GigaScience
Draft Genome of the Protandrous Chinese Black Porgy, Acanthopagrus schlegelii
--Manuscript Draft--

Manuscript Number: GIGA-D-17-00137R2
Full Title: Draft Genome of the Protandrous Chinese Black Porgy, Acanthopagrus schlegelii
Article Type: Data Note

Funding Information:

<table>
<thead>
<tr>
<th>Funding Information</th>
<th>Dr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aquatic Sanxin Engineering Major Project of Jiangsu Province (D2015-17)</td>
<td>Zhiyong Zhang</td>
</tr>
<tr>
<td>Jiangsu Innovation Ability Construction Program (BM2015017)</td>
<td>Zhiyong Zhang</td>
</tr>
<tr>
<td>Nantong Applied Basic Research Program (MS12015071)</td>
<td>Zhiwei Zhang</td>
</tr>
<tr>
<td>Key Research and Development (Modern Agriculture) Program of Jiangsu Province (BE2016326)</td>
<td>Zhiwei Zhang</td>
</tr>
<tr>
<td>Aquatic Sanxin Engineering Project of Jiangsu Province (Y2016-23)</td>
<td>Zhiyong Zhang</td>
</tr>
<tr>
<td>Nantong Applied Basic Research Program (MS12015070 &amp; MS12016029)</td>
<td>Zhiyong Zhang</td>
</tr>
<tr>
<td>Zhenjiang Leading Talent Program for Innovation and Entrepreneurship</td>
<td>Qiong Shi</td>
</tr>
</tbody>
</table>

Abstract: Background: As one of the most popular and valuable commercial marine fishes in China and East Asian countries, the Chinese black porgy (Acanthopagrus schlegelii) also known as the blackhead seabream, has some attractive characteristics, such as fast growth rate, good meat quality, resistance to diseases and excellent adaptability to various environments. Furthermore, the black porgy is a good model for investigating sex changes in fish due to its protandrous hermaphrodite. Here, we obtained a high-quality genome assembly of this interesting teleost species and performed a genomic survey on potential genes associated with the sex-change phenomenon.

Findings: We generated 175.4 gigabases (Gb) of clean sequence reads using a whole-genome shotgun sequencing strategy. The final genome assembly is approximately 688.1 megabases (Mb), accounting for 93% of the estimated genome size (739.6 Mb). The achieved scaffold N50 is 7.6 Mb, reaching a relatively high level among sequenced fish species. Meanwhile, we identified 19,465 protein-coding genes, which had an average transcript length of 17.3 kb. By performing a comparative genomic analysis, we found three types of genes potentially associated with sex change, which are useful for the prediction of related genetic basis for the interesting protandrous hermaphrodite.

Conclusions: We provided a draft genome assembly of the Chinese black porgy and discussed about the potential genetic mechanisms of sex change. These data are also an important resource for studying the biology and facilitating the molecular breeding of this economically important fish.

Corresponding Author: Qiong Shi, PhD
BGI
Shenzhen, CHINA

Corresponding Author's Institution: BGI

First Author: Qiong Shi, PhD

Order of Authors: Qiong Shi, PhD
Dear Editor,

Thanks for your kind help. We also appreciate the instructive comments from the two reviewers. According to their suggestions, we performed additional analyses and made a careful revision on our previous manuscript. Our point-by-point responses are attached for your consideration.

Please also include the details on your BUSCO analysis in the manuscript (and we will also ask you to upload BUSCO results to GigaDB). Why was Busco v2 used, when the latest version is v3?

Answer: Thanks for your good advice. We also ran BUSCO v3 on the genome assembly, and the final BUSCO score was 89.1% (C:89.1% [S:86.2%, D:2.9%], F:2.5%, M:8.4%, n:4584). For the predicted gene set, the new BUSCO score reached 85.5%, (C:85.5% [S:82.3%, D:3.2%], F:2.8%, M:11.7%, n:4584). Please find more details about these revisions on lines 118-121 and 148-150 in the revised manuscript. In addition, we uploaded corresponding BUSCO results to the GigaDB.

Highlighting revisions to the manuscript in color can help us and the reviewers to quickly assess your next version.

Answer: Yes, it’s done. We highlighted all the revisions in yellow for your convenience.

Finally, please note that GigaScience allows a maximum number of 3 indications of co-first authorship and a maximum number of 2 co-corresponding authors. Please revise accordingly.

Answer: Thanks for your comments. We revised the authorship information in accordance with your instructions. Three co-first authors (Prof. Zhang, Mr. Zhang and
Dr. Chen) and two co-corresponding authors (Profs. Xu and Shi) are marked in the revised manuscript.

Best regards,
Qiong Shi, PhD, Professor
BGI
Shenzhen 518083
China

P.S. point-by-point responses to the reviewers’ comments.

Reviewer #1: I thank the authors for addressing my comments in a satisfactory fashion. I have a couple of comments, more regarding choice of words. Lines 111-112, and elsewhere in the manuscript: I am unsure if I would have chosen to discuss the N50 length of the scaffolds as "reaching a relatively high level among sequenced fish species". Do you mean length instead of level?

Answer: Thanks for your question and advice. You are right. We changed “level” to “length” on line 110 of the revised manuscript.

While the manuscript discusses the manner of whole-genome duplication well, I found the wording in the answer a bit strange: "Black porgy is a diploid species, hence multiple copies of wnt4, vasa and JNK1 in the genome assembly may be resulted from the teleost-specific whole genome duplication." Most gene assemblers output a haploid assembly, based on a diploid genome, but sometimes both haplotypes are output. Do you mean that this has happened in this case? Or that the multiple copies have been retained since the whole-genome duplication in the ancestor to the teleosts?

Answer: Thanks for your good questions. Yes, you are right. The multiple copies may be retained since the whole-genome duplication in the ancestor to the teleost. Please find corresponding changes on lines 222-223 in the revised manuscript.

Reviewer #2: The revised version has been improved, but several of the reviewer comments have not been addressed appropriately. I therefore encourage the authors to take the following points into full consideration when revising the current version of the study:

1. Comparison to other fish genome assemblies: I think that the assembly should only be compared to other teleost fishes. Elephant shark (cartilaginous fish) and coelacanth (lobe-finned fish), although being ‘fish’ in the broader sense, are no more justifiable to compare to than tetrapods.

Answer: Thanks for your comments and advice. We removed elephant shark and coelacanth in the revised manuscript. Please find more details on lines 111-114 in the revised manuscript.

2. BUSCO analysis: You mention that you used the BUSCOv2 actinopterygii dataset in the response to the reviewers, but this information needs to be included in the manuscript as well.

Answer: Thanks for your advice. According to the reviewer one’s opinions, we updated BUSCO v2 to v3, although the final results are similar. We hence added the updated information to the revised manuscript (on lines 118-121 and 148-150 of the revised manuscript).

3. Phylogenetic analysis:
3.1 The phylogenetic tree obtained with MrBayes should be included as a supplementary figure.

Answer: Thanks for your advice. However, please be informed that a Data Note paper do not provide space for any supplementary document. We rebuilt the phylogenetic tree using MrBayes (Version 3.2, with the GTR+gamma model), which is the same in topology as our Figure 2 in the main content.
3.2 In the response to the reviewers you mention that you removed the statement of a ‘close relationship’ of black porgy to fugu, but it is still in the manuscript (l. 174).

Answer: Sorry for the mistake. We removed it now in the new manuscript. Please see more details on lines 172-173 of the revised manuscript.

3.3 Furthermore, in the revised version it is newly stated that the phylogenetic tree ‘suggests a lower neutral evolutionary rate than any other investigated teleost’ (l. 175). It is unclear how this statement is substantiated by the data presented. Is this referring to the branch lengths in the phylogenetic tree? Porgy and tilapia do not seem to be very different. The conclusion of a specifically slow rate of molecular evolution in black porgy has to be supported with a statistical tests such as relative rate and two-cluster tests.

Answer: Thanks for the nice comments and advice. Yes, you are right. We hence removed the conclusions due to the lack of sufficient evidence in the revised manuscript. Please see more details on lines 172-173.

4. Survey of sex related genes: This part still needs major improvement.

4.1. First of all, it still remains unresolved why there are multiple copies of many of these genes in the genome assembly. In the response to reviewer 1, it is stated that ‘… hence multiple copies of wnt4, vasa and JNK1 in the genome assembly may be resulted from the teleost-specific whole genome duplication.’; in response to reviewer 2: ‘Several genes, such as Wnt4, vasa and JNK1, with multiple copies in the assembly may be copy number variants.’ As mentioned in the previous reviews, this really needs to be addressed in more detail. For example, the table states that there are 15 copies of wnt4 in porgy. As the authors correctly state, there is wnt4a and wnt4b in vertebrates, as well as potential additional paralogs from the teleost genome duplication etc. All genes reported here need to be analyzed by phylogenetic methods to appropriately establish their orthology to other fishes and vertebrates. While I appreciate that the authors are in the process of cloning these individual copies, their sequences from the genome assembly need be submitted to NCBI and/or provided as supplements as part of the current article.

Answer: Thanks for your nice comments and advice. Sorry for the misleading statements. In fact, the multiple copies of wnt4, vasa and jnk1 in the genome assembly may be retained since the whole-genome duplication in the ancestor to the teleost. Please find more details of the corrections on lines 222-249 of the revised manuscript. According to your advice, we employed all the predicted protein sequences to rebuild a new phylogenetic tree using PhyML, and we still observed that they were clustered with each corresponding homologue from other vertebrates (data not shown). Please find corresponding changes on lines 186-188 of the revised manuscript. We also uploaded these data, including the ML tree and the predicted coding sequences and protein sequences of these genes, to the GigaDB.

4.2 Secondly, as mentioned in my previous report, the reasoning for the existence of a putative sex chromosome in black porgy needs to be supported beyond the fact that some of the sex-related genes are located on the same scaffold. Is the linkage of these specific genes in the black genome assembly unique or are these general blocks of conserved synteny across teleosts? The authors state: "our data demonstrate that the distribution of these 3 types of genes in the black porgy genome is similar to that in Chinese ricefield eel" but give no further explanation in which ways the porgy is similar to the rice eel. Again, if this refers to linkage of specific genes, it would need to be demonstrated that the linkage of these genes is unique to both porgy and ricefield eel. In their response to the reviewer, the authors further mention that the data are similar to arowana, which has identified sex chromosomes, but give no further explanation in which ways porgy and arowana are similar to support the conclusion that a sex chromosome might be present in porgy. Without further details, such conclusion remains unsubstantiated and should be removed.

Answer: Thanks for your nice comments and advice. The linkage of these specific genes in the black genome assembly seems to be unique. However, we don’t have solid evidence to support existence of a sex chromosome in the porgy. We hence
removed these conclusions in the revised manuscript (on lines 243-249). Meanwhile, more details about the similarity to ricefield eel were provided on lines 246-248 in the revised manuscript.

5. It would be better to mention the information on medaka dmrt1 after the statement on tongue sole dmrt1 (l. 198).

Answer: Thanks for your advice. It's done on lines 200-202 of the revised manuscript.

6. Gene duplicates should be called paralogs, not 'isotypes'.

Answer: Sorry for our mistake. We corrected it on line 217 in the revised manuscript.

7. Citations in the later part of the main text seem to be out of order (starting at about citation 48 and following). Please double-check that the citations in the text and the bibliography correspond to each other.

Answer: Thanks for your careful checking. We corrected the citation orders in the revised manuscript.

8. Throughout the manuscript and tables, again, please follow the gene nomenclature conventions for teleost fish (see https://wiki.zfin.org/display/general/ZFIN+Zebrafish+Nomenclature+Guidelines).

Answer: Thanks for your comments. It's done for all the gene names.

Additional Information:

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are you submitting this manuscript to a special series or article collection?</td>
<td>No</td>
</tr>
<tr>
<td><strong>Experimental design and statistics</strong></td>
<td>Yes</td>
</tr>
<tr>
<td>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</td>
<td></td>
</tr>
<tr>
<td>Have you included all the information requested in your manuscript?</td>
<td></td>
</tr>
<tr>
<td><strong>Resources</strong></td>
<td>Yes</td>
</tr>
<tr>
<td>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</td>
<td></td>
</tr>
<tr>
<td>Have you included the information requested as detailed in our Minimum</td>
<td></td>
</tr>
<tr>
<td>Standards Reporting Checklist?</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>Availability of data and materials</td>
<td>Yes</td>
</tr>
</tbody>
</table>

All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?
Draft Genome of the Protandrous Chinese Black Porgy, *Acanthopagrus schlegelii*

Zhiyong Zhang†, Kai Zhang2,3,4†, Shuyin Chen1†, Zhiwei Zhang1, Jinyong Zhang5,
Xinxin You3, Chao Bian3,6, Jin Xu1, Chaofeng Jia1, Jun Qiang2, Fei Zhu1, Hongxia Li2,
Hailin Liu1, Dehua Shen1, Zhonghong Ren1, Jieming Chen3, Jia Li3, Tianheng Gao7,
Ruobo Gu3,6, Junmin Xu3,6, Qiong Shi3,4,6*, Pao Xu2*

1 Jiangsu Marine Fishery Research Institute, Nantong, Jiangsu 226007, China
2 Freshwater Fishery Research Center, Chinese Academy of Fishery Sciences, Wuxi, Jiangsu 214081, China
3 Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, Shenzhen, Guangdong 518083, China
4 BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, Guangdong 518083, China
5 State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei 430000, China
6 BGI-Zhenjiang Institute of Hydrobiology, Zhenjiang, Jiangsu 212000, China
7 College of Oceanography, Hohai University, Nanjing, Jiangsu 210098, China

* Correspondence address. Pao Xu, Freshwater Fishery Research Center, Chinese Academy of Fishery Sciences, Wuxi, Jiangsu 214081, China (tel: +86-138 0619 0669; email: xup@ffrc.cn); Qiong Shi, Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI, Shenzhen 518083, China (tel: +86-185 6627 9826; email: shiqiong@genomics.cn)

† Contributed equally to this work.
Abstract

Background: As one of the most popular and valuable commercial marine fishes in China and East Asian countries, the Chinese black porgy (*Acanthopagrus schlegelii*) also known as the blackhead seabream, has some attractive characteristics, such as fast growth rate, good meat quality, resistance to diseases and excellent adaptability to various environments. Furthermore, the black porgy is a good model for investigating sex changes in fish due to its protandrous hermaphrodite. Here, we obtained a high-quality genome assembly of this interesting teleost species and performed a genomic survey on potential genes associated with the sex-change phenomenon.

Findings: We generated 175.4 gigabases (Gb) of clean sequence reads using a whole-genome shotgun sequencing strategy. The final genome assembly is approximately 688.1 megabases (Mb), accounting for 93% of the estimated genome size (739.6 Mb). The achieved scaffold N50 is 7.6 Mb, reaching a relatively high level among sequenced fish species. Meanwhile, we identified 19,465 protein-coding genes, which had an average transcript length of 17.3 kb. By performing a comparative genomic analysis, we found three types of genes potentially associated with sex change, which are useful for the prediction of related genetic basis for the interesting protandrous hermaphrodite.

Conclusions: We provided a draft genome assembly of the Chinese black porgy and discussed about the potential genetic mechanisms of sex change. These data are also an important resource for studying the biology and facilitating the molecular breeding of this economically important fish.

Keywords: Chinese black porgy; *Acanthopagrus schlegelii*; whole genome sequencing; genome assembly; sex-change related genes

Data description

Background information

As one of the most popular and valuable commercial marine fishes in China and East
Asian countries, the Chinese black porgy (Acanthopagrus schlegelii), also known as the blackhead seabream, has some interesting characteristics, such as fast growth rate, good meat quality, resistance to diseases and good adaptability to various environments. It is often farmed for food in the South China Sea and the coastal waters around Japan and Korea [1,2]. In addition, it is an eurythermal and euryhaline fish, living in a wide range of water temperatures and salinities. Recently, some basic studies on the genetic improvement for its growth and disease resistance have been increasingly performed in order to increase efficiency of farming [3].

The Chinese black porgy is also a good model for investigating the genetic mechanisms of sex change due to its interesting life cycle. It is a functional male during the first 2 years and a subsequent female during the next couple of years. Recently, a good hybrid of the Japanese seabream (Pagrosomus major; ♀) and the Chinese black porgy (♂) has become available [4,5], with better growth performance and higher tolerance against low temperature than its parents. However, related genetic mechanisms for these interesting biological characteristics are still unclear.

Here, we sequenced and assembled the whole genome of the Chinese black porgy, before performing a genomic survey on potential genes associated with the sex-change phenomenon.

### Sample and Sequencing

The wild black porgy (NCBI Taxonomy ID: 72011; Fishbase ID: 6531) individuals (Figure 1) were collected from Laizhou Bay in Yantai, Shandong Province, China. Genomic DNA was extracted from the muscle of a female fish using Qiagen GenomicTip100 (Qiagen, Hilden, USA). We employed the whole-genome shotgun sequencing strategy and constructed the subsequent three short-insert libraries (250-bp, 500-bp and 800-bp) and four long-insert libraries (2-kb, 5-kb, 10-kb and 20-kb) in accordance with the standard protocol from Illumina (San Diego, USA). All these constructed libraries were sequenced on the Illumina HiSeq 2000 system [6] (the read length is 125 bp). Finally, we generated a total of 272.9-Gb raw reads from all seven libraries.
Before assembly of the sequencing reads, SOAPfilter v2.2 software [7] with default parameters (-y -p -g 1 -o clean -M 2 -f 0) was utilized to remove low-quality raw reads (including reads with 10 or more non-sequenced/low-quality bases), PCR duplicates and adaptor sequences. Subsequently, we obtained approximately 175.4 Gb of clean reads for further genome size prediction and assembling. A $k$-mer analysis with the formula $G = \frac{k_{\text{num}}}{k_{\text{depth}}}$ [8] was performed to estimate the genome size of Chinese black porgy. In our current study, the achieved total number of $k$-mers and $k_{\text{depth}}$ was $2.81 \times 10^{10}$ and 38, respectively. Therefore, the genome size of Chinese black porgy is estimated to be 739.6 Mb. Based on this result, the retained reads were calculated to cover approximately 238-fold of the whole genome.

**Assembly and Evaluation**

To obtain a genome assembly, we employed the SOAPdenovo2 v2.04.4 [9] with optimized parameters (pre-graph -K 27 -p 16 -d 1; contig -M 3; scaffold -F -b 1.5 -p 16) using these clean reads. In brief, the reads from short-insert libraries were applied for the contig assembly, before alignment of all the filtered reads onto the contigs for linking these contigs to generate scaffolds. GapCloser v1.12 software [7] with default parameters was subsequently used to fill some intra-scaffold gaps in the local assembly, in which the reads were equipped with one end uniquely mapped to a contig and the other end located within a gap. Meanwhile, SSPACE (version 2.0) [10] with default parameters was employed to obtain super scaffolds with the reads from the long-insert libraries (2-kb, 5-kb, 10-kb and 20-kb). The final genome assembly was approximately 688.1 Mb, which accounts for 93.0% of the estimated genome size (739.6 Mb; Table 1).

The achieved scaffold N50 is 7.64 Mb, reaching a relatively high length among sequenced fish species. In comparison, other scaffolds have levels of 1.55 Mb for the zebrafish [11], 1.1 Mb for platy fish [12], 867 kb for half-smooth tongue sole [13], 1 Mb for common carp [14], 6.4 Mb for grass carp [15], 2.97 Mb for Atlantic salmon [16], 1.8 Mb for a seahorse [17] and 1.15 Mb for a Chinese barbel fish [18]. Core Eukaryotic Genes Mapping Approach (CEGMA; version 2.5) [19] with a set of 248
conserved Core Eukaryotic Genes (CEGs) was employed to assess the completeness of the final assembly. The estimates suggest that 90.7% CEGs are complete and 92.3% are partial. Meanwhile, Benchmarking Universal Single-Copy Orthologs (BUSCO: version 3) [20] was applied to evaluate the quality of the generated genome assembly. The final BUSCO score reached 89.1%, (C:89.1% [S:86.2%, D:2.9%], F:2.5%, M:8.4%, n:4584). These results from CEGMA and BUSCO suggested that the assembled genome covers the majority of the gene space.

Annotation

We used RepeatProteinMask (version 4.0.6) [21] in RepeatMasker to identify the repetitive sequences, before employing RepeatModeller (version 1.05) [22] and LTR_FINDER.x86_64-1.0.6 to construct a de novo repeat library. Additionally, repetitive elements were predicted using Tandem Repeat Finder (TRF, version 4.04). Finally, we observed that the identified repeat sequences cover 19.78% of the assembled genome (Table 2). Prediction of protein-coding genes was performed based on the integration of ab initio prediction, homologue prediction and transcriptome-based prediction. The ab initio prediction was carried out with Augustus (version2.5) [23] and GENSCAN (version1.0) [24] on the repeat-masked assembly. For the homology-based gene prediction, homologous proteins of several reported fishes (zebrafish, Japanese puffer, stickleback and medaka) were downloaded from Ensembl release 75 and aligned to the assembled genome using tBlastn (version2.2.19) with e-value ≤ 1e-5. Subsequently, all the achieved alignments were analyzed using Genewise (version2.2.0) software [25] to search for precise gene structures. We further filtered out these short (less than 150 bp), prematurely terminated or frame-shifted genes. For the transcriptome-based prediction, we obtained transcriptome data from a mixture of liver, muscle, skin, gill and brain of a female fish at cDNA level. Those with low-quality bases, adapter sequences and duplicated sequences were removed and we acquired approximately 8 Gb of high-quality clean reads. Subsequently, TopHat2.1.1 [26] and Cufflinks (version 2.2.1) [27] were applied to predict gene structures using
these retained reads. Eventually, the three gene sets generated from the prediction approaches were integrated into a comprehensive and non-redundant gene set using GLEAN [28]. As summarized in Table 1, the final gene set contains 19,465 genes, with an average transcript length of 17.3 kb. In addition, we ran BUSCO v3 [20] on the predicted coding sequences (CDS), and the final BUSCO score was up to 85.5%, (C:85.5% [S:82.3%, D:3.2%], E:2.8%, M:11.7%, n:4584).

Simultaneously, all the protein sequences from the GLEAN analysis were mapped onto several public databases, including Pfam [29], PRINTS [30], ProDom [31] and SMART [32], to detect the known motifs and domains within our genome assembly. The data demonstrated that 99.3% of the predicted genes from the assembled genome contain at least one related functional assignment from other public databases, including Swiss-Prot [33], Interpro [34], TrEMBL [35] and KEGG [36].

**Phylogenetic Analysis**

In order to examine the phylogenetic position of the Chinese black porgy, we downloaded protein sequences of seven reported fishes, including spotted gar (*Lepisosteus oculatus*), stickleback (*Gasterosteus aculeatus*), Japanese fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), zebrafish (*Danio rerio*), platyfish (*Xiphophorus maculatus*), and Nile tilapia (*Oreochromis niloticus*) from Ensembl (release 83) [37]. These sequences were used to construct gene families by OrthoMCL [38] and eventually generated a total of 17,431 gene families by the all-to-all BLASTP strategy with an E-value of 1e⁻⁵. In additional, 65 gene families were only presented in the black porgy genome.

Subsequently, 3,239 single-copy orthologous genes from these gene families were selected. These single-copy genes were further aligned using MUSCLE (version 3.8.31) with default parameters [39], before the protein alignments were changed to corresponding CDS using an in-house perl script. All these nucleotide sequences of each species were integrated into a supergene, which were used to build a phylogenetic tree with PhyML [40]. Our final data orientated the phylogenetic position of the black porgy in teleost (Figure 2).
Analysis of Three Types of Genes for Sex Change

Sex change (secondary sex determination) is a universal phenomenon in fish, but it usually does not occur in amphibians and mammals. The black porgy is a good model for the study on the molecular mechanisms of sex change. For providing a genomic survey on these genes in the assembled genome, protein sequences of three main types of genes potentially associated with sex change, including sex determination and differentiation genes, pluripotency factors and apoptosis factors [41–43], were downloaded from the NCBI database and used for homology searches against the black porgy genome with tBlastn (version2.2.19) [44]. We chose alignments with coverage > 70% and identity > 70% for further prediction of gene structures using Genewise (version 2.2.0) [25]. Finally, we obtained homologous sequences of 26 genes in the genome assembly of Chinese black porgy (see more details in Table 3).

All these predicted protein sequences were employed to build a phylogenetic tree using PhyML [40], and we eventually observed that they were clustered with each corresponding homologue from other vertebrates.

Previous studies have revealed that multiple genes, including dmrt1, cyp19a1a, wnt4, sox9, sf1, foxl2, figla, amhr2 and dax1, are associated with sex change in the black porgy [41,45-47]. These sex determination and differentiation genes were also identified in our assembled scaffolds (in the first batch of Table 3). In the current study, the important male-related dmrt1 and the steroidogenesis-suppressing factor dax1 were mapped on the scaffolds 56 and 14 of the black porgy genome, respectively.

It was reported that dmrt1 may play a key role in the sex change of the black porgy, while the male-phase maintenance of male development was regulated by the brain–pituitary–gonadal axis via the GnRH-GtH-Dmrt1 pathway [41]. In the economically important half-smooth tongue sole (Cynoglossus semilaevis), dmrt1 has been proven to be a necessary male sex-determining gene [48,49]. Moreover, previous findings suggest that a duplicate of dmrt1 is the male sex determinant in medaka and dmrt1 mutation causes a male–female sex reversal [50,51]. We also validated the existence...
of *foxl2* and *cyp19a1a*, two putative female-related genes, in the black porgy genome. Previous findings revealed that *cyp19a1a* plays dual roles in gonadal development, while both *cyp19a1a* and *foxl2* are related to the sex change of the black porgy [47]. However, *foxl2* has functions in sex differentiation, but it is not essential for sex determination and sex change in the tongue sole [52]. *figla*, with only one copy in the black porgy, is a germ-cell-specific transcription factor related to ovary development and differentiation [53]. However, two isoforms (figla_tv1 and figla_tv2) were reported in the tongue sole. *figla_tv1* possesses a conserved function in folliculogenesis as found in other vertebrates, while *figla_tv2* may play a role in the spermatogenesis of pseudo-males by regulating the synthesis and metabolism of steroid hormones [53]. *sf1*, also identified with one gene in the black porgy (*Table 3*), was reported to act as an essential transcriptional factor for steroidogenesis and for development of the reproductive axis [54].

Interestingly, five copies of *sox9* were also identified in the black porgy genome. Nevertheless, previous findings reported that only 2 paralogs of *sox9* (*sox9a* and *sox9b*) are present in zebrafish [55] and catfish [56]. *Sox9a* is usually associated with testicular development, while this may be linked with sex reversal in the tongue sole [52]. In comparison, *sox9b* possesses a new function in the ovary [55]. In addition, we noticed that female-related genes (*wnt4*, *vasa* and *jnk1*) have multiple copies in our current study, which may be retained since the whole-genome duplication in the ancestor to the teleost. These genes have been proven to play important roles in ovarian growth and natural sex changes in fishes [57–60]. It was reported that two *wnt4* genes (*wnt4a* and *wnt4b*) are present in most teleost fish, while other vertebrates and invertebrates possess only a single *wnt4* gene. Furthermore, two copies of the *wnt4a*, *wnt4a1* and *wnt4a2*, exist in some teleost species resulting from the additional duplication of *wnt4* gene [61]. It has been shown that *wnt4a* was mainly expressed in the gonad, gill and brain of teleost fish (such as zebrafish [62] and rainbow trout [63]), and it was confirmed to be associated with sex reversal in the tongue sole [61]. The *vasa* gene, also called *ddx4*, was reported to
play an important role in gametogenesis and germ cell development [64]. Previous findings showed that vasa was a single copy gene in the majority of chordates, such as zebrafish [65,66]. However, 3 vasa genes were also reported in Nile tilapia (Oreochromis niloticus) [67]. Ink1 is closely associated with ovarian differentiation and development in fish. A previous finding [58] reported that jnk1 highly transcribed in the ovary of the female ricefield eel (Monopterus albus), another teleost with natural sex-change from female to male, and reduced to a substantial level at the subsequent stage of intersex; hence, the data demonstrated that jnk1 may play a key role in the sexual reversal. Surprisingly, two jnk1 genes (jnk1a and jnk1b) were reported in the polyploid hybrids of red crucian carp (Carassius auratus red var.) and common carp (Cyprinus carpio L.) [68].

Interestingly, our data demonstrate that the distribution of these 3 types of genes in the black porgy genome is similar to that in ricefield eel (our unpublished results; Data of the Monopterus Whole Genome Shotgun project have been deposited at DDBJ/EMBL/GenBank under the accession number of AONE00000000). For example, 2 male-related genes (piwi1 and piwi2) are clustered together, while lin28a and rspo1 are adjacent to each other. We also observed that most of these genes are congregated on the scaffolds 1, 2, 3, 11 and 15 (Table 3).

Conclusions

In summary, we sequenced and assembled the whole genome of Chinese black porgy. This is the first genomic report of Sparidae fish. Furthermore, we provided a genomic survey on the 26 genes potentially associated with sex change. The achieved genome data will be helpful for further biological and evolutionary studies. Furthermore, it will be valuable for implementation of molecular breeding, with substantial support from our genomic data, to obtain genetic improvement of this economically important teleost fish.
Table 1. Summary of the achieved genome assembly and annotation.

<table>
<thead>
<tr>
<th>Genome assembly</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>contig N50 size (kb)</td>
<td>17.2</td>
</tr>
<tr>
<td>contig number (&gt; 100 bp)</td>
<td>115,091</td>
</tr>
<tr>
<td>scaffold N50 size (Mb)</td>
<td>7.6</td>
</tr>
<tr>
<td>scaffold number (&gt; 100 bp)</td>
<td>31,359</td>
</tr>
<tr>
<td>Total length (Mb)</td>
<td>688.1</td>
</tr>
<tr>
<td>Genome coverage (×)</td>
<td>257.6</td>
</tr>
<tr>
<td>The longest scaffold (bp)</td>
<td>22,574,836</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome annotation</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding gene number</td>
<td>19,465</td>
</tr>
<tr>
<td>Mean transcript length (kb)</td>
<td>17.3</td>
</tr>
<tr>
<td>Mean exons per gene</td>
<td>11.1</td>
</tr>
<tr>
<td>Mean exon length (bp)</td>
<td>170.2</td>
</tr>
<tr>
<td>Mean intron length (bp)</td>
<td>1519.2</td>
</tr>
</tbody>
</table>
Table 2. Detailed classification of repeat sequences in the assembled genome.

<table>
<thead>
<tr>
<th>Type</th>
<th>Repbase TEs</th>
<th>TE proteins</th>
<th>Denovo</th>
<th>Combined TEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length (Mb)</td>
<td>In genome (%)</td>
<td>Length (Mb)</td>
<td>In genome (%)</td>
</tr>
<tr>
<td>DNA</td>
<td>20.93</td>
<td>3.041</td>
<td>2.200</td>
<td>0.320</td>
</tr>
<tr>
<td>SINE</td>
<td>1.120</td>
<td>0.163</td>
<td>2.340</td>
<td>0.000</td>
</tr>
<tr>
<td>LTR</td>
<td>7.200</td>
<td>1.046</td>
<td>35.410</td>
<td>0.340</td>
</tr>
<tr>
<td>Other</td>
<td>0.020</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Total</td>
<td>35.300</td>
<td>5.130</td>
<td>11.480</td>
<td>1.669</td>
</tr>
</tbody>
</table>
Table 3. Three types of genes potentially related to sex change in the black porgy genome

<table>
<thead>
<tr>
<th>Sex determination and differentiation genes</th>
<th>Gene</th>
<th>Copy number</th>
<th>Scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fst</td>
<td>2</td>
<td>10, 17</td>
</tr>
<tr>
<td></td>
<td>sox9</td>
<td>5</td>
<td>11, 13, 16, 19, 27</td>
</tr>
<tr>
<td></td>
<td>vasa</td>
<td>10</td>
<td>11, 14, 16, 20, 27, 34, 37, 47, 53, 68</td>
</tr>
<tr>
<td></td>
<td>ctnnb1</td>
<td>4</td>
<td>2, 16, 64, 115</td>
</tr>
<tr>
<td></td>
<td>piwill1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>piwil2</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>sf1</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>rspl</td>
<td>2</td>
<td>2, 74</td>
</tr>
<tr>
<td></td>
<td>foxl2</td>
<td>2</td>
<td>1, 22</td>
</tr>
<tr>
<td></td>
<td>cyp19a1a</td>
<td>2</td>
<td>8, 28</td>
</tr>
<tr>
<td></td>
<td>gsdf</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>figla</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>dmrt1</td>
<td>1</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>wnt4</td>
<td>15</td>
<td>1, 2, 5, 6, 7, 8, 9, 18, 19, 20, 32, 34, 62, 67, 122</td>
</tr>
<tr>
<td></td>
<td>daxl</td>
<td>4</td>
<td>2, 3, 14, 43</td>
</tr>
<tr>
<td></td>
<td>cyp11a1</td>
<td>2</td>
<td>8, 33</td>
</tr>
<tr>
<td></td>
<td>hsd3b1</td>
<td>2</td>
<td>7, 36</td>
</tr>
<tr>
<td></td>
<td>amhr2</td>
<td>2</td>
<td>9, 185</td>
</tr>
<tr>
<td></td>
<td>jnk1</td>
<td>9</td>
<td>1, 3, 4, 5, 16, 17, 38, 79, 117</td>
</tr>
<tr>
<td>Pluripotency factors</td>
<td>klf4</td>
<td>5</td>
<td>1, 3, 17, 96, 142</td>
</tr>
<tr>
<td></td>
<td>nr5a2</td>
<td>3</td>
<td>8, 19, 28</td>
</tr>
<tr>
<td></td>
<td>lin28a</td>
<td>2</td>
<td>2, 3</td>
</tr>
<tr>
<td></td>
<td>oct4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Apoptosis factors</td>
<td>iraf2</td>
<td>2</td>
<td>3, 15</td>
</tr>
<tr>
<td></td>
<td>casp2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>infr1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 1. Image of a Chinese black porgy. It was captured from Laizhou Bay in Yantai, Shandong Province, China.

Figure 2. Phylogeny of ray-finned fishes. Spotted gar was used as the outgroup. The bootstrap supporting value for the topology is 100. These pictures in the phylogenetic tree were downloaded from the Fishbase.

Ethics approval and consent to participate

All animal experiments in this study were implemented in the light of the guidelines...
of the Animal Ethics Committee and ratified by the Institutional Review Board on
Bioethics and Biosafety of BGI, China

**Availability of supporting data**

The raw sequencing reads of all libraries and the transcriptome data have been
deposited in the NCBI SRA database with accession numbers SRA541936 and
SRA587358. Supporting data are available in the GigaScience database, GigaDB.

**Author's contributions**

ZyZ, QS, and PX conceived the project. JX, CJ, JQ, FZ, HxL, HIL, DS, ZR and JC
extracted the genomic DNA and performed genome sequencing. KZ, SC, ZwZ, XY,
JZ, CB and JL assembled the genome and analyzed the data. TG, RG and JX
participated discussions and provided valuable advice for revision. KZ, QS, ZyZ, PX,
ZwZ and SC prepared the manuscript.

**Acknowledgements**

This work was supported by Aquatic Sanxin Engineering Major Project of Jiangsu
Province (No. D2015-17), Key Research and Development (Modern Agriculture)
Program of Jiangsu Province (No. BE2016326), Fund for Independent Innovation of
Agricultural Science and Technology of Jiangsu Province (No. CX(17)2021), Aquatic
Sanxin Engineering Project of Jiangsu Province (No.Y2016-23), Jiangsu Innovation
Ability Construction Program (No. BM2015017), Nantong Applied Basic Research
Program (No.MS12015071), Nantong Applied Basic Research Program (No.
MS12015070 & MS12016029), and Zhenjiang Leading Talent Program for
Innovation and Entrepreneurship.

**Competing interests**

The authors declare that they have no competing interests.

**References**

1. Gonzalez EB, Umino T, Nagasawa K. Stock enhancement programe for black sea
   bream, *Acanthopagrus schlegelii* (Bleeker), in Hiroshima Bay, Japan: a Review.


insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle.


sex-determining region of the Y chromosome of the medaka, Oryzias latipes.

male-to-female sex reversal after the sex determination by Dmy in the medaka.

52. Dong X, Chen S, Ji X et al. Molecular cloning, characterization and expression
analysis of sox9a and foxl2 genes in half-smooth tongue sole (Cynoglossus

53. Li H, Xu W, Zhang N et al. Two Figla homologues have disparate functions
during sex differentiation in half-smooth tongue sole (Cynoglossus semilaevis).
Scientific Reports 2016;6:28219.

54. Xie QP, He X, Sui YN et al. Haploinsufficiency of sf1 Causes Female to Male

pattern of zebrafish Anti-Mu‘llerian hormone (Amh) relative to sox9a, sox9b, and

ethynyl estradiol-induced sex differentiation on catfish, Clarias gariepinus:
expression profiles of dmrt1, Cytochrome P450 aromatases and 3
beta-hydroxysteroid dehydrogenase. Fish Physiology and Biochemistry 2005;31(2):143-147.

gene cDNA and its expression in gonads during natural sex transformation.
Biochemical Genetics 2007;45(3-4):211-224.

58. Xiao YM, Chen L, Liu J et al. Contrast expression patterns of jnk1 during sex
reversal of the rice field eel. Journal of Experimental Zoology Part B
2010;314(3):242-256.


