First draft Genome of Loach (Orenectus shuilongensis; Cypriniformes: Nemacheilidae) provide insights into the Evolution of Cavefish

Zhi-Jin Liu (6888@cnu.edu.cn)
Capital Normal University
Xiong-Fei Zhang
Capital Normal University
Hua-Mei Wen
Guizhou Normal University
Ling Han
Guizhou Normal University
Jiang Zhou
Guizhou Normal University

Research Article

Keywords: Oreonectes shuilongensis, cavefish, genome assembly, adaptation, Nemacheilidae, Cobitidea

Posted Date: December 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-192229/v2

Abstract

Background

Loaches from the superfamily Cobitoidea (Cypriniformes, Nemacheilidae) are small elongated bottom-dwelling freshwater fishes with several barbels near the mouth, and some species of loach inhabit the underground drainage. The genus *Oreonectes* with 18 currently recognized loach species represent the three key stages of the evolutionary process (a surface-dwelling lifestyle, facultative cave persistence, and permanent cave dwelling). Some *Oreonectes* species show typical cave dwelling-related traits, such as partial or complete leucism and regression of the eyes, rendering them as suitable study objects of micro-evolution. Genome information of *Oreonectes* species is therefore an indispensable research resource of the evolution of cavefishes.

Result

We assembled the genome sequence of *O. shuilongensis*, a surface-dwelling species, using an integrated approach that combined PacBio single-molecule real-time sequencing and Illumina X-ten paired-end sequencing. The genome assembly contains 803 contigs with N50 values of 5.58 Mb. 25,247 protein-coding genes were predicted, of which 95.65% have been functionally annotated. Meanwhile,we found that dozens of genes related to eye development and melanogenesis were pseudogenised during the evolutionary process in cave environment, providing novel insights into complex phenotypic adaptations of animals in specific environment.

Conclusion

Here we report the first draft genome assembly of *Oreonectes* fishes, which is also the first genome reference for Cobitidea fishes. This genome assembly will contribute to the study of the evolution and adaptation of cavefishes within *Oreonectes* and beyond (Cobitidea) and provid valuable genomic resources for studies on the evolutionary history of the rapid speciation processes of family Nemacheilidae.

Background

Cavefishes are successful vertebrate colonizers in subterranean habitats and usually process some regressive features, such as the rudimentary eyes and loss of pigmentation. Meanwhile, some compensative traits, such as elongated appendages and reinforced non-visual sensory systems, have evolved in cavefishes. Since uncovering the genetic basis of phenotypic adaptations of animals to a specific environment is a key goal in evolutionary study, cavefishes have attracted interests from biologist and certain cavefishes (*Astyanax mexicanus* and *Sinocyclocheilus* spp.) have been well studied. However, there is few data available to unravel the genomic mechanism under the evolution and adaptation to subterranean life among other groups of fishes.

Loaches (Cypriniformes: Cobitoidea, Nemacheilidae) are small elongated bottom-dwelling freshwater fishes with several barbels near the mouth, distributed in Eurasia and Africa. Within this group, *Oreonectes* fishes are distributed only in southwestern China and northern Vietnam, most of which dwell in underground rivers in the karst environment [1]. *Oreonectes* fishes contain 18 species representing the three key stages of the evolutionary process including a surface-dwelling lifestyle, facultative cave persistence, and permanent cave dwelling. Almost all *Oreonectes* species show some cave-related traits, such as part or complete eye degeneration and leucism, which makes this genus a good study system of micro-evolution [2] . *O. shuilongensis* is a surface-dwelling species which was newly discovered in the Shuilong Township in Guizhou Province of China [3]. A genome

assembly of *O. shuilongensis* would facilitate research into key aspects of the evolutionary history of cave versus surface dwelling in *Oreonectes*, including the role of environmental changes in the seemingly rapid diversification and speciation in underground caves [2].

Results

Whole Genome and RNA Sequencing

After removal of <500 bp PacBio subreads, 5 million subreads (total 50.9 Gb) remained, with an average length of 10.2 kb (Table 1 and Suppl. Tables S2 and S3). Additionally, a total of 11.7 Gb transcriptome data were obtained from RNA-sequencing (Table 1).

Estimation of the Genome Size and Sequencing Coverage

The genome size of *O. shuilongensis* was estimated at approximately 515.66 Mb based on k-mer analysis (Suppl. Figure S1), and our *O. shuilongensis* genome assembly spans 521.68 Mb (803 contigs, contig N50 of 5.58 Mb; Table 2 and Suppl. Table S4). The completeness of the *O. shuilongensis* genome assembly was evaluated using CEGMA v2.5 [4] and BUSCO v2 [5]. CEGMA analysis suggested that 99.78 % of conserved Core Eukaryotic Genes (CGEs) proteins are present in our assembled genome, and BUSCO analysis showed that 97.58% of vertebrate Benchmarking Universal Single-Copy Orthologs have been assembled, implying a high completeness of our *O. shuilongensis* genome assembly (Suppl. Tables S5 and S6). We found around 99.52% of the reads properly mapped to the genome assembly (Suppl. Table S7). All these results indicate that the assembly of the *O. shuilongensis* genome is characterized by a high level of accuracy.

Annotation of Repeat Sequences, Protein Coding Genes and Noncoding RNA

We found that repetitive elements comprised 23.42% of the *O. shuilongensis* genome (Table 3). 25,247 protein-coding genes were identified. The average transcript length, CDS length, and intron length were 9,744 bp, 2,010 bp, and 7,734 bp, respectively (Suppl. Table S8). Among these annotated genes, 64.51% of encoded proteins showed homology to proteins in the KOG database, 95.53% were identified in the NCBI non-redundant database, 46.72% were identified in the KEGG database, 94.90% were identified in the TrEMBL database, and 95.65% could be mapped onto the functional databases (Suppl. Table S9). Finally, 947 miRNAs, 561 rRNAs and 417 tRNAs were discovered from the *O. shuilongensis* (Suppl. Table S10).

Phylogenetic relationship and genomic comparison

As a result, 16,708 gene families were constructed for the *O. shuilongensis*. Among the families, there were 144 families unique to *O. shuilongensis* (Figure 2 and Suppl. Table S11). The constructed phylogenetic tree indicated that *O. shuilongensis* were clustered closely to Cyprinidae species, which is inconsistent with their putative evolutionary relationships (Figure 3). The divergence between *O. shuilongensis* and cyprinid fishes (Cyprinidae) occurred ca. 91.31 million years ago (95% HPD, 82.58-108.26). When comparing this with the other seven fish (*S. salar*, *I. punctatus*, *A. mexicanus*, *C. carpio*, *S. rhinocerous*, *D. rerio* and *L. crocea*), the expansion and contraction of gene ortholog clusters showed 77 gene families were expanded and 282 gene families contracted significantly in the *O. shuilongensis* (Figure 3).

Genomic mechanism underlying the degeneration of eyes and body color

Genome re-sequencing were performed for facultative cave-dwelling *O. jiarongensis* (three individuals) and cave-dwelling *O. daqikongensis* (two individuals) and *O. dongliangensis* (one individual) at a high average depth of 28.06±5.08×, with an overall average genome coverage of 93.77% of the *O. shuilongensis* genome assembly (Table 4). A total of 12,534,348 SNPs was identified in these three species, and the number of SNPs per individual ranged from 6.0 to 6.2 M (Table 4). The reconstructed phylogeny indicates that the ancestor of cave-dwelling *O. daqikongensis*, *O. dongliangensis* and facultative cave-dwelling *O. jiarongensis* first diverged from the surface-dwelling *O. shuilongensis* about 9.31 million years ago (95% HPD, 12.54 – 7.12) (Figure 4).

Annotation of all sequence variants in SnpEff [6] (Figure 5) suggested that 1,541 SNPs and 438 indels were located in 401 genes, likely resulting in pseudogenization in semi cave-dwelling and cave-dwelling species. Twenty-nine pseudogenes annotated using DAVID [7] showed significant enrichment for the GO terms of “eye development (0001654)” and “retina development in camera-type eye (0060041)” (Table 5, Figure 6 and Suppl. Table S12). For example, the expression of *six7*, *six6a* and *six6b* is required for optic primordium development and the specification and proliferation of the eye field in vertebrate embryos [8]. The function-lost mutation of *aldh1a3* and *tfap2a* caused eye and retinal defects in zebrafish [9]. It is presumed that these pseudogenes might lead to eye degeneration of semi/complete cave-dwelling *Oreonectes* species. Furthermore, *Mc1r* (melanocortin-1 receptor), a key gene regulating the biosynthesis of melanin in most vertebrates, is a pseudogenization caused by an deletion in *O. daqikongensis* (Figure 7), likely blocking biosynthesis of melanin and leading to the albino phenotype (Fig. 1). Remaining pseudogenes are enriched for the GO terms of “potassium channel activity”, “regulation of axon extension involved in axon guidance”, “G-protein coupled receptor activity” and KEGG pathway of “neuroactive ligand-receptor interaction” (Table 5).

Discussion

Organisms that have colonized underground caves encounter vastly different selective pressures than their relatives in above-ground habitats. In the present study, we report the first whole genome sequencing, assembly, and annotation of the *O. shuilongensis*, encompassing a total of predicted 25,247 protein-coding genes and 7,041 noncoding RNAs. We anticipate that this genome assembly will serve as a basis for in-depth biological studies of evolution and adaptation of cavefishes. With the availability of these genomic data, genomic/transcriptomics differences between surface-dwelling and cave-dwelling loaches can be studied at the genomic scale. More broadly, our assembly will facilitate evolutionary and genomics research of Cobitoidea fishes and beyond.

Conclusion

We reported the first draft genome assembly of Oreonectes fish, which is also the first genome reference for Cobitidea fish. The genome sequence of *O. shuilongensis*, a surface-dwelling species, was assembled with an integrated method combining PacBio single-molecule real-time sequencing and Illumina paired-end sequencing. The genome re-sequencing of another three cave-dwelling Oreonectes fishes were also performed. Dozens of genes related to eye development and melanogenesis are found to be pseudogenised during the evolutionary process in cave environment, providing novel insights into complex phenotypic adaptations of animals in specific environment. This genome assembly could facilitate the study of the diversification and adaptation of cavefishes.

Materials And Methods

Whole Genome Sequencing

O. shuilongensis (Figure1) was captured from Shuilong Township, Sandu County, Guizhou Province, China. The sample was quickly frozen in liquid nitrogen for one hour before storing at −80°C. Genomic DNA was extracted from a muscle sample using DNeasy Blood &Tissue Kit (Qiagen). Three small-insert libraries (270 bp) were constructed by using Illumina's paired-end kits according to the manufacturer's instructions. The libraries were sequenced on Illumina Hiseq X Ten platform. For the raw reads, sequencing adaptors were removed. Contaminated reads (such as chloroplast, mitochondrial, bacterial and viral sequences, etc.) were screened by alignment to the NCBI-NR database using BWA [10] with default parameters. Duplicated read pairs were removed by FastUniq v1.12 [11], and low-quality reads were filtered under the following conditions: (1) reads with ≥10% unidentified nucleotides (N), (2) reads with >10 nucleotides aligned to the adapter, allowing ≤10% mismatches, (3) reads with >50% bases having Phred quality <5.

RNA Sequencing

Tissues of skin, muscle, intestine, liver and kidney of the same loach individual were collected and RNAs were extracted with TRIZOL Reagent (Invitrogen, USA). RNAs were then balanced mixed for the sequencing. The absorbance of 1.90 at 260 nm/280 nm and the RIN of 9.1 were obtained for the purified RNA sample by Nanodrop ND-1000 spectrophotometer (LabTech, USA) and 2100 Bioanalyzer (Agilent Technologies, USA), respectively. One microgram of RNA was reverse transcribed using Clontech SMARTer cDNA synthesis kit, and was further fragmented using divalent cations for the sequencing. The paired-end library was prepared following the manual of the Paired-End Sample Preparation Kit (Illumina Inc., San Diego, CA, USA). Then the library with an insert length of 270 bp was sequenced by Illumina HiSeq X Ten in 150 bp paired-end mode (Illumina Inc., San Diego, CA, USA).

Estimation of the Genome size and Sequencing Coverage

The 19-mer frequency distribution analysis was performed on the remaining clean reads to estimate the genome size of the *O. shuilongensis* using the formula: Genome size = kmer_Number/Peak_Depth. Corrected Illumina reads were selected to perform genome size estimation.

De novo Genome Assembly and Quality Assessment of O. shuilongensis Genome

The single-molecule sequencing (SMS) data are assembled through Canu v1.5 [12] a comprehensive and scalable pipeline for SMS data assembly, and the draft assembly polished through Pilon [13]. In the correction step, we used Canu to first select longer seed reads with the settings 'genomeSize = 520M' and 'corOutCoverage = 90', then to detect raw reads overlapping through a highly sensitive over lapper MHAP (mhap-2.1.2, option 'corMhapSensitivity = normal'), and finally to perform error correction through the falcon sense method (option 'corrected Error Rate = 0.045'). Next, using default parameter settings, error-corrected reads were trimmed of unsupported bases and hairpin adapters to obtain their longest supported range. In the last step, Canu generated the draft assembly using trimmed reads. The draft assembly was polished to obtain the final assembly, adopting Pilon [13] using Illumina data with the parameters '–mindepth 10 –changes –threads 4 –fix bases'. To further evaluate the accuracy of the *O. shuilongensis* genome, we aligned the high-quality short reads from short insert size pair-ended libraries against the genome assembly using BWA [10].

Annotation of Repeat Sequences

The repeat composition of the assemblies was estimated by building a repeat library employing the *de novo* prediction programs LTR FINDER v1.05[14], MITE-Hunter [15] , RepeatScout v1.0.5 [16] and PILER-DF v2.4 [17]. The

database was classified using PASTEClassifier [18] and was combined with the Repbase [19] database to create the final repeat library. Repeat sequences in the *O. shuilongensis* genome were identified and classified using the RepeatMasker v4.0.6 [20] program.

Annotation of Protein Coding Genes

Protein-coding genes were predicted based on *de novo*, protein homology-based and RNA-Seq method. We first used five *de novo* gene prediction tools including Genscan [21], Augustus v2.4 [22], GlimmerHMM v3.0.4 [23], GeneID v1.4 [24] and SNAP [25] to predict protein-coding genes in the *O. shuilongensis* genome. For homology-based gene prediction, protein sequences from four closely related teleost species including *Cyprinus carpio*, *Danio rerio*, *Sinocyclocheilus rhinocerous* and *Astyanax mexicanus* were downloaded from Ensembl database and aligned against to the *O. shuilongensis* genome using GeMoMa v1.3.1 [26] software; the RNA-Seq reads were assembled into contigs *de novo* into unigenes using Trinity and the unigenes were aligned to the repeat-masked assemblies using BLAT [27], and subsequently the gene structures of BLAT alignment results were modeled using PASA v2.02 [28]. Additionally, the RNA-Seq reads were also assembled into transcripts through mapping to the assembled genome using Hisat2 v2.0.4 [29] and StringTie v1.3.0 [30], and the protein-coding regions were identified with TransDecoder v2.0 [31] and GeneMarkS-T v5.1 [32], respectively. All these consensus gene models were generated by integrating the *de novo* predictions, protein alignments and transcripts data using EvidenceModeler [33].

Gene Functional Annotation

Annotation of the predicted genes was performed by local BLAST [27] programs blasting their sequences against a number of nucleotide and protein sequence databases, including NCBI-NR [34] , KOG [35], KEGG [36], and TrEMBL [37] with an E-value cutoff of 1e-5. We then searched the Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway databases using the software Blast2GO [38].

Noncoding RNA Annotation

Non-coding RNAs play important roles in a great variety of processes, such as the rRNAs and tRNAs involved in mRNA translation. The rRNA fragments were identified by aligning the rRNA template sequences using BLAST with E-value at 1e-10 and identity cutoff at 95% or more. The tRNAScan-SE v1.3.1 [39] algorithms with default parameters were applied to the prediction of tRNA genes. The rRNA and miRNA genes were predicted by Infernal v1.1 [40] against the Rfam [41] database and miRbase [42] with cutoff score at 30 or more. The minimum a cutoff score was based on the settings which yield a false positive rate of 30 bits.

Global Gene Family Classification

In order to identify gene families among fish species in this work, proteins of the longest transcripts of each individual gene from *O. shuilongensis* and other sequenced species, including *Salmo salar*, *Ictalurus punctatus*, *A. mexicanus*, *C. carpio*, *S. rhinocerous*, *D. rerio* and *Larimichthys crocea* were analyzed. All data was downloaded from NCBI [43]. Gene family analysis based on the homolog of gene sequences in related species was initially implemented by the alignment of an “all against all” BLASTP [44] with a cutoff of 1e-5 and subsequently followed by alignments with high-scoring segment pairs conjoined for each gene pair by Solar. To identify homologous gene pairs, we required more than 30% coverage of the aligned regions in both homologous genes. Finally, homologous genes were clustered into gene families by OrthoMCL [45] with the inflation parameter set at 1.5.

Phylogenetic Relationship and Genomic Comparison

Evolutionary analysis was performed using the single-copy protein-coding genes among all species. Amino acid and nucleotide sequences of the ortholog genes were aligned using the multiple alignment software MUSCLE [46] with default parameters. A total number of 724 single-copy ortholog alignments were concatenated into a super alignment matrix of 1,692,085 nucleotides. A maximum likelihood method deduced tree was inferred based on the matrix of nucleotide sequences using PhyML package with the JTT+G+F model. Clade support was assessed using bootstrapping algorithm in the PhyML package with 100 alignment replicates. The constructed phylogenetic tree indicated that *O. shuilongensis* were clustered closely to Cyprinidae species, which is inconsistent with their putative evolutionary relationships (Figure 3).

We determined the expansion and contraction of the orthologous gene families by comparing the cluster size differences between the ancestor and each of the *O. shuilongensis* and seven other fish species using the CAFÉ [47] program. A random birth and death model were used to study changes of gene families along each lineage of the phylogenetic tree. A probabilistic graphical model (PGM) was introduced to calculate the probability of transitions in gene family size from parent to child nodes in the phylogeny. Using conditional likelihoods as the test statistics, we calculated the corresponding *P*-values in each lineage. A *P*-value of 0.05 was used to identify families that were significantly expanded in *O. shuilongensis* genome.

Dating the Divergence among Oreonectes Fishes and Assessing the Genomic Mechanism Underlying the Degeneration of Eyes and Body Color

O. shuilongensis is a surface-dwelling species with intact eyes and a unique color pattern consisting of fine black marks on the body except on the abdomen [3]. To explore the genomic changes resulting in the degeneration of eyes and body color of semi cave-dwelling and cave-dwelling *Oreonectes* fishes, the whole genome re-sequencing was performed for facultative cave-dwelling *O. jiarongensis* (three individuals) and cave-dwelling *O. daqikongensis* (two individuals) and *O. dongliangensis* (one individual). For each individual, ~3μg of DNA was sheared into fragments of 350 bp with the Covaris v1.8 system. DNA fragments were then processed and sequenced using the Illumina HiSeq 4000 platform. Filtered sequence reads were mapped to the *O. shuilongensis* reference genome using BWA-MEM with default parameters (0.7.10-r789)[10]. Alignment bam files were imported to SAMtools v 0.1.19 [48] for sorting and Picard v1.92 [49] was used to remove duplicated reads. Following mapping, we performed variant calling using the GATK [50] package with default parameters. To explore the phylogenetic relationships among *Oreonectes* fishes, phylogenetic tree was inferred using the Neighbor-Joining (NJ) algorithm as implemented in RaxML [51] software with 1000 bootstraps and the divergence times of the taxa analyzed were estimated with mcmctree [14]. The outgroup sequences were chosen from the zebrafish genome assembly GRCz11, with the genome alignment to the obtained *Oreonectes* genome assembly using LASTZ [52]. We employed calibration points from the dated Cyprinidae-Nemacheilidae divergence to place a lognormally distributed prior on the age of the root of a tree containing all samples of *Oreonectes* species and outgroup. To explore the genomic mechanism underlying the degeneration of eyes and body color, the SNPs and Indels obtained through GATK pipeline were annotated using the software SnpEff [6].

Abbreviations

BLAST: Basic Local Alignment Search Tool;

KEGG: Kyoto Encyclopedia of Genes and Genomes;

NCBI: National Center for Biotechnology Information;

CGEs: Core Eukaryotic Genes;

TrEMBL: translated European Molecular Biology Laboratory;

SMS: single-molecule sequencing;

GO: Gene ontology.

Declarations

Ethics approval and consent to participate

No regulated invertebrates were involved in this study.

Consent for publication

Not applicable.

Availability of data and materials

Supporting data is available in the Genome Sequence Archive (GSA, https://ngdc.cncb.ac.cn/gsa/). Raw data has been deposited in GSA with the project accession CRA004953. BioSample accessions: SAMC452834-452840.

Competing interests

The authors declare no competing interests.

Funding

This project was supported by the key project of Science-technology basic condition platform from The Key Project of Science and Technology Program of Guizhou Province “Study of biological and ecological vales for Fanjingshan World Natural Heritage Nomination of Tongren City (3052 2015 Qiankehe SY)”, Science-technology basic condition platform from The Ministry of Science and Technology of the People’s Republic of China (Grant No. 2005DKA21402), Provincial Governor Fund of Guizhou (Qian-Sheng-Zhuan-He-Zi (2010) 14) project and the Youth Innovation Promotion Association of Chinese Academy of Sciences (2011077).

Authors' contributions

ZJ L, XF Z and HM W performed the experiments, analyzed the data, wrote the paper, and participated in the design of the study. L H participated in data statistics and the draft manuscript. J Z and ZJ L conceived the study, participated in its design and coordination of the research. All authors read and approved the final manuscript.

Acknowledgements

We are further grateful to Kai Gao, Jian Chen, Zhonglong Cen and Gang Li for their help in sampling and laboratory work. We thank Martin Burrows and Wenhua Xiong for assistance in writing and data analysis.

References

1. Zhu S. The Loaches of the Subfamily Nemacheilinae in China (Cypriniformes: Cobitidae)[M]. Jiangsu Science & Technology Publishing House.; 1989.

2. Liu Z, Wen H, Hailer F, Dong F, Yang Z, Liu T, et al. Pseudogenization of Mc1r gene associated with transcriptional changes related to melanogenesis explains leucistic phenotypes in Oreonectes cavefish (Cypriniformes, Nemacheilidae). J Zool Syst Evol Res. 2019;900.0-909.

3. Deng H, Xiao N, Hou X, Zhou J. A new species of the genus Oreonectes (Cypriniformes: Nemacheilidae) from Guizhou, China. Zootaxa. 2016;143.0-150.

4. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;1061–7.

5. Simão AF, Waterhouse MR, Ioannidis P, Kriventseva VE, Zdobnov ME. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;3210.0-3212.0.

6. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;80–92.

7. Huang WD, Sherman TB, Lempicki AR. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;44.0-57.0.

8. Ogawa Y, Shiraki T, Asano Y, Muto A, Kawakami K, Suzuki Y, et al. Six6 and Six7 coordinately regulate expression of middle-wavelength opsins in zebrafish. Proc Natl Acad Sci U S A. 2019;

9. Pittlik S, Domingues S, Meyer A, Begemann G. Expression of zebrafish aldh1a3 (raldh3) and absence of aldh1a1 in teleosts. Gene Expr Patterns. 2008;141–7.

10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;1754–60.

11. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. PloS One. 2012;e52249–e52249.

12. Koren S, Walenz PB, Berlin K, Miller RJ, Bergman HN, Phillippy MA. Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. Genome Res. 2017;722–36.

13. Walker JB, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One. 2014;e112963–e112963.

14. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;W265-8.

15. Han Y, Wessler RS. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;e199–e199.

16. Price LA, Jones CN, Pevzner AP. De novo identification of repeat families in large genomes. ISMB Suppl Bioinforma. 2005;351–8.

17. Edgar CR, Myers WE. PILER: identification and classification of genomic repeats. ISMB Suppl Bioinforma. 2005;152–8.

18. Wicker T, Sabot F, Hua-Van A, Bennetzen LJ, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;973–82.

19. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;462–7.

20. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinforma. 2004;Unit 4.10-Unit 4.10.

21. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;78–94.

22. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;II215–25.

23. Majoros HW, Pertea M, Salzberg S. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;2878–9.

24. Blanco E, Parra G, Guigó R. Using geneid to identify genes. Curr Protoc Bioinforma. 2007;Unit 4.3-Unit 4.3.

25. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;59–59.

26. Keilwagen J, Wenk M, Erickson LJ, Schattat HM, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;

27. Altschul FS, Gish W, Miller CW, Myers WE, Lipman JD. Basic local alignment search tool. J Mol Biol. 1997;403–10.

28. Campbell AM, Haas JB, Hamilton PJ, Mount MS, Buell RC. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics. 2006;327–327.

29. Kim D, Langmead B, Salzberg LS, Langmead B, Langmead B. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;357-U121.

30. Pertea M, Pertea MG, Antonescu MC, Salzberg LS, Antonescu MC, Chang T-C, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;290-+.

31. Haas B, Zimmermann B, Crusoe MR, MacManes M, Plessy C. TransDecoder (Find Coding Regions Within Transcripts). 2015. Available from: https://github.com/TransDecoder/TransDecoder.wiki.git

32. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 2015;e78–e78.

33. Haas JB, Salzberg LS, Zhu W, Pertea M, Allen EJ, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;R7–R7.

34. Marchler-Bauer A, Lu S, Anderson BJ, Chitsaz F, Derbyshire KM, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2011;D225-9.

35. Tatusov LR, Natale AD, Garkavtsev VI, Tatusova AT, Shankavaram TU, Rao SB, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001;22–8.

36. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;27–30.

37. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;365–70.

38. Conesa A, Götz S, García-Gómez MJ, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;3674–6.

39. Lowe MT, Eddy RS. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;955–64.

40. Nawrocki PE, Eddy RS. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;2933–5.

41. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy RS, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 2005;D121-4.

42. Griffiths-Jones S, Grocock JR, Dongen van S, Bateman A, Enright JA. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;D140-4.

43. Wheeler LD, Church MD, Lash EA, Leipe DD, Madden LT, Pontius UJ, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018;6–17.

44. Altschul FS, Madden LT, Schäffer AA, Zhang J, Zhang Z, Miller CW, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Faseb J. 1998;A1326–A1326.

45. Fischer S, Brunk PB, Chen F, Gao X, Harb SO, Iodice BJ, et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinforma. 2011;Unit 6.12.1-19.

46. Edgar CR. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;1792–7.

47. Bie DT, Cristianini N, Demuth PJ, Hahn WM. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;1269–71.

48. Heng L, Bob H, Alec W, Tim F, Jue R, Nils H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;2078–9.

49. Picard toolkit. Broad Institute, GitHub repository. Broad Institute; 2019. Available from: https://broadinstitute.github.io/picard/

50. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;1297–303.

51. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;1312–3.

52. Harris R. Improved Pairwise Alignment of Genomic DNA [PhD Thesis]. ProQuest. 2007.

Tables

Table 1. Summary of sequence data from *Orenectus shuilongensis*. The sequencing data used in this work. Note that read length for PacBio Sequel were measured for subreads.

Type	Libraries	Insert size (bp)	Read length (bp)	Clean data (Gb)	Coverage (X)
DNA	HiSeq X Ten	270	150	120.94	234.55
DNA	PacBio	20,000	10,187	50.94	98.79
RNA	HiSeq X Ten	270	150	11.7	-

The coverage was calculated using an estimated genome size of 515.64 Mb.

Table 2. Summary of the loach (*O. shuilongensis*) genome assembly and annotation.

Genome quality	values
Contig N50 size (Mb)	5.58
Estimated genome size (Mb)	515.64
Assembled genome size (Mb)	521.68
Genome coverage (fold)	234.55 X
Content of transposable elements (%)	23.42
Total GC content (%)	38.51
Protein-coding gene numbers	25,247
Average gene length (kb)	234.62
Average CDS length (bp)	2,010

Table 3. Annotation of repeat sequences in the loach (*O. shuilongensis*) genome assembly.

Type	Number	Length（bp）	Percentage (%)
ClassI/DIRS	1,566	1,563,184	0.3
ClassI/LINE	6,300	11,594,987	2.22
ClassI/LTR	5,947	4,877,346	0.93
ClassI/LTR/Copia	909	528,739	0.1
ClassI/LTR/Gypsy	9,340	6,473,497	1.24
ClassI/PLE|LARD	7,579	15,588,001	2.99
ClassI/SINE	4,322	591,812	0.11
ClassI/TRIM	3,358	3,025,034	0.58
ClassI/Unknown	342	35,272	0.01
ClassII/Crypton	11,782	2,048,568	0.39
ClassII/Helitron	8,994	1,095,311	0.21
ClassII/MITE	4,176	596,316	0.11
ClassII/Maverick	2,145	374,478	0.07
ClassII/TIR	71,957	21,304,786	4.08
ClassII/Unknown	178,045	22,476,697	4.31
PotentialHostGene	32,601	4,535,251	0.87
SSR	797	112,798	0.02
Unknown	178,496	25,369,612	4.86
Total without overlap:	528,656	122,191,689	23.42

Table 4. Re-sequencing analysis of cave-dwelling *Orenectus* fishes.

Species	*O. jiarongensis*			*O. daqikongensis*		*O. dongliangensis*
Sample ID	R01	R02	R03	R04	R05	R06
Total_reads	154919048	145879318	136865354	147160730	126896868	123201446
Clean_Reads	77459524	72939659	68432677	73580365	63448434	61600723
GC(%)	39.06	39.02	38.92	38.92	38.95	39.38
Mapping_rate	90.49%	91.15%	91.83%	90.82%	91.63%	92.26%
Ave_depth	29	28	26	30	26	24
Cov_ratio_1X(%)	87.44	87.76	87.62	87.26	86.98	87.46
Cov_ratio_5X(%)	83.46	83.74	83.5	83.37	82.76	83.01
Cov_ratio_10X(%)	80.35	80.56	80.07	80.44	79.3	79.05
Total_SNP	6201592	6207932	6201733	6064030	6049874	6189968
Transition	3544204	3547586	3544063	3485910	3478106	3539101
Transversion	2657388	2660346	2657670	2578120	2571768	2650867
Ti/Tv	1.33	1.33	1.33	1.35	1.35	1.33
Het_snp	199131	193488	191382	183450	182330	248330
Hom_snp	6002461	6014444	6010351	5880580	5867544	5941638
Het-ratio	3.21%	3.11%	3.08%	3.02%	3.01%	4.01%
Total_Indel	2335694	2340858	2336570	2218308	2208883	2321897
Het_Indel	113531	111329	108840	112158	110204	135160
Hom_Indel	2222163	2229529	2227730	2106150	2098679	2186737

*Ave_depth is average sequencing depth; Cov_ratio_1X, Cov_ratio_5X and Cov_ratio_10X are described respectively; Ti/Tv is Transition/Transversion; Het_snp is heterozygosis SNP number; Hom_snp is homozygosis SNP number; Het_ratio is the percentage of heterozygosis SNPs in total SNPs; Het_Indel is heterozygosis indel numbers; Hom_Indel is homozygosis indel numbers.

Table 5. Functional classification and enrichment analysis of pseudogenes in cave-dwelling *Oreonectes* fishes. Gray boxed GOs are associated with retina and eye development.

GO_ID	GO_Term	GO Class	Gene num	Adjusted *P*-value
GO:0005267	potassium channel activity	MF	5	0.00012
GO:0000276	mitochondrial proton-transporting ATP synthase complex, coupling factor	CC	2	0.049
GO:0046332	SMAD binding	MF	2	0.044
GO:1902287	semaphorin-plexin signaling pathway involved in axon guidance	BP	2	0.033
GO:0002116	semaphorin receptor complex	CC	2	0.075
GO:0048841	regulation of axon extension involved in axon guidance	BP	2	0.033
GO:0021785	branchiomotor neuron axon guidance	BP	2	0.025
GO:0006813	potassium ion transport	BP	4	0.013
GO:0005251	delayed rectifier potassium channel activity	MF	3	0.011
GO:0005244	voltage-gated ion channel activity	MF	4	0.0038
GO:0034765	regulation of ion transmembrane transport	BP	4	0.0085
GO:0004930	G-protein coupled receptor activity	MF	10	0.0083
GO:0007165	signal transduction	BP	13	0.0043
GO:0005249	voltage-gated potassium channel activity	MF	4	0.0038
GO:0007186	G-protein coupled receptor signaling pathway	BP	11	0.0033
GO:0051260	protein homooligomerization	BP	4	0.0031
GO:0008076	voltage-gated potassium channel complex	CC	4	0.0038
GO:0071805	potassium ion transmembrane transport	BP	5	0.013
GO:0060041	retina development in camera-type eye	BP	9	2.8E-9
GO:0001654	eye development	BP	8	8.9E-9
dre04080*	Neuroactive ligand-receptor interaction	KEGG	5	0.0088

Figures

Figure 1

Photo of (a) O. shuilongensis, (b) O. jiarongensis, (c) O. dongliangensis and (d) O. daqikongensis.

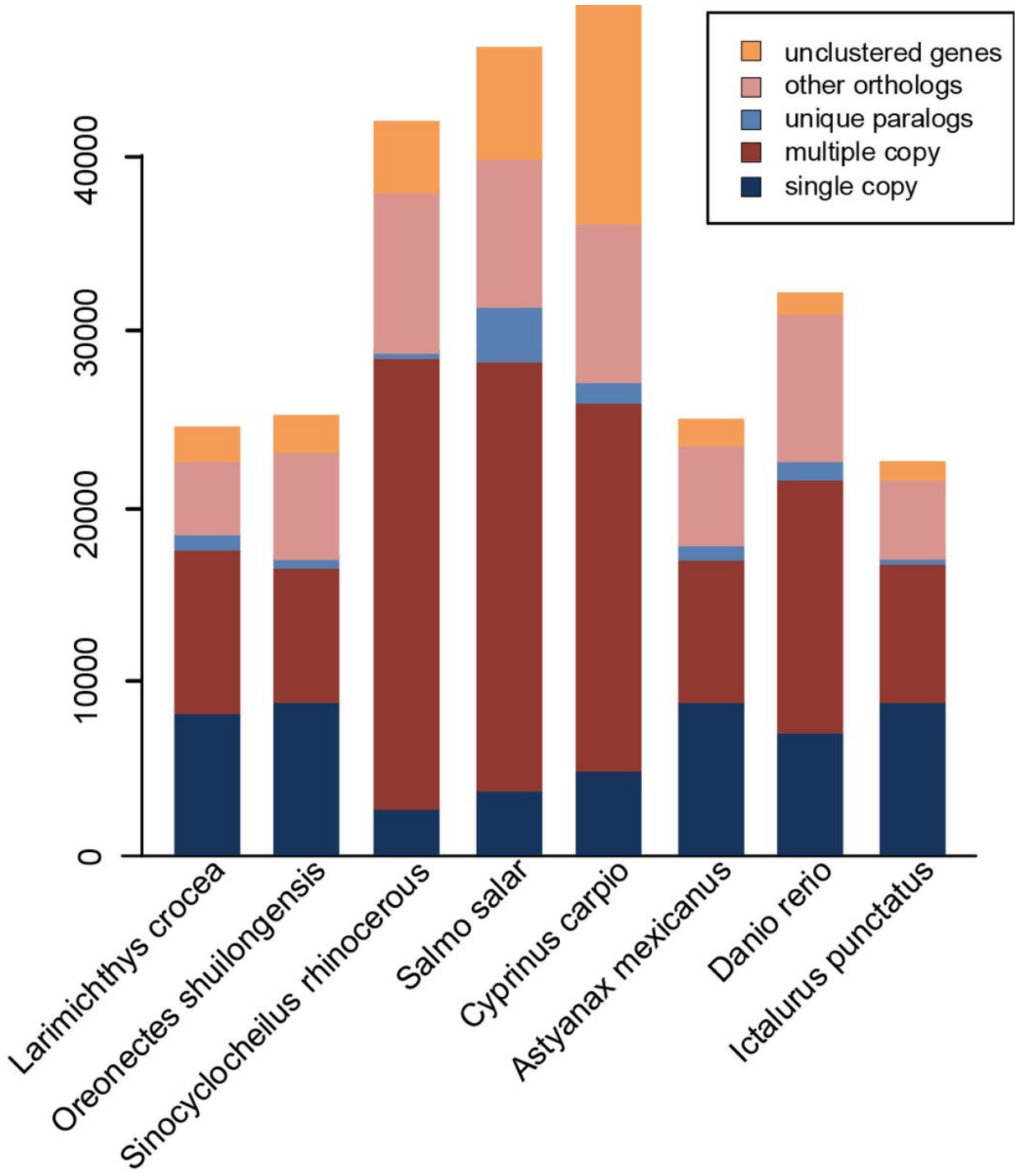

Figure 2

Gene family comparison between O. shuilongensis and other fish species. The Y-axis represented the gene number for each class: single-copy (one gene for each species), multiple-copy (more than one gene for each species), unique paralogs (no genes in other species), other orthologs (other cases in gene clusters) and un-clustered genes (genes that did not clustered with other genes).

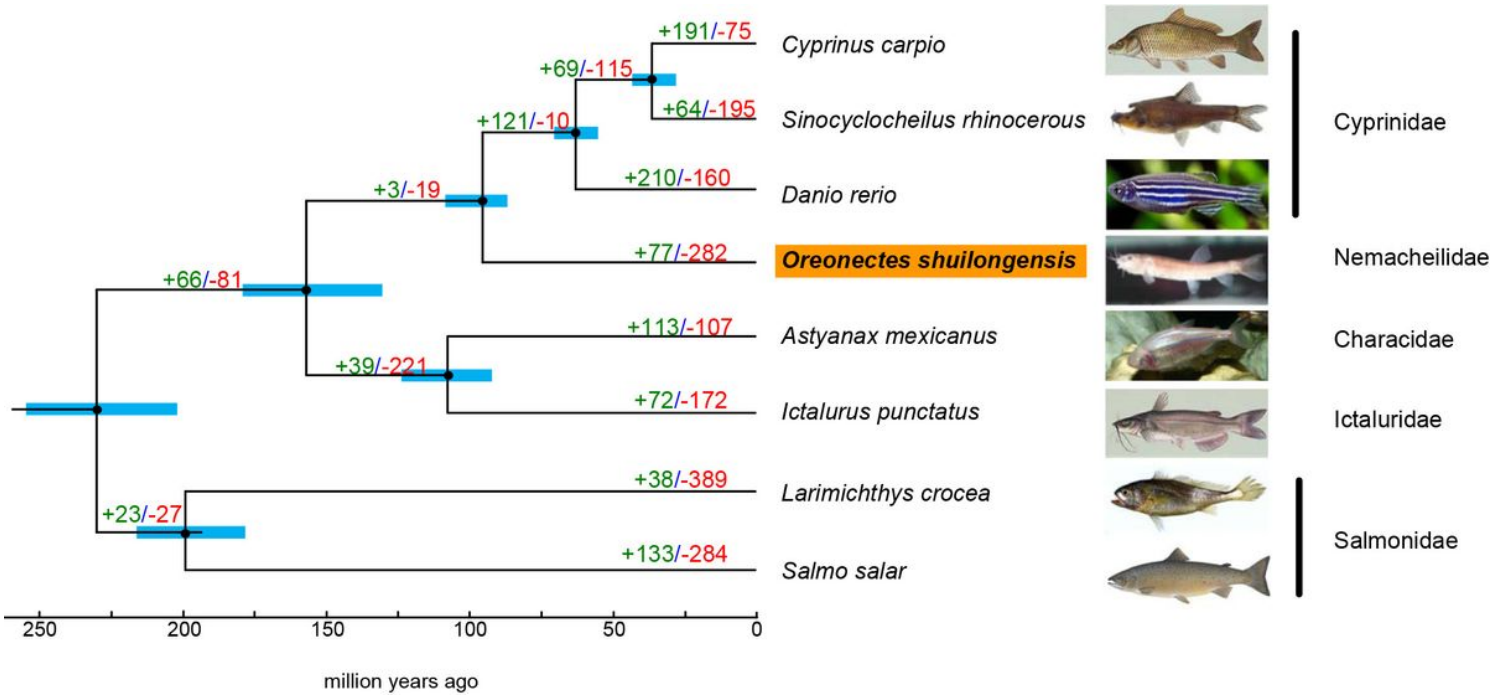

Figure 3

Phylogenetic relationships and estimated divergence times in million years ago (MYA) of O. shuilongensis and other fish. The numbers on each branch correspond to the numbers of gene families that have expanded (green) or contracted (red).

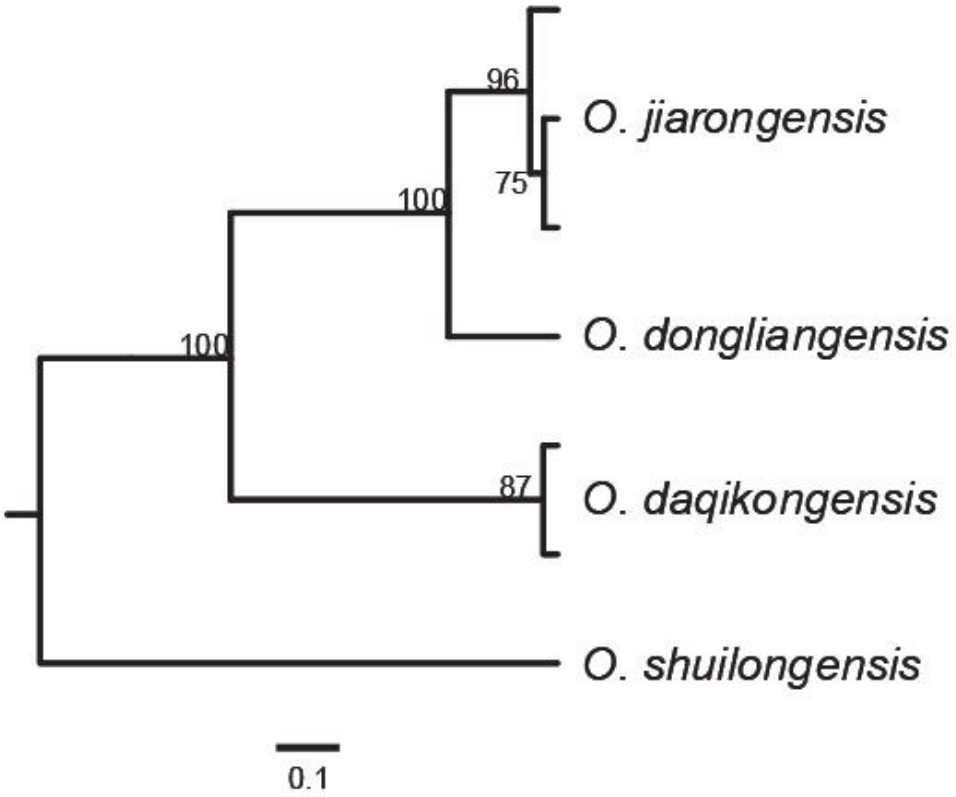

Figure 4

Neighbor-joining phylogenomic tree of Orenectus fishes, based on ML distances calculated from 1,364,638 SNPs. Numbers on branches denote support estimated from 1000 bootstrap replicates. The root was obtained from the orthologous sequences of Zebrafish.

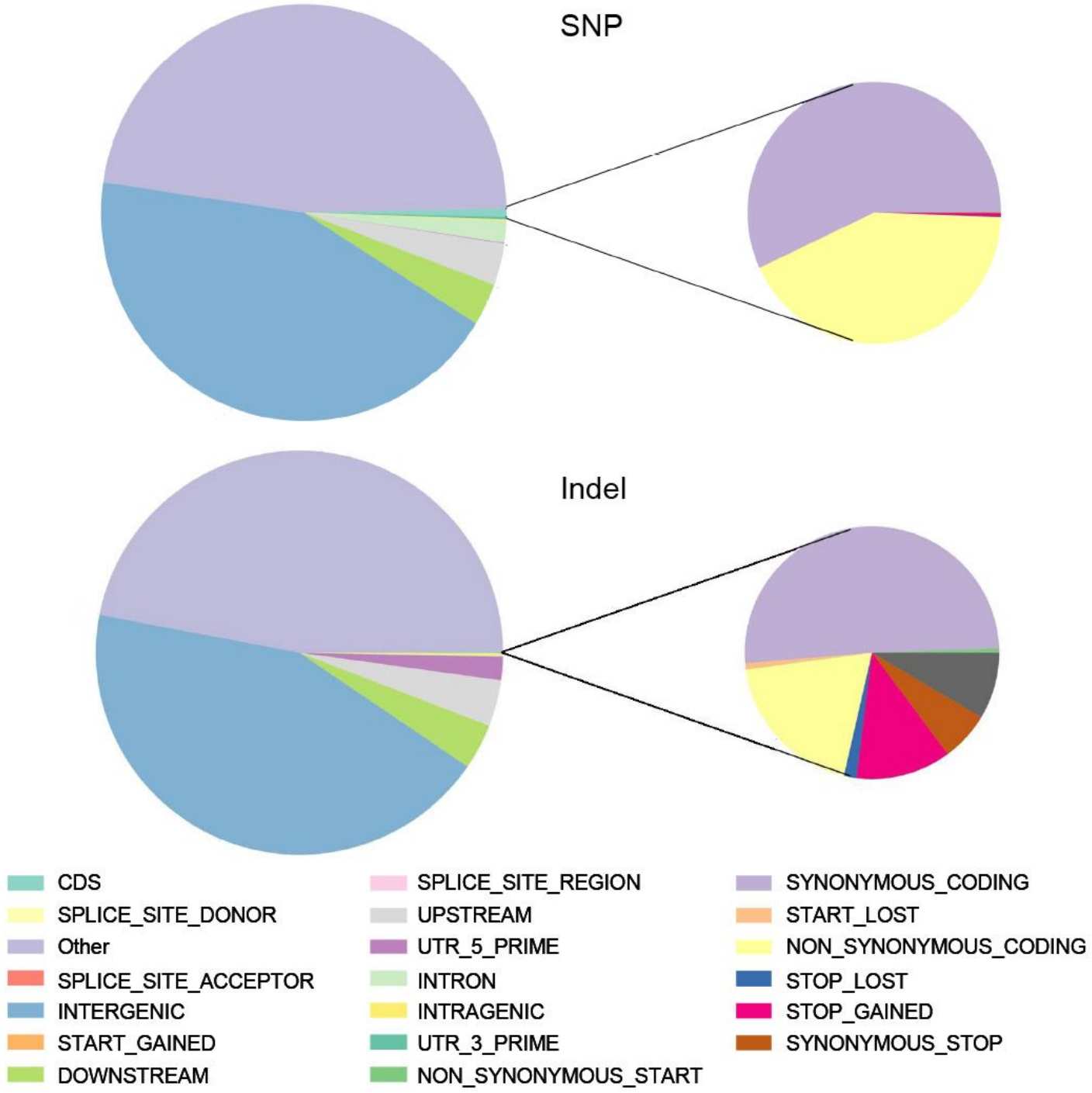

Figure 5

SNPs (a) and Indels (b) annotation of cave-dwelling O. jiarongensis, O. daqikongensis and O. shuilongensis.

Figure 6

Pseudogenes related to the GO terms “retina development in camera-type eye” and “eye development” among cave-dwelling O. jiarongensis, O. daqikongensis and O. shuilongensis.

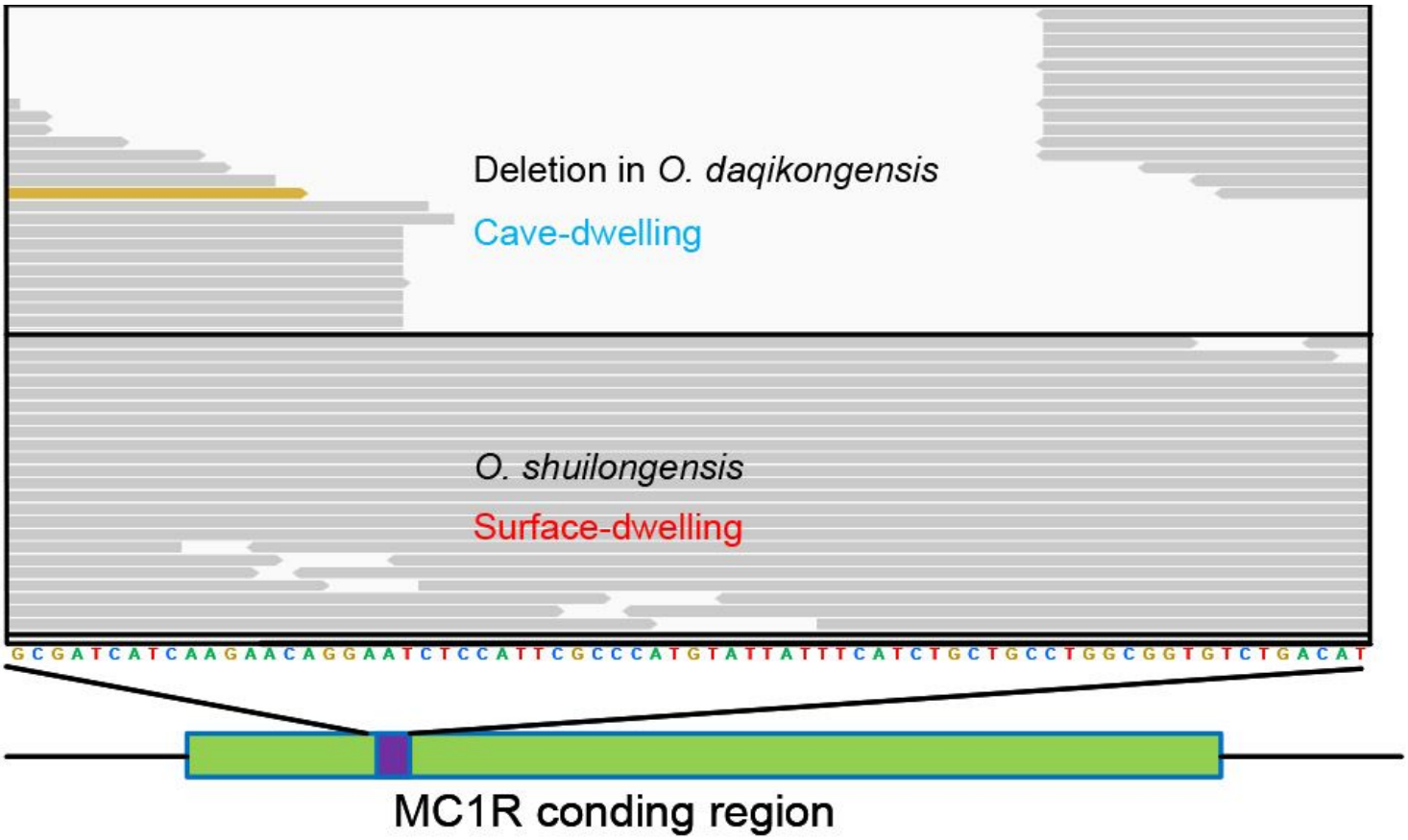

Figure 7

Pseudogene of MC1R in O. daqikongenis caused by 29 bp deletion in coding region.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- cavefishgenomeSI.docx