








## RESEARCH ARTICLE

# Whole genome sequencing of a snailfish from the Yap Trench (~7,000 m) clarifies the molecular mechanisms underlying adaptation to the deep sea

Yinnan Mu<sup>1</sup> , Chao Bian<sup>2</sup> , Ruoyu Liu<sup>1</sup> , Yuguang Wang<sup>3</sup>, Guangming Shao<sup>1</sup>, Jia Li<sup>2</sup> , Ying Qiu<sup>2</sup> , Tianliang He<sup>1</sup> , Wanru Li<sup>1</sup>, Jingqun Ao<sup>3</sup>, Qiong Shi<sup>2\*</sup> , Xinhua Chen<sup>1,4\*</sup> 

**1** Key Laboratory of Marine Biotechnology of Fujian Province, Institute of Oceanology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China, **2** Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Economic Animals, BGI Academy of Marine Sciences, BGI Marine, Shenzhen, Guangdong, China, **3** Key Laboratory of Marine Biogenetic Resources, Third Institute of Oceanography, Ministry of Natural Resources, Xiamen, Fujian, China, **4** Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, Guangdong, China

 These authors contributed equally to this work.

\* [shiqiong@genomics.cn](mailto:shiqiong@genomics.cn) (QS); [chenxinhua@tio.org.cn](mailto:chenxinhua@tio.org.cn) (XC)



## OPEN ACCESS

**Citation:** Mu Y, Bian C, Liu R, Wang Y, Shao G, Li J, et al. (2021) Whole genome sequencing of a snailfish from the Yap Trench (~7,000 m) clarifies the molecular mechanisms underlying adaptation to the deep sea. *PLoS Genet* 17(5): e1009530. <https://doi.org/10.1371/journal.pgen.1009530>

**Editor:** Gregory P. Copenhaver, The University of North Carolina at Chapel Hill, UNITED STATES

**Received:** September 23, 2020

**Accepted:** April 5, 2021

**Published:** May 13, 2021

**Copyright:** © 2021 Mu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The data underlying the results presented in the study are available from NCBI with the accession numbers of PRJNA512070 and PRJNA660667, and CNSA with the accession number of CNP0000315.

**Funding:** This work was supported by grants from National Key R&D Program of China (2018YFD0900602; <https://service.most.gov.cn>) to Y.M., and National Program on the Key Basic Research Project (2015CB755903; <https://service.most.gov.cn>), China Agriculture Research System

## Abstract

Hadal environments (depths below 6,000 m) are characterized by extremely high hydrostatic pressures, low temperatures, a scarce food supply, and little light. The evolutionary adaptations that allow vertebrates to survive in this extreme environment are poorly understood. Here, we constructed a high-quality reference genome for Yap hadal snailfish (YHS), which was captured at a depth of ~7,000 m in the Yap Trench. The final YHS genome assembly was 731.75 Mb, with a contig N50 of 0.75 Mb and a scaffold N50 of 1.26 Mb. We predicted 24,329 protein-coding genes in the YHS genome, and 24,265 of these genes were successfully functionally annotated. Phylogenetic analyses suggested that YHS diverged from a Mariana Trench snailfish approximately 0.92 million years ago. Many genes associated with DNA repair show evidence of positive selection and have expanded copy numbers in the YHS genome, possibly helping to maintain the integrity of DNA under increased hydrostatic pressure. The levels of trimethylamine N-oxide (TMAO), a potent protein stabilizer, are much higher in the muscles of YHS than in those of shallow-water fish. This difference is perhaps due to the five copies of the TMAO-generating enzyme flavin-containing monooxygenase-3 gene (*fmo3*) in the YHS genome and the abundance of trimethylamine (TMA)-generating bacteria in the YHS gut. Thus, the high TMAO content might help YHS adapt to high hydrostatic pressure by improving protein stability. Additionally, the evolutionary features of the YHS genes encoding sensory-related proteins are consistent with the scarce food supply and darkness in the hadal environments. These results clarify the molecular mechanisms underlying the adaptation of hadal organisms to the deep-sea environment and provide valuable genomic resources for in-depth investigations of hadal biology.

(CARS-47; <http://www.cars.ren>), China Ocean Mineral Resources R&D Association Program (DY135-B2-16; <http://www.comra.org>), and Special Fund for Marine Economic Development of Fujian Province (FJHJF-L-2019-2; <https://hyyj.fujian.gov.cn/>) to X.C. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Author summary

Hadal environments (depths below 6,000 m) are characterized by extremely high hydrostatic pressures, low temperatures, a scarce food supply, and little light. Fish are the only vertebrates inhabiting the hadal zone, and hadal snailfishes have been found in at least five geographically separated marine trenches. However, the genetic mechanisms that allow vertebrates to live in such extreme conditions are not well understood. Here, we constructed a high-quality reference genome for Yap hadal snailfish (YHS) captured at a depth of ~7,000 m in the Yap Trench, using long reads obtained by Pacific Biosciences Sequel sequencing. Comparative genomic analyses revealed that many genes associated with DNA repair show evidence of positive selection and have expanded copy numbers in the YHS genome, which potentially reflect the difficulty of maintaining DNA integrity under high hydrostatic pressure. Moreover, the five copies of the trimethylamine N-oxide (TMAO)-generating enzyme flavin-containing monooxygenase-3 gene (*fmo3*) and the abundance of trimethylamine (TMA)-generating bacteria in the YHS gut could provide enough TMAO to improve protein stability under hadal conditions. In addition, characteristics of the YHS sensory system genes were consistent with the scarce food supply and darkness in the hadal zone. Our results provide new insights into the molecular mechanisms underlying the adaptation of hadal organisms to the deep-sea environment and valuable genomic resources that will help further clarify hadal adaptations.

## Introduction

The hadal zone (6,000–11,000 m deep) is composed mainly of deep trenches, and is usually characterized by extremely high hydrostatic pressures, low temperatures, a scarce food supply, a lack of light, and geographical isolation [1,2]. Hydrostatic pressure is the most conspicuous environmental gradient in the deep sea, increasing by approximately 0.1 megapascals (MPa) with every ten meters of depth, and reaching ~100 MPa in the deepest hadal zone [3]. The temperatures in the hadal zone range from 1 to 2.5°C, and the light intensity there is too low to sustain photosynthetic production [4]. The food resources in the hadal zone are mainly supplemented by surface-derived carrion falls, which implies that the food in the hadal zone is much more limited than that in shallower regions [1,5]. These harsh living conditions form a unique deep ocean trench ecosystem with an endemic faunal community distinct from those in surrounding deep-sea environments [6].

Environmental extremes in the hadal zone influence the physiological and biochemical processes of marine organisms [7]. Comparative studies have revealed that the pressure sensitivities of some structural proteins, membrane-based systems, and enzymes differ markedly between deep- and shallow-living species [2,3,8]. For example, the volume associated with  $\alpha$ -actin polymerization in two deep-sea fish species (*Coryphaenoides yaquinae* and *C. armatus*) is markedly smaller than that in congeneric shallow-living fishes (*C. cinereus* and *C. acrolepis*) [2]. Amino acid substitutions also occur during adaptation to hydrostatic pressures, and these substitutions affect protein-protein interactions or ligand binding [9]. Substitutions in lactate dehydrogenase in deep-sea fishes help the enzyme better tolerate high hydrostatic pressures [7]. Another potential mechanism of pressure adaptation involves several small organic solutes, which are referred to as “pyrolites” [10,11]. Pyrolites enhance the structural stability and binding ability of proteins by altering water molecule structure to reduce its tendency to pressurize [12]. Trimethylamine N-oxide (TMAO), a potent protein stabilizer commonly found in

the muscles of marine fish species, can counteract the effects of hydrostatic pressure on enzyme kinetics and protein stability [10]. In deep-sea teleosts, a striking correlation exists between capture depth and TMAO content [13,14]. Moreover, at low temperatures, the DNA and RNA strand structures tend to tighten, which hinders interaction with enzymes involved in DNA replication, transcription, and translation and thus disrupts the transcription and translation processes [9].

The Liparidae (the snailfish; Scorpaeniformes), which includes more than 400 species, is probably the most bathymetrically and geographically widespread family of marine fish [15]. Liparidae has the widest depth range of all marine fish species, with habitats ranging from intertidal to depths greater than 8,000 m [16]. Hadal snailfishes have been found in at least five geographically separated marine trenches. The deepest snailfish recorded to date is *Abyssobrotula galathea*, which was retrieved from the Puerto Rico Trench at a depth of 8,370 m in 1970 [17]. Other hadal snailfishes have been found in the Kermadec, Mariana, Kuril-Kamchatka, and Japan trenches at depths of ~6,660–7,966 m [15,18]. Snailfishes from various hadal trenches possess many similar characteristics, such as transparent skin and peritoneum, a pinkish-white body, internal organs that are visible through the skin and thin abdominal walls [15]. However, the molecular mechanisms that allow snailfishes to survive in extreme hadal environments are poorly understood, mainly due to the limited availability of genetic data. To date, only the genome assembly of the Mariana hadal snailfish (*Pseudoliparis swirei*) captured at a depth of 7,034 m in the Mariana Trench has been reported, exploring the adaptive characteristics related to the hadal environment [19].

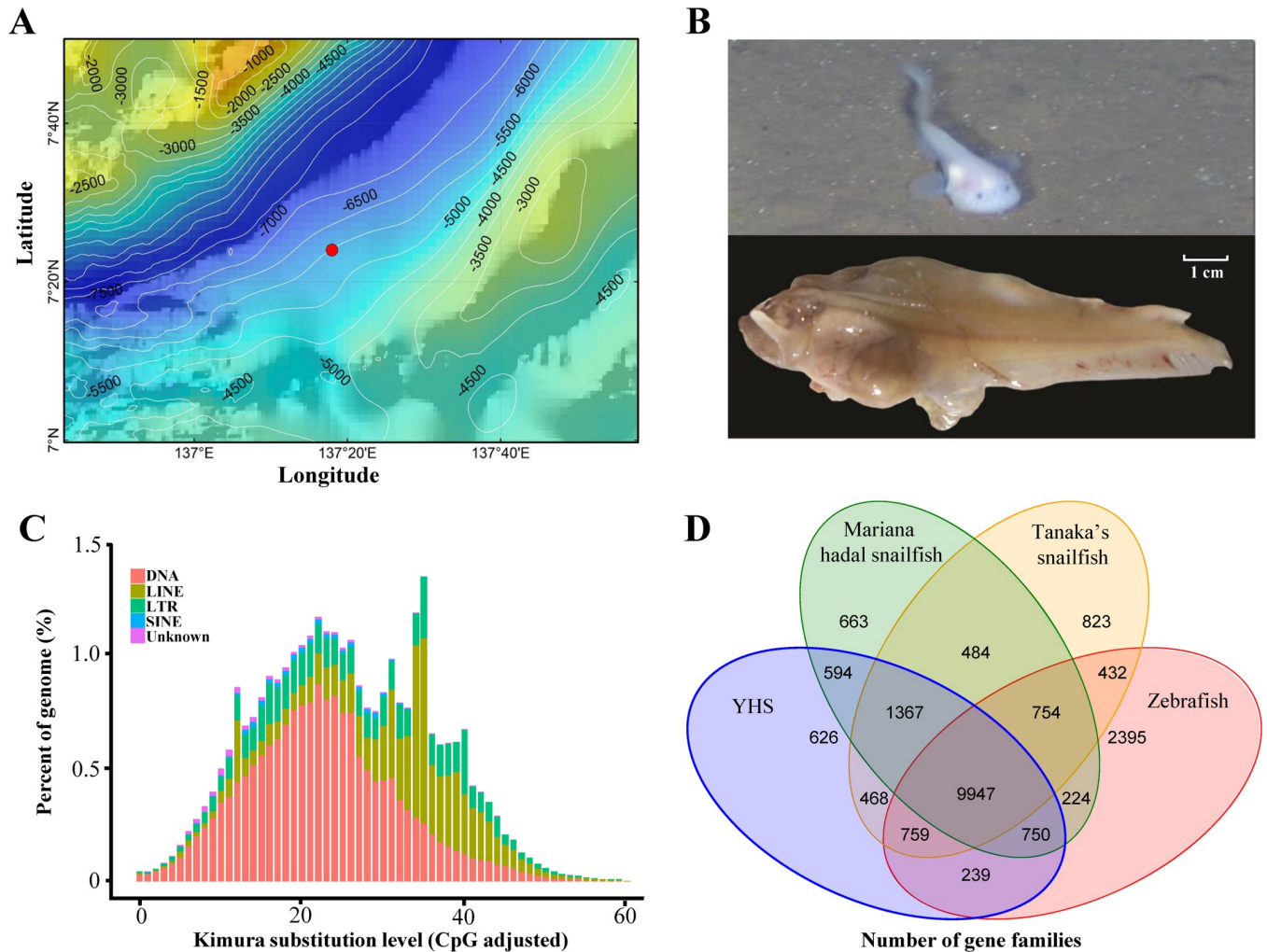
In this study, we sequenced and assembled a reference genome for a species of hadal snailfish, collected at a depth of 6,903 m in the Yap Trench (tentatively named Yap hadal snailfish; YHS) using advanced single-molecule real-time (SMRT) sequencing. Comparative genomic analyses among YHS, Mariana hadal snailfish, and other sequenced shallow-water fish species were performed to identify genetic changes associated with adaptations to the hadal environment. The gut microbiota of YHS was also analyzed to explore its correlation with deep-sea adaptation. Our results provide new insights into the molecular mechanisms underlying adaptation of hadal organisms to the deep-sea environment and valuable genomic resources for further in-depth investigations of hadal biology.

## Results

### Characterization of the hadal snailfish from the Yap Trench and its genome sequencing

Two hadal snailfishes were collected at a depth of 6,903 m in the Yap Trench (Western Pacific Ocean; 137.3°E, 7.4°N; Fig 1A) during an expedition of the Chinese manned submersible Jiaolong (National Deep Sea Center, China). Both individuals had enlarged stomachs and livers, and these internal organs were visible through the skin and peritonea (Fig 1B). Their eyes are markedly smaller than the orbit and almost enter the dorsal profile of the head. The two individuals share many morphological characteristics with other hadal snailfishes, including pinkish-white bodies, transparent skins and peritonea, and absent pseudobranchia [15,18,19]. Phylogenetic analyses based on 16S rRNA and cytochrome c oxidase subunit I (COI) genes indicated that the snailfish from the Yap Trench fall into the same clade as the hadal snailfish from the Mariana Trench (S1 Fig). Both morphological observations and phylogenetic analyses showed that our specimens are highly similar to the Mariana hadal snailfish [15,19], and tentatively named Yap hadal snailfish (YHS).

We sequenced the genome of a YHS specimen and obtained 81.82 gigabases (Gb) of PacBio reads and 44.83 Gb of Illumina raw reads, which correspond to 99.34- and 59.87-fold coverage



**Fig 1. Collection and morphology of Yap hadal snailfish and its genomic characteristics.** (A) Bathymetric map of the Yap Trench. The red dot indicates the location at which the two hadal snailfish were caught by the Chinese manned submersible Jiaolong. The bathymetric map was obtained from GEBCO Compilation Group (2020) GEBCO 2020 Grid (doi:10.5285/a29c5465-b138-234d-e053-6c86abc040b9). (B) Yap hadal snailfish (YHS) in situ at 6,903 m (above) and after capture (below). (C) Distribution of TE families across the YHS genome: DNA transposons (DNA), long interspersed nuclear elements (LINEs), long terminal repeats (LTRs), short interspersed nuclear elements (SINEs), and unknown TEs (unknown). (D) Venn diagram showing shared and unique gene families across YHS, Mariana hadal snailfish, Tanaka's snailfish, and zebrafish.

<https://doi.org/10.1371/journal.pgen.1009530.g001>

of the entire genome (S1 Table), respectively. The estimated size of the YHS genome is approximately 815.59 megabases (Mb), and the genome exhibits 0.61% heterozygosity (S2 Table). A total 129.70 Gb of PacBio subreads (80.87 Gb) and Illumina clean reads (44.08 Gb) were used for the *de novo* assembly of the YHS genome. The final genome assembly is approximately 731.75 Mb, which includes 1,271 scaffolds with a contig N50 of 0.75 Mb and a scaffold N50 of 1.26 Mb (Tables 1 and S3). Importantly, about 95.70% of the high-quality clean reads map to the genome assembly, which accounts for 96.17% of the complete assembly (S4 Table). In addition, 228 of the 248 highly conserved core proteins identified with the Core Eukaryotic Genes Mapping Approach (CEGMA; 91.94%), as well as 90.30% of all complete Actinopterygii Benchmarking Universal Single-Copy Orthologs (BUSCOs), were identified in our assembly (S5 and S6 Tables).

**Table 1. Statistics of the genome assembly and annotation of Yap hadal snailfish.**

Genome assembly	Data
Contig N50 size (Mb)	0.75
Scaffold N50 size (Mb)	1.26
Estimated genome size (Mb)	815.59
Assembled genome size (Mb)	731.75
Genome coverage (×)	159.21
Longest scaffold (bp)	8,357,545
Genome annotation	
Number of protein-coding genes	24,329
Number of annotated functional genes	24,265
Repeat content (%)	53.61

<https://doi.org/10.1371/journal.pgen.1009530.t001>

## Characterization of the Yap hadal snailfish genome

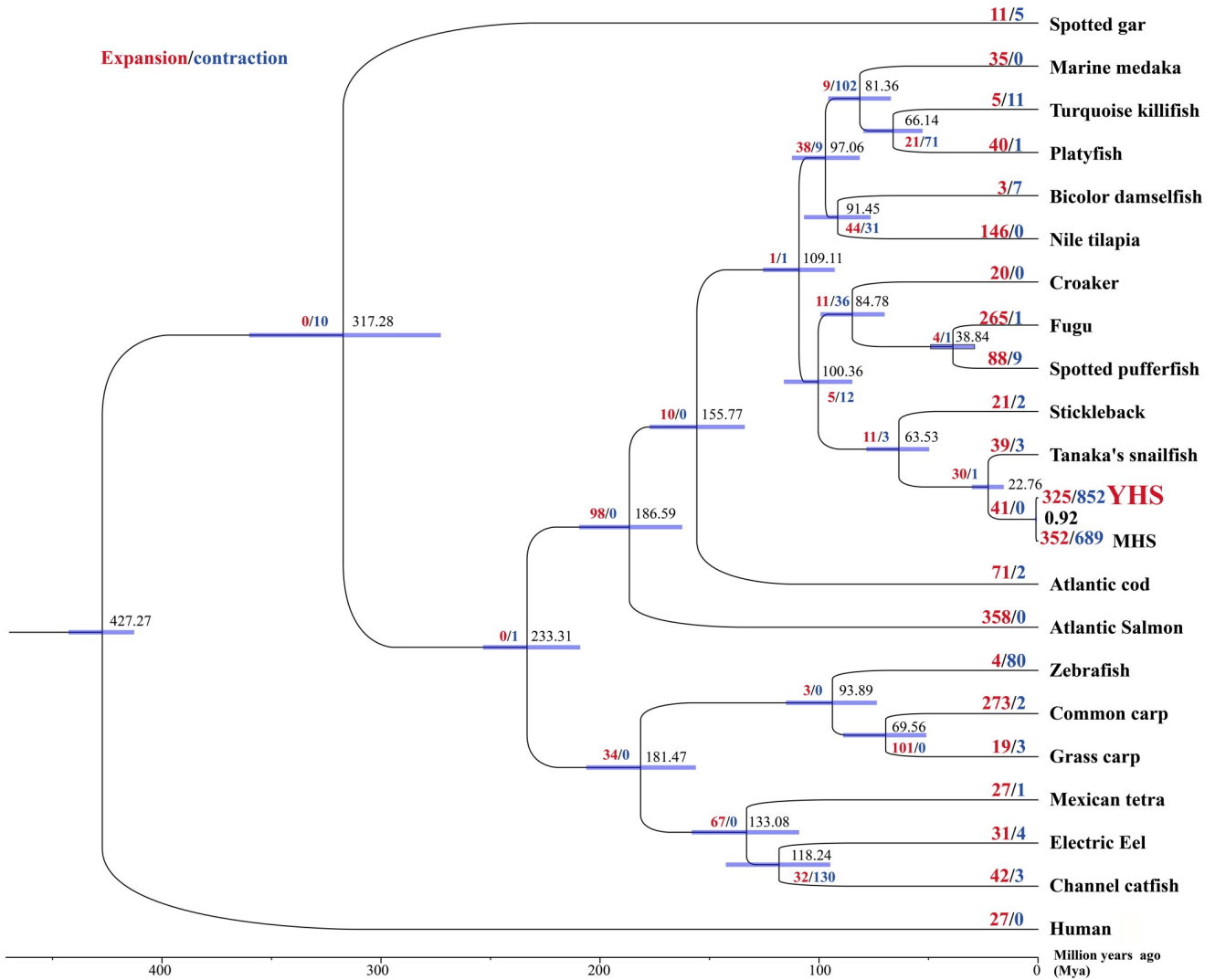
The GC content of the YHS genome assembly is 43.92% (S7 Table), the heterozygous single nucleotide polymorphism (SNP) rate is 0.2078%, and the homologous SNP rate is 0.0022% (S8 Table). The repetitive elements comprise 53.61% of the genome assembly (S9 Table), which is a much higher percentage of the genome than those found for the other sequenced snailfishes, including Mariana hadal snailfish (36.38%) and Tanaka's snailfish (24.09%) [19]. The repetitive elements are comprised mostly of transposable elements (TEs), which account for 48.47% of the YHS genome (Fig 1C and S10 Table). Most of the TEs are long interspersed elements (LINEs; 23.98% of the genome), DNA transposons (13.54%), long terminal repeats (LTRs; 13.03%), or short interspersed elements (SINES; 0.40%). We identified 2.14 Mb of noncoding RNAs (ncRNAs) in the YHS genome, including microRNAs (miRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), and transfer RNAs (tRNAs), which account for 0.29% of the genome assembly (S11 Table).

After characterizing the repetitive sequences and ncRNAs, we predicted 24,329 protein-coding genes in the final assembled genome (S12 Table), and 18,537 of these genes (76.19%) are supported by transcriptome data (S2 Fig and S13 Table). The number of protein-coding genes identified in the YHS genome is similar to previously published numbers reported from other diploid teleost genomes (S14 Table), including Mariana hadal snailfish (25,262 genes), Tanaka's snailfish (*Liparis tanakae*; 23,776 genes), large yellow croaker (*Larimichthys crocea*; 22,274 genes), and zebrafish (*Danio rerio*; 25,619 genes). On average, the transcript length, coding sequence length, and intron length of the protein-coding genes in the YHS genome are 10,675 bp, 1,420 bp, and 1,249 bp, respectively. Moreover, each gene contains an average of 8.41 exons, which is similar to reports from other teleosts (S3 Fig). Finally, we successfully annotated 24,265 of the predicted protein-coding genes (99.74%; S15 Table).

## Gene family and genome evolution

Across 22 representative vertebrate genomes (21 teleosts and human, S16 Table and S4 Fig), we identified 25,787 gene families and 259 shared single-copy gene families (S5 Fig). A total of 144 gene families were found only in the YHS genome. We found a high degree of gene-family overlap among YHS, Mariana hadal snailfish, Tanaka's snailfish, and zebrafish with a core set of 9,947 gene families (Fig 1D). A coalescent species tree based on 45 partitioned datasets of the nonrecombinant single-copy orthologous genes and a dataset of concatenated single-copy orthologous genes (10,697 amino acids), recovered YHS and Mariana hadal snailfish as sister to a clade comprising Tanaka's snailfish (Fig 2). The estimated time of divergence between our



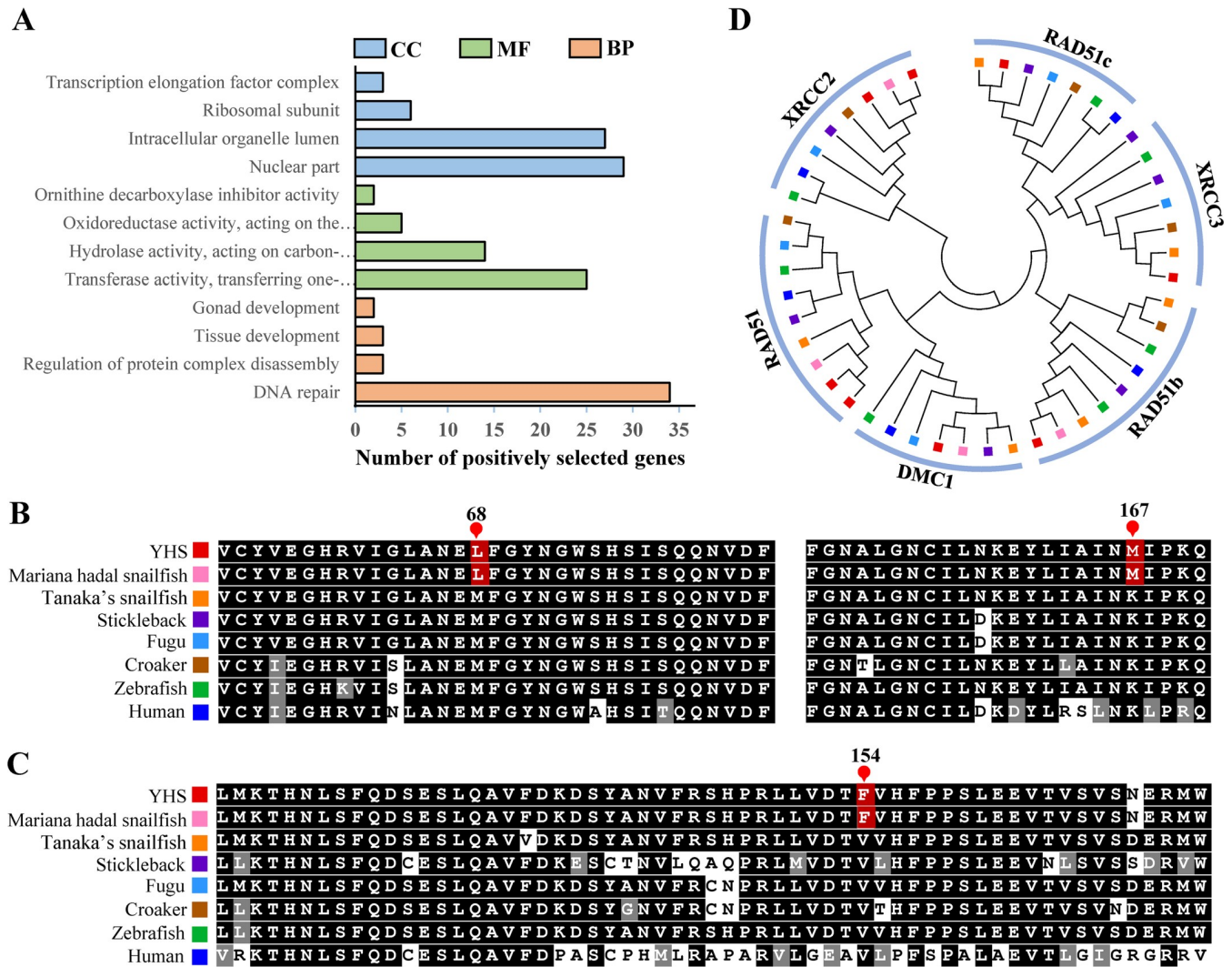


**Fig 2. Coalescent species tree and divergence time estimation for Yap hadal snailfish and 20 other teleost species.** Human served as the outgroup species. The purple rectangle bar at each node indicates the 95% confidence interval. The numbers at nodes represent estimated divergence times (Mya), and the numbers on branches indicate the event of gene family expansion (red) and contraction (blue). YHS: Yap hadal snailfish, MHS: Mariana hadal snailfish.

<https://doi.org/10.1371/journal.pgen.1009530.g002>

YHS and Mariana hadal snailfish was approximately 0.92 million years ago (Mya), whereas that between the two hadal snailfishes and Tanaka’s snailfish was approximately 22.76 Mya (Fig 2), which is consistent with a previous report [19]. Furthermore, 852 gene families are significantly contracted and 325 are significantly expanded in the YHS genome (Fig 2).

Compared with four shallow-water teleosts, 1,621 positively selected genes (PSGs) were identified in the YHS genome. Gene Ontology (GO) enrichment analysis revealed that 19 GO terms are overrepresented among these PSGs (level 4; S17 Table), and these terms primarily include cellular nitrogen compound metabolic process (356 genes), cellular aromatic compound metabolic process (348 genes), organic cyclic compound metabolic process (352 genes), heterocycle metabolic process (349 genes), nucleic acid binding (304 genes), and DNA repair (34 genes). Only one Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, “cytokine-cytokine receptor interaction,” is significantly enriched with six PSGs (*bmp2*, *csf3r*, *cxcr3*, *il11ra*, *il12rb2*, and *ngfr*).



**Fig 3. Genomic indications of enhanced DNA repair mechanisms in the Yap hadal snailfish (YHS).** (A) Gene ontology enrichment of the positively selected genes from YHS (level 4). CC, cellular component, MF, molecular function, BP, biological process. (B, C) Partial alignment of the (B) RAD52 and (C) RAD9A amino acid sequences from various representative teleosts and human. Amino acids unique to two hadal snailfishes (YHS and Mariana hadal snailfish) are highlighted in red; positions within each sequence are given above. (D) A maximum-likelihood tree showing RAD51 and RAD51 paralogs (RAD51b, RAD51c, XRCC2, XRCC3, and DMC1). The leaf-node colors correspond to the species given in panels (B) and (C).

<https://doi.org/10.1371/journal.pgen.1009530.g003>

### DNA repair capacity

High hydrostatic pressure can cause DNA breakage and damage [9,20]. Consistent with this, the GO term “DNA repair” is enriched with 34 PSGs (Fig 3A), including *rad52*, *rad9a*, *ercc1*, *exo1*, *pms1*, and *polk* (S18 Table). Specifically, both YHS and Mariana hadal snailfish have the same two high-confidence amino acid substitutions in the DNA repair protein RAD52 homolog (RAD52) compared with the corresponding completely conserved amino acids in Tanaka’s snailfish, large yellow croaker, zebrafish, stickleback (*Gasterosteus aculeatus*), and human (*Homo sapiens*): a methionine-to-leucine substitution at position 68 and a lysine-to-methionine substitution at position 167 (Fig 3B). RAD9A, a DNA damage checkpoint protein, is also different in the genomes of YHS and Mariana hadal snailfish compared with those of other examined vertebrates (a valine-to-phenylalanine substitution at position 154; Fig 3C).

Moreover, a Pfam domain analysis revealed that the YHS genome includes eight RAD51 paralog genes (S19 Table) and more copies of *rad51* and *xrcc2* than other teleost genomes (Fig 3D).

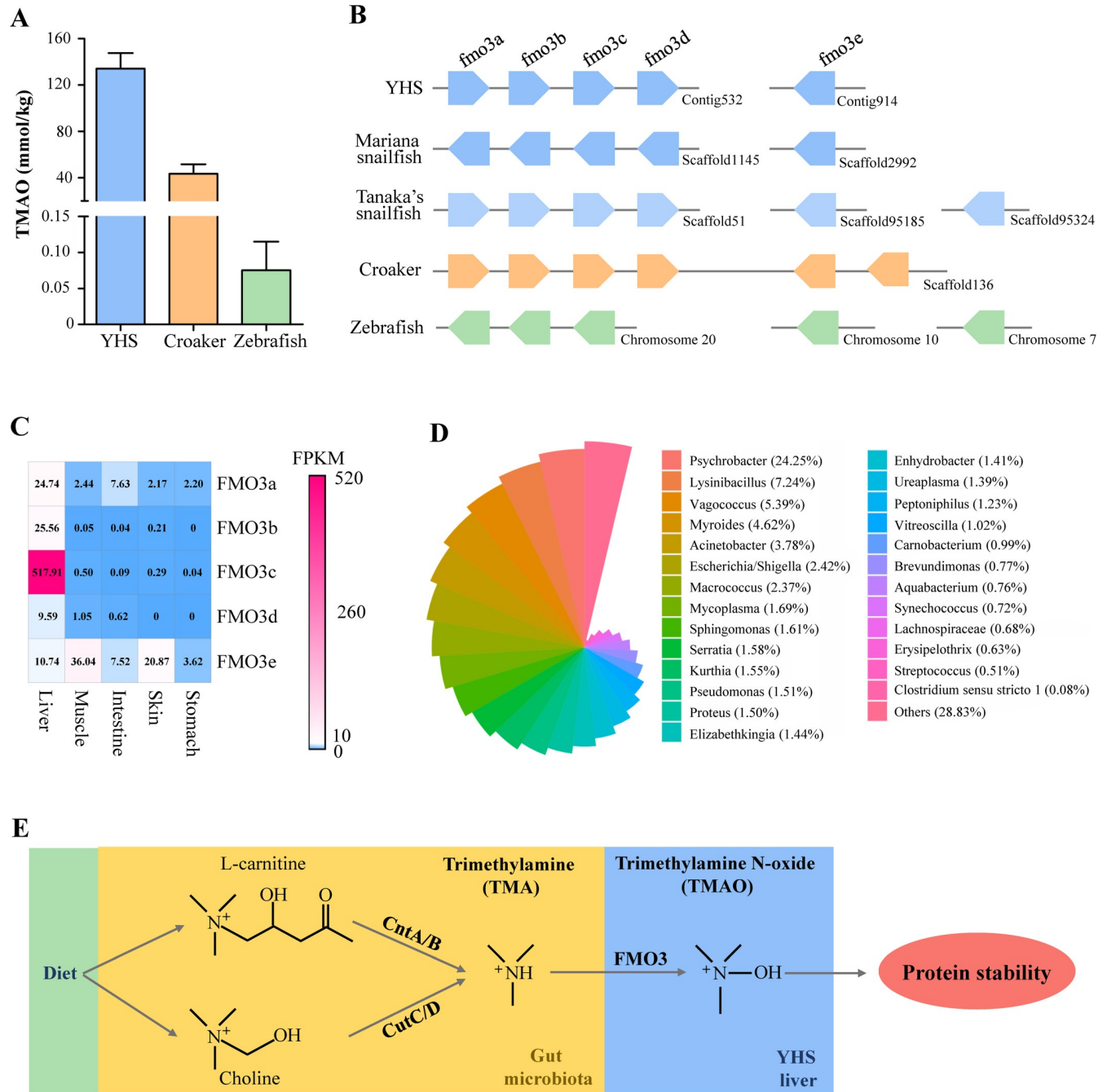
### Protein stabilization

TMAO is a potent protein stabilizer commonly found in the muscle tissues of marine fish species, and this stabilizer can alleviate the effects of hydrostatic pressure on protein stability and restore denatured proteins to their native structures [10,21]. In fish, TMAO is formed via the oxygenation of TMA by a hepatic enzyme, flavin-containing monooxygenase-3 (FMO3) [22]. However, TMA must be synthesized by the gut microbiota via two major pathways: TMA-lyase (CutC) in conjunction with its activator CutD utilizing choline as a substrate and a two-component Rieske-type oxygenase/reductase (CntA/B) acting on L-carnitine and its derivative gamma-butyrobetaine [23]. Consistent with expectations, the TMAO content in the YHS muscles (134 mmol/kg) is much higher than that in the large yellow croaker (43 mmol/kg) and zebrafish (0.07 mmol/kg; Fig 4A). Similar to Mariana hadal snailfish, five copies of the TMAO-generating gene (*fmo3*) were identified in the YHS genome, and four of them are tandem repeats (Fig 4B). The abundance of FMO3c transcripts in the YHS liver was greater than the abundance of all other FMO3 genes across all organs tested (Fig 4C), suggesting that FMO3c might be a major oxidase for TMAO production in YHS. An analysis of the gut microbiota of YHS showed that bacterial genera containing species carrying the *cutC* gene account for 8.8% of all microbes (*Escherichia/Shigella*, 2.42%; *Serratia*, 1.58%; *Vitreoscilla*, 1.02%; and *Acinetobacter*, 3.78%) and that bacterial genera containing species carrying the *cntA* gene account for 4.59% (*Escherichia/Shigella*; *Clostridium sensu stricto* 1, 0.08%; *Serratia*; and *Streptococcus*, 0.51%; Figs 4D and S6). The abundance of TMA, in conjunction with the five copies of the *fmo3* gene, might maintain a high level of TMAO in Yap hadal snailfish to improve protein stability under hadal conditions (Fig 4E).

### Sensory systems

Hadal environments are characterized by a low food supply and darkness, which are similar to the characteristics of underwater cave habitats [5,24]. We compared the genes associated with the sensory systems in hadal snailfishes to those of cave-restricted fishes (*Sinocyclocheilus anshuiensis*, *S. grahami*, and *S. rhinocerosus*) and other shallow-water fish species to explore the genetic basis for their unique sensory characteristics. A syntenic analysis identified two copies of the sour taste receptor gene (polycystic kidney disease 2-like 1, *pkd2l1*) in both the YHS and Mariana hadal snailfish genomes (Fig 5A). In comparison, only one copy of *pkd2l1* was identified in other diploid teleosts, including zebrafish, stickleback, large yellow croaker, pufferfish (*Tetraodon nigroviridis*), and tuna (*Thunnus atlanticus*), whereas two copies were identified in tetraploid cavefishes (e.g., *Sinocyclocheilus* species) due to genome duplication [25]. We also found that the bitter taste gene (taste receptor type 2, *tas2r*) has been lost in the hadal snailfish genomes, even though reference genes have been found in zebrafish and *Sinocyclocheilus* cavefishes (Fig 5A). In addition, we predicted 40 olfactory receptor (OR) genes in the YHS genome, including 25 functional genes and 15 pseudogenes, similar to the results found for Mariana hadal snailfish (S20 Table). The genomes of hadal snailfishes have fewer olfactory receptor genes than the genomes of shallow-water teleosts, such as *Sinocyclocheilus* cavefishes (26–33 functional genes and 17–35 pseudogenes), pufferfish (44 functional genes and 54 pseudogenes), and zebrafish (102 functional genes and 35 pseudogenes; S20 Table). More specifically, only 14 functional genes encoding  $\delta$  group olfactory receptors that are important for the perception of water-borne odorants [26], were identified in the YHS genome, and this number is far fewer than that found in zebrafish (Figs 5B and S7). The expansion of certain taste receptor

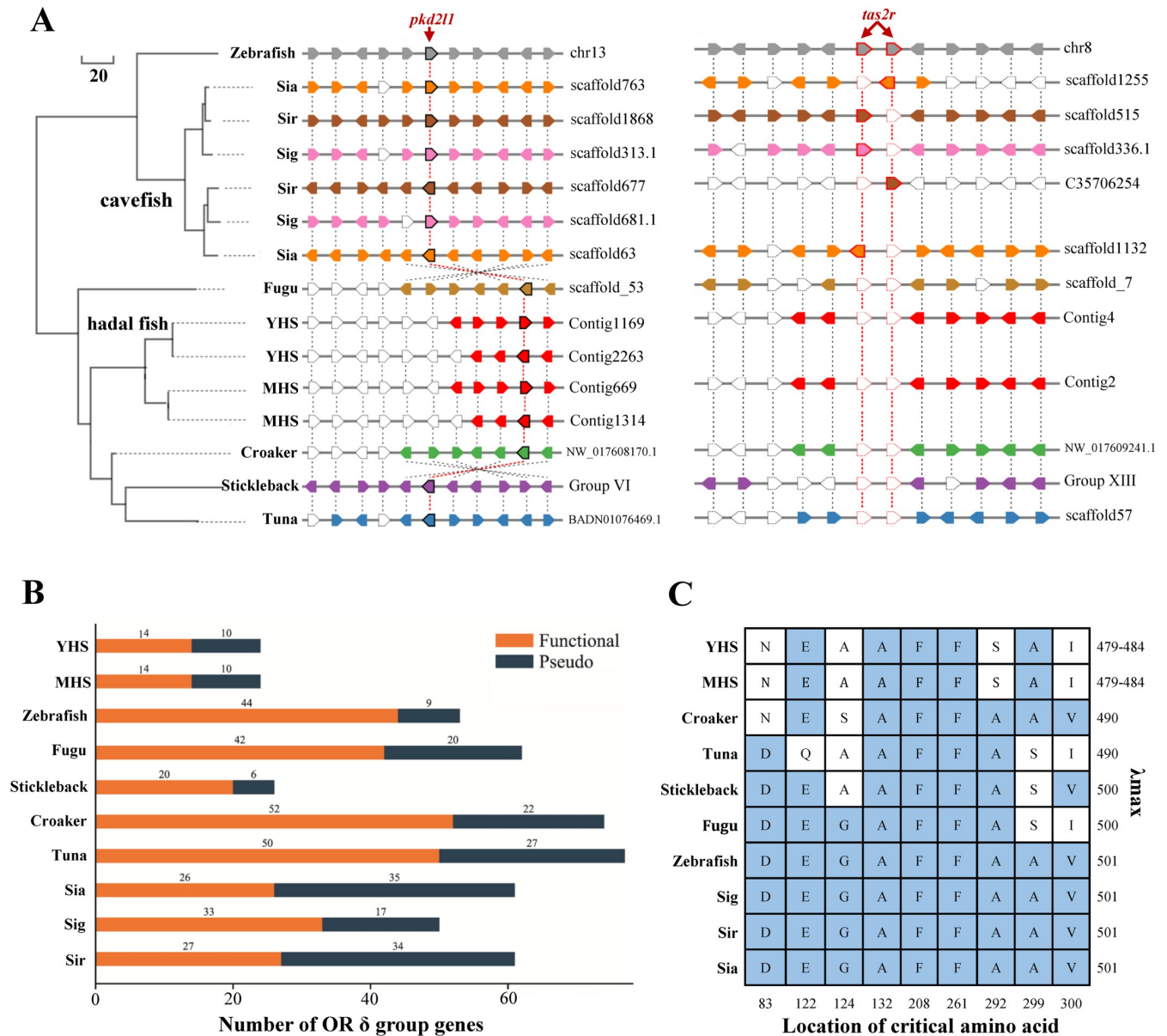




**Fig 4. Potential mechanism of TMAO-mediated protein stabilization in Yap hadal snailfish.** (A) Muscle TMAO contents (mmol/kg wet mass) in three teleosts. (B) Arrangement of the TMAO-generating enzyme *fmo3* genes in the genomes of Yap hadal snailfish (YHS), Mariana hadal snailfish, Tanaka's snailfish, large yellow croaker, and zebrafish. (C) Relative expression of *fmo3* genes in the liver, muscle, intestine, skin, and stomach of YHS. Gene expression was quantified as fragments per kilobase of transcript per million fragments mapped (FPKM) values. (D) Bacterial communities in the YHS gut identified using the Silva database. (E) Proposed TMAO biosynthesis pathway in the YHS. Through this pathway, both the gut microbiota and five copies of *fmo3* help to maintain a high TMAO levels and thus improve protein stability to ameliorate the destabilizing effects of hydrostatic pressure.

<https://doi.org/10.1371/journal.pgen.1009530.g004>

genes and the massive loss of functional olfactory receptors in hadal snailfish genomes might be due to their specific dietary habits in hadal trenches. Trench communities are typically considered as nutrition-limited systems, and food resources are mainly supplemented by surface-



**Fig 5. Genetic features of the unique sensory systems of Yap hadal snailfish.** (A) Evolution and synteny of taste genes across several representative teleosts. The reference gene positions were based on the zebrafish genome. The white pentagons indicate lost genes. *Tas2r*: taste receptor type 2, *pkd21l1*: polycystic kidney disease 2-like 1. (B) Numbers of functional genes or pseudogenes encoding  $\delta$  group olfactory receptors in various fish species. (C) Amino acid residues at the nine critical sites in rhodopsin across representative fish species and  $\lambda_{max}$  values of rhodopsin in different fish species. YHS: Yap hadal snailfish, MHS: Mariana hadal snailfish, Sia: *Sinocyclocheilus anshuiensis*, Sig: *Sinocyclocheilus grahami*, Sir: *Sinocyclocheilus rhinoceros*.

<https://doi.org/10.1371/journal.pgen.1009530.g005>

derived carrion falls [1,5]. However, most of the sinking materials are consumed and intercepted by plankton and heterotrophic bacteria in shallower and bathyal waters. As the top predator in the hadal ecosystem, the snailfish possesses an inflated stomach that is typically filled with only one dominant crustacean species, *Hirondellea gigas* [27]. The relatively simple diet of the hadal snailfish and the limited food sources available to this species might have driven adaptive alterations in its taste sensation and olfaction. Therefore, the changes in the gene number of taste and olfaction receptors might facilitate the foraging of hadal snailfish in the nutrition-limited hadal environment.

Light in the hadal zone is extremely faint and has two primary natural sources: residual sunlight and bioluminescence [28]. Previous video observations have shown that the hadal snailfishes do not respond to strong light [19,29]. We therefore performed a comparative genomic analysis and found that the YHS had fewer copies of crystallin genes relative to those found in other sequenced teleosts (S8 Fig and S21 Table). Notably, the number of  $\gamma$ -crystallin genes in YHS was markedly lower than that in any of the shallow-water teleosts examined (S21 Table). The crystallins of the vertebrate eye lens are the predominant structural proteins that maintain the transparency and high refractive index of the lens, which enables the focusing of light on the retina [30]. The loss of crystallin genes suggests that the visual system of the YHS might have degenerated during life in the dark. We also identified two opsin genes, rhodopsin *rh1* and shortwave-sensitive *sws2*, in the YHS genome. The  $\lambda_{\max}$  of YHS rhodopsin is 479–484 nm, which is lower than the levels found in shallow-water teleosts (Fig 5C). Moreover, the same frameshift insertion in aralkylamine N-acetyltransferase 2 (*aanat2*), the critical gene for melatonin biosynthesis, was observed in both YHS (S9 Fig) and Mariana hadal snailfish [31], which led to the inactivation of AANAT2. This inactivation in the two hadal snailfishes may result in low levels of blood melatonin, which potentially reflects an adaptation to the darkness in hadal environments. Thus, Yap hadal snailfish might principally sense shortwave light, similar to fish with degenerated eyes, such as cavefish [25].

## Discussion

High hydrostatic pressures, low temperatures, and a scarce food supply are thought to be the major barriers to survival in the deep-sea environment, but a specialized fauna thrive in the hadal zone at depths exceeding 6000 m [32]. The most common vertebrate species in the hadal zone is snailfish, and the deepest snailfish recorded to date was captured from a depth exceeding 8,100 m [17]. Recently, the genetic basis and mechanisms of vertebrate adaptation to such an extreme environment have attracted more attention. Genomic analyses of a hadal snailfish from the Mariana Trench have revealed its adaptation to the extreme hydrostatic pressure, which possibly involve enhancing the cell membrane fluidity, transport protein activity, and protein stability [19]. Here, we sequenced and assembled a genome of another hadal snailfish from the Yap Trench. The final genome assembly is approximately 731.75 Mb, with a contig N50 of 0.75 Mb and a scaffold N50 of 1.26 Mb (Table 1), which are much higher than those found in the Mariana hadal snailfish genome; the genome assembly of Mariana hadal snailfish is 684 Mb, with a contig N50 of 0.34 Mb and a scaffold N50 of 0.42 Mb [19]. Thus, our assembly improved the genome quality of hadal snailfish. Coalescent species tree analysis recovered a sister relationship between Yap hadal snailfish and Mariana hadal snailfish, with a divergence time of approximately 0.92 Mya (Fig 2). In combination with their similar morphology and genome structure characteristics, Yap hadal snailfish and Mariana hadal snailfish might possess similar genetic features and adaptive mechanisms to the hadal ecosystem. Further genomic analyses revealed significant alterations in several gene families, such as taste receptors, olfactory receptors, and vision-related genes, in Yap hadal snailfish (Fig 5), and these alterations might provide the genetic basis for the adaptation of Yap hadal snailfish to the nutrition-limited and dark hadal environments.

DNA is vulnerable to high hydrostatic pressure, and hadal organisms must employ efficient DNA repair mechanisms to alleviate hydrostatic pressure-associated DNA damage [9,20]. Homologous recombination is a high-fidelity process that uses homologous DNA sequences as templates for the repair of damaged DNA [33]. RAD52 is an important member of the homologous recombination pathway, which promotes annealing between two complementary single-stranded DNA (ssDNA) molecules or between one ssDNA molecule and a

complementary ssRNA molecule [34]. RAD9A, a DNA damage checkpoint protein, is essential for the DNA damage response [35]. Compared with shallow-water teleosts, substitutions with longer branched amino acids were detected in RAD52 (M68L and K167M) and RAD9A (V154F) of Yap hadal snailfish and Mariana hadal snailfish (Fig 3), which might contribute to the maintenance of the structure and function of these two proteins under ultrahigh hydrostatic pressure [2,36,37]. Similar circumstances were also observed in the hadal amphipod *Hirondellea gigas* [4], the hadal holothurian *Paelopatides* sp. [37] and the deep-sea fish *Aldrovandia affinis* [9]. In *H. gigas*, positive selection was observed in the replication factor A1 (RFA1) gene that is associated with DNA repair and maintenance of chromosomal stability [4]. Eight positively selected genes involved in DNA repair were identified in *Paelopatides* sp., including RAD9A [37]. Therefore, the amino acid substitutions in deep-sea organisms might reflect hydrostatic pressure-associated positive selection, suggesting that deep-sea species may share similar adaptive strategies. RAD51, a major eukaryotic homologous recombinase, plays a key role in homologous recombination by promoting the search for homologous double-stranded DNA (dsDNA) templates and repairing DNA double-strand breaks [38]. The expansion of *rad51* in Yap hadal snailfish might increase the DNA repair rates and thus facilitate the maintenance of DNA integrity under adverse environmental conditions. Therefore, the positive selection and expansion of genes needed for DNA repair imply that a stronger DNA repair capacity might be essential for the adaptation of hadal organisms to the high hydrostatic pressures in hadal environments.

Hydrostatic pressure affects protein folding and function [8]. Consequently, the organisms living in the hadal zone must use various mechanisms to maintain the stability of protein structures under elevated hydrostatic pressures [3]. TMAO is a physiologically strong protein stabilizer that plays a role in the stabilization of protein structures [10,21,39]. In this study, we found that the TMAO content in the muscle of Yap hadal snailfish was markedly higher than that in the muscle of shallow-water fish (Fig 4A), which is similar to the results found for other deep-sea fish [10,14,40]. TMAO is synthesized from TMA by host hepatic flavin monooxygenase 3 (FMO3) [22]. However, TMA cannot be synthesized in fish and is mainly produced by microbes in the fish gut [29]. An analysis of the gut microbiota showed that TMA-generating bacteria were abundant in the gut of Yap hadal snailfish (Fig 4D), suggesting that hadal snailfish might largely depend on these TMA-generating bacteria for their TMA supply. The synthesis of TMA by gut bacteria, in conjunction with the presence of five copies of *fmo3* in Yap hadal snailfish (Fig 4B), could facilitate the synthesis of TMAO in this species. The effective TMAO synthesis would help Yap hadal snailfish adapt to high hydrostatic pressure by improving protein stability under hadal conditions. To the best of our knowledge, this study constitutes the first demonstration of the role of the gut microbiota in the adaptation of hadal snailfish to high hydrostatic pressure. Though the abundance of TMA-generating bacteria in the gut was not analyzed, Mariana hadal snailfish also contain five copies of *fmo3* in their genomes and five putative promoters were predicted upstream of the *fmo3a* [19]. Thus, the enhancement of TMAO synthesis might represent a unique adaptation mechanism of hadal snailfishes to hadal environments, which would help the snailfishes live in the harsh hadal trenches.

In conclusion, we assembled a high-quality reference genome for Yap hadal snailfish. Comparative genomic analyses revealed significant alterations in several gene families associated with sensory systems and DNA repair in Yap hadal snailfish, providing the genetic basis for the adaptation of Yap hadal snailfish to hadal environments. The high levels of TMAO found in Yap hadal snailfish, which might be due to the presence of five copies of the TMAO-generating enzyme *fmo3* in the genome and the abundance of TMA-generating bacteria in the gut, might facilitate its adaptation to high hydrostatic pressure. Our results provide new insights



into the molecular mechanisms underlying the adaptation of hadal organisms to the deep-sea environment. However, the exact functions of these positively selected and expanded genes in hadal snailfish adaptation require further investigation.

## Materials and methods

### Ethics statement

The studies were performed in strict accordance with the Regulations of the Administration of Affairs Concerning Experimental Animals established by the Fujian Provincial Department of Science and Technology. The animal experiments were approved by the Animal Care and Use Committee of the Fujian Agriculture and Forestry University (PZCASFAFU2019019). All efforts were made to minimize suffering.

### Sample collection and DNA/RNA extraction

One YHS individual was randomly selected for genomic analysis and genomic DNA was extracted from its liver tissue using the conventional sodium dodecyl sulfate (SDS) method [41]. The DNA quality was measured using a Bioanalyzer (Agilent Technologies, USA) and by agarose gel electrophoresis. Two mitochondrial fragments, the 16S rRNA gene (830 bp) and cytochrome c oxidase subunit I gene (COI, 645 bp), were amplified from the genomic DNA using polymerase chain reaction (PCR) using optimized primer pairs (16S rRNA-F: CTATT AATACCCCAAAATACCCC, 16S rRNA-R: CGATGTTTTTGGTAAACAGGCG; COI-F: TCAACCAACCACAAAGACATTGGCAC, COI-R: TAGACTTCTGGGTGGCCAAAGAA TCA) [42]. The amplicons were sequenced using traditional Sanger sequencing. The 16S rRNA and COI gene sequences of four other hadal snailfishes (*P. swirei*, *Notoliparis kermadecensis*, *Rhodichthys regina*, and *Careproctus marginatus*) and two shallow-water snailfishes (*L. tanakae* and *Liparis ochotensis*) were retrieved from the GenBank database for comparison. Maximum likelihood (ML) trees were constructed based on each gene using MEGA 6.06 [43] with default parameters. Genetic distances were computed using the Kimura 2-parameter (K2P) algorithm. The rate variation among sites was modeled using a gamma distribution (shape parameter = 1), and all positions containing gaps and missing data were eliminated (i.e., the complete deletion option) [44]. Six tissues (eye, gut, liver, muscle, skin, and stomach) were then collected from the selected YHS for transcriptome sequencing.

### Library construction and sequencing

Genomic DNA was fragmented using a Covaris sonication system. Short-insert paired-end libraries (350 bp) were constructed according to Illumina's protocol with end repair, poly-A tail base addition, sequencing-adaptor ligation, amplification, and purification. Paired-end sequencing was performed with an Illumina HiSeq X Ten platform (Illumina, USA). The Illumina raw reads were evaluated using FASTQC v0.11.6 [45] and filtered with fastp v0.20 [46] using the default parameters to produce the Illumina clean reads for subsequent analysis.

Simultaneously, the extracted genomic DNA was sheared using Megaruptor2 (Diagenode, Ougrée, Belgium), and then used to construct SMRT bell libraries via the ligation of universal hairpin adaptors onto double-stranded DNA fragments in accordance with the 20-kb preparation protocol (Pacific Biosciences, USA). The MagBead kit (Pacific Biosciences, USA) was used to remove adaptor dimers. The failed ligation products were removed using exonucleases. The sequencing primer was annealed to each SMRT bell template for subsequent sequencing with a PacBio Sequel instrument using Sequel Sequencing Kit 1.2.1 (Pacific Biosciences, USA).

The PacBio polymerase reads were filtered using the RS\_Subreads protocol with minimum length > 300 bp to produce the PacBio subreads for genome assembly.

### Genome-size estimation

We estimated the genome size of YHS based on the 17-mer frequency distribution of the 44.08-Gb Illumina clean reads using the following formula: genome size = (total number of 17-mers) / (position of the peak depth) [47].

### Genome assembly

All of the Illumina clean reads were subsequently split into small K-mers, and low-occurrence K-mers were removed. A de Bruijn graph was constructed using Platanus assembler v1.2.1 [48] with the following optimized parameters: input type = raw, genome size = 828360000, seed coverage = 50, and length cutoff pr = 10000. PacBio subreads longer than 300 bp were retained, corrected using Canu v1.5 [49], and assembled into an initial genome using Falcon v1.8.8 [50]. Pilon v1.22 [51] was used to polish the genome assembly using the clean reads. Finally, the PacBio subreads were scaffolded using SSPACE-LongRead v1-1 [52] with the default parameters. Gaps in the constructed scaffolds were filled based on the PacBio subreads using PBJelly v14.9.9 [53] with the default parameters.

### Assessment of the constructed genome assembly

Three approaches were integrated to evaluate the quality of the YHS genome assembly: read alignments, BUSCO (version 3.03) [54], and CEGMA (version 2.5) [55]. To assess the assembly based on read alignments, we mapped the Illumina clean reads onto scaffolds using BWA v0.6.2 [56] with the optimized parameters "-o 1 -i 15". Subsequently, to evaluate the completeness of the YHS genome assembly, we assessed the assembled scaffolds using BUSCO v3.03 (RRID: SCR\_015008) [54] with the default parameters, against the conserved core genes in the Actinopterygii\_odb9 database (4,584 orthologs). Finally, known genes in the genome assembly were aligned against a reliable set of 248 highly conserved proteins from a wide range of eukaryotes using CEGMA [55].

### Detection of repetitive sequences and noncoding RNAs

Repeat annotations were performed to clarify the genome assembly of YHS. We identified TEs using a combination of homology-based and *de novo* prediction methods. First, we used RepeatModeler v1.0.8 [57], LTR\_FINDER v1.06 [58], and RepeatScout v1.0.5 [59] to build a *de novo* repeat library. Subsequently, we performed homology-based gene predictions using RepeatMasker [57] with the default parameters against Repbase and the *de novo* repeat libraries. Simultaneously, we used RepeatProteinMask [57] with the default parameters to identify repeated amino acid sequences. Tandem Repeats Finder [60] was used to identify tandem repeats with the following parameters: Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, and MaxPeriod = 2000. After integration of all repeat annotation results using an in-house Perl script, we calculated the sequence divergence rate for each family of TEs. The genes associated with noncoding RNAs (ncRNAs) in the YHS genome, including microRNAs (miRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), and transfer RNAs (tRNAs), were predicted using Infernal v1.1.2 [61] based on alignments to the Rfam ncRNA database [62].

## Prediction of protein-coding genes

We used a combination of *de novo*, homology-based, and transcriptome-based methods for the prediction of protein-coding genes. We performed *de novo* prediction of the repeat-masked genome using Augustus v3.0.2 [63], Genescan v1.0 [64], Geneid v1.4 [65], GlimmerHMM v3.0.2 [66], and SNAP v2.0 [67]. To predict protein-coding genes based on homology, we obtained the longest available protein-coding sequences of eight representative vertebrate species from GenBank: human, large yellow croaker, spotted green pufferfish, three-spined stickleback, zebrafish, Nile tilapia (*Oreochromis niloticus*), fugu (*Takifugu rubripes*), and mouse (*Mus musculus*). These protein sequences were aligned to the assembled YHS genome using TBLASTN with an e-value < 1e-5, and the alignment results were integrated using SOLAR v0.96 [68,69]. GeneWise v2.2.0 was then used to predict gene models based on this alignment [70]. Protein-coding genes were also predicted based on the transcriptome data from six tissues (eye, intestine, liver, muscle, skin, and stomach) of YHS. We mapped the transcriptome data to the YHS genome assembly using TopHat v2.0.13 [71] and obtained gene structures using Cufflinks v2.1.1 [72]. In parallel, we assembled the transcriptome data using Trinity v2.1.1 [73] and obtained gene structures using PASA v2.3.3 [74]. Gene expression levels were determined based on fragments per kilobase of transcripts per million fragments mapped (FPKM) values using RNA-Seq by Expectation Maximization (RSEM) with the default settings [75]. The gene structures obtained using these three approaches were integrated with EVIDENCEModeler v1.1.1 [76] to yield a nonredundant gene set. Finally, we used PASA v2.3.3 [74] to adjust the gene models based on the assembled transcripts in order to obtain a final reference protein-coding gene set.

## Annotation of protein-coding genes

We functionally annotated the genes in the final protein-coding gene set to better understand their biological roles. First, we annotated the deduced protein sequences using InterProScan v4.7 [77] with the default settings, to identify motifs and domains. Then, Gene Ontology (GO) terms for each gene were assigned based on the corresponding InterPro descriptions. We then searched for the deduced protein sequences in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database using the bidirectional best hit (BBH) method to identify associated pathways. Finally, we annotated the deduced protein sequences in the Swiss-Prot and TrEMBL databases with an e-value of 1e-5 for prediction of protein function.

## Genome evolution

Gene family clusters were determined using OrthoMCL v1.1 [78] based on 22 representative vertebrate genomes: Yap hadal snailfish, Mariana hadal snailfish, Tanaka's snailfish, Atlantic salmon (*Salmo salar*), Atlantic cod (*Gadus morhua*), bicolor damselfish (*Stegastes partitus*), common carp (*Cyprinus carpio*), channel catfish (*Ictalurus punctatus*), electric eel (*Electrophorus electricus*), grass carp (*Ctenopharyngodon idella*), fugu, large yellow croaker, medaka (*Oryzias melastigma*), Mexican tetra (*Astyanax mexicanus*), Nile tilapia, spotted gar (*Lepisosteus oculatus*), spotted green pufferfish, Southern platyfish (*Xiphophorus maculatus*), stickleback, turquoise killifish (*Nothobranchius furzeri*), zebrafish, and human. For each species, we retained the longest transcript of each gene, and removed the genes encoding a protein consisting of less than 30 amino acids. We then identified all possible matches among the retained protein sequences through All-vs-All Blast with an e-value of 1e-7. Finally, we clustered the alignments into gene families using OrthoMCL [78] with an inflation index of 1.5.

The phylogenetic relationships among 22 representative vertebrates were then evaluated. First, the recombination for each single-copy gene shared across all representatives was tested

with RDP4 [79] in two rounds using the methods RDP, CHIMERA, GENECONV, and Max-Chi with Bonferroni correction (significance parameter set to 0.01). Second, the amino acid sequence of each remaining single-copy gene was separately aligned using MAFFT v7.237 [80] with the linsi package, and each served as a partitioned dataset. All of the aligned sequences were also concatenated and then served as the concatenated dataset. Third, the model for each dataset was suggested by ProtTest v3.4 [81]. A maximum likelihood (ML) phylogenetic tree of each dataset was then constructed based on the alignment using RAxML v8.2.12 [82]. Finally, a coalescent species tree was built with all of the passed dataset ML trees using ASTRAL v5.7.5 [83].

The divergence times among 22 representative vertebrates were estimated based on the constructed phylogenetic tree and the coding sequences using the MCMCTree module in PAML v4.8 [84], with the following primary parameters: clock = 2, alpha = 0.5, ncatG = 5, kappa\_gamma = 6 2, alpha\_gamma = 1 1, rgene\_gamma = 2 20 1, sigma2\_gamma = 1 10 1, burn-in = 4,000,000, sample-number = 100,000, and sample-frequency = 100. Two calibration points from TimeTree [85] were used as time priors: the divergence between human and zebrafish (413–443 Mya) [86] and the divergence between zebrafish and stickleback (206–252 Mya) [87].

Positive selection was inferred using the branch-site Ka/Ks test using the codeml module in PAML v4.8 [84] based on YHS (designated as the foreground phylogeny) and three shallow-water fish, including stickleback, large yellow croaker, and zebrafish (designated as the background phylogeny). The phylogenetic relationship across these four species was reconstructed with the concatenated sequences as described above. ParaAT v2.0 [88] with a “-g” parameter was used to align the coding DNA sequences of each ortholog according to its amino acid sequence alignment. An alternative branch site model (Model = 2, NSsites = 2, and fix\_omega = 0) and a neutral branch site model (Model = 2, NSsites = 2, fix\_omega = 1 and omega = 1) were configured. *P*-values were first computed by a chi-squared test [89] and then corrected by a multiple testing correction [90]. Genes with Bayesian Empirical Bayes (BEB) sites exceeding 90% and corrected *P*-values lower than 0.1 were considered as positively selected genes. The InterProScan annotation results of positively selected genes were used to obtain the GO term assignments with the BLAST2GO v3.1.3 (e-value <1e-6, *P*-value < 0.05) [91].

### Contraction and expansion of gene families

To examine the evolutionary history of the identified gene families, we estimated their expansion and contraction in the YHS genome and then identified those that were substantially altered across Mariana hadal snailfish, Tanaka’s snailfish, the other 18 other shallow-water teleost species and human (serving as the outgroup). The expanded and contracted gene families were identified using CAFÉ v4.2.1 [92] after the removal of those gene families with more than 200 copies in a single species but fewer than two copies in any other species. By comparing each branch to its ancestral branch, we calculated *P*-values using Fisher’s exact test and then adjusted these *P*-values based on the false discovery rate [93]. Gene families with adjusted *P*-values less than 0.05 were considered to have undergone significant contraction or expansion during evolution. The expanded domains in the YHS genome were identified using Pfam [94] with an e-value <1e-5.

### Identification of the sensory genes in teleost genomes

We downloaded the sequences of zebrafish taste receptor genes (sour, sweet, umami, bitter, and salty), olfactory receptor (OR) genes, visual opsin genes, and crystallin genes from the



GenBank or Ensembl databases. We used tBlastn v2.26 [95] to search these genes against the genomes of YHS, Mariana hadal snailfish, Fugu, stickleback, large yellow croaker, tuna, and three cave-restricted fishes (*Sinocyclocheilus anshuiensis*, *S. grahami*, and *S. rhinoceros*) with an e-value < 1e-5. We also collected the upstream and downstream genes to finish the syntenic analysis. The protein sequences were aligned using the *in*si package in MAFFT v7.237 [80], and phylogenetic trees were constructed using FastTree v2.1.10 [96]. The rhodopsin  $\lambda_{\text{max}}$  values were estimated according to a previous study [97].

### Determination of the TMAO content by LC-MS/MS

The TMAO contents (per kilogram of wet tissue) in the muscle tissues of Yap hadal snailfish, large yellow croaker (a shallow-sea fish), and zebrafish were measured by high-performance liquid chromatography and mass spectrometry (LC-MS/MS). In brief, 50 mg of muscle sample was mixed with methanol containing 50 ng/mL deuterium-labeled methyl d9-TMAO (d9-TMAO) as an internal standard. The mixture was homogenized for 2 min and centrifuged at 4°C and 13000 rpm for 10 min. The supernatant (2  $\mu$ L) was then injected into an Agilent 1290 Infinity II UHPLC System (Santa Clara, CA, USA) coupled to Agilent 6470 Triple Quadrupole MS/MS. Analytes were separated on an XBridge BEH HILIC Column (100 mm x 2.1 mm, particle size of 2.5  $\mu$ m) at room temperature. Mobile phase A was 0.15% formic acid and 10 mM ammonium acetate in water; mobile phase B was 100% acetonitrile. The mobile phase was run isocratically at a flow rate of 0.35 mL/min based on the following program: 0–1 min, 98% B; 1–6 min, 98–90% B; 6–7 min, 90% B; 7–10 min, 90–85% B; 10–12 min, 85–70% B; and 12–13 min, 70–60% B. The compounds were ionized by electrospray ionization operated in the positive mode. The capillary temperature was set to 300°C, the gas source temperature was 350°C, and the ESI voltage was 3500 V. The ion pairs used for the qualitative analysis were m/z 76–58 for TMAO and m/z 85–66 for d9-TMAO. The concentration of TMAO was calculated using QQQ quantitative analysis software (Agilent, USA).

### Cloning and classification of 16S rRNA gene sequences

Total DNA was also extracted from the YHS gut for bacterial 16S rRNA gene sequencing. The V3–V4 region of the microbial 16S rRNA gene was amplified using universal primers (341F: CCTACGGGNGGCWGCAG; 806R: GGACTACHVGGGTATCTAAT). Purified amplicons were pooled in equimolar volumes and used for paired-end sequencing (2  $\times$  250) with an Illumina platform (HiSeq 2500) following standard protocols. After quality control, the paired-end clean reads were merged using FLASH (v1.2.11) [98] and filtered with QIIME (v1.9.1) [99]. Taxonomic assignments were performed based on the SILVA database ([www.arb-silva.de/](http://www.arb-silva.de/)) using the RDP classifier v2.2 [100] with the default settings.

### Supporting information

**S1 Fig. Phylogenetic relationships among the hadal snailfish (YHS) from the Yap Trench and other snailfish species.** Maximum likelihood (ML) trees were constructed based on 16S rRNA (A) and cytochrome c oxidase subunit I (COI; B) genes. The 16S rRNA and COI gene sequences of the two Yap hadal snailfish specimens were consistent. ML bootstrap support values (> 50%) are shown. Branches with < 50% support have been collapsed. (PDF)

**S2 Fig. Comparison of the gene sets obtained using three prediction methods.** Genes in the Yap hadal snailfish genome were predicted using a combination of three approaches: *de novo*, homolog-based, and transcriptome-based methods. More than 98% of the 23,853 predicted

genes were supported by at least two methods.  
(PDF)

**S3 Fig. Comparison of gene structure characteristics among Yap hadal snailfish and other vertebrates.** The lines with different colors represent different species.  
(PDF)

**S4 Fig. BUSCO completeness assessment of 21 teleost species.** The genome completeness of 21 teleost species was estimated using BUSCO v3.03 against the Actinopterygii\_odb9 database. YHS: Yap hadal snailfish, MHS: Mariana hadal snailfish.  
(PDF)

**S5 Fig. Gene family characteristics of the genomes of Yap hadal snailfish (YHS) and other representative vertebrates.** Gene family clusters were determined using OrthoMCL v 1.1.1. For each species, the longest transcript of each gene was retained, whereas the genes encoding a protein consisting of less than 30 amino acids were removed.  
(PDF)

**S6 Fig. Tag distribution in the Yap hadal snailfish gut at different levels of classification.**  
(PDF)

**S7 Fig. Phylogenetic tree of the functional  $\delta$  group olfactory receptor genes of Yap hadal snailfish (YHS) and zebrafish.**  
(PDF)

**S8 Fig. Phylogenetic tree of the  $\gamma$ -crystallin genes of Yap hadal snailfish (YHS) and zebrafish.**  
(PDF)

**S9 Fig. Frameshift mutation in the ANNAT gene of Yap hadal snailfish (YHS).** A frameshift mutation was identified in the YHS genome that led to a severe mutation in the open reading frame (ORF) of the ANNAT gene. The translated protein according to the mutated ORF is completely distinct from that of Medaka ANNAT.  
(PDF)

**S1 Table. Summary of sequencing data for Yap hadal snailfish (YHS).**  
(PDF)

**S2 Table. Estimation of Yap hadal snailfish genome size (Kmer = 17).**  
(PDF)

**S3 Table. Summary of the genome assembly of Yap hadal snailfish.**  
(PDF)

**S4 Table. Statistics of the genome reads coverage.**  
(PDF)

**S5 Table. The CEGMA evaluation of genome assembly.**  
(PDF)

**S6 Table. The BUSCO evaluation of genome assembly.**  
(PDF)

**S7 Table. Statistics of the genome base content.**  
(PDF)

**S8 Table. SNP results of Yap hadal snailfish genome.**

(PDF)

**S9 Table. Summary of repeats in Yap hadal snailfish genome.**

(PDF)

**S10 Table. Transposable elements in Yap hadal snailfish genome.**

(PDF)

**S11 Table. Statistics of noncoding RNA of Yap hadal snailfish genome.**

(PDF)

**S12 Table. Prediction of gene structure in Yap hadal snailfish genome.**

(PDF)

**S13 Table. Information for the RNA-seq data from different Yap hadal snailfish tissues.**

(PDF)

**S14 Table. Gene structures of Yap hadal snailfish and other teleost genomes.**

(PDF)

**S15 Table. Functional annotation of genes in Yap hadal snailfish genome.**

(PDF)

**S16 Table. The accession numbers for genome assemblies used in this study.**

(PDF)

**S17 Table. GO enrichment (Level 4) of positively selected genes from the genome assembly of Yap hadal snailfish.**

(PDF)

**S18 Table. The positively selected genes involved in DNA repair from Yap hadal snailfish.**

(PDF)

**S19 Table. List of Pfam domains with more copies in Yap hadal snailfish than in other species.**

(PDF)

**S20 Table. Numbers of olfactory receptor (OR) genes in examined fish species.**

(PDF)

**S21 Table. The copy number of crystalline genes in Yap hadal snailfish and other species.**

(PDF)

## Author Contributions

**Conceptualization:** Qiong Shi, Xinhua Chen.

**Data curation:** Chao Bian, Ruoyu Liu, Guangming Shao.

**Investigation:** Yinnan Mu, Chao Bian, Guangming Shao, Tianliang He, Jingqun Ao.

**Methodology:** Yinnan Mu, Jia Li, Ying Qiu.

**Resources:** Yuguang Wang.

**Supervision:** Qiong Shi, Xinhua Chen.

**Validation:** Yinnan Mu, Wanru Li, Jingqun Ao.

**Writing – original draft:** Yinnan Mu, Chao Bian.

**Writing – review & editing:** Ruoyu Liu, Qiong Shi, Xinhua Chen.

## References

1. Jamieson AJ. Ecology of Deep Oceans: Hadal Trenches. Chichester: In: eLS. John Wiley & Sons, Ltd; 2011.
2. Morita T. Comparative sequence analysis of myosin heavy chain proteins from congeneric shallow- and deep-living rattail fish (genus *Coryphaenoides*). *J Exp Biol*. 2008; 211(Pt 9): 1362–1367. <https://doi.org/10.1242/jeb.017137> PMID: 18424669.
3. Somero GN. Adaptations to high hydrostatic pressure. *Annu Rev Physiol*. 1992; 54: 557–577. <https://doi.org/10.1146/annurev.ph.54.030192.003013> PMID: 1314046.
4. Lan Y, Sun J, Tian R, Bartlett DH, Li R, Wong YH, et al. Molecular adaptation in the world's deepest-living animal: Insights from transcriptome sequencing of the hadal amphipod *Hirondellea gigas*. *Mol Ecol*. 2017; 26(14): 3732–3743. <https://doi.org/10.1111/mec.14149> PMID: 28429829.
5. Jorgensen BB, Boetius A. Feast and famine—microbial life in the deep-sea bed. *Nat Rev Microbiol*. 2007; 5(10): 770–781. <https://doi.org/10.1038/nrmicro1745> PMID: 17828281.
6. Gerrerger ME, Andrews AH, Huss GR, Nagashima K, Popp BN, Linley TD, et al. Life history of abyssal and hadal fishes from otolith growth zones and oxygen isotopic compositions. *Deep-Sea Res Pt I*. 2018; 132: 37–50. <https://doi.org/10.1016/j.dsr.2017.12.002> WOS:000429293500005.
7. Brindley AA, Pickersgill RW, Partridge JC, Dunstan DJ, Hunt DM, Warren MJ. Enzyme sequence and its relationship to hyperbaric stability of artificial and natural fish lactate dehydrogenases. *PLoS One*. 2008; 3(4): e2042. <https://doi.org/10.1371/journal.pone.0002042> PMID: 18446214; PubMed Central PMCID: PMC2323112.
8. Yancey PH, Siebenaller JF. Co-evolution of proteins and solutions: protein adaptation versus cytoprotective micromolecules and their roles in marine organisms. *J Exp Biol*. 2015; 218(Pt 12): 1880–1896. <https://doi.org/10.1242/jeb.114355> PMID: 26085665.
9. Lan Y, Sun J, Xu T, Chen C, Tian R, Qiu JW, et al. *De novo* transcriptome assembly and positive selection analysis of an individual deep-sea fish. *BMC Genomics*. 2018; 19(1): 394. <https://doi.org/10.1186/s12864-018-4720-z> PMID: 29793428; PubMed Central PMCID: PMC5968573.
10. Yancey PH, Gerrerger ME, Drazen JC, Rowden AA, Jamieson A. Marine fish may be biochemically constrained from inhabiting the deepest ocean depths. *P Natl Acad Sci USA*. 2014; 111(12): 4461–4465. <https://doi.org/10.1073/pnas.1322003111> WOS:000333341100037. PMID: 24591588
11. Yancey PH, Blake WR, Conley J. Unusual organic osmolytes in deep-sea animals: adaptations to hydrostatic pressure and other perturbants. *Comp Biochem Phys A*. 2002; 133(3): 667–676. [https://doi.org/10.1016/s1095-6433\(02\)00182-4](https://doi.org/10.1016/s1095-6433(02)00182-4) WOS:000180019200022. PMID: 12443924
12. Sarma R, Paul S. Crucial importance of water structure modification on trimethylamine N-oxide counteracting effect at high pressure. *J Phys Chem B*. 2013; 117(2): 677–689. <https://doi.org/10.1021/jp311102v> WOS:000313920300021. PMID: 23268746
13. Gillett MB, Suko JR, Santoso FO, Yancey PH. Elevated levels of trimethylamine oxide in muscles of deep-sea gadiform teleosts: A high-pressure adaptation? *J Exp Zool Part A*. 1997; 279(4): 386–391.
14. Samerotte AL, Drazen JC, Brand GL, Seibel BA, Yancey PH. Correlation of trimethylamine oxide and habitat depth within and among species of teleost fish: An analysis of causation. *Physiol Biochem Zool*. 2007; 80(2): 197–208. <https://doi.org/10.1086/510566> WOS:000243966400004. PMID: 17252516
15. Gerrerger ME, Linley TD, Jamieson AJ, Goetze E, Drazen JC. *Pseudoliparis swirei* sp nov.: A newly-discovered hadal snailfish (Scorpaeniformes: Liparidae) from the Mariana Trench. *Zootaxa*. 2017; 4358(1): 161–177. <https://doi.org/10.11646/zootaxa.4358.1.7> WOS:000416782400007. PMID: 29245485
16. Linley TD, Gerrerger ME, Yancey PH, Drazen JC, Weinstock CL, Jamieson AJ. Fishes of the hadal zone including new species, in situ observations and depth records of Liparidae. *Deep-Sea Res Pt I*. 2016; 114: 99–110. <https://doi.org/10.1016/j.dsr.2016.05.003> WOS:000381531500009.
17. Nielsen JG. The deepest living fish *Abyssobrotula galathea*: A new genus and species of oviparous ophiroids (Pisces, Brotulidae). *Galathea Report*. 1977; (14): 41–48.
18. Fujii T, Jamieson AJ, Solan M, Bagley PM, Priede IG. A Large aggregation of Liparids at 7703 meters and a reappraisal of the abundance and diversity of hadal fish. *Bioscience*. 2010; 60(7): 506–515. <https://doi.org/10.1525/bio.2010.60.7.6> WOS:000279879600010.



19. Wang K, Shen YJ, Yang YZ, Gan XN, Liu GC, Hu K, et al. Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation. *Nat Ecol Evol*. 2019; 3(5): 823–833. <https://doi.org/10.1038/s41559-019-0864-8> WOS:000466498300021. PMID: 30988486
20. Rothschild LJ, Mancinelli RL. Life in extreme environments. *Nature*. 2001; 409(6823): 1092–1101. <https://doi.org/10.1038/35059215> PMID: 11234023.
21. Ma JQ, Pazos IM, Gai F. Microscopic insights into the protein-stabilizing effect of trimethylamine N-oxide (TMAO). *P Natl Acad Sci USA*. 2014; 111(23): 8476–8481. <https://doi.org/10.1073/pnas.1403224111> WOS:000336976000052. PMID: 24912147
22. Subramaniam S, Fletcher C. Trimethylamine N-oxide: breathe new life. *Brit J Pharmacol*. 2018; 175(8): 1344–1353. <https://doi.org/10.1111/bph.13959> WOS:000428313400018. PMID: 28745401
23. Rath S, Heidrich B, Pieper DH, Vital M. Uncovering the trimethylamine-producing bacteria of the human gut microbiota. *Microbiome*. 2017; 5(1): 54. <https://doi.org/10.1186/s40168-017-0271-9> PMID: 28506279; PubMed Central PMCID: PMC5433236.
24. You XX, Bian C, Zan QJ, Xu X, Liu X, Chen JM, et al. Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nat Commun*. 2014; 5: 5594. ARTN 559410.1038/ncomms6594. WOS:000347224500001. <https://doi.org/10.1038/ncomms6594> PMID: 25463417
25. Yang JX, Chen XL, Bai J, Fang DM, Qiu Y, Jiang WS, et al. The *Sinocyclocheilus* cavefish genome provides insights into cave adaptation. *Bmc Biol*. 2016; 14: 1. ARTN 110.1186/s12915-015-0223-4. WOS:000367435000001. <https://doi.org/10.1186/s12915-015-0223-4> PMID: 26728391
26. Niimura Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol Evol*. 2009; 1: 34–44. <https://doi.org/10.1093/gbe/evp003> PMID: 20333175; PubMed Central PMCID: PMC2817399.
27. Gerring ME, Popp BN, Linley TD, Jamieson AJ, Drazen JC. Comparative feeding ecology of abyssal and hadal fishes through stomach content and amino acid isotope analysis. *Deep Sea Research Part I: Oceanographic Research Papers*. 2017; 121: 110–120. <https://doi.org/10.1016/j.dsr.2017.01.003>
28. Partridge JC, Douglas RH, Marshall NJ, Chung WS, Jordan TM, Wagner HJ. Reflecting optics in the diverticular eye of a deep-sea barreleye fish (*Rhynchohyalus natalensis*). *Proc Biol Sci*. 2014; 281(1782): 20133223. <https://doi.org/10.1098/rspb.2013.3223> PMID: 24648222; PubMed Central PMCID: PMC3973263.
29. Jamieson AJ, Fujii T, Solan M, Matsumoto AK, Bagley PM, Priede IG. Liparid and macrourid fishes of the hadal zone: *in situ* observations of activity and feeding behaviour. *Proc Biol Sci*. 2009; 276(1659): 1037–1045. <https://doi.org/10.1098/rspb.2008.1670> PMID: 19129104; PubMed Central PMCID: PMC2679086.
30. Ao J, Mu Y, Xiang LX, Fan D, Feng M, Zhang S, et al. Genome sequencing of the perciform fish *Larimichthys crocea* provides insights into molecular and genetic mechanisms of stress adaptation. *PLoS Genet*. 2015; 11(4): e1005118. <https://doi.org/10.1371/journal.pgen.1005118> PMID: 25835551; PubMed Central PMCID: PMC4383535.
31. Lv YY, Li YP, Li J, Bian C, Qin CJ, Shi Q. A comparative genomics study on the molecular evolution of serotonin/melatonin biosynthesizing enzymes in vertebrates. *Front Mol Biosci*. 2020; 7: 11. ARTN 1110.3389/fmolb.2020.00011. WOS:000556777700001.
32. Blankenship LE, Yayanos AA, Cadien DB, Levin LA. Vertical zonation patterns of scavenging amphipods from the Hadal zone of the Tonga and Kermadec Trenches. *Deep-Sea Res Pt I*. 2006; 53(1): 48–61. <https://doi.org/10.1016/j.dsr.2005.09.006> WOS:000235477300004.
33. Jones NJ, Cox R, Thacker J. Isolation and cross-sensitivity of X-ray-sensitive mutants of V79-4 hamster cells. *Mutat Res*. 1987; 183(3): 279–286. [https://doi.org/10.1016/0167-8817\(87\)90011-3](https://doi.org/10.1016/0167-8817(87)90011-3) PMID: 3106801.
34. Mazina OM, Keskin H, Hanamshet K, Storici F, Mazin AV. Rad52 inverse strand exchange drives RNA-templated DNA double-strand break repair. *Mol Cell*. 2017; 67(1): 19–29. <https://doi.org/10.1016/j.molcel.2017.05.019> PMID: 28602639; PubMed Central PMCID: PMC5547995.
35. Sierant ML, Davey SK. Identification and characterization of a novel nuclear structure containing members of the homologous recombination and DNA damage response pathways. *Cancer Genet*. 2018; 228–229: 98–109. <https://doi.org/10.1016/j.cancergen.2018.10.003> PMID: 30553479.
36. Mozhaev VV, Heremans K, Frank J, Masson P, Balny C. High pressure effects on protein structure and function. *Proteins: Structure, Function, and Bioinformatics*. 1996; 24(1): 81–91. [https://doi.org/10.1002/\(SICI\)1097-0134\(199601\)24:1<81::AID-PROT6>3.0.CO;2-R](https://doi.org/10.1002/(SICI)1097-0134(199601)24:1<81::AID-PROT6>3.0.CO;2-R) PMID: 8628735
37. Liu R, Liu J, Zhang H. Positive selection analysis provides insights into the deep-sea adaptation of a hadal sea cucumber (*Paelopatides* sp.) to the Mariana Trench. *Journal of Oceanology and Limnology*. 2021; 39(1): 266–281. <https://doi.org/10.1007/s00343-020-0241-0>

38. Sullivan MR, Bernstein KA. RAD-ical new insights into RAD51 regulation. *Genes-Basel*. 2018; 9(12): 629. ARTN 62910.3390/genes9120629. WOS:000454717800063. <https://doi.org/10.3390/genes9120629> PMID: 30551670
39. Downing AB, Wallace GT, Yancey PH. Organic osmolytes of amphipods from littoral to hadal zones: Increases with depth in trimethylamine N-oxide, scyllo-inositol and other potential pressure counteractants. *Deep-Sea Res Pt I*. 2018; 138: 1–10. <https://doi.org/10.1016/j.dsr.2018.05.008> WOS:000445985100001.
40. Yancey PH, Blake WR, Conley J. Unusual organic osmolytes in deep-sea animals: adaptations to hydrostatic pressure and other perturbants. *Comp Biochem Physiol A Mol Integr Physiol*. 2002; 133(3): 667–676. [https://doi.org/10.1016/s1095-6433\(02\)00182-4](https://doi.org/10.1016/s1095-6433(02)00182-4) PMID: 12443924.
41. Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrlund TF, Gupta V, et al. Spider genomes provide insight into composition and evolution of venom and silk. *Nature communications*. 2014; 5: 3765. <https://doi.org/10.1038/ncomms4765> PMID: 24801114; PubMed Central PMCID: PMC4273655.
42. Becker S, Hanner R, Steinke D. Five years of FISH-BOL: Brief status report. *Mitochondrial DNA*. 2011; 22: 3–9. <https://doi.org/10.3109/19401736.2010.535528> WOS:000295726100002. PMID: 21271850
43. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*. 2013; 30(12): 2725–2729. <https://doi.org/10.1093/molbev/mst197> WOS:000327793000019. PMID: 24132122
44. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*. 1980; 16(2): 111–120. <https://doi.org/10.1007/BF01731581> PMID: 7463489
45. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
46. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018; 34(17): 884–890. <https://doi.org/10.1093/bioinformatics/bty560> WOS:000444317200035. PMID: 30423086
47. Binghang L, Shi Y, Yuan J, Galaxy Y, Zhang H, Li N, et al. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. 2013; arXiv:1308.2012.
48. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*. 2014; 24(8): 1384–1395. <https://doi.org/10.1101/gr.170720.113> PMID: 24755901; PubMed Central PMCID: PMC4120091.
49. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*. 2017; 27(5): 722–736. <https://doi.org/10.1101/gr.215087.116> PMID: 28298431; PubMed Central PMCID: PMC5411767.
50. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*. 2015; 12(8): 780–786. <https://doi.org/10.1038/nmeth.3454> PMID: 26121404; PubMed Central PMCID: PMC4646949.
51. Walker BJ, 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; 9(11): e112963. <https://doi.org/10.1371/journal.pone.0112963> PMID: 25409509; PubMed Central PMCID: PMC4237348.
52. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics*. 2014; 15(1): 211. <https://doi.org/10.1186/1471-2105-15-211> PMID: 24950923
53. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS one*. 2012; 7(11): e47768. <https://doi.org/10.1371/journal.pone.0047768> PMID: 23185243; PubMed Central PMCID: PMC3504050.
54. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015; 31(19): 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351> PMID: 26059717.
55. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007; 23(9): 1061–1067. <https://doi.org/10.1093/bioinformatics/btm071> PMID: 17332020.
56. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25(14): 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324> PMID: 19451168; PubMed Central PMCID: PMC2705234.

57. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics / editorial board, Andreas D Baxevanis et al. 2004; Chapter 4: Unit 4 10. <https://doi.org/10.1002/0471250953.bi0410s05> PMID: 18428725.
58. Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research. 2007; 35(Web Server issue): W265–268. <https://doi.org/10.1093/nar/gkm286> PMID: 17485477; PubMed Central PMCID: PMC1933203.
59. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005; 21 Suppl 1: i351–358. <https://doi.org/10.1093/bioinformatics/bti1018> PMID: 15961478.
60. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research. 1999; 27(2): 573–580. <https://doi.org/10.1093/nar/27.2.573> PMID: 9862982; PubMed Central PMCID: PMC148217.
61. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013; 29(22): 2933–2935. <https://doi.org/10.1093/bioinformatics/btt509> WOS:000326643600018. PMID: 24008419
62. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Research. 2018; 46(D1): D335–D342. <https://doi.org/10.1093/nar/gkx1038> WOS:000419550700051. PMID: 29112718
63. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic acids research. 2006; 34 (Web Server issue): W435–439. <https://doi.org/10.1093/nar/gkl200> PMID: 16845043; PubMed Central PMCID: PMC1538822.
64. Salamov AA, Solovyev VV. *Ab initio* gene finding in Drosophila genomic DNA. Genome research. 2000; 10(4): 516–522. <https://doi.org/10.1101/gr.10.4.516> PMID: 10779491; PubMed Central PMCID: PMC310882.
65. Parra G, Blanco E, Guigo R. GeneID in Drosophila. Genome research. 2000; 10(4): 511–515. <https://doi.org/10.1101/gr.10.4.511> PMID: 10779490; PubMed Central PMCID: PMC310871.
66. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. Bioinformatics. 2004; 20(16): 2878–2879. <https://doi.org/10.1093/bioinformatics/bth315> WOS:000225250100057. PMID: 15145805
67. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004; 5: 59. <https://doi.org/10.1186/1471-2105-5-59> PMID: 15144565; PubMed Central PMCID: PMC421630.
68. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997; 25 (17): 3389–3402. <https://doi.org/10.1093/nar/25.17.3389> PMID: 9254694; PubMed Central PMCID: PMC146917.
69. Posada D. jModelTest: phylogenetic model averaging. Molecular biology and evolution. 2008; 25(7): 1253–1256. <https://doi.org/10.1093/molbev/msn083> PMID: 18397919.
70. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome research. 2004; 14(5): 988–995. <https://doi.org/10.1101/gr.1865504> PMID: 15123596; PubMed Central PMCID: PMC479130.
71. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9): 1105–1111. <https://doi.org/10.1093/bioinformatics/btp120> PMID: 19289445; PubMed Central PMCID: PMC2672628.
72. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012; 7(3): 562–578. <https://doi.org/10.1038/nprot.2012.016> PMID: 22383036; PubMed Central PMCID: PMC3334321.
73. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011; 29(7): 644–652. <https://doi.org/10.1038/nbt.1883> PMID: 21572440; PubMed Central PMCID: PMC3571712.
74. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr., Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research. 2003; 31(19): 5654–5666. <https://doi.org/10.1093/nar/gkg770> PMID: 14500829; PubMed Central PMCID: PMC206470.
75. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. Bmc Bioinformatics. 2011; 12: 323. Artn 32310.1186/1471-2105-12-323. WOS:000294361700001. <https://doi.org/10.1186/1471-2105-12-323> PMID: 21816040
76. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology. 2008; 9(1): R7. <https://doi.org/10.1186/gb-2008-9-1-r7> PMID: 18190707; PubMed Central PMCID: PMC2395244.

77. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*. 2000; 16(12): 1145–1150. 11159333. <https://doi.org/10.1093/bioinformatics/16.12.1145> PMID: 11159333
78. Li L, Stoekert CJ Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*. 2003; 13(9): 2178–2189. <https://doi.org/10.1101/gr.1224503> PMID: 12952885; PubMed Central PMCID: PMC403725.
79. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution*. 2015; 1(1): vev003. <https://doi.org/10.1093/ve/vev003> PMID: 27774277; PMCID: PMC5014473.
80. Rozewicki J, Li SL, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research*. 2019; 47(W1): W5–W10. <https://doi.org/10.1093/nar/gkz342> WOS:000475901600002. PMID: 31062021
81. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*. 2011; 27(8): 1164–1165. <https://doi.org/10.1093/bioinformatics/btr088> PMID: 21335321
82. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006; 22(21): 2688–2690. <https://doi.org/10.1093/bioinformatics/btl446> PMID: 16928733.
83. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. 2018; 19(6): 153. <https://doi.org/10.1186/s12859-018-2129-y> PMID: 29745866.
84. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007; 24(8): 1586–1591. <https://doi.org/10.1093/molbev/msm088> PMID: 17483113.
85. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*. 2015; 32(4): 835–845. <https://doi.org/10.1093/molbev/msv037> PMID: 25739733; PubMed Central PMCID: PMC4379413.
86. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 2013; 496(7446): 498–503. <https://doi.org/10.1038/nature12111> WOS:000329441500045. PMID: 23594743
87. Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc Natl Acad Sci U S A*. 2018; 115(24): 6249–6254. Epub 2018/05/16. <https://doi.org/10.1073/pnas.1719358115> PMID: 29760103; PubMed Central PMCID: PMC6004478.
88. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, et al. ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem Biophys Res Commun*. 2012; 419(4): 779–781. <https://doi.org/10.1016/j.bbrc.2012.02.101> WOS:000302335800033. PMID: 22390928
89. Zhang JZ, Nielsen R, Yang ZH. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 2005; 22(12): 2472–2479. <https://doi.org/10.1093/molbev/msi237> WOS:000233361500014. PMID: 16107592
90. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*. 1995; 57(1): 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x> WOS:A1995QE45300017.
91. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005; 21(18): 3674–3676. <https://doi.org/10.1093/bioinformatics/bti610> PMID: 16081474
92. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*. 2006; 22(10): 1269–1271. <https://doi.org/10.1093/bioinformatics/btl097> WOS:000237319300018. PMID: 16543274
93. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. The evolution of mammalian gene families. *PloS one*. 2006; 1: e85. <https://doi.org/10.1371/journal.pone.0000085> PMID: 17183716; PubMed Central PMCID: PMC1762380.
94. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016; 44(D1): D279–285. <https://doi.org/10.1093/nar/gkv1344> PMID: 26673716; PubMed Central PMCID: PMC4702930.
95. Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *Bmc Biol*. 2006; 4: 41. Artn 410.1186/1741-7007-4-41. WOS:000243655200001. <https://doi.org/10.1186/1741-7007-4-41> PMID: 17156431

96. Price MN, Dehal PS, Arkin AP. FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One*. 2010; 5(3): e9490. ARTN e949010.1371/journal.pone.0009490. WOS:000275328800002. <https://doi.org/10.1371/journal.pone.0009490> PMID: 20224823
97. Musilova Z, Cortesi F, Matschiner M, Davies WIL, Patel JS, Stieb SM, et al. Vision using multiple distinct rod opsins in deep-sea fishes. *Science*. 2019; 364(6440): 588–892. <https://doi.org/10.1126/science.aav4632> WOS:000467631800043. PMID: 31073066
98. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011; 27(21): 2957–2963. <https://doi.org/10.1093/bioinformatics/btr507> WOS:000296099300005. PMID: 21903629
99. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. 2010; 7(5): 335–336. <https://doi.org/10.1038/nmeth.f.303> WOS:000277175100003. PMID: 20383131
100. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*. 2007; 35(21): 7188–7196. <https://doi.org/10.1093/nar/gkm864> WOS:000251868800024. PMID: 17947321