Novel genome sequence of Chinese cavefish (*Triplophysa rosa*) reveals pervasive relaxation of natural selection in cavefish genomes

Qingyuan Zhao [1,2], Feng Shao [1], Yanping Li [1,3], Soojin V. Yi [4,*], Zuogang Peng [1,5,*]

[1] Key Laboratory of Freshwater Fish Reproduction and Development (Ministry of Education), Southwest University School of Life Sciences, Chongqing 400715, China

[2] Department of Laboratory Animal Science, College of Basic Medical Sciences, Army Medical University (Third Military Medical University), Chongqing 400038, China

[3] Key Laboratory of Sichuan Province for Fish Conservation and Utilization in the Upper Reaches of the Yangtze River, Neijiang Normal University College of Life Sciences, Neijiang 641000, China

[4] Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA 93106, USA

[5] Academy of Plateau Science and Sustainability, Qinghai Normal University, Xining, 810008, China

Running title: Comparative genomics of cavefishes

* To whom correspondence may be addressed. Email: soojinyi@ucsb.edu or pzg@swu.edu.cn

Materials and Methods

Sample collection and sequencing

Two healthy female *Triplophysa rosa* loaches were collected from an underground cave in WuLong (29°23'57.27"N, 107°54'26.84") Chongqing, China. Of them, one was used for Illumina and SMRT sequencing, whereas the other was used for Hi-C sequencing. To better store samples for Hi-C analyses, live muscle tissues collected from *T. rosa* were segmented and placed in 90% fetal bovine serum and 10% DMSO, and stored at low temperature as follows: after incubating for 30 min in a refrigerator at 4 ℃, the sample was transferred to a refrigerator at -20 ℃ for 2 h and finally stored at -80 ℃. In addition, normal-molecular-weight genomic DNA for Illumina sequencing was extracted from the muscle tissues using the CTAB method (Murray & Thompson, 1980). DNA concentrations were quantified using Qubit dsDNA HS Assay Kit (Life technologies, Carlsbad, CA, USA), followed by 1.5% agarose gel electrophoresis to check its integrity.

Genomic DNA extracted from the muscle tissues was mechanically sheared using sonication. The fragmented DNA was purified using a 1.8-fold sample volume of Beckman AMPure XP Beads (Beckman Coulter, Beverly, MA, USA). The purified DNA fragments were blunt-end converted and ligated using VAHTS TM Turbo DNA Library Prep Kit for Illumina (Vazyme Biotech Co., Ltd, NanJing, China). Purified ligation products were then separated using electrophoresis in 2% agarose gel (BioRad, Hercules, California, USA). Subsequently, we obtained 350–400 bp (insert size 270 bp) and 650 bp (insert size 500 bp) fragments that were further purified using QIAquick Gel Extraction Kit (QIAGEN, Duesseldorf, Germany). Quality was tested and concentration was quantified via 7,500 DNA LabChip by Agilent Technologies 2100 Bioanalyzer (Waldbronn, Germany). The libraries of insert size 3k bp, 4k bp, 8k bp, 10k bp, 15k bp, 17k bp were constructed using Nextetra® Mate Pair Library Preparation Kit (Illumina, San Diego, CA, USA). Libraries were pooled accordingly following qPCR results. The 270 bp library was sequenced on an Illumina HiSeq 4000 sequencer with a pair-end of 150 bp, whereas the other libraries were sequenced on an Illumina HiSeq 2500 sequencer with a

pair-end of 125 bp. FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was used for Illumina read filtering, with the following parameters: percent cutoff of PHRED>30 bp as 80%, reads more than 100 bp with no adapter contamination, and less than 3 bp aligned to adapter sequence. We constructed a total of 13 Illumina sequencing libraries as mentioned above, three of which were short-fragment paired-end sequencing libraries (two with 270 bp insert size and one with 500 bp insert size). The results of Illumina sequencing from paired-end and mate-pair libraries are shown in Table S2. A total of 141 Gb of Illumina sequencing data was obtained after filtering low-quality reads (Table S2). All libraries had a Q20 base ratio of > 93% and a Q30 base ratio of > 87%. The total amount of sequencing data exceeded 207X coverage.

To improve the continuity and integrity of genome assembly sequences, we performed the PacBio (Single Molecule Real Time, SMRT) sequencing. Genomic DNA was dissolved at 37 ℃ for 30 min in 50 μL RNaseA-containing ddH2O (Thermo Fisher Scientific, PuDong, ShangHai, China). Quantification and integrity were determined using Qubit dsDNA HS Assay Kit and 0.75% agarose gel electrophoresis, respectively. Further, the DNA was purified by Beckman AMPure XP Beads. Its concentration was estimated using Nanodrop 2000 (Thermo Scientific, PuDong, ShangHai, China) and Qubit (Thermo Fisher Scientific, PuDong, ShangHai, China) (ratio of Nanodrop/Qubit readings less than 2 were used further sequencing). The purified DNA was subjected to electrophoresis to confirm that the main band was clearly larger than 23Kb. Following the Covaris g-Tube (Covaris, Woburn, MA, USA) shearing, SMRTbell libraries were prepared with insert sizes in the range of 10–20 Kb, followed by the quantification and size detection using Qubit2.0 and Agilent Technologies 2100 Bioanalyzer, respectively. BluePippin Size-Selection System (Sage Science, Beverly, MA, USA) was applied to the libraries. Further, the libraries were recovered using PB AMPure beads (Pacific Biosciences of California, Menlo Park, CA, USA), quantified using Qubit dsDNA HS Assay Kit, and assessed for quality using DNA12000 kit (Agilent, Waldbronn, Germany) on an Agilent Technologies 2100 Bioanalyzer. SMRT bell templates were bound to P6 polymerase using the DNA polymerase binding kit (Pacific Biosciences of California, Menlo Park, CA, USA) P6 v2 primers. Polymerase-template

complexes were bound to magnetic beads using the Magbead Binding Kit (Pacific Biosciences of California, Menlo Park, CA, USA) and sequenced on PacBio RS II sequencer using C4v1 sequencing reagents with movie length of 360 min on SMRT cells. All CLR type PacBio sequencing was based on nine cells. The sequencing data were filtered using SMRT analysis with the quality<0.75 and length<500 bp. Consequently, we acquired high-quality reads with read N50 of 26,821 bp, mean read length of 17,994 bp, and read quality of 0.826. After further filtering shorter-length reads and removing adapter and primer sequences, we acquired 1,209,649 subreads with a total length of 11.60 Gb, of which the longest read length was 74.60 Kb, average length was above 9.59 Kb, and read N50 was 15.40 Kb (Table S3, S4).

We prepared the Hi-C library according to the procedures of Lieberman-Aiden et al (Lieberman-Aiden et al., 2009). Briefly, we fixed the samples in 37% formaldehyde with cross-linked intracellular protein with DNA and DNA with DNA, preserved the interactions, and obtained the 3D structure inside the cell. Genomic DNA was further digested using the restriction enzyme Hind III to produce sticky ends on both sides of the crosslink. Terminal repair mechanism was used to introduce biotin-labeled bases for subsequent DNA purification and capture; end-repaired DNA was then cyclized, de-crosslinked, purified, and fragmented into 300-700 bp fragments. DNA fragments containing interactions were captured using streptavidin magnetic beads for library construction. After the library was constructed, its concentration and insert size were detected using Qubit2.0 and Agilent Technologies 2100 Bioanalyzer, respectively, and the effective concentration of the library was accurately quantified using qPCR to ensure library quality. Consequently, we constructed two Hi-C sequencing libraries and obtained 111.29 Gb clean data (approximately 371.56 M reads) using the Illumina sequencing platform. A total of 241 M (64.88% of all pair-end reads) uniquely mapped pair-end reads were screened out using bowtie v1.0.0 to align all the reads to genome sequences. Based on the uniquely mapped pair-end reads, we evaluated the effective data rate of Hi-C sequencing and obtained 93,08 M valid interaction pairs (mean effective Hi-C data, 38.61% of unique mapped pair-end reads; Table S5).

Genome size estimation based on next-generation sequencing (NGS) data

To determine whether the extracted sample DNA was contaminated, we randomly selected 10,000 single-end reads from the sequenced 270 bp library and performed BLAST v 2.2.28+ (Altschul et al., 1990) with the Nucleotide Sequence Database (NT, https://ncbi.nlm.nih.gov) To assess the mitochondrial content in the *T. rosa* sequencing library, we performed SOAP alignment (SOAP2, https://github.com/ShujiaHuang/SOAPaligner) (Li et al., 2008) with the mitochondrial genome (from *T. rosa*: NC_019587.1 and *T. bleekeri*: NC_018774.1) of the 270 bp library. Approximately 50 paired-end and 3,699 singleton reads were aligned to the mitochondrial sequences; both the ratios of the paired-end and singleton reads alignment were near 0%. Therefore, we deduced that the mitochondrial content in *T. rosa* sequencing library was extremely low; thereby not affecting/interfering the evaluation of the genome size Characteristics of the *T. rosa* genome were evaluated using the K-mer based method (Liu et al., 2013). From the frequencies of 17-mers in result of jellyfish (https://github.com/gmarcais/Jellyfish) (Marcais & Kingsford, 2011) (Fig. S1; Total number of K-mers: 31496014642), We estimated that the genome size of *T. rosa* was 732 Mb (mid-sized), with 39.87% GC content, 0.15% heterozygosity (low), and 42% repeat content. The currently estimated genome size is smaller than the previous estimate based on flow cytometry (Niu et al., 2017). Estimated genome sizes are known to vary according to different methods (Dolezel & Greilhuber, 2010; Pflug et al., 2020; Pucker, 2019). Further studies are needed to yield an accurate estimate of the genome size of *T. rosa*. The low heterozygosity might imply a low population genetic diversity of the loach in the underground cave, potentially reflecting a founder effect. We have previously suggested higher population diversities than the current estimate (Liu et al., 2017; Zhao et al., 2014). The difference may be related to the different methods used: previous studies employed highly polymorphic microsatellite markers.

Genome annotations

We constructed a specific repetitive sequence database for *T. rosa* as follows: we used LTR_FINDER v1.05 (parameters: -S 6) (https://github.com/xzhub/LTR_Finder) (Xu & Wang, 2007), MITE-Hunter v1.0.10 (http://target.iplantcollaborative.org/mite_hunter.html) (Han & Wessler, 2010), RepeatScout v1.0.5 (https://github.com/mmcco/RepeatScout/) (Price et al., 2005), and PILER-DF v2.4 (Edgar & Myers, 2005) tools to construct a repetitive sequence database of *T. rosa* genome based on the principles of structural prediction and *de novo* prediction. Further, using the PASTEClassifier v1.0 (https://urgi.versailles.inra.fr/Tools/PASTEClassifier/) (Wicker et al., 2007)), we classified the database and merged it with the Repbase (https://www.girinst.org/repbase) (Jurka et al., 2005) as the final repetitive sequence database. Subsequently, we used the RepeatMasker v4.0.5 software (http://repeatmasker.org/) (Tarailo-Graovac & Chen, 2009) (with the parameters: -s, -q, -species Animals) and predicted repetitive sequences of *T. rosa* based on database constructed by us as mentioned above. Consequently, we obtained 324.63 Mb of the repeat sequence.

Motif is a local conserved region in a sequence, or a small sequence pattern shared by a group of sequences. Generally, it refers to the basic structure that constitutes any one of the characteristic sequences, but in most cases, refers to any sequence pattern that may have molecular function, structural properties, or gene family member correlation. In this report, we used InterProScan (www.ebi.ac.uk/interpro/) (Jones et al., 2014) to predict motifs based on the PROSITE (http://prosite.expasy.org/) (Bairoch, 1991), HAMAP (http://hamap.expasy.org/) (Lima et al., 2009), Pfam (http://pfam.xfam.org/) (Finn et al., 2014), PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php) (Attwood & Beck, 1994), ProDom (http://www.toulouse.inra.fr/prodom.html) (Bru et al., 2005), SMART (http://smart.embl.de) (Letunic et al., 2004), TIGRFAMs (http://www.jcvi.org/cgi-bin/tigrfams/index.cgi) (Haft et al., 2003), PIRSF (https://proteininformationresource.org/) (Wu et al., 2004), SUPERFAMILY (http://supfam.org/) (Gough & Chothia, 2002), CATH-Gene3D (https://cathdb.info/) (Lees et al., 2012), and PANTHER (http://www.pantherdb.org/) (Thomas et al., 2003) databases. Proteins smaller than

20 amino acids were excluded. Through these analyses, we retrieved 1,481 motifs and 13,178 domains from 26,027 genes (Table S9).

Non-coding RNAs, such as microRNAs, rRNAs, and tRNAs, are those that do not encode proteins but perform various other known functions such as participate in the regulation of gene expression, participate in translation of mRNA, and transport amino acids. Using Rfam (http://rfam.xfam.org/) (Griffiths-Jones et al., 2005) database, Blastn (Altschul et al., 1990) was used to identify microRNAs and rRNAs with "E-value 1e-5," whereas tRNAs were identified using tRNAscan-SE (http://trna.ucsc.edu/tRNAscan-SE/) (Lowe & Eddy, 1997). Afterwards, we obtained 6,834 non-coding genes (Table S10).

To acquire the functional information of the genes, we aligned the predicted gene sequences to functional databases, such as non-redundant protein sequence database (NR, https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/) (Marchler-Bauer et al., 2011), Clusters of orthologous groups for eukaryotic complete genomes database (KOG, https://www.hsls.pitt.edu/obrc/index.php) (Tatusov et al., 2001), Gene Ontology database (GO, http://geneontology.org/2) (Dimmer et al., 2012), Kyoto Encyclopedia of Genes and Genomes database (KEGG, https://www.genome.jp/kegg/kegg1.html) (Kanehisa & Goto, 2000), and TrEMBL (Boeckmann et al., 2003) using BLAST with an E-value cutoff of 1e-5. Further, annotation analyses of the gene using KEGG pathway, KOG function, and GO function were performed. Eventually, 24,872 genes (95.27%) were annotated from at least one of these databases (Table S11).

Assembly validation

We used Trinity v2.0.5 (https://github.com/trinityrnaseq/trinityrnaseq/) to *de novo* assemble the transcriptome sequencing data, whereby contigs with lengths greater than 1,000 bp were screened for subsequent analysis. We aligned the contigs to the assembled genomic sequences and evaluated the coverage of the gene region using the BLAST-like Alignment Tool (BLAT http://www.kentinformatics.com) (Kent, 2002) software with default parameters. A total of

25,556 contigs were longer than 1,000 bp, wherein 25,453 of these could be aligned to the genome, accounting for 99.60% of the total contigs.

We evaluated the single base error rate as follows: the sequencing reads were aligned to genomic sequence by BWA v0.7.17-r1188 (https://sourceforge.net/projects/bio-bwa/files/) software. This was followed by calling SNPs using Samtools v1.3 (https://sourceforge.net/projects/samtools/) (Li et al., 2009), Picard tools (https://github.com/broadinstitute/picard) and GATK (https://software.broadinstitute.org/gatk/). We filtered SNPs using the following criteria: read depth (DP) > 10, quality by depth (QD) > 10.0, mapping quality (MQ) > 30.0, phred score of strand bias (FS) < 13.0, HaplotypeScore < 13.0, MQRankSum > -1.96, and ReadPosRankSum > -1.96. Rate of heterozygosity was estimated as the density of heterozygous SNPs for the whole genome. Single base error rate of the genome was estimated using the following formula:

Single base error rate = (Total SNP - Heterozygosity SNP)/genome size)

On evaluating the base error rate of the genome, we found that the number of inconsistent bases was 1,608, accounting for 0.0002% of the total length of genome.

Subsequently, we performed integrity evaluation of the genome sequence using the Merqury software (Rhie et al., 2020) by parameters k=20 and tolerable collision rate: 0.001. The analysis results showed that the genome assembly integrity was 93.33%, consistent with the result of the genome size estimation above. Furtherly, to understand the completeness of gene set, we used CEGMA v2.3 (http://korflab.ucdavis.edu/Datasets/cegma) (Parra et al., 2007) with default parameters to assess the integrity of the genome assembly of *T. rosa* based on 458 core genes and 248 highly conserved genes with default parameters and found 457 (99.78%) high similarity genes (identity 70%) among the 458 core genes (Table S12). Out of the 248 highly conserved sequences, we found 245 (98.79 %) in the genome (Table S12). Subsequently, BUSCO software v1.22 (https://busco.ezlab.org/) (Simao et al., 2015) was used along with the Actinopterygii database (actinopterygii_odb9) to assess the predicted gene set. The genome mode results showed that 98.2% of all 4,584 orthologs were assembled, whereby 95.3% and 2.9%

were completely and partially assembled, respectively (Table S12). This implies a high level of completeness for the *de novo* assembly of the present study.

To check the accuracy of the chromosome-scale assembly of *T. rosa*, we performed a collinearity analysis with *T. tibetana*. MUMmer (http://mummer.sourceforge.net/) was used for aligning entire genomic DNA sequences from the *T. rosa* and *T. tibetana* chromosomes. Circos plot was based on the homologous sequence pairs which length was greater than 2,000 bp. The results showed that the 25 pseudo-chromosomes of *T. rosa* had well matched with the 25 chromosomes of *T. tibetana* (Fig. S3). Therefore, the chromosome-scale assembly of *T. rosa* is accurate.

Effective population size estimation

The reads from 270 bp_1 sequencing library and the reads download from NCBI (SRA ID: SRR8118711) were used to estimate the effective population size of *T. rosa* and *T. tibetana*. Since we were unable to obtain data from additional individuals for efficient population estimation, the two sequencing data which used for their respective genome assembly were used for our analysis. The method is as follows: First, we use bwa (Li et al., 2009) mem to align the sequencing data to their respective genomes with default parameters. Use Samtools (Li et al., 2009) for file conversion and sorting. Use bcftools (Li, 2011) to obtain SNP sites with the parameters “mpileup -C50” and “call -c”. Then, use the vcfutils.pl program with parameters vcf2fq -d 10 -D 100 to convert the VCF file to fq format. Finally, the PSMC software (Nadachowska-Brzyska et al., 2016) was used to estimate the historical effective population size with the following parameters: “fq2psmcfa -q20”, “psmc -N25 -t15 -r5 -p "4+25*2+4+6"" and “psmc_plot.pl -u 2e-09 -g 1”.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410.

doi:10.1016/S0022-2836(05)80360-2

Attwood, T. K., & Beck, M. E. (1994). PRINTS--a protein motif fingerprint database. *Protein Eng, 7*(7), 841-848. doi:10.1093/protein/7.7.841

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res, 19 Suppl*, 2241-2245. doi:10.1093/nar/19.suppl.2241

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., . . . Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res, 31*(1), 365-370. doi:10.1093/nar/gkg095

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res, 33*(Database issue), D212-215. doi:10.1093/nar/gki034

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., . . . Apweiler, R. (2012). The UniProt-GO Annotation database in 2011. *Nucleic Acids Res, 40*(Database issue), D565-570. doi:10.1093/nar/gkr1048

Dolezel, J., & Greilhuber, J. (2010). Nuclear genome size: are we getting closer? *Cytometry A, 77*(7), 635-642. doi:10.1002/cyto.a.20915

Edgar, R. C., & Myers, E. W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics, 21 Suppl 1*, i152-158. doi:10.1093/bioinformatics/bti1003

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., . . . Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res, 42*(Database issue), D222-230. doi:10.1093/nar/gkt1223

Gough, J., & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res, 30*(1), 268-272. doi:10.1093/nar/30.1.268

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., & Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res, 33*(Database issue), D121-124. doi:10.1093/nar/gki081

Haft, D. H., Selengut, J. D., & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res, 31*(1), 371-373. doi:10.1093/nar/gkg128

Han, Y., & Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res, 38*(22), e199. doi:10.1093/nar/gkq862

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240. doi:10.1093/bioinformatics/btu031

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res, 110*(1-4), 462-467. doi:10.1159/000084979

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30. doi:10.1093/nar/28.1.27

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664. doi:10.1101/gr.229202

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., & Orengo, C. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res, 40*(Database issue), D465-471. doi:10.1093/nar/gkr1181

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., . . . Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res, 32*(Database issue), D142-144. doi:10.1093/nar/gkh088

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754-1760. doi: 10.1093/bioinformatics/btp324.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. 2011. Bioinformatics27:2987-2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics, 24*(5), 713-714. doi:10.1093/bioinformatics/btn025

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326*(5950), 289-293. doi:10.1126/science.1181369

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., . . . Bairoch, A. (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res, 37*(Database issue), D471-478. doi:10.1093/nar/gkn661

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., . . . Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint arXiv:1308.2012*.

Liu, S., Ludwig, A., & Peng, Z. (2017). Nine novel microsatellites for the cavefish (*Triplophysa rosa* Chen & Yang, 2005). *J Appl Ichthyol, 33*(1), 119-120. doi:https://doi.org/10.1111/jai.13231

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res, 25*(5), 955-964.

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics, 27*(6), 764-770. doi:10.1093/bioinformatics/btr011

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., . . . Bryant, S. H. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res, 39*(Database issue), D225-229.

doi:10.1093/nar/gkq1189

Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res, 8*(19), 4321-4325. doi:10.1093/nar/8.19.4321

Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. 2016. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Mol Ecol*. 25:1058-1072. doi: 10.1111/mec.13540.

Niu, Y., Zhao, Q., Zhao, H., Ludwig, A., & Peng, Z. (2017). Karyotype and genome size of an endangered cavefish (Triplophysa rosa Chen & Yang, 2005). *J Appl Ichthyol*, *33*(1), 124-126. doi:https://doi.org/10.1111/jai.13209

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics, 23*(9), 1061-1067. doi:10.1093/bioinformatics/btm071

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., & Maddison, D. R. (2020). Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). *G3 (Bethesda), 10*(9), 3047-3060. doi:10.1534/g3.120.401028

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics, 21 Suppl 1*, i351-358. doi:10.1093/bioinformatics/bti1018

Pucker, B. (2019). Mapping-based genome size estimation. *bioRxiv*, 607390. doi:10.1101/607390

Rhie, A., Walenz, B. P., Koren, S., Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*, 21(1), 245

Simao, 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. doi:10.1093/bioinformatics/btv351

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in

genomic sequences. *Curr Protoc Bioinformatics, Chapter 4*, Unit 4 10. doi:10.1002/0471250953.bi0410s25

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., . . . Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res, 29*(1), 22-28. doi:10.1093/nar/29.1.22

Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., . . . Doremieux, O. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res, 31*(1), 334-341. doi:10.1093/nar/gkg115

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., … Schulman A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 8(12), 973-982. doi: 10.1038/nrg2165

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L. S., Natale, D. A., Vinayaka, C. R., . . . Barker, W. C. (2004). PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res, 32*(Database issue), D112-114. doi:10.1093/nar/gkh097

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res, 35*(Web Server issue), W265-268. doi:10.1093/nar/gkm286

Zhao, J., Zhao, K., & Peng, Z. (2014). Development and characterization of eleven microsatellite markers for an endangered cavefish (Triplophysa rosa Chen and Yang, 2005) using 454 sequencing. *J Appl Ichthyol, 30*(5), 1056-1058. doi:https://doi.org/10.1111/jai.12474

Table S1. The source of genome sequences used in this study.

Species	Download links
Anabarilius grahami	https://www.ncbi.nlm.nih.gov/assembly/GCA_003731715.1/
*Astyanax mexicanus cave**	https://www.ncbi.nlm.nih.gov/assembly/GCA_004802775.1
Astyanax mexicanus surface	https://www.ncbi.nlm.nih.gov/assembly/GCF_000372685.2
Bagarius yarrelli	https://www.ncbi.nlm.nih.gov/assembly/GCA_005784505.1/
Chanos chanos	https://www.ncbi.nlm.nih.gov/assembly/GCF_902362185.1
Clupea harengus	https://www.ncbi.nlm.nih.gov/assembly/GCA_900700415.2
Ctenopharyngodon idellus	https://www.ncgr.ac.cn/grasscarp/
Cyprinus carpio#	https://bigd.big.ac.cn/bioproject/browse/PRJCA001408
Danio rerio	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002035.6
Danionella translucida	https://www.ncbi.nlm.nih.gov/assembly/GCA_007224835.1
Denticeps clupeoides	https://www.ncbi.nlm.nih.gov/assembly/GCA_900700345.2
Electrophorus electricus	http://efishgenomics.zoology.msu.edu/?q=node/1
Gadus morhua	https://www.ncbi.nlm.nih.gov/assembly/GCF_902167405.1
Gasterosteus aculeatus	http://hgdownload.soe.ucsc.edu/downloads.html#stickleback
Ictalurus punctatus	https://www.ncbi.nlm.nih.gov/assembly/GCA_001660625.2
Lepisosteus oculatus	https://www.ncbi.nlm.nih.gov/assembly/GCF_000242695.1/
Megalobrama amblycephala	https://www.ncbi.nlm.nih.gov/assembly/GCA_009869865.1/
Oreochromis niloticus	https://www.ncbi.nlm.nih.gov/assembly/GCF_001858045.2
Oryzias latipes	https://www.ncbi.nlm.nih.gov/assembly/GCF_000313675.1
Pangasianodon hypophthalmus	https://www.ncbi.nlm.nih.gov/assembly/GCF_003671635.1
Paramormyrops kingsleyae	https://www.ncbi.nlm.nih.gov/assembly/GCF_002872115.1/
Perca flavescens	https://www.ncbi.nlm.nih.gov/assembly/GCF_004354835.1
Pygocentrus nattereri	https://www.ncbi.nlm.nih.gov/assembly/GCF_001682695.1
Scleropages formosus	https://www.ncbi.nlm.nih.gov/assembly/GCF_900964775.1
Sinocyclocheilus anshuiensis#*	https://www.ncbi.nlm.nih.gov/assembly/GCF_001515605.1/
Sinocyclocheilus grahami#	https://www.ncbi.nlm.nih.gov/assembly/GCF_001515645.1/
Tachysurus fulvidraco	https://www.ncbi.nlm.nih.gov/assembly/GCF_003724035.1/
Takifugu rubripes	https://www.ncbi.nlm.nih.gov/assembly/GCF_000180615.1
Triplophysa tibetana	https://www.ncbi.nlm.nih.gov/assembly/GCA_008369825.1/

Note: *: cavefish; #: allotetraploid species, its genomic gene set is divided into two sub-genomic parts with A and B.

Table S2. The results of Illumina sequencing (clean data).

Library	Data (Gb)	Depth (×)	Q20 (%)	Q30 (%)
270 bp_1	43.04	63.01	95.41	87.46
270 bp_2	43.05	63.02	97.76	93.07
500 bp	12.01	17.58	96.38	88.95
3 K_1	5.30	7.78	93.41	87.25
3 K_2	5.11	7.50	94.01	88.03
4 K_1	4.22	6.20	94.01	88.06
4 K_2	4.80	7.05	94.13	88.21
8 K_1	4.12	6.05	93.97	87.99
8 K_2	5.47	8.03	93.60	87.95
10 K_1	2.67	3.92	94.01	88.02
10 K_2	2.59	3.81	93.57	87.84
15 K	4.25	6.24	93.95	88.64
17 K	4.80	7.04	94.25	89.04
Total	141.43	207.23	--	--

Note: **Q20 (%)**: The ratio of bases which quality over 20; **Q30 (%):** The ratio of bases which quality above 30.

Table S3. The statistics of Pacbio sequencing (row data) results.

Type	Filter	Read Bases (bp)	Read Num	Read N50	Mean Read Length(bp)	Read Quality
Polymerase reads	Pre-Filter	13,138,204,084	1,352,628	26,705	9,713	0.413
	Post-Filter	11,628,811,925	646,245	26,821	17,994	0.826

Note: **Filter**: **Pre-Filter** before filtering; **Post-Filter** after filter low-quality, short reads (less than 500 bp); **Read base (bp)**: Total bases amount of sequencing data; **Read Quality**: The average quality of the sequenced data.

Table S4. The statistics of Pacbio clean reads.

Length	Total_num	Total_length (bp)	Aver_length (bp)
0~2000	115,960	160,405,627	1383.28
2000~4000	153,893	447,998,092	2911.10
4000~6000	142,433	712,758,863	5004.17
6000~8000	137,513	960,979,201	6988.28
8000~10000	127,277	1,144,444,449	8991.76
10000~12000	136,903	1,509,532,487	11026.29
12000~14000	138,906	1,800,477,714	12961.84
14000~16000	94,656	1,412,186,863	14919.15
16000~18000	56,835	961,179,616	16911.76
18000~	105,273	2,486,275,163	23617.41
Total	1,209,649	11,596,238,075	9,586.45

Table S5. Summary of pairs-end reads generated by Hi-C sequencing.

Type	Number	Ratio (%)
Unique Mapped Read Pairs	241,100,314	100
Valid Interaction Pairs	93,078,312	38.61
Dangling End Pairs	81,550,373	33.82
Re-ligation Pairs	46,970,151	19.48
Self-cycle Pairs	1,113,062	0.46
Dumped Pairs	18,388,416	7.63

Table S6. Summary of assembled 25 pseudo-chromosomes of *T. rosa*.

Chromosomes	Number of scaffolds	Length (Mbp)
Lachesis Group0	74	41.75
Lachesis Group1	79	38.13
Lachesis Group2	33	18.71
Lachesis Group3	63	34.92
Lachesis Group4	58	31.87
Lachesis Group5	53	29.61
Lachesis Group6	56	29.43
Lachesis Group7	56	28.87
Lachesis Group8	59	29.09
Lachesis Group9	61	28.44
Lachesis Group10	68	29.24
Lachesis Group11	50	26.82
Lachesis Group12	52	26.96
Lachesis Group13	57	25.61
Lachesis Group14	53	26.29
Lachesis Group15	46	24.68
Lachesis Group16	40	25.29
Lachesis Group17	46	25.00
Lachesis Group18	45	23.64
Lachesis Group19	38	22.84
Lachesis Group20	44	23.24
Lachesis Group21	41	22.43
Lachesis Group22	46	22.30
Lachesis Group23	41	17.86
Lachesis Group24	57	20.33
Total unlinked (Ratio %)	1034 (53.24%)	53.24 (7.70%)
Total linked (Ratio %)	908 (46.76%)	638.47 (92.30%)
Total	1059	691.71

Table S7. Genomic assembly information statistics.

Type	Number	Length(bp)	N50(bp)	N90(bp)	max (bp)	Gap length(bp)
scaffold	1,059	691.71M	24.84M	16.00M	40.13M	9,92M
contig	10,700	681.79M	201K	35K	1.92M	--

Table S8. The statistics of gene prediction.

Method	Software	Species	Gene number
Ab initio	Genscan	---	24,377
	Augustus	---	31,155
	GlimmerHMM	---	48,855
	GeneID	---	27,669
	SNAP	---	53,999
	FGENESH	---	34,963
Homology-based	GeMoMa	*Oryzias latipes*	27,073
		Danio rerio	32,154
		Gasterosteus aculeatus	9,392
		Xiphophorus maculatus	23,810
		Tetraodon nigroviridis	8,373
		Poecilia formosa	31,003
		Gadus morhua	3,445
		Lepisosteus oculatus	22,935
		Oreochromis niloticus	32,842
		Takifugu rubripes	23,291
		Astyanax mexicanus	27,027
		Homo sapiens	25,451
EST/Unigene	PASA	---	19,186
Integration	EVM	---	26,027

Table S9. The statistics of motif prediction.

Type	Count
Motif	1,481
Domain	13,178

Table S10. The statistics of non-coding genes.

RNA classification	Number	Family
miRNA	368	118
rRNA	378	5
tRNA	6,088	25

Table S11. The statistics of gene annotation.

Annotation database	Annotated number	Percentage (%)
GO	13,640	52.25
KEGG	11,999	45.96
KOG	16,536	63.34
Swissprot	16,556	63.42
TrEMBL	24,628	94.33
NR	24,831	95.11
All Annotated	24,872	95.27

Table S12. Details of accuracy and completeness of genome assembly.

Transcripts alignment		
Total Transcripts (> 1kbp)	25,556	
Hitted transcripts	25,453	
Proportion of Hitted transcripts(%)	99.60	
CEGMA		
Total number of reference genes	457	
Number of completely assembled CEGs	456	
Proportion of completely assembled CEGs (%)	99.78	
Total number of reference genes	245	
Number of completely assembled CEGs	244	
Proportion of completely assembled CEGs (%)	98.79	
BUSCO (protein mode)	**Number**	**Proportion (%)**
All orthologues used	4584	100
Complete and fragmented orthologues	4506	98.2
Complete orthologues	4371	95.3
Fragmented orthologues	135	2.9
Missing orthologues	78	1.8

Table S13. The number of genes where (dN/dS) was higher or lower in cavefish compared to surface fish.

Genomes	Higher	Lower	All orthologous genes	Significance*
Astyanax mexicanus	7604	5570	13328	$p < 0.0001$
Sinocyclocheilus A	4636	3930	8566	$p < 0.0001$
Sinocyclocheilus B	5525	4690	10215	$p < 0.0001$
Triplophysa	7813	6605	14623	$p < 0.0001$

Note: The statistics are based on the gene pairs that exist in both cavefish and their closed relative surface fish at the same time.

*Fisher's exact test

Table S14. The number of genes under significantly relaxed selection and intensified selection.

Genomes	Surface fish		Cavefish		Significance*
	Relaxed	Intensified	Relaxed	Intensified	
Astyanax mexicanus	25	71	113	41	$p < 0.0001$
Sinocyclocheilus A	14	12	19	13	$p = 0.79$
Sinocyclocheilus B	16	17	18	18	$p = 1$
Triplophysa	27	21	36	25	$p = 0.85$

Relaxed: $k<1$, FDR<0.05; Intensified: $k>1$, FDR<0.05; FDR was estimate by Benjamini-Hochberg correction. * Fisher's exact test.

Table S15. Pseudogenes in *Triplophysa rosa* and *Triplophysa tibetana*.

Gene set	Species	Genes' symbol
Vision	*Triplophysa rosa*	Cryba1l1, gc3, gcap7, grk7a, Opn4x2, Opn6b, pde6b, Tmt2a, Tmt2b, Va2, Parapinopsin-1, Parietopsin
	Triplophysa tibetana	gc3, opn4m1, Opn6b, Opn9, Tmt2b
Clock	*Triplophysa rosa*	cry2a, nfil3, nr1d1
	Triplophysa tibetana	cry2a, nfil3, per1a
Pigmentation	*Triplophysa rosa*	ap3b1, atrn, gfpt1, lyst, pax3b, pcbd2, rab38a, trpm7
	Triplophysa tibetana	gfpt1, gpr143, hps3, lyst, nf1b, rab38a, shroom2a, slc45a2, trpm7, wnt3a

Supplementary Figures

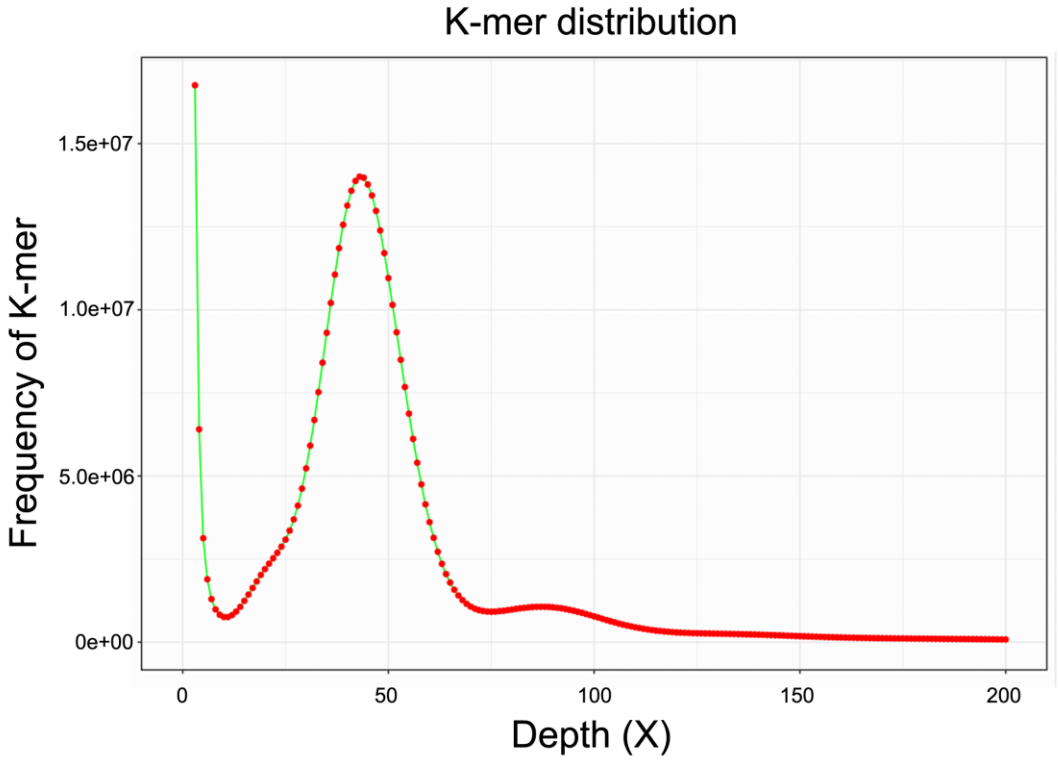

Figure S1. Distribution of 17-mer frequency in *Triplophysa rosa* genome. The depth at the peak of the frequency distribution is 43 ×that is used for genome size estimation. Another peak with depth 21 ×indicates low heterozygous rate in *T. rosa*. ×: the depth of coverage.

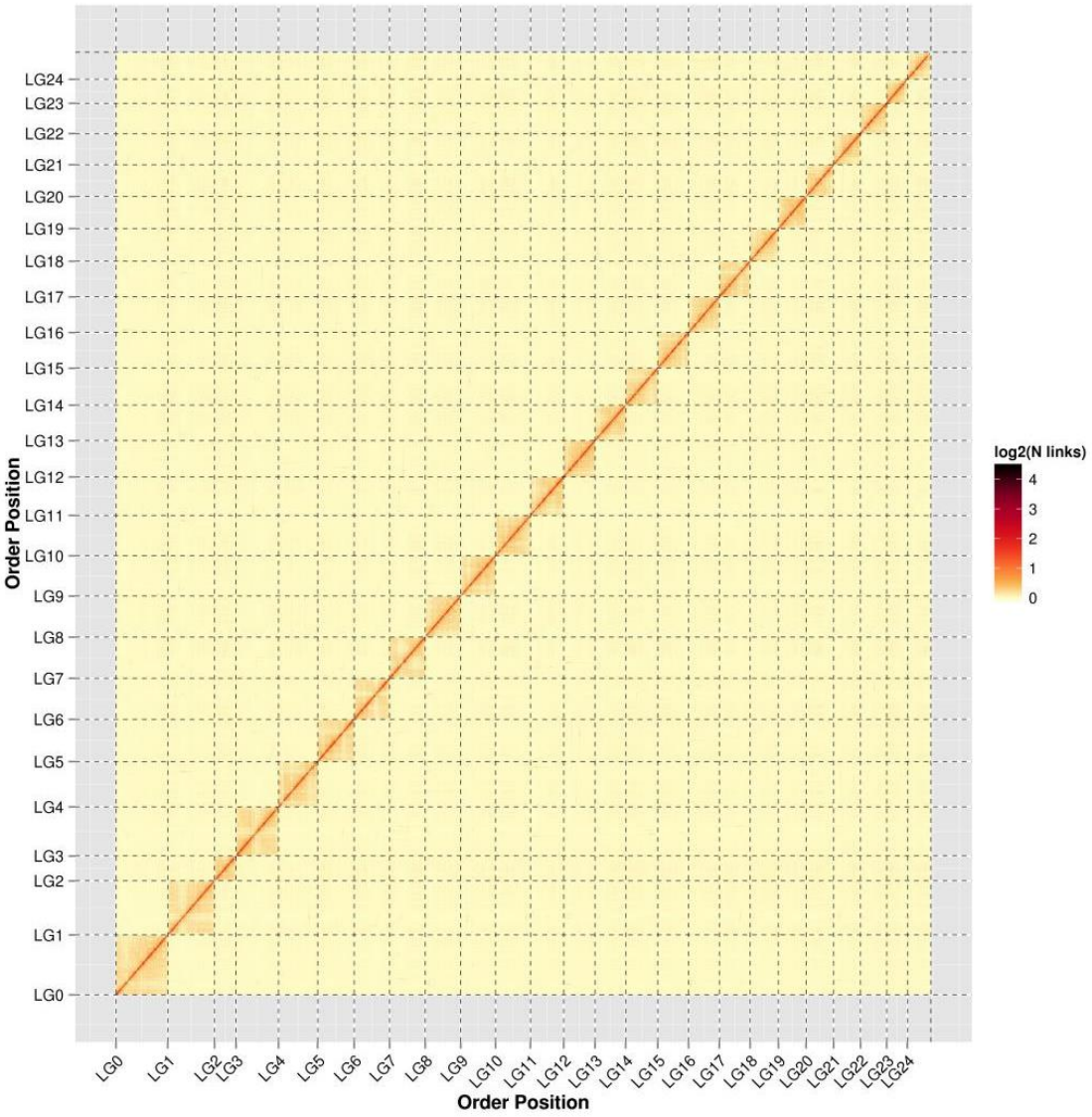

Figure S2. Hi-C interaction relationships between chromosome regions across the genome. The color bar indicates contact density from red (high) to white (low) (bin length: 100 Kb).

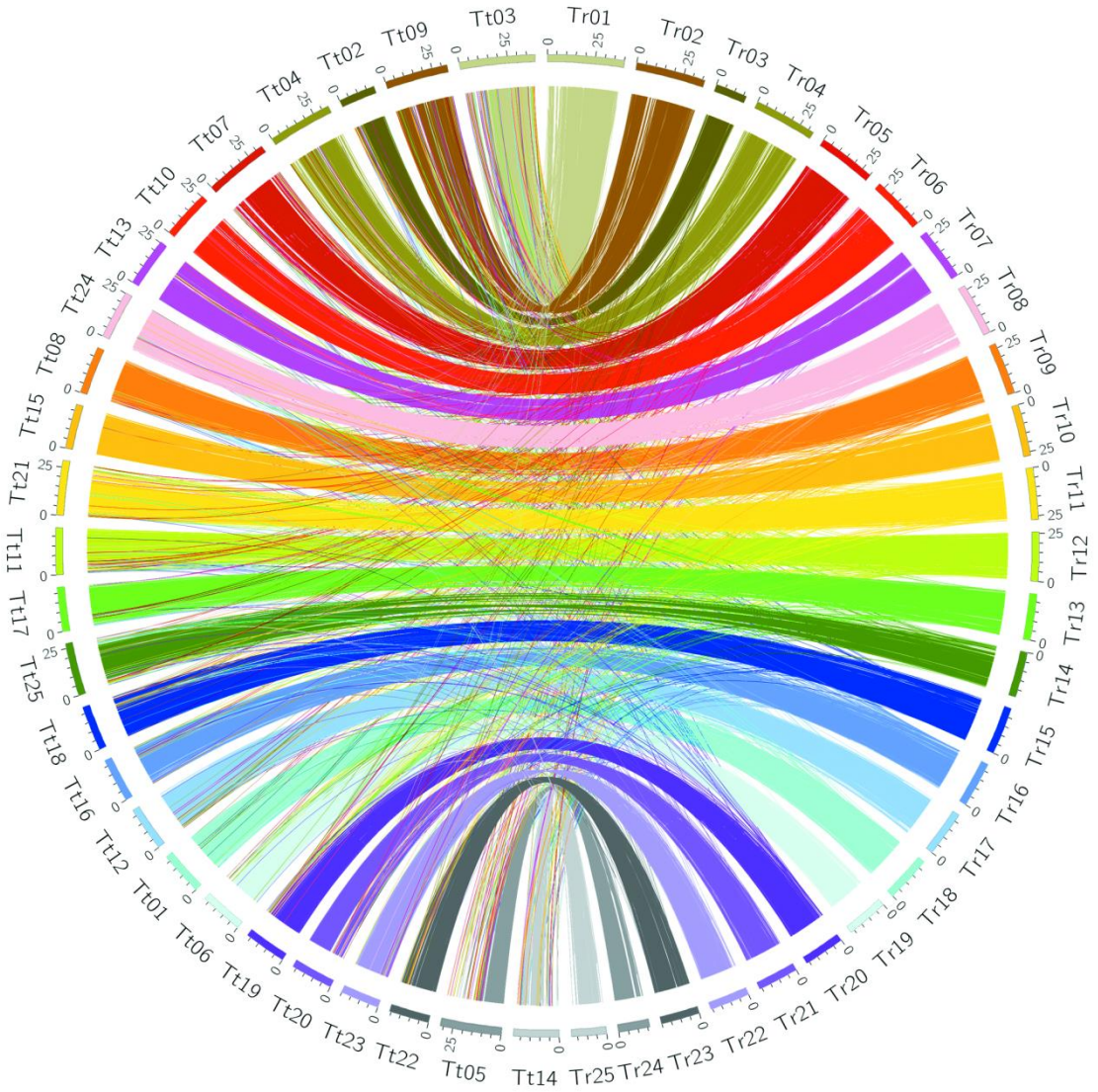

Figure S3. Genomic synteny of *Triplophysa rosa* (Tr) and *T. tibetana* (Tt). Each colored line represents a best match between two species.

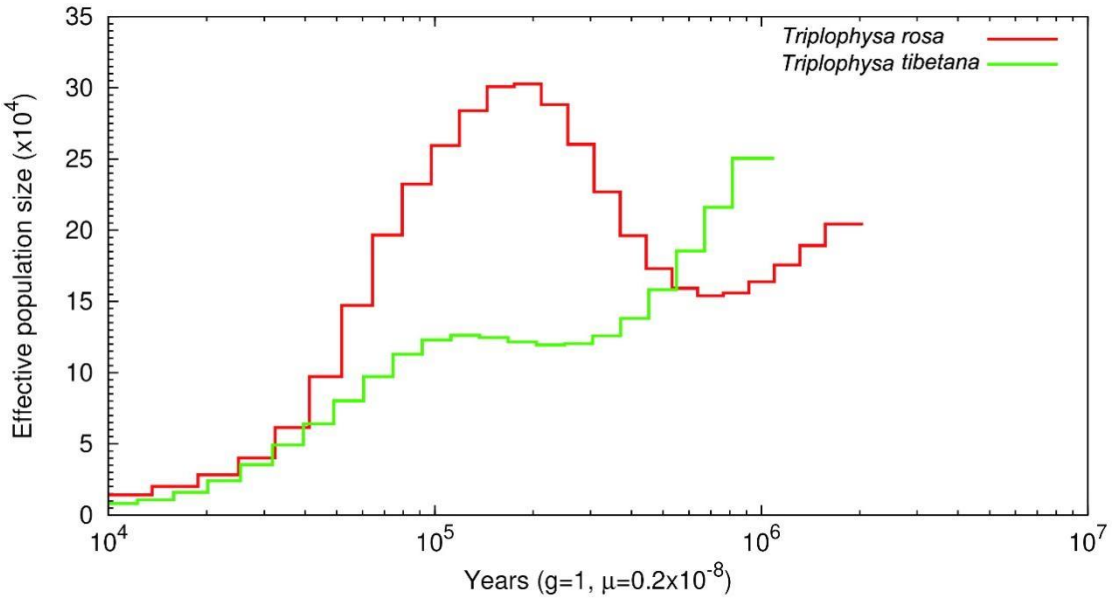

Figure S4. Estimation of the effective population size of *Triplophysa rosa* and *Triplophysa tibetana*.

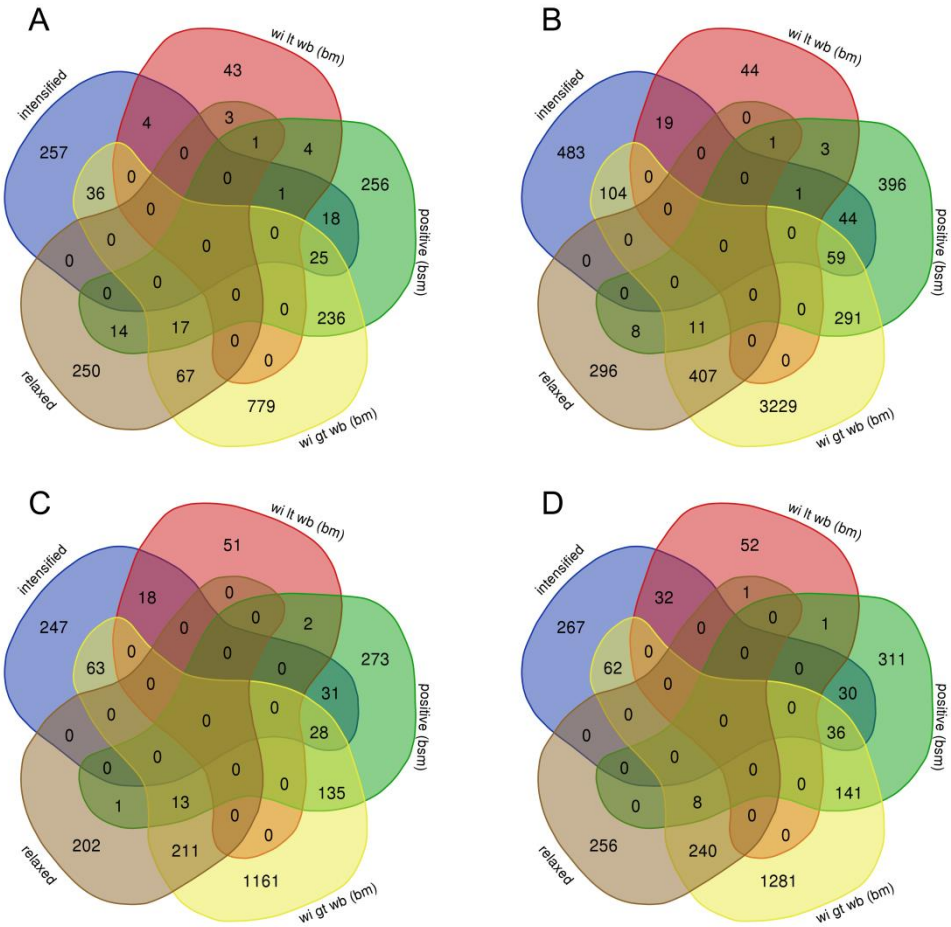

Figure S5. Venn diagram of results from Codeml and RELAX analysis. (A) *Astyanax mexicanus* cavefish, (B) *Triplophysa rosa*, (C) *S. anshuiensis* A subgenome, and (D) *S. anshuiensis* B subgenome. wi: ω in the target species; wb: background ω; lt: less than; gt: great than. bm: branch model; bsm: branch site model.

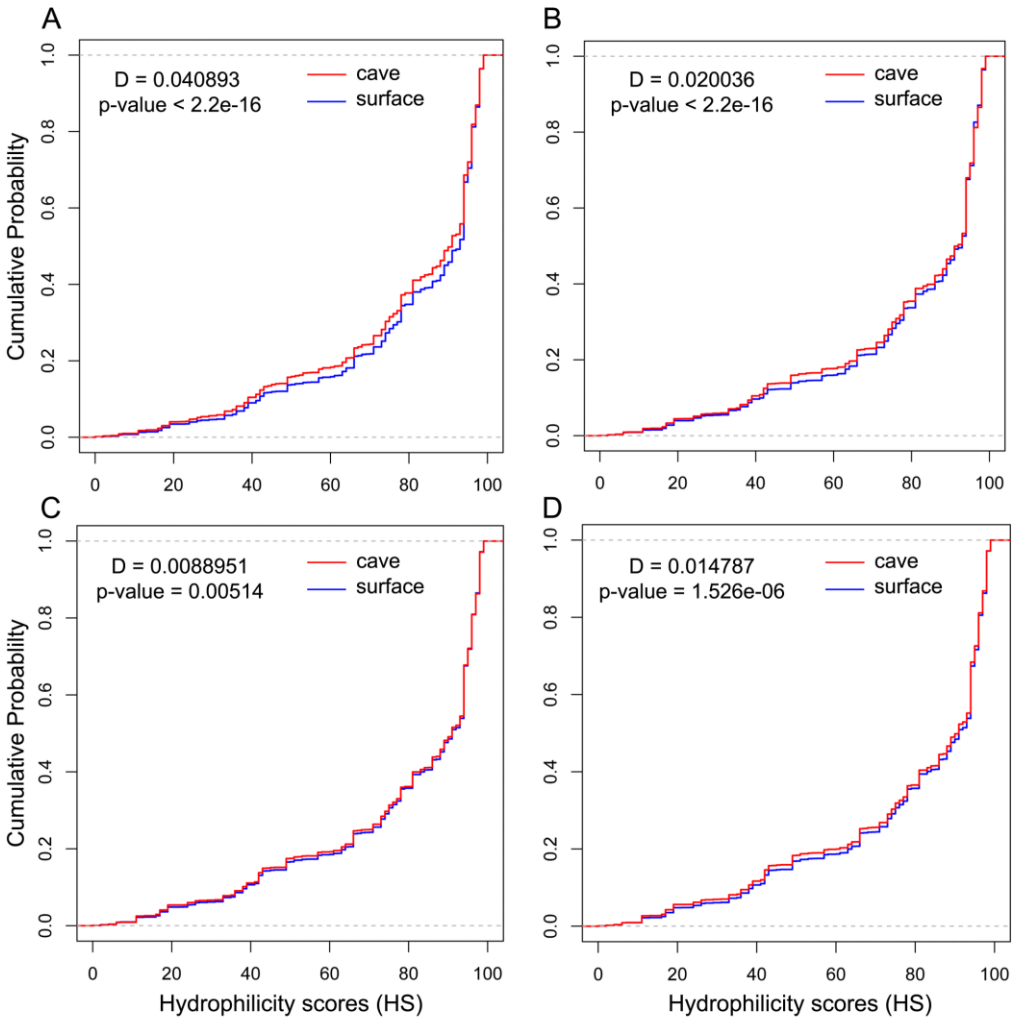

Figure S6. The empirical cumulative density plots of HS scores from cavefish and surface fish. p-value was calculated by using Kolmogorov-Smirnov test. D: Maximum vertical difference between two cumulative distribution curves. The cavefish for each comparison are: (A) *Astyanax mexicanus* cavefish, (B) *Triplophysa rosa*, (C) *S. anshuiensis* A subgenome, and (D) *S. anshuiensis* B subgenome.

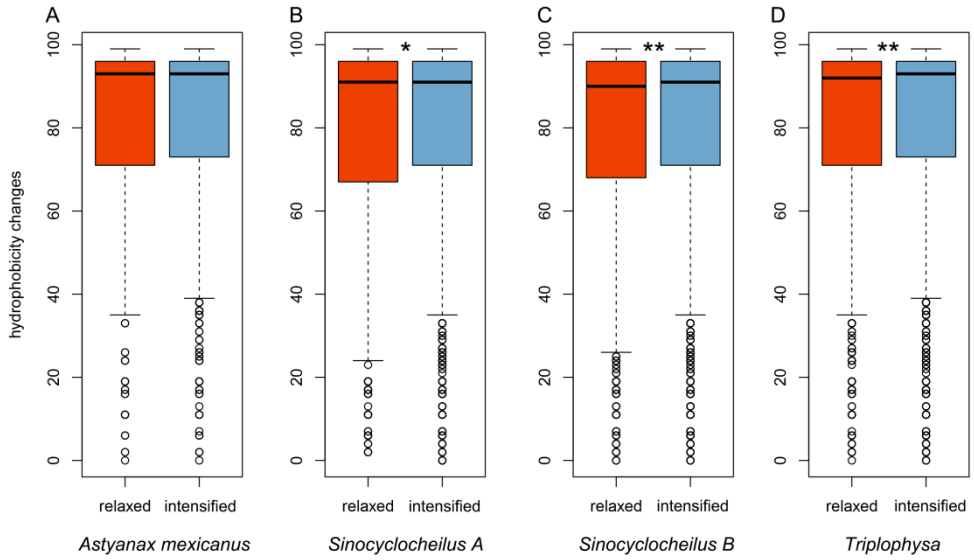

Figure S7. Hydrophilicity scores (HSs) at intensified selection and relaxed selection genes. The HS at the intensified selection (blue) and relaxed selection (red) genes are shown as boxplots (significant differences: *p < 0.05; **p < 0.01; generalized linear mixed model). The horizontal line in the center of the box chart represents the median value of the HS. (A) *Astyanax mexicanus*. (B) *Sinocyclocheilus* A subgenome. (C) *Sinocyclocheilus* B subgenome. (D) *Triplophysa*. A lower score indicates a more dissimilar hydrophobicity.

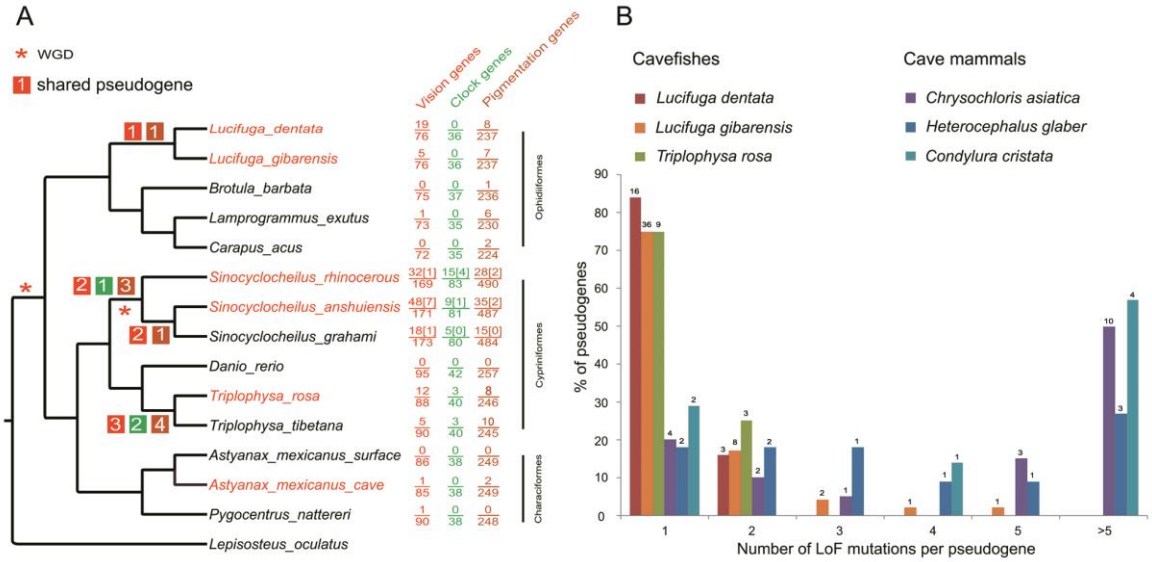

Figure S8. Pseudogenes in cavefish. (A) The tree topology is based on Policarpo et al, (2021) and Figure 1A. For each gene set, the number of pseudogenes found and the number of genes examined in a species are given to the right of the species name. The number of pairs of pseudogenes that are ohnologs is shown in square brackets for tetraploid species. (B) Distributions of the number of LoF mutations per vision pseudogene in cavefishes and cave mammals. The number of pseudogenes is given above the bar.

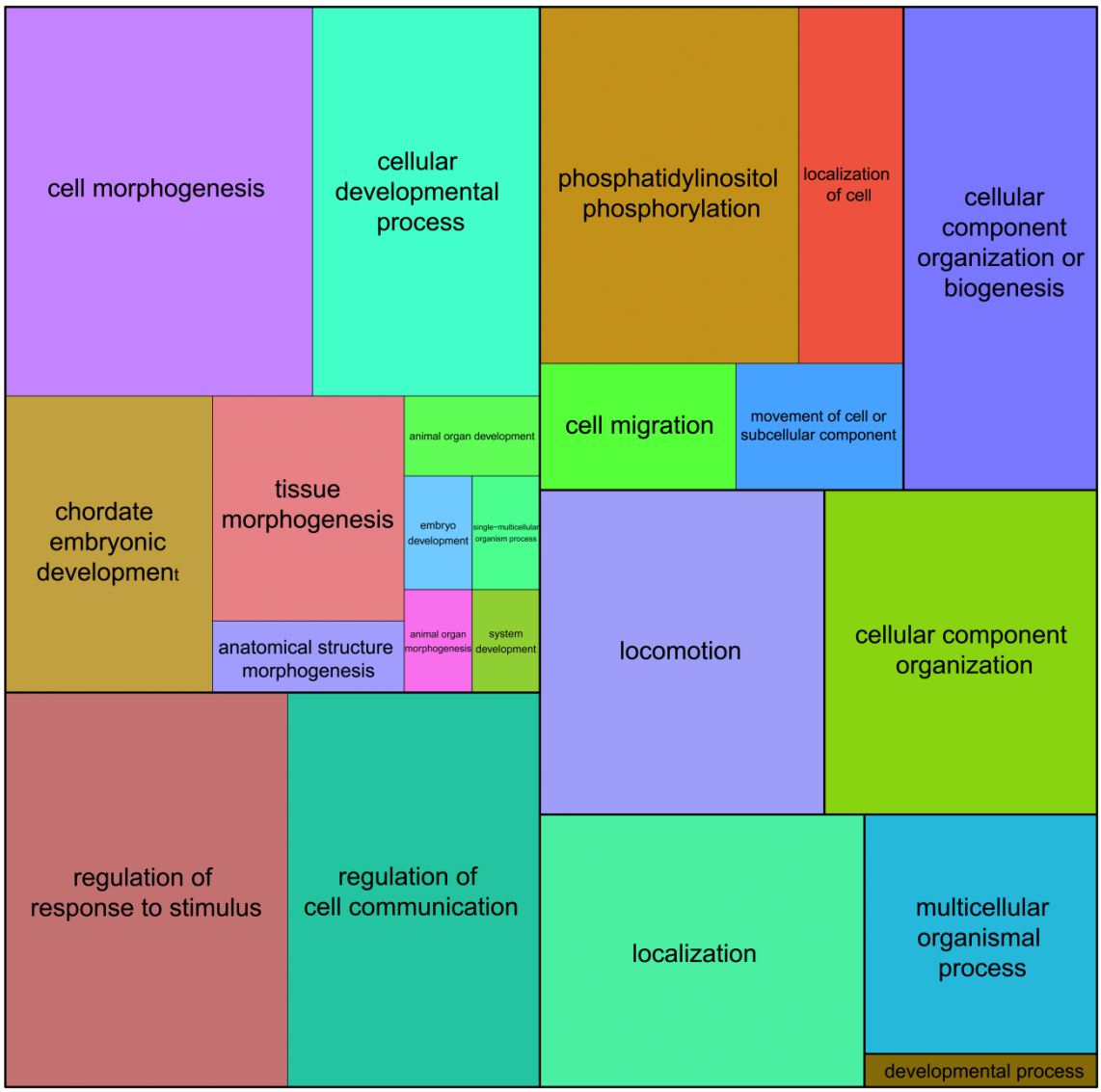

Figure S9. REVIGO TreeMap for GO biological processes categories present in the genes under positive selection in the *Astyanax mexicanus* cavefish. Rectangle size reflects semantic uniqueness of the respective GO term that measures the degree to which the term is an outlier when compared semantically to the list of terms present in the zebrafish.

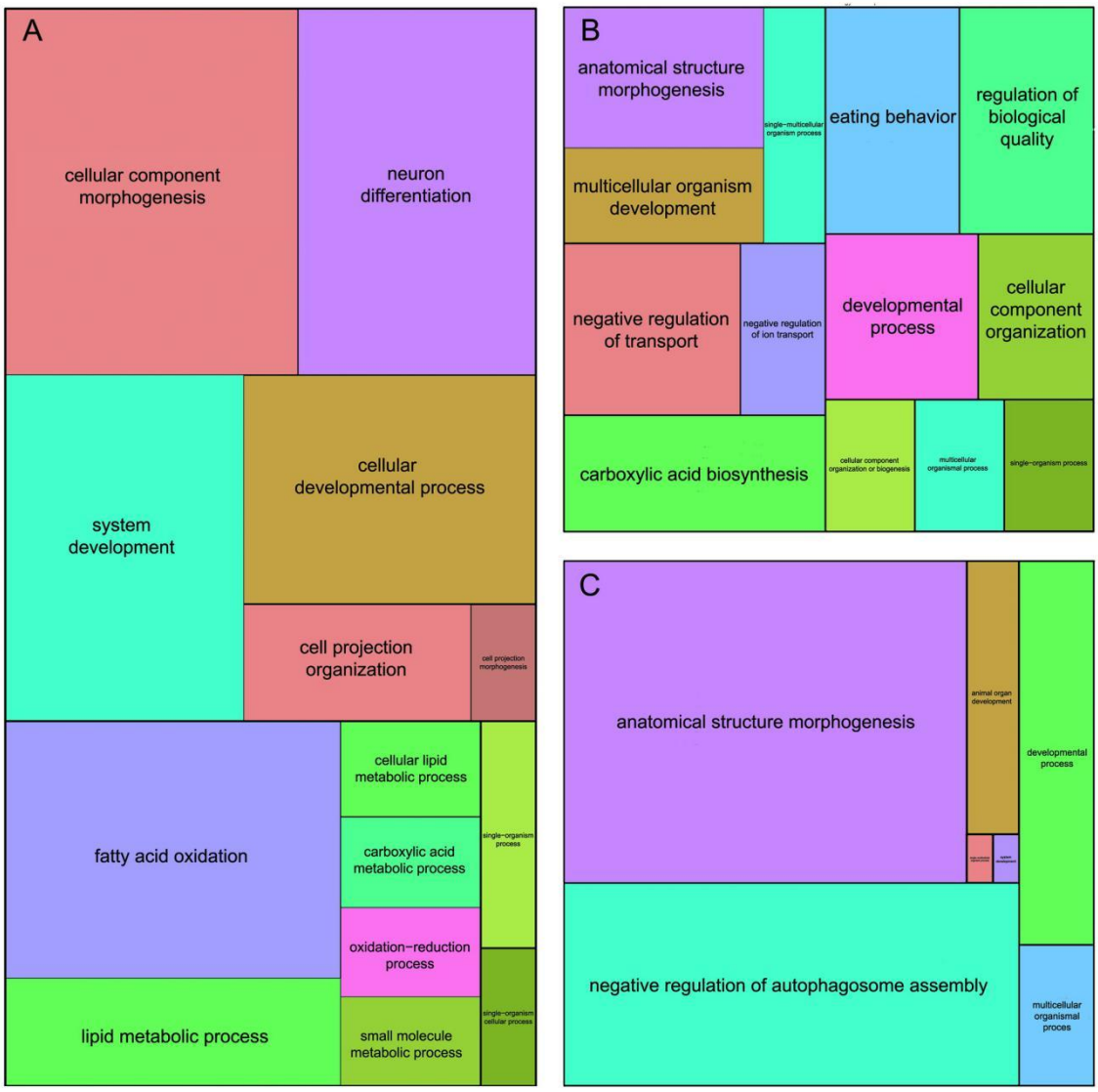

Figure S10. GO biological process categories of the genes under relaxed selection summarized and visualized as a REVIGO TreeMap: (A) *Triplophysa*. (B) *Astyanax mexicanus* cavefish. (C) *Sinocyclocheilus* B subgenome. Each rectangle or color represents a cluster of related GO categories, with a single category chosen by REVIGO as the cluster representative. The rectangle size reflects the negative average similarity of a category to all other categories.