Chromosome-level genome assembly of *Pterygoplichthys pardalis* reveals genetic basis of extensive invasion

Wangxiao Xia[1,†], Haorong Li[3,†], Yaowen Liu[4,†], Hui Jiang[5,†], Yonghong Wu[1], Yuanwei Zhang[6,*], Lixian Xu[2,*], Xingchun Gou[1,*]

1. Shaanxi Key Laboratory of Brain Disorders, Institute of Basic Translational Medicine, Xi'an Medical University, Xi'an 710021, China.
2. State Key Laboratory of Military Stomatology & National Clinical Research Center for Oral Diseases & Shaanxi Engineering Research Center for Dental Materials and Advanced Manufacture, Department of Anesthesiology, School of Stomatology, Fourth Military Medical University, Xi'an 610100, China.
3. School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710072, China.
4. College of Veterinary Medicine, Yunnan Agricultural University, Kunming 650231, China.
5. College of Life Sciences, Hainan Normal University, Haikou 571158, China.
6. State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650199, China.

[†]These authors contributed equally to this work.

[*]Corresponding author:
Xingchun Gou (gouxingchun@189.cn);
Lixian Xu (xlx116@fmmu.edu.cn);
Yuanwei Zhang (zhangyuanwei@mail.kiz.ac.cn).

Running title: Chromosome-level genome of *P. pardalis*

Abstract

The Amazon sailfin catfish (*Pterygoplichthys pardalis*), which belongs to the *Loricariidae* family, is an invasive species that has caused massive damage to the ecological environment. However, a high-quality reference genome for this catfish species has not yet been reported. Here, we successfully assembled the first chromosome-level high-quality genome of *P. pardalis* using data produced from multiple sequencing platforms. The assembled genome contains 26 chromosomes, with a scaffold N50 of 49.47 Mb. Different evaluation methods indicated the high connectivity and accuracy genome we got. In total, 23 859 protein-coding genes were predicted in the genome, 22 169 (92.92%) of which were functionally annotated in public databases. Phylogenetic analysis showed that *P. pardalis* was clustered with all other catfish studied and diverged from their common ancestor 132.5 million years ago. Whole-genome collinearity analysis indicated the chromosome 6 of *P. pardalis* was aligned to two distinct chromosomes in *Ameiurus melas*, *Pangasianodon hypophthalmus*, and *Ictalurus punctatus*, suggesting the occurrence of potential chromosomal fusion/fission events. Furthermore, many immune system-related genes were expanded in the *P. pardalis* genome, which may have contributed to their adaptive traits to highly polluted environments and successful invasion. This study not only provides insights into the genetic basis of the successful invasion of *P. pardalis*, but also provides important data for comparative genomic analysis of *P. pardalis* in Siluriformes in the future.

Keywords

Invasive species, Catfish, *Pterygoplichthys pardalis*, Genome assembly, Adaptive evolution, Immune system.

Introduction

Pterygoplichthys pardalis is a neotropical sucker mouth catfish belonging to the family *Loricariidae* (Bowen, 2016). As an omnivorous species, these catfish feed on algae, organic material, small invertebrates, and sediment (Bowen, 1983; Delariva & Agostinho, 2010). Significant changes in their gastric system, which acts as an

additional respiratory organ, allow them to survive in highly polluted environments with low dissolved oxygen (Da Cruz et al., 2013; Hussan et al., 2016). These catfish can also withstand cold temperatures and drought by burrowing, even when the water level is below the burrow opening, and can survive out of water for many hours (Burgess, 1989; Nico & Martin, 2001). These characteristics are highly beneficial for survival under harsh conditions. However, although this species is classified as non-edible and non-commercial due to the bony plates covering its skin and low muscle content, it has been exported to many countries as an exotic ornamental aquarium fish, with subsequent accidental or careless release into various wild environments (Ebenstein et al., 2015; Krishnakumar et al., 2009; Nurubhasha et al., 2019). Of concern, their rapid growth and high reproductive capacity, combined with a lack of natural predators, have allowed these catfishes to increase their populations within a short period of time and successfully invade many areas (Hui et al., 2017; Nico & Martin, 2001). With their expanded distribution into many tropical, subtropical, and intercontinental warm waters, these catfishes have become a considerable threat to aquatic biodiversity and environmental and economic health (Anguebes et al., 2019; Mohammad et al., 2018; Wakida-Kusunoki et al., 2007). Not only do they compete with native species for food resources and reduce native populations by predating on eggs and young fish, threatening the local ecological chain, but they also cause damage to lake and levees when burrowing to spawn (Hill & Sowards, 2015; Orfinger & Goodding, 2018). Once these invaders establish a population, they are very difficult to eradicate.

As a well-known representative species in *Loricariidae* and a damaging invasive pest, *P. pardalis* has attracted intense interest among ecological biologists (Anguebes et al., 2019; Chaichana et al., 2012; Hoover, Killgore & Cofrancesco, 2004; Krishnakumar et al., 2009). New insights into this species should help identify effective ways to control the population size within a reasonable range. To date, however, only one mitochondrial genome sequence and one draft nuclear genome (only with a short contig N50 length of 4.15 kb) have been published, with no high-quality genome yet reported (Liu et al., 2016). That's may the reason that studies

associated with the genetic basis of extensively invasive ability for *P. pardalis* were still in the single gene level (Baldissera et al., 2019; Barra et al., 1981; Bossa et al., 1982). Therefore, systematic and comprehensive analyses of the species at the whole-genome level are required.

In this study, we aimed to elucidate the genetic basis for the successful invasion ability of *P. pardalis*. We assembled the first chromosome-level reference genome of *P. pardalis* and annotated its protein-coding genes. Phylogenetic analysis indicated that *P. pardalis* diverged from the common ancestor of all other catfishes studied 132.5 million years ago (Mya). In addition, *P. pardalis* exhibited a much faster evolutionary rate relative to the other catfish examined except *B. yarrelli*. Genome-wide collinearity analysis suggested a potential chromosomal fusion/fission event on chromosome 6 of *P. pardalis*. Furthermore, comparative genomic analyses indicated that various immune-related genes were expanded in the *P. pardalis* genome, which may contribute to their robust adaptation to harsh conditions (especially highly polluted environments) and worldwide invasion. Taken together, this study not only reveals the genetic basis underlying the ability of *P. pardalis* to survive in harsh benthic environments, but also provides an important genetic resource for further studies on the development and environmental adaptation of this species.

Materials and Methods

Data acquisition

Genomic DNA was extracted from the muscle tissue of one armored catfish (*P. pardalis*) individual using a Qiagen Blood & Cell Culture DNA Mini Kit. To obtain a high-quality chromosome-level genome assembly, data from multiple sequencing platforms were acquired: 1) A short-insert paired-end library was prepared and sequenced on the Illumina NovaSeq 6000 platform; 2) A Nanopore library was prepared and sequenced on 26 flow-cells using Nanopore PromethION 48 (Oxford Nanopore, Oxford, UK); 3) A Hi-C library was constructed and sequenced using the Illumina NovaSeq 6000 platform; 4) To aid genome annotation, total RNA from the

muscle was extracted using a TRIzol Kit (Life Technologies) and used for library construction and sequencing on the Illumina NovaSeq 6000 platform.

Quality control of sequencing data

Data from the Illumina and Nanopore sequencing platforms were acquired for high-quality genome assembly. For Illumina reads, adaptor sequences and polymerase chain reaction (PCR) duplicates of all paired-end reads were removed. Any Illumina reads with more than 5% unknown bases or more than 30 low-quality bases along with their paired-end reads were discarded. For Nanopore reads, reads with a mean quality score > 7 were retained and used for further analysis.

Genome size estimation

The *P. pardalis* genome size was estimated using *k*-mer analysis. All cleaned short-insert Illumina reads were used for 17-mer frequency analysis to investigate genome size with the following formula: G = Knum/Kdepth, where G is the estimated genome size, Knum is the total number of 17-mers, and Kdepth is the depth of the 17-mer peak.

Genome assembly

The genome was assembled using the following steps: 1) Nanopore long reads were assembled into contigs using NextDenovo (v2.2) with parameters: read_cutoff = 1 k, seed_cutoff = 59754 and blocksize = 5g. 2) Clean short reads produced by the Illumina short-insert library were mapped to the assembled contigs using BWA (v0.7.17) (Li & Durbin, 2009) and further correction was performed using Pilon (v1.22) (Walker et al., 2014) with two iterations. 3) Hi-C sequencing reads were mapped to the corrected contigs, and chromosome construction was performed with Juicer (v1.5.7) (Durand et al., 2016) and 3D *de novo* assembly (v180922) (Dudchenko et al., 2017).

Genome annotation

Tandem repeats in the genome were identified using Tandem Repeat Finder (v4.07) (Benson, 1999). Non-interspersed repeats in the genome were annotated using RepeatMasker (v4.1.0) (Nishimura, 2000). Transposable elements (TEs) in the genome were annotated at both the DNA and protein levels. The *de novo* repeat library at the DNA level was constructed using RepeatModeler (v1.0.4) (http://www.repeatmasker.org/ RepeatModeler). The genome assembly was searched against Repbase (v23.06) using RepeatMasker to identify homologous repeats. RM-BLASTX within RepeatProteinMask (v4.1.0) was used to query the TE protein database at the protein level.

Three methods were used to predict protein-coding gene sequences based on the masked genome of repetitive sequences. For *de novo*- based prediction, the transcripts were *de novo* assembled based on the RNA-seq data using Bridger (r2014-12-01) (Chang et al., 2015), with the assembled sequences then filtered and primary predication performed using the Program to Assemble Spliced Alignment (PASA) (v2.1.0) (Haas et al., 2003) pipeline and AUGUSTUS (v2.5.5) (Stanke et al., 2003). Protein sequences, including *Bagarius yarrelli* (GCA_005784505.1), *Ameiurus melas* (GCA_012411365.1), *Ictalurus punctatus* (GCF_001660625.1), *Pangasianodon hypophthalmus* (GCF_009078355.1), *Tachysurus fulvidraco* (GCF_003724035.1), *Hemibagrus wyckioides* (GCA_019097595.1), *Silurus meridionalis* (GCF_014805685.1), *Clarias magur* (GCA_013621035.1), *Danio rerio* (GCF_000002035.6), *Pelteobagrus fulvidraco* (http://gigadb.org/dataset/100506), and *Glyptosternon maculatum* (https://doi.org/10.1093/gigascience/giy104), were downloaded for homology-based prediction of the coding genes. The longest transcript of each gene was selected and those genes with early termination sites were removed. Using the Basic Local Alignment Search Tool (BLAST) (v2.2.26) with an *e-value* threshold of 1e-5, we performed homology-based annotation and then conjoined by Solar (v0.9.6) software. GeneWise (v2.4.1) (Birney, 2004) was then used to predict the exact gene structure of each BLAST hit. For transcripts-based prediction, RNA-seq reads were directly mapped to the assembled genome using Blat (v34) (Kent, 2002) and spliced alignments were linked using the PASA pipeline

(v2.1.0) (Haas et al., 2003). Finally, the predicted coding genes derived from the three methods were integrated using EvidenceModeler (r2012-06-25) (Haas et al., 2008).

All predicted protein-coding genes were aligned to multiple databases, including InterPro, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), UniProt/SwissProt, UniProt/TrEMBL, and Non-Redundant Protein Sequence Database (NR database of NCBI), for functional annotation.

Evaluation of genome quality

Genome assembly quality was evaluated using multiple methods. (1) The Illumina short-read sequencing data were aligned to the assembled genome, and mapping ratio statistics were then determined using BWA (v0.7.17) (Li & Durbin, 2009). (2) The assembled transcripts were also mapped to the genome assembly using Blat (v34) (Kent, 2002), with mapping ratio statistics then determined. (3) The *P .pardalis* genome was aligned to the Vertebrata_odb9 database in Benchmarking Universal Single-Copy Orthologs (BUSCO) (v2.0) (Simão et al., 2015) to evaluate the percentage of conserved core genes assembled. (4) Genomic synteny between the genomes of *P. pardalis* and closely related species was analyzed using LAST (v1066) (Kielbasa et al., 2011).

Identification of orthologous genes

Orthologous gene relationships were analyzed using the OrthoMCL pipeline (v2.0.9) (Li et al., 2003). The protein-coding sequences of 11 species (including *B. yarrelli*, *A. melas*, *I. punctatus*, *P. fulvidraco*, *G. maculatum*, *P. hypophthalmus*, *T. fulvidraco*, *H. wyckioides*, *S. meridionalis*, *C. magur*, and *D. rerio*) were first downloaded from NCBI and the GigaScience Database (gigaDB). For genes with multiple isoforms, we selected the longest transcripts to represent the gene and removed genes with early termination sites. Subsequently, we used all retained gene sequences to perform a reciprocal BLAST search of the 11 downloaded species and *P. pardalis*. We then used *orthmcl* to compute pairwise relationships. We considered the reciprocal best similarity pairs as putative orthologs and reciprocal better similarity pairs as paralogs.

Finally, the one to one orthologous genes among species were identified as single-copy orthologous genes.

Phylogenetic relationships and divergence time

All single-copy genes of the above 12 species were extracted from OrthoMCL analysis. The sequences of each single-copy gene among the 12 species were aligned using MUSCLE (v3.8.31) (Edgar, 2004) with default parameters and concatenated into one super-sequence of one species with the same order. Phylogenetic analysis was then conducted using the maximum-likelihood algorithm in RAxML (v8.2.10) (Stamatakis, 2014). Divergence times among the species were estimated using the Bayesian relaxed molecular clock approach with MCMCtree in the PAML package (v4.8) (Yang, 2007), and fossil times obtained from the TIMETREE database (http://www.timetree.org/) were used for calibration of the results.

Relative evolutionary rates

We analyzed the relative evolutionary rate of each taxon using the *tpcv* module in LINTRE. The relative evolutionary rates between *P. pardalis* and other species were calculated using zebrafish as the outgroup.

Expansion/contraction of gene families

Phylogenetic relationships, divergence times, and gene family relationships were used to analyze changes in gene families. The expansion and contraction of gene family in ancestor nodes and each of the 12 species mentioned above were analyzed using CAFE (v3.1) (De Bie et al., 2006).

Results

Chromosome-level genome assembly and annotation

We generated a total of 146.15 Gb of Illumina paired-end short reads to investigate the size and characteristics of the *P. pardalis* genome (Table S1). Based on 17-mer analysis, the genome was 1.48 Gb in size, with a heterozygous (first) peak indicating

239 a high level of genome heterozygosity (Fig. 1A). To obtain a high-quality genome
240 assembly, 218.07 Gb of data produced by the third-generation Nanopore sequencing
241 platform were used to assemble contigs (Table S2), with potential single-base
242 sequencing errors then corrected using Illumina short reads (Table S1). To further
243 improve the continuity of the genome assembly, chromosomes were constructed by
244 3D *de novo* assembly using 149.24 Gb of clean Hi-C reads (Table S3). Finally, a total
245 of 26 chromosomes were successfully anchored with the scaffold N50 of 49.47 Mb
246 and genome size of 1.51 Gb (Fig. 1B and 1C, Tables S4 and S5). The assembled
247 genome size was close to the estimated genome size based on *k*-mer analysis (1.48
248 Gb), indicating that an appropriate assembly size was obtained in this study. To better
249 estimate assembled genome quality, we determined the read mapping ratio (98.52%)
250 (Table S6), transcript mapping ratio (99.61%) (Tables S7-S9), and percentage of
251 conserved core genes (BUSCO score, 98.8%) (Table S10, Eukaryota database), with
252 the conserved *Hox* gene clusters (Fig. 1D) indicating that the chromosome-level
253 reference genome of *P. pardalis* exhibited high integrity and accuracy.
254 Based on genome annotation, 64.47% (0.97 Gb) of the assembled *P. pardalis*
255 genome sequences were repetitive sequences (Table S11), belonging to four major
256 repeat classes, including DNA elements, long interspersed nuclear elements (LINEs),
257 short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs) (Fig.
258 1C, Table S12). Results showed that DNA elements accounted for the highest
259 proportion (33.15%; total length: 499.77 Mb) of repetitive sequences in the genome,
260 while SINEs (1.94%; total length: 29.27 Mb) accounted for the lowest proportion
261 (Table S12). Using repeat-masked genome sequences, we successfully predicted
262 23,859 protein-coding genes in the *P. pardalis* genome (Table S13). To assess the
263 quality of the predicated protein-coding genes, we compared the length distributions
264 of genes (Fig. 2A), coding sequences (CDS) (Fig. 2B), exons (Fig. 2C), and introns
265 (Fig. 2D) between *P. pardalis* and other species. Results showed that the predicted
266 quality of the protein-coding genes in *P. pardalis* was comparable to previously
267 published species (Fig. 2). Furthermore, functional annotation analysis showed that
268 most of the predicted genes (22,169; 92.92%) had homologous genes in public

databases, including InterPro, GO, KEGG, COG, SwissProt, TREMBL, and NR (Table S13). Taken together, these results indicate that a high-quality protein-coding gene set was obtained.

Phylogenetic relationships and evolutionary rate analysis of *P. pardalis*

To systematically explore the evolution of *P. pardalis*, we analyzed potential gene families in *B. yarrelli*, *A. melas*, *I. punctatus*, *P. fulvidraco*, *G. maculatum*, *P. hypophthalmus*, *T. fulvidraco*, *H. wyckioides*, *S. meridionalis*, *C. magur*, *D. rerio* and *P. pardalis* and identified 9 239 gene families and 699 single-copy genes among these species (Fig. 3A, Table S14). Based on the 699 single-copy genes, we constructed a phylogenetic tree using zebrafish as the outgroup. Results showed that all catfish examined were clustered on one large branch, indicating a common ancestor and origin (Fig. 3B). With calibration of known fossil records downloaded from the TIMETREE database, we determined divergence time among the species. Result showed that *P. pardalis* diverged from the common ancestor of the other examined catfish 132.5 Mya (Fig. 3B). We also analyzed the relative evolutionary rates of the above species, with *P. pardalis* evolving relatively rapidly compared to the other catfish species except *B. yarrelli*, which may be related to its ability to adapt to harsh environments (Fig. 3C, Tables S15 and S16).

Chromosomal evolution of *P. pardalis*

Chromosomal fusion/fission events usually occur in a key period of species evolution. Therefore, we analyzed the genome syntenic relationships between *P. pardalis* and three catfish species with chromosome-level genome assemblies, i.e., *A. melas*, *I. punctatus*, and *P. hypophthalmus*. According to analysis, the *P. pardalis* genome exhibited high genomic collinearity with the three other catfish species, consistent with their close phylogenetic relationship (Figs. 4A-4C). Interestingly, we observed multiple chromosomal fusion/fission phenomena. Typically, chromosome 6 of *P. pardalis* showed the best synteny to chromosomes CM022778 and CM022752 in *A. melas* (Fig. 4D), chromosomes NC030444 and NC030436 in *I. punctatus* (Fig. 4E),

and chromosomes NC047622 and NC047614 in *P. hypophthalmus* (Fig. 4F). Furthermore, we remapped the Nanopore long reads to the assembled genome, successfully spanning the break point locations across the three species. We also analyzed the sequencing depths of the break point regions, with most sites showing a sequencing depth of more than 50 (the right panels in Figs. 4D-4F). These results further indicated high confidence in the assembly of chromosome 6 in the *P. pardalis* genome. Previous studies have found that chromosomal fusion/fission events play important roles in the adaptive evolution of a species in a specific environment (Biello et al., 2021). Our results suggest that Siluriformes exhibits great diversification in chromosomal organization and karyotype variability, especially in *P. pardalis*, which may contribute to species evolution through chromosomal fusion/fission events. However, whether these changes also play a key role in the environmental adaptation of *P. pardalis* requires further study in the future.

Expansion of immune system-related genes contributes to robust invasion capacity of *P. pardalis*

Changes in gene families can cause serious defects or enhance environmental adaptations. Therefore, identifying significantly expanded and contracted gene families should provide valuable insights into the adaptive evolution of *P. pardalis*. In this study, we analyzed and identified gene families that were expanded or contracted in the *P. pardalis* genome compared with the genomes of *B. yarrelli*, *A.melas*, *I. punctatus*, *P. fulvidraco*, *G. maculatum*, *P. hypophthalmus*, *T. fulvidraco*, *H. wyckioides*, *S. meridionalis*, *C. magur*, and *D. rerio* based on cluster analysis. We identified 1,182 expanded and 2,286 contracted gene families in the *P. pardalis* genome. Furthermore, functional enrichment analysis of the expanded gene families identified 67 significantly enriched KEGG pathways ($P < 0.05$) (Table S17) and 78 significantly enriched Gene Ontology (GO) terms ($P < 0.05$) (Table S18). Further examination revealed enrichment in several immune-related pathways, including antigen processing and presentation ($P < 1 \times 10^{-9}$), intestinal immune network for IgA production ($P < 1 \times 10^{-9}$), toll-like receptor signaling pathway ($P = 4.7 \times 10^{-3}$), and T

cell receptor signaling pathway (P= 4.86 × 10^{-2}) (Table S17). The expanded genes in the immune-related pathways may help explain the adaptation of *P. pardalis* to complex feeding habits and highly polluted environments. Both KEGG and GO categories also identified several significantly enriched metabolism-related pathways, including nitrogen metabolism ($P < 1 \times 10^{-9}$), cellular lipid metabolic process (P = 3.2 × 10^{-2}), and protein metabolic process ($P = 1.96 \times 10^{-4}$) (Tables S17 and S18), suggesting that a powerful metabolism may help this species cope with harsh food conditions. We manually determined gene copy number in *P. pardalis* and identified several markedly expanded genes, including *CASP6* (Figs. 5A and 5C) and *TLR7* (Figs. 5B and 5C). These two genes play critical roles in the immune response, participating in host defense against viruses and parasites. For example, *CASP6* acts as a key regulator in host defense against influenza A virus and gram-negative bacterial infections, and in mediating innate immunity and inflammasomes (Zheng et al., 2020). *TLR7* exhibits antiviral activities through sensing endosomal single-stranded RNA to recognize influenza genomic RNA and activate immune cells (Arpaia et al., 2011; Diebold et al., 2004; Takeuchi et al., 2010), which may also play a central defense role in *P. pardalis*. Thus, the expansion of these two genes in *P. pardalis* may contribute to their adaptation to the complex benthic microbial environment.

Conclusions

The strong invasive capacity of *P. pardalis* has had considerable ecological and economic impact in invaded regions. However, little is known regarding the genetic basis of this strong invasive ability due to the lack of a high-quality reference genome. Here, we assembled the first chromosome-level reference genome of *P. pardalis* (1.51 Gb) using data from multiple sequencing platforms. The genome of *P. pardalis* contained 26 chromosomes and showed high sequence continuity (scaffold N50: 49.47 Mb). Combining *de novo*-, homology-, and transcription-based strategies, we obtained 0.94 Gb of repetitive sequences and successfully annotated 23 859 protein-coding genes in the *P. pardalis* genome. Genome quality assessment based on

conserved core genes, read and transcript mapping ratios, and genome synteny indicated high genomic accuracy and continuity. The successful assembly and annotation of the *P. pardalis* genome not only provides a valuable genomic resource for this species but should also facilitate systematic comparative genomic studies of Siluriformes. Divergence time analysis indicated that *P. pardalis* diverged from the common ancestor of all examined catfish 132.5 Mya. Evolutionary rate analysis showed that *P. pardalis* evolved relatively rapidly compared to the other species, except for *B. yarrelli*. Analysis of expanded and contracted gene families showed that many immune- and metabolism- related gene families were expanded in the *P. pardalis* genome, which may contribute to their adaptation to harsh conditions and successful invasion. The genome assembly and comparative genomic analyses performed in this study not only provide a valuable genomic resource for the study of Siluriformes, but also shed light on the genetic basis of the invasive ability of *P. pardalis*.

References

Anguebes, F., Bassam, A., Abatal, M., Tzuc, O. M., & Pedro, L. S. (2019). P hysical and chemical properties of biodiesel obtained from amazon sailfin catfish (*pterygoplichthys pardalis*) biomass oil. *Journal of Chemistry*, *2019* (3).

Arpaia, N., Godec, J., Lau, L., Sivick, K. E., McLaughlin, L. M., Jones, M. B., ... & Barton, G. M. (2011). TLR signaling is required for Salmonella typhimurium virulence. *Cell*, *144*(5), 675-688.

Baldissera M.D., Souza, C., Val, A.L., & Baldisserotto, B. (2019). Involvement of purinergic signaling in the amazon fish *pterygoplichthys pardalis* subje cted to handling stress: relationship with immune response. *Aquaculture*, *5 14*, 734481.

Barra, D., Bossa, F., & Brunori, M. (1981). Structure of binding sites for heterotropic effectors in fish haemoglobins. *Nature*, *293*(5833), 587-588.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, *27*(2), 573-580.

Biello, R., Singh, A., Godfrey, C. J., Fernández, F. F., Mugford, S. T., Powell, G., ... & Mathers, T. C. (2021). A chromosome-level genome assembly o f the woolly apple aphid, Eriosoma lanigerum Hausmann (Hemiptera: Aphi didae). *Molecular ecology resources*, *21*(1), 316-326.

Birney E., Clamp M., Durbin R. (2004) GeneWise and Genomewise. *Genome Res*, *14*(5):988-95.

Bossa, F., Savi, M. R., Barra, D., & Brunori, M. (1982). Structural comparison of the haemoglobin components of the armoured catfish Pterygoplichthys pardalis. Evolutionary considerations. *Biochemical Journal*, *205*(1), 39-42.

Bowen, J. S., Grande, T. C., & Wilson, M. V. (2016). *Fishes of the World*. John Wiley & Sons.

Bowen, S.H. Detritivory in neotropical fish communities. *Environ Biol Fish*, 9, 137–144 (1983).

Burgess, W. (1989). An atlas of freshwaters and marine catfishes, a preliminar y survey of the siluriformes.

Chaichana, R., & Jongphadungkiet, S. (2012). Assessment of the invasive catfis h *pterygoplichthys pardalis* (castelnau, 1855) in thailand: ecological impact s and biological control alternatives. *Tropical Zoology*, *25*(4), 173-182.

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., ... & Huang, X. (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome biology*, *16*(1), 1-10.

Da Cruz, A. L., Da Silva, H. R., Lundstedt, L. M., Schwantes, A. R., Moraes, G., Klein, W., & Fernandes, M. N. (2013). Air-breathing behavior and physiological responses to hypoxia and air exposure in the air-breathing loricariid fish, *Pterygoplichthys anisitsi*. *Fish Physiology and Biochemistry*, *39*(2), 243-256.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a c omputational tool for the study of gene family evolution. *Bioinformatics*, *2 2*(10), 1269-1271.

Delariva, R. L., & Agostinho, A. A. (2010). Relationship between morphology and diets of six neotropical loricariids. *Journal of Fish Biology*, *58*(3), 832 -847.

Diebold, S. S., Kaisho, T., Hemmi, H., Akira, S., & Sousa, C. (2004). Innate antiviral responses by means of tlr7-mediated recognition of single-stranded rna. *Science*, *303*(5663), 1529-1531.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., ... & Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92-95.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*, *3*(1), 95-98.

Ebenstein, D., Calderon, C., Troncoso, O. P., & Torres, F. G. (2015). Characte rization of dermal plates from armored catfish *pterygoplichthys pardalis* re veals sandwich-like nanocomposite structure. *Journal of the Mechanical Be havior of Biomedical Materials*, *45*, 175-182.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792-1797.

Haas B. J., Delcher A. L., Mount S. M., et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, 1;*31*(19):5654-66.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., ... & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology*,*9*(1), 1-22.

Hill, J., & Sowards, J. (2015). Successful eradication of the non-native loricariid catfish *pterygoplichthys* disjunctivus from the rainbow river, florida. *Management of Biological Invasions*, *6*(3), 311-317.

Hoover, J. J., Killgore, K. J., & Cofrancesco, A. F. (2004). Suckermouth catfishes: threats to aquatic ecosystems of the united states?. Aquatic Nuisance Species Res Prog Bull, 04–1.

Hui, W., Copp, G. H., Vilizzi, L., Fei, L., & Hu, Y. (2017). The distribution, establishment and life-history traits of non-native sailfin catfishes *pterygoplichthys spp.* in the guangdong province of china. *Aquatic Invasions*, *12*(2), 241-249.

Hussan, A., Choudhury, T. G., Das, A., & Gita, S. (2016). Suckermouth sailfin catfishes: A future threat to aquatic ecosystems of India. *Aquaculture times*, *2*(6), 20-22.

Kent W.J. (2002).BLAT--the BLAST-like alignment tool. *Genome Res*,*12*(4):656-64.

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3).

Krishnakumar, K., Raghavan, R., Prasad, G., Bijukumar, A., Sekharan, M., Pereira, B., & Ali, A. (2009). When pets become pests–exotic aquarium fishes and biological invasions in Kerala, India. *Current science*, *97*(4), 474-476.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, *13*(9), 2178-2189.

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., ... & Waldbieser, G. C. (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nature communications*, *7*(1), 1-13.

Mohammad, H., Robert, V., Ramon, R. C., & Galib, S. M. (2018). Amazon sailfin catfish *pterygoplichthys pardalis* (loricariidae) in bangladesh: a critical review of its invasive threat to native and endemic aquatic species. *Fishes*, *3*(1), 14-.

Nico, L. G., Martin,N. T. (2001). The south american suckermouth armored catfish, *pterygoplichthys anisitsi* (pisces: loricaridae), in texas, with comments on foreign fish introductions in the american southwest. *Southwestern Naturalist*, *46*(1), 98-104.

Nishimura, D. (2000). RepeatMasker. *Biotech Software & Internet Report*, *1*(1-2), 36-39.

Nurubhasha, R., Sampath Kumar, N. S., Thirumalasetti, S. K., Simhachalam, G., & Dirisala, V. R. (2019). Extraction and characterization of collagen from the skin of *Pterygoplichthys pardalis* and its potential application in food industries. *Food science and biotechnology*, *28*(6), 1811-1817.

Orfinger, A. B., & Goodding, D. D. (2018). The global invasion of the sucker mouth armored catfish genus *pterygoplichthys* (siluriformes: loricariidae): an notated list of species, distributional summary, and assessment of impacts. *Zoological Studies*, 57.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*, *34*(suppl_2), W435-W439.

Takeuchi O, Akira S. (2010). Pattern recognition receptors and inflammation. *C ell*, 140:805–820.

Wakida-Kusunoki, A. T., Ruiz-Carus, R., & Amador-Del-Angel, E. (2007). Am azon sailfin catfish, *pterygoplichthys pardalis* (castelnau, 1855) (loricariida e), another exotic species established in southeastern mexico. *The Southwes tern Naturalist*, (1), 52.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*,*9*(11), e112963.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, *24*(8), 1586-1591.

Zheng, M., Karki, R., Vogel, P., & Kanneganti, T. D. (2020). Caspase-6 is a key regulator of innate immunity, inflammasome activation, and host defense. *Cell*, *181*(3), 674-687.

Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; RNA-seq: RNA sequencing; BWA: Burrows-Wheeler Aligner; BLAST: Basic Local Alignment Search Tool; KEGG: Kyoto Encyclopedia of Genes and Genomes; PacBio: Pacific Biosciences.

Conflicts of interest

The authors declare no conflicts of interest.

Funding

This research was funded by the National Natural Science Foundation of China (32100406), Chinese Academy of Sciences, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology (GREKF21-07), Scientific Research Fund of Shaanxi Provincial Education Department (21JK0888), and Natural Science Basic Research Plan in Shaanxi Province of China (2021JQ-775).

Author contributions

X. G., L. X., and Y. Z. supervised the project. X. G. designed the study. W. X., H. L., and H. J. performed the data analysis. W. X. wrote the manuscript. Y. L. and Y.W. revised the manuscript.

Data accessibility

The genome assembly and annotation data were deposited in the Dryad database (doi:10.5061/dryad.bk3j9kdgh).

Figure legends:

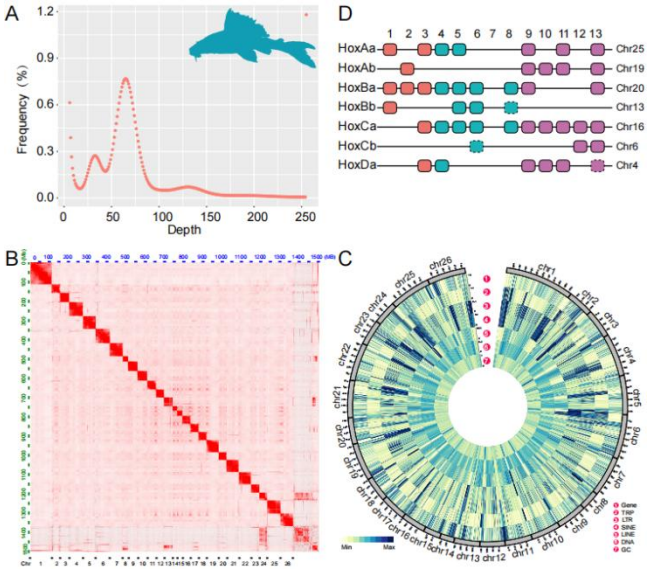


Figure 1. Genomic information of *P. pardalis*. (A) Survey of genomic characteristics. X-axis represents 17-mer depth, y-axis represents 17-mer frequency. (B) Heatmap of chromosomal interactions. Blocks represent contact between

corresponding locations. (C) Distributions of genomic elements in *P. pardalis* genome. Outer to inner ring are distributions of protein-coding genes, tandem repeats (TRP), long terminal repeats (LTR), short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), DNA elements, and GC content, respectively. (D) Hox gene clusters in *P. pardalis* genome. Solid line represents functionally annotated gene in the database, dotted line represents that only the gene fragment could be found.

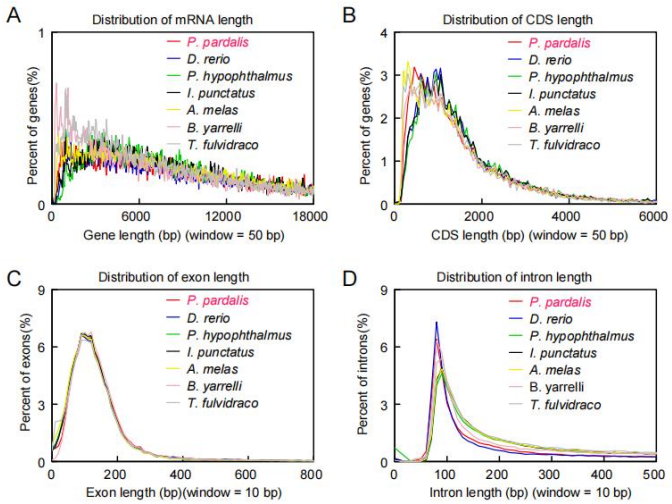
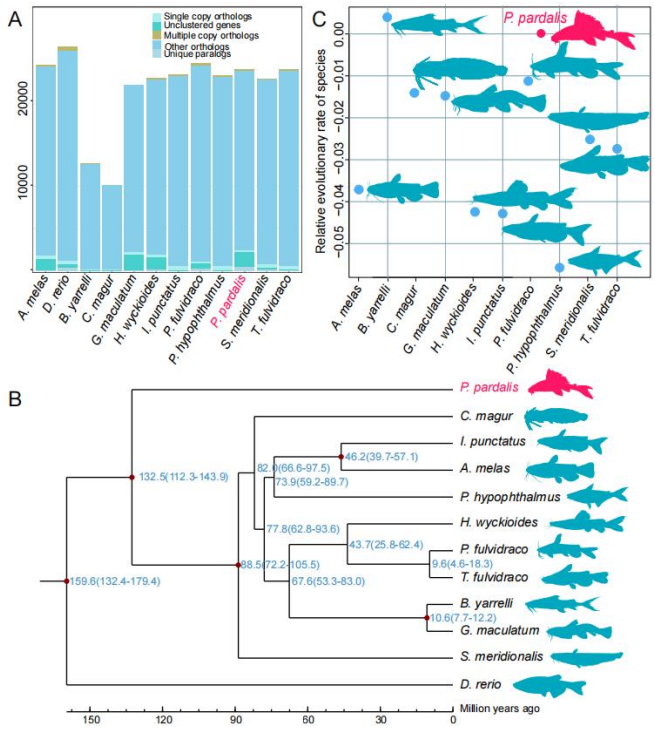


Figure 2. Quality comparison of protein-coding genes between *P. pardalis* and other species. Quality of gene annotation based on (A) gene length, (B) CDS length, (C) exon length, and (D) intron length, respectively.



552

553 **Figure 3. Comparative genomic analyses of *P. pardalis* and closely related species.**

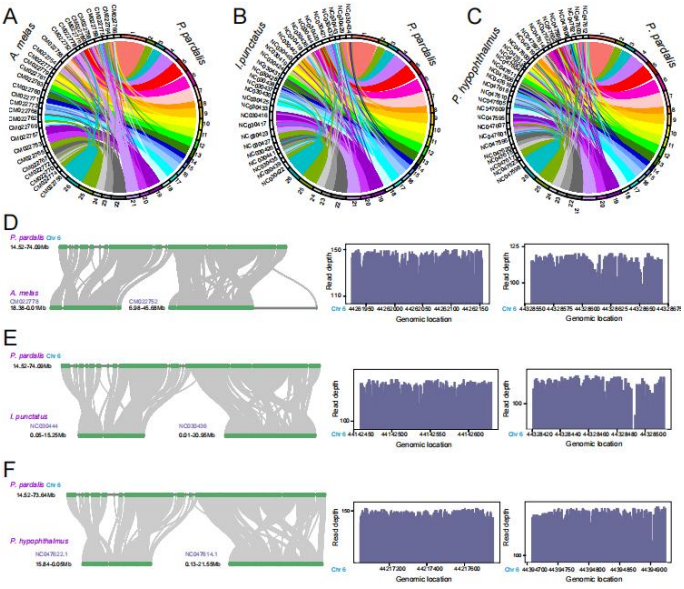
554 (A) Numbers of orthologous and paralogous genes. (B) Phylogenetic relationship and

555 divergence time between species, with *D. rerio* used as the outgroup. Red dots

556 represent use of fossil evidence to adjust divergence times. Blue numbers represent

557 divergence time. (C) Relative evolutionary rates of species. Analysis was performed

558 with *P. pardalis* as the reference species and *D. rerio* as the outgroup species.



559

560 **Figure 4. Whole-genome synteny analysis between *P. pardalis* and three other**

561 **catfish species with chromosome-level genomes.** (A) Whole-genome synteny

between *P. pardalis* and *A. melas*. (B) Whole-genome synteny between *P. pardalis* and *I. punctatus*. (C) Whole-genome synteny between *P. pardalis* and *P. hypophthalmus*. (D) Collinear blocks between chromosome 6 of *P. pardalis* and CM022778/CM022752 of *A. melas.* Figure on right shows depth of Nanopore long reads in *P. pardalis* genome. (E) Collinear blocks between chromosome 6 of *P. pardalis* and NC030444/NC030436 of *I. punctatus.* Figure on right shows depth of Nanopore long reads in *P. pardalis* genome. (F) Collinear blocks between chromosome 6 of *P. pardalis* and NC047622.1/NC047614.1 of *P. hypophthalmus.* Figure on right shows depth of Nanopore long reads in *P. pardalis* genome.

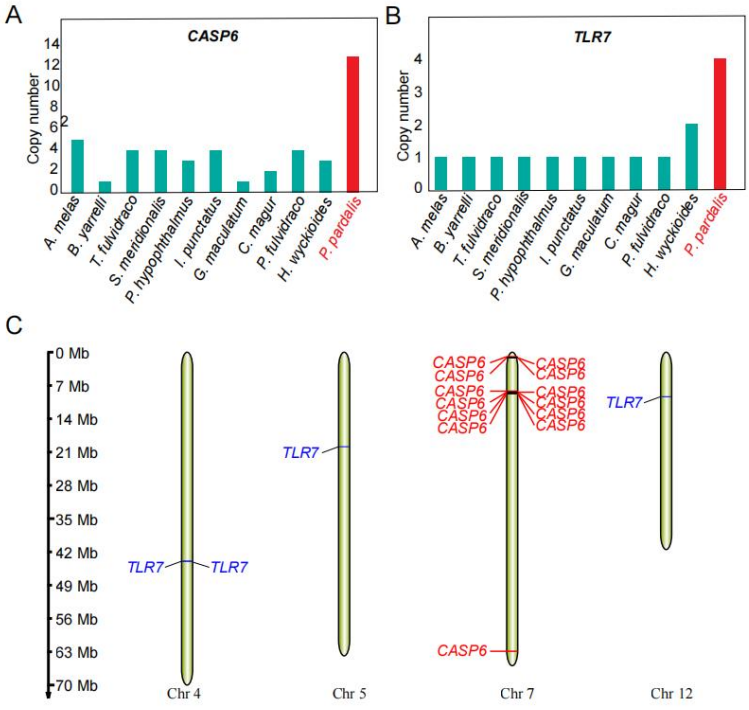


Figure 5. Expansion of gene copy number in *P. pardalis* genome. (A) *CASP6* gene copy number in annotated gene sets among species. (B) *TLR7* gene copy number in annotated gene sets among species. (C) Distributions of both genes in *P. pardalis* genome.