

1Chromosomal-level genome assembly of silver sillago (*Sillago sihama*)

2Changxu Tian¹, Xinghua Lin¹, Yang Huang, Huapu Chen, Dongneng Jiang, Hongjuan

3Shi, Siping Deng, Tianli Wu, Yulei Zhang, Mouyan Jiang, Tao Du, Chunhua Zhu and

4Guangli Li*

5Fisheries College, Guangdong Ocean University, Guangdong Research Center on Reproductive Control

6and Breeding Technology of Indigenous Valuable Fish Species, Guangdong Provincial Engineering

7Laboratory for Mariculture Organism Breeding, Guangdong Provincial Key Laboratory of Pathogenic

8Biology and Epidemiology for Aquatic Economic Animals, Southern Marine Science and Engineering

9Guangdong Laboratory (Zhanjiang), Zhanjiang, 524088, China

10*Corresponding author: Guangli Li, Fisheries College, Guangdong Ocean University, Zhanjiang, China. E-

11mail: ligl@gdou.edu.cn.

12¹ These authors contributed equally to this work.

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 *S. sinica* about 7.92 million years ago. Comparative genomic analysis of silver sillago with other teleosts showed that 45 unique and 100 expansion gene families were identified in silver sillago. Expansion gene families were involved in immune and olfactory receptors. In this study, the genomic resources provide valuable reference genomes for functional genomics research of silver sillago.

29

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

31

321 INTRODUCTION

33 Sillaginidae family (also known as smelt-whittings or sand borers) belongs to order
34 Perciformes, are bottom-dwelling fishes and widely distributed in the shallow sea regions of

35 Indo-West-Pacific Ocean (S. Y. Xu et al., 2018). Sillaginidae consists of 31 species in 3
36 genera and 3 subgenera, of which the genus *Sillago* comprises 24 species. *Sillago* species drill
37 sand to avoid seine-net and other environmental hazards (Lou, Zhang, Song, Ji, & Gao, 2020).
38 *Sillago* flesh is white and very tender, with excellent flavor. Steamed whiting fillet of *Sillago*
39 fishes contains little fat content, which is easy to digest. Due to its ecological and economic
40 importance, the inshore fishing of *Sillago* has developed rapidly in the past decades. However,
41 the natural population of *Sillago* spp. has reduced in recent years due to overfishing and
42 demersal environmental deterioration, such as localized oxygen depletion, sulfide
43 accumulation and high turbidity (Lou et al., 2020). Therefore, it is necessary to develop
44 genomic resources to protect their natural resources and to accelerate the process of genome-
45 assisted improvement of important economic traits.

46 Silver sillago, *S. sihama* (Figure 1) is found in all Chinese waters, including beaches,
47 sandbars, mangrove creeks and estuaries (Guo et al., 2014). This fish species has been widely
48 cultured in China due to its high meat quality. However, the reduction of natural population of
49 *S. sihama* and a low survival rate in artificial breeding decrease the development of the
50 marine aquaculture of *S. sihama*. To date, complete mitogenome (Siyal, Xiao, Song, & Gao,
51 2015), simple sequence repeat (Guo et al., 2014; Qiu, Fang, Ikhwanuddin, Wong, & Ma,
52 2020), transcriptome (Saetan et al., 2020; Tian et al., 2019) and draft genomic survey data (Z.
53 Li et al., 2019) have been reported for *S. sihama*.

54 The genome of *S. sinica* was the first and only reference genome for Sillaginidae (Lou et
55 al., 2020). However, large-scale genomic analysis at the chromosome level has not been well-
56 characterized in *Sillago* due to the fragmented assemblies. Our study reported the

57chromosome-level genome of *Sillago*, which is the first chromosome-level genome of *S.*
58*sihama*. Genomic and comparative genomic analyses provide insights into the genes related to
59environmental stress. The genome can be used as a basis for the research on the evolution and
60biology of *S. sihama*.

61

62 **MATERIALS AND METHODS**

63 **2.1 Ethics statement**

64 All experimental protocols were approved by the Animal Research and Ethics
65Committees of the Institute of Aquatic Economic Animals of Guangdong Ocean University,
66Zhanjiang, Guangdong, China (201903003). The study does not involve endangered or
67protected species.

68

69 **2.2 Sample collection and sequencing**

70 *S. sihama* (length of 19.3 cm) was obtained from Donghai Island, Guangdong, China.
71Genomic DNA (gDNA) was extracted from muscle samples and constructed two Pacific
72Biosciences (PacBio) sequencing libraries (insert size of 20 kb). DNA samples were
73interrupted by g-TUBE, and the adaptor was connected to the DNA. The libraries were
74purified by an exonuclease, and the sequencing fragments were screened by BluePippin.
75Sequencing was conducted using the PacBio platform. Adaptors, low-quality reads and short
76fragments were filtered to obtain high-quality subreads.

77 The high-throughput chromosome conformation capture (Hi-C) library (insert size of
78350 bp) was constructed for sequencing to obtain the chromosome-level assembly of the

79genome. The samples were fixed by formaldehyde, and restriction enzyme was added to
80digest DNA, followed by repairing the 5'-end by biotin residues. Sequencing was done using
81the Illumina platform. Adapter sequences of raw reads were trimmed, and low-quality paired-
82end (PE) reads were removed to get clean data.

83 RNA was extracted from eight tissues, including liver, heart, head kidney, gonad, muscle,
84brain, stomach and gill of *S. sihama*. Illumina HiSeq platform was used for transcriptome
85sequencing.

86

872.3 Genome assembly

88 The filtered data were corrected by Canu (Koren et al., 2017), and then the corrected data
89were used to assemble the primary genome by WTDBG. After completing the primary
90assembly, the chromosomal-level genome was assembled from HI-C data. The clean data
91were compared with preliminary assembly results by Burrows-Wheeler Aligner (H. Li &
92Durbin, 2009). HiC-Pro (Rusk, 2014) was used to filter and evaluate the quality of Hi-C data.
93The genome sequence was divided into groups, and then sorted and oriented. The assembly
94results were evaluated by LACHESIS (Servant et al., 2015).

95

962.4 Genome prediction and annotation

97 Based on structural prediction and *de novo*, a repetitive sequence database of *S. sihama*
98genome was constructed by LTR FINDER v1.05 (Z. Xu & Wang, 2007), RepeatScout v1.0.5
99(Price, Jones, & Pevzner, 2005) and PILER-DF v2.4 (Edgar & Myers, 2005).
100PASTECClassifier (Wicker et al., 2006) was used to classify the repetitive sequence database

101and then merged with the Repbase (Jurka et al., 2005) database as the final repetitive
102sequence database. The repetitive sequence of *S. sihama* was predicted by RepeatMasker
103v4.0.6 (Tarailo-Graovac & Chen, 2009).

104 Based on *ab initio*, homologous alignment and transcriptome data were used to predict
105protein-coding genes in the genome. The *ab initio* prediction was done using Genscan (Burge
106& Karlin, 1997), Augustus v2.4 (Stanke & Waack, 2003), GlimmerHMM v3.0.4 (Majoros,
107Perte, & Salzberg, 2004), GeneID v1.4 (Alioto, Blanco, Parra, & Guigó, 2018) and
108Supplemental Nutrition Assistance Program (SNAP) (version 2006-07-28) (Korf, 2004). The
109protein sequences of *Larimichthys crocea*, *Oreochromis niloticus*, *Oryzias latipes*, *Danio*
110*rerio* and *S. sinica* were downloaded from the National Center for Biotechnology Information
111(NCBI) and GIGA databases. The homologous alignment was constructed using GeMoMa
112v1.3.1 (Keilwagen et al., 2016) to predict protein-coding genes. The reference transcripts were
113assembled by Hisat v2.0.4, Stringtie v1.2.3 (Perte, Kim, Perte, Leek, & Salzberg, 2016),
114TransDecoder v2.0 (Haas et al., 2013) and GeneMarkS-T v5.1 (Tang, Lomsadze, &
115Borodovsky, 2015) were used for gene prediction. Based on transcriptome data, unigene
116sequences were predicted by PASA v2.0.2 (Campbell, Haas, Hamilton, Mount, & Buell,
1172006). EVM v1.1.1 (Haas et al., 2008) was used to integrate the prediction results obtained by
118the above three methods.

119 We performed homology searches in public gene databases, including NCBI Refseq
120(Marchler-Bauer et al., 2011), Kyoto Encyclopedia of Genes and Genomes (KEGG,(Ogata et
121al., 1999), Clusters of orthologous groups for eukaryotic complete genomes (Tatusov et al.,
1222001), Translation of EMBL nucleotide sequence database (Boeckmann et al., 2003) and

123Gene Ontology (Dimmer et al., 2012). Function annotation was performed on the predicted
124gene sequences by BLAST v2.2.31 (Altschul, Gish, Miller, Myers, & Lipman, 1990) (-evalue
1251e-5). Based on the comparison results of the NR database, the functional annotation of the
126GO database was performed by Blast2GO (Conesa et al., 2005).

127 The rRNA and microRNA sequences were predicted by Infernal 1.1 (Nawrocki & Eddy,
1282013) on the Rfam (Griffiths-Jones et al., 2005) and miRBase (Griffiths-Jones, Grocock, van
129Dongen, Bateman, & Enright, 2006) databases. The tRNA was identified by tRNAscan-SE
130v1.3.1 (Lowe & Eddy, 1997).

131

1322.5 Assessment of completeness of the genome assembly

133 The core eukaryotic gene mapping approach was used to assess the completeness of
134assembly and gene annotation (CEGMA, v2.5) (<http://korflab.ucdavis.edu/datasets/cegma/>)
135(Parra, Bradnam, & Korf, 2007) and benchmarking universal single-copy orthologs (BUSCO,
136v2) (<http://busco.ezlab.org/>) (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015)
137were used.

138

1392.6 Genome evolution analysis

140 Based on the protein sequences of the *S. sihama* and 10 other teleosts, including *Takifugu*
141*rubripes* (Accession no.: GCA_000180615.2), *Gasterosteus aculeatus* (Accession no.:
142GCA_006229165.1), *O. latipes* (Accession no.: GCA_004347445.1), *D. rerio*
143(GCA_000002035.4), *O. niloticus* (Accession no.: GCA_001858045.3), *Latimeria chalumnae*
144(Accession no.: GCF_000225785.1), *S. sinica* (Accession no.: PRJNA437933), *L. crocea*

145(Accession no.: GCA_003845795.1), *Lepisosteus oculatus* (Accession no.:
146GCA_000242695.1) and *Xiphophorus maculatus* (Accession no.: GCA_002775205.2). The
147evolution between species and the classification of gene families were analyzed. The protein
148sequences of 11 teleosts were classified into gene families, and single-copy genes were
149extracted by OrthoMCL (L. Li, Stoeckert, & Roos, 2003). In order to study the evolutionary
150relationship between 11 teleosts, the single-copy protein sequences of 11 teleosts were used to
151construct the maximum-likelihood (ML) phylogenetic tree by PHYML (Guindon et al., 2010).
152The divergence time was predicted by McMcree in PAML and timetree databases
153(<http://www.timetree.org/>) to correct divergence time. *L. crocea* was phylogenetically closely
154related to *S. sihama*. The 24 *S. shama* chromosomes were aligned with *L. crocea*
155chromosomes by MCScanX to visualize the consistency between the genomes of *S. sihama*
156and *L. crocea* (Wang et al., 2012).

157

1582.7 Gene family expansion and contraction analysis

159 The expansion and contraction gene families among *T. rubripes*, *G. aculeatus*, *O. latipes*,
160*D. rerio*, *O. niloticus*, *L. chalumnae*, *S. sinica*, *L. crocea*, *L. oculatus*, *X. maculatus* and *S.*
161*sihama* were identified by CAFÉ (De Bie, Cristianini, Demuth, & Hahn, 2006). The number
162of gene families of each ancestor was estimated by the birth mortality model, thereby
163predicting the number of gene family expansion and contraction gene families.

164

1652.8 Selective pressure analysis

166 Multiple alignments were constructed based on the single-copy gene sequences of each

teleost by ClustalW to identify possible positively selected genes (PSGs). We used the Branch Site model to analyze the selection pressure of single-copy genes of each teleost by CodeML (Schabauer et al., 2012) in PAML.

170

1713 RESULTS AND DISCUSSION

1723.1 Genome sequencing and assembly

173 After quality filtering, 66.16 Gb subread data was obtained from two long-insert (20 kb) libraries (sequence coverage: $\sim 126\times$; subread N50: 15,715 bp; Table S1). A total of 89.08 Gb Hi-C data was obtained from the Hi-C sequencing library (sequence coverage: $\sim 170\times$; GC content: 43.95%; Q30: 90.92%; Table S1).

177 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 was 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 (Figures 2, Figure 3, Figure S1). 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 (Table S2, Figure S2).

186 According to BUSCO results, the genome contained 4463 (97.36%) complete BUSCOs, including 4345 single-copy BUSCOs and 118 duplicated BUSCOs (Table S3). The CEGMA v2.5 database contained 248 conserved core genes of eukaryotes, and there were 246

189 conserved core genes (99.19%) in this genome (Table S4). The results indicated that the
190 genome assembly had high coverage and completeness.

191

192 3.2 Genome annotation

193 *de novo* prediction and Repbase database results showed that the repeated sequences
194 accounted for 13.78% of *S. sihama* genome, which is lower than *D. rerio* (63.12%), *O. latipes*
195 (42.83%) and *L. crocea* (20.31%), and higher than *S. sinica* (10.92%) and *T. rubripes*
196 (9.37%). DNA transposons (3%) were the most common among transposons of *S. sihama*
197 genome, followed by long interspersed repeated segments (LINEs, 1.44%) and long terminal
198 repeats (LTR, 1.33%) (Table S5, Figure 3).

199 A total of 23,959 protein-coding genes (Table S6) were predicted in the *S. sihama*
200 genome by *ab initio*, homologous prediction and RNA-seq prediction methods, with an
201 average length of 11241.51 bp. Comparing the length distribution of genes, coding sequences
202 (CDS), exons and introns, the gene distribution of *S. sihama* was similar to other teleosts. *S.*
203 *sihama* gene proportions were lower than other fishes but similar to *S. sinica* (Figure 4). The
204 functions of the protein-coding genes were annotated in NR, TrEMBL, KOG, KEGG and GO
205 databases. A total of 23,123 genes were annotated, accounting for 96.5% of all protein-coding
206 genes (Table S7).

207 Rfam, miRBase and tRNAscan-SE databases were used to predict non-coding RNA, and
208 a total of 1,587 tRNAs, 67 rRNAs, 419 miRNAs and 301 snRNAs were predicted (Table S8).

209

210 3.3 Comparative genome analysis

211 The genomes of 11 teleosts were compared to study the phylogenetic relationships
212 between *S. sihama* and other teleosts. A total of 16,856 gene families and 5,950 single-copy
213 orthologs were identified (Table S9, Figure 5). The ML phylogenetic tree was constructed
214 from single-copy orthologs. The phylogenetic tree showed that *S. sinica* was closely related to
215 *S. sihama*, and the divergence time was about 7.92 (2.45–16.57) million years ago (Figure 6).
216 The genomes of *S. sihama* and *L. crocea* were compared to analyze chromosomal
217 evolutionary events (Figure 7). The results showed that the 24 chromosomes of *S. sihama*
218 were aligned with 22 chromosomes of *L. crocea*. The chromosomes III and XII of *L. crocea*
219 were compared to LG2, LG10, LG5 and LG16 of *S. sihama*, respectively. The common
220 ancestor of *L. crocea* and *S. sihama* undergone a chromosome break recombination event
221 during the evolution process, which increases the number of chromosomes.

222

223 3.4 Gene family analysis

224 The expansion and contraction of gene families are one of the most important factors for
225 the evolution of phenotypic diversity and environmental adaptation. *S. sihama* is sensitive to
226 environmental factors such as sound, vibration, light and shadow. In order to explore the
227 adaptability of environmental factors in *S. sihama*, the gene families of 11 teleost fishes (*T.*
228 *rubripes*, *G. aculeatus*, *O. latipes*, *D. rerio*, *O. niloticus*, *L. chalumnae*, *S. sinica*, *L. crocea*, *L.*
229 *oculatus*, *X. maculatus* and *S. sihama*) were compared. A total of 57 unique, 100 expanded (P
230 < 0.05) and 25 contracted ($P < 0.05$) gene families were identified in *S. sihama* (Table S10),
231 including immune-related gene families (immunoglobulin domain, immunoglobulin V-set
232 domain, immunoglobulin I-set domain and NACHT domain) and olfactory receptor gene

233family (7 transmembrane receptor). The immune-related gene families were also expanded in
234Perciformes genome, such as *Epinephelus akaara* (Ge et al., 2019), *Epinephelus lanceolatus*
235(Zhou et al., 2019), *Oreochromis aureus* (Bian et al., 2019), *Miichthys miiuy* (T. Xu et al.,
2362016) and *Larimichthys crocea* (Mu et al., 2018). In addition, the expansion of olfactory
237receptor genes were also found in the *Miichthys miiuy* (T. Xu et al., 2016) and *Larimichthys*
238*crocea* (Mu et al., 2018) genomes.

239 Novel immune-type receptor 1 (*nitr1*), t-cell receptor alpha (*tra*), polymeric
240immunoglobulin receptor (*pigr*) and signal-regulatory protein beta-2 (*sirpβ2*) were expanded.
241*Nitr1* was presented in the V-set of immunoglobulin and T cells, which plays a vital role in
242innate and adaptive immunities (Litman, Hawke, & Yoder, 2001). The *pigr* gene is an
243essential part of the mucosal immune system, which is related to innate and adaptive
244immunities (Rombout et al., 2008). *Sirpβ2* binds to T cells with CD47, which results in
245enhanced proliferation of antigen-specific T cells (Seiffert et al., 2001). The *nitr1* and *pigr*
246genes were associated with adaptive and innate immunities, and their expansions indicate the
247enhancement of adaptive immunity of *S. sihama*. In the NACHT gene family, NACHT, LRR
248and PYD domains containing protein 12 (NLRP12) and NLRP3 were expanded. NLRP3
249plays a vital role in innate immunity and inflammation (Shao, Xu, Han, Su, & Liu, 2015).
250NLRP12 inhibits the release of inflammation-related molecules, thereby reducing the damage
251of the inflammatory response to the cells and tissues (Normand et al., 2018). Stressed fish
252stimulates the autonomic nervous system to secrete large amounts of catecholamine hormones
253and activates the hypothalamus-pituitary-interrenal (HPI), thereby regulating the defense
254response of fish immune system (Wendelaar Bonga, 1997). The expansion of immune-related

255families in *S. sihama* adapt to the stress response. It is speculated that *S. sihama* reduces the
256impact of the stress response by improving the capacity of the immune system.

257 Olfactory receptors were divided into main olfactory receptors (*mor*), olfactory receptors
258related to class A (*ora*), olfactory receptors related to class C (*olfc*), and trace amine-
259associated receptors (*taar*) in fish. The *mor* genes were divided into type-I and type-II
260according to their functions. Each type included several subtypes. Type-I was subdivided into
261 α , β , γ , δ , ϵ and ζ subfamilies, while type-II was subdivided into η , θ and κ subfamilies. Type-I
262*mor* was used to identify water-soluble odorant molecules, while type-II *mor* was applied to
263identify volatile odorant molecules (Glusman et al., 2000; Malnic, Godfrey, & Buck, 2004). In
264this study, the olfactory receptor 142 (or142, δ subfamily), olfactory receptor family 2
265subfamily a member 12 (or2a12, η subfamily), olfactory receptor family 2 subfamily ag
266member 2 (or2ag2, δ subfamily), and olfactory receptor family 52 subfamily n member 5
267(or52n5, ϵ subfamily) of main olfactory receptors were expanded. It is possible that the
268expansion of δ and ϵ subfamilies caused the significant functional differentiation of *S. sihama*
269in the recognition of specific water-soluble odors. However, the reasons for the expansion of η
270subfamily in *S. sihama* are still unknown, which needs to be examined in further studies.
271Olfactory receptors play an important role in the recognition of stressors. Fish recognizes
272amino acids, steroids, prostaglandins, cholic acid and other odorous molecules in the
273surrounding water environment through olfactory receptor (OR) proteins (Freitag, Ludwig,
274Andreini, RoÈssler, & RoÈssler, 1998), and can accurately detect the changes in the
275environment. The expansion of the olfactory receptor gene family of *S. sihama* possibly
276enabled *S. sihama* to more accurately detect subtle environmental changes and search for

277food.

278

2793.5 Selective pressure analysis

280 Positively selected gene (PSG) is the result of adaptive evolution, and it is usually related
281to the selected function during the evolution of an organism. A total of 5950 single-copy
282genes from 11 species of teleost fishes were used to identify the PSGs of *S. sihama* for
283conducting selective pressure analysis. A total of 177 significantly different ($P < 0.05$) positive
284selection genes were obtained (Table S11). Transcription factor *hivep2* (*hivep2*) served as a
285transcriptional factor regulating NF- κ B and diverse genes that are essential in neural
286development (Srivastava et al., 2016). JmjC domain-containing protein 8 (*jmjd8*) is recently
287shown to be involved in angiogenesis and TNF-induced NF- κ B signaling pathway (Yeo et al.,
2882016). PDZ and LIM domain protein 7 (*pdlim7*) was involved in the formation of heart valves
289and pectoral fins in zebrafish (Camarata, Krcmery, et al., 2010; Camarata, Snyder, et al.,
2902010). GDP-mannose 4,6 dehydratase (*gmds*) encoded a short-chain mannose dehydrogenase
291enzyme involved in the regulation of hindbrain neural migration (Haliburton, McKinsey, &
292Pollard, 2016). The *hivep2*, *jmjd8*, *pdlim7* and *gmds* genes were under positive selection
293pressure, suggesting that *S. sihama* has a comprehensive nervous system and circulatory
294system. In addition, comm domain-containing protein 5 (*commd5*) was a transcription factor
295affecting adaptive immunity, apoptosis, and oncogenesis (Burstein et al., 2005). The *hivep2*
296and *jmjd8* genes regulated the NF- κ B signaling pathway, which play a key role in regulating
297the immune response to infection. The evolution of immune-related genes may enhance the
298resistance of organisms to external stimuli in *S. sihama*.

3004 CONCLUSIONS

301 This study was determined the chromosomal-level genome assembly of *S. sihama*. The
302continuity and completeness of the *S. sihama* genome was reached the level of other high-
303quality teleost fish genomes, which provides a useful reference for system biology and
304comparative genome evolution analysis. Genome evolution analysis showed the insights into
305the high irritability of *S. sihama*, and found significant changes in immune, olfactory receptor
306and stimulus-response. This reference genome is important for aquaculture and artificial
307breeding of *S. sihama*, which provides a basis for further research.

308

309ACKNOWLEDGEMENTS

310 This study was supported by grants from the National Natural Science Foundation of
311China (Nos. 41706174 and 31702326); Natural Science Foundation of Guangdong Province
312(2016A030313743, 2017A030313101, 2018B030311050, 2019A1515010958,
3132019A1515110619); Independent Project of Guangdong Province Laboratory (ZJW-2019-
31406); the Department of Education of Guangdong Province (2018KQNCX111,
3152019KTSCX060) and Program for Scientific Research Start-up Funds of Guangdong Ocean
316University (R19026).

317

318CONFLICTS OF INTERESTS

319 The authors declare that they have no competing interests.

320

321AUTHOR CONTRIBUTIONS

322 C.X. T., and X.H. L., designed research, performed research, analyzed data and wrote
323the paper. Y. H., H.P. C., D.N. J., S.H. J, S.P. D., T.L. W., Y.L. Z, M.Y. J., T. D., and C.H.
324Z., collected the samples for sequencing and obtained funding. G.L. L., obtained funding,
325conceived and managed the project. All authors reviewed the manuscript.

326

327**DATA AVAILABILITY**

328 The raw genome and RNA sequencing data have been submitted in the SRA under
329Bioproject number PRJNA642704. The final chromosome assembly and gene annotation of *S.*
330*sihama* has been submitted the Genome Warehouse in National Genomics Data Center
331(<https://bigd.big.ac.cn/gwh>) under accession number GWHAOSB000000000.

332

333**References**

334Figure legends

335Figure 1 *Sillago sihama*.

336Figure 2 The chromosome contact maps of *S. sihama* genome. LG0-LG23 represent
337Lachesis Groups 0-23; the abscissa and ordinate represent the order of each bin on the
338corresponding chromosome group.

339Figure 3 Genome landscape of *S. sihama*. (A) chromosome length, (B) GC content,
340(C) gene density, (D) repeat sequence, (E) long terminal repeated (LTE), (F) long
341interspersed nuclear elements (LINE) and (G) simple sequence repeat (SSR).

342Figure 4 The length distribution of (A) annotated genes, (B) coding sequences (CDS),
343(C) exons and (D) introns between *S. sihama* and other teleosts.

344Figure 5 Statistics of gene family clustering. Clusternum: genes that have not been
345clustered into any family; other gene: all other genes; special gene: genes in the
346species-specific gene family; multi-copy: multi-copy homologous genes in common
347gene family of species; one-copy: single-copy homologous genes in common gene
348families of species.

349Figure 6 Phylogenetic analysis of 11 teleost fishes. At each branch point, the predicted
350species divergence time (million years ago) is marked. The red number on each
351evolutionary branch represents the number of expanding gene families, and the blue
352number represents the number of contracting gene families.

353Figure 7 Collinearity analysis of *S. sihama* and *L. crocea* genomes. Blue and orange
354outer circles represent the chromosome of *S. sihama* and *L. crocea*, respectively.

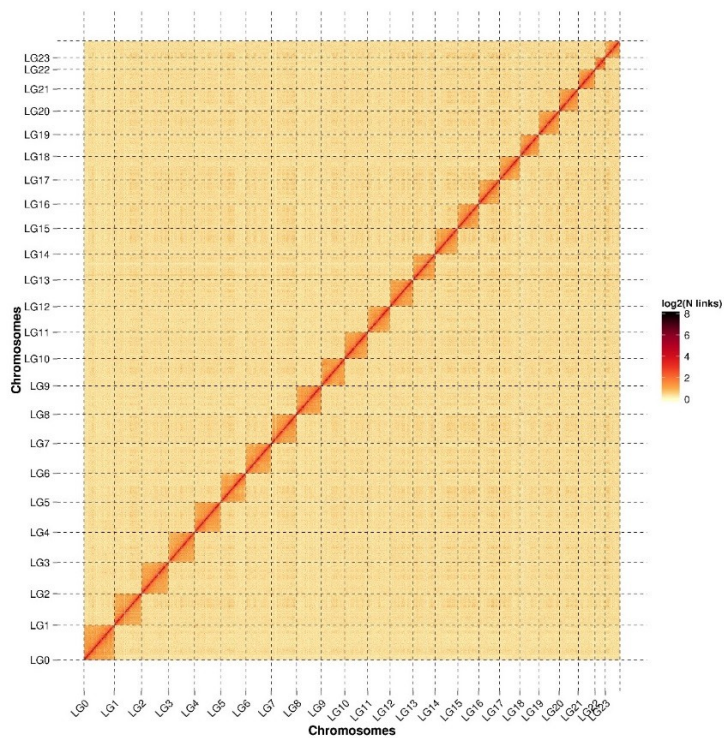
355Figure S1 The length distribution of contig in the genome of *S. sihama*.

356Figure S2 The distribution of gaps in the chromosomes of *S. sihama*. The red line on
357the chromosome represents the gap.



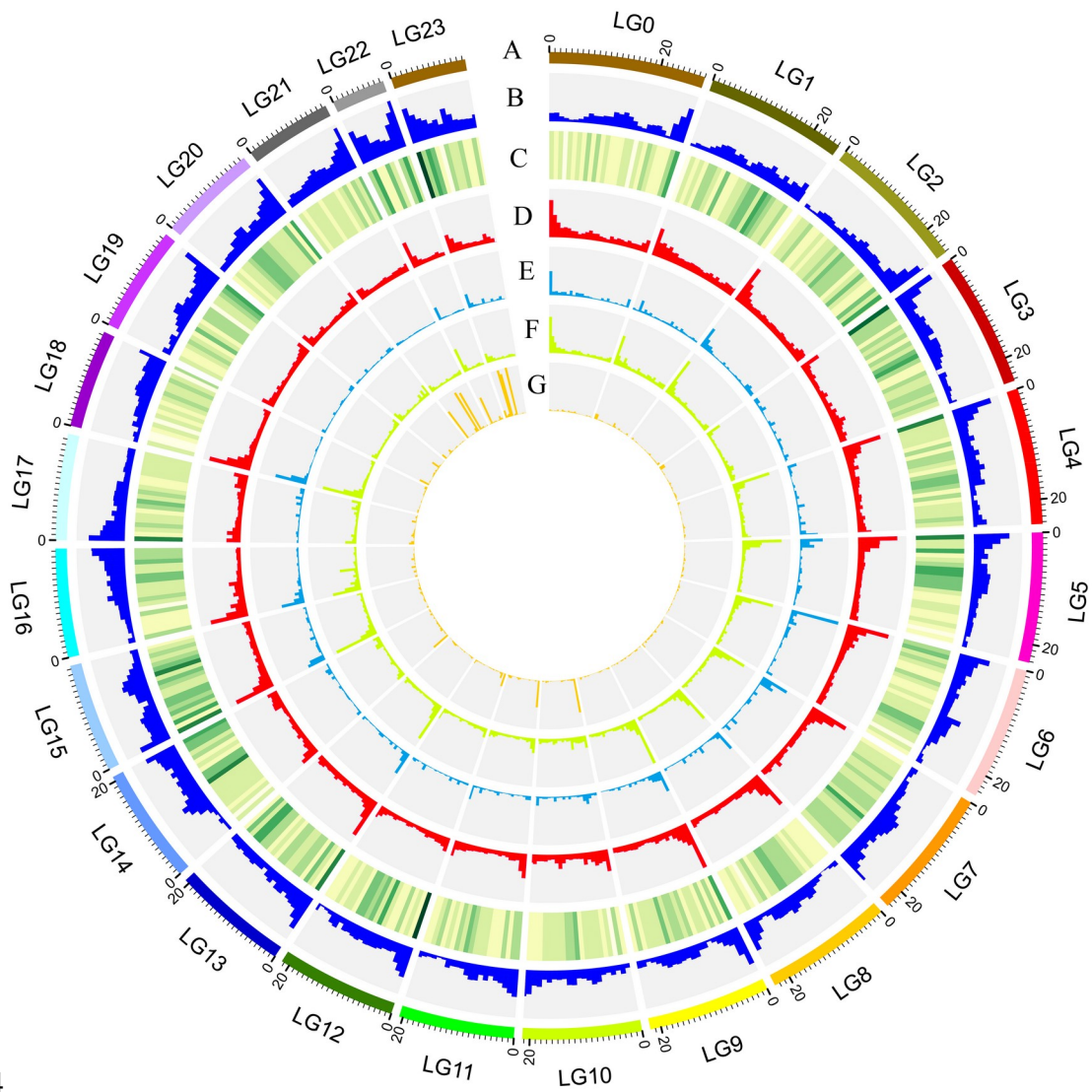
358

359Figure 1 *Sillago sihama*.



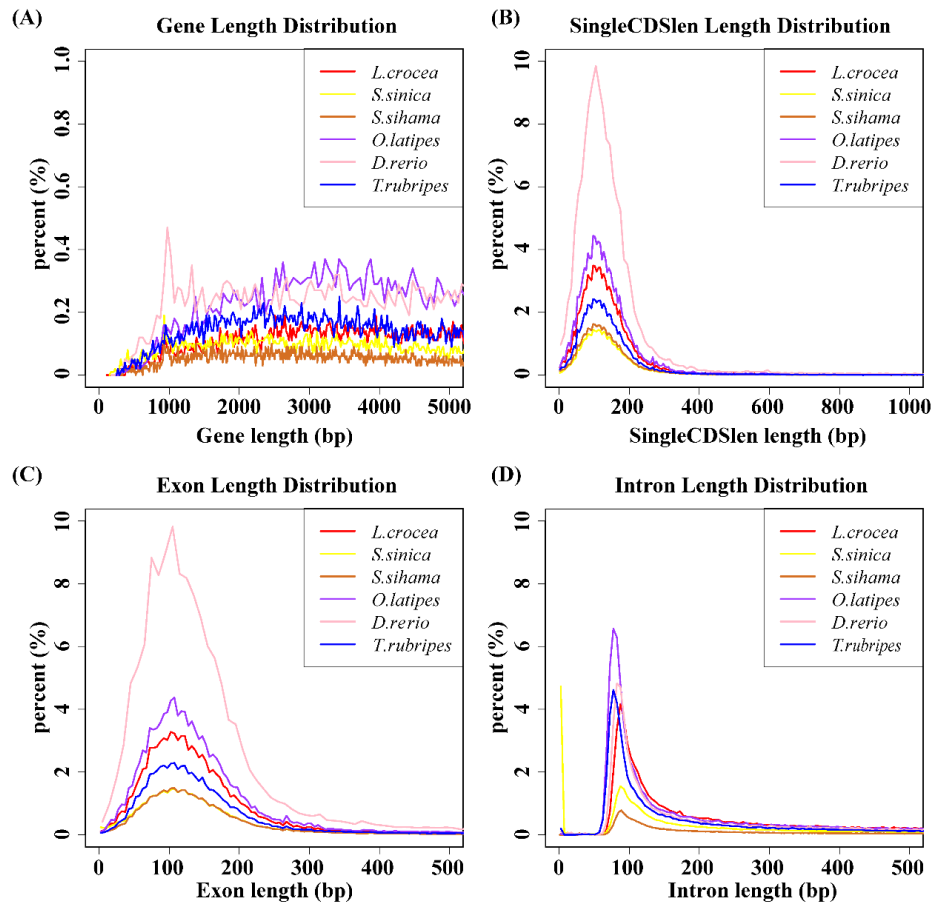
360

361Figure 2 The chromosome contact maps of *S. sihama* genome. LG0-LG23 represent
362Lachesis Groups 0-23; the abscissa and ordinate represent the order of each bin on the
363corresponding chromosome group.



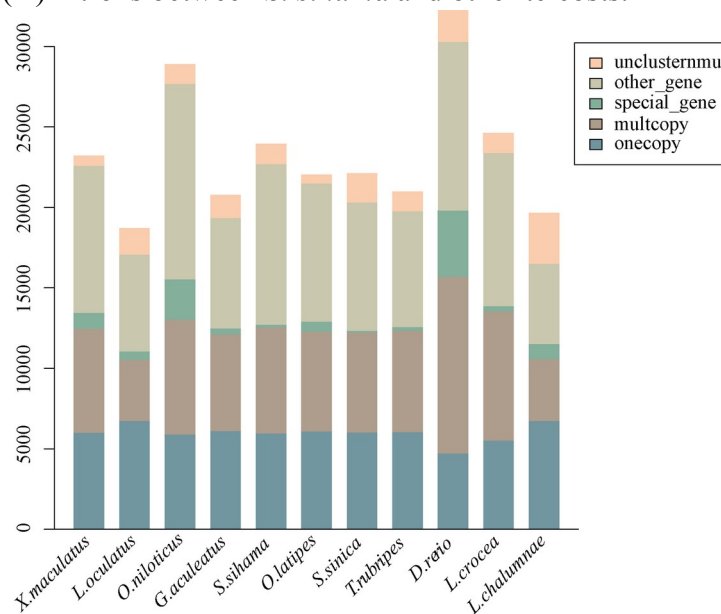
364

365 Figure 3 Genome landscape of *S. sihama*. (A) chromosome length, (B) GC content,
 366 (C) gene density, (D) repeat sequence, (E) long terminal repeated (LTE), (F) long
 367 interspersed nuclear elements (LINE) and (G) simple sequence repeat (SSR).



368

369 Figure 4 The length distribution of (A) annotated genes, (B) coding sequences (CDS),
370 (C) exons and (D) introns between *S. siham* and other teleosts.



371

372 Figure 5 Statistics of gene family clustering. Clusternum: genes that have not been
373 clustered into any family; other gene: all other genes; special gene: genes in the
374 species-specific gene family; multi-copy: multi-copy homologous genes in common
375 gene family of species; one-copy: single-copy homologous genes in common gene
376 families of species.

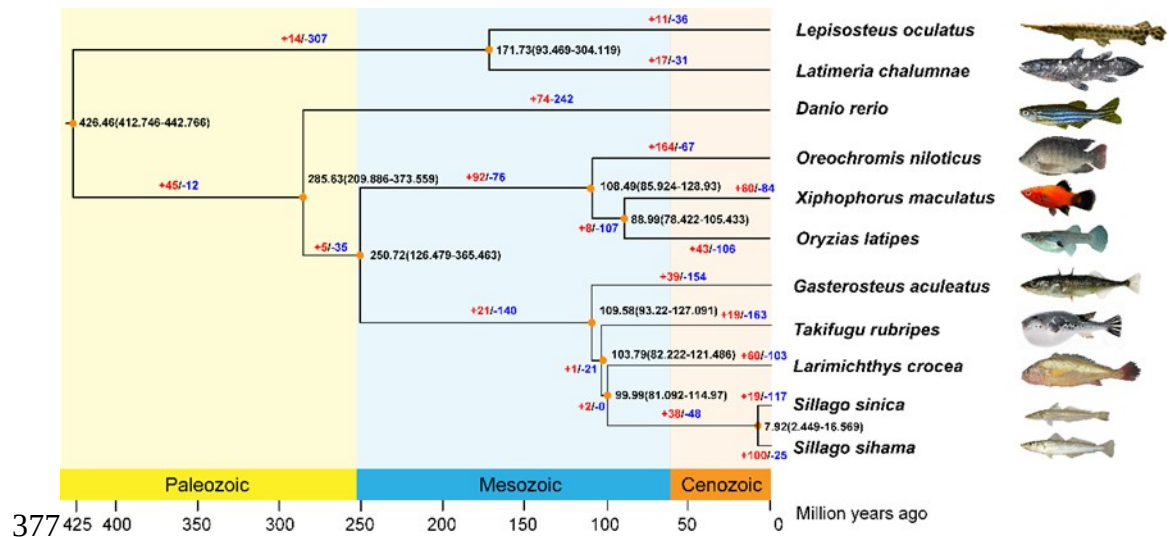


Figure 6 Phylogenetic analysis of 11 teleost fishes. At each branch point, the predicted species divergence time (million years ago) is marked. The red number on each evolutionary branch represents the number of expanding gene families, and the blue number represents the number of contracting gene families.

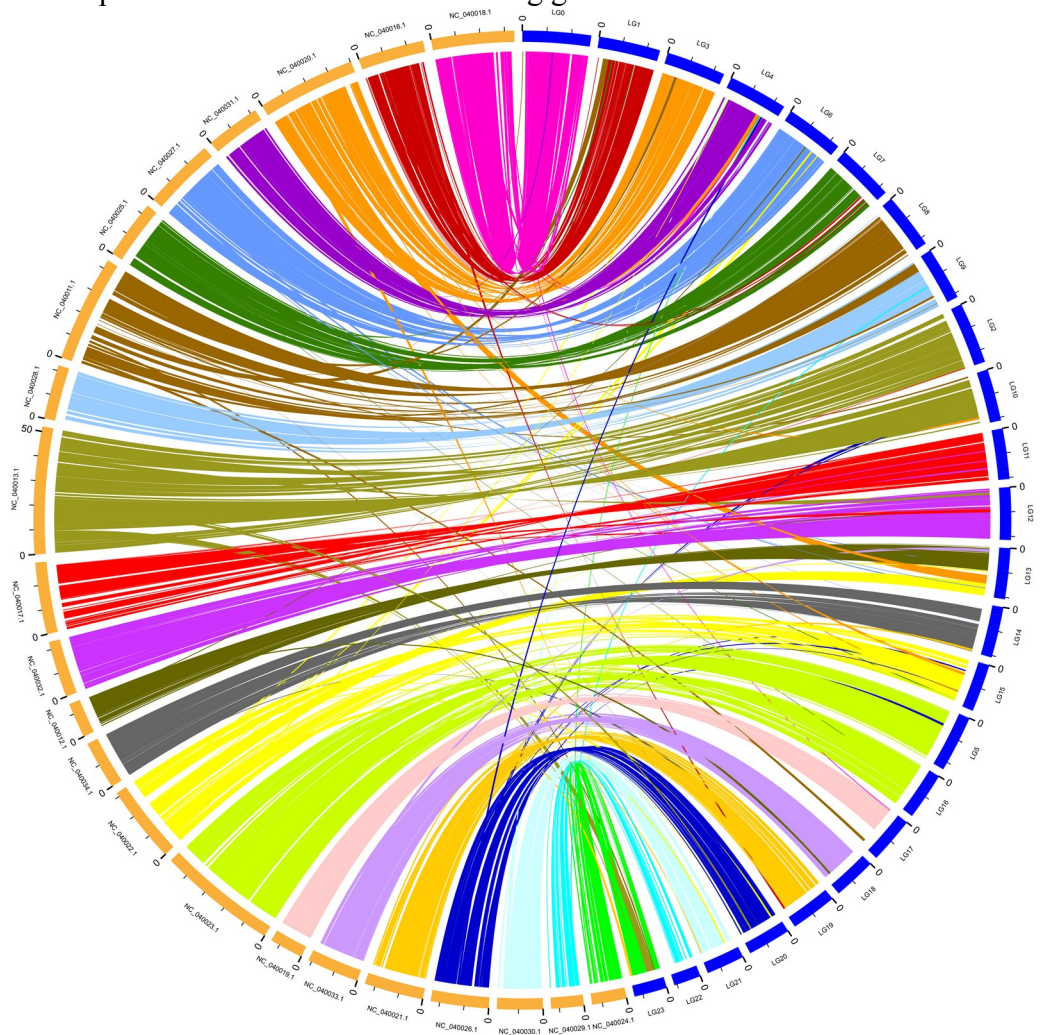


Figure 7 Collinearity analysis of *S. sihama* and *L. crocea* genomes. Blue and orange outer circles represent the chromosome of *S. sihama* and *L. crocea*, respectively.

385Table legends

386Table 1 Statistics of *S. sihama* genome assembly data.

387Table S1 Statistics of *S. sihama* genome sequencing data.

388Table S2 Statistics of chromosome contigs and gap data of *S. sihama* genome.

389Table S3 Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of *S.*
390*sihama* genome.

391Table S4 Core Eukaryotic Genes Mapping Approach (CEGMA) analysis of *S. sihama*
392genome.

393Table S5 Comparison of repetitive genomic sequences between *S. sihama* and five
394teleosts.

395Table S6 Statistics of the protein-coding gene prediction.

396Table S7 Statistical analysis of the protein-coding gene functional annotation.

397Table S8 Statistical analysis of non-coding protein genes.

398Table S9 Statistics of gene family clustering.

399Table S10 Statistics of expansion and contraction gene family in *S. sihama*.

400Table S11 Positive selection genes in *S. sihama*.

401

402 Table 1 Statistics of *S. sihama* genome assembly data.

| | Primary genome assembly* | Chromosome-level genome assembly** |
|--------------------|--------------------------|------------------------------------|
| Number of contigs | 551 | 556 |
| Contig N50 (bp) | 13,559,141 | 13,543,514 |
| Contig N90 (bp) | 1,284,248 | 1,283,116 |
| Contig max (bp) | 22,127,184 | 22,111,180 |
| GC content (%) | 44.67 | 44.66 |
| Contig length (bp) | 522,064,597 | 521,631,495 |

403Note: * The PacBio data was used to construct the primary assembly.

404** Re-assemble after correcting the errors of the primary assembled genome by the
405high-throughput chromosome conformation capture (Hi-C data).