

## DATA NOTE

# Draft genome of the Northern snakehead, *Channa argus*

Jian Xu<sup>1,†</sup>, Chao Bian<sup>2,3,4,†</sup>, Kunci Chen<sup>5,†</sup>, Guiming Liu<sup>6</sup>, Yanliang Jiang<sup>1</sup>, Qing Luo<sup>5</sup>, Xinxin You<sup>2,3</sup>, Wenzhu Peng<sup>1,7</sup>, Jia Li<sup>3</sup>, Yu Huang<sup>3</sup>, Yunhai Yi<sup>3</sup>, Chuanju Dong<sup>1,8</sup>, Hua Deng<sup>9</sup>, Songhao Zhang<sup>1</sup>, Hanyuan Zhang<sup>1</sup>, Qiong Shi<sup>2,3,10,\*</sup> and Peng Xu<sup>1,7,\*</sup>

<sup>1</sup>Key Laboratory of Aquatic Genomics, Ministry of Agriculture, CAFS Key Laboratory of Aquatic Genomics and Beijing Key Laboratory of Fishery Biotechnology, Chinese Academy of Fishery Sciences, Fengtai, Beijing, 100141, China, <sup>2</sup>BGI Research Center for Aquatic Genomics, Chinese Academy of Fishery Sciences, Shenzhen, Guangdong, 518083, China, <sup>3</sup>Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI, Shenzhen, Guangdong, 518083, China, <sup>4</sup>Centre of Reproduction, Development and Aging, Faculty of Health Sciences, University of Macau, Taipa, Macau, China, <sup>5</sup>Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, Guangdong, 510380, China, <sup>6</sup>CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Chaoyang, Beijing, 100029, China, <sup>7</sup>Fujian Collaborative Innovation Center for Exploitation and Utilization of Marine Biological Resources, Xiamen University, Xiamen, Fujian, 361102, China, <sup>8</sup>College of Fishery, Henan Normal University, Xinxiang, Henan, 453007, China, <sup>9</sup>Research Institute of Forestry Policy and Information, Chinese Academy of Forestry, Haidian, Beijing, 100091, China and <sup>10</sup>Laboratory of Aquatic Genomics, College of Ecology and Evolution, School of Life Sciences, Sun Yat-Sen University, Guangzhou, Guangdong, 510275, China

\*Correspondence address. Qiong Shi: BGI Research Center for Aquatic Genomics, Chinese Academy of Fishery Sciences, Shenzhen, Guangdong, 518083, China. E-mail: [shiqiong@genomics.cn](mailto:shiqiong@genomics.cn); Peng Xu: Key Laboratory of Aquatic Genomics, Ministry of Agriculture, CAFS Key Laboratory of Aquatic Genomics and Beijing Key Laboratory of Fishery Biotechnology, Chinese Academy of Fishery Sciences, Fengtai, Beijing, 100141, China. E-mail: [xupeng77@xmu.edu.cn](mailto:xupeng77@xmu.edu.cn)

<sup>†</sup>Contributed equally to this work.

## Abstract

**Background:** The Northern snakehead (*Channa argus*), a member of the Channidae family of the Perciformes, is an economically important freshwater fish native to East Asia. In North America, it has become notorious as an intentionally released invasive species. Its ability to breathe air with gills and migrate short distances over land makes it a good model for bimodal breath research. Therefore, recent research has focused on the identification of relevant candidate genes. Here, we performed whole genome sequencing of *C. argus* to construct its draft genome, aiming to offer useful information for

Received: 17 August 2016; Revised: 5 January 2017; Accepted: 25 February 2017

© The Author 2017. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

further functional studies and identification of target genes related to its unusual facultative air breathing. Findings: We assembled the *C. argus* genome with a total of 140.3 Gb of raw reads, which were sequenced using the Illumina HiSeq2000 platform. The final draft genome assembly was approximately 615.3 Mb, with a contig N50 of 81.4 kb and scaffold N50 of 4.5 Mb. The identified repeat sequences account for 18.9% of the whole genome. The 19877 protein-coding genes were predicted from the genome assembly, with an average of 10.5 exons per gene. Conclusion: We generated a high-quality draft genome of *C. argus*, which will provide a valuable genetic resource for further biomedical investigations of this economically important teleost fish.

**Keywords:** *Channa argus*; genome assembly; annotation; gene prediction

## Data description

### Introduction of *C. argus*

The Northern snakehead (*Channa argus*) is a special snakehead fish cultivated mainly in Asia and Africa for food, especially in China with an annual production of about 510 000 tons (worth ~1.6 billion US dollars) (Fig. 1). Genetic degradation caused by inbreeding of *C. argus* cultivation has led to higher susceptibility to diseases. Furthermore, *C. argus* is considered a serious invasive species in North America, due to its wide-range diet, parental care, and rapid colonization and expansion [1]. *C. argus* has a specialized aerial breathing organ, the suprabranchial chamber, which facilitates its aquatic-aerial bimodal breathing. Because of its aggressive status in ecosystem of rivers, lakes, and ponds, and little consumption of the *C. argus* in America for food, this leads to threats to the balance of ecosystems. For both economic and ecological consideration, it is vital to develop genomic resources for further genetic breeding studies or ecological research. So far, the genome sequence of *C. argus* has not been reported, and hence in our current study we performed genome sequencing, assembly, and annotation of this teleost species.

### *C. argus* genome sequencing on the Illumina platform

Genomic DNA was extracted from blood sample of a single female *C. argus* (Fishbase ID: 4799) using Qiagen GenomicTip100 (Qiagen). The fish was obtained from the Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, China. A whole-genome shotgun sequencing strategy was applied, and short-insert libraries (180, 500, and 800 bp) and long-insert libraries (3 and 5 kb) were constructed using the standard protocol provided by Illumina (San Diego, CA, USA). Paired-end sequencing with a  $2 \times 100$ -bp read length was

performed on the short-insert and long-insert libraries using the Illumina HiSeq2000 platform. In total, we generated about 140.3 Gb of raw reads, including 33.0, 36.9, 17.4, 26.5, and 26.5 Gb of reads from the 180-, 500-, 800-, 3-, and 5-kb libraries. After removal of low-quality and redundant reads, we obtained about 138.2 Gb of clean data for further *de novo* assembling of the *C. argus* genome.

### Estimation of *C. argus* genome size and sequencing coverage

All the cleaned reads were subjected to 17-mer frequency distribution analysis [2]. As the total number of *k*-mers was about  $5.90 \times 10^{10}$  and the peak of *k*-mers at a depth of 88, the genome size of *C. argus* was calculated to be 670.4 Mb using the following formula: genome size = *k*-mer number / peak.depth. Therefore, the sequencing coverage was found to be  $\sim 124.5 \times$  based on the estimated genome size.

### *De novo* genome assembly and quality assessment

For whole genome assembly, SOAPdenovo2 [3] was used with optimized parameters (-K 75) to construct contigs and original scaffolds by using the reads from short-insert libraries. All reads were then mapped onto contigs for scaffold construction by utilizing the paired-end information of long-insert libraries. Some intra-scaffold gaps were filled by local software using read-pairs in which one end uniquely mapped to a contig and the other end was located within a gap. Finally, a draft *C. argus* genome of 615.3 Mb was assembled, with a contig N50 size of 81.4 kb and a scaffold N50 size of 4.5 Mb (Table 1).

Subsequently, the Core Eukaryotic Genes Mapping Approach software [4] (version 2.3) with 248 conserved Core Eukaryotic



Figure 1: the Northern snakehead fish, *Channa argus*.

**Table 1:** summary of the *Channa argus* genome assembly and annotation

Genome assembly	
Contig N50 size (kb)	81
Contig number (>100 bp)	29 146
Scaffold N50 size (Mb)	4.5
Scaffold number (>100 bp)	5297
Total length (Mb)	615.3
Genome coverage (X)	224.6
The longest scaffold (bp)	18 736 006
Genome annotation	
Protein-coding gene number	19 877
Mean transcript length (kb)	16.5
Mean exons per gene	10.5
Mean exon length (bp)	175.0
Mean intron length (bp)	1537.3

Genes was utilized to evaluate completeness of genes. Our results demonstrated that the generated genome assembly covered 242 of the 248 Core Eukaryotic Gene sequences, suggesting a high level of completeness within the genome assembly. Alongside this, we also used BUSCO (version 1.22) [5] (the representative vertebrate gene set containing 3023 single-copy genes that are highly conserved in vertebrates) software to assess the quality of the generated genome assembly. The assessment demonstrated that the BUSCO value is 82.9%, containing C: 66% [D: 1.4%], F: 16%, M: 17%, n: 3023 (C: complete [D: duplicated], F: fragmented, M: missed, n: genes), suggesting a high quality of the generated assembly.

### Repeat sequence within the *C. argus* genome assembly

To analyze the *C. argus* genome, we employed Tandem Repeats Finder [6] (version 4.04) with core parameters set as “Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, and MaxPeriod = 2000” to identify tandem repeats. Simultaneously, RepeatModeler (version 1.04) and LTR.FINDER [7] were utilized to construct a *de novo* repeat library with default parameters. Subsequently, we used RepeatMasker [8] (version 3.2.9) to map our assembled sequences on the Repbase TE (version 14.04) [9] and the *de novo* repeat libraries to search for known and novel transposable elements (TEs). In addition, the TE-related proteins were annotated by using RepeatProteinMask software [8] (version 3.2.2). In summary, the total identified repeat sequences accounted for 18.94% of the *C. argus* genome (Table 2). Among them, long interspersed nuclear elements were the most abundant type of repeat sequences and occupy 8.92% of the whole genome.

**Table 2:** the detailed classification of repeat sequences of *Channa argus*

Type	Repbse TEs		TE protiens		De novo		Combined TEs	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	17 984 515	2.92	6 784 728	1.10	25 663 752	4.17	35 435 946	5.76
LINE	16 799 343	2.73	17 563 763	2.85	54 890 557	8.92	60 651 866	9.86
SINE	4 512 385	0.73	0	0	6 672 552	1.08	9 026 285	1.47
LTR	4 421 728	0.72	3 031 607	0.49	24 144 657	3.92	26 983 318	4.39
Other	8125	0.001	0	0	0	0	8125	0.001
Unknown	0	0	0	0	9 413 375	1.53	9 413 375	1.53
Total	41 585 442	6.76	27 363 267	4.45	103 162 115	16.77	116 545 270	18.94

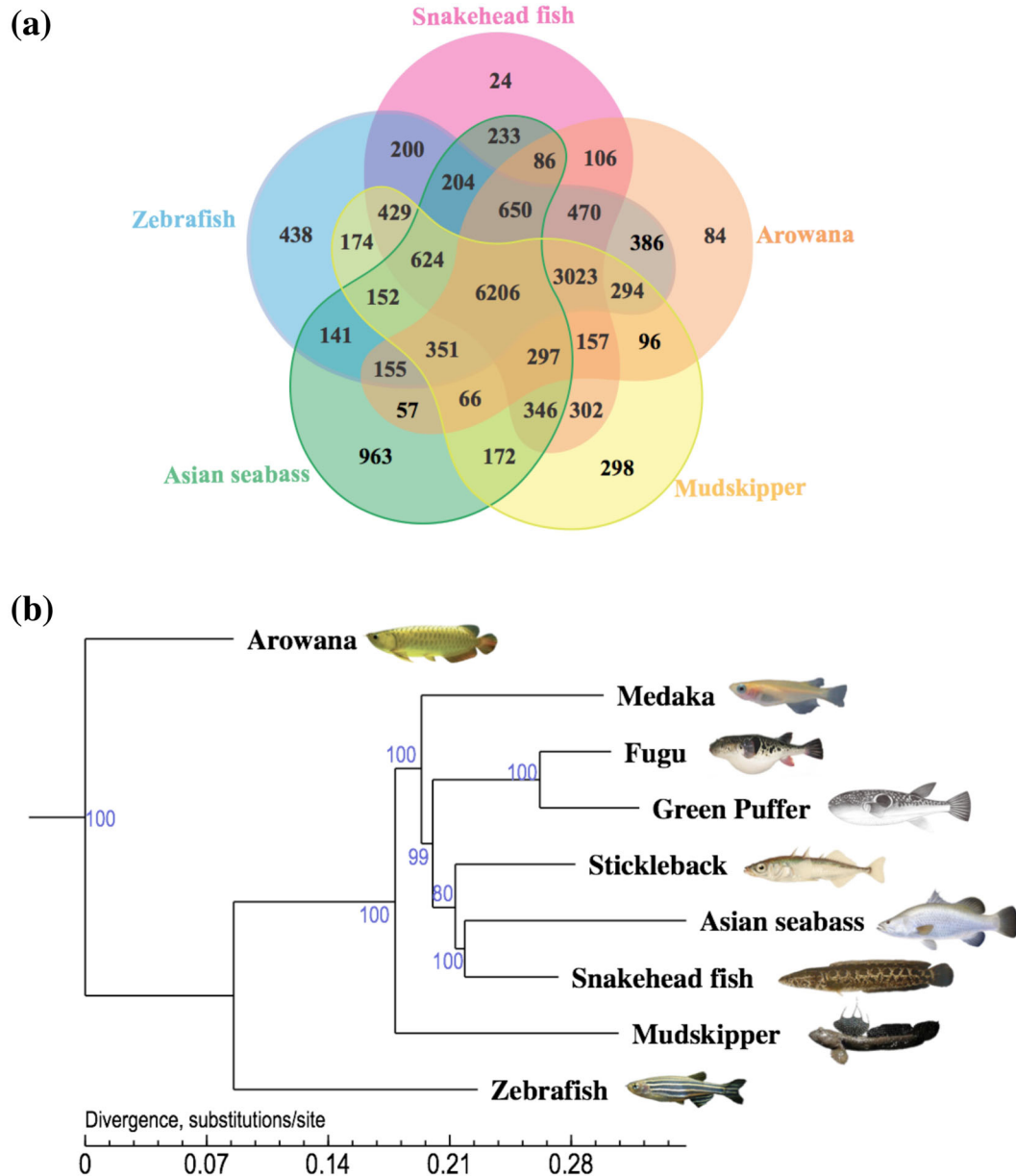
### Gene annotation

Gene annotation of the *C. argus* genome was conducted using several approaches, including transcriptome-based prediction, *de novo* prediction, and homology-based prediction. RNA-seq datasets of pooled 13 tissues were obtained from our previous work [10]. We mapped these RNA reads onto our genome assembly using TopHat1.2 software [11], and then we employed Cufflinks (version 2.2.1) [12] to predict the gene structures. Furthermore, we performed Augustus (version 2.5.5) [13], GlimmerHMM (version 3.0.1) [14], and GenScan (version 1.0) [15] softwares for *de novo* prediction on the repeat-masked *C. argus* genome assembly. The protein sequences of zebrafish (*Danio rerio*) [16], Japanese puffer (*Fugu rubripes*) [17], medaka (*Oryzias latipes*) [18], spotted green pufferfish (*Tetraodon nigroviridis*) [19] (the above 5 species were downloaded from Ensembl release 75), blue spotted mudskipper (*Boleophthalmus boddarti*) [20], and golden arowana (*Scleropages formosus*) [21] were mapped on the *C. argus* genome using TblastN with e-value  $\leq 1e-5$ . Subsequently, Genewise2.2.0 software [22] was employed to predict the potential gene structures on all alignments. Finally, the above three datasets were integrated to yield a comprehensive and nonredundant gene set using GLEAN (<https://sourceforge.net/projects/glean-gene/>) [23] with several filter steps (removing partial sequences or genes shorter than 150 bp or prematurely terminated/frame-shifted genes). The final total gene set was composed of 19 877 genes, with an average of 10.5 exons per gene (Table 1).

### Construction of gene families and phylogenetic tree

We downloaded the protein sequences of zebrafish [17], Japanese puffer [18], stickleback (*Gasterosteus aculeatus*) [24], spotted green pufferfish [20], and medaka [19] from the Ensembl Core database (release 75), and we also obtained the protein sequences of Asian seabass (*Lates calcarifer*) [25], blue spotted mudskipper [21], and golden arowana [22] from their corresponding ftp websites, respectively. The consensus proteome set of the above eight species and snakehead fish was filtered to remove those protein sequences <50 amino acids and resulted in a dataset of 190 566 protein sequences, which was used as the input file for OrthoMCL [26] to construct gene families. A total of 17 954 OrthoMCL families were built utilizing an effective database size of 190 566 sequences for all-to-all BLASTP strategy with an E-value of  $1e-5$  and a Markov Chain Clustering default inflation parameter. We further identified 24 gene families that were specific in the snakehead fish (Fig. 2a).

Subsequently, we selected 1918 single-copy (only one gene from each species) orthogroups from the above-mentioned 9 teleost species. We used MUSCLE (version 3.8.31) [27] to align the



**Figure 2:** genome evolution. (a) Orthologous gene families across five fish genomes (Snakehead fish, Zebrafish, Asian seabass, Mudskipper, and Arowana). (b) Phylogeny of ray-finned fishes (the arowana as the outgroup species).

protein sequences from the 1918 orthogroups, respectively. We also converted protein alignments to their corresponding coding DNA sequence alignments using an in-house perl script. All the translated coding DNA sequence sequences were then combined into one “supergene” for each species. Nondegenerated sites (4D) extracted from the supergenes were then joined into new sequence of each species to construct a phylogenetic tree (Fig. 2b) using MrBayes [28] (Version 3.2, with the GTR+gamma model).

## Conclusion

We report the first whole genome sequencing, assembly, and annotation of the Northern snakehead (*Channa argus*). The final draft genome assembly is approximately 615.3 Mb, accounting for 91.8% of the estimated genome size (670.4 Mb). We

also predicted 19877 protein-coding genes from the generated assembly.

The draft genome assembly will be valuable resource for genetic breeding, environmental DNA detection of invasive species, and biological studies on this economically important teleost fish. Based on these genomic data, researchers will be able to develop genetic markers for further quantitative trait locus and genome-wide association studies on growth traits. These markers will also be very useful for DNA barcoding in screening invasive *C. argus* for ecological protection.

## Availability of supporting data

The raw sequencing reads of all libraries have been deposited at NCBI (SRP078899). Further supporting data are available in the GigaScience database, GigaDB [29].

## Abbreviation

TE: transposable element.

## Author contributions

PX designed the study. JX, CB, GL, JL, HD, YH, YX, and QS assembled and annotated the genome. CB and YY performed the evolution analysis. JX, YJ, XY, QL, and HZ analyzed the data. WP, CD, SZ, and KC collected the sample and prepared the quality control. JX, CB, QS, and PX wrote the manuscript. QS and PX participated in discussions and provided advice. All authors read and approved the final manuscript.

## Acknowledgements

This work was supported by Central Public-interest Scientific Institution Basal Research Fund, CAFS (No. 2015C005, No. 2016HY-JC0301), the National Natural Science Foundation of China (No. 31422057, No.31402291), the National Infrastructure of Fishery Germplasm Resources of China (No. 2017DKA30470), Special Project on the Integration of Industry, Education and Research of Guangdong Province (No. 2013B090800017), Quality Inspection Programs of Scientific Research Project (No. 2015IK246), and Shenzhen Special Program for Future Industrial Development (No. JSGG20141020113728803).

## Competing interests

The authors declare that they have no competing interests.

## References

- Jiang Y, Feng S, Xu J et al. Comparative transcriptome analysis between aquatic and aerial breathing organs of *Channa argus* to reveal the genetic basis underlying bimodal respiration. *Mar Genomics* 29:89–96.
- Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27(6):764–70.
- Luo R, Liu B, Xie Y et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 2012;1(1):18.
- Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 2007;23(9):1061–67.
- Simao FA, Waterhouse RM, Ioannidis P et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31(19):3210–12.
- Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 1999;27(2):573–80.
- Xu Z, Wang H. LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 2007;35(Web Server issue):W265–68.
- Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. In: Editorial board, Baxevanis Andreas D et al. (eds.), *Current Protocols in Bioinformatics* 2009, Chapter 4:Unit 4 10.
- Jurka J, Kapitonov VV, Pavlicek A et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic Genome Res* 2005;110(1–4):462–67.
- Jiang Y, Feng S, Xu J et al. Comparative transcriptome analysis between aquatic and aerial breathing organs of *Channa argus* to reveal the genetic basis underlying bimodal respiration. *Mar Genomics* 2016: DOI: 10.1016/j.margen.2016.1006.1002.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;25(9):1105–11.
- Trapnell C, Williams BA, Pertea G et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;28(5):511–15.
- Stanke M, Steinkamp R, Waack S et al. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 2004;32(Web Server issue):W309–12.
- Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 2004;20(16):2878–79.
- Cai Y, Gonzalez JV, Liu Z et al. Computational systems biology methods in molecular biology, chemistry biology, molecular biomedicine, and biopharmacy. *BioMed Res Int* 2014;2014:746814.
- Howe K, Clark MD, Torroja CF et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 2013;496(7446):498–503.
- Aparicio S, Chapman J, Stupka E et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 2002;297(5585):1301–10.
- Kasahara M, Naruse K, Sasaki S et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 2007;447(7145):714–19.
- Jaillon O, Aury JM, Brunet F et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 2004;431(7011):946–57.
- You X, Bian C, Zan Q et al. Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nat Commun* 2014;5:5594.
- Bian C, Hu Y, Ravi V et al. The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Sci Rep* 2016;6:24501.
- Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res* 2004;14(5):988–95.
- Elsik CG, Mackey AJ, Reese JT et al. Creating a honey bee consensus gene set. *Genome Biol* 2007;8(1):R13.
- Jones FC, Grabherr MG, Chan YF et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 2012;484(7392):55–61.
- Vij S, Kuhl H, Kuznetsova IS et al. Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS genetics* 2016;12(4):e1005954.
- Li L, Stoeckert CJ, Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;13(9):2178–89.
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32(5):1792–97.
- Ronquist F, Teslenko M, van der Mark P et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biol* 2012;61(3):539–42.
- Xu J, Bian C, Chen K et al. Supporting data for the draft genome of the Northern snakehead, *Channa argus*. *Giga-Science Database*. 2017; <http://dx.doi.org/10.5524/100279>.