Chromosome-level genome assembly of the endangered humphead wrasse Cheilinus undulates insight into unexpected expansion of opsin genes in fishes

liu dong[1], Xinyang Wang[1], hongyi guo[2], Xuguang Zhang[1], Ming Zhang[1], and wenqiao tang[2]

[1]Affiliation not available
[2]Shanghai Ocean University

September 11, 2020

Abstract

Wrasses are distributed worldwide in coral reef environments, and display a specialized feature of paired pharyngeal bones united into a single jawbone. Among wrasses, Cheilinus undulatus is an endangered species with high economic and ecological values. Here, we present genome assembly of C. undulatus, using Illumina, Nanopore, and Hi-C sequencing. The 1.17 Gb genome was generated from 328 contigs with an N50 length of 16.5 Mb, and anchored to 24 chromosomes. A total of 22,218 genes were functionally annotated, and 96.36% of BUSCO genes were completely represented. Transcriptomic analyses showed to express 96.79% of the predicted gene. Transposons were most abundant, accounting for 39.88% of genome, with low divergence, owing to evolve with close species approximately 58.37 million years ago, and 560/1,848 gene families were expanded and contracted in the reconstructed phylogeny, respectively. Additionally, 46 genes underwent positive selection. Comparative genomic analyses with other fishes revealed unexpected expansion of opsins SWS2, LWS1, and Rh2, showing single-gene expansion up to five copies in tandem arrays. Gene conversion was responsible for abundance of opsin specific for C. undulatus, and the uneven distribution of transposons in opsin windows and adjacent windows probably contributed to gene conversion, providing gene function fluidity. Divergence of opsin expression in tissues indicated alternative adaptations of the increased opsin copies for the visual foraging and sexual behavior. Genome sequencing of the humphead wrasse provides valuable resources for future investigation of conservation, evolution, and functional morphology of fishes.

Keywords

Cheilinus undulatus , transcriptome, visual gene, gene conversion, transposon

1 Introduction

Adaptation of coral reef fish plays an important role in sustaining marine ecological environments. The family Labridae presents a unique opportunity to gain insight into adaptation. Species of Labridae originated in the late Cretaceous to early Paleogene periods (Alfaro *et al.* 2018) and quickly diversified into over 519 species in 71 genera with an extensive variety of inter- and intra- specific color, morphs, body shapes, and feeding behavior to adapt to various reef environments (Liu D *et al.*2019). In the feeding apparatus, the paired pharyngeal bones are united into a single jawbone, which is derived from a pair of gill arch bones, whereas the other widespread fishes display left and right separated pharyngeal bones (Cowman *et al.* 2009). The united pharyngeal bones allowed labrid fish to generate a great bite force for efficient capture of prey (Wainwright *et al.* 2012). The earliest Labridae fossils demonstrate this pharyngeal apparatus (Bannikov & Sorbini 1990). Among labrid fishes, the humphead wrasse, *Cheilinus undulatus* Rüppell, is an endangered species found on coral reefs and inshore habitats and is distributed in much of the tropical Indo-Pacific Ocean. Moreover, it is one of the most valued and high-priced fish (Russell 2004). *C. undulatus* , one of the

few predators of sea hares, boxfish, and starfish, controls excess reproduction of such toxic animals in coral reef environments, maintaining the stability of reef ecology (Sadovy 1998). Therefore, international trade has been limited to conserve this species (Sadovy *et al.* 2003). This species has been listed as a vulnerable species in the IUCN 1996 Red Data Book and a threatened species in the IUCN 2001 Red List (Donaldson & Sadovy 2001).*C. undulatus* is characterized by several prominent features, including a large hump on the forehead of adult individuals, large fleshy lips, and a pair of distinctive lines running through the eyes. Body color varies at different developmental stages. *C. undulatus*is the largest member of the family Labridae, with a maximum size of 2.3 m in length and over 190 kg in weight (Graham *et al.* 2015).

C. undulatus adults inhabit steep outer reef slopes and benthal at 2-60 m, whereas juveniles are typically found in shallower waters adjacent to coral reefs (Sadovy *et al.* 2003). Little is known about the mechanism underlying the habitat change related to its diet, probably because the whole genome is unknown to date, the genetic architecture could not be provided, and there may be associations with genes coding for visual, olfactory, and feeding parameters. In morphological evolution, the specialized pharyngeal jaw apparatus functions chiefly to collect, manipulate, and transport food into the esophagus. Meanwhile, visual sensitivity could be useful for fish to detect potential prey through the water column. Therefore, *C. undulatus* must have co-evolved a set of visual adaptations for food gathering; however, this remains to be answered. A particularly widespread and well-studied example of this adaptation is the expression of opsin genes. For example, in rainbow trout, the short-wavelength sensitive 1 (*SWS1*) gene may be nonfunctional in adults, but functions in juveniles for foraging zooplankton, which is an important developmental factor (Cheng & Flamarique 2007). Interestingly, diverse expression of opsin genes provides alternative mechanisms for feeding ecology of Labrid fish (Phillips *et al.* 2016). Opsins in fish are keys to the successful colonization of habitats, ranging from the dark deep sea to clear mountain streams (Cortesi *et al.* 2015).

Fish possess five opsins composed of a monophyletic gene family, including one rhodopsin (*Rh1*), *SWS1* , *SWS2* , one middle-wavelength sensitive (*Rh2*), and one long-wavelength sensitive (*LWS*) opsin gene, with a total of five subfamilies that are sensitive to dim vision, ultraviolet, blue, green, and red wavelengths, respectively (Collin *et al.* 2003). Synteny analysis of opsin genes indicated that a local duplication produced *LWS*and *SWS* ; subsequently, two rounds of whole-genome duplication expanded visual opsin into five subfamilies in early vertebrates (Lagman*et al.* 2014). A five-gene repertoire of opsin can be found in the lamprey (*Geotria australis*) without jaws, suggesting that the opsin gene is the ancestral state in jaws (Davies *et al.* 2007). The majority of ray-finned fishes display several copies within each opsin subfamily due to tandem duplications or whole-genome duplication events (Cortesi *et al.* 2015; Rennison *et al.* 2012).*LWS* and *SWS2* duplications in Cypininae and *Rh2*duplication in salmonids were regarded as a consequence of tetraploidy (Lin *et al.* 2017). Tandem duplication is a major contributor to*LWS* subfamily amplification (Rennison *et al.* 2012). However, little is known about the molecular mechanism of opsin tandem duplication. Opsin duplicates could be divergent or display loss of function. Color sensitivity may have been restored through gene duplications (Sharkey *et al.* 2017) or inactivation of one opsin, resulting in retinal monochromacy (Springer *et al.* 2016). The duplicates of opsin gains and losses are believed to correlate with the evolutionary adaptation of fish under different living environments (Lin*et al.* 2017). Opsin gene repertoires in deep-water fish differ from those living closer to the surface, and *LWS* genes are lost in some deep-water species (Rennison *et al.* 2012). Such events dictate whether fish are successful in catching prey or escaping from predators (Phillips *et al.* 2016).

C. undulatus is an ideal candidate for the investigation of opsin evolution in coral reef fish based on the visual system and the united pharyngeal bones. However, a genome with chromosomal assembly of*C. undulatus* has not been reported. To our knowledge, the mitochondrial genome (Qi *et al.* 2013) and a few transcriptomes (Liu H *et al.* 2019) have been reported for humphead wrasse. From an evolutionary perspective, genomic resources of *C. undulatus*provide insight into the mechanism of the visual system for food foraging. In this study, we present the first genome assembly at the chromosomal level for endangered humphead wrasse using Illumina short reads, Nanopore long-read DNA sequencing platform, Hi-C technologies, and a genome assembly strategy. In comparison with other known fish genomes, we found that *C. undulatus* has five *LWS1*genes, four *SWS2* genes, and five *Rh2* genes, the most reported number of any fish yet. The multiple

genes were initially produced via whole-genome duplication, subsequently expanded by gene conversion, while transposons contributed to opsin gene conversion. PAML analyses showed positive selection sites in *Rh2* genes. RNA sequencing (RNA-seq) analyses showed variation in opsin expression. Our results indicate that the sudden increase in opsin copies may play an important role in prey strategy, behavior ecology, sexual change, and evolution of this species. We believe that the annotated draft genome assembly will serve as a resource for future studies of ecology and conservation of the humphead wrasse.

2 Materials and Methods

2.1 Sample collection

A wild female *C. undulatus*(Fig. 1), caught in Guangzhou, Guangdong province, China, was used for genome sequencing and assembly. The fish was determined to be 4 years old, based on annuli otolith interpretation. The living fish was transported to the laboratory. The brain, muscle, liver, spleen, olfactory organ, gonad, and retina tissues of the fish were collected, quickly rinsed with 1× phosphate buffered saline (PBS) solution, and then frozen in liquid nitrogen for 24 h. All samples were stored at -80 degC before sample preparation.

2.2 DNA extraction and genome size estimation

High-quality DNA was extracted from fresh muscle tissues using DNeasy Blood & Tissue Kits (Qiagen, Halden, Germany). The genome size of *C. undulatus* was estimated based on Illumina DNA sequencing technology, as performed in a previous study (Xiao *et al.* 2019). In brief, DNA was randomly sheared to 300–500 bp fragments using Covaris 2000, purified, end-repaired, and amplified using PCR. The constructed DNA library was sequenced using the Illumina NovaSeq 6000 platform in 150 PE mode (Illumina Inc., San Diego, CA, USA). After removal of low-quality and redundant reads, the clean reads were obtained for *de novo* assembly to estimate the genome size. All clean reads were subjected to 17-mer frequency distribution analysis. We obtained a k-mer frequency distribution for *C. undulatus* (Fig. S1). The heterozygosity of the genome was not significantly different from the k-mer distribution of *C. undulatus* at the half-expected depth site (Fig. S1). Therefore, we did not perform heterozygosity analysis in the next step. Genome size was calculated using the formula with amendment: $G = N_{k\text{ -mer\_num}} / D_{\text{k\_mer\_depth}}$, where G is the genome size, $N_{\text{k-mer\_num}}$ is the number of$k$ -mers, and D is the k -mer expected depth, as described (Xiao *et al.* 2019).

2.3 Long DNA library construction and sequencing

Long fragments of extracted high-quality DNA from fresh muscle were selected using a Blue Pippin System (Sage Science, MA, USA), with a peak value of 20 kb. After the short reads were reduced, the long fragments were used for nanopore sequencing. The sequencing adapters were ligated to the ends of the long fragments, according to the manufacturer's instructions for the 1D Ligation Sequencing Kit (SQK-LSK109, Nanopore, Oxford, UK). Finally, the 20 kb genomic DNA libraries were quantified using a Qubit 3.0 Fluorometer (Invitrogen, Camarillo, USA). Meanwhile, three 20 kb libraries (three 1D prep) were prepared and sequenced on one flow cell using the Nanopore PromethION DNA sequencer (Oxford Nanopore, Oxford, UK), according to the manufacturer's instructions.

2.4 RNA extraction and sequencing

To estimate the coverage rate of the assembled genome over gene regions and to predict gene models, RNA-seq was performed to generate transcript data for multiple tissues, including the brain, muscle, liver, spleen, olfactory organ, retina, and gonad tissues collected from the same individual. RNA was extracted separately using Trizol Reagent (Invitrogen, Camarillo, USA), and RNA quality was checked using a Nanodrop spectrophotometer (Labtech, Ringmer, UK). The RNA was used to construct the Illumina RNA-seq library as described in a previous study (Zhu *et al.* 2014). These transcript libraries were sequenced using an Illumina NovaSeq 6000 in PE150 mode (Illumina Inc., San Diego, CA, USA).

2.5 Genome assembly

The error of the nanopore clean reads was first corrected using NextDenovo (https://github.com/ Nextomics/NextDenovo), with the seed cutoff set at 22 k. The *C. undulatus* genome was assembled using

SMARTdenovo (https://github.com/ruanjue/smartdenovo). The nanopore- assembled genome was polished in two runs for error-corrected long reads using the Illumina DNA short reads by NextPolish (Hu *et al.* 2020). The genes in the assembled genome were predicted using BUSCO (Simao *et al.* 2015) with the vertebrata_odb9 database. The integrity of the genome was assessed using the Illumina short reads by BWA (Li & Durbin 2010) and the Illumina RNA-seq reads of multiple tissues by hisat2 (Kim *et al.*2015) aligned against the assembled genome. The accuracy rate of single base was validated by SNP calling using FreeBayes (Garrison & Marth 2012).

2.6 Chromosome assembly using Hi-C library

To obtain a chromosomal assembly of*C. undulatus* , the Hi-C technique was applied to obtain the interaction information among contigs, which are strongly dependent on the one-dimensional distance between a pair of loci (Xiao *et al.*2019). The Hi-C library was constructed using 1 g of muscle tissue from the same one individual. The steps involved in the process, as previously described (Xiao *et al.* 2019), include tissue fixation with formaldehyde, lysis, chromatin digestion (DpnII), biotin marking, proximity ligation, DNA purification, physical shearing, and DNA amplification. The Hi-C library was sequenced using the Illumina NovaSeq 6000 platform with PE150 mode. After the low-quality reads were filtered using Fastp (Chen *et al.* 2018) with default parameters, the clean read pairs were mapped to the polished *C. undulatus* genome using Bowtie2 (Langmead & Salzberg 2012) in end-to-end mode. Clean read pairs that did not provide interaction information were excluded by alignment to the sequences at the restriction site of DpnII. With valid interaction information, the contigs from nanopore sequencing of*C. undulatus* were clustered into 24 groups using Lachesis (Burton*et al.* 2013), which were further ordered and oriented into chromosomes.

2.7 Gene functional annotation and genome assembly validation

To build consensus gene models of the *C. undulatus* genome assembly, gene predictions were performed using Augustus (Stanke *et al.* 2006), GeMoMa (Jens *et al.*2016), and PASA (Haas *et al.* 2008), with *de novo* model, homology sequence, and transcript data, respectively. For homology-based prediction, protein sequences of *Danio rerio* , *Gasterosteus aculeatus* , *Labrus bergylta* , *Lateolabrax maculatus* ,*Symphodus melops* , and *Takifugu rubripes* were downloaded from the NCBI database (https://www.ncbi.nlm.nih. gov). All gene models were merged into a gene set, and transposon-including genes were removed using Transposon PSI (http:// transposon- psi. sourceforge.net).

The predicted protein-coding genes were functionally annotated based on two combined methods. First, the SWISS-PROT database (Bairoch *et al.* 2010), the NCBI non-redundant protein (NR) database, the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (Kanehisa *et al.* 2000) were used to annotate the protein-coding genes using BLAST with an e value of 1 x 10^{-5}. GO terms were assigned to genes based on NR annotation information using Blast2GO. Second, we performed functional annotation using InterProScan (Zdobnov & Rolf 2001) to examine motifs, domains, and other signatures in the secondary structure of the protein-coding genes by searching the ProDom, PRINTS, Pfam, SMART, PANTHER, and PROSITE databases in InterPro (Sarah *et al.* 2009). The quality of gene annotation was evaluated by checking the number of expressed genes using Cufflinks (Trapnell *et al.* 2012) with FPKM values >0, based on the Illumina short reads from tissue transcripts.

2.8 Repetitive element annotation

Simple sequence repeats (SSRs) of the *C. undulatus* genome were first identified using GMATo (Wang*et al.* 2013). After SSRs were soft-masked, tandem repetitive elements (TREs) were annotated using Tandem Repeat Finder (Benson 1999) with default parameters. When TREs were soft-masked in the genome assembly, the miniature transposable elements (MITE) were identified using MITE-hunter (Han & Wessler 2010) to produce a *de novo* MITE library, and the long terminal repeat retrotransposons (LTR) were identified using LTR_Finder (Flicek *et al.* 2014) to obtain a*de novo* LTR library. Thereafter, MITE and LTR libraries were merged into a *de novo* transposon library (TE-lib). The genome assembly was hard-masked by TE-lib and annotated to generate a *de novo* RepMod Library (RepM-lib) using RepeatModeler (http://www.repeatmasker. org). Finally, RepeatMasker (http://www.repeatmasker.org) was used to screen

for repeats with homology-based annotation in the TE-lib, RepM-lib, and RepBase databases (Bao *et al.* 2015).

2.9 Non-coding RNA (ncRNA) annotation

ncRNAs can appear anywhere in the genome and are defined as non-coding transcripts that are not translated into proteins (Willingham *et al.* 2005). The ncRNAs of *C. undulatus* genome were identified using BLAST against the Rfam ncRNA database (Griffiths-Jones *et al.* 2005) by aligning with the known rRNA and tRNA. The tRNAscan-SE (Lowe & Eddy 1997) and RNAmmer tools (Karin *et al.* 2007) were used to predict tRNAs and rRNAs in the genome, respectively.

2.10 Gene family identification

The predicted proteomes in the *C. undulatus* genome and those from other genomes of 13 teleost fishes, including ballan wrasse (*L. bergylta*), corkwing wrasse (*S. melops*), nile tilapia (*Oreochromis niloticus*), clownfish (*Amphiprion percula*), zebrafish (*D. rerio*), cave fish (*Sinocyclocheilus anshuiensis*), tongue sole (*Cynoglossus semilaevis*), stickleback (*G. aculeatus*), medaka (*Oryzias latipes*), mudskipper (*Boleophthalmus pectinirostris*), spotted seabass (*L. maculates*), pufferfish (*T. rubripes*), and ghost shark (*Callorhinchus milii*), were filtered to obtain the longest script per gene, subjected to an all-vs-all Blastp (E-value [?]1e-5), and then clustered to identify gene family using OrthoMCL (Li *et al.*2003), with the inflation index set at 1.5 to find orthologs. In the predicted gene repertoires of the compared genomes, orthologs that could not be found were ascribed to species-specific genes.

2.11 Phylogenetic analyses and divergence time

A total of 619 single-copy orthologous gene clusters were extracted from the OrthoMCL clustering results. Protein sequences from each cluster were aligned using Mafft (Katoh & Standley 2013) with default parameters. Protein-aligned sequences were translated into coding sequence (CDS), and the CDS regions were filtered using Gblocks (Castresana 2000). Thereafter, ghost shark was used as an outgroup to construct the phylogenetic tree using RaxML (Stamatakis 2006) with Gtrgamma model and bootstrap of 100. With the calibration divergence times deposited in the Timetree database (Hedges *et al.* 2006), we selected calibration times at four sites, 471.34, 239.84, 98.25, and 28.46 million years ago (Mya), and MCMCTREE in PAML packages (Yang 1997) was used to confirm the estimated divergence time.

2.12 Expansion and contraction of gene families

According to the orthologs obtained by gene family clustering and the phylogenetic tree constructed based on single-copy orthologous genes, the gene families that expanded or contracted in 14 species were analyzed using CAFE (De Bie *et al.* 2006). A random birth and death model was used to predict gene family variations along each lineage of the phylogenetic tree. The P -values were used to determine the significance of each gene family by comparing conditional likelihoods derived from a probabilistic graphical model. Significantly expanded gene families (P values < 0.05) were performed using a GO term enrichment analysis with EnrichPipeline32 (Beissbarth & Speed 2004; Huang da *et al.* 2009).

2.13 Positive selection genes

Proteins from 14 species were subjected to all paired alignment using Blastp (e value <= 1e-5), and orthologous genes were inferred from the aligned results. Positive selection occurs when the number of nonsynonymous substitutions divided by the number of synonymous substitutions for each site (?) is greater than 1. Positive selection is common in amino acid-level changes to determine functional constraints on proteins (Fay & Wu 2003). To identify the positive selection genes of *C. undulatus* , the average ? among orthologous genes was calculated using Codeml in the PAML package (Yang 1997) with the branch-site model. A likelihood ratio test was conducted on each model pair to determine whether there were significant positive selection genes.

2.14 Identification of opsin genes

The zebrafish genome annotation file (Zv10) was downloaded from NCBI RefSeq, and zebrafish opsin genes were subtracted (Table S1) and used as reference sequences. For each of the other 13 fish genomes, the opsin genes were identified as follows: 1) The zebrafish and 13 fish protein sequences were aligned with Pfam database using Hmmer (http://hmmer.org/ download.html) to find conserved motifs of opsin protein sequences. 2) The opsin genes with conserved motifs in 13 fishes were BLASTed against the zebrafish protein sequences (e-value $< 1x10^{-5}$). Only the protein sequences with the best hits to annotated zebrafish opsins were retained. 3) To reduce false negatives, the coding sequences of these protein sequences and the genomic locations of opsin genes were retrieved from their genome annotation files (Table S2), and only genes annotated as opsin genes/light sensitive genes were retained (Table S1), while the coding sequences were used for follow-up studies.

2.15 Synteny and phylogenetic analyses of opsin genes

To examine opsin duplications by both gene synteny and gene trees, we first observed the locations of opsin genes in the studied genome annotations. The gene order and orientation in the syntenic region were defined based on the original genome annotation. We used a sliding window approach to check adjacent genes that appeared in nearby regions. The size of the sliding window was set as the opsin gene plus three upstream and three downstream genes. If none of the adjacent genes on the same chromosome/contig could be found, we reported no adjacent upstream/ downstream gene of the opsin genes. Second, we extracted the coding sequences of opsin genes, and single-exon coding sequences were aligned using Muscle, and gene trees were constructed using MEGAX v10 (Kumar *et al.* 2018) for maximum likelihood methods, and the model GTR was chosen for the likelihood ratio. The reliability of the clades in the gene trees was assessed by bootstrap probabilities computed using 1000 replicates.

2.16 Gene conversion and positive selection sites of opsin genes

To investigate the mechanism of opsin gene expansion via gene conversion, with the duplication events obtained by gene synteny and gene trees, we grouped these opsin genes, and each group included at least three sequences. These groups were merged into one where distance values between groups were less than 0.4, as described (Sawyer 1989). Gene conversion per group was checked using GENECONV (Sawyer 1989). A P value <0.05 indicated statistical significance. To measure the divergence between genes with gene conversion, we obtained ? values along the coding sequences of genes using a sliding window of 30 with a step size of 1 in DnaSP 6 (Rozas *et al.* 2017). To avoid bias toward gene clades, we calculated ? values for per subfamily with more than three genes. Sequence identity of the sequence flanking the local gene was used to test if gene conversion was induced by whole-genome duplication.

2.17 Content of transposon related to opsin gene

To examine if transposons flanking the opsin genes contribute to gene conversion, we used a 100 kb sliding window along the chromosome to statistically determine the gene number and transposon number per window (Perl script). We plotted gene numbers and transposon numbers per window and drew a distribution of gene/transposon (R script). A one-sample Poisson test was used to validate significant differences in the transposon numbers of the opsin window and its flanking window, and a P value <0.01 was regarded as the transposon hotspot region.

2.18 Opsin gene expression

Retina-specific transcriptomes were analyzed for *C. undulatus* to determine the expression of opsin genes. The filtered reads of transcriptomes from retina tissue, together with the reads of brain, muscle, liver, spleen, olfactory organs, and gonad tissues collected from the same individual were mapped against the coding sequences of opsin genes identified above in *C. undulatus* using Blast, and the mapping reads were input into Cufflinks (Trapnell *et al.* 2012) for the estimation of opsin gene expression. The confidence intervals for the estimation of fragments per kilobase of transcript per million mapped reads (FPKM) were calculated using a Bayesian inference method (Trapnell *et al.* 2012). The criterion of expressed abundance was a statistic FPKM value of 0.1-3.75 for lower expression, 3.75-15 for model expression, and more than

15 for high expression (Pertea *et al.* 2015). Meanwhile, the difference in gene expression was identified by DESeq2 (Love *et al.* 2014), using the false discovery rate (FDR) to calculate the P value at a significant level.

3 Results and Discussion

3.1 Genome assembly and quality assessment

The genome size of *C. undulatus* was estimated to be 1.18-1.27 Gb based on 17-mer frequencies (Fig. S1), and the total number of k-mers was approximately 4.2 x 10^{10} using findGSE (Sun*et al.* 2018) and 3.9 x 10^{10} using GenomeScope (Vurture *et al.* 2017) at the k-mer peak with a depth of 33x. We sequenced approximately 49.9 Gb data via Illumina short-read sequencing, and 90.7 Gb data via nanopore long-read sequencing, indicating 77-fold coverage of the genome (Table S3). The low-quality reads and adapter sequences were filtered from raw genome data from nanopore sequencing of three 20-kb libraries, and we obtained 86.4 Gb clean reads with an N50 length of 31.69 kb for the following genome assembly (Table S4). As a result, a total length of 1164.9 Mb and a contig N50 length of 16.4 Mb were obtained for genome assembly of *C. undulatus* . The size of the assembled genome was slightly lower than the genome size estimated by 17-mer analysis. The nanopore-assembled genome was polished in two runs. The final draft genome assembly was 1173.4 Mb from 328 contig number, which reached a high level of continuity with a contig N50 length of 16.5 Mb (Table S5), and the whole-genome average GC content was 42%. The genome of this species is larger than the known genomes of other marine fishes, usually ranging from 366 to approximately 900 Mb (Xiao *et al.* 2019; Xu *et al.* 2018). We evaluated the quality of the assembled *C. undulatus* genome against the BUSCO database, and 96.36% of complete BUSCO genes were found in the assembled genome. Meanwhile, the entire genome was covered by more than 98% of Illumina short reads, and the base accuracy of the genome was more than 99.99% (Table 1). Furthermore, the transcriptome of multiple tissues from Illumina RNA-seq showed high map-read rates from 89.74% to 94.98% (Table S6). Therefore, we have provided thorough genome assembly for *C. undulatus* .

3.2 Chromosome-level genome assembly

We obtained 145.8 Gb raw reads, 124-fold coverage of the genome (Table S3) via Hi-C sequencing at the chromosomal level, which produced 497.8 million total clean read pairs with Q30 of 93.2%, and 380.2 million clean read pairs, accounting for 76.4% of the total clean read pairs, which uniquely mapped the polished *C. undulatus* genome. After exclusion of the clean read pairs that could not provide interaction information, we obtained 320.6 million clean read pairs (64.4%), which provided valid interaction information for chromosome assembly. With the valid interaction information, the contigs were clustered, ordered, and oriented into chromosomes. Finally, 308 contigs with an N50 length of 3.7 Mb were clustered into 24 scaffolds with an N50 length of 51.5 Mb (Table S3), reliably anchored on the 24 chromosomes, and a final genome size of 1173.2 Mb, representing a 99.98% draft genome. The size of the 24 chromosomes ranges from 27.2 Mb to 59.6 Mb (Fig. 2), providing the chromosomal genome assembly for the humphead wrasse.

3.3 Genome annotation

Homology-based methods were used to predict gene models, together with transcriptome data, and we obtained a total of 22,286 protein-coding genes (Table S7). After functional annotation, 22,218 genes of the predicted protein-coding genes were functional, accounting for 99.69% of the total predicted genes (Table S8), and were distributed in chromosome ranges from 460 to 1246 (Fig. 2). Functional annotation in public databases, including KOG, KEGG, NR, SWISS-PROT, and GO, indicated that at least 61.27% (13,654) of the genes displayed homologues in one database (Table S8) and a total of 9,190 genes could be annotated in all databases (Fig. 3A). Compared to seven species with available annotated genomes, no abnormal length distribution of genes, exons, and introns was observed (Fig. 3B).

A total of 21,572 genes expressed in tissue transcripts were obtained based on FPKM values >0, accounting for 96.79% of the total predicted protein-coding genes. When the expression of genes in muscle was used as a criterion and an FDR value [?]0.005, we obtained differentially expressed genes from multiple tissues

(Fig. 4A). We focused on the intersection size between tissues, and there were mostly 965 genes shared by the muscle and spleen (2,509/2,232 genes expressed up/down, Fig. 4B), and the smallest 227 genes shared by muscle and retina (1,632/1,172 genes expressed up/down, Fig. 4B).

Transposons (RNA and DNA types) and simple sequence repeats (SSRs) were identified in the *C. undulatus* genome. We found 540.85 Mb of the repeat sequences, which accounted for 46.07% of the genome, and transposons accounted for 39.88% of the genome (Table 2). A total of 711 ncRNAs, 111 rRNAs, and 2,618 tRNAs were annotated in the *C. undulatus* genome (Table S9). The divergence rates of the transposons were mostly lower than 30% (Fig. 5A), suggesting recent activity and a burst in the genome. In contrast, ray-finned fishes display the highest diversity, such as the zebrafish, which displays 27 transposon super families (Sotero-Caio *et al* . 2017). Transposon activity and diversity are associated with the evolutionary history of species. Zebrafish originated about 230 Mya (Tine *et al.* 2014), whereas*C. undulatus* diverged from a common ancestor with Cheilines around 50 Mya (Cowman *et al.* 2009). In comparison with ten ray-finned fish genomes with annotated transposons, such as zebrafish (Howe *et al.* 2013), spotted sea bass (Shao *et al.* 2018),*Takifugu rubripes* (Aparicio *et al.* 2002), corkwing wrasse (Mattingsdal *et al.* 2018), Nile tilapia (Brawand *et al.* 2014), the orange clownfish (Lehmann *et al.* 2018), *S. anshuiensis* (Yang *et al.* 2016), flatfish (Chen *et al.*2014), and mudskipper (You *et al.* 2014), we found that transposon content contributed to genome size, with larger genomes exhibiting richer transposon content (Fig. 5B). Transposon content is highly present in the genome of *C. undulatus* , suggesting importantly roles of transposon in genomic evolutions.

3.4 Genome evolution

To better understand the evolutionary history of *C. undulatus* , we identified single-copy orthologs by clustering homologous gene sequences from *L. bergylta* , *S. melops* , *O. niloticus* , *A. percula* , *D. rerio* , *S. anshuiensis* , *C. semilaevis* ,*G. aculeatus* , *O. latipes* , *B. pectinirostris* ,*L. maculates* , *T. rubripes* , *and C. milii* . As a result, 619 single-copy genes and 22,286 genes from 15,410 families were identified in *C. undulatus* (Fig. 6A). Next, we used the coding sequences of 619 single-copy genes to construct a phylogenetic tree and determine divergence times. According to the phylogenetic analysis,*C. undulatus* was clustered together with *S. melops* and*L. bergylta* , belonging to the Labrida family. *C. undulatus*diverged from the common ancestor with *G. aculeatus* and *L. maculates* around 92.28 Mya (Fig. 6B). Our results are consistent with major percomorph subclade (Labrida family included) diversification, which occurred approximately 85-100 Mya in the Late Cretaceous period (Alfaro *et al.* 2018). *C. undulatus* diverged from the common ancestor with *L. bergylta* and *S. melops* around 58.37 Mya. This calibration time is believed to be the initial diversification of Labridae; subsequently, the pharyngeal jaw accelerated adaptive radiation of the Labrid ecomorphological diversity (Alfaro *et al.* 2009).

3.5 Expansion and contraction of gene families and positively selected genes

To investigate the adaptive evolution of *C. undulatus* , we estimated the expansion and contraction of gene families. A total of 560 expanded gene families and 1,848 contracted gene families were identified in the *C. undulatus* genome (Fig. 6B). A total of 430 genes belonging to 199 significantly expanded gene families ($P< 0.05$) were subjected to a GO term enrichment analysis. We found that the significantly expanded gene families were mainly associated with oxidation- reduction processes, oxidoreductase activity, and catalytic activity (Fig. S2). According to the branch site of*C. undulatus* in the evolutionary tree, we found that 46 genes were subjected to significantly positive selection ($P<= 0.05$), and these genes function in fatty acid synthase, elongator complex protein 1 (ELP1), and hepatocyte growth factor-like protein, identified by SWISS-PROT for function annotation (Table S10). Interestingly, ELP1 is the largest subunit of the evolutionarily conserved elongator complex, which catalyzes translational elongation, and the loss of function variants leads to protein misfolding and aggregation, which predisposes tumor development (Waszak *et al.*2020). Positively selected ELP1 in *C. undulatus* may play an important role in development, and its impact could be investigated in the future.

3.6 Expansion of opsin genes in ray-finned fishes

To test whether the diversity of opsin genes is useful for behavioral functions, we investigated copy number of opsin genes in 13 well-assembled genomes of ray-finned fishes, representing several visual types, and *C. milli* as an outgroup (Fig. 6B). For example,*L. maculates* usually rely on eye development and the visual system for prey capture (Shao *et al.* 2018). *B. pectinirostris* , an amphibious fish adapted to terrestrial environments, is adapted for aerial vision in order to avoid terrestrial predators (You *et al.* 2014). In contrast, *S. anshuiensis* is a cavefish, and its eyes are completely lost (Yang *et al.* 2016). Opsin genes in the zebrafish genome were used as a sequence reference. We finally identified 122 opsin genes (Table S11), including 117 complete genes, 5 incomplete genes (missing DNA segments less than 100 bp in four, and inserting DNA fragments of 480 bps in one), and sorted into five subfamilies. The *LWS2* and *Rh2-1* genes were lost in these fishes. Genes in the Rh1 subfamily were lost in *G. aculeatus* ; the *Rh1-2* gene was only found in *S. anshuiensis* . *Rh2-2* and three genes were found only in three species. The *SWS1* gene was lost in *S. melops* , *T. rubripes* , *C. semilaevis* , and *B. pectinirostris* . Species lost opsin genes that were not related to that species' visual sensitivity variation.

Synteny analyses of opsin genes, together with evolutionary trees, showed *LWS1* as a single ancient opsin, and duplicated an immediately adjacent component of *SWS2* in evolutionary origins (Fig. 7A), which was a result of the second run of whole-genome duplication events (2R) (Lin *et al.* 2017). A duplicate of the SWS2- LWS1 complexity in different chromosomes of *S. anshuiensis*was believed to be a result of the 4R that occurred in the common ancestor of Cypriniformes (Lin *et al.* 2017). The *SWS2* gene has two duplicated copies in *B. pectinirostris* , *L. maculates* , and *O. niloticus,* and up to four adjacent copies in*C. undulatus* , which could improve visual acuity of these species by gene duplications to adapt to dim-light environments (Yokoyama & Jia 2020). *SWS2* is a single-gene duplication that often acts as a springboard for adaptive diversification in a genomic region (˜ 30 kb), where two additional duplication events occurred to produce up to three *SWS2* genes in some percomorph fish lineages (Cortesi *et al.* 2015). Interestingly, we observed the retention of up to four *SWS2* genes in *C. undulatus* , suggesting an unexpected expansion of the *SWS2* gene in the percomorph group. However, the *SWS2* gene in *L. maculates*showed divergence due to an inserted fragment, and a gene that was located between *SWS2* and *LWS1* could be diversified from the *SWS2* gene in *O. latipes* (Fig. 7A). Our results provide evidence that *SWS2* could diversify after gene duplication, subsequently causing gene functional changes (Porath-Krause *et al.* 2016).

Surprisingly, the *LWS1* gene has displayed single-gene duplications to produce four copies, and a retrotransposed duplicate in*C. undulatus* , whereas other species showed either one *LWS1*gene, or one *LWS1* gene plus a retrotransposed duplicate, such as*B. pectinirostris* and *L. bergylta* (Fig. 7A). Traditionally, expansion events of the *LWS1* gene have not been observed in approximately 100 ray-finned fish genomes (Cortesi *et al.* 2015; Lin *et al.* 2017; Phillips *et al.* 2016; Rennison*et al.* 2012). However, an insect species, *Xenos vesparum* , with compound eyes possessed five unique *LWS* opsin genes, and these *LWS* duplications were used to restore the longer-wavelength sensitivity due to *SWS* gene loss (Sharkey *et al.* 2017).*LWS1* expansions were notably important in *C. undulatus* , with behaviors often guided by visual cues. *C. undulatus*juveniles inhabit sandy inshore regions, shallow reefs, and murky outer river areas to capture zooplankton (Sadovy *et al.* 2003), and the*LWS1* gene is required for the prevalent wavelengths under shallow water conditions, such as depths less than 30 m (Lin *et al.*2017). Therefore, duplication of the *LWS1* gene could increase the ability of juveniles to detect prey against a light background.*LWS1* gene duplications indicated an independent opsin expansion event, rather than whole-genome duplication. *C. undulatus* has 48 chromosomes (Huo *et al.* 2009), suggesting that this species underwent the 3R, a fish-specific genome duplication, similar to that of most percomorph fishes (Cortesi *et al.* 2015). We inferred that*LWS1* expansion occurred in *C. undulatus* after 3R with very few changes in copy number, except for one copy translocated to another chromosome (Fig. 7A).

Synteny analyses of *Rh2* showed that gene expansions have occurred to produce three to five tandem copies in the branches composed of Labridae fishes, including *C. undulatus* , *S. melops* , and*L. bergylta* (Fig. 7B). One copy was translocated to another chromosome in *L. bergylta* , and one copy was diversified to acquire a novel function different from the opsin gene in *S. melops* . In contrast, other species displayed no more than three*Rh2* copies in tandem arrays, except zebrafish. *C. undulatus* has five *Rh2* copies, the most reported of

any fish (Phillips *et al.* 2016). *Rh2* duplications may coincide with eye evolution in which they have played an important role in expanding the photoreceptive capabilities of organisms by opsin copies (Davies *et al.* 2007). Labridae fish are more commonly observed inhabiting offshore habitats along steep outer reef slopes, reef flats, and lagoon reefs to depths of up to 60 m (Sadovy *et al.* 2003). It is reported that marine fish below 50 m possess more *Rh2* genes than those living above 30 m (Lin *et al.* 2017). Green-sensitive Rh2 helps vertebrate species to better discriminate wavelengths in this environment (Yokoyama & Jia 2020). Multiple Rh2 genes imply good visual adaptation in predating sea hares, boxfish, and starfish, which employ color disguises, similar to reef environments. In contrast, the Rh2 gene was reduced to one in *C. semilaevis* and *A. percula* (Fig. 7B), suggesting that these species likely adapted alternative mechanisms instead of gene copy to contribute to visual sensitivity, such as the known opsin sequence tuning sites (Phillips *et al.* 2016).

To detect duplication events in the *SWS2* , *LWS1* , and*Rh2* genes, we constructed phylogenetic trees of these genes. Four duplications of the SWS1 gene were present in *C. undulatus* , and one duplication in *L. maculates* (Fig. S3). As divergence after gene duplication, *B. pectinirostris* and *O. niloticus* did not show *SWS1* gene duplication, although synteny analyses revealed two adjacent genes in one chromosome (Fig. 7A). In the*LWS1* gene tree, gene duplications were observed in two species (Fig. S4). Due to 4R, the *S. anshuiensis* genome showed*LWS1* gene duplications, whereas four gene duplications plus a retrotransposed duplicate was found in *C. undulatus* (Fig. 7A, Fig. S4). Phylogenetic analysis was applied to infer *Rh2* gene duplication, and many species showed duplication events (Fig. S5). In the *C. undulatus* genome, the *Rh2* gene showed one duplication and two copies were separated into different closed clades, which was regarded as diversification after gene duplication. To our knowledge, it is not known whether opsin genes, such as *SWS1* ,*LWS1* , and *Rh2* , expanded their copy numbers in one species. Opsin genes can change copies by genome duplication, gene duplication, and gene conversion (Sawyer 1989). Interestingly, *C. undulatus*showed unexpected opsin copies, revealing multiple genetic mechanisms of opsin expansion.

3.7 Mechanisms of opsin expansion

We determined gene conversion, which could be used to explain opsin gene expansion. Gene conversion is any process that causes a segment of DNA to be copied onto another segment of DNA, and plays an important role in evolution (Guttman & Dykhuizen 1994). We found that there were two gene conversions of the *SWS2* gene in *C. undulatus* , *SWS2* a vs. *SWS2* d (314 bp, $P = 0.016$), and *SWS2* b vs.*SWS2* c (79 bp, $P = 0.044$), one gene conversion of*LWS1* gene, *LWS1* a vs. *LWS1* d (70 bp, $P = 0.047$). Two gene conversions occurred in the *Rh2* gene,*Rh2* a vs. *Rh2* b (96 bp, $P = 0.0006$) and *Rh2* d vs. *Rh2* e (197 bp, $P = 0.0001$), suggesting opsin expansion by gene conversion. We also tested genes with gene conversions by sliding windows using the rate of non-synonymous to synonymous substitutions (?). The ? value was much less than 1.0 for *Rh2* a vs. *Rh2* b, and *Rh2* d vs. *Rh2* e (Fig. 8A), *SWS2* a vs. *SWS2* d (Fig. S6A), *SWS2* b vs. *SWS2* c (Fig. S6B), and *LWS1* a vs. *LWS1* d (Fig. S6C), indicating that the opsin gene underwent purifying selection ($? < 1.0$). To uncover positive selection sites in the opsin gene, we used PAML to find five positive selection codon sites with a probability greater than 95% in the *Rh2* gene (Fig. 8A), but no site was found in the *LWS1*or *SWS2* genes (Table 3). Our results suggest that opsin gene conversions occurred during post-speciation of *C. undulatus* at the evolutionary level.

We then determined whether the sudden increase in opsin copies in*C. undulatus* is the result of an increased rate of local gene expansion events rather than entire genome duplication. In this case, duplicates should share a flanking sequence (Lagman *et al.* 2013), we determined the flank sequence of the local gene where gene conversion occurred in *SWS2* , *LWS1* , and *Rh2* genes. The identifications of the flank sequences were very low, and no more than 48% (Table 4). Our results demonstrated that, after the 3R (Lagman *et al.* 2013), gene conversions have contributed to the number of opsin genes. Besides, the retrotransposed duplicates also gave rise to opsin gene copies. Based on our results, it is difficult to interpret opsin gene expansion based on a single factor. It has been reported that the genomic environment, such as genomic architecture, can affect opsin gene conversion (Sandkam *et al.* 2017).

The activity of transposons can shape genomic architecture (Mat Razali *et al.* 2019), and the *C. undulatus* genome showed a high content of transposons (39.88% of the entire genome); therefore, we determined the

transposon content of a 100 kb window along chromosome 3, owing to opsin gene expansion mainly occurring in this chromosome. We found that the number of transposons in the *LWS1 -SWS2*window (111) was lower than that in the adjacent windows (up 131 and down 130) along the negative strand (Fig. 8B), and the number was significantly different between them ($P = 0.04$). For the*Rh2* gene, the number of transposons in the *Rh2* windows (82) was significantly lower than that in the adjacent windows (up 102 and down 111) (Fig. 8B), with a significant difference ($P = 0.02$ for up and 0.002 for down). The average means of transposons per window is 113 in the negative strand and 116 in the positive strand. The average means of transposons per gene window was 107 in the negative strand and 111 in the positive strand. Transposons play important roles in genome plasticity to adaptive behavior in evolution (Liu *et al.* 2020; Robert *et al.* 2008). It is reasonable to believe that transposons may be ascribed to opsin expansion.

3.8 Divergent expression of opsin genes

Opsin expression plays a key role in facilitating fish ecological adaption and evolution (Hofmann & Carleton 2009). To determine functional changes in the expanded genes, we performed expression analysis of opsin via RNA-seq of the retina tissue of *C. undulatus* . A total of 44,028,650 clean reads were obtained, and 3,849 genes showed higher expression (FPKM > 15), accounting for 17.3% of the total retina-expressed genes. Genes highly expressed in the retina include genes encoding retinal dehydrogenase 5 (FPKM: 15.34), which catalyzes the final step in the biosynthesis of 11-cis retinaldehyde to produce a light-sensitive chromophore (Skorczyk- Werner*et al.* 2015). Genes encoding retinal dehydrogenase 9 (FPKM: 26.98) are capable of converting 9-cis retinal to corresponding retinol with high efficiency. In contrast, the gene encoding retinol dehydrogenase 10 (FPKM: 18.63) converts all-trans-retinol to all-trans-retinal, which plays a profound role in chromophore generation at the level of rhodopsin (Tian *et al.* 2013). The cluster analysis of opsin genes showed divergent expression (Fig. 9).*SWS2* a, *SWS2* d, and *Rh1* b are expressed in the retina, whereas *SWS1* , *SWS2* c, *Rh1* a, and *Rh2* b were expressed in other areas instead of the retina, suggesting functional changes after gene duplication, and acquisition of novel functions (Porath-Krause *et al.* 2016).

It is interesting to note that *LWS1* was not expressed in any tissue. In many fishes, *LWS1* and *SWS2* spectrally reside in a head-to-tail pattern (Mackin *et al.* 2019), while in beetles,*SWS2* was lost recently, and the sensitivity of *SWS2* to blue wavelengths was restored by *LWS1* extra copies (Sharkey*et al.* 2017). It has been reported that labrid species have shown diverse expression of opsin genes in adaption of variable visual sensitivities to drive phenotypic diversity and behavioral ecologies (Phillips *et al.* 2016). A recent study found that the expression of the *LWS1/SWS2* gene is a stochastic event, and exhibits a switch between opsin genes, even copies of which are triggered by the endocrine signal thyroid hormone (Mackin *et al.* 2019). This mechanism of opsin expression is useful to provide alternative adaption of visual cues in the developmental stages of organisms, the environment, and behavior. The alternative expression of opsin genes may be used to adapt to different photic environments for juveniles growing into *C. undulatus* adults, which live in different habitats (Sadovy *et al.* 2003), and opsin expression variation is also highly correlated to feeding strategy in damsels with herbivorous feeders (Stieb *et al.* 2017). Furthermore, *C. undulatus* is a protogynous hermaphrodite, changing sex from female to male around 8-9 years of age (Sadovy *et al.* 2003), and environmental stimuli, such as a unique mating behavior during spawning aggregations, is a primary trigger of sex change (Todd *et al.* 2019). Opsin gene expression divergence is believed to be responsible for detecting and discriminating between mating partners (Sandkam *et al.* 2017). Our results showed that the opsin gene expression pattern plays an important role in visual plasticity, development, behavior, and reproduction.

4 Conclusions

Wrasses are marine fish, and the humphead wrasse is an endangered species that plays an important role in maintaining the stability of reef ecology. Due to the lack of high-quality and high-continuity reference genomes, the understanding and conservation efforts of this species remain limited. In this study, we first present a draft genome assembly of the humphead wrasse generated via Nanopore long read sequencing, achieving a 1173.4 Mb genome with a contig N50 length of 16.5 Mb based on raw reads of 90.7 Gb, a 77-fold coverage of the genome. Using Hi-C sequencing technology, the raw reads that represented a 124-fold

coverage of the genome anchored into 24 chromosomes and produced a genome size of 1173.2 Mb with a contig N50 length of 3.7 Mb and a scaffold N50 length of 51.5 Mb. The genome was annotated with 22,180 functional genes and 97% of the complete BUSCO genes. Transposable elements accounted for 39.88% of the entire genome. Comparisons with other fishes reveal that a larger genome correlates with increased content of transposable elements. A specialized feature of the united jawbone, which increased the ability to manipulate food, allows us to investigate the functional morphology of this species in visual clues. We have found a sudden increase in number of opsins, *SWS2* , *LWS* 1, and*Rh2,* in tandem pattern by comparative genomics, possibly highlighting alternative adaptation after gene duplication. The increased opsin copies were specific for the humphead wrasse, owing to gene conversion, which was contributed by the uneven distribution of transposable elements in a special genome region. The divergent expression of opsin in the retina indicated visual plasticity for function divergence in efficient foraging, transition into adulthood, specific mate-searching behavior, sexual reversal, and reproduction of this species. The chromosome-level genome assembly of humphead wrasse will provide valuable resources to further understand behavior, gene fluidity, and evolution in fishes.

Acknowledgements

This work was supported by the National Key R&D Program of China (No.2018YFD0900802 & 2018YFD0900905) and the Shanghai Universities First-class Disciplines Project of Fisheries.

References

Alfaro ME, Brock CD, Banbury BL, Wainwright PC (2009) Does evolutionary innovation in pharyngeal jaws lead to rapid lineage diversification in labrid fishes? *BMC Evolutionary Biology* **9** (1), 255.

Alfaro ME, Faircloth BC, Harrington RC*, et al.* (2018) Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary.*Nat Ecol Evol* , 2(4):688-696

Aparicio S, Chapman J, Stupka E*, et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.*science* **297** , 1301-1310.

Bairoch A, Bougueleret L, Altairac S, Amendolia V, Zhang J (2010) The Universal Protein Resource (UniProt) 2009. *Nucleic acids research***37** , D169-D174.

Bannikov AF, Sorbini L (1990) Coris bloti, a new genus and species of labrid fish (Perciformes, Labroidei) from the Eocene of Monte Bolca,Italy. *Studi E Ricerche Sui Giacimenti Terziari Di Bolca* , 133-148.

Bao W, Kojima K, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* **6** , 11.

Beissbarth T, Speed T (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics***20** , 1464-1465.

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

Brawand D, Wagner CE, Li YI*, et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature***513** , 375-381.

Burton JN, Adey A, Patwardhan RP*, et al.* (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* **31** , 1119-1125.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol***17** , 540-552.

Chen S, Zhang G, Shao C, QF H (2014) Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* **46** , 253-260.

Chen S, Zhou Y, Chen Y, Jia G (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34** , i884-i890.

Cheng C, Flamarique I (2007) Chromatic organization of cone photoreceptors in the retina of rainbow trout: single cones irreversibly switch from UV (SWS1) to blue (SWS2) light sensitive opsin during natural development. *J Exp Biol* **210** , 4123-4135.

Collin SP, Knight MA, Davies WL*, et al.* (2003) Ancient colour vision: multiple opsin genes in the ancestral vertebrates. *Curr Biol* **13** , R864-865.

Cortesi F, Musilova Z, Stieb S, Hart N (2015) Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. *PNAS***112** , 1493-1498.

Cowman PF, Bellwood DR, Herwerden LV (2009) Dating the evolutionary origins of wrasse lineages (Labridae) and the rise of trophic novelty on coral reefs. *Molecular phylogenetics and evolution* **52** , 621-631.

Davies WL, Cowing JA, Carvalho LS*, et al.* (2007) Functional characterization, tuning, and regulation of visual pigment gene expression in an anadromous lamprey. *FASEB J* **21** , 2713-2724.

De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics***22** , 1269-1271.

Donaldson TJ, Sadovy Y (2001) Threatened fishes of the world: Cheilinus undulatus Ruppell, 1835 (Labridae). *Environmental Biology of Fishes* **62** , 428.

Fay JC, Wu CI (2003) Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet***4** , 213-235.

Flicek P, Amode M, Barrell D, Beal K, Billis K (2014) Ensembl 2014.*Nucleic acids research* **42** , 749-755.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *Quantitative Biology* , 1-9.

Graham KS, Boggs CH, DeMartini EE, Schroeder RE, Trianni MS (2015) Status review report: humphead wrasse (Cheilinus undulatus), pp. 1-123, Report to National Marine Fisheries Service, Office of Protected Resources.

Griffiths-Jones S, Moxon S, Marshall M*, et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research* **33** , 121-124.

Guttman D, Dykhuizen DE (1994) Clonal divergence in Escherichia coli as a result of recombination, not mutation. *science* **266** , 1380-1383.

Haas BJ, Salzberg SL, Zhu W, Pertea... M (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology***9** , 7.

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences.*Nucleic acids research* **38** , e199.

Hedges S, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22** , 2971-2972.

Hofmann CM, Carleton KL (2009) Gene duplication and differential gene expression play an important role in the diversification of visual pigments in fish. *Integr Comp Biol* **49** , 630-643.

Howe K, Clark MD, Torroja CF*, et al.* (2013) The zebrafish reference genome sequence and its relationship to the human genome.*Nature* **496** , 498-503.

Hu J, Fan J, Sun Z, Liu S (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics***36** , 2253-2255.

Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37** , 1-13.

Huo R, Zhang B, Chen GH*, et al.* (2009) The karyotype of Cheilinus undulates. *Marine Sciences* **33** , 94-97.

Jens K, Michael W, Erickson JL*, et al.* (2016) Using intron position conservation for homology- based gene prediction. *Nucleic acids research* , e89.

Kanehisa M, Goto S, Kawashima S, Nakaya A, S,, S. K. (2000) KEGG: kyoto encyclopaedia of genes and genomes. *Nucleic acids research***28** , 27-30.

Karin L, Peter H, Andreas RE*, et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* **35** , 3100-3108.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30** , 772-780.

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12** , 357-360.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35** , 1547-1549.

Lagman D, Daza D, Widmark J, Abalo X (2013) The vertebrate ancestral repertoire of visual opsins, transducin alpha subunits and oxytocin/vasopressin receptors was established by duplication of their shared genomic region in the two rounds of early vertebrate genome duplications. *BMC evolutionary biology* **13** , 238.

Lagman D, Daza D, Widmark J, Abalo X, Sundstro"m G (2014) The vertebrate ancestral repertoire of visual opsins, transducin alpha subunits and oxytocin/vasopressin receptors was established by duplication of their shared genomic region in the two rounds of early vertebrate genome duplications. *BMC Evol Biol* **13** , 238.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2.*Nature Methods* **9** , 357-359.

Lehmann R, Lightfoot DJ, Schunter C, Michell CT (2018) Finding Nemo's Genes: A chromosome-scale reference assembly of the genome of the orange clownfish Amphiprion percula. *Mol Ecol Resour* **00** , 1-16.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26** , 589-595.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13** , 2178-2189.

Lin J, Wang F, Li W, Wang T (2017) The rises and falls of opsin genes in 59 ray-finned fish genomes and their implications for environmental adaptation. *Scientific reports* **7** , 15568.

Liu D, Huang X, Tang W (2019) Advances in systematics of the Labridae.*Marine Fisheries* **41** , 107-117.

Liu D, Yang J, Tang W*, et al.* (2020) SINE Retrotransposon variation drives Ecotypic disparity in natural populations of Coilia nasus. *Mob DNA* **11** , 4.

Liu H, Liu J, Yang M, He Y, Wang Y (2019) SSR and SNP Polymorphic Feature Analysis Based on Cheilinus undulatus Transcriptome.*Genomics and Applied Biology* **3** , 1-12.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol***15** , 550.

Lowe TM, Eddy SR (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic acids research***25** , 955-964.

Mackin RD, Frey RA, Gutierrez C*, et al.* (2019) Endocrine regulation of multichromatic color vision. *PNAS* **116** , 16882-16891.

Mat Razali N, Cheah B, Nadarajah K (2019) Transposable Elements Adaptive Role in Genome Plasticity, Pathogenicity and Evolution in Fungal Phytopathogens. *Int J Mol Sci* **20** , 3597.

Mattingsdal M, Jentoft S, Torresen OK*, et al.* (2018) A continuous genome assembly of the corkwing wrasse (Symphodus melops).*Genomics* **110** , 399-403.

Pertea M, Pertea GM, Antonescu CM*, et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.*Nat Biotechnol* **33** , 290-295.

Phillips GA, Carleton KL, Marshall NJ (2016) Multiple Genetic Mechanisms Contribute to Visual Sensitivity Variation in the Labridae. *Mol Biol Evol* **33** , 201-215.

Porath-Krause A, Palrett A, Fagglonato D, Birla B (2016) Structural differences and differential expression among rhabdomeric opsins reveal functional change after gene duplication in the bay scallop, Argopecten irradians (Pectinidae). *BMC evolutionary biology* **16** , 250.

Qi XZ, Yin SW, Luo J, Huo R (2013) Complete mitochondrial genome sequence of the humphead wrasse, Cheilinus undulatus. *Genetics & Molecular Research Gmr* **12** , 1095-1105.

Rennison DJ, Owens GL, Taylor JS (2012) Opsin gene duplication and divergence in ray-finned fish. *Mol Phylogenet Evol* **62** , 986-1008.

Robert VJ, Davis MW, Jorgensen EM, Bessereau JL (2008) Gene conversion and end-joining- repair double-strand breaks in the Caenorhabditis elegans germline. *Genetics* **180** , 673-679.

Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S (2017) DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. *Mol. Biol. Evol.* **34** , 3299-3302.

Russell B (2004) Cheilinus undulatus. *The IUCN Red List of Threatened Species 2004* , e.T4592A11023949.

Sadovy Y (1998) Ciguatera hits Hong Kong live food-fish trade. *SPC Live Reef Fish Information Bulletin* **4** , 51-53.

Sadovy Y, Kulbicki M, Labrosse P*, et al.* (2003) The humphead wrasse, Cheilinus undulatus: Synopsis of a threatened and poorly known giant coral reef fish. *Reviews in Fish Biology & Fisheries***13** , 327-364.

Sandkam BA, Joy JB, Watson CT, Breden F (2017) Genomic Environment Impacts Color Vision Evolution in a Family with Visually Based Sexual Selection. *Genome Biol Evol* **9** , 3100-3107.

Sarah H, Rolf A, Attwood TK*, et al.* (2009) InterPro: the integrative protein signature database. *Nucleic acids research***37** , 211-215.

Sawyer S (1989) Statistical tests for detecting gene conversion.*Mol Biol Evol* **6** , 526-538.

Shao C, Li C, Wang N*, et al.* (2018) Chromosome-level genome assembly of the spotted sea bass, Lateolabrax maculatus.*GigaScience* **7** .

Sharkey CR, Fujimoto MS, P.Lord NP, Shin S (2017) Overcoming the loss of blue sensitivity through opsin duplication in the largest animal group, beetles. *Scientific reports* **7** , 8.

Simao FA, Waterhouse RM, Panagiotis I, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31** , 3210-3212.

Skorczyk-Werner A, Pawlowski P, Michalczuk M*, et al.* (2015) Fundus albipunctatus: review of the literature and report of a novel RDH5 gene mutation affecting the invariant tyrosine (p.Tyr175Phe).*J Appl Genet* **56** , 317-327.

Sotero-Caio C, Platt II RN, Suh A, Ray D (2017) Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol***91** , 161-177.

Springer MS, Emerling CA, Fugate N, Patel R (2016) Inactivation of Cone-Specific Phototransduction Genes in Rod Monochromatic Cetaceans.*Front. Ecol. Evol* **4** , 61.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics***22** , 2688-2690.

Stanke M, Schoffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7** , 62.

Stieb SM, Cortesi F, Sueess L*, et al.* (2017) Why UV vision and red vision are important for damselfish (Pomacentridae): structural and expression variation in opsin genes. *Mol Ecol* **26** , 1323-1342.

Sun H, Ding J, Piednoe M, Schneeberger K (2018) findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **34** , 550-557.

Tian Y, Li T, Sun M*, et al.* (2013) Neurexin regulates visual function via mediating retinoid transport to promote rhodopsin maturation. *Neuron* **77** , 311-322.

Tine M, Kuhl H, Gagnaire P, Louro B (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat Commun* **5** , 5770.

Todd EV, Ortega-Recalde O, Liu H*, et al.* (2019) Stress, novel sex genes, and epigenetic reprogramming orchestrate socially controlled sex change. *Sci Adv* **5** , eaaw7006.

Trapnell C, Roberts A, Goff L*, et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7** , 562-578.

Vurture GW, Sedlazeck FJ, Nattestad M*, et al.* (2017) GenomeScope: fast reference-free genome profiling from short reads.*Bioinformatics* **33** , 2202-2204.

Wainwright P, Smith W, Price S, Tang K (2012) The Evolution of Pharyngognathy: A Phylogenetic and Functional Appraisal of the Pharyngeal Jaw Key Innovation in Labroid Fishes and Beyond.*Systematic Biology* **61** , 1001-1027.

Wang X, Lu P, Luo Z (2013) GMATo: A novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformation***9** , 541-544.

Waszak SM, Robinson GW, Gudenas BL*, et al.* (2020) Germline Elongator mutations in Sonic Hedgehog medulloblastoma. *Nature***580** , 396-401.

Willingham A, Orth A, Batalov S*, et al.* (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT.*science* **309** , 1570-1573.

Xiao Y, Xiao Z, Ma D, Liu J, Li J (2019) Genome sequence of the barred knifejaw Oplegnathus fasciatus (Temminck & Schlegel, 1844): the first chromosome-level draft genome in the family Oplegnathidae.*GigaScience* **8** , 1-8.

Xu S, Xiao S, Zhu S*, et al.* (2018) A draft genome assembly of the Chinese sillago (Sillago sinica), the first reference genome for Sillaginidae fishes. *GigaScience* **7** .

Yang J, Chen X, Bai J, Fang D (2016) The Sinocyclocheilus cavefish genome provides insights into cave adaptation. *BMC Biology***14** .

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13** , 555-556.

Yokoyama S, Jia H (2020) Origin and adaptation of green-sensitive (RH2) pigments in vertebrates. *FEBS Open Bio* **10** , 873-882.

You X, Bian C, Zan Q, Xu X (2014) Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nature Communications* **5** , 5594.

Zdobnov EM, Rolf A (2001) InterProScan – an integration platform for the signature-recognition methods in InterProScan. *Bioinformatics***17** , 847-848.

Zhu G, Wang L, Tang W, Liu D, Yang J (2014) De Novo Transcriptomes of Olfactory Epithelium Reveal the Genes and Pathways for Spawning Migration in Japanese Grenadier Anchovy (Coilia nasus). *PloS one***9** , e103832.

Data Accessibility

The BioProject from genomic sequencing was assigned the NCBI accession number PRJNA622923. The BioSample data were assigned the NCBI accession number SAMN14532944. The raw sequences and genome annotation files were deposited in the NCBI Sequence Read Archive under the accession number SRA7239481.

Author Contributions

Dong Liu wrote the manuscript; Xinyang Wang collected the samples, performed Illunima sequencing, and estimated the genome size; Hongyi Guo took a picture of the fish, determined annuli on otolith, and assembled the genome; Wenqiao Tang designed the projects of this study and performed Nanopore sequencing, and transcript sequencing; Ming Zhang performed the bioinformatic analyses and assessed the assembly quality. Xuguang Zhang carried out the genome annotation and functional gene analysis. All authors read, edited, and approved the final manuscript.

Figure Legends

Figure 1. A picture of *Cheilinus underlatus* used for the genome sequencing

Figure 2. Chromosome size and gene density of*Cheilinus underlatus* . Genes were plotted in 100 kb windows, and chromosomes size was showed on the left axis in Mb.

Figure 3. Gene annotation and length distribution for*Cheilinus undulatus* . **(A)** Gene annotated in the public databases. **(B)** Comparison of gene, exon, and intron length distributions between *C. undulatus* and other six species.

Figure 4 . Express of genes based on the transcriptome of *Cheilinus underlatus* . **(A)** Numbers of the expressed genes muscle shared, and tissue-specific expressed genes.**(B)** Different express of genes in tissue and muscle used as background. mu: muscle; re: retina; li: liver; or: olfactory organ; br: brain; sp: spleen; gl: gonad.

Figure 5 . Divergence and percentage of transposons in the genome of *Cheilinus* ***underlatus*** . **(A)**Divergence distribution of known transposons in *C. underlatus* .**(B)** Pie charts comparing the percentage of transposons presented in a genome. The area of the pie chart noted genome size.

Figure 6. Phylogenetic tree, divergence time and gene families of *Cheilinus underlatus* . **(A)** Gene family comparison among *C. undulatus* and other fishes, and single-copy orthologs were used to construct phylogenetic tree**.(B)** The phylogenetic relationship of *C. underlatus* with other fishes. The red blot in the internal nodes indicated fossil calibration times. All nodes had support values of 100%. Numbers under each species showed gene families that have been expanded (green) and contracted (red) since the split of species from the most recent common ancestor. The numbers near each node correspond to the estimated divergence time of these species. The circled number on the right is the same as picture.

Figure 7. Opsin gene synteny and orientation in genomes from species tree based on 619 single-copy orthologs . **(A)** synteny of SWS2 and LWS1opsin genes in 13 reprsentative species and*Callorhinchus milii* used as outgroup. **(B)** synteny of Rh2 opsin gene in species the same as that of the left. The genomic location of each synteny is summarized in Table S1. The files of the genome annotations could be used in Table S2.

Figure 8. Gene conversions of Rh2 opsin, transposons and genes distribution . **(A)** Gene conversion analyses by sliding window. Pairwise rate of synonymous substitutions between Rh2 copies calculated with a window of 30 and a step size of 1, and vertical dotted lines depicted significantly positive selection sites of Rh2 opsin with P<0.05. **(B)** The numbers in transposons and genes distribution along positive/negative strands in Chromosome 3. The vertical dotted lines depicted locations of opsin genes. X axis depicted a 100 kb window, and transposon numbers in Y axis have been divided by 10.

Figure 9. Cluster analyses of opsin express based on transcriptome of *Cheilinus underlatus*.

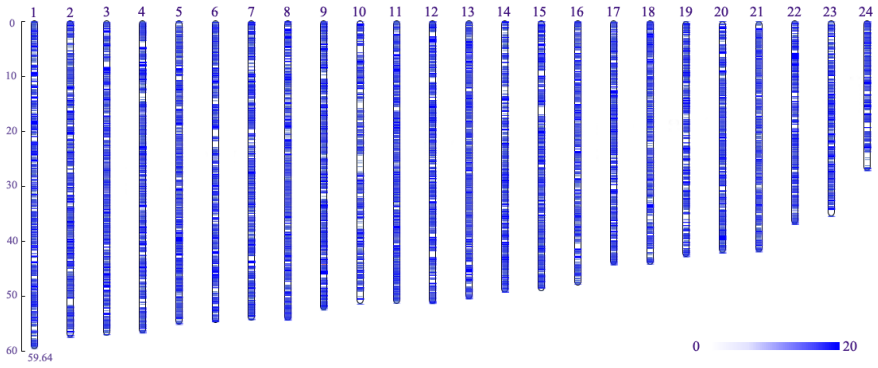
Tables Legends

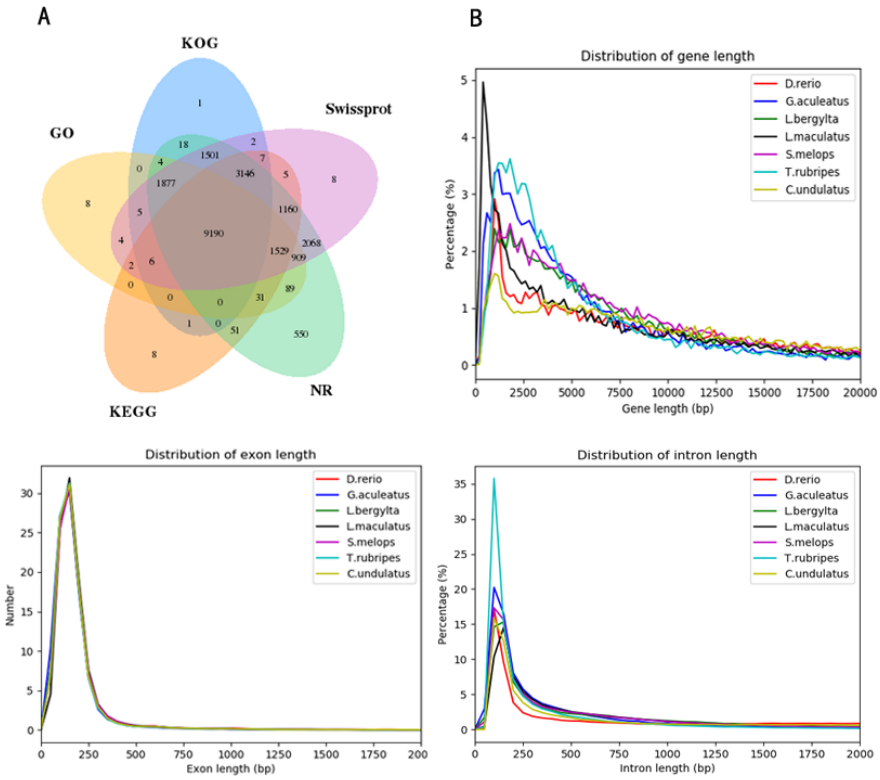
Table 1 Summary of the assembled genome assessed

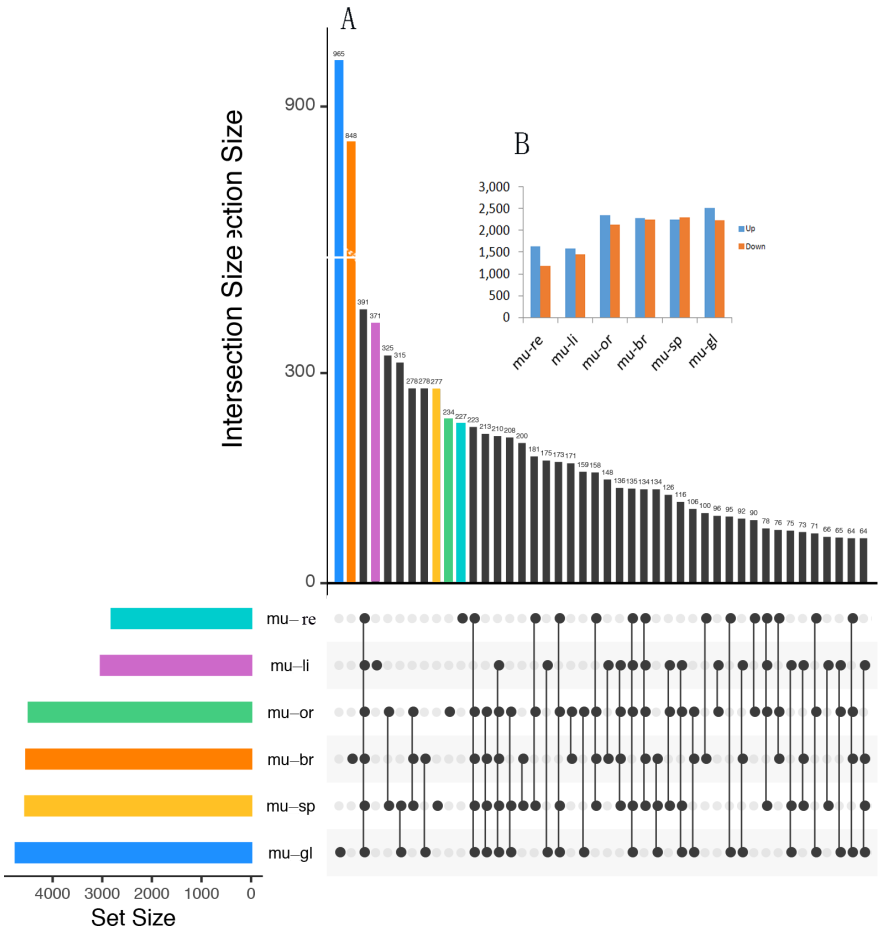
Table 2 Repetitive element annotations in genome of the humphead wrasse

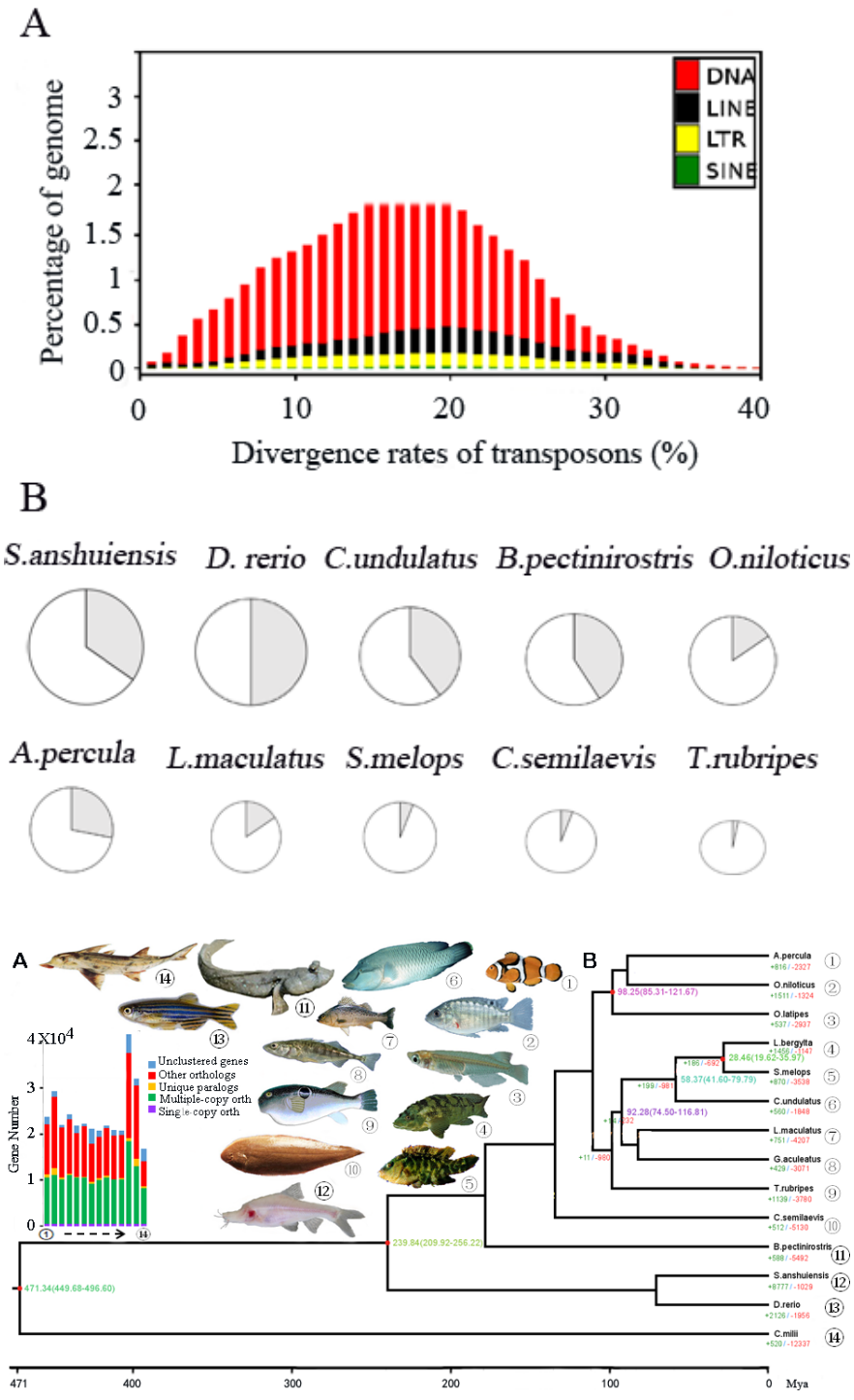
Table 3 Number of positive selection sites of opsin genes

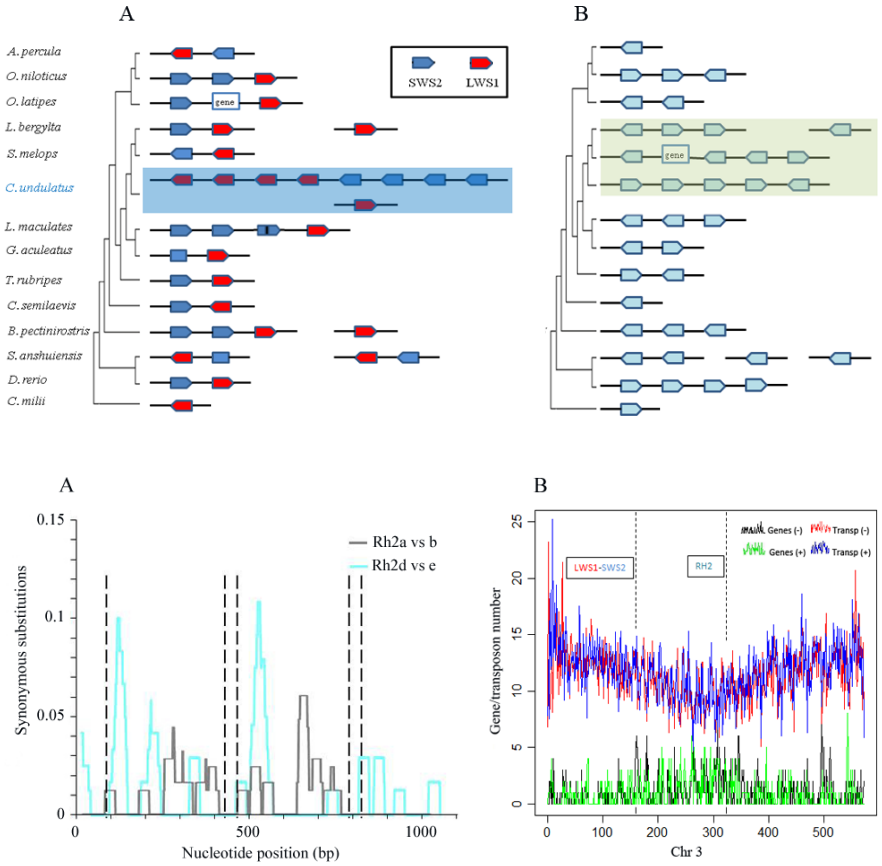
Table 4 Identity of sequence flanking the local gene in gene conversion

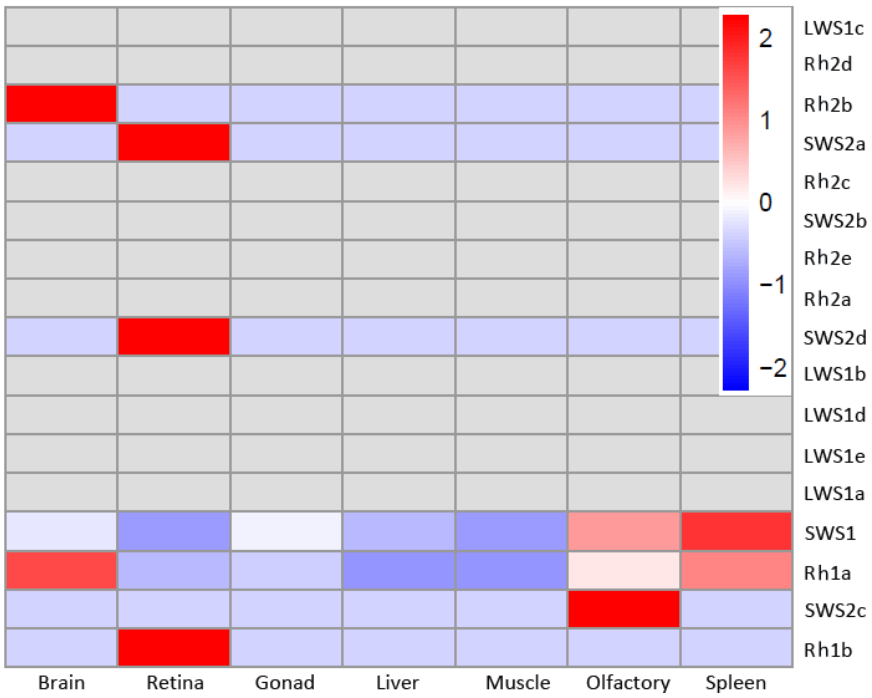












Hosted file

Table 1 Summary of the assembled genome assessed.docx available at https://authorea.com/users/357254/articles/479883-chromosome-level-genome-assembly-of-the-endangered-humphead-wrasse-cheilinus-undulates-insight-into-unexpected-expansion-of-opsin-genes-in-fishes

Hosted file

Table 2 Type of repeats.docx available at https://authorea.com/users/357254/articles/479883-chromosome-level-genome-assembly-of-the-endangered-humphead-wrasse-cheilinus-undulates-insight-into-unexpected-expansion-of-opsin-genes-in-fishes

Hosted file

Table 3 Number of positive selection sites.docx available at https://authorea.com/users/357254/articles/479883-chromosome-level-genome-assembly-of-the-endangered-humphead-wrasse-cheilinus-undulates-insight-into-unexpected-expansion-of-opsin-genes-in-fishes

Hosted file

Table 4 identification of flank sequence.docx available at https://authorea.com/users/357254/articles/479883-chromosome-level-genome-assembly-of-the-endangered-humphead-wrasse-cheilinus-undulates-insight-into-unexpected-expansion-of-opsin-genes-in-fishes