Chromosomal-Level Genome Assembly of Silver Sillago (*Sillago sihama*)

Xinghua Lin[1,2,3,4,], Yang Huang[1,2,3,4,5], Dongneng Jiang[1,2,3,4,5], Huapu Chen[1,2,3,4,5], Siping Deng[1,2,3,4,5], Yulei Zhang[1,3,4,5], Tao Du[1,2,3,4,5], Chunhua Zhu[1,2,3,4,5], Guangli Li[1,2,3,4,5,]*, and Changxu Tian[1,2,3,4,5,]*

[1]Fisheries College, Guangdong Ocean University, Zhanjiang, China

[2]Guangdong Research Center on Reproductive Control and Breeding Technology of Indigenous Valuable Fish Species, Fisheries College, Guangdong Ocean University, Zhanjiang, China

[3]Guangdong Provincial Engineering Laboratory for Mariculture Organism Breeding, Fisheries College, Guangdong Ocean University, Zhanjiang, China

[4]Guangdong Provincial Key Laboratory of Pathogenic Biology and Epidemiology for Aquatic Economic Animals, Fisheries College, Guangdong Ocean University, Zhanjiang, China

[5]Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang), Zhanjiang, China

*Corresponding authors: E-mails: tiancx@gdou.edu.cn; ligl@gdou.edu.cn.

Accepted: 22 December 2020

Abstract

Silver sillago, *Sillago sihama* is a member of the family Sillaginidae and found in all Chinese inshore waters. It is an emerging commercial marine aquaculture species in China. In this study, high-quality chromosome-level reference genome of *S. sihama* was first constructed using PacBio Sequel sequencing and high-throughput chromosome conformation capture (Hi-C) technique. A total of 66.16 Gb clean reads were generated by PacBio sequencing platforms. The genome-scale was 521.63 Mb with 556 contigs, and 13.54 Mb of contig N50 length. Additionally, Hi-C scaffolding of the genome resulted in 24 chromosomes containing 96.93% of the total assembled sequences. A total of 23,959 protein-coding genes were predicted in the genome, and 96.51% of the genes were functionally annotated in public databases. A total of 71.86 Mb repetitive elements were detected, accounting for 13.78% of the genome. The phylogenetic relationships of silver sillago with other teleosts showed that silver sillago was separated from the common ancestor of *Sillago sinica* ~7.92 Ma. Comparative genomic analysis of silver sillago with other teleosts showed that 45 unique and 100 expansion gene families were identified in silver sillago. In this study, the genomic resources provide valuable reference genomes for functional genomics research of silver sillago.

Key words: silver sillago, chromosomal assembly, genome, PacBio, Hi-C.

Significance

Sillago sihama is a commercial marine fish species with an important economic value in China. A high-quality chromosome-level reference genome of *S. sihama* was constructed in this study. The genome is an important resource for advancing research on physiology, reproduction, and breeding research.

Introduction

Sillaginidae family (also known as smelt-whitings or sand borers) belongs to order Perciformes, are bottom-dwelling fishes and widely distributed in the shallow sea regions of Indo-West-Pacific Ocean (Xu et al. 2018). Sillaginidae consists of 31 species in three genera and three subgenera, of which the genus *Sillago* comprises 24 species. *Sillago* species drill sand to avoid seine-net and other environmental hazards (Lou et al. 2020). *Sillago* flesh is white and very tender, with excellent flavor. Steamed whiting fillet of *Sillago* fishes contains little fat content, which is easy to digest. Due to its ecological and economic importance, the inshore fishing of *Sillago* has developed rapidly in the past decades. However, the natural population of *Sillago* spp. has reduced in recent years due to overfishing and demersal environmental deterioration, such as localized oxygen depletion, sulfide accumulation, and high turbidity (Lou et al. 2020). Therefore, it is necessary to develop genomic resources to protect their natural resources and to accelerate the process of genome-assisted improvement of important economic traits.

Silver sillago, *Sillago sihama* is found in all Chinese waters, including beaches, sandbars, mangrove creeks, and estuaries (Guo et al. 2014). This fish species has been widely cultured in China due to its high meat quality. However, the reduction of natural population of *S. sihama* and a low survival rate in artificial breeding decrease the development of the marine aquaculture of *S. sihama*. To date, complete mitogenome (Siyal et al. 2016), simple sequence repeat (Guo et al. 2014; Qiu et al. 2020), transcriptome (Tian et al. 2019; Saetan et al. 2020), and draft genomic survey data (Li et al. 2019) have been reported for *S. sihama*.

The genome of *S. sinica* was the first and only reference genome for Sillaginidae (Lou et al. 2020). However, large-scale genomic analysis at the chromosome level has not been well-characterized in *Sillago* due to the fragmented assemblies. Our study reported the chromosome-level genome of *Sillago*, which is the first chromosome-level genome of *S. sihama*. Genomic and comparative genomic analyses provide insights into the genes related to environmental stress. The genome can be used as a basis for the research on the evolution and biology of *S. sihama*.

Materials and Methods

Ethics Statement

All experimental protocols were approved by the Animal Research and Ethics Committees of the Institute of Aquatic Economic Animals of Guangdong Ocean University, Zhanjiang, Guangdong, China (201903003). The study does not involve endangered or protected species.

Sample Collection and Sequencing

Sillago sihama (length of 19.3 cm) was obtained from Donghai Island, Guangdong, China. Genomic DNA (gDNA) was extracted from muscle samples and constructed two Pacific Biosciences (PacBio) sequencing libraries (insert size of 20 kb). DNA samples were interrupted by g-TUBE, and the adaptor was connected to the DNA. The libraries were purified by an exonuclease, and the sequencing fragments were screened by BluePippin. Sequencing was conducted using the PacBio platform. Adaptors, low-quality reads and short fragments were filtered to obtain high-quality subreads.

The high-throughput chromosome conformation capture (Hi-C) library (insert size of 350 bp) was constructed for sequencing to obtain the chromosome-level assembly of the genome. The samples were fixed by formaldehyde, and restriction enzyme was added to digest DNA, followed by repairing the 5′-end by biotin residues. Sequencing was done using the Illumina platform. Adapter sequences of raw reads were trimmed, and low-quality paired-end (PE) reads were removed to get clean data.

RNA was extracted from eight tissues, including liver, heart, head kidney, gonad, muscle, brain, stomach, and gill of *S. sihama*. Illumina HiSeq platform was used for transcriptome sequencing.

Genome Assembly

The filtered data were corrected by Canu (Koren et al. 2017), and then the corrected data were used to assemble the primary genome by WTDBG. After completing the primary assembly, the chromosomal-level genome was assembled from HI-C data. The clean data were compared with preliminary assembly results by Burrows–Wheeler Aligner (Li and Durbin 2009). HiC-Pro (Rusk 2014) was used to filter and evaluate the quality of Hi-C data. The genome sequence was divided into groups, and then sorted and oriented. The assembly results were evaluated by LACHESIS (Servant et al. 2015).

Genome Prediction and Annotation

Based on structural prediction and de novode novo, a repetitive sequence database of *S. sihama* genome was constructed by LTR FINDER v1.05 (Xu and Wang 2007), RepeatScout v1.0.5 (Price et al. 2005), and PILER-DF v2.4 (Edgar and Myers 2005). PASTEClassifier (Wicker et al. 2007) was used to classify the repetitive sequence database and then merged with the Repbase (Jurka et al. 2005) database as the final repetitive sequence database. The repetitive sequence of *S. sihama* was predicted by RepeatMasker v4.0.6 (Tarailo-Graovac and Chen 2009).

Based on ab initioab initio, homologous alignment and transcriptome data were used to predict protein-coding genes in the genome. The ab initioab initio prediction was done using Genscan (Burge and Karlin 1997), Augustus v2.4 (Stanke and Waack 2003), GlimmerHMM v3.0.4 (Majoros et al. 2004), GeneID v1.4 (Alioto et al. 2018), and Supplemental Nutrition Assistance Program (SNAP) (version 2006-07-28) (Korf 2004). The protein sequences of *Larimichthys crocea*, *Oreochromis niloticus*, *Oryzias latipes*, *Danio rerio*, and *Sillago sinica* were downloaded from the National Center for Biotechnology Information (NCBI) and GIGA databases. The homologous alignment was constructed using GeMoMa v1.3.1 (Keilwagen et al. 2016) to predict protein-coding genes. The reference transcripts were assembled by Hisat v2.0.4, Stringtie v1.2.3 (Pertea et al. 2016), TransDecoder v2.0 (Haas et al. 2013), and GeneMarkS-T v5.1 (Tang et al. 2015) were used for gene prediction. Based on transcriptome data, unigene sequences were predicted by PASA v2.0.2 (Campbell et al. 2006). EVM v1.1.1 (Haas et al. 2008) was used to integrate the prediction results obtained by the above three methods.

We performed homology searches in public gene databases, including NCBI Refseq (NR, Marchler-Bauer et al. 2011), Kyoto Encyclopedia of Genes and Genomes (KEGG, Ogata et al. 1999), Clusters of orthologous groups for eukaryotic complete genomes (KOG, Tatusov 2001), Translation of EMBL nucleotide sequence database (TrEMBL, Boeckmann 2003) and Gene Ontology (GO, Dimmer et al. 2012). Function annotation was performed on the predicted gene sequences by BLAST v2.2.31 (Altschul et al. 1990) (-evalue 1e-5). Based on the comparison results of the NR database, the functional annotation of the GO database was performed by Blast2GO (Conesa et al. 2005).

The rRNA and microRNA sequences were predicted by Infenal 1.1 (Nawrocki and Eddy 2013) on the Rfam (Griffiths-Jones et al. 2005) and miRBase (Griffiths-Jones et al. 2006) databases. The tRNA was identified by tRNAscan-SE v1.3.1 (Lowe and Eddy 1997).

Assessment of Completeness of the Genome Assembly

The core eukaryotic gene mapping approach was used to assess the completeness of assembly and gene annotation (CEGMA, v2.5) (http://korflab.ucdavis.edu/Datasets/cegma/, last accessed January 13, 2021) (Parra et al. 2007) and benchmarking universal single-copy orthologs (BUSCO, v2) (http://busco.ezlab.org/, last accessed January 13, 2021) (Simao et al. 2015) were used.

Genome Evolution Analysis

Based on the protein sequences of the *S. sihama* and 10 other teleosts, including *Takifugu rubripes* (GCA_000180615.2), *Gasterosteus aculeatus* (GCA_006229165.1), *O. latipes* (GCA_004347445.1), *D. rerio* (GCA_000002035.4), *O. niloticus* (GCA_001858045.3), *Latimeria chalumnae* (GCF_000225785.1), *S. sinica* (http://dx.doi.org/10.5524/100490, last accessed January 13, 2021), *L. crocea* (GCA_003845795.1), *Lepisosteus oculatus* (GCA_000242695.1), and *Xiphophorus maculatus* (GCA_002775205.2). The evolution between species and the classification of gene families were analyzed. The protein sequences of 11 teleosts were classified into gene families, and single-copy genes were extracted by OrthoMCL (Li et al. 2003). In order to study the evolutionary relationship between 11 teleosts, the single-copy protein sequences of 11 teleosts were used to construct the maximum-likelihood (ML) phylogenetic tree by PHYML (Guindon et al. 2010). The divergence time was predicted by McMctree in PAML and timetree databases (http://www.timetree.org/, last accessed January 13, 2021) to correct divergence time. *L. crocea* was phylogenetically closely related to *S. sihama*. The 24 *S. shama* chromosomes were aligned with *L. crocea* chromosomes by MCScanX to visualize the consistency between the genomes of *S. sihama* and *L. crocea* (Wang et al. 2012).

Gene Family Expansion and Contraction Analysis

The expansion and contraction gene families among *T. rubripes*, *G. aculeatus*, *O. latipes*, *D. rerio*, *O. niloticus*, *L. chalumnae*, *S. sinica*, *L. crocea*, *L. oculatus*, *X. maculatus*, and *S. sihama* were identified by CAFÉ (De Bie et al. 2006). The number of gene families of each ancestor was estimated by the birth mortality model, thereby predicting the number of gene family expansion and contraction gene families.

Results and Discussion

Genome Sequencing and Assembly

After quality filtering, 66.16 Gb subread data were obtained from two long-insert (20 kb) libraries (sequence coverage: ~126×; subread N50: 15,715 bp; supplementary table S1, Supplementary Material online). A total of 89.08 Gb Hi-C data were obtained from the HI-C sequencing library (sequence coverage: ~170×; GC content: 43.95%; Q30: 90.92%; supplementary table S1, Supplementary Material online).

The PacBio data were used to construct the primary assembly. The primary genome assembly size was 522.06 Mb, and contig N50 was 13.55 Mb. The efficiency of comparing HI-C sequence data with the primary assembled genome was 90.79% (Unique Mapped Read Pair was 77.18%). Total effective Hi-C data were 153.18 Mb. Re-assemble after correcting the errors of the primary assembled genome by Hi-C data. The chromosome-level genome size was 521.63 Mb, and contig N50 was 13.54 Mb (table 1). Using Hi-C data, 556 contigs were mapped to 24 chromosomes (supplementary fig. S1, Supplementary Material online). A total length of 498.82 Mb of the genomic sequence was anchored to 24 chromosomes, accounting for 96.93% of the entire genomic sequence (supplementary table S2 and fig. S2, Supplementary Material online).

According to BUSCO results, the genome contained 4,463 (97.36%) complete BUSCOs, including 4,345 single-copy BUSCOs and 118 duplicated BUSCOs (table 1). The CEGMA v2.5 database contained 248 conserved core genes of eukaryotes, and there were 246 conserved core genes (99.19%) in this genome (table 1). The results indicated that the genome assembly had high coverage and completeness.

Genome Annotation

De novoDe novo prediction and Repbase database results showed that the repeated sequences accounted for 13.78% of *S. sihama* genome, which is lower than *D. rerio*

Table 1
Statistics of *Sillago sihama* Genome Assembly and Annotation Data

	Chromosome-Level Genome Assembly
Assembly	
Assembly size (bp)	521,631,495
Number of scaffolds	470
Scaffold N50 (bp)	21,469,626
Longest scaffold (bp)	28,013,376
Number of contigs	556
Contig N50 (bp)	13,543,514
Longest contig max (bp)	22,111,180
GC (%)	44.66
BUSCO (% of total BUSCO)	
Complete	4,463 (97.36%)
Single-copy	4,345 (94.79%)
Duplicated	118 (2.57%)
Fragmented	27 (0.6%)
Missing	94 (2.05%)
CEGMA	
CEGs (% of all CEGs)	453 (98.97%)
Highly conserved CEGs (% of all highly conserved CEGs)	246 (99.16%)
Repetitive sequences (% of genome)	
SINE (bp)	60,396 (0.01%)
LINE (bp)	7,497,699 (1.44%)
LTR (bp)	6,955,457 (1.33%)
DNA (bp)	17,803,273 (3.00%)
SSR (bp)	101,169 (0.02%)
Unclassified (bp)	39,549,161 (7.58%)
Total (bp)	71,864,242 (13.78%)
Gene annotations (% of all genes)	
GO annotation	12,408 (51.79%)
KEGG annotation	14,510 (60.59%)
KOG annotation	15,991 (66.74%)
TrEMBL annotation	22,953 (95.8%)
NR annotation	23,101 (96.42%)
All annotated	23,123 (96.51%)
Noncoding protein genes (% of genome)	
Number of miRNA	419
Number of tRNA	1,587
Number of rRNA	67
Length of miRNA	34,211 (0.00656%)
Length of tRNA	160,051 (0.03068%)
Length of rRNA	60,018 (0.00575%)

(63.12%), *O. latipes* (42.83%), and *L. crocea* (20.31%), and higher than *S. sinica* (10.92%) and *T. rubripes* (9.37%). DNA transposons (3%) were the most common among transposons of *S. sihama* genome, followed by long interspersed repeated segments (LINEs, 1.44%) and long terminal repeats (LTR, 1.33%) (table 1, supplementary table S3, Supplementary Material online, fig. 1*A*).

A total of 23,959 protein-coding genes (supplementary table S4, Supplementary Material online) were predicted in the *S. sihama* genome by ab initio, homologous prediction and RNA-seq prediction methods, with an average length of 11,241.51 bp. Comparing the length distribution of genes, coding sequences (CDS), exons and introns, the gene distribution of *S. sihama* was similar to other teleosts. *Sillago sihama* gene proportions were lower than other fishes but similar to *S. sinica* (supplementary fig. S3, Supplementary Material online). The functions of the protein-coding genes were annotated in NR, TrEMBL, KOG, KEGG, and GO databases. A total of 23,123 genes were annotated, accounting for 96.5% of all protein-coding genes (table 1).

Rfam, miRBase, and tRNAscan-SE databases were used to predict noncoding RNA, and a total of 1,587 tRNAs, 67 rRNAs, and 419 miRNAs were predicted (table 1, supplementary table S5, Supplementary Material online).

Comparative Genome Analysis

The genomes of 11 teleosts were compared with study the phylogenetic relationships between *S. sihama* and other teleosts. A total of 16,856 gene families and 5,950 single-copy orthologs were identified (supplementary table S6 and fig. S4, Supplementary Material online). The ML phylogenetic tree was constructed from single-copy orthologs. The phylogenetic tree showed that *S. sinica* was closely related to *S. sihama*, and the divergence time was ∼7.92 (2.45–16.57) Ma (fig. 1*B*). The genomes of *S. sihama* and *L. crocea* were compared with analyze chromosomal evolutionary events (fig. 1*C*). The results showed that the 24 chromosomes of *S. sihama* were aligned with 22 chromosomes of *L. crocea*. The chromosomes III and XIII of *L. crocea* were compared with LG2, LG10, LG5, and LG16 of *S. sihama*, respectively. The common ancestor of *L. crocea* and *S. sihama* undergone a chromosome break recombination event during the evolution process, which increases the number of chromosomes.

Gene Family Analysis

The expansion and contraction of gene families are one of the most important factors for the evolution of phenotypic diversity and environmental adaptation. *S. sihama* is sensitive to environmental factors, such as sound, vibration, light, and shadow. In order to explore the adaptability of environmental factors in *S. sihama*, the gene families of 11 teleost fishes (*T. rubripes*, *G. aculeatus*, *O. latipes*, *D. rerio*, *O. niloticus*, *L. chalumnae*, *S. sinica*, *L. crocea*, *L. oculatus*, *X. maculatus*, and *S. sihama*) were compared. A total of 57 unique, 100 expanded ($P < 0.05$) and 25 contracted ($P < 0.05$) gene families were identified in *S. sihama* (supplementary table S7, Supplementary Material online), including immune-related gene families (immunoglobulin domain, immunoglobulin V-set domain, immunoglobulin I-set domain and NACHT domain) and olfactory receptor gene family (seven transmembrane receptor).

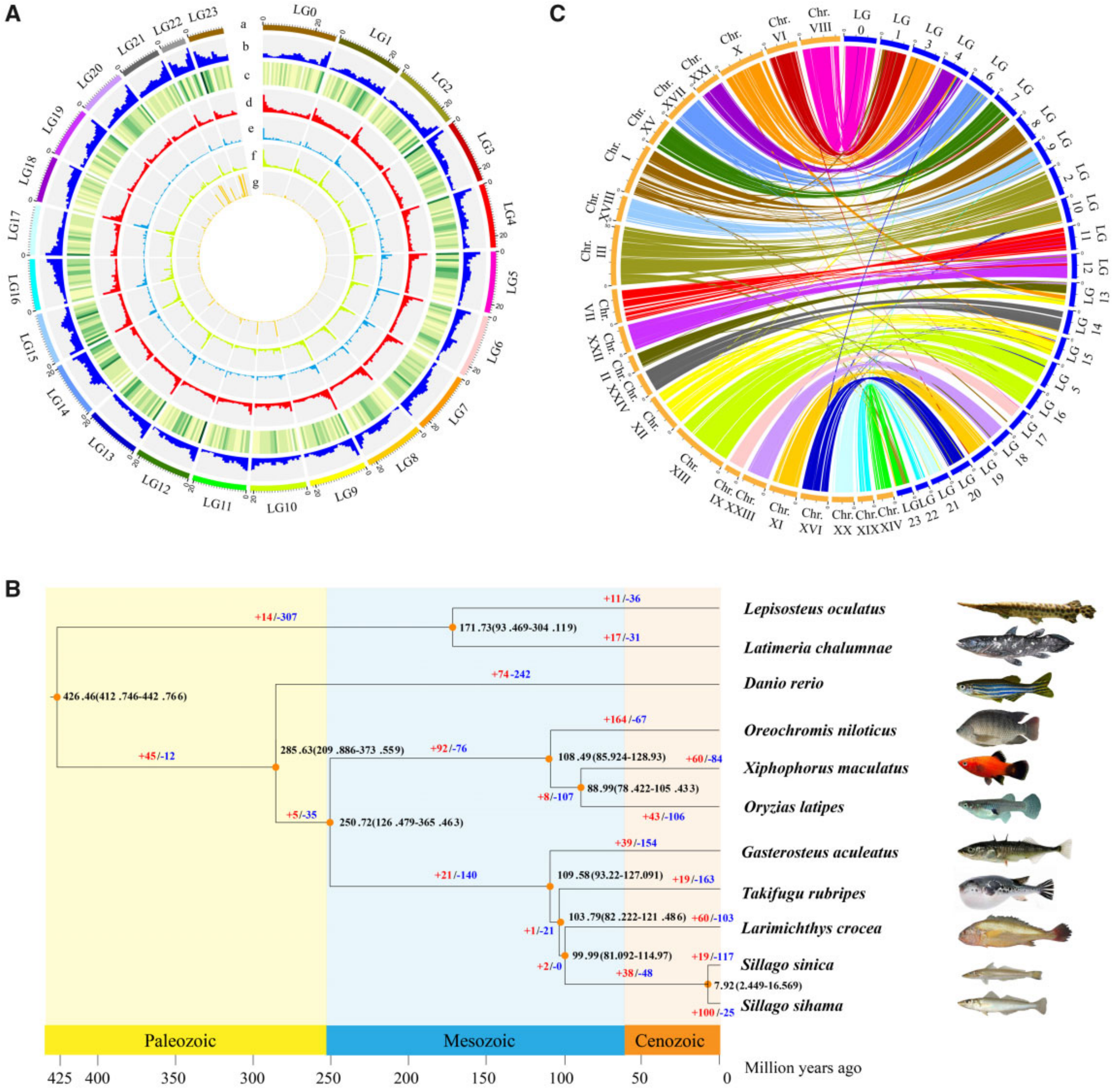

Fig. 1.—Genome landscape and evolutionary analysis of *Sillago sihama*. (*A*) Genome landscape of *S. sihama.* (*a*) Chromosome length, (*b*) GC content, (*c*) gene density, (*d*) repeat sequence, (*e*) long terminal repeated (LTE), (*f*) long interspersed nuclear elements (LINE), and (*g*) simple sequence repeat (SSR). (*B*) Phylogenetic analysis of 11 teleost fishes. At each branch point, the predicted species divergence time (million years ago) is marked. The red number on each evolutionary branch represents the number of expanding gene families, and the blue number represents the number of contracting gene families. (*C*) Collinearity analysis of *S. sihama* and *Larimichthys crocea* genomes. Blue and orange outer circles represent the chromosome of *S. sihama* and *L. crocea*, respectively.

Conclusions

This study was determined the chromosomal-level genome assembly of *S. sihama*. The continuity and completeness of the *S. sihama* genome was reached the level of other high-quality teleost fish genomes, which provides a useful reference for system biology and comparative genome evolution analysis. Genome evolution analysis showed the insights into the high irritability of *S. sihama*. This reference genome is important for aquaculture and artificial breeding of *S. sihama*, which provides a basis for further research.

Supplementary Material

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

Acknowledgments

This study was supported by grants from the National Natural Science Foundation of China (Nos. 41706174 and 31702326); Natural Science Foundation of Guangdong Province (2016A030313743, 2017A030313101, 2018B030311050, 2019A1515010958, 2019A1515110619); Independent Project of Guangdong Province Laboratory (ZJW-2019-06); the Department of Education of Guangdong Province (2018KQNCX111, 2019KTSCX060); and Program for Scientific Research Start-up Funds of Guangdong Ocean University (R19026).

Author Contributions

X.H.L. designed research, performed research, analyzed data, and wrote the paper. Y.H., D.N.J., H.P.C., S.P.D., Y.L.Z., T.D., and C.H.Z. collected the samples for sequencing and obtained funding. C.X.T. and G.L.L. obtained funding, conceived, and managed the project. All authors reviewed the manuscript.

Data Availability

The raw genome and RNA sequencing data have been submitted in the SRA under Bioproject number PRJNA642704. The final chromosome assembly and gene annotation of *Sillago sihama* has been submitted the Genome Warehouse in National Genomics Data Center (https://bigd.big.ac.cn/gwh) under accession number GWHAOSB00000000.

Literature Cited

Alioto T, Blanco E, Parra G, Guigó R. 2018. Using geneid to identify genes. Curr Protoc Bioinformatics. 64(1):e56.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Boeckmann B. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31(1):365–370.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 268(1):78–94.

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. 2006. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 7(1):327.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22(10):1269–1271.

Dimmer EC, et al. 2012. The UniProt-GO annotation database in 2011. Nucleic Acids Res. 40(D1):D565–570.

Edgar RC, Myers EW. 2005. PILER: identification and classification of genomic repeats. Bioinformatics 21(Suppl 1):i152–i158.

Griffiths-Jones S, et al. 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 33(Database issue):D121–D124.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34(90001):D140–D144.

Guo YS, et al. 2014. Isolation and characterization of microsatellite DNA loci from *Sillago sihama*. J Genet. 93(S1):32–36.

Haas BJ, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8):1494–1512.

Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9(1):R7.

Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110(1–4):462–467.

Keilwagen J, et al. 2016. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 44(9):e89.

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.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5(1):59.

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

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.

Li Z, et al. 2019. A first insight into a draft genome of silver sillago (*Sillago sihama*) via genome survey sequencing. Animals (Basel) 9:756.

Lou FR, Zhang Y, Song N, Ji DP, Gao TX. 2020. Comprehensive transcriptome analysis reveals insights into phylogeny and positively selected genes of *Sillago* species. Animals (Basel) 10(4):633.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20(16):2878–2879.

Marchler-Bauer A, et al. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39(Database):D225–229.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935.

Ogata H, et al. 1999. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27(1):29–34.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061–1067.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 11(9):1650–1667.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):i351–i358.

Qiu B, Fang S, Ikhwanuddin M, Wong L, Ma H. 2020. Genome survey and development of polymorphic microsatellite loci for *Sillago sihama* based on Illumina sequencing technology. Mol Biol Rep. 47(4):3011–3017.

Rusk N. 2014. Genomes in 3D improve one-dimensional assemblies. Nat Methods. 11(1):5.

Saetan W, et al. 2020. Comparative transcriptome analysis of gill tissue in response to hypoxia in silver sillago (*Sillago sihama*). Animals (Basel) 10(4):628.

Servant N, et al. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16(1):259.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.

Siyal FK, Xiao JG, Song N, Gao TX. 2016. The complete mitochondrial genome of *Sillago sihama* (Perciformes: Sillaginidae). Mitochondrial DNA A 27(4):2933–2934.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2):ii215–ii225.

Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 43(12):e78.

Tarailo-Graovac M, Chen NS. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 25(1):4.10.11–14.10.14.

Tatusov RL. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29(1):22–28.

Tian CX, et al. 2019. Transcriptome analysis of male and female mature gonads of silver sillago (*Sillago sihama*). Genes (Basel) 10(2):129.

Wang Y, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40(7):e49

Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8(12):973–982.

Xu SY, et al. 2018. A draft genome assembly of the Chinese sillago (*Sillago sinica*), the first reference genome for Sillaginidae fishes. Gigascience 7(9):giy108.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35(Web Server):W265–268.

Associate editor: Bonnie Fraser