Chromosome-level genome assembly of *Acanthopagrus latus* using PacBio and Hi-C technologies

Dong Gao¹, Wenyu Fang¹, Ying Sims³, Joanna Collins³, James Torrance³, Genmei Lin¹, Jingui Xie¹, Kerstin Howe³, Jianguo Lu^{1, 2*}

1. School of Marine Sciences, Sun Yat-sen University, Zhuhai, 519082, China

2. Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519080, China

3. Wellcome Sanger Institute, Cambridge, UK

*Corresponding author(s): Jianguo Lu (lujianguo@mail.sysu.edu.cn)

The yellowfin seabream, *Acanthopagrus latus*, is widely distributed throughout the Indo-West Pacific. This fish is an ideal model species in which to study the mechanism of sex reversal since it exhibits a specific feature: sequential hermaphrodite. Here, we report a chromosome-scale assembly of the *A. latus* based on PacBio and Hi-C data. 22,485 protein-coding genes were annotated in whole genome level using transcriptome data. Taken together, this highly accurate, chromosome-level reference genome can provide a valuable resource to elucidate the mechanism of sex reversal for *A. latus*.

Background & Summary

Evolution of sex, especially the evolution of different sexual systems, is a fascinating subject in evolutionary biology. The Sparidae, commonly known as seabreams or progies, is a family of fishes of the order Perciformes. And this family consist about 150 species in the world, which are mainly coastal fish¹. Previous researchers mentioned that Sparidae is an ideal taxon to study the evolution history and adaptive significance of sexual systems, particularly for both types of sequential hermaphroditism, given that this group contains many protogyny, protandry and genochorist species².

The yellowfin seabream, *Acanthopagrus latus* is a protandry species which belongs to the Sparidae family. It is widely distributed in Indo-West Pacific area³. It has a great relevance for marine aquaculture and its biology is well focused on reproductive physiology and nutrition⁴. Interestingly, *A. latus* has a special gender feature is that it belongs to protandrous sexual system (initially as male and change later to female)⁵. Most of the past studies of *A. latus* mainly focused on the reproductive biology, population structure, aquaculture and taxonomy^{3,4,6-8}. Although some sex reversal related genes were found in *A. latus*, the lack of genomic resources still limit us to elucidate the mechanism

of sex reversal for this species.^{9,10}. In addition, this lack was also limited the studies of evolution of sexual systems for Sparidae.

In this study, long-read (PacBio SMRT) sequencing and Hi-C sequencing technologies were applied to construct a high quality reference genome for yellowfin seabream. This high-quality genome can provide a valuable resource to elucidate the mechanism of sex reversal for *A. latus*. Furthermore, this genome can also facilitate the studies of evolution of sexual systems for Sparidae.

Methods

Ethics statement

All experimental procedures in our study with *A. latus* were approved by the Ethics Committee of Sun Yat-sen University.

Sample collection, library construction and sequencing

A wild healthy female yellowfin seabream was captured from the area of Xiangzhou Bay, Zhuhai, Guangdong Province, China (Fig. 1). Seven tissues were collected respectively for genome sequencing and genome annotation, including brain, heart, liver, spleen, kidney, gonad and muscle. These samples were immediately frozen using liquid nitrogen for 30 minutes and then stored at -80°C for later usage. For high-molecular-weight (HMW) genomic DNA (gDNA) extraction, frozen samples were lysed in SDS digestion buffer with proteinase K. Then, the lysates were purified using AMPure XP beads to obtain HMW gDNA. Meanwhile, normal-molecular-weight (NMW) gDNA was extracted from the same samples using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, USA).

A whole-genome shotgun sequencing strategy was performed for genome size estimation and polishing of preliminary contigs. A library with 350 bp insert size was constructed from NMW gDNA using the standard protocol provided by Illumina (San Diego, CA, USA). Paired-end sequencing was employed using the Illumina NovaSeq platform with a read length of 2×150 bp. 38.44 Gbp raw reads were generated. Adaptor, low-quality and duplicated reads were trimmed with fastp (v0.20.0)¹¹. In total, 34.97 Gbp clean reads were used to genome size estimation and preliminary contig polishing (Table 1).

Table 1 Summary of obtained data using multiple sequencing technologies. Note: The genome size of *A. latus* used to calculate sequencing coverage was 656.61 Mbp, which was estimated using a K-mer analysis of the short reads.

Library Type	Clean data (Gbp)	Average Read Length of Raw Reads (bp)	Sequencing Coverage (×)
Illumina	34.97	150	51.05
Pac-Bio	180.74	16,490	263.85
Hi-C	60.81	150	88.77
Total	276.52	-	403.67

HWM gDNA was used in DNA template preparation for sequencing on the PacBio System following the “Template Preparation and Sequencing Guide” provided by Pacific Biosciences (Menlo Park, CA, USA). The main steps were as follows: extracted DNA was first sheared into large fragments (10 Kbp on average) and then purified and concentrated using AMPure PB beads; DNA damage and ends induced in the shearing step were repaired; blunt hairpins were subsequently ligated to the repaired fragment ends; prior to sequencing, the primer was annealed to the SMRTbell template, and then, DNA polymerase was bound to the annealed templates; finally, DNA sequencing polymerases were bound to the primer-annealed SMRTbell templates.

After sequencing, a total of 180.74 Gbp long reads were generated from the PacBio SEQUEL platform. The average length of reads was 16.49 Kbp. The long reads covered the genome about $263.85 \times$ (Table 1).

Hi-C sequencing was performed parallel to the PacBio sequencing. We used formaldehyde to fix the conformation of the HMW gDNA. Then, the fixed DNA was sheared with DpnII restriction enzyme. The 5' overhangs induced in the shearing step were repaired using biotinylated residues. Following the ligation of blunt-end fragments *in situ*, the isolated DNA was reverse-crosslinked, purified, and filtered to remove biotin-containing fragments. Subsequently, DNA fragment end repair, adaptor ligation, and polymerase chain reaction (PCR) were performed successively. Sequencing was performed on the Illumina NovaSeq platform and yielded a total of 67.99 Gbp paired-end reads, with an average sequencing coverage of $88.77 \times$ (Table 1).

In addition, total RNA was extracted from each tissue using TRIZOL (Invitrogen, Carlsbad, CA, USA). The RNA samples were then treated by Dnase I. The integrity and size distribution were checked with Bioanalyzer 2100 (Agilent technologies, santa Clara, CA, USA). The high-quality RNA samples were mixed and then sequenced on Illumina NovaSeq platforms with the manufacturer's instructions. At last, 69.36 Gbp raw reads were generated for transcriptome-base gene prediction.

***De novo* assembly of the *A. latus* genome**

In summary, whole-genome shotgun sequencing data were used in estimation of genome size and polishing of preliminary contigs; PacBio sequencing data were used for preliminary contain assembly; and Hi-C reads were used in chromosome-level scaffolding.

The shotgun sequencing data were used to estimate the genome size with Jellyfish (v2.1.3)¹². As a result, the genome size of *A. latus* was estimated to be approximately 656.61 Mbp. All raw long-read sequences were aligned to each other using ‘dalinger’ executed by the main script of the FALCON assembler¹³. The overlap data and raw subheads were then processed to generate consensus sequences. After the error-correction step, FALCON identified the overlaps between all pairs of the preassembled error-corrected reads. The read overlaps were used to construct a directed string graph that contains sites of ‘haplotype-fused’ contigs as well as bubbles representing divergent regions between homologous sequences. Next, FALOCN-Unzip identified read haplotypes using phasing information from heterozygous positions. Phased reads were then used to assemble haplotigs and primary contigs. The shotgun sequencing data and PacBio long-reads were used to polish the preliminary contigs with Nextpolish and Arrow (version 1.21, Pacific BioSciences). The draft genome of *A. latus* was assembled in 215 contigs. The genome size equivalents to 685.14 Mbp with contig N50 of 14.88 Mbp (Table 2).

131 **Table 2 Summary of the *A. latus* genome assembly and structural annotation.**

Genome Assembly	
Contig N50 length (Mbp)	14.88
Number of contigs longer than N50	18
Number of contigs	215
Scaffold N50 length (Mbp)	30.72
Number of contigs longer than N50	11
Number of contigs	65
Total contain length (Mbp)	685.14
Structural Annotation	
Number of protein-coding genes	29,227
Average genes length (bp)	10,797.36
Average exons per gene	8.51
Average exons length (bp)	172.77
Average CDS length (bp)	1,470.37
Average intro length (bp)	1, 087.79

132
133 To obtain chromosome-level scaffolds, Hi-C reads were filtered in the same way as the shotgun
134 sequencing reads. Subsequently, filtered Hi-C reads were mapped to *de novo* assembled contigs to
135 construct contacts among the contigs using BWA (version 0.7.17)¹⁴ with the default parameters.
136 BAM files containing Hi-C linking messages were processed by another round of filtering, in which
137 reads were removed if they were not mapped to the reference genome within 500 bp from the nearest
138 restriction enzyme site. Then, LACHESIS¹⁵ was used for ultra-long-range scaffolding of *de novo*
139 genome assemblies using the signal of genomic proximity provided by the Hi-C data (Fig. 2).
140 The parameter CLUSTER_N was used to specify the number of chromosomes. For yellowfin
141 seabream, this number was determined to be 24 in previous studies¹⁶. Ultimately, we obtained 24
142 chromosome-level scaffolds with length of 680.74 Mbp (99.36 % of the total length of genome) (Fig.
143 2 and Table 3).

144
145
146
147
148

149 **Table 3 Detailed results of chromosome-level scaffolding using Hi-C technology.**

Chromosomes	Length (Mbp)	Number of Scaffolds	GC%
Chr1	33.53	1	41.8
Chr2	32.02	1	42.1
Chr3	26.36	1	42.4
Chr4	35.91	1	41.7
Chr5	31.40	1	41.9
Chr6	34.04	1	41.8
Chr7	31.19	1	42.2
Chr8	31.34	1	42.1
Chr9	30.72	1	41.8
Chr10	22.41	1	42.3
Chr11	31.32	1	41.9
Chr12	26.68	1	41.9
Chr13	28.11	1	42
Chr14	22.16	1	42.7
Chr15	28.96	1	42.1
Chr16	25.62	1	42.6
Chr17	31.84	1	42
Chr18	31.74	1	41.8
Chr19	24.96	1	42.3
Chr20	24.49	1	42.3
Chr21	28.39	1	42
Chr22	26.17	1	42.1
Chr23	25.35	1	42.6
Chr24	16.04	1	43
Total	680.74	24	-

150
151 **Gene annotation**

152 To obtain a fully annotated *A. latus* genome, three different approaches were employed to predicted
153 protein-coding gens. *Ab initio* gene prediction was performed on the repeat-masked *A. latus* genome
154 assembly using Augustus (version 3.3.1)¹⁷ and GeneMark-ES (version 4)¹⁸. Furthermore, homology-
155 based prediction was performed using protein sequences of two common model species (*Danio rerio*

and *Nile tilapia*) and two related species (*Sparus aurata* and *Larimichthys corcea*). Subsequently, these protein sequences were mapped onto the generated assembly using GeMoMa (version 1.6.1)¹⁹. In addition, transcriptome-based prediction was also applied by RNA-seq data. The RNA-seq reads were mapped onto the genome assembly using STAR (version 2.7.3a)²⁰, and the structures of all transcribed genes were predicted by Stringtie (version 1.3.4d)²¹ with the default parameters. Based on the results of Stringtie, PASA (version 2.3.3)²² was used to predicted genes of genome of *A. latus*. The predicted gene sets, generated from these three approaches, were integrated to produce a non-redundant gene set using EvidenceModeler (version 1.1.1)²³. As a result, a total of 29,227 protein-coding genes were predicted. The average number of exons per gene was 8.51. The average CDS length was 1,470.37 bp.

Gene function annotations were conducted against the NCBI nr and SwissProt protein databases, and homologs were called with E values of $< 1 \times 10^{-5}$. The functional classification of Gene Ontology (GO) categories was performed using the InterProScan program (version 5.3.2)²⁴. Blastp (version 2.7.1)²⁵ were performed to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Eukaryotic Orthologous Groups of protein (KOG) annotation analysis. As a result, a total of 22,485 genes were annotated (Fig. 3 and Table 2).

Repetitive element characterization

Two software: GMATA (version 2.2)²⁶ and Tandem Repeats Finder (version 4.07b)²⁷ were employed to detect tandem repeats (TRs) in assembly of yellowfin seabream. In addition, tandem repeats were masked before searching transposable elements (TEs) to avoid the conflicts between TRs and TEs. A MITE database was constructed, based on masked TR genome, using MITE-hunter²⁸. Meanwhile, a long terminal repeat (LTR) database was obtained by LTR_FINDER²⁹ and LTRharvest³⁰. Next, these two databases were combined into a TE library (TE.lib) and repeat sequences of genome were marked again. A *de novo* repeat sequence library (RepMod.lib) was generated with RepeatModeler³¹ and TEclass³². TE.lib, RepMod.lib and Repbase were then integrated as a non-redundant repeat sequence library (nrRep.lib). Subsequently, *A. latus* genome was annotated with nrRep.lib using RepeatMasker (version 1.331, <http://repeatmasker.org>) to search repeat sequences.

Combining the annotation results of TRs and TEs, ~21.24% sequences of the *A. latus* genome were identified as repetitive elements, including 0.51% long terminal repeats (LTRs), 1.68% long interspersed nuclear elements (LINEs), 0.1% short interspersed nuclear elements (SINEs) and 5.49% of DNA transposons, (Fig. 3 and Table 4).

Table 4 Detailed classification of repeat sequences. Note: “Unknown” represents transposable elements that could not be classified by RepeatMasker.

Type	Number of elements	Length of sequence (bp)	Percentage of sequence (%)
LTR	8,942	3,512,788	0.51
LINE	44,400	11,480,725	1.68
SINE	6,094	686,582	0.10
DNA	206,827	37,633,093	5.49
Tandem Repeats	188,984	7,124,823	1.04
Unknown	508,488	81,118,285	11.84
Other	17,306	3,984,546	0.58
Total Repeats	981,041	145,540,842	21.24

Data Records

All sequencing data, including Illumina short reads, PacBio long reads, Hi-C reads were submitted to the NCBI Sequence Read Archive (SRA) database under Bioproject accession PRJEB40702. The assembled genome was deposited at DDBJ/ENA/GenBank under the accession GCA_904848185.1.

Technical Validation

The completeness and accuracy of the assembly further assessed in multiple ways. First, the Illumina reads were re-mapped onto the assembly using BWA. As a result, 99.57% of the reads were accurately mapped with a coverage of 99.70%. Subsequently, Benchmarking Universal Single-Copy Orthologs (BUSCO) software (version 3.0.1)³³ was executed using actinopterygii_odb9 database to assess the predicted gene set. The result showed that 98.3% of all 4584 BUSCOs were assembled, including 97.50% and 0.80% of all BUSCOs were completely and partially assembled, also implying a high level of completeness for the *de novo* assembly (Table 5).

Table 5 Details of accuracy and completeness validation of genome assembly.

Illumina Reads Mapping

Mapping ratio	97.09%
Mapping coverage	99.89%

BUSCO	Number	Proportion (%)
Total BUSCO groups searched	4,584	100.00
Complete BUSCOs (C)	4,469	97.50
Fragmented BUSCOs (F)	37	0.80
Missing BUSCOs (M)	78	1.70

Code availability

No custom computer codes were generated in this work.

References

- 1 Nelson, J. S., Grande, T. C. & Wilson, M. V. *Fishes of the World*. (John Wiley & Sons, 2016).
- 2 Pla, S., Benvenuto, C., Capellini, I. & Piferrer, F. A phylogenetic comparative analysis on the evolution of sequential hermaphroditism in seabreams (Teleostei: Sparidae). *Scientific Reports* **10**, 3606, doi:10.1038/s41598-020-60376-w (2020).
- 3 Iwatsuki, Y. Review of the *Acanthopagrus latus* complex (Perciformes: Sparidae) with descriptions of three new species from the Indo-West Pacific Ocean. *Journal of fish biology* **83**, 64-95 (2013).
- 4 Abol-Munafi, A. & UMEDA, S. The gonadal cycle of the yellowfin porgy, *Acanthopagrus latus* (Houttuyn) reared in the net cage at Tosa Bay, Japan. *Aquaculture Science* **42**, 135-144 (1994).
- 5 Buxton, C. D. & Garratt, P. A. Alternative reproductive styles in seabreams (Pisces: Sparidae). *Environmental Biology of Fishes* **28**, 113-124, doi:10.1007/BF00751031 (1990).
- 6 Abu-Hakima, R. Some aspects of the reproductive biology of *Acanthopagrus* spp.(Family: Sparidae). *Journal of fish biology* **25**, 515-526 (1984).
- 7 Syazni, K. A., Tomano, S., Ueno, K., Ohara, K. & Umino, T. J. A. S. Genetic structure of yellowfin black seabream *Acanthopagrus latus* in western Japan based on microsatellite and mtDNA marker analyses. **63**, 17-27 (2015).
- 8 Karimi, S., Kochinian, P. & Salati, A. J. I. J. o. V. R. The effect of sexuality on some haematological parameters of the yellowfin seabream, *Acanthopagrus latus* in Persian Gulf. **14**, 65-68 (2013).
- 9 Li, S. *et al.* Identification and Comparison of microRNAs in the Gonad of the Yellowfin Seabream (*Acanthopagrus Latus*). **21**, 5690 (2020).
- 10 Li, S. *et al.* Gonadal Transcriptome Analysis of Sex-Related Genes in the Protandrous Yellowfin Seabream (*Acanthopagrus latus*). **11**, doi:10.3389/fgene.2020.00709 (2020).
- 11 Chen, S., Zhou, Y., Chen, Y. & Jia, G. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
- 12 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770, doi:10.1093/bioinformatics/btr011 (2011).
- 13 Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050-1054, doi:10.1038/nmeth.4035 (2016).
- 14 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
- 15 Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119-1125, doi:10.1038/nbt.2727 (2013).
- 16 Jiahan, L. & Lisha, L. Studies on karyotype of *Sparus latus* Houttuyn. *J. Oceanogr. Taiwan Strait* **8**, 162-166 (1989).
- 17 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644, doi:10.1093/bioinformatics/btn013 (2008).
- 18 Alexandre, L., Vardges, T. H., Chernoff, Y. O. & Mark, B. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 20 (2005).
- 19 Keilwagen, J. *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**, e89-e89, doi:10.1093/nar/gkw092 (2016).
- 20 Valencia, A. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (2014).
- 21 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290-295, doi:10.1038/nbt.3122 (2015).
- 22 Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654-5666, doi:10.1093/nar/gkg770 (2003).
- 23 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7, doi:10.1186/gb-2008-9-1-r7 (2008).
- 24 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).
- 25 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410, doi:[https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) (1990).

- 26 Wang, X. & Wang, L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. *Frontiers in Plant Science* **7**, doi:10.3389/fpls.2016.01350 (2016).
- 27 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
- 28 Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research* **38**, e199-e199, doi:10.1093/nar/gkq862 (2010).
- 29 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265-W268, doi:10.1093/nar/gkm286 (2007).
- 30 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* **9**, 18, doi:10.1186/1471-2105-9-18 (2008).
- 31 Bedell, J. A., Korf, I. & Gish, W. MaskerAid : a performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040-1041, doi:10.1093/bioinformatics/16.11.1040 (2000).
- 32 Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329-1330, doi:10.1093/bioinformatics/btp084 (2009).
- 33 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

Acknowledgements

This work was supported by Guangzhou Science and Technology Project [No. 201803020017], National Natural Science Foundation of China [No. 91858208], [No. 31902427], R&D Project for Jinwan Yellowfin Seabream Breeding System Construction [No. K20-42000-018], Science and Technology Project of Zhanjiang [No. 2019A03011], and Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) [No. 311020005].

Author Contributions

J. L. conceived the study. D. G., W. F. and J. X. collected the samples. D. G. and G. L. extracted the genomic DNA. D. G., Y. S., J. T., J. C. and K. H. assembled and curated the genome. D. G. and J. L. wrote, reviewed and edited the manuscript.

Competing Interests

The authors declare no competing interests.

Figures

Fig. 1

Fig. 2

395 **Fig. 3**

Figure Legends

Fig. 1 A healthy female yellowfin seabream was collected.

Fig. 2 Genome-wide all by all Hi-C interaction of *A. latus*.

Fig.3 Circos plot showing 24 chromosomes of *A. latus*. Note: chromosome length in Mb unit; the blue histogram presents the GC content for 1 Mbp window; heatmap of repeats density within 1 Mbp window, ranging from 27 to 2,198 repeat sequences per million base pairs; line plot of gene density for 1 Mbp windows and the innermost pie chart displays the completeness validation results of genome assembly using BUSCO.