Chromosome-level genome assembly of the humpback puffer, *Tetraodon palembangensis*

Rui Zhang[1,†], Chang Li[1,†], Mengjun Yu[1,†], Xiaoyun Huang[1], Mengqi Zhang[1], Shanshan Liu[1], Shanshan Pan[1], Weizhen Xue[1], Congyan Wang[1], Chunyan Mao[1], He Zhang[1,2,*], Guangyi Fan[1] *

[1]BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China.

[2]Department of Biology, Hong Kong Baptist University, Hong Kong, China.

[†]These authors contributed equally to this work.

[*]For correspondence (e-mails fanguangyi@genomics.cn or zhanghe@genomics.cn), office phone number: 0532-55711134

ORCIDs number for authors:

Rui Zhang https://orcid.org/0000-0003-1921-9435

Chang Li, https://orcid.org/0000-0002-7823-3467

Mengjun Yu, https://orcid.org/0000-0003-2394-9692

Guangyi Fan, https://orcid.org/0000-0001-7365-1590

He Zhang, https://orcid.org/0000-0001-9294-1403

Abstract

The humpback puffer, *Tetraodon palembangensis*, also known as *Pao palembangensis*, is a species of poisonous freshwater pufferfish mainly distributed in Southeast Asia (Thailand, Laos, Malaysia and Indonesia). Despite interesting biological features, such as its very inactive nature, tetrodotoxin production and body expansion mechanisms, molecular research on the humpback puffer is still rare because of the lack of a high-quality reference genome. Here, we reported a first chromosome-level genome assembly of an adult humpback puffer, of which the genome size is 362 Mb with ~1.78 Mb contig N50 and ~15.8 Mb scaffold N50s. Based on the genome, ~61.5Mb (18.11%) repeat sequences were also identified, and totally 19,925 genes were annotated, 99.20%

of which could be predicted with function using protein-coding function databases. Finally, a phylogenetic tree was constructed with single-copy gene families from ten teleost fishes. The humpback puffer genome will be a valuable genomic resource to illustrate possible mechanisms of tetrodotoxin synthesis and tolerance, providing clues for future detailed studies of biological toxins.

Data Description

Background and Context

The humpback puffer, *Tetraodon palembangensis* (NCBI Taxonomy ID: 1820603, Fishbase ID: 25179), is widely distributed in Southeast Asian and prefers to live in alkalescent, warm (24-28 ℃), and slow-flowing rivers [1] (**Fig.1**). The female and male humpback puffers have a similar body size, but the male's hump at the back is much bigger than that of the female [2]. Becasuse of its beautiful skin colouration and patterns, the humpback puffer is a popular ornamental fish. Different from other species of predatory pufferfish, the humpback puffer is so inactive that it only moves when the food is right in front of it [1]. Furthermore, its body contains deadly toxins known as tetrodotoxin (TTX), and it can swell up to three times than its normal size as a defense mechanism when it feel threatened [1]. Previous studies have proved that the content of the toxicity in the humpback puffer varies greatly in different seasons, so it can be edible when its skin and internal organs are removed[3]. However, the wild population of the humpback puffer has declined in recent years due to the destruction of living conditions caused by the pressures of global warming and human fishing [4].

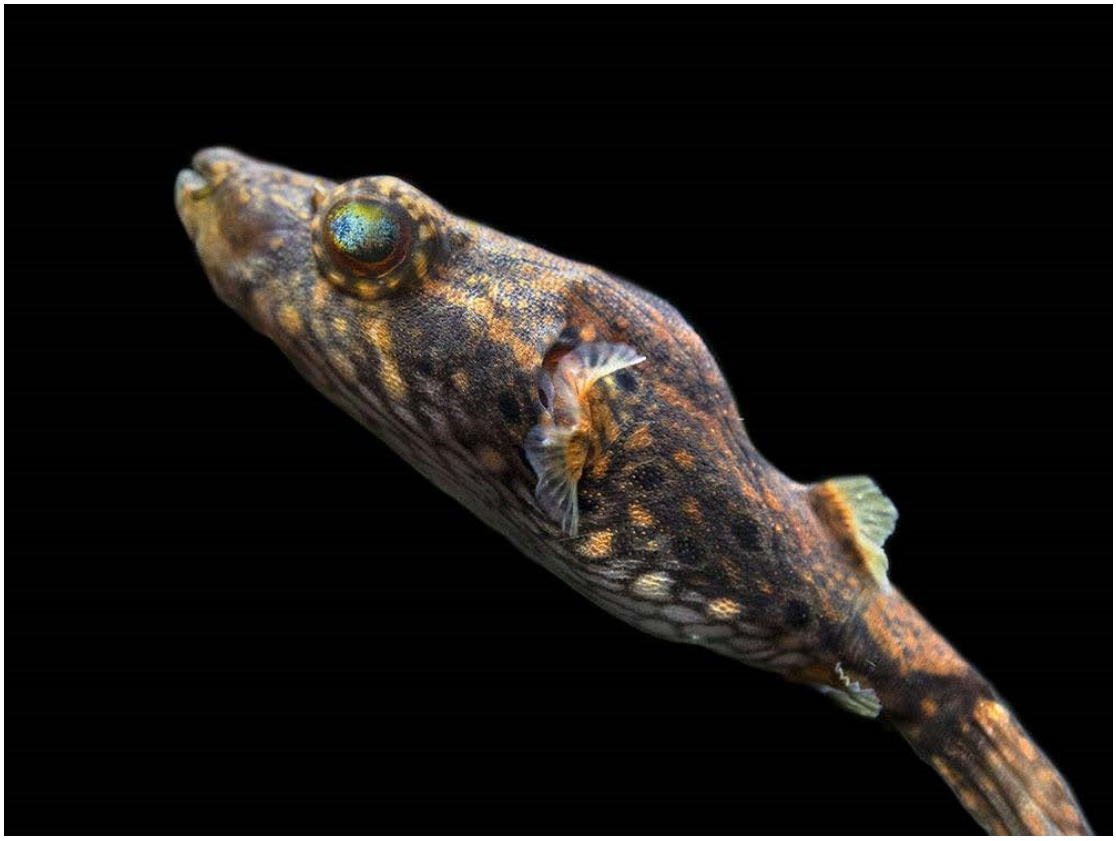

Figure 1 Photograph of *Tetraodon palembangensis.*

In addition to these biological characteristics, the compact genome size of humpback puffer is roughly about 385 Mb, which alongside other pufferfish species which have been used to study intron evolution[5], makes it an ideal model species for genetic study[6]. However, molecular research in the organism has been hampered due to the lack of a high-quality reference genome. In this study, we provided a chromosome-scale genome of an adult humpback puffer that will allow us to study features such as mechanisms of tetrodotoxin synthesis, expansion defense, body differences between males and females, and genome size. Comparative genomics analysis can help to better understand the phenotypic evolution and special gene families of the Tetraodontidae.

Methods

Sample collection and sequencing

The sample (CNGB ID: CNS0224034) used in this study came from an adult humpback pufferbought from YueHe Flower-Bird-Fish market in Guangzhou Province, China. Following DNA extraction protocols (available via protocols.io) [7], genomic DNA was

extracted from muscle tissues to construct libraries for sequencing. A paired-end single tube long fragment reads (stLFR) library [8] and a Hi-C library were constructed based on published protocols.io [7, 9], and sequenced on the DNBSEQ-G50 (formely know as BGISEQ-500) platform[10]. A PacBio library was constructed and sequenced on the Pacbio Sequel I system [11]. In total, we obtained 120 Gb (~312 X) raw stLFR data, 19 GB (~ 49X) raw Hi-C data, and 12 GB (~ 32X) raw Pacbio data (**Table 1**). All resources of this study were approved by the Institutional Review Board of BGI (IRB approval NO.FT17007).

To improve the assembly quality, low-quality reads with obvious sequencing error rate and adapters were filtered out from raw stLFR data by SOAPnuke (v1.6.5, RRID: SCR 015025)[12], and 62 Gb (152×) clean data were retained for further assembly. Raw Hi-C data was produced with a quality control using HiC-Pro (v. 2.8.0) [13], generating 5.4 Gb validated data which accounted for 28.81% of all data (**Table 1**).

Table 1. **Statistics of DNA sequencing data.**

Libraries	Reads length	Raw data		Valid data	
		Total bases (Gb)	Sequencing depth (X)	Total bases (Gb)	Sequencing depth (X)
stFLR	PE100 bp	120.4	311.69	62.6	162.60
Hi-C	PE 100bp	19.01	49.43	5.4	14.04
Pacbio	CN50: 32kb	12.3	31.98	--	--

Note: Sequencing depth = Total bases / Genome size, where the genome size is the result of K-mer estimation in Table 2.

Genome assembly

We used Jellyfish (v2.2.6, RRID: SCR_005491) with 58 Gb clean stLFR reads to perform the k-17mer analysis [14], estimating the humpback buffer genome size about 385 Mb (**Table 2 and Fig.1**). To assemble the humpback puffer genome, we firstly converted the format of stLFR reads and used Supernova assembler(v. 2.0.1, RRID: SCR_016756) to perform the draft assembly. Then we used GapCloser (v. 1.12, RRID: SCR_015026) [15] to fill gaps with stLFR reads. Next, to futher improve the assembly

quality, TGSgapFiller [16] was used to re-fill gaps with PacBio reads and Pilon (v. 1.22, RRID: SCR_014731) [17] was used to polish twice of the assembly. At this stage the draft genome assembly was about 362 Mb with 7.1Mb scaffold N50 and 1.8 Mb contig N50 (**Table 3**). Finally, we perfom the chromosomal-level assembly using the 3dDNA pipeline (v. 170123) [18] with Hi-C data, which anchored 91.2% of total sequences to 18 chromosomes, the length ranging from 11 Mb to 35 Mb (**Figure 2, Table 4**).

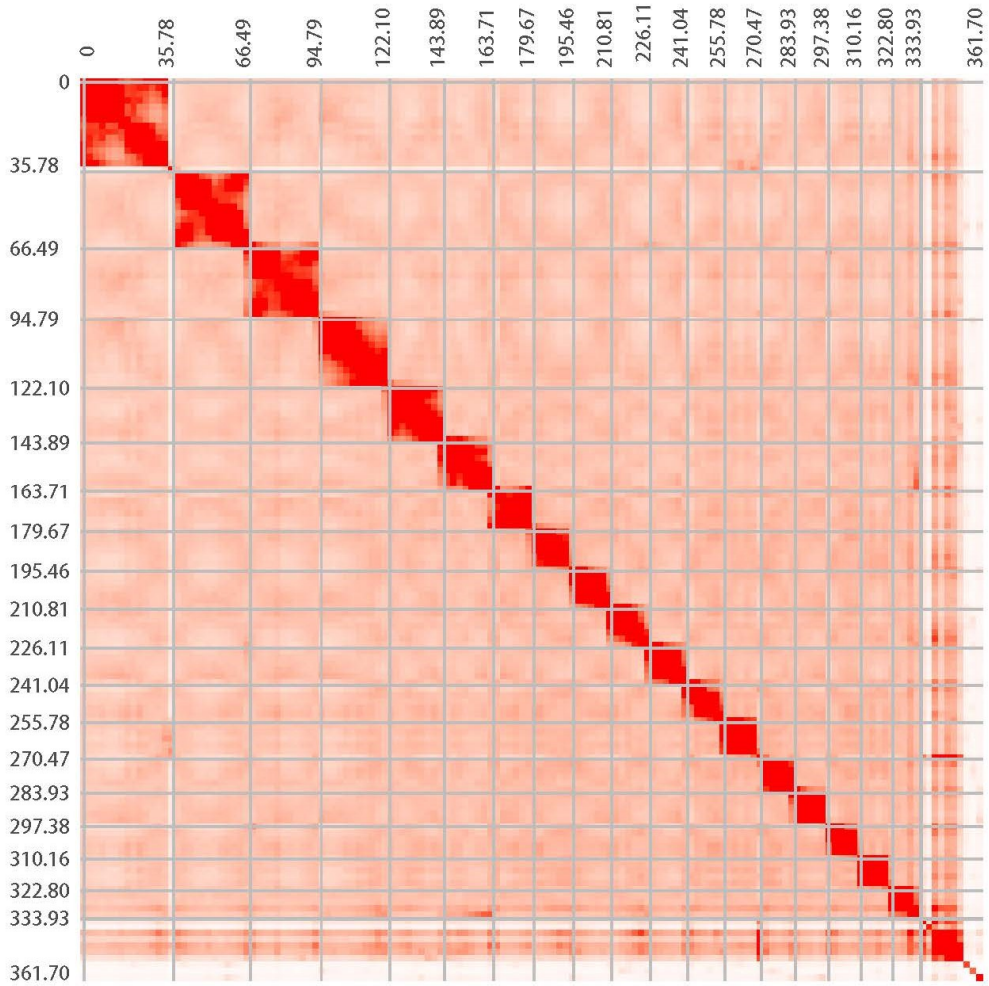

Figure 2 Heat map of chromosomal interaction of Hi-C assembly.

Table 2. Statistical of 17-mer analysis.

K-mer	K-mer Number	K-mer Depth	Heterozygosity	Genome Size (bp)
17	48,458,703,762	126	0.205%	384,592,887

Note: The genome size, G, was defined as G=K_num / K_depth, where the K_num is the total number of K-mers, and K_depth is the most frequently occurring frequency.

Table 3. Statistics of the draft assembly of the humpback puffer genome.

Statistics	Contig	Scaffold
Total Number (#)	5,291	6,190
Total length (bp)	361,704,206	360,427,744
Gap(N)(bp)	1,276,462	0
Average Length (bp)	68,362	58,227
N50 Length (bp)	7,059,990	1,830,664
N90 Length (bp)	453,057	157,209
Maximum Length (bp)	19,534,197	9,842,180
Minimum Length (bp)	682	48
GC content	44.66%	44.66%

Table 4. Statistics of the Hi-C assembly of the humpback puffer genome.

Statistics	Scaffold	Contig
Total Number (#)	5,366	6,435
Total length (bp)	361,698,760	360,427,744
Gap(N)(bp)	1,271,016	0
Average Length (bp)	67,406	56,011
N50 Length (bp)	15,808,960	1,794,775
N90 Length (bp)	11,014,520	117,115
Maximum Length (bp)	34,916,285	9,792,502
Minimum Length (bp)	682	48
GC content	44.66%	44.66%

Genomic annotation

For the annotation of repetitive sequences, we used two methods: one is aligning the genometo the Repbase library, by TRF (v.4.09) [19], RepeatMasker (v. 3.3.0, RRID: SCR 012954) and RepeatProteinMask (v. 3.3.0) [20] were then used to predict and classify

the repeatitive sequences; and the other is constructing the repeat library by RepeatModeler (v1.0.8, RRID: SCR_015027) and classifying transposable elements (TEs) by RepeatMasker (v. 3.3.0) [20]. The results of two methods were integrated to a total of 65 Mb repeat sequences and 59 Mb TEs, accounting for 18.11% (**Figure 3a, Table 5**) and 16.62% of the entire genome, respectively (**Figure 3a, Table 6**). In addition, the genes in the mitochondria was also annotated by MitoZ [21] (**Figure 3b**).

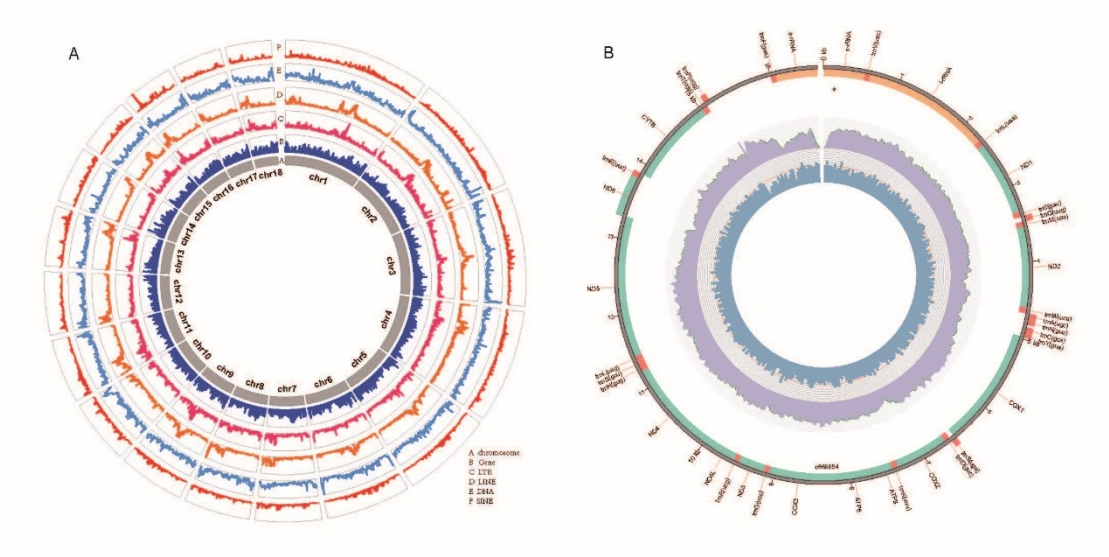

Figure 3 (A) Basic genomic elements of *Tetraodon palembangensis* genome. LTR, long terminal repeat; LINE, long interspersed nuclear elements; SINE, short interspersed elements. **(B) Physical map of mitochondrial assembly.**

Table 5. Statistics of repeat sequence.

Type	Repeat Size(bp)	% of genome
TRF	9,050,571	2.52
RepeatMasker	34,142,529	9.50
RepeatProteinMask	17,674,660	4.92
De novo	57,492,865	16.00
Total	65,080,476	18.11

Table 6. Sstatistics of transposable elements (TE).

Type	RepBase TEs	TE Proteins	*De novo*	Combined TEs

	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome
DNA	12,412,491	3.45	1,086,262	0.30	16,089,219	4.48	22,470,373	6.25
LINE	18,430,929	5.13	13,695,154	3.81	29,418,621	8.19	33,421,782	9.30
SINE	524,061	0.15	0	0.00	289,252	0.08	789,086	0.22
LTR	5,393,600	1.50	2,906,451	0.81	12,758,934	3.55	15,803,098	4.40
Other	8,290	0.00	228	0.00	0	0.00	8,518	0.00
Unknown	0	0.00	0	0.00	3,202,764	0.89	3,202,764	0.89
Total	34,142,529	9.50	17,674,660	4.92	55,052,617	15.32	59,729,335	16.62

For the gene structural annotation, we performed *de novo* prediction using AUGUSTUS (v3.1, RRID: SCR_008417)[22], GlimmerHMM(v3.0.4, RRID: SCR_002654) [23] and Genscan (RRID: SCR_013362) [24]. We also used TRITNITY (v2.8.5, RRID: SCR_013048) [25] to assemble draft transcriptome with RNA-seq reads, then used HISAT2(v2.1.0, RRID: SCR_015530)-StringTie (v1.3.4, RRID: SCR_016323)[26] and PASA (v2.3.3, RRID: SCR 014656)-TransDecoder (RRID: SCR_017647) [27] to predict transcripts. Furthermore, we used GeneWise (v2.4.1, RRID: SCR_015054) [28] for homologous annotation with protein data obtained from National Center for Biotechnology Information (NCBI) database of following eight species: *Danio rerio* (NCBI, GenBank ID:50), *Cynoglossus semilaevis* (NCBI, GenBank ID:11788), *Gasterosteus aculeatus* (NCBI, GenBank ID:146), *Gadus morhua* (NCBI, GenBank ID:2661), *Larimichthys crocea* (NCBI, GenBank ID:12197), *Oreochromis niloticus* (NCBI, GenBank ID:197), *Oryzias latipes* (NCBI, GenBank ID:542), and *Takifugu rubripes* (NCBI, GenBank ID:63). Finally, the above three kinds of evidence were integrated by EVidenceModeler (v1.1.1, RRID: SCR_014659)[29], generating 19,925 non-redundant coding genes, each containing an averge of 11 exons and 1,945 bps coding region (**Table 7**).

For the gene function annotation, we aligned the 19,925 genes to the databases of TrEMBL (UniProtKB, RRID: SCR_004426) [30], Swissprot [31], Kyoto Enyclopedia of

Genes and Genomes (KEGG, RRID: SCR_012773) [32], Gene Ontology (GO, RRID: SCR 002811) [33] and InterProScan (RRID: SCR_005829)[34]. Overall, 90.1% of all genes could be annotated with function (**Table 8 and Fig 2**).

Table 7. Statistics of the predicted genes in the humpback puffer genome.

	Gene set	Gene number	Average transcript length (bp)	Average CDS length (bp)	Average intron length (bp)	Average exon length (bp)	Average exons per gene
Homolog	*Cynoglossus semilaevis*	19,686	9,136.12	1,715.14	856.1	177.4	9.67
	Danio rerio	19,348	15,066.80	1,577.39	1,718.92	178.28	8.85
	Gadus morhua	20,361	7,040.85	1,441.77	744.62	169.23	8.52
	Gasterosteus aculeatus	26,630	6,896.85	1,474.53	686.88	165.79	8.89
	Larimichthys crocea	21,220	9,425.27	1,690.06	902.2	176.53	9.57
	Oreochromis niloticus	24,562	9,494.62	1,789.15	829.18	173.82	10.29
	Oryzias latipes	23,332	8,859.46	1,467.62	962.5	169.08	8.68
	Takifugu rubripes	19,635	7,762.47	1,645.04	707.22	170.47	9.65
De novo	Augustus	21,662	7,149.08	1,725.00	659.42	186.98	9.23
	Genscan	25,933	9,855.53	1,791.43	990.72	196.01	9.14
	GlimmerHMM	99,722	1,192.96	594.53	378.42	230.31	2.58
Transcript	Pasa & Transdecoder	33,965	4,856.71	1,186.88	558.33	156.73	7.57
	Hisat & Stringtie	31,664	5,551.59	1,303.52	608.39	163.3	7.98
EVM		19,925	9,418.80	1,945.48	757.65	179.08	10.86

Note: The EVM gene set contains the integrated result of *De novo* genes predictions, Homolog genes predictions and Transcript annotation by EVM software.

Table 8. Statistics of the functional annotation.

Database	Number	Percentage (%)
Total	20,057	100.00%
SwissProt	17,333	86.42%
KEGG	16,182	80.68%
TrEMBL	18,037	89.93%
Interpro	17,108	85.30%
Overall	18,064	90.06%

Genome Evolution

To study the evolutionary status of humpback puffer among bony fishes, we clustered gene families by alignment using protein sequences of the humpback puffer and other 9 teleosts (*Xiphophorus maculatus, Gasterosteus aculeatus*, *Sebastes schlegelii*, *Oryzias latipes*, *Gadus morhua*, *Oreochromis niloticus*, *Tetraodon nigroviridis*, *Danio rerio*, *Takifugu rubripes*) with an evalue cutoff of 1e-7. All these species' protein-coding genes were downloaded from NCBI except *S. schlegelii*' [35] was obtained from CNSA (Accession ID: CNP0000222). To improve analysis quality, we removed genes with frame shift or less than 50 amino acids as well as redundancy copies, only kepping the longest transcripts for comparative genomic analysis. A total of 21,022 gene families were identified, among which 40 gene families were unique to the humpback puffer (**Table 9**).

Of all gene families, we identified 4,461 single-copy ortholog genes containing 3,584,782 amino acids for further evolutionary analyses. We firstly used MUSCLE (v3.8.31, RRID: SCR_011812)[36] to align these ortholog gene sequences to each other, constructing a alignment matrix. Then we utilized the matrix to build a Maximum Likelihood method (ML) tree by RAxML (v8.2.4, RRID: SCR_006086)[37], applying nucleotide substitution model- JTT and 100 replicates[38] **(Fig 4).** Next, we calucated the divergence time among these teleosts by MCMCTree included in PAML (v4.7a, RRID: SCR_014932)[39] with parameters of '--rootage 500 -clock 3 -alpha 0.431879' and the fossil correction time obtained from Timetree[40]. The evolutionary tree showed

that *T. palembangensis* and *T. nigroviridis* diverged about 18.1 millon years ago **(Fig 4).**

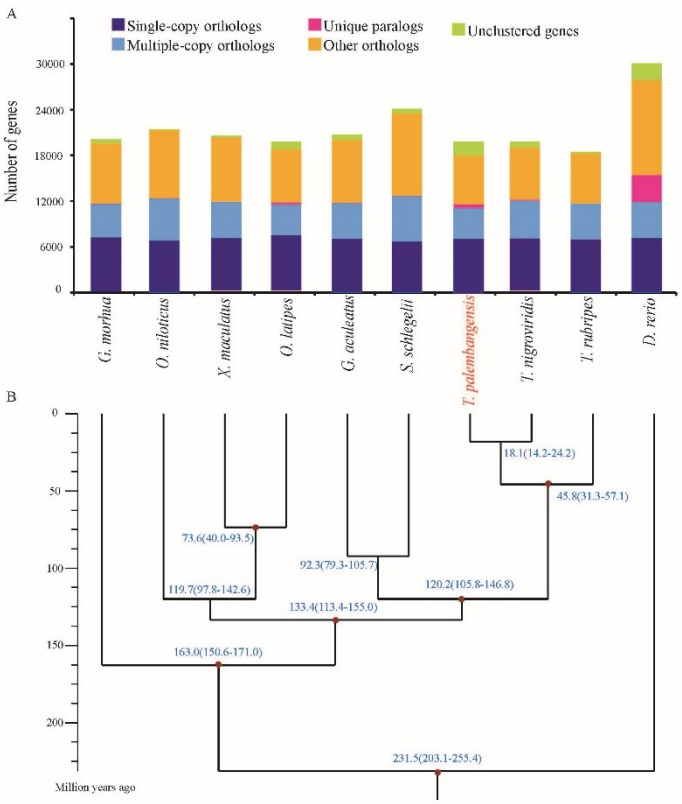

Figure 4 Comparative analysis of the *Tetraodon palembangensis* and 9 teleosts. **(A)** Clustering of gene families. **(B)** Phylogenetic tree constructed with the single-copy gene families.

Table 9. Statistics of gene family clustering.

Species	Total genes	Unclustered genes	Families	Unique families	Average genes per family
D. rerio	30,067	2,171	18,635	735	1.5
G. aculeatus	20,756	728	15,995	11	1.25
G. morhua	19,987	525	15,650	11	1.24
S. schlegelii	24,094	558	16,991	30	1.39
O. latipes	19,535	984	14,873	71	1.25
O. niloticus	21,431	160	15,811	13	1.35
T. nigroviridis	19,544	805	14,916	50	1.26
T. palembangensis	19,796	690	15,830	40	1.21
T. rubripes	18,459	207	14,733	6	1.24
X. maculatus	20,356	271	16,446	3	1.22

Data Validation and quality control

To demonstrate the quality of genome assembly and gene set, we performed qulity evaluation using the vertebrata database from Benchmarking Universal Single-Copy Orthologs (BUSCO v.3.0.2, RRID: SCR_015008) [41]. The results showed that 96.2% and 93.1% complete BUSCOs were covered by the genome assembly and gene set, respectively. **(Table 10)**.

Table 10. Statistics of the BUSCO assessment

	Gene Set		Assembly	
Types of BUSCOs	Number	Percentage	Number	Percentage
Complete BUSCOs	2408	93.1%	2486	96.2%
Complete single-copy BUSCOs	2348	90.8%	2438	94.3%
Fragmented BUSCOs	81	3.1%	64	2.5%
Missing BUSCOs	97	3.8%	36	1.3%
Total BUSCO groups searched	2586	100%	2586	100%

Re-use potential

In summary, we assemble a first chromosome-level genome and conduct genomic annotation of humpback puffer. These resources will be helpful to study the body

expansion mechanism, the synthesis mechanism and treatment of the tetrodotoxin, as well as the evolution of freshwater puffer. Futhermore, as part of the Fish 10K program[42], the genome of humpback puffer will contribute to fill a gap in the phylogenetic tree of life.

Availability of supporting data

We have deposited the project at CNGB Nucleotide Sequence Archive (CNSA) where the accession ID is CNP0001025. The genomic data can be obtained in *GigaScience* Database[43]. The sequencing data have been deposited at National Center for Biotechnology Information (NCBI) where the bioproject accession ID is PRJNA597275.

Abbreviations

bp: base pair; Gb: giga base; kb: kilo base; Mb: mega base; stLFR: single tube long fragment reads; TE: transposable element. transposable elements (TEs); NCBI: National Center for Biotechnology Information; KEGG: Kyoto Enyclopedia of Genes and Genomes; GO: Gene Ontology; ML: Maximum Likelihood; BUSCO: Benchmarking Universal Single-Copy Orthologs; CNSA: CNGB Nucleotide Sequence Archive.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

He Zhang and Guangyi Fan designed this project. Mengqi Zhang prepared the samples. Shanshan Liu, Shanshan Pan, Weizhen Xue, Congyan Wang and Chunyan Mao conducted the experiments. Rui Zhang, Chang Li, Mengjun Yu and Xiaoyun Huang did the analyses. Rui Zhang, Chang Li and Mengjun Yu wrote and revised the manuscript.

Acknowledgements

This work was supported by the special funding of "Blue granary" scientific and technological innovation of China (2018YFD0900301-05). We also thank for the technical supports of stLFR library construction and sequencing from China National Genebank.

Reference

1. Saitanu, K., et al., Toxicity of the freshwater puffer fish Tetraodon fangi and T. palembangensis from Thailand. Toxicon, 1991. **29**(7): p. 895-897.
2. Subamia, I.W., S. Sudarto, and W. Purbowasito, SEX DETERMINATION IN INDONESIAN PUFFERFISH Tetraodon palembangensis Bleeker, 1852: IMPLICATION FOR AQUACULTURE AND CONSERVATION. Indonesian Aquaculture Journal, 2011. **6**(1): p. 37-45.
3. Jaillon, O., et al., Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature, 2004. **431**(7011): p. 946-957.
4. Sutaria, D., et al., Humpback dolphins (Genus Sousa) in India: an overview of status and conservation issues, in Advances in Marine biology. 2015, Elsevier. p. 229-256.
5. Loh, Y.-H., S. Brenner, and B. Venkatesh, Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon). Molecular biology and evolution, 2008. **25**(3): p. 526-535.
6. Hedges and S. Blair, The origin and evolution of model organisms. Nature Reviews Genetics, 2002. **3**(11): p. 838-849.
7. Shao CW, L.C., Wang N, et al., Protocols for "Hi-C library preparation for the Lateolabrax maculatus genome." protocols.io, 2018. http://dx.doi.org/10.17504/protocols.io.ss4eegw.
8. Wang, O., et al., Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome research, 2019. **29**(5): p. 798-808.
9. Belton, J.M., et al., Hi–C: A comprehensive technique to capture the conformation of genomes. Methods, 2012. **58**(3).
10. Huang, J., et al., BGISEQ-500 WGS library construction. protocols.io, 2018. https://dx.doi.org/10.17504/protocols.io.ps5dng6.
11. Rhoads, A. and K.F. Au, PacBio sequencing and its applications. Genomics, proteomics & bioinformatics, 2015. **13**(5): p. 278-289.
12. Chen, Y., et al., SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience, 2018. **7**(1): p. gix120.
13. Servant, N., et al., HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome biology, 2015. **16**(1): p. 259.
14. Li, R., et al., The sequence and de novo assembly of the giant panda genome. Nature, 2010. **463**(7279): p. 311-317.

15. Luo, R., et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 2012. **1**(1): p. 2047-217X-1-18.
16. Liu, X., TGS-GapCloser: fast and accurately passing through the Bermuda in large genome using error-prone third-generation long reads. bioRxiv, 2019: p. 831248.
17. Walker, B.J., et al., Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one, 2014. **9**(11).
18. Liu, X., Protocols for "The pipeline of Hi-C assembly." 2018, protocols.io. https://dx.doi.org/10.17504/protocols.io.qradv2e.
19. Benson, G., Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research, 1999. **27**(2): p. 573-580.
20. Tarailo-Graovac, M. and N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics, 2009. **25**(1): p. 4.10. 1-4.10. 14.
21. Meng, G., et al., MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic acids research, 2019. **47**(11): p. e63-e63.
22. Stanke, M., et al., AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research, 2006. **34**(suppl_2): p. W435-W439.
23. Majoros, W.H., M. Pertea, and S.L. Salzberg, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics, 2004. **20**(16): p. 2878-2879.
24. Burge, C. and S. Karlin, Prediction of complete gene structures in human genomic DNA. Journal of molecular biology, 1997. **268**(1): p. 78-94.
25. Grabherr, M.G., et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 2011. **29**(7): p. 644.
26. Pertea, M., et al., Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols, 2016. **11**(9): p. 1650.
27. Campbell, M.A., et al., Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC genomics, 2006. **7**(1): p. 327.
28. Doerks, T., et al., Systematic identification of novel protein domain families associated with nuclear functions. Genome research, 2002. **12**(1): p. 47-56.
29. Haas, B.J., et al., Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology, 2008. **9**(1): p. R7.
30. Elsik, C.G., et al., Creating a honey bee consensus gene set. Genome biology, 2007. **8**(1): p. R13.
31. Bairoch, A. and R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research, 2000. **28**(1): p. 45-48.
32. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 2000. **28**(1): p. 27-30.
33. Consortium, G.O., The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 2004. **32**(suppl_1): p. D258-D261.
34. Jones, P., et al., InterProScan 5: genome-scale protein function classification. Bioinformatics, 2014. **30**(9): p. 1236-1240.
35. He, Y., et al., A chromosome__evel genome of black rockfish, Sebastes schlegelii, provides insights into the evolution of live birth. Molecular Ecology Resources, 2019. **19**(5).
36. Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 2004. **32**(5): p. 1792-1797.

37. Stamatakis, A., RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 2014. **30**(9): p. 1312-1313.
38. Jones, D.T., W.R. Taylor, and J.M. Thornton, The rapid generation of mutation data matrices from protein sequences. Bioinformatics, 1992. **8**(3): p. 275-282.
39. Yang, Z., PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics, 1997. **13**(5): p. 555-556.
40. Sudhir, K., et al., TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology & Evolution, 2017(7): p. 1812.
41. Simão, F.A., et al., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 2015. **31**(19): p. 3210-3212.
42. Fan, G., et al., Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). GigaScience, 2020. **9**(8).
43. Zhang R; Li C; Yu M; Huang X; Zhang M; Liu S; Pan S; Xue W; Wang C; Mao C; Zhang H; Fan G. Genome data for the chromosome-level assembly of the humpback puffer, Tetraodon palembangensis. *GigaScience* Database 2020. http://dx.doi.org/10.5524/100755.