G3, 2021, 11(1), jkaa032
DOI: 10.1093/g3journal/jkaa032
Advance Access Publication Date: 14 December 2020
Genome Report

A high-quality *de novo* genome assembly of one swamp eel (*Monopterus albus*) strain with PacBio and Hi-C sequencing data

Hai-Feng Tian, Qiao-Mu Hu, and Zhong Li*

Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, Hubei 430223, China

*Corresponding author: Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, No.8, 1st Wudayuan Road, Donghu Hi-TechDevelopment Zone, Wuhan, Hubei 430223, China. lizhong@yfi.ac.cn

Abstract

The swamp eel (*Monopterus albus*) is one economically important fish in China and South-Eastern Asia and a good model species to study sex inversion. There are different genetic lineages and multiple local strains of swamp eel in China, and one local strain of *M. albus* with deep yellow and big spots has been selected for consecutive selective breeding due to superiority in growth rate and fecundity. A high-quality reference genome of the swamp eel would be a very useful resource for future selective breeding program. In the present study, we applied PacBio single-molecule sequencing technique (SMRT) and the high-throughput chromosome conformation capture (Hi-C) technologies to assemble the *M. albus* genome. A 799 Mb genome was obtained with the contig N50 length of 2.4 Mb and scaffold N50 length of 67.24 Mb, indicating 110-fold and ~31.87-fold improvement compared to the earlier released assembly (~22.24 Kb and 2.11 Mb, respectively). Aided with Hi-C data, a total of 750 contigs were reliably assembled into 12 chromosomes. Using 22,373 protein-coding genes annotated here, the phylogenetic relationships of the swamp eel with other teleosts showed that swamp eel separated from the common ancestor of Zig-zag eel ~49.9 million years ago, and 769 gene families were found expanded, which are mainly enriched in the immune system, sensory system, and transport and catabolism. This highly accurate, chromosome-level reference genome of *M. albus* obtained in this work will be used for the development of genome-scale selective breeding.

Keywords: *Monopterus albus*; swamp eel; genome assembly; PacBio; Hi-C

Introduction

The swamp eel (Fishbase ID: 4663; NCBI Taxonomy ID: 43700), *Monopterus albus*, is an economically important freshwater fish species in China, Japan, and Southeast Asia countries (Khanh and Ngan 2010; Nhan *et al.* 2019). As *M. albus* has a sex transition process during its life (Liu 1944), and it is becoming an emerging model species in development, genetics, and evolution (Cheng *et al.* 2003). Due to high nutritional level and impressive economic value of this species, the annual production of swamp eel has increased rapidly and exceeded 319,000 tons in China in 2018 (Zhang 2019), and its artificial propagation has been studied since last century in China (Guan *et al.* 1996) and corresponding techniques have been succeeded and promoted in Hubei, Sichuan provinces in China recently. Four or five different genetic lineages were recognized based on mitochondrial control region (Cai *et al.* 2013; Liang *et al.* 2016), and different local strains with different body colors (yellow, cyan, and grey) were cultivated in China (Yang *et al.* 2009; Wu *et al.* 2014). Due to the superiority in growth rate and fecundity (Chen *et al.* 2009; Yang *et al.* 2009), we have carried out selective breeding of one local strain of swamp eels (deep yellow and big spots). Though a genome of *M. albus* using second-generation sequencing technology (Illumina HiSeq 2000 plat-form) was published, which only covers 81.3% (Zhao *et al.* 2018), a chromosome-level, highly accurate reference genome for *M. albus* is still lacking, which hindering genome-scale genetic breeding for sustainable aquaculture of this species. Compared to the second-generation sequencing technologies, which constructs a complete genome by assembling numerous small sequence reads, thus leaving many gaps, third-generation sequencing technologies produce long reads and enable the production of a genome with a high level of completeness. Here, we report the chromosome-level genome assembly of *M. albus* using PacBio long-read sequencing and Hi-C technology. The assembly resulted in excellent continuity at the contig and scaffold levels, which facilitated subsequent investigations on genomic-based breeding studies targeting its economic traits.

Materials and methods

Ethics statement

All experiments in the present study were approved by the Animal Care and Use Committee of Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences (CAFS), China.

Received: August 27, 2020. Accepted: November 22, 2020

Sample collection and sequencing

The local strain (deep yellow and big spots) of *M. albus* was bred and cultivated at the Yangtze River Fisheries Research Institute (Jingzhou, Hubei Province, China). The F_1 full-sibling families were established from breeding population by brother–sister mating, and the F_2 generation full-sibling families were obtained from the F_1 generation with brother–sister mating. The larvas were hatched in plastic incubation boxes and fed with red worms, and then they were transferred to netcages in traditional ponds about 6–7 days post-hatching. The juvenile and adult fish were fed with artificial diets. Two 1-year-old healthy female F_2 individuals were used to collect sample tissues (Figure 1). The fish was immediately dissected after anesthesia with MS-222. White muscle and liver tissues of one fish were collected for whole genome sequencing and Hi-C library construction, respectively. Six different tissues (white muscle, liver, spleen, intestine, pituitary, and ovary) of another fish were collected and used for transcriptome sequencing. Genomic DNA from muscle tissue was extracted using the standard phenol/chloroform extraction method for DNA sequencing library construction. The integrity of the genomic DNA molecules was checked using agarose gel electrophoresis. Both the Illumina NovaSeq platform and the PacBio SEQUEL II platform were applied for genomic sequencing to generate short and long genomic reads, respectively. For the Illumina NovaSeq platform (San Diego, CA, USA), a pair-end library was constructed with an insert size of 350 base pairs (bp) according to the protocol provided by the manufacture. For the PacBio SEQUEL II platform, 10 μg genomic DNA was used for 20 kb SMRTbell library construction according to the manufacturer's protocol (Pacific Biosciences). Liver tissue collected from the same individual was used for Hi-C library construction. One gram of fresh liver was first fixed using formaldehyde with a final concentration of 1%. The fixed tissue was then homogenized with tissue lysis, digested with the restriction enzyme (*MboI*), *in situ* labeled with a biotinylated residue, and end-repaired. Then the Hi-C libraries were quantified and sequenced using the Illumina NovaSeq 6000 instrument (Illumina).

We also performed RNA sequencing to generate transcriptome data on the Illumina NovaSeq platform for gene model prediction. TRIzol reagent (Invitrogen, USA) was used to separately extract RNA from collected six tissues (white muscle, liver, spleen, intestine, pituitary, and ovary). RNA purity was checked using the kaiaoK5500®Spectrophotometer (Kaiao, Beijing, China), and RNA integrity and concentration were assessed using the RNA Nano 6000 Assay Kit of Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Then, a total amount of 2 μg RNA per sample was mixed for RNA sequencing. Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (#E7530L, NEB, USA) following the manufacturer's recommendations and the libraries were sequenced on an Illumina platform and 150 bp paired-end reads were generated.

Figure 1 Picture of the local strain of swamp eel (deep yellow and big spots) used in the genome sequencing and assembly.

Genome size estimation and assembly

For the next-generation sequencing (NGS) short reads, the Kmer-based method (Liu *et al.* 2013) was used to perform genome survey analysis to estimate the genome size, heterozygosity and repeat content of the swamp eel genome. We countered the number of each 17-mer with Jellyfish (Marcais and Kingsford 2011), and the frequency distribution is plotted. The PacBio long reads were assembled with Canu package v1.8 (Koren *et al.* 2017)(corrected error rate = 0.045, cor out coverage = 40). The obtained draft assembly was then polished with two steps: The assembly was first polished with arrow (Chin *et al.* 2013) using long reads, and then was polished with Illumina short reads with Pilon (Walker *et al.* 2014). Lastly, redundant sequences were removed with Purge Haplotigs (Roach *et al.* 2018).

The reads from the Hi-C library sequencing were mapped to the polished swamp eel genome with Bowtie (v2.3.4.3) (Langmead and Salzberg 2012). We independently aligned the two read ends to the genome and only selected the read pairs for which both ends were uniquely aligned to the genome. The hiclib Python library (Imakaev *et al.* 2012) and a previously reported method (Servant *et al.* 2015) were applied to filter the Hi-C reads, and the interaction frequency was quantified and normalized among contigs. Based on the interaction matrix, Lachesis (Burton *et al.* 2013) with default parameters was then applied to anchor the contig to the chromosomes using an agglomerative hierarchical clustering method.

The old genome assembly (GCF_001952655.1) was aligned to our genome assembly by using minimap2 v2.14-r883(Li 2018)with the "asm5" preset, and the output results were plotted as dotplots using an R script called dotPlotly (https://github.com/tpoorten/dotPlotly) with default parameters. The completeness and accuracy were also assessed via short-read mapping and Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis. The reads generated by Illumina NovaSeq platform were mapped to the *M. albus* genome assembly using BWA (version 0.7.17) (Li and Durbin 2009), and variant calling was performed with SAMTOOLS (Li *et al.* 2009). We queried the assembled swamp eel genome against the Actinopterygii (actinopterygii_odb9, containing 4584 highly conserved single-copy core Actinopterygian genes) by using BUSCO v3.0.2 (Simao *et al.* 2015). Besides, the completeness and consensus accuracy of our new assembly were also computed using Merqury version 1.0 (Rhie *et al.* 2020).

Gene model prediction and functional annotations

Repeat elements were annotated in the swamp eel before gene model annotation. We applied Tandem Repeat Finder (TRF) (Benson 1999), LTR_FINDER (Xu and Wang 2007), PILER (Edgar and Myers 2005), and RepeatScout (Price *et al.* 2005) for the *ab initio* prediction of repeat elements in the genome. Thereafter, RepeatMasker and RepeatProteinMask (v4.0.9) (http://www.repeatmasker.org, last accessed April 10, 2019) were used to search the genome sequences for known repeat elements, with the genome sequences used as queries against the Repbase database (version 23.08) (Jurka *et al.* 2005; Bao *et al.* 2015).

Gene annotation was performed on the repetitive-element-masked genome. A combined strategy of homology-based, *ab initio* and transcriptome-based gene prediction methods was used. For homology-based prediction, protein sequences of closely related fish, including *Acanthochromis polyacanthus* (GCF_002109545.1), *Amphiprion ocellaris* (GCF_002776465.1), *Anabas testudineus* (GCF_900324465.1), *Astatotilapia calliptera* (GCF_900246225.1), *Mastacembelus armatus* (GCF_900324485.1), and *M. albus*

(GCF_001952655.1) were downloaded from National Center for Biotechnological Information (NCBI). Proteins from the closely related fish species were mapped to the swamp eel genome using TBLASTN (Gertz *et al.* 2006). The alignment hits were joined with Solar (Yu *et al.* 2006). Next, GeneWise (Birney *et al.* 2004) was used to predict the exact gene structure of the corresponding genomic regions. In addition, RNA-seq reads were directly mapped to the assembled genome to identify putative exon regions using TopHat package (v2.1.1) (Trapnell *et al.* 2009) and Cufflinks (v2.2.1) (Ghosh and Chan 2016). For *ab initio* gene prediction, AUGUSTUS v3.3.2 (Stanke *et al.* 2006) was first trained using the BUSCO annotation of core *Actinopterygii* genes and then used for the prediction of genes in the repeat-masked genome. All of the gene models were merged, and redundancy was removed by MAKER (Cantarel *et al.* 2007). To annotate protein-coding genes functionally, gene sequences were searched using DIAMOND (v 0.9.19) with an e-value threshold of 1e–5 against the NCBI non-redundant protein (nr), TrEMBL, KOG, PFAM, and the SwissProtdatabases. Functional ontology and pathway information from the Gene Ontology (GO) and the Kytoto Encyclopedia of Genes and Genomes (KEGG) databases (Ogata *et al.* 1999) were assigned to the genes using BLAST2GO (Conesa *et al.* 2005) and KAAS (v2.1) (Moriya *et al.* 2007). Noncoding RNAs, including rRNAs, snRNAs, miRNAs, and tRNAs, were annotated by using TRNASCAN-SEv1.3.1 (Lowe and Eddy 1997) and the RFAM database (release 13.0) (Griffiths-Jones *et al.* 2003) using INFERNAL (v1.1.2) (Nawrocki and Eddy 2013).

Phylogenetic analysis and divergence time estimation

To identify gene families among *M. albus* and other species, the protein sequences of other nine fish species were downloaded from NCBI. These species included four close relative species of Anabantaria [*Anabas testudineus* (GCF_900324465.2), *Betta splendens* (GCF_900634795.2), *Mastacembelus armatus* (GCF_900324485.2), *and Channa argus* (GCA_004786185.1)], and five other teleost species [*Oryzias latipes* (GCF_002234675.1), *Seriola lalandi* (GCF_002814215.1), *Takifugu rubripes* (GCF_901000725.2), *Danio rerio* (GCF_000002035.6), and *Gadus morhua* (GCF_902167405.1)] as the outgroup. After removing those sequences less than 30 amino acids, these protein sequences were aligned to each other with BLASTP (Altschul *et al.* 1990) programs with an e-value threshold of 1e–5. Then, OrthoMCL (1.4) (Li *et al.* 2003) were used to cluster gene families with a Markov inflation index of 1.5 and a maximum *e* value of 1e–5. All gene families were ascertained from at least two genomes.

Protein sequences of the obtained one-to-one orthologous genes obtained from OrthoMCL analysis were used for phylogenetic tree reconstruction. MUSCLE (Edgar 2004) was used to generate multiple sequence alignments for protein sequences in each single-copy family with default parameters. Then we used Gblocks (Talavera and Castresana 2007) to extract the well-aligned regions of each gene family alignment and converted protein alignments to the corresponding coding DNA sequence alignments using an in-house script. Then, the alignments of each family were concatenated to a super alignment matrix. The super alignment matrix was subjected to phylogenetic analysis using RAxML (v8.2.11) (Stamatakis 2014) with the GTRGAMMA model and 100 bootstrap replicates.

Bayesian molecular dating was adopted to estimate the neutral evolutionary rate and species divergence time with MCMCTREE from PAML (version 4.4b) (Yang and Rannala 2006) with the options "correlated molecular clock" and "JC69" model. Three calibration times were obtained from TimeTree database (http://www.timetree.org/, last accessed June 26, 2020) (Hedges *et al.* 2015) and previous report (Benton and Donoghue 2007). The MCMC (Markov chain Monte Carlo) chain length was set to 20,000 generations using a burn-in of 2000 iterations.

Expansion and contraction of gene families

The orthologous genes and phylogenetic tree topology inferred from the OrthoMCL analysis were taken into CAFÉ v4.2 (De Bie *et al.* 2006), which used a random birth and death model, to estimate the size of each family at each ancestral node and obtain a family-wise p-value (based on a Monte–Carlo re-sampling procedure) to indicate whether a significant expansion or contraction occurred in each gene family across species.

Data availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Wang *et al.* 2017) in National Genomics Data Center (National Genomics Data Center and Partners 2020), Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, under accession number CRA003062 that are publicly accessible at https://bigd.big.ac.cn/gsa.

Supplementary material is available at figshare DOI: https://doi.org/10.6084/m9.figshare.13228784.

Results and discussion

Genome size estimation and assembly of the *M. albus* genome

To estimate the genome size and heterozygosity of *M. albus*, 89.78 gigabyte (Gb) Illumina clean reads were obtained by Illumina paired-end libraries (Table 1). The mean reads length for Illumina data was 150 bp. The paired-end reads were primarily used for genome properties estimation, to assess and improve the base-level quality of the assembly. Based on the total number of 68,732,704,793 17-mers and a peak 17-mer depth of 83, the estimated genome size of *M. albus* was calculated to be 785 Mb, with 42.07% GC content (Supplementary Figure S1). In addition, the estimated heterozygosity rate and repeats were approximately 0.49% and 47.32%, respectively (Supplementary Figure S1).

We used the PacBio SEQUEL II platform to generate long genomic reads for the reference genome construction. A final total of 136.89 Gb sequencing data (~174 X) were obtained (Table 1). The mean length of read and the N50 were 16 kb and 22.7 kb, respectively (Supplementary Tables S1). Then the PacBio long reads were assembled with CANUpackage and polished by Arrow and Pilon. We obtained a final assembled genome of 799 megabase (Mb) with a contig N50 length of 2.48 Mb (Supplementary Table S2) and GC content of 41.4%. The genome contained 809 contigs, and the longest contig was 14.73 Mb (Table 2; Supplementary Table S2). The GC content and sequencing depth for the genomic contigs were consistent with the estimation from the *k-mer*-based analysis (Supplementary Figure S2).

The Hi-C reads were used to generate a chromosome-level assembly of the genome. Using Hi-C library sequencing and read-quality filtering, we obtained approximately 107 Gb of clean bases with Q20 of 97.73% and Q30 of 93.22% (Table 2; Supplementary Table S3). Using our mapping strategy, we found that more than 95% of the total reads properly mapped to the genome, and 149 million read pairs (39.7%) provided valid interaction information for chromosome assembly. As a result, 750 contigs were successfully clustered, ordered, and oriented in 12 chromosomes

Table 1 Statistics of the DNA sequence data used for *M. albus* genome assembly

Library resource	Sequencing platform	Insert size	Raw data (Gb)	Clean data (Gb)	Sequence coverage (X)
Genome	Illumina NovaSeq	350 bp	91.02	89.78	115.9
genome	PacBio SEQUEL II	20 kb	136.89	136.89	174.49
Hi-C	Illumina NovaSeq	350 bp	114.21	107.93	145.49
Transcriptome	Illumina NovaSeq	350 bp	8.93	8.67	11.38

Table 2 Assembly statistics for the swamp eel

	Length				Number			
	Contig (Mb)		Scaffold (Mb)		Contig		Scaffold	
Content	new	GCF_001952655.1	new	GCF_001952655.1	new	GCF_001952655.1	new	GCF_001952655.1
Total	799.64	635	799.72	689.52	809	117,579	71	62,978
Max	14.73	0.16	88.66	11.7	–	–	–	–
Number ≥2 kb	–	–	–	–	805	44,314	63	2,360
N50	2.44	0.02	67.24	2.11	97	8438	97	87
N90	0.54	0.005	49.70	0.37	348	33115	348	379

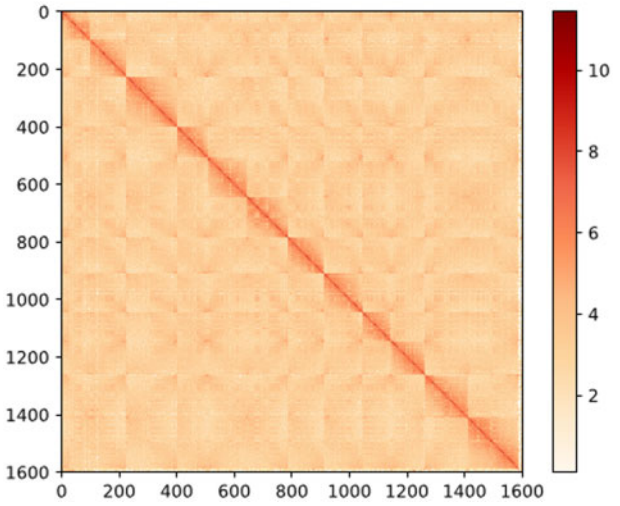

Figure 2 Swamp eel genome contig contact matrix using Hi-C data, color bar indicates contact density from red (high) to white (low).

(Figure 2), which represented 99.26% and 88.57% of all of the contigs at the base and sequence number level, respectively. The contig and scaffold N50 length of the final assembly reached 2.44 and 67.24 Mb, respectively, which represent ~110-fold and ~31.87-fold improvement compared with previously reported ~22.24 Kb and 2.11 Mb (Zhao *et al.* 2018), respectively (Table 2). There were still 59 unanchored contigs after the Hi-C-based chromosome construction with an N50 length of 88.2 kb, which was significantly smaller than that of the anchored contigs. Although a primary genome assembly of *M. albus* was available (Zhao *et al.* 2018), its scaffold N50 (2.11 Mb) and contig N50 (22.23 kb) were shorter. Therefore, in this study, we improved the genome work with a high-quality assembly.

We evaluated the accuracy of the initial assembly by mapping the 606,791,208 Illumina reads for genome survey to the assembly using BWA. As a result, 99.57% of the Illumina reads were successfully mapped to the assembled genome with a coverage of 99.84% (Supplementary Table S4). We also obtained 0.001% homozygous single nucleotide polymorphisms (SNP), suggesting the high accuracy of this assembly.

The comparisons between the two assemblies obtained 120,662 aligned blocks, which account for 94.85% of the previous assembly (GCF_001952655.1), but only account for 83.96% of our new assembly, and the dot plot for the whole-genome alignment shows good collinearity between the two assemblies though some gaps and inversions are also found (Supplementary Figure S3). This result indicates great improvements in terms of contiguity and gaps in our newly obtained genome assembly. To evaluate the completeness of the assembly, the assembled swap-eel genome was assessed by using BUSCO (Simao *et al.* 2015) with the actinopterygii_odb9 database (4584 core genes). We found that 95.99% of core genes were identified in full-length in this swamp eel genome assembly and 49 (1.07%) core genes were captured as fragments (Table 3). These results suggest that less than 2.95% of the core Ray-finned fish genes were missing in our swamp eel assembly. Furthermore, when assessed by using Merqury, an optimal $k = 19$ was calculated and the spectra-cn plot indicates that our assembly is a complete haplotype-resolved assembly (Supplementary Figure S4), and the *k-mer* completeness of our assembly is 92.6% and QV scores is 41.35 (Supplementary Table S5), indicating our assembled genome with high sequencing coverage and assembly accuracy. All these results suggested that these sequences represent a significant improvement in contiguity in contrast to previously published swamp eel genome (Zhao *et al.* 2018) and the genome assembly was complete and robust.

Genome annotation

Using *de novo* searching and homolog prediction, we found that ~46.76% of genome contents were repetitive elements (Supplementary Table S6), which was comparable with our estimation from the *k-mer*-based method. The identified repetitive elements in this assembly are higher than that previous report identified by Illumina sequencing (Zhao *et al.* 2018) and those of *Betta spenden* (15%) and *Channa argus* (18.9%). With regard to transposable elements (TEs), DNA transposons (20.98%), long interspersed nuclear elements (LINEs, 17.39%) and long terminal repeats (LTRs, 9.64%) were the top three categories of repetitive elements in the *M. albus* genome (Table 4). Thus, altogether, 46.23% of the genome was predicted to be repeated.

In the current assembly, a total of 22,373 protein-coding genes with an average coding DNA Sequence (CDS) of 1707 bp and 10.4 exons per gene were identified based on *de novo*, homology- and

transcriptome sequencing-based methods (Supplementary Table S7; File S1). By comparing the distribution of genes, coding sequences (CDS), exon and intron lengths of *M. albus* and closely related species, we found that the distributions in the *M. albus* genome were comparable with those of other teleosts (Supplementary Figure S5). For clarity, the distributions of GC density, gene density and repeat density across the 12 chromosomes of *M. albus* genome were further illustrated in Figure 3. To understand gene function, all of the predicted genes were mapped against several public databases, and 21,769 genes were functionally annotated in at least one of the databases, and up to 97.3% of *M. albus* genes were functionally annotated (Table 5). Finally, 670 miRNAs, 296 rRNAs, 503 snRNAs, and 16,279 tRNAs were also identified (Supplementary Table S8).

Gene families and divergence time estimation

The consensus proteome set of the nine teleost species and the swamp eel were composed of a final data set of 239,948 protein sequences. A total of 19,705 gene families were predicted, including 4777 were one-to-one single-copy (Supplementary Table S9). *M. albus* contained 15,171 gene families. We studied the ortholog profiles of the five closely related species of Anabantaria (*Anabas testudineus, Betta splendens, Mastacembelus armatus, Channa argus*, and *M. albus*), a total of 12,353 (62.7%) gene families were shared by all five species (Figure 4) and 285 (~0.1%) were *M. albus* specific ones.

In the phylogenetic tree obtained using 4777 single-copy ortholog gene families from the 10 teleost species, *M. albus* was recovered to be clustered with *M. armatus*, and then grouped with other teleosts, which was consistent with the fish species taxonomy (Hughes *et al.* 2018). *Monopterus albus* diverged from the common ancestor with *M. armatus* around 32.7–58.6 million years ago (MYA) (Figure 5).

Gene family expansion and contraction analysis

We analyzed gene family expansion and contraction in the *M. albus* lineage. There were 769 and 2414 gene families that expanded and contracted, respectively, after speciation from *M. armatus* (Figure 6A). Using the Gene Ontology (GO) and KEGG databases, we observed that 1514 genes from expanded gene families were significantly enriched in 479 and 179 ($q < 0.05$) GO terms (Supplementary Table S10) and KEGG pathways (Supplementary Table S11), respectively. The expanded gene

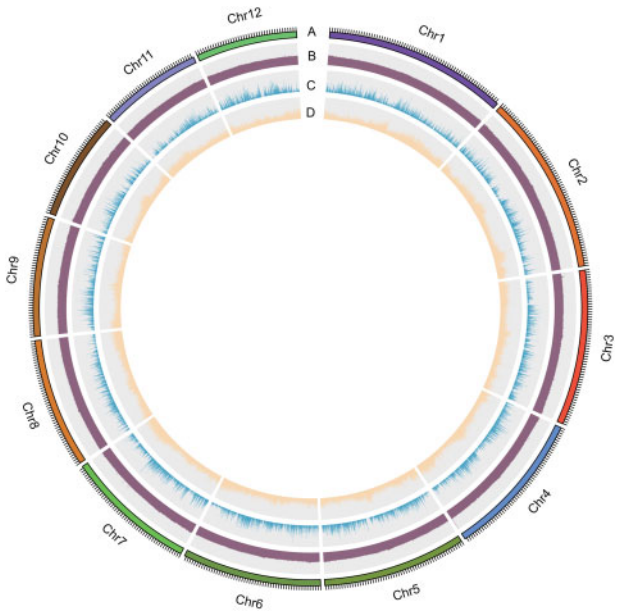

Figure 3 Genome landscape of *M. albus*. From outer to inner circles: (A) 19 chromosomes; (B) GC density; (C) gene density; and (D) repeat density. Chr: chromosome. All statistics are based on 50 kb nonoverlapping windows.

Table 3 BUSCO results for analysis of genome completeness for *M. albus*

Type	Number of genes	%
Complete BUSCOs (C)	4,400	95.99
Complete and single-copy BUSCOs (S)	4,247	92.65
Complete and duplicated BUSCOs (D)	153	3.34
Fragmented BUSCOs (F)	49	1.07
Missing BUSCOs (M)	135	2.95
Total BUSCO groups searched	4,584	100.00

Table 5 Statistics of functional annotation of protein-coding genes

Database	Number	%
InterPro	19,893	88.92
GO	15,058	67.3
KEGG_ALL	21,371	95.52
KEGG_KO	13,605	60.81
Swissprot	18,961	84.75
TrEMBL	21,492	96.06
TF	3,601	16.1
Pfam	19,252	86.05
NR	21,684	96.92
KOG	17,519	78.3
At least one database	21,769	97.3
Total	22,373	–

Note that "at least one database" here refers to genes with at least one hit in multiple database.

Table 4 Repetitive element annotations in the swamp eel

	Repbase TEs		TE proteins		De novo		Combined TEs	
Type	Length (Bp)	% in genome	Length (Bp)	% in genome	Length (Bp)	% in genome	Length (Bp)	% in genome
DNA	64154403	8.02	22201254	2.78	140659060	17.59	167727458	20.98
LINE	65337162	8.17	48523109	6.07	111587057	13.95	139038129	17.39
SINE	6049424	0.76	0	0	5559625	0.7	11190306	1.4
LTR	21457697	2.68	18986592	2.37	68782334	8.6	77048713	9.64
Satellite	1279944	0.16	0	0	623725	0.08	1897697	0.24
Other	4012	0	0	0	0	0	4012	0
Unknown	1177982	0.15	870	0	22004421	2.75	23151650	2.9
Total	155589278	19.46	89653198	11.21	315767592	39.49	369647460	46.23

families were mainly found on immune system pathways, especially on Intestinal immune network for IgA production (q = 1.96E–33); Hematopoietic cell lineage (q = 7.05E–24); Th1 and Th2 cell differentiation (q = 3.90E–20); Th17 cell differentiation (q = 1.92E–16) on KEGG pathways; Transport and catabolism, including Phagosome (q = 1.39E–32), Autophagy-animal (q =2.66e–21); Sensory system, including Olfactory transduction (q =2.77E–29); signal transduction pathways, including NF-kappa B signaling pathway (q =1.19E–18), Notch signaling pathway (q = 1.84E–12), Calcium signaling pathway (q = 1.82E–9), and Rap1 signaling pathway (q = 6.44E–8); Metabolism of cofactors and vitamins (q = 1.82E–9). In addition, a total of 195 genes in the contracted families enriched in 412 GO terms and 49 KEGG pathways ($P < 0.05$) were enriched, respectively (Supplementary Tables S12 and S13). These enriched terms were mainly involved in Cellular community-eukaryotes, including Tight junction (q = 1.59E–7), Gap junction (q = 9.13E–3); Carbohydrate metabolism, including Ascorbate and aldarate metabolism (q = 1.59E–7), Pentose and glucuronate interconversions (q = 1.59E–7). These biological processes may be related to the special characteristics of the swamp eel.

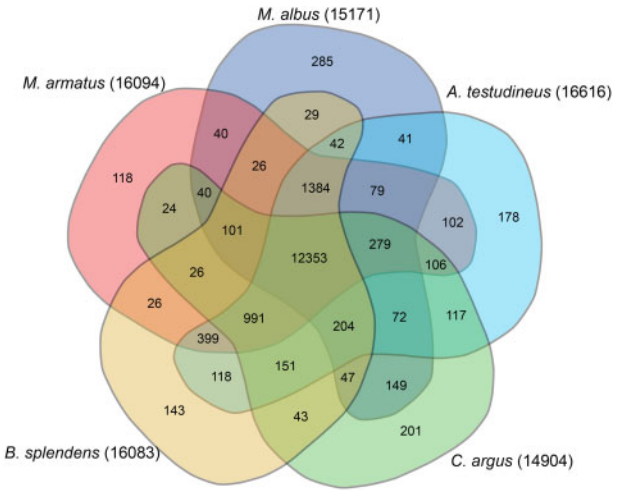

Figure 4 Venn diagram showing the distribution of shared gene families and their distribution between *M. albus*, *Anabas testudineus*, *Betta splendens*, *Mastacembelus armatus*, and *Channa argus*. The intersections between species indicate the numbers of shared gene families, whereas unique family numbers are shown in species-specific areas. The center represents the number of families shared by all species simultaneously. Analysis of families showed 12,535 families shared by all species simultaneously. Total number of gene families for each species is given in parenthesis.

Conclusions

Due to the unique protogynous hermaphroditism of *M. albus* (Liu 1944) and high economic value in fisheries in China and Southeast Asia (Matsumoto *et al.* 2010), highly accurate, chromosome-level reference genome would provide a sound support for subsequent investigations of its sexual reversion, genome-scale selective breeding, and future investigations of population and conservation genetics.

In the present study, combing Illumina, PacBio long-read sequencing, and Hi-C technologies, we reported a high-quality chromosome-level genome assembly for the localstrain (deep yellow and big spots) of swamp eel. The contig and scaffold N50 reached 2.44 and 67.24 Mb, respectively, suggesting this genome assembly improved substantially in the primary assembly contiguity in contrast to previous assembly (Zhao *et al.* 2018). More than 95.99% of the complete *Actinopterygii* BUSCO genes (n = 4,400) were identified in the genome and 99.57% of the high-quality Illumina reads can be mapped onto the *de novo* genome assembly, suggesting the quality of our *de novo* assembled genome was high for both completeness and base-level accuracy. Besides, good *k-mer* completeness (92.6%) and high QV value (41.35) calculated with Merqury suggesting high coverage and accuracy in our assembly. More repetitive sequences (46.23%) were identified than those in the previous report (~28%)(Zhao *et al.* 2018), which might be attributed to the higher contiguity and

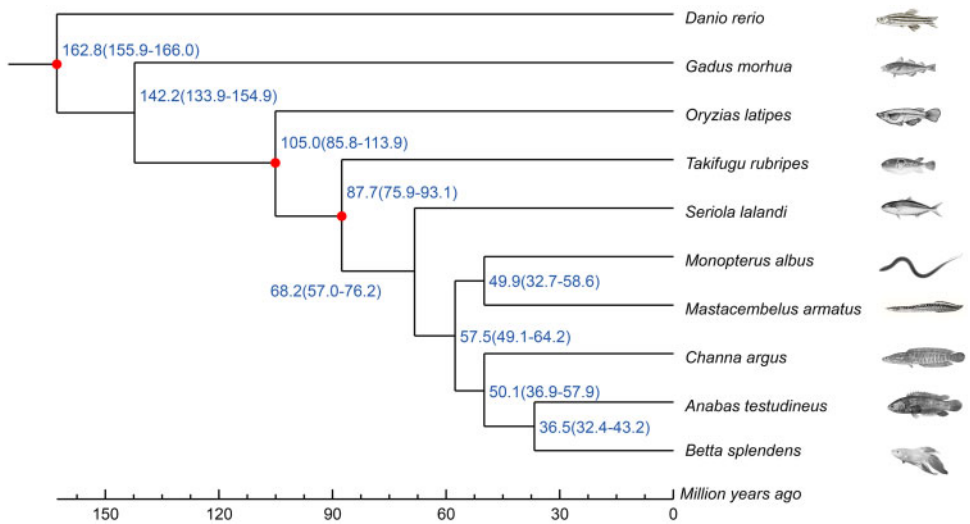

Figure 5 Phylogenetic analysis of *M. albus* and closely related teleost fish species. The species divergence time was shown at the branches of the phylogenetic tree, and the confidence intervals are given in parentheses. Except the figure of *M. albus*, all of the other species figures are retrieved from Wikimedia Commons. The species figures of *Danio rerio* (Mintern 1878), *Gadus morhua*, *Seriola lalandi* (Culloch 1911), *Mastacembelus armatus* (Ford 1878), *Channa argus*, and *Anabas testudineus* (France Day 1878) are distributed in public domain, and the species figures of *Oryzias latipes* (by Minami Kawasaki) (Kawasaki 2016), *Takifugu rubripes* [by DataBase Center for Life Science (DBCLS) ((DBCLS) DCFLS 2012)], and *Betta splendens* [by Pharaoh Hound (Hound 2007)] are distributed under Creative Commons Attribution 4.0 International license and Creative Commons Attribution 2.5 Generic license respectively.

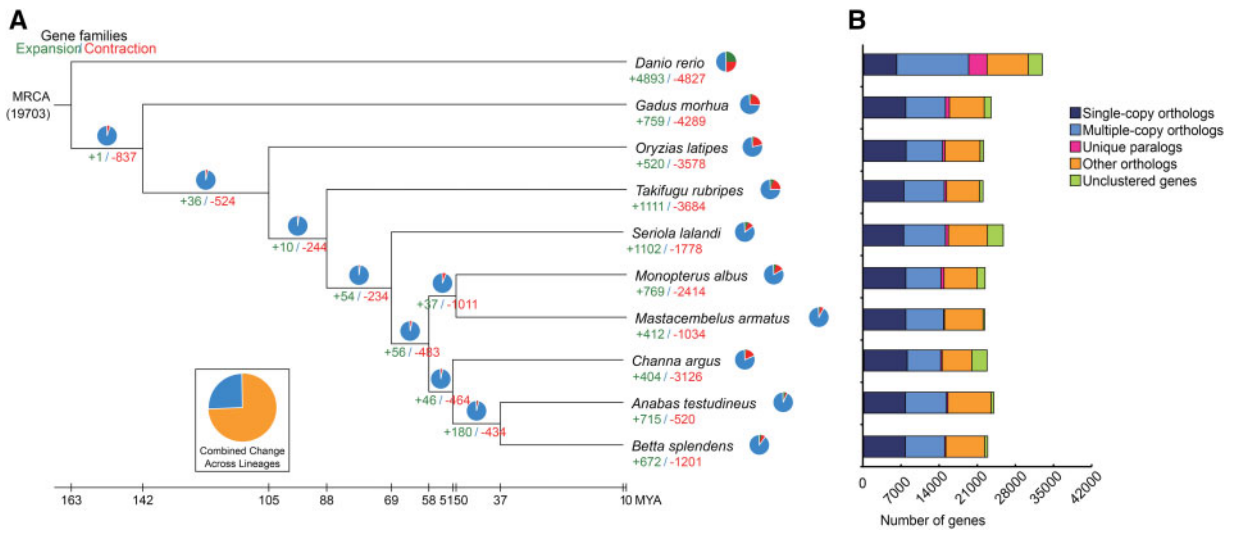

Figure 6 (A) Dynamic evolution and distribution of gene families among 10 teleost species, including *Monopterus albus*, *Anabas testudineus*, *Betta splendens*, *Mastacembelus armatus*, and *Channa argus*, *Takifugu rubripes*, *Oryzias latipes*, *Gadus morhua*, and *Danio rerio*. The green and red numbers represent the expanded and extracted gene families, respectively. MRCA: most recent common ancestor. (B) The distribution of single-copy, multiple-copy, unique, other orthologs, and unclustered genes in 10 teleost species.

accuracy of our assembly. Such high percent of repetitive elements identified in this species is also higher than many of the sequenced teleost genomes (Reichwald *et al.* 2009), but lower than that of the zebrafish (54.3%) (Howe *et al.* 2013) and Atlantic salmon (58%) (Lien *et al.* 2016). Besides, a total of 22,373 protein-coding genes were annotated based on this assembly.

The phylogenetic analysis of related species showed that swamp eel was diverged ~49.9 MYA from the common ancestor of *M. armatus*. Expanded gene families were significantly enriched in several important biological pathways, mainly in immune system, sensory system, and Transport and catabolism.

In summary, we have completed a chromosome genome assembly of *M. albus*. Given the economic importance of swamp eel for aquaculture production and the increasing research interests for its unique sex reversal, we expect this well-assembled genomic data could offer a valuable resource for future genomic-scale selective breeding programs and provide valuable information for improving economically relevant culture traits, such as disease resistance.

Acknowledgments

We thank two anonymous reviewers for their valuable suggestions.

Funding

This research was funded by the Central Public-Interest Scientific Institution Basal Research Fund, Chinese Academy of Fishery Sciences (CASF) (grant numbers 2018JBF01 and 2020XT08).

Conflict of interest: The authors declare that they have no competing interests.

Literature cited

(DBCLS) DCFLS 2012. Torafugu. https://commons.wikimedia.org/wiki/File:201207_torafugu.svg

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

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.

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

Benton MJ, Donoghue PC. 2007. Paleontological evidence to date the tree of life. Mol Biol Evol. 24:26–53.

Birney E, Clamp M, Durbin R. 2004. Genewise and genomewise. Genome Res. 14:988–995.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, *et al.* 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 31:1119–1125.

Cai X, Yu SM, Mipam T, Zhang XY, Yue BS. 2013. Phylogenetic lineages of *Monopterus albus* (synbranchiformes: Synbranchidae) in China inferred from mitochondrial control region. J Zoolog Syst Evol Res. 51:38–44.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, *et al.* 2007. Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188–196.

Chen F, Yang D-Q, Su Y-B. 2009. Comparative study on growth speed of *Monopterus albus* of different body colours. J Yangtze Univ (Nat Sci Edit). 6:33–38.

Cheng H, Guo Y, Yu Q, Zhou R. 2003. The rice field eel as a model system for vertebrate sexual development. Cytogenet Genome Res. 101:274–277.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, *et al.* 2013. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. Nat Methods. 10:563–569.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, *et al.* 2005. Blast2go: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674–3676.

Culloch ARM. 1911. *Seriola lalandi*. https://commons.wikimedia.org/wiki/File:Seriola_lalandi.jpg

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

Edgar RC. 2004. Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edgar RC, Myers EW. 2005. Piler: identification and classification of genomic repeats. Bioinformatics. 21:i152–i158.

Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF. 2006. Composition-based statistics and translated nucleotide searches: improving the tblastn module of blast. BMC Biol. 4:41.

Ghosh S, Chan C-KK. 2016. Analysis of RNA-seq data using tophat and cufflinks. In: D Edwards, editor. Plant Bioinformatics: Methods and Protocols. New York, NY: Springer New York. p. 339–361.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. Nucleic Acids Res. 31:439–441.

Guan R, Zhou L, Cui G, Feng X. 1996. Studies on the artificial propagation of *Monopterus albus* (zuiew). Aquaculture Res. 27:587–596.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. Mol Biol Evol. 32: 835–845.

Hound P. 2007. Https://commons.wikimedia.org/wiki/file:Betta_splendens_pale.Jpg.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, *et al.* 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 496:498–503.

Hughes LC, Orti G, Huang Y, Sun Y, Baldwin CC, *et al.* 2018. Comprehensive phylogeny of ray-finned fishes (actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci USA. 115:6249–6254.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, *et al.* 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 9:999–1003.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, *et al.* 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Kawasaki M. 2016. Oryzias latipes in j-phenome. https://commons.wikimedia.org/wiki/File:201606_08_medaka.png.

Khanh N, Ngan H. 2010. Current practices of rice field eel *Monopterus albus* (zuiew, 1973) culture in vietnam. Aquaculture Asia Mag. 15: 26–29.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.* 2017. Canu: Scalable and accurate long-read assembly via adaptive *k-mer* weighting and repeat separation. Genome Res. 27:722–736.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. Nat Methods. 9:357–359.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Genome Project Data Processing S, *et al.* 2009. The sequence alignment/map format and samtools. Bioinformatics. 25:2078–2079.

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

Liang HW, Guo SS, Li Z, Luo XZ, Zou GW. 2016. Assessment of genetic diversity and population structure of swamp eel *Monopterus albus* in china. Biochem Syst Ecol. 68:81–87.

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, *et al.* 2016. The Atlantic salmon genome provides insights into rediploidization. Nature. 533:200–205.

Liu B, Shi Y, Yuan J, Hu X, Zhang H, *et al.* 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Quantitative Biol. 35:62–67.

Liu CK. 1944. Rudimentary hermaphroditism in the symbranchoid eel. Sinensia. 15:1–8.

Lowe TM, Eddy SR. 1997. Trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. Nucleic Acids Res. 25:955–964.

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

Mastacembelus armatus. 1878. https://commons.wikimedia.org/wiki/File:Mastacembelus_armatus_Ford_73.jpg.

Matsumoto S, Kon T, Yamaguchi M, Takeshima H, Yamazaki Y, *et al.* 2010. Cryptic diversification of the swamp eel *Monopterus albus* in east and southeast asia, with special reference to the ryukyuan populations. Ichthyol Res. 57:71–77.

Mintern R. 1878. Danio rerio. Day, Francis (1878) the Fishes of India. 2: https://commons.wikimedia.org/wiki/File:Danio_rerio_Mintern_151.jpg.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. Kaas: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35(Web Server issue):W182–W185.

National Genomics Data Center and Partners. 2020. Database resources of the national genomics data center in 2020. Nucleic Acids Res. 48:D24–D33.

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

Nhan HT, Tai NT, Liem PT, Ut VN, Ako H. 2019. Effects of different stocking densities on growth performance of asian swamp eel *Monopterus albus*, water quality and plant growth of watercress nasturtium officinale in an aquaponic recirculating system. Aquaculture. 503:96–104.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, *et al.* 1999. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27:29–34.

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

Reichwald K, Lauber C, Nanda I, Kirschner J, Hartmann N, *et al.* 2009. High tandem repeat content in the genome of the short-lived annual fish nothobranchius furzeri: a new vertebrate model for aging research. Genome Biol. 10:R16.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21:245.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 19:460.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, *et al.* 2015. Hic-pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16:259.

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

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

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, *et al.* 2006. Augustus: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34(Web Server issue):W435–W439.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biol. 56:564–577.

Trapnell C, Pachter L, Salzberg SL. 2009. Tophat: discovering splice junctions with rna-seq. Bioinformatics. 25:1105–1111.

Walker B J, Abeel T, Shea T, Priest M, Abouelliel A, *et al.* 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. **9**:e112963.

Wang Y, Song F, Zhu J, Zhang S, Yang Y, *et al.* 2017. Gsa: genome sequence archive. Genom Proteom Bioinform. 15:14–18.

WuL X-L, Ding W-D, Cao Z-M, Bing X-W. 2014. Research status and prospects on biological characteristics of *Monopterus albus* with different colors. J Huaihai Inst Technol (Nat Sci Ed). 23: 80–87.

Xu Z, Wang H. 2007. Ltr_finder: an efficient tool for the prediction of full-length ltr retrotransposons. Nucleic Acids Res. 35(Web Server issue):W265–W268.

Yang D-Q, Chen F, Ruan G-L, Li T-M. 2009. Comparative study on fecundity of different strains of *Monopterus albus*. J Hydroecol. 2: 133–135.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol. 23:212–226.

Yu XJ, Zheng HK, Wang J, Wang W, Su B. 2006. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics. 88:745–751.

Zhang X-L. 2019. China Fishery Statistical Yearbook. Beijing: China Agriculture Press.

Zhao X, Luo M, Li Z, Zhong P, Cheng Y, *et al.* 2018. Chromosome-scale assembly of the *Monopterus genome*. Giga Sci. **7**:giy046.

Communicating editor: A. Whitehead