and gene structure was predicted using Cufflinks (Trapnell et al. 2009; Ghosh and Chan 2016). Finally, all gene models were merged, and redundancy was removed wit MAKER (Cantarel et al. 2007).

Phylogenetic Analysis of *A. japonicus*

To reveal the phylogenetic relationships with other species, we downloaded the protein sequences of *C. lucidus* (GCA_004119915.1), *L. crocea* (GCF_000972845.2), *Takifugu rubripes* (GCF_901000725.2), *Oryzias latipes* (GCF_002234675.1), *Lates calcarifer* (GCF_001640805.1), *Gasterosteus aculeatus* (GCA_006229165.1), *Dicentrarchus labrax* (GCA_000689215.1), and *Danio rerio* (GCF_000002035.6) from the NCBI database, and *Sillago sinica* (http://dx.doi.org/10.5524/100490) from GigaScience database. The proteome sets of all species were analyzed

Table 1
Gene Structure Prediction of the *Argyrosomus japonicus* Genome

Gene Set		Number	Average Gene Length (bp)	Average CDS Length (bp)	Average Exons Per Gene	Average Exon Length (bp)	Average Intron Length (bp)
De novo prediction	AUGUSTUS	28,447	11,628.52	1,434.27	8.16	175.77	1,423.80
Homologue prediction	*Collichthys lucidus*	63,173	7,412.39	883.72	4.09	216.31	2,115.92
	Larimichthys crocea	51,693	7,396.15	1,077.62	5.27	204.66	1,481.33
	Amphiprion ocellaris	48,436	7,640.05	1,056.38	5.36	196.95	1,508.73
	Acanthochromis polyacanthus	50,682	7,140.94	993.37	5.12	194.15	1,493.39
	Astatotilapia calliptera	50,080	8,179.76	1,107.74	5.28	209.8	1,652.34
	Anabas testudineus	50,589	8,103.9	1,063.8	5.32	199.79	1,627.96
RNA-seq	Cufflinks	14,679	16,664.61	1,895.99	11.91	310.89	1,188.27
Final	MAKER	23,730	15,378.24	1,664.78	9.81	269.66	1,445.04

with OrthoMCL to construct different types of orthologues (Li 2003). These single-copy orthologues were aligned using MUSCLE, and phylogenetic trees were generated with RAxML with 500 bootstrap replicates (Edgar 2004; Stamatakis 2014). MCMCTREE was applied to obtain estimates according to the divergence time-based approximate likelihood calculation method using molecular clock data from the TimeTree database (http://www.timetree.org/), including data for *L. crocea* and *D. labrax* (87–105 Ma), *L. crocea* and *L. calcarifer* (94–115 Ma), *L. crocea* and *O. latipes* (104–145 Ma) (Yang 1997; Hedges et al. 2006).

Results and Discussion

Genome Assembly and Assessment

The details of the sequencing data obtained in the present study are listed in supplementary table S2, Supplementary Material online. The 17-mer frequency of short reads followed a Poisson distribution, with the highest peak occurring at a depth of 93 (supplementary fig. S1, Supplementary Material online). The estimated genome size was ~675 Mb, the heterozygosity rate of the genome was 0.21%, and the repeat content of the genome was 35.47%.

Using Canu, we obtained an *A. japonicus* genome of 791 Mb with 1,984 contigs and a contig N50 of 13.1 Mb. After correcting the random sequencing errors in the assembled genome, the genome assembly of *A. japonicus* contained 674 Mb of sequences within 282 polished contigs, with a contig N50 of 18.4 Mb, and the overall GC content was 41.20%. Furthermore, 271 assembled contigs (98.8% of genome) were successfully clustered into 24 chromosome groups. Finally, we obtained a high-quality chromosomal-level genome with a total size of 674 Mb (supplementary table S3, Supplementary Material online), and the contig N50 and scaffold N50 lengths were 18.4 and 29.4 Mb, respectively, which were longer than most of other fish in Sciaenidae. Furthermore, 98.22% of short reads were mapped to the assembled genome, which covered 99.95% of the assembly. Then, the PacBio long reads were mapped with BLASR, and 95.57% of long reads covered 99.86% of the assembled genome. Finally, BUSCO was also used to evaluate the completeness of the genome assembly; 97.77% of the "complete BUSCOs" were successfully identified in the assembly, and the proportion of "missing BUSCOs" was only 1.4%. These results demonstrate the high reliability and completeness of the reported genome assembly.

Genome Repetitive Elements and Gene Prediction

After removing redundancies, a total of 217.2 Mb of sequences (32.24% of the *A. japonicus* genome) were identified as repeat elements. Among these repeat elements, DNA transposons were the main type, accounting for 16.24% (100.43 Mb) of the repeat elements. A total of 23,730 protein-coding genes were predicted in the present study (table 1). Among all protein-coding genes, 22,938 protein-coding genes, corresponding to 97.34% of the total predicted genes in the *A. japonicus* genome, were functionally annotated in at least one public database (supplementary table S4, Supplementary Material online).

Phylogenetic Relationships of *A. japonicus*

Using OrthoMCL, we identified a set of 2,502 single-copy orthologues (fig. 1*a*). Then, the data matrix was applied to construct a phylogenetic tree (supplementary fig. S2,

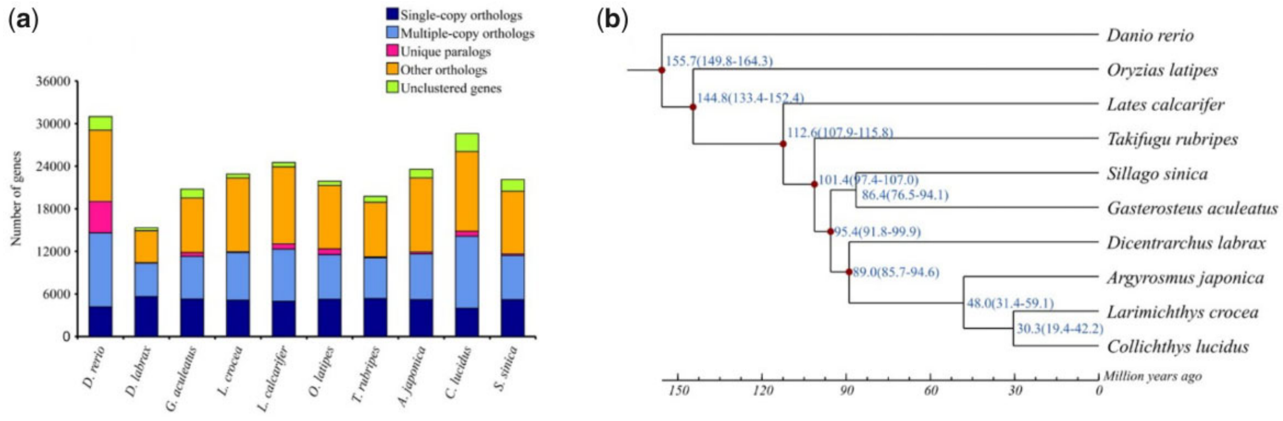

FIG. 1.—(*a*) Distribution of different types of orthologues in the selected representative. (*b*) Time tree and estimated divergence times of representative species based on single-copy orthologues.

Supplementary Material online) and estimate the divergence time (fig. 1*b*). *Argyrosomus japonicus* diverged from the common ancestor with *C. lucidus* and *L. crocea* ~48 Ma, and the divergence time between Sciaenidae and other teleosts (89.0 Ma) was similar to that reported in a previous study (~90 Ma; Betancur et al. 2017).

Conclusions

In this work, we applied a combined strategy involving Illumina, PacBio and Hi-C technologies for the de novo assembly of a chromosome-level genome for *A. japonicus*. We assembled the genome sequences into 282 contigs with a total length of 673.7 Mb and a contig N50 length of 18.4 Mb. By using Hi-C data, the contigs were further ordered and clustered into 24 chromosomes with a total length of 665.4 Mb. This well-annotated chromosome-level whole-genome sequence should be a valuable resource for studies on the genomics, evolution, and conservation biology of *A. japonicus*.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (41706187, 41776171) and the National Key R&D Program of China (2019YFD0901301).

Data Availability

The sequencing data, including the Illumina genome data (SRA accession: SRR11461811), PacBio data (SRA accession: SRR11566291), Hi-C data (SRA accession: SRR11566289/SRR11566290), and Illumina transcriptome data (SRA accession: SRR11349394), were submitted to the NCBI Sequence Read Archive (SRA) database. The final chromosome assembly and genome annotation were submitted to NCBI database with accession number: SAMN16453595. The functional annotation files are also available at figshare (https://doi.org/10.6084/m9.figshare.13076369.v1).

Literature Cited

Bao W, et al. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580.

Betancur RR, et al. 2017. Phylogenetic classification of bony fishes. BMC Evol Biol. 17(1):162.

Bolton JJ, et al. 2013. Innovative processes and products involving marine organisms in South Africa. Afr J Mar Sci. 35(3):449–464.

Burton JN, et al. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 31(12):1119–1125.

Cantarel BL, et al. 2007. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18(1):188–196.

Chin CS, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 10(6):563–569.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Gertz EM, et al. 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 4(1):41.

Ghosh S, Chan CK. 2016. Analysis of RNA-Seq data using TopHat and Cufflinks. Methods Mol Biol. 1374:339–361.

Gong G, et al. 2018. Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. Gigaence 11:giy120.doi:10.1093/gigascience/giy120.

Hedges SB, et al. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22(23):2971–2972.

Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760.

Li L. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.

Lo PC, et al. 2015. A multi-gene dataset reveals a tropical new world origin and early Miocene diversification of croakers (Perciformes: Sciaenidae). Mol Phylogenet Evol. 88:132–143.

Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6): 764–770.

Nakabo T. 2013. Fish of Japan with pictorial keys to the species. Tokyo, Japan: Tokai University Press. p. 969–973.

Nelson JS. 2006. Fishes of the world. 4th ed. New Jersey: John Wiley & Sons. doi:10.1111/j.1467-2979.2006.00227.x.

Seppey M, et al. 2019. BUSCO: assessing genome assembly and annotation completeness. Meth Mol Biol. 1962:227–245.

Shen SC, Wu GY. 1993. Fishes of Taiwan. Taiwan: National Taiwan University. p. 383–392.

Silberschneider V, et al. 2009. Age, growth, maturity and the overfishing of the iconic sciaenid, *Argyrosomus japonicus*, in south-eastern, Australia. Fish Res. 95(2–3):220–229.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9): 1312–1313.

Stanke M, et al. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 24(5):637–644.

Tarailo M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 25(1):4.10.1–4.10.14.

Trapnell C, et al. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.

Wei MX, et al. 2012. Research progress on culture biology of Japanese croaker *Nibea japonica*. Fish Sci. 4:245–248.

Xu JY. 2010. Progress on research of *Nibea japonica*. Mod Fish Infor. 5:6–8.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555–556.

Zhu YD, et al. 1963. A study on the classification of the sciaenoid fishes of China, with description of new genera and species. Shanghai: Shanghai Science and Technology Press. p. 6–18.

Associate editor: B. TE Venkatesh