





# Whole-genome sequencing reveals sex determination and liver high-fat storage mechanisms of yellowstripe goby (*Mugilogobius chulae*)

Lei Cai <sup>1,3</sup>✉, Guocheng Liu<sup>2,3</sup>, Yuanzheng Wei<sup>1,3</sup>, Yabing Zhu <sup>2,3</sup>, Jianjun Li<sup>1</sup>, Zongyu Miao<sup>1</sup>, Meili Chen<sup>1</sup>, Zhen Yue <sup>2</sup>, Lujun Yu<sup>1</sup>, Zhensheng Dong<sup>2</sup>, Huixin Ye<sup>1</sup>, Wenjing Sun<sup>2</sup> & Ren Huang <sup>1</sup>✉

As a promising novel marine fish model for future research on marine ecotoxicology as well as an animal model of human disease, the genome information of yellowstripe goby (*Mugilogobius chulae*) remains unknown. Here we report the first annotated chromosome-level reference genome assembly for yellowstripe goby. A 20.67-cM sex determination region was discovered on chromosome 5 and seven potential sex-determining genes were identified. Based on combined genome and transcriptome data, we identified three key lipid metabolic pathways for high-fat accumulation in the liver of yellowstripe goby. The changes in the expression patterns of *MGLL* and *CPT1* at different development stage of the liver, and the expansion of the *ABCA1* gene, innate immune gene *TLR23*, and *TRIM* family genes may help in balancing high-fat storage in hepatocytes and steatohepatitis. These results may provide insights into understanding the molecular mechanisms of sex determination and high-fat storage in the liver of marine fishes.

<sup>1</sup>Guangdong Provincial Key Laboratory of Laboratory Animals, Guangdong Laboratory Animals Monitoring Institute, Guangzhou, China. <sup>2</sup>BGI Genomics, BGI-Shenzhen, Shenzhen, China. <sup>3</sup>These authors contributed equally: Lei Cai, Guocheng Liu, Yuanzheng Wei, Yabing Zhu. ✉email: [cailei17@163.com](mailto:cailei17@163.com); [1649405216@qq.com](mailto:1649405216@qq.com)

Compared with most mammals, sex determination mechanisms in fish have exhibited a high degree of plasticity and complexity<sup>1,2</sup>, which may be related to genetic or environmental factors or both. So far, several sex determination systems have been identified in fishes, including male-heterogametic gonochorism (XY)<sup>3</sup>, female-heterogametic gonochorism (ZW)<sup>4</sup>, hermaphroditism<sup>5</sup>, and environmental dependency<sup>6</sup>. However, the reasons for the evolution of so many diverse sex determination mechanisms and the key factor for transition of different sex determination mechanisms remain unknown<sup>7</sup>. Especially in gobiid fish, one of the largest fish families, comprising more than 2000 species, little information about the sex determination mechanism has been discovered, and no sex determination genes or sex determination regions have been identified to date<sup>8–10</sup>. Considering that gobiid fishes are one of the most important taxa in the marine ecosystem, understanding their sex determination mechanism is of great significance in revealing the adaptive strategies and providing invaluable insights on the evolution of sex determination in teleosts.

Lipids and their constituent fatty acids are major organic constituents in various organisms from worms (*Caenorhabditis elegans*) to humans, and the ability to store fats is conserved<sup>11,12</sup>. Interestingly, starting with the primitive teleosts (jawless vertebrates such as lampreys), the lipid-storing cells have evolved into a tissue that has distinct functions underneath the skin<sup>13</sup>, while the type and sites of fat storage is species-specific in fish and depend on the nutritional state, life-stage, and the physiological state<sup>14,15</sup>. Most fishes, such as zebrafish (*Danio rerio*)<sup>14</sup> and cavefish (*Astyanax mexicanus*)<sup>16</sup>, show deposition of lipids, mainly triglycerides (TAG), in mesentery and viscera. However, in majority of the gobiids, the lipids are only stored in the liver<sup>17–20</sup>. This phenomenon is also exhibited in some other marine fishes, such as pufferfish (*Takifugu rubripes*) and cultured flounder (*Paralichthys olivaceus*)<sup>21</sup>. The special lipid deposition organ in these fish might be an evolutionary adaptation to cope with the typical living environments, such as rapid energy mobilization, migration, or benthic adaptation. This is similar to the situation observed among the Inuits, who display genetic and physiological adaptations to a diet rich in polyunsaturated fatty acids (PUFAs)<sup>22</sup> and stickleback lineages (*Gasterosteus aculeatus* species complex), which have evolved different copy numbers of lipid metabolism-related genes, such as docosahexaenoic acid biosynthesis-related genes, to achieve transitions between marine and freshwater environments<sup>23</sup>. In the majority of gobiids, large quantities of neutral fat are stored in the liver. In most gobiid fishes, fat represents more than 70% of the liver wet weight, and 90% of the total lipids are triglycerides<sup>18</sup>. In humans, the liver fat content is less than 5% of the wet weight<sup>24</sup>. When this percentage is exceeded, non-alcoholic fatty liver disease (NAFLD) can develop, and more than 20% of patients with NAFLD develop non-alcoholic steatohepatitis (NASH)<sup>25</sup> along with inflammation and varying degrees of fibrosis<sup>26</sup>. However, gobiids can maintain a high level of fat storage in the liver lifelong without developing steatohepatitis, suggesting the existence of a specialized mechanism for maintaining a balance between high-fat storage and inflammation in the liver in these fishes. Hence, elucidating the high-fat storage mechanism of gobiids is of great value for understanding the evolution of lipid storage in teleosts, and the pathogenesis of human NASH as well as for promoting aquaculture health in the mariculture industry.

Gobiids (Teleostei, Gobiidae), commonly known as gobies, are a diverse and fascinating group with worldwide distribution<sup>27</sup>, and is one of the most diverse families of vertebrates on earth<sup>28</sup>. Therefore, gobies represent a potential excellent model for adaptation studies. Unfortunately, neither the genome sequence

and phylogenetic relationships of many groups of gobies nor the laboratory breeding and rearing methods are resolved. Only a few gobiid genomes have been sequenced, such as those of round goby (*Neogobius melanostomus*)<sup>8</sup>, mudskippers (*Boleophthalmus pectinirostris*, *Periophthalmodon schlosseri*, *Periophthalmus magnuspinnatus*, *Scartelaos histophorus*)<sup>10</sup>, and sand goby (*Pomatoschistus minutus*)<sup>29</sup>. Currently, only a few small fish species [e.g., zebrafish<sup>30</sup>, Japanese medaka (*Oryzias latipes*)<sup>31</sup>, and platyfish (*Xiphophorus maculatus*)<sup>32</sup>] have been widely used in the laboratory; however, the large majority of these inhabit freshwater environments. Laboratory models of marine species are limited to the three-spined stickleback species (*Gasterosteus aculeatus*)<sup>33</sup>, pufferfish (*Takifugu rubripes*)<sup>34</sup>, and the Atlantic silverside (*Menidia menidia*)<sup>35</sup>. Considering the specific needs of laboratory animals, such as convenient large-scale indoor cultivation and controlled year-round spawning, it is necessary to develop a representative marine fish to supplement the marine model fish species. Yellowstripe goby (*Mugilogobius chulae*) is a representative fish of the Gobiidae family that is widely distributed along the western Pacific coast<sup>36</sup>. This species has a moderate body size (adult body length, 3–5 cm), short sexual maturity period, strong reproductive capacity, short spawning interval, annual reproduction, easy indoor rearing, and easy genetic manipulation<sup>37</sup>. In addition, a Chinese national quality-control standard, including genetic, microorganism, parasite, nutrition, and environment quality control, has been established for yellowstripe goby (draft national standard no. 20091329-T-469)<sup>38</sup>. Hence, the yellowstripe goby is a quality-controlled laboratory fish and a promising novel marine fish model for future research on genetic evolution and marine ecotoxicology as well as an animal model of human disease.

In this study, a 7th generation inbred line of yellowstripe goby was subjected to whole-genome sequencing using Illumina HiSeq and PacBio RSII sequencing platforms. A high-density genetic linkage map was constructed based on single-nucleotide polymorphism (SNP) markers, using restriction site-associated DNA (RAD) sequencing. The constructed genetic map was used to assemble the genome at the chromosomal level. A sex determination region was identified by quantitative trait locus analysis. Further, based on combined genome and transcriptome data, we aimed to identify the potential mechanism underlying neutral fat storage and lipid homeostasis in the liver of yellowstripe goby.

## Results

**Genome assembly and annotation.** Using a 7th generation inbred line (female), we generated 134.9 giga base pairs (Gb) of clean reads (135× coverage) by Illumina short-read sequencing and 28.7 Gb of clean reads (29X coverage) by PacBio long-read sequencing (Supplementary Table 1), which were corrected and hybrid-assembled into 7098 contigs and 1776 scaffolds, respectively. The yellowstripe goby reference genome was 1.002 Gb, with a scaffold N50 of 1.57 Mb, and a contig N50 of 261 kb (Table 1).

The GC distribution of the genome was relatively concentrated and unbiased (Supplementary Fig. 1). The GC content of yellowstripe goby (39%) was similar to that of the great blue-spotted mudskipper (*Boleophthalmus pectinirostris*), zebrafish, and Japanese medaka (Supplementary Fig. 2). BUSCO evaluation revealed that the assembled genome contained 87% of the known fish orthologous genes (Supplementary Table 2). When mapping the assembled transcriptome<sup>37</sup> to the assembly, 99% of the sequences were mappable (Supplementary Table 3), indicating a high-quality assembly. Using 65.29 Gb of sequencing data from the HiSeq platform for 17-mer analysis, the heterozygosity of the yellowstripe goby genome was calculated to be 1.2% (Supplementary Fig. 3).

**Table 1** Statistical analysis of the genome-assembly results.

Genome assembly		N50	Max length(bp)	Total length(bp)	Number	
	Contigs	260,505	2,174,452	988,921,936	7098	
	Scaffolds	1,569,707	9,301,748	1,002,319,200	1776	
Protein-coding genes		Total number		annotated	unannotated	
		20,531		19,729	802	
Non-coding RNAs		Copy	Average length (bp)	Percentage of genome (%)		
	miRNA	367	84.11	0.0031		
	tRNA	1273	74.46	0.0095		
	rRNA	5328	504.75	0.0306		
	snRNA	226	394.22	0.0028		
TEs		Trf	Repeatmasker	Proteinmask	De novo	Total
	Number	49,843,024	66,868,036	26,540,334	438,150,562	467,641,374

We identified 20,531 protein-coding genes (Supplementary Tables 4 and 5) in yellowstripe goby, which was similar in number to the 20,798 protein-coding genes in great blue-spotted mudskipper<sup>10</sup>, but lower than those in zebrafish (26,260)<sup>30</sup> and round goby (*Neogobius melanostomus*) (38,773)<sup>8</sup>. The lengths of mRNAs, coding sequences, exons, and introns in the yellowstripe goby genome were consistent with those in mudskipper, zebrafish, and Japanese medaka (Supplementary Fig. 4). The yellowstripe goby genome had an overall repeat content of 42.56%, which is similar to that in mudskipper<sup>10</sup> and round goby<sup>8</sup>, but lower than that in zebrafish<sup>30</sup> (52.2%) and Atlantic salmon<sup>39</sup> (58%). Long interspersed nuclear elements (LINEs) (15.88%) and DNA transposons (15.61%) were the most enriched repeat elements, whereas short interspersed nuclear elements (SINEs) were the least prevalent (2.26%) (Supplementary Table 6 and Supplementary Fig. 5).

**Construction of a high-density SNP-based genetic linkage map and assisted genome assembly.** RAD sequencing was carried out for the two parents and 225 F1 progenies, and 381 Gb of clean data, with an average sequencing depth of approximately 30× were obtained. Reads were aligned to the reference genome to identify SNP sites. We identified 627,394 and 666,860 SNPs in the female and male parent, respectively. After filtration, we constructed a high-density genetic linkage map of yellowstripe goby (Supplementary Fig. 6) based on 9534 SNP markers, representing the first genetic linkage map for the family Gobiidae. The total map length was 3098.2 centimorgans (cM), and the average genetic distance between markers was 0.32 cM, which is higher than that in most fishes at present. Using the genetic linkage map, we anchored the assembly to 22 chromosomes (Fig. 1a), containing 1065 scaffolds and 922 Mb (92%) of the total length of the assembled sequences, representing the first genome assembled at chromosomal scale in the family Gobiidae.

**Comparative genome analysis.** Nineteen representative species were selected for phylogenetic analysis; 17,347 gene families were identified, and 572 single-copy orthologous genes were selected for phylogenetic tree construction and divergence time estimations (Fig. 1b). The phylogenetic trees revealed that yellowstripe goby and mudskipper diverged 77 million years (Myr) ago and yellowstripe goby diverged from other teleosts ~120 Myr ago, which was later than the time of mudskipper divergence (140 Myr)<sup>10</sup>.

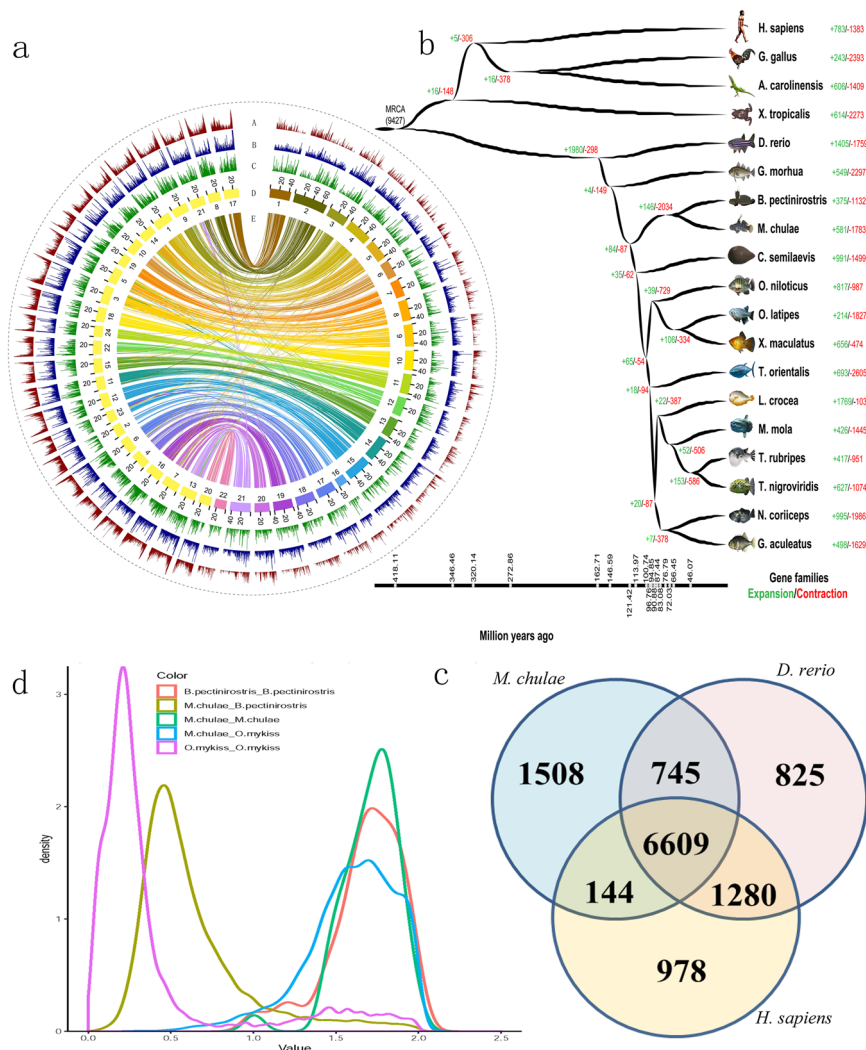
We identified 12,088 orthologous genes among yellowstripe goby, zebrafish, and humans, and 6609 genes were common among the three species (Supplementary Data 1, Fig. 1c). Yellowstripe goby shared more orthologous genes (7354) with zebrafish than did grass carp (*Ctenopharyngodon idella*) (7227)<sup>40</sup>.

The spectrum of synonymous substitutions (Ks) among yellowstripe goby, mudskipper, and *Oncorhynchus mykiss* showed peaks at 0.5 for gobiids (yellowstripe goby versus mudskipper, yellow curve in Fig. 1d), which were close to those for trout (*O. mykiss* versus *O. mykiss*, pink curve in Fig. 1d), that has undergone four whole-genome duplication events (WGD) (Fig. 1d). However, the peak with yellow line continued into another peak with green line (yellowstripe goby versus yellowstripe goby) (Fig. 1d). Thus, yellowstripe goby has just undergone three WGD events. We compared the reference genome of yellowstripe goby with Japanese medaka; most chromosomes (20/22) of the yellowstripe goby showed a one-to-one relationship with the medaka chromosomes, and only two chromosomes of yellowstripe goby showed a one-to-two correspondence with the medaka chromosomes (Fig. 1a), indicating that the relationship between their genomes was high.

**Sex determination mechanism.** We compared RAD sequencing data for males and females to identify sex determination regions in yellowstripe goby. Notably, we detected a strong signal (log of the odds score = 12.5) with a 20.67-cM-broad peak on chromosome 5 (Fig. 2), representing 49.2 Mb physical size. This is the first sex determination region to be discovered in the family Gobiidae. In this region, 58 SNPs and 102 genes were identified.

We selected the 25 most associated SNPs (Supplementary Table 7) to genotype 200 random wild fish samples ( $n = 100$  males and  $n = 100$  females) (Supplementary Methods). The locus numbered S247-888402 was homozygous (GG) in all female fish and had three genotypes (AA, GG, AG) in male fish (Supplementary Table 8). Sequence analysis showed that locus S247-888402 was located in the second intron of the *GALNT10*-like (polypeptide N-acetylgalactosaminyltransferase 10-like) gene. *GALNT10*-like expression was higher in the ovary than in the testis (Supplementary Table 9, Supplementary Fig. 7a). Functional annotation of the 102 genes revealed two more genes that may be related to sex determination, namely *MSL3* (male-specific lethal 3 homolog) and *H2AFY* (core histone macro-H2A.1) (Supplementary Table 10).

Illumina sequencing was used to identify genes differentially expressed between the testis (SA) and ovary (OV) of yellowstripe goby (Supplementary Methods). We found two male-determining genes and one female-determining gene, namely, doublesex and mab-3-related transcription factor 1 (*DMRT1*), gonadal somatoderm-derived factor (*Gsdf*), and forkhead box L2 (*FOXL2*), respectively. *DMRT1* was expressed only in the testis, while *Gsdf*, which is a downstream-regulated gene of *DMRT1*, was highly expressed in the testis (Supplementary Table 9). *FOXL2* is a critical gene for female determination and was highly expressed in the ovary but was also expressed at low levels in the testis (Supplementary Table 9). In addition, a *FOXL2* homolog (*FOXL3*; forkhead box L3), which may



**Fig. 1 Yellowstripe goby genome features.** **a** Landscape of the 22 assembled yellowstripe goby chromosomes. From the outer to the inner: GC\_content, ssr density, gene density, chromosomes, maps of the 22 yellowstripe goby chromosomes and of the 24 Japanese medaka chromosomes based on the positions of 11,756 orthologous pairs demonstrate highly conserved synteny for the 2 species. **b** The divergence-time tree of single-copy genes. **c** Gene families of *M. chulae*, *D. rerio*, and *H. sapiens*. **d** The third genome duplications in the yellowstripe goby genome was identified by Ks analyses; the pink curve represents the fourth genome duplication event of *O. mykiss*, green and orange curve represents the third genome duplication event of yellowstripe goby and mudskipper, yellow and blue curve represents the interspecific differentiation.

be related to testis and ovary development, was expressed only in the testis. The expression data are shown in Supplementary Table 9.

The expression of all these genes was validated by quantitative real-time PCR (RT-qPCR). Concordance between RNA-seq and RT-qPCR results was observed for all seven genes (Supplementary Fig. 7, Supplementary Methods).

**Global upregulation of lipid synthetic pathway genes might underlie liver high-fat storage.** Histological observation of the liver of yellowstripe goby showed that high-fat deposition in the liver might be a normal physiological phenomenon, according to histologic features in different developmental stages (Supplementary Fig. 8) and a 28-day starvation assay (Supplementary Fig. 9). The lipid component represented up to 77% of the liver wet weight (as indicated by mass spectrometry), which exceeds that in other species such as zebrafish, medaka, and humans by a large degree (Supplementary Table 11). Triglyceride was the main liver lipid, accounting for 92.59% of the total lipid dry weight (Supplementary Fig. 10; Supplementary Methods). Glucose

tolerance tests showed that glucose is cleared quickly from the blood in yellowstripe goby, as impaired glucose tolerance was not observed (Supplementary Fig. 11; Supplementary Methods). In addition, we did not observe inflammatory gene expression in normal hepatic tissues. Thus, the liver is a natural energy-storage organ in yellowstripe goby.

Liver fat rapidly accumulated between 10 and 60 days of age, and intrahepatic lipid droplets filled the hepatocytes after 3 months of age. Transcriptome analysis of livers from 2-month-old (G2M) and 3-month-old (G3M) fish revealed that lipid synthetic genes in the G2M group were globally upregulated compared with that in the G3M group (because of the low correlation of G3M-2 when compared with G3M-1 or G3M-3, the data for G3M-2 were discarded; Supplementary Fig. 12, Supplementary Tables 12 and 13, Supplementary Methods). These genes are mainly involved in the synthesis of TAG and cholesterol (CHOL), the two most important components in hepatocyte lipid droplets.

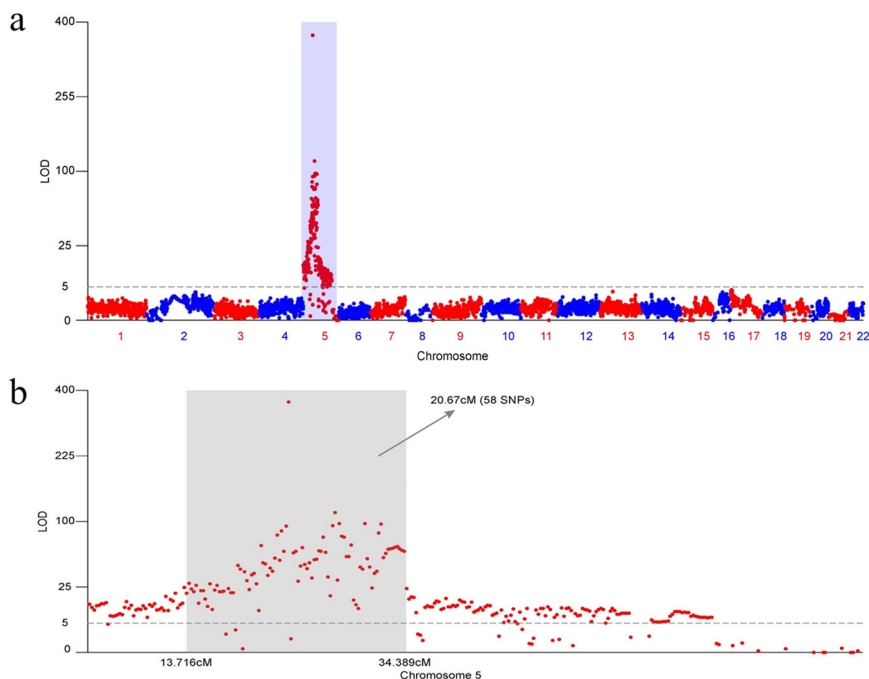
TAG is synthesized in hepatocytes via two major pathways, namely glycerol-3-phosphate (G3P) → lysophosphatidic acid



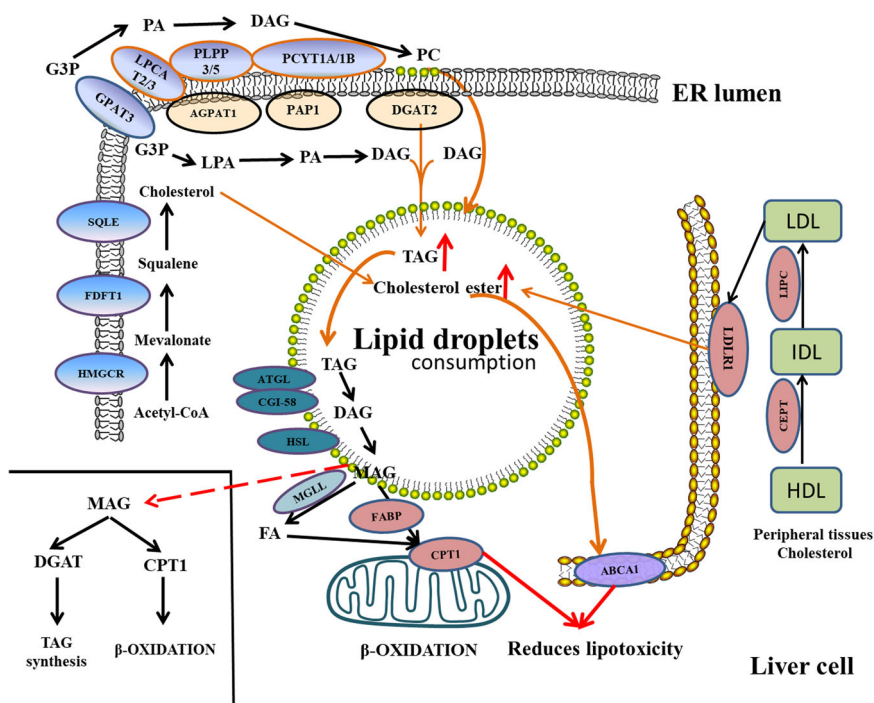
(LPA) → phosphatidic acid (PA) → diacylglycerol (DAG) → TAG and monoacylglycerol (MAG) → DAG → TAG. Our analysis revealed that some key genes involved in these two synthetic pathways, including glycerol-3-phosphate acyltransferase 3 (*GPAT3*), lipid phosphate phosphohydrolase 1 (*PAP1*), and diacylglycerol O-acyltransferase 2 (*DGAT2*) were upregulated in the livers from the G2M group compared with those from the

G3M group (Fig. 3, Supplementary Table 13). Concordance between RNA-seq and RT-qPCR results was observed (Supplementary Fig. 13, Supplementary Methods).

Another major component in lipid droplets is CHOL ester, which is mainly synthesized in hepatocytes or transported to hepatocytes from peripheral tissues. The main synthetic pathway in the liver is acetyl-CoA → mevalonate → squalene → CHOL,



**Fig. 2 QTL mapping of sex linkage in yellowstripe goby.** **a** Genome-wide LOD (Likelihood ratio statistic) score for tests of genotype difference between sexes, arranged by chromosome. The gender-related signal is occurred only at chromosome 5. **b** A 20.67-cM-broad peak (sex-association QTL search) on chromosome 5.



**Fig. 3 Formation and maintenance mechanisms of lipid droplets in the liver of yellowstripe goby.** Genes include *GPAT3*, *PAP*, *DGAT2* in TAG synthesis; *FDFT1*, *SQLE*, *LIPC*, *LDLR1* in CHOL synthesis; *LPCAT3* in phospholipid synthesis; *MGLL*, *CPT1*, and *ABCA1* in maintaining the balance between liver high-fat storage and steatohepatitis.

and the associated key genes *FDFT1* and *SQLE* were upregulated in G2M livers compared with those in the G3M livers (Fig. 3, Supplementary Table 13). High-density lipoprotein (HDL) is necessary for reverse cholesterol transport (RCT) from the peripheral to hepatic tissues. We found that genes associated with HDL transport, including those encoding hepatic lipase (*LIPC*), and low-density lipoprotein receptor adapter protein 1 (*LDLR1*) were all upregulated in G2M livers (Fig. 3, Supplementary Table 13). The increase in HDL RCT function in the G2M group might promote CHOL deposition in lipid droplets. Concordance between RNA-seq and RT-qPCR results was observed (Supplementary Fig. 13, Supplementary Methods).

**Differential expression of monoglyceride lipase (*MGLL*) gene might promote lipid accumulation in the early stage of the liver.** TAG decomposition in lipid droplets is mainly driven by patatin-like phospholipase domain-containing 2 (*ATGL*) and abhydrolase domain-containing 5 (*CGI-58*) on lipid droplet membranes to generate DAG, which is then degraded by the lipid droplet membrane protein, hormone-sensitive lipase (*HSL*), to produce MAG. MAG is finally degraded by intracellular *MGLL* to free fatty acids (FFAs) and glycerol. Excessive intracellular FFAs are essential for activating Kupffer cells, which induce cellular inflammation and NASH. Surprisingly, *MGLL* gene expression was low in the livers from the G2M group, whereas it was highly expressed in the livers from the G3M group. The low *MGLL* expression in the G2M group might contribute to the promotion of lipid accumulation in the early stage of the liver (Fig. 3).

Low expression of *MGLL* in G2M group hepatocytes may lead to drastic accumulation of the decomposition target, MAG. We found that MAG in hepatocytes could be returned to lipid droplets via  $\text{MAG} \rightarrow \text{TAG}$  pathway. *DGAT2*, a key gene in the  $\text{MAG} \rightarrow \text{TAG}$  pathway, was highly expressed in the G2M livers. In the G2M livers, MAG mostly was re-metabolized into TAG through *DGAT2* for storage. Moreover, the results of RNA-seq analysis could be verified by RT-qPCR (Supplementary Fig. 13, Supplementary Methods).

**Increased phospholipid synthesis contributes to maintaining lipid droplet membrane homeostasis.** Phospholipids play a key role in maintaining lipid droplet membrane homeostasis. Lecithin (PC) and cephalin (PE) are the main components of the monolayer phospholipid membrane of lipid droplets. We found that the key genes required for PC and PE synthesis, including those encoding *GPAT3* and lysophosphatidylcholine acyltransferase 3 (*LPCAT3*) were upregulated in G2M compared to those in G3M (Fig. 3, Supplementary Table 13). Furthermore, concordance between RNA-seq and RT-qPCR results was observed (Supplementary Fig. 13, Supplementary Methods).

**Carnitine palmitoyltransferase 1 (*CPT1*) reduces lipotoxicity by promoting free fatty acid consumption in hepatocytes.** We found that *CPT1* expression was much higher in the G3M group than in the G2M group (Supplementary Table 13, Supplementary Fig. 13). After the TAG storage in hepatocytes reached saturation in the G3M group, MAG mainly underwent  $\beta$ -oxidation through mitochondrial *CPT1* for energy supply (Fig. 3). *CPT1* might reduce lipotoxicity by promoting FFA consumption in hepatocytes. Concordance was observed between RNA-seq and RT-qPCR results (Supplementary Fig. 13).

***ABCA1* expansion reduces lipotoxicity by increasing RCT function in hepatocytes.** Sequence analysis revealed that *ABCA1* of yellowstripe goby has the largest copy number among known species (Supplementary Table 14), with four copies distributed

over four chromosomes (Fig. 4a). All *ABCA1* genes had two nucleotide-binding domains and two characteristic extracellular domains. Analysis of the tissue distribution of *ABCA1* revealed that *ABCA1a-1* and *ABCA1c* are the most highly expressed genes in the yellowstripe goby liver (Fig. 4b). Phylogenetic analysis revealed that *ABCA1c* was the ancestral copy (Fig. 4c), and most fishes had lost it during evolution, as had humans and zebrafish. *ABCA1* is a key RCT gene in hepatocytes. Expansion of yellowstripe goby *ABCA1* and the high *ABCA1c* expression in the liver, in particular, could reduce the lipotoxicity of CHOL to hepatocytes by increasing RCT in hepatocytes, thus maintaining the balance between liver high-fat storage and steatohepatitis.

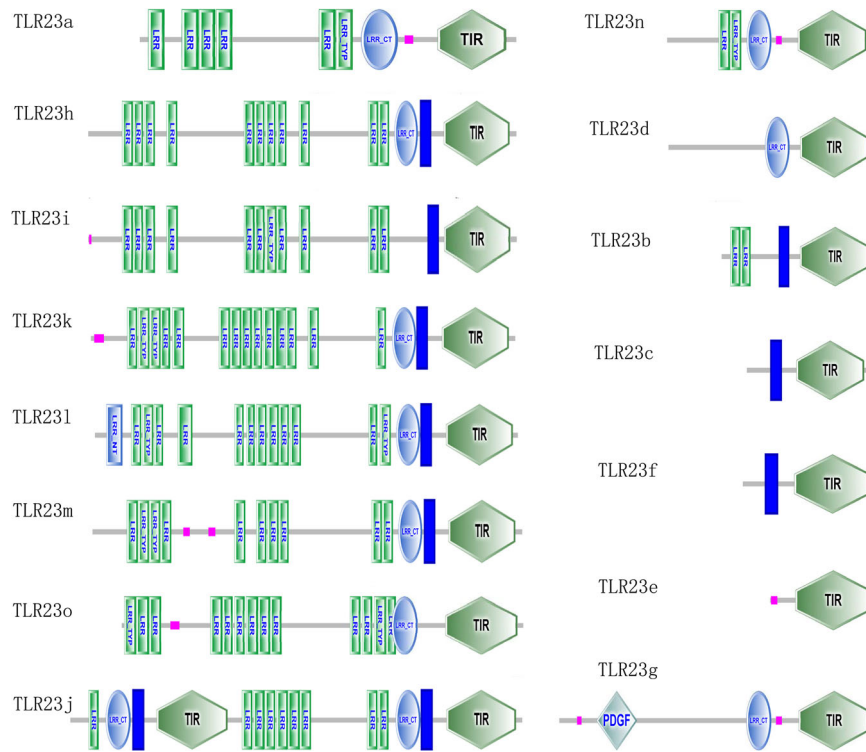
**A robust innate immune system helps reduce the amplification effect of the external environment on steatohepatitis.** We discovered an expansion of *TLR23* (15 copies) and tripartite motif containing (*TRIM*) family members (234 members) in yellowstripe goby (Supplementary Tables 15 and 16). All *TLR23* copies had a Toll/IL-1 receiver (TIR) domain, and one of them (*TLR23h*) had two TIR domains, representing the first fish TLR gene with these characteristics (Fig. 5). A phylogenetic tree was constructed for *TLR23* genes, using the neighbor-joining method, based on the amino acid sequences of yellowstripe goby and mudskipper (Fig. 6). Two distinct clades, representing two subfamilies were clearly distinguished. The tripartite motif containing (*TRIM*) family is another important gene family involved in innate immunity. The *TRIM* family encodes E3 ubiquitin ligases, which are involved in several important biological processes, especially, antiviral responses. We identified 234 *TRIM* genes in yellowstripe goby, which is the largest number among known species (Supplementary Table 16), with *TRIM14*, *TRIM16*, *TRIM21*, *TRIM25*, *TRIM35*, and *TRIM39* being the most abundant (Fig. 7). Among them, *TRIM21* had the largest copy number (58) (Supplementary Table 17). Expansion of *TLR23* and the *TRIM* family might be beneficial in the context of high-fat storage.

## Discussion

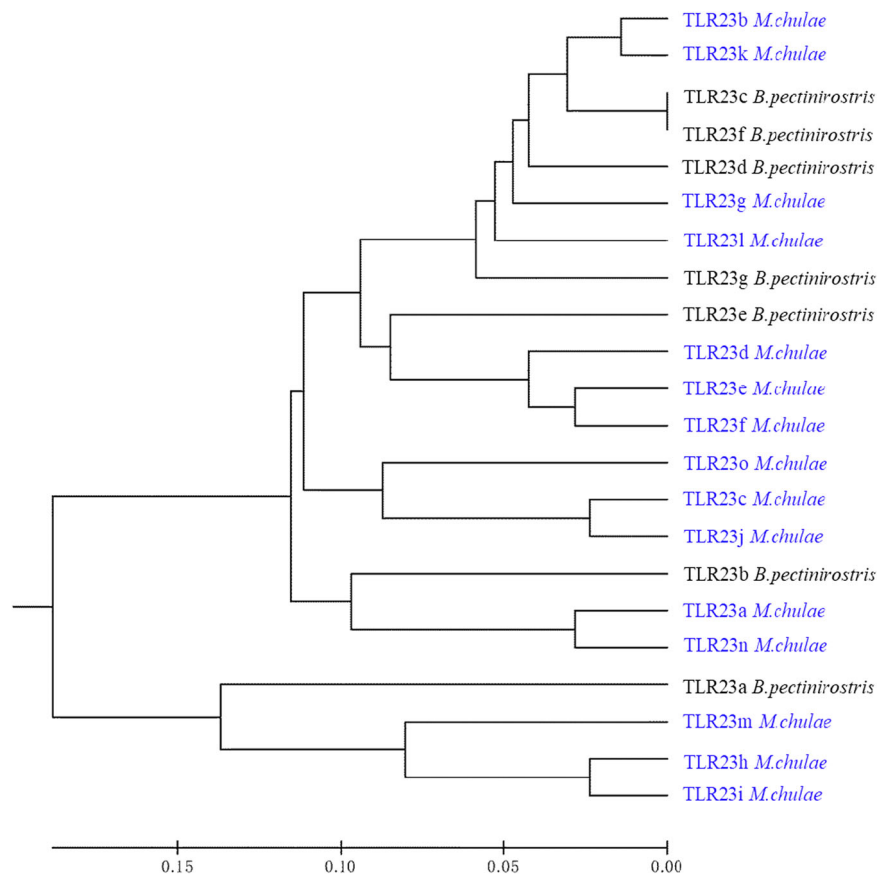
We generated a whole-genome sequence of yellowstripe goby, using Illumina short-read and PacBio long-read sequencing. This represents the first annotated chromosome-level reference genome assembly for yellowstripe goby, providing important basic information for future research on the genetic evolution, sex determination, diseases, and marine ecotoxicology of yellowstripe goby as a potential marine model fish. Further, we constructed a high-density SNP genetic linkage map for yellowstripe goby, the first in the family Gobiidae (>2000 species). The average marker distance was 0.32 cm, which is lower than that for most non-model and non-aquaculture species, including *Cyprinus carpio haematopterus* (0.57 cm)<sup>41</sup>, *Nibeia albiflora* (0.47 cm)<sup>42</sup>, *Larmichthys crocea* (0.36 cm)<sup>43</sup>, and *Pseudobagrus ussuriensis* (0.36 cm)<sup>44</sup>. Using the constructed genetic map, the yellowstripe goby genome was assembled at the chromosome level, and >92% of the assembled sequences were anchored. This is the first genome assembled at the chromosomal level in the family Gobiidae<sup>8–10</sup>.

We discovered that chromosome 5 of yellowstripe goby carries a 20.67-cm sex determination region that contains three genes that might be related to sex determination, namely, *MSL3*, *H2AFY*, and *GALNT10*-like genes. *MSL3* contributes to overexpression of genes on the X chromosome of male *Drosophila*<sup>45</sup>. Research on the structure of human *MSL3* has shown that it functions similar to the *Drosophila MSL3*<sup>46</sup>, by binding to lysine 20 on the N-terminal tail of histone H4 to regulate the male-specific lethal complex on the X chromosome. *H2AFY* helps maintain X-chromosome inactivation<sup>47</sup>. Thus, *MSL3* and *H2AFY* may be involved in yellowstripe goby sex determination. *GALNT10*-like participates in



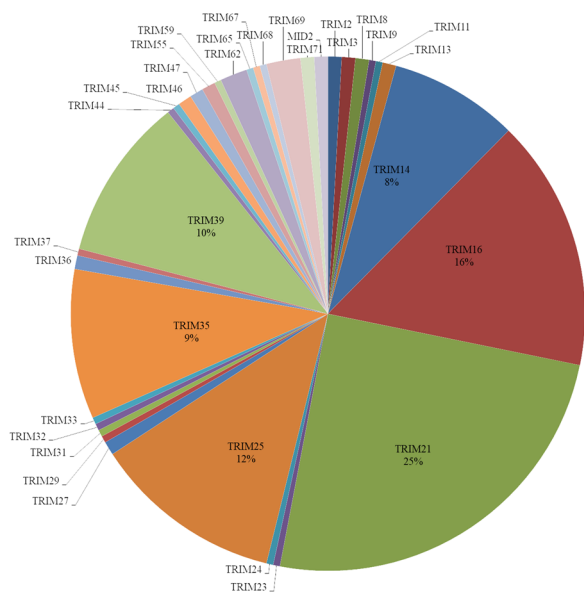


**Fig. 5 Protein domain structures of *TRL23* in yellowstripe goby.** LRR leucine-rich repeat, LRR-TYP leucine-rich repeat typical subfamily, TIR Toll/IL-1 receptor, LRR-NT leucine-rich repeat N-(nitrogen) terminal, LRR-CT leucine-rich repeat C-(carboxyl) terminal, PDGF platelet-derived growth factor.



**Fig. 6 Phylogenetic tree of *TRL23* between yellowstripe goby and *B. pectinirostris*.** *TRL23* genes of yellowstripe goby are highlighted in blue. Phylogeny of *TLR23* family between yellowstripe goby and *B. pectinirostris* showing the expansion of *TLR23* in yellowstripe goby.





**Fig. 7 Distribution of gene numbers in the TRIM gene family in yellowstripe goby.** A total of 34 TRIM gene families (234 members) were identified in yellowstripe goby, which is the largest number among known species (Supplementary Table 16), with TRIM14, TRIM16, TRIM21, TRIM25, TRIM35, and TRIM39 being the most abundant.

genome and transcriptome data, we identified three key metabolic pathways for lipid accumulation in the liver of yellowstripe goby, namely TAG synthesis, CHOL synthesis, and PC synthesis. TAG and CHOL are the two major components in the lipid droplets of hepatocytes, and PC is a major structural component of biomembranes, including lipid droplet membranes. Key genes involved in TAG, CHOL, and PC synthesis were globally up-regulated in the G2M livers; especially, the key genes *GPAT3*, *PAP1*, *DGAT2*, *FDFT1*, *SQLE*, *LIPC*, *LDLR1*, and *LPCAT3* were all expressed at higher levels in the G2M phase than in the G3M phase (Supplementary Fig. 13). Increased TAG and CHOL synthesis can promote lipid droplet accumulation, while increased PC synthesis is crucial for regeneration and maintaining the stability of lipid droplets<sup>56,57</sup>. High hepatic expression of genes associated with TAG, CHOL, and PC synthesis and transportation possibly promotes lipid droplet formation in the liver in yellowstripe goby.

In yellowstripe goby, the liver starts to accumulate TAG after hatching and maintains a high-fat accumulation state throughout life without an accompanying inflammatory reaction (Supplementary Fig. 8 and Fig. 9). Yellowstripe goby faces great challenges in maintaining normal physiological function and energy-storage balance in the liver. Genome and transcriptome analysis revealed that the yellowstripe goby might maintain the balance between lipid accumulation and intracellular inflammatory response by modulating the differential expression of *MGLL* and *CPT1*. Usually, TAG is decomposed into MAG by *ATGL*, *CGI-58*, and *HSL* on the surface of lipid droplets<sup>58,59</sup> (Fig. 3), and MAG is further decomposed into FFAs by *MGLL*<sup>60</sup>. Surprisingly, *MGLL* gene expression was low in the livers from the G2M group. The low *MGLL* expression in the G2M group might contribute to the promotion of lipid accumulation in the early stage of the liver (Fig. 3). In contrast, *MGLL* was highly expressed in the livers from the G3M group and might produce vast quantities of FFAs in the hepatocytes. FFAs are derived from lipids stored in lipid droplets and can produce severe lipotoxicity in cells<sup>61</sup>. We found that *CPT1* expression was much higher in the G3M group than in the G2M group (Supplementary Table 13, Supplementary

Fig. 13). *CPT1* is the rate-limiting enzyme for the uptake of FFAs and their subsequent beta-oxidation in mitochondria<sup>62</sup>. *CPT1* might reduce lipotoxicity by promoting FFA consumption in hepatocytes. The newly identified *MGLL* and *CPT1* expression pattern in the yellowstripe goby liver may provide a new insight for future research on treating and preventing human NAFLD.

Yellowstripe goby had the highest *ABCA1* copy number among the species evaluated. *ABCA1* participates in the transport of CHOL from hepatocytes to the extracellular space<sup>63</sup>. Yellowstripe goby *ABCA1a-1* has similar functions to human *ABCA1*, whereas *ABCA1c* is an ancient subtype especially preserved in yellowstripe goby. Yellowstripe goby *ABCA1a-1* and *ABCA1c* are both highly expressed in the liver, which might increase the extrusion of free CHOL from hepatocytes and play an important role in reducing the lipotoxicity related to free CHOL.

Genome analysis revealed that the innate immunity gene *TLR23* is expanded in yellowstripe goby, with 15 copies, which is higher than that in most other fish, except round goby (40 copies)<sup>8</sup>, Glacier lanternfish (*Benthoosema glaciale*) (49 copies)<sup>8</sup>, and European perch (*Perca fluviatilis*) (17 copies)<sup>8</sup>. In teleosts, TLR expansions might correlate with the survival and successful radiation of this lineage<sup>64</sup>. The expansion of *TLR23* in yellowstripe goby may help it to adapt to the complex environment, for instance, by reducing the amplification effect of the external environment on steatohepatitis. During the progression from simple NAFLD to NASH in humans, hepatocytes usually undergo multiple hits and eventually produce inflammatory reactions<sup>63,65</sup>. Pathogen- and damage-associated molecular patterns (PAMPs and DAMPs, respectively) are the main inducers of NASH inflammation. Both PAMPs and DAMPs can trigger inflammatory reactions in hepatocytes through TLRs<sup>66</sup>. However, the major function of *TLR23* is its involvement in identifying bacterial 23 S ribosomal RNA<sup>67</sup>. Therefore, the increased *TLR23* copy number in yellowstripe goby may be relevant in the inhibition of NASH caused by PAMPs, weakening the amplification effect of pathogens on NASH, and providing a stable innate immune environment in the liver to allow higher fat storage.

## Methods

### Genome sequencing and assembly

**DNA library construction and sequencing.** For sequencing, genomic DNA was isolated from the 7th generation of an inbred line. For whole-genome shotgun sequencing, four female full siblings were used: one to construct short-insert libraries of 270-bp and 800-bp, two for long-insert libraries of 20-kb, and one for a 40-kb long-insert library. Paired-end sequencing of the short-insert libraries was performed using the Illumina HiSeq 4000 system. After removing the adapter sequences, ambiguous and low-quality reads were filtered out using SOAPnuke<sup>68</sup> software, version 1.5.4 (<https://github.com/BGI-flexlab/SOAPnuke>) with the parameters: '-n 0.05 -l 7 -q 0.2 -d -i -Q 2'. The clean, high-quality data were used for genome assembly. Long-read libraries were sequenced using the Pacific Biosciences RSII system. After removing reads with a length of <500 bp or a score of <80, a total of 30.0 Gb of high-quality data was obtained. The study was approved by Institutional Animal Care and Use Committee (IACUC) of Guangdong Laboratory Animals Monitoring Institute, Guangzhou.

**Genome size estimation.** Filtered short reads were used to estimate the genome size and heterozygosity of yellowstripe goby by performing 17-mer analysis using the KmerFreq software, version 5.0 (<https://github.com/fanagislab/kmerfreq>). The genome size was estimated using the formula  $G = \frac{N_{k-mer}}{C_{k-mer}} = \frac{N_{read} \times (L-k+1)}{C_{k-mer}}$ , where  $G$  is the genome size,  $N_{k-mer}$  and  $N_{read}$  are the respective numbers of K-mers and reads,  $C_{k-mer}$  is the average coverage depth of the K-mers, and  $L$  and  $K$  represent the read and K-mer lengths, respectively. A series of curve simulations was used to estimate the genome heterozygosity rate.

**Genome hybrid assembly.** The high-quality paired-end sequencing reads from the small-insert libraries were used to construct short, but accurate De Bruijn graph contigs using Platanus<sup>69</sup> (<http://platanus.bio.titech.ac.jp/platanus-assembler>, version 1.2.4) with the parameters: '-k 31 -s 10 -n 2 -c 3 -a 10.0 -u 0.2 -d 0.4 -t 16 -m 200'. The Celera Assembler PBcR pipeline (version 8.3rc2)<sup>70</sup> was used to correct the sequencing errors of the PacBio SMRT reads. The short, accurate contigs were then mapped to PacBio long reads to generate a hybrid assembly using

DBG2OLC<sup>71</sup> (<https://github.com/yechengxi/DBG2OLC>). The initial consensus sequences of DBG2OLC were polished to correct erroneous sequences due to the high error rates of the PacBio reads. SSPACE software (version 1.2.4)<sup>72</sup> was used for scaffolding the hybrid contigs (<https://www.baseclear.com/services/bioinformatics/basetools/sspace-standard/>). Finally, we used TrimDup, which is part of the Rabbit Genome Assembler (<https://github.com/gigascience/rabbit-genome-assembler>, version 2.6) with a percentage of 0.3 to remove redundant sequences.

**BUSCO evaluation.** We used BUSCO<sup>73</sup> software, version 3.0.2 to evaluate completeness and accuracy of the genome assembly. We selected actinopterygii\_odb9 as the database, which contained 4584 highly conserved genes of fishes.

### Genome characterization

**Repeat detection.** We utilized both known and de novo methods for detecting repetitive DNA sequences. We used RepeatMasker (version 4.0.6)<sup>74</sup>, RepeatProteinMask, and Tandem Repeats Finder (version 4.07)<sup>75</sup> to detect known transposable elements, transposable element-related proteins, and tandem repeats, respectively. In addition, we constructed a de novo repeat library using RepeatModeler (version 1.0.8) and LTR\_FINDER (version 1.0.6)<sup>76</sup>, and then we employed RepeatMasker to find de novo transposable elements.

**Gene-structure prediction.** To identify candidate protein-coding genes, we first aligned great blue-spotted mudkipper (*Boleophthalmus pectinirostris*), zebrafish (*Danio rerio*), and Japanese medaka (*Oryzias latipes*) protein-coding genes against the yellowstripe goby genome using TblastN with an *E* value of 1e-5. Then, we used GeneWise (version 2.4.1)<sup>77</sup> for precise alignments and gene-structure predictions. We used AUGUSTUS software (version 3.2.1)<sup>78</sup> and the GENSCAN<sup>79</sup> web server for ab initio gene-structure predictions. Furthermore, Illumina HiSeq RNA-Seq transcriptome data was used. HiSeq RNA-Seq reads from two liver tissues<sup>37</sup> were mapped on to the yellowstripe goby genome using Hisat2 (version 2.0.2)<sup>80</sup>. Finally, all the above data were combined to generate a comprehensive gene set using GLEAN<sup>81</sup>.

**Gene function annotation.** Gene functions were annotated by aligning yellowstripe goby protein sequences to public databases (NT, NR, COG, KEGG, Swiss-Prot, and TrEMBL) using BlastP with an *E* value of 1e-5. InterProScan analysis was performed by running the ProDom, PRINTS, HAMAP, and Pfam applications. Gene Ontology annotations were extracted from the NR database using Blast2GO<sup>82</sup>.

**Non-coding RNA predictions.** Non-coding RNAs (including microRNAs, ribosomal RNAs, small nuclear RNAs, and transfer RNAs) were predicted by comparing the yellowstripe goby genome against public libraries.

**Ks analysis.** For the Ks analysis, we first blasted the proteins sequences of three species (*M.chuluae*, *Boleophthalmus pectinirostris*, *Oncorhynchus mykiss*) using blastp with themselves or between two species (Parameter: -m 8 -e 1e-5 -b 5 -v 5). The alignment was then processed by MCScanX to get collinear blocks (Parameter: -k 200 -g -2 -m 15 -s 5); each block contained no less than 5 gene pairs. Next, we connected gene pairs of each block into two supergenes and carried out sequence alignment with MUSCLE. Finally, we converted the protein sequences into nucleotides and carried out selective pressure analysis by PAML, and calculated Ks values of every block.

### Comparative genome analysis

**Gene family construction.** We assembled gene sets from 18 species (*Anolis carolinensis*, *B. pectinirostris*, *Cynoglossus semilaevis*, *D. rerio*, *Takifugu rubripes*, *Gadus morhua*, *Gallus gallus*, *Gasterosteus aculeatus*, *Homo sapiens*, *Larimichthys crocea*, *Mola mola*, *Notothenia coriiceps*, *Oreochromis niloticus*, *O. latipes*, *Tetraodon nigroviridis*, *Thunnus orientalis*, *Xenopus tropicalis*, and *Xiphophorus maculatus*), in addition to yellowstripe goby. All-to-all BlastP was performed using all protein sequences with *E* value of 1e-7. Hcluster\_sg (<https://github.com/douglascosfield/hcluster>) was used for protein clustering.

**Phylogenetic tree construction.** Using the clustered families, single-copy protein-encoding genes were extracted and multiple-sequence alignments were performed using MUSCLE (version 3.8.31)<sup>83</sup>. Corresponding coding sequence alignments were determined from protein alignments and were joined to form a 'supergene' for every species. We removed poorly aligned positions and divergent regions using Gblocks software (version 0.91b)<sup>84</sup> with default parameters, before constructing a phylogenetic tree using RAXML software (version 8.2.4)<sup>85</sup> with the GTRGAMMA model.

**Divergence time estimation.** The MCMCTree module from the PAML package was used to estimate the divergence time of yellowstripe goby from *B. pectinirostris* and other species. We selected several reference divergence times (marked by red dots in several branches) from the TimeTree database<sup>86</sup> (<http://www.timetree.org/>) to calibrate the divergence times for other nodes.

**Gene family expansion and extraction.** We used Computational Analysis of gene Family Evolution (CAFE) software<sup>87</sup>, version 2.1 to analyze the changes in family sizes that occurred during the phylogenetic history. Prior to this analysis, we removed gene families with changes that were either too large ( $\geq 200$ ) or too small ( $\leq 2$ ) in size, as these could lead to wrong parameter estimations in CAFE.

**Gene copy-number scanning.** Protein sequences were downloaded from the NCBI or KEGG database. All-to-all BlastP was performed to assess the criteria of coverage and identity cutoffs. Then, these proteins were used to scan all genomes to determine gene copy numbers for each species.

### Genetic map construction

**RAD sequencing.** RAD sequencing libraries from the two parents and 225 F1-generation offspring were constructed and pooled into 22 final libraries with equal amounts of products, which were then sequenced on an Illumina HiSeq 4000 sequencing platform.

**RAD sequencing data analysis.** After removing adapters and low-quality bases, the clean reads were assigned to each individual based on specific barcodes and the EcoRI recognition site (GAATTC). Reads that did not contain a matching, unique barcode were discarded. All reads were aligned against the reference yellowstripe goby genome using BWA software (version 0.7.12)<sup>88</sup>. Single-nucleotide polymorphisms (SNPs) were detected using SAMtools software (version 1.2.1)<sup>89</sup> with the parameters: 'mpileup -g -d 100 -q 20 -Q 15', and bcftools software (version 1.3.1). Subsequently, we identified divergent SNP sites that differed between parents and sites that were heterozygous in either parent. Then, the basetypes of each individual offspring were extracted to construct a final basetype table.

**Linkage map construction.** SNP markers were filtered before they were used for linkage map construction by removing the following: (1) markers whose genotypes were the same between the parents or homozygous in both parents, (2) offspring samples in which <80% of the SNP sites were genotyped, (3) markers that could not be genotyped in  $\geq 1\%$  of the offspring samples, (4) markers with significantly distorted segregation ( $P < 0.05$ ) in  $\chi^2$  goodness-of-fit tests, and (5) redundant markers linkage disequilibrium  $< 0.8$ . Paternal- and maternal-specific linkage maps were constructed using both JoinMap (version 4.0)<sup>90</sup> and LepMap2<sup>91</sup> with cross-pollinator population type codes and a logarithm of odds score limit of 20. After removing markers that came from the same sequence but were located on different linkage groups or with contradictory orders, an integrated sex-averaged linkage map was obtained.

**Sex chromosome identification.** To identify sex chromosomes, we conducted gender identification on the F1 individuals using a microarray method, because of the tiny size of the samples. Of the 225 F1 individuals, 67 were identified as males, 72 were identified as females (Supplementary Table 18), and 86 were unidentifiable. We used mapQTL6 to identify the sex determination region by interval mapping (assigning a value of -1 for females and a value of 1 for males).

**Liver tissue sections of yellowstripe goby.** Samples were anesthetized in MS222 (50 mg/L) and dissected to obtain the livers. Each liver was preserved in 4% neutral formaldehyde fixative for sectioning. The liver samples were dehydrated for routine pathology, fixed by paraffin embedding, and sliced using an automatic slicer (thickness, 4  $\mu$ m). The sections were subjected to heating, pasting, drying, dewaxing, hematoxylin staining for 7 min, bluing with warm water for 1 min, and soaking in 1% hydrochloric acid alcohol for differentiation for approximately 60 s. Subsequently, they were subjected to eosin staining for 5 min, dehydration through an alcohol gradient, xylene hyalinization, neutral-resin sealing, observation under a microscope, and photographed.

### Determination of the hepatic lipid composition

**Experimental fish.** The experimental fish comprised a closed group of 12-month-old yellowstripe goby bred in our laboratory, totaling 300 individuals, which were divided into three groups.

**Hepatic lipid assays.** Total lipids, saturated fatty acids, monounsaturated fatty acids, polyunsaturated fatty acids, N-3, and N-6 were subjected to methyl esterification and analyzed using gas chromatography-mass spectrometry. Approximately 2 g of each sample was taken in a test tube, to which 1 mL of potassium hydroxide-methanol solution was added. Each test tube was sealed, placed in an oscillator to shake for approximately 30 min, and centrifuged at 5000 rpm in a high-speed centrifuge for approximately 5 min. The supernatants were used for testing.

After dehydration, total CHOL ester was quantitatively evaluated by sulfur, phosphorus, and iron indicator-spectrophotometry. Free CHOL ester was measured by staining with Coomassie brilliant blue. Glycerol ester and diglyceride were determined by spectrophotometry. The free and total glycerol contents after diglyceride hydrolysis were determined separately under different hydrolysis conditions. Total fatty acid contents were measured by acid-base titration. Diphosphatidylglycerol, lysolecithin, phosphatidylcholine,

phosphatidylethanolamine, phosphatidylserine, and sphingomyelin were measured by high-performance liquid chromatography.

#### Validation of differentially expressed genes by quantitative real-time PCR.

Samples used for sex determination and lipid metabolism genes validation had been described in the Materials subsection under Transcriptome Sequencing (Supplementary Methods). Reverse transcription of total RNA was conducted following the manufacturer's protocol (Takara Bio, China). Quantitative real-time PCR amplification was performed on an ABI7500 system (Thermo, USA) using a TB Green kit (Takara Bio, China). The relative expression levels of target genes were calculated by the  $2^{-\Delta\Delta C_t}$  method<sup>38</sup>.

**Statistics and reproducibility.** Statistical analysis was performed by Student's *t* test (between two groups) and one-way ANOVA (among three or more groups) using SPSS 17.0 software (SPSS Inc.). A value of *P* value < 0.05 was used to indicate a significant difference.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

#### Data availability

The yellowstripe goby whole-genome project has been deposited in NCBI under project PRJNA598084. RAD sequencing reads have been deposited in the NCBI Sequence Read Archive under project PRJNA642226, and RNA-Seq sequencing reads have been deposited in the NCBI Sequence Read Archive under project PRJNA641222.

Received: 26 February 2020; Accepted: 1 December 2020;

Published online: 04 January 2021

#### References

- Muralidhar, P. & Veller, C. Sexual antagonism and the instability of environmental sex determination. *Nat. Ecol. Evol.* **2**, 343–351 (2018).
- Capel, B. Vertebrate sex determination: evolutionary plasticity of a fundamental switch. *Nat. Rev. Genet.* **18**, 675–689 (2017).
- Oliveira, C. & Toledo, L. F. A. Evidence of an XX/XY sex chromosome system in the fish *Dormitator maculatus* (Teleostei, Eleotrididae). *Genet. Mol. Biol.* **29**, 653–655 (2006).
- Chen, S. et al. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nat. Genet.* **46**, 253–262 (2014).
- Kelley, J. L. et al. The genome of the self-fertilizing mangrove rivulus fish, *Kryptolebias marmoratus*: a model for studying phenotypic plasticity and adaptations to extreme environments. *Genome Biol. Evol.* **8**, 2145–2154 (2016).
- Mank, J. E., Promislow, D. E. & Avise, J. C. Evolution of alternative sex-determining mechanisms in teleost fishes. *Biol. J. Linn. Soc.* **87**, 83–93 (2006).
- Li, X. Y. et al. Origin and transition of sex determination mechanisms in a gynogenetic hexaploid fish. *Heredity* **121**, 64–74 (2018).
- Adrian-Kalchauer, I. et al. The round goby genome provides insights into mechanisms that may facilitate biological invasions. *BMC Biol.* **18**, 1–33 (2020).
- Malmström, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S. & Jentoft, S. Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Sci. Data* **4**, 160132 (2017).
- You, X. et al. Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nat. Commun.* **5**, 5594 (2014).
- Birsoy, K., Festuccia, W. T. & Laplante, M. A comparative perspective on lipid storage in animals. *J. Cell Sci.* **126**, 1541–1552 (2013).
- Gesta, S., Tseng, Y. H. & Kahn, C. R. Developmental origin of fat: tracking obesity to its source. *Cell* **131**, 242–256 (2007).
- Lethbridge, R. C. & Potter, I. C. Quantitative studies on the skin of the paired species of lampreys, *Lampetra fluviatilis* (L.) and *Lampetra planeri* (BLOCH). *J. Morphol.* **164**, 39–46 (1980).
- Flynn, E. J. 3rd, Trent, C. M. & Rawls, J. F. Ontogeny and nutritional control of adipogenesis in zebrafish (*Danio rerio*). *J. Lipid Res.* **50**, 1641–1652 (2009).
- Salmerón, C. Adipogenesis in fish. *J. Exp. Biol.* **221**, jeb161588 (2018).
- Xiong, S., Krishnan, J., Peuss, R. & Rohner, N. Early adipogenesis contributes to excess fat accumulation in cave populations of *Astyanax mexicanus*. *Dev. Biol.* **441**, 297–304 (2018).
- Tamura, E. Histological changes in the organs and tissues of the gobiid fishes throughout the life span-IV. Digestive organs of the ice-goby. *Nippon Suisan Gakkaishi* **37**, 831–839 (1971).
- Akiyoshi, H. & Inoue, A. Comparative histological study of teleost livers in relation to phylogeny. *Zool. Sci.* **21**, 841–850 (2004).
- Louiz, I., Palluel, O., Ben-Attia, M., Ait-Aissa, S. & Hassine, O. K. B. Liver histopathology and biochemical biomarkers in *Gobius niger* and *Zosterisessor* *ophiocephalus* from polluted and non-polluted Tunisian lagoons (Southern Mediterranean Sea). *Mar. Pollut. Bull.* **128**, 248–258 (2018).
- Cuevas, N., Zorita, I., Franco, J., Costa, P. M. & Larreta, J. Multi-organ histopathology in gobies for estuarine environmental risk assessment: a case study in the Ibaizabal estuary (SE Bay of Biscay). *Estuar. Coast. Shelf Sci.* **179**, 145–154 (2016).
- Ando, S., Mori, Y., Nakamura, K. & Sugawara, A. Characteristics of lipid accumulation types in five species of fish. *Nippon Suisan Gakkaishi* **59**, 1559–1564 (1993).
- Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).
- Ishikawa, A. et al. A key metabolic gene for recurrent freshwater colonization and radiation in fishes. *Science* **364**, 886–889 (2019).
- Kwiterovich, P. O. Jr., Sloan, H. R. & Fredrickson, D. S. Glycolipids and other lipid constituents of normal human liver. *J. Lipid Res.* **11**, 322–330 (1970).
- Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of NAFLD development and therapeutic strategies. *Nat. Med.* **24**, 908–922 (2018).
- Wree, A., Broderick, L., Canbay, A., Hoffman, H. M. & Feldstein, A. E. From NAFLD to NASH to cirrhosis—new insights into disease mechanisms. *Nat. Rev. Gastroenterol. Hepatol.* **10**, 627–636 (2013).
- Thacker, C. E. & Roje, D. M. Phylogeny of *Gobiidae* and identification of gobiid lineages. *Syst. Biodivers.* **9**, 329–347 (2011).
- Patzner, R., Van Tassell, J. L., Kovacic, M. & Kapoor, B. *The Biology of Gobies* (CRC Press, 2011).
- Leder, E. H. et al. Post-glacial establishment of locally adapted fish populations over a steep salinity gradient. *J. Evol. Biol.* **00**, 1–19 (2020).
- Howe, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
- Kasahara, M. et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714–719 (2007).
- Schartl, M. et al. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nat. Genet.* **45**, 567–572 (2013).
- Peichel, C. L. et al. The genetic architecture of divergence between threespine stickleback species. *Nature* **414**, 901–905 (2001).
- Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
- Therkildsen, N. O. et al. Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science* **365**, 487–490 (2019).
- Larson, H. K. A revision of the gobiid fish genus *Mugilogobius* (Teleostei: Gobioidae), and its systematic placement. *Rec. West. Aust. Mus. Suppl.* **62**, 1–233 (2001).
- Cai, L. et al. De novo transcriptome assembly of the new marine fish model of goby, *Mugilogobius chulae*. *Mar. Genom.* **40**, 18–20 (2018).
- Cai, L. et al. Characterization of transcriptional responses mediated by benzo [a]pyrene stress in a new marine fish model of goby, *Mugilogobius chulae* genes. *Genomics* **41**, 113–123 (2019).
- Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
- Wang, Y. et al. The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. *Nat. Genet.* **47**, 625–631 (2015).
- Feng, X. et al. A high-resolution genetic linkage map and QTL fine mapping for growth-related traits and sex in the Yangtze River common carp (*Cyprinus carpio haematopterus*). *BMC Genom.* **19**, 230 (2018).
- Qiu, C. et al. A high-density genetic linkage map and QTL mapping for growth and sex of yellow drum (*Nibea albiflora*). *Sci. Rep.* **8**, 17271 (2018).
- Kong, S. et al. Constructing a high-density genetic linkage map for large yellow croaker (*Larimichthys crocea*) and mapping resistance trait against ciliate parasite *Cryptocaryon irritans*. *Mar. Biotechnol.* **21**, 262–275 (2019).
- Zhu, C. et al. Construction of a high-density genetic linkage map and QTL mapping for growth traits in *Pseudobagrus ussuriensis*. *Aquaculture* **511**, 734213 (2019).
- Gorman, M., Franke, A. & Baker, B. S. Molecular characterization of the male-specific lethal-3 gene and investigations of the regulation of dosage compensation in *Drosophila*. *Development* **121**, 463–475 (1995).
- Moore, S. A., Ferhatoglu, Y., Jia, Y., Al-Jiab, R. A. & Scott, M. J. Structural and biochemical studies on the chromo-barrel domain of male specific lethal 3 (MSL3) reveal a binding preference for mono- or dimethyllysine 20 on histone H4. *J. Biol. Chem.* **285**, 40879–40890 (2010).
- Hernandez-Munoz, I. et al. Stable X chromosome inactivation involves the PRC1 polycomb complex and requires histone MACROH2A1 and the CULLIN3/SPOP ubiquitin E3 ligase. *Proc. Natl. Acad. Sci. USA* **102**, 7635–7640 (2005).
- Bennett, E. P. et al. Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* **22**, 736–756 (2012).



49. Sano, K., Kawaguchi, M., Yoshikawa, M., Iuchi, I. & Yasumasu, S. Evolution of the teleostean zona pellucida gene inferred from the egg envelope protein genes of the Japanese eel, *Anguilla japonica*. *FEBS J.* **277**, 4674–4684 (2010).
50. Giulianini, P. G. & Ferrero, E. A. Ultrastructural aspects of the ovarian follicle and egg envelope of the sea-grass goby *Zosterisessor ophiocephalus* (Osteichthyes, Gobiidae). *Ital. J. Zool.* **68**, 29–37 (2001).
51. Smith, C. A. et al. The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* **461**, 267–271 (2009).
52. Yoshimoto, S. et al. A W-linked DM-domain gene, DM-W, participates in primary ovary development in *Xenopus laevis*. *Proc. Natl Acad. Sci. USA* **105**, 2469–2474 (2008).
53. Matsuda, M. et al. DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417**, 559–563 (2002).
54. Lau, E. S., Zhang, Z., Qin, M. & Ge, W. Knockout of zebrafish ovarian aromatase gene (*cyp19a1a*) by TALEN and CRISPR/Cas9 leads to all-male offspring due to failed ovarian differentiation. *Sci. Rep.* **6**, 37357 (2016).
55. Garcia-Ortiz, J. E. et al. Foxl2 functions in sex determination and histogenesis throughout mouse ovary development. *BMC Dev. Biol.* **9**, 1–21 (2009).
56. Olzmann, J. A. & Carvalho, P. Dynamics and functions of lipid droplets. *Nat. Rev. Mol. Cell Biol.* **20**, 137–155 (2019).
57. Gluchowski, N. L., Becuwe, M., Walther, T. C. & Farese, R. V. Jr. Lipid droplets and liver disease: from basic biology to clinical implications. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 343–355 (2017).
58. Lord, C. C. et al. Regulation of hepatic triacylglycerol metabolism by CGI-58 does not require ATGL co-activation. *Cell Rep.* **16**, 939–949 (2016).
59. Thiam, A. R., Farese, R. V. Jr. & Walther, T. C. The biophysics and cell biology of lipid droplets. *Nat. Rev. Mol. Cell Biol.* **14**, 775–786 (2013).
60. Labar, G., Wouters, J. & Lambert, D. M. A Review on the monoacylglycerol lipase: at the interface between fat and endocannabinoid signalling. *Curr. Med. Chem.* **17**, 2588–2607 (2010).
61. Li, Z., Berk, M., McIntyre, T. M., Gores, G. J. & Feldstein, A. E. The lysosomal-mitochondrial axis in free fatty acid-induced hepatic lipotoxicity. *Hepatology* **47**, 1495–1503 (2008).
62. Qu, Q., Zeng, F., Liu, X., Wang, Q. J. & Deng, F. Fatty acid oxidation and carnitine palmitoyltransferase I: emerging therapeutic targets in cancer. *Cell Death Dis.* **7**, 1–9 (2016).
63. Takaki, A., Kawai, D. & Yamamoto, K. Multiple hits, including oxidative stress, as pathogenesis and treatment target in non-alcoholic steatohepatitis (NASH). *Int. J. Mol. Sci.* **14**, 20704–20728 (2013).
64. Solbakken, M. H., Voje, K. L., Jakobsen, K. S. & Jentoft, S. Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system. *Proc. R. Soc. B* **284**, 1–9 (2017).
65. Buzzetti, E., Pinzani, M. & Tsochatzis, E. A. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). *Metabolism* **65**, 1038–1048 (2016).
66. Roh, Y. S. & Seki, E. Toll-like receptors in alcoholic liver disease, non-alcoholic steatohepatitis and carcinogenesis. *J. Gastroenterol. Hepatol.* **28** (Suppl.), 38–42 (2013).
67. Oldenburg, M. et al. TLR13 recognizes bacterial 23S rRNA devoid of erythromycin resistance-forming modification. *Science* **337**, 1111–1115 (2012).
68. Chen, Y. et al. SOAPnuka: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1–6 (2018).
69. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
70. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
71. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. S. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6**, 31900 (2016).
72. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
73. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
74. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* Chapter 4, Unit 4.10 (2009).
75. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
76. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
77. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
78. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
79. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
80. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
81. Elsik, C. G. et al. Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
82. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
83. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
84. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
85. Stamatakis, A. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
86. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
87. Hahn, M. W., Demuth, J. P. & Han, S. G. Accelerated rate of gene gain and loss in primates. *Genetics* **177**, 1941–1949 (2007).
88. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
89. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
90. Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.* **3**, 739–744 (1993).
91. Rastas, P., Calboli, F. C., Guo, B., Shikano, T. & Merila, J. Construction of ultradense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example. *Genome Biol. Evol.* **8**, 78–93 (2015).

## Acknowledgements

We acknowledge Y. Zhang for their support on the yellowstripe goby genome project. We thank M. H. Wu, X. Q. Chen, S. Q. Lai, Z. T. Lin, J. Zeng, and X. Y. Lin for their assistance in sample collection and W. L. Wu for preparing the graphs. We acknowledge grant support from the National Key Technologies R & D Program of China (Grant No. 2015BAI09B05).

## Author contributions

R.H. initiated the yellowstripe goby genome project. R.H., L.C., Y.W., G.L., and Y.Z. conceived the study. L.C., Y.W., G.L., and Y.Z. wrote and revised the manuscript and the supplementary materials. L.C. and Y.W. conducted the genome and transcriptome analysis and performed the comparative genomic and genome evolution studies. G.L. and Y.Z. conducted the genome assembly, annotation, comparative genomics analysis, genetic map construction, and transcriptomic analysis. J.L., Z.M., L.Y., M.C., and H.Y. conducted the sample preparation, and gene validation. Z.Y., Z.D. and W.S. supervised the sequencing, assembly, and bioinformatics analysis. M.C. coordinated the project.

## Competing interests

The authors declare no competing interests.

## Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s42003-020-01541-9>.

Correspondence and requests for materials should be addressed to L.C. or R.H.

Reprints and permission information is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021