GigaScience, 6, 2017, 1–13

doi: 10.1093/gigascience/gix039
Advance Access Publication Date: 23 May 2017
Research

RESEARCH

The draft genome of blunt snout bream (*Megalobrama amblycephala*) reveals the development of intermuscular bone and adaptation to herbivorous diet

Han Liu[1,†], Chunhai Chen[2,†], Zexia Gao[1,†], Jiumeng Min[2,†], Yongming Gu[3,†], Jianbo Jian[2,†], Xiewu Jiang[3], Huimin Cai[2], Ingo Ebersberger[4], Meng Xu[2], Xinhui Zhang[1], Jianwei Chen[2], Wei Luo[1], Boxiang Chen[1,3], Junhui Chen[2], Hong Liu[1], Jiang Li[2], Ruifang Lai[1], Mingzhou Bai[2], Jin Wei[1], Shaokui Yi[1], Huanling Wang[1], Xiaojuan Cao[1], Xiaoyun Zhou[1], Yuhua Zhao[1], Kaijian Wei[1], Ruibin Yang[1], Bingnan Liu[3], Shancen Zhao[2], Xiaodong Fang[2], Manfred Schartl[5,6,*], Xueqiao Qian[3,*] and Weimin Wang[1,*]

[1]College of Fisheries, Key Lab of Freshwater Animal Breeding, Ministry of Agriculture, Key Lab of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong Agricultural University, Wuhan 430070, China, [2]Beijing Genomics Institute (BGI)–Shenzhen, Shenzhen 518083, China, [3]Guangdong Haid Group Co., Ltd., Guangzhou 511400, China, [4]Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Frankfurt D-60438, Germany, [5]Physiological Chemistry, University of Würzburg, Biozentrum, Am Hubland, and Comprehensive Cancer Center Mainfranken, University Clinic Würzburg, Würzburg 97070, Germany and [6]Texas A&M Institute for Advanced Study and Department of Biology, Texas A&M University, College Station, TX 77843, USA

[*]**Correspondence address:** Weimin Wang, College of Fisheries, Huazhong Agricultural University, Wuhan 430070, China. E-mail: wangwm@mail.hzau.edu.cn; Xueqiao Qian, Guangdong Haid Group Co., Ltd., Guangzhou 511400, China. E-mail: xueqiaoqian@263.net; Manfred Schartl, Physiological Chemistry, University of Würzburg Biozentrum, Am Hubland, and Comprehensive Cancer Center Mainfranken, University Clinic Würzburg, Würzburg 97070, Germany. E-mail: phch1@biozentrum.uni-wuerzburg.de
[†]These authors contributed equally to the work.

Received: 8 September 2016; **Revised:** 23 November 2016; **Accepted:** 20 May 2017

Abstract

The blunt snout bream *Megalobrama amblycephala* is the economically most important cyprinid fish species. As an herbivore, it can be grown by eco-friendly and resource-conserving aquaculture. However, the large number of intermuscular bones in the trunk musculature is adverse to fish meat processing and consumption. As a first towards optimizing this aquatic livestock, we present a 1.116-Gb draft genome of *M. amblycephala*, with 779.54 Mb anchored on 24 linkage groups. Integrating spatiotemporal transcriptome analyses, we show that intermuscular bone is formed in the more basal teleosts by intramembranous ossification and may be involved in muscle contractibility and coordinating cellular events. Comparative analysis revealed that olfactory receptor genes, especially of the beta type, underwent an extensive expansion in herbivorous cyprinids, whereas the gene for the umami receptor *T1R1* was specifically lost in *M. amblycephala*. The composition of gut microflora, which contributes to the herbivorous adaptation of *M. amblycephala,* was found to be similar to that of other herbivores. As a valuable resource for the improvement of *M. amblycephala* livestock, the draft genome sequence offers new insights into the development of intermuscular bone and herbivorous adaptation.

Keywords: *Megalobrama amblycephala*; whole genome; herbivorous diet; intermuscular bone; transcriptome; gut microflora

Background

Fishery and aquaculture play an important role in global alimentation. Over the past decades, food fish supply has been increasing, with an annual rate of 3.6%, about 2 times faster than the human population [1]. This growth of fish production is meanwhile solely accomplished by an extension of aquaculture as over the past 30 years the total mass of captured fish has remained almost constant [1]. As a consequence of this emphasis on fish breeding, the genomes of various economically important fish species, e.g., Atlantic cod (*Gadus morhua*) [2], rainbow trout (*Oncorhynchus mykiss*) [3], European sea bass (*Dicentrarchus labrax*) [4], yellow croaker (*Larimichthys crocea*) [5], half-smooth tongue sole (*Cynoglossus semilaevis*) [6], tilapia (*Oreochromis niloticus*) [7], and channel catfish (*Ictalurus punctatus*) [8], have been sequenced. Yet, the majority of these species are carnivorous, requiring large inputs of protein from wild caught fish or other precious feed. Reports on draft genomes of herbivorous and omnivorous species, in particular cyprinid fish, are scarce. It is well known that cyprinids are currently the economically most important group of teleosts for sustainable aquaculture. They grow to large population sizes in the wild and already now account for the majority of freshwater aquaculture production worldwide [1]. Among these, the herbivorous blunt snout bream, *Megalobrama amblycephala* (Fishbase Sp. ID: 285) (Fig. 1), a particularly eco-friendly and resource-conserving species, is predominant in aquaculture and has been greatly developed in China (Additional file 1: Fig. S1) [1]. However, most cyprinids, including *M. amblycephala*, have a large number of intermuscular bones (IBs) in the trunk musculature, which have an adverse effect on fish meat processing and consumption. IBs—a unique form of bone occurring only in the more basal teleosts—are completely embedded within the myosepta and are not connected to the vertebral column or any other bones [9, 10]. Our previous study on IB development of *M. amblycephala* revealed that some miRNA-mRNA interaction pairs may be involved in regulating bone development and differentiation [11]. However, the molecular genetic basis and the evolution of this unique structure are still unclear. Unfortunately, the recent sequencing of 2 cyprinid genomes, common carp (*Cyprinus carpio*) [12] and grass carp (*Ctenopharyngodon idellus*) [13], which provided valuable information for their genetic breeding, contributed little to the understanding of IB formation.

In an initial genome survey of *M. amblycephala*, we identified 25 697 single-nucleotide polymorphisms (SNPs) [14], 347 conserved miRNAs [15], and 1136 miRNA-mRNA interaction pairs [11]. However, lack of a whole-genome sequence resource limited a thorough investigation of *M. amblycephala*. Here we report the first high-quality draft genome sequence of *M. amblycephala*. Integrating this novel genome resource with tissue- and developmental stage–specific gene expression information, as well as with meta-genome data to investigate the composition of the gut microbiome (workflow shown in Additional file 1: Fig. S2), provides relevant insights into the function and evolution of 2 key features characterizing this species: The formation of IB and the adaptation to herbivory. By that our study lays the foundation for genetically optimizing *M. amblycephala* to further increase its relevance for securing human food supply.

Data Description

Genome assembly and annotation

The *M. amblycephala* genome was sequenced and assembled by a whole-genome shotgun strategy using genomic DNA from a double haploid fish (Additional file 1: Table S1). We assembled a 1.116-Gb reference genome sequence from 142.55 Gb (approximately 130-fold coverage) of clean data (Additional file 1: Tables S1 and S2, Fig. S3) [16]. The contig and scaffold N50 lengths reached 49 Kb and 839 Kb, respectively (Table 1). The largest scaffold spans 8951 Kb, and the 4034 largest scaffolds cover 90% of the assembly. To assess the quality of genome assembly, the short-insert size paired-end library reads and published expressed sequence tags (ESTs) [14] (Additional file 1: Tables S3 and S4) were mapped onto the genome. The results indicated that the assembled error is low. To further estimate the completeness of the assembly and gene prediction, benchmarking universal single-copy orthologs (BUSCO; RRID:SCR_015008) [17] analysis was used, and the results showed that the assembly contains 81.4% complete and 9.1% partial vertebrate BUSCO orthologs (Additional file 1: Table S5).

The *M. amblycephala* genome has an average GC content of 37.3%, similar to cyprinid *C. carpio* and *Danio rerio* (Additional file 1: Fig. S4). Using a comprehensive annotation strategy combining RNA-seq-derived transcript evidence, *de novo* gene prediction, and sequence similarity to proteins from 5 further fish species, we annotated a total of 23 696 protein-coding genes (Additional file 1: Table S6). Of the predicted genes, 99.44% (23 563 genes) are annotated by functional database. In addition, we identified 1796 non-coding RNAs including 474 miRNAs,

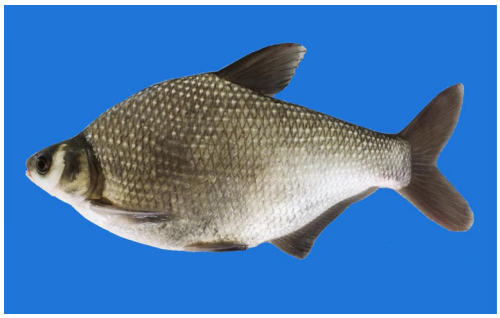

Figure 1: Image of an adult blunt snout bream (*Megalobrama amblycephala*).

Table 1: Features of the *M. amblycephala* whole-genome sequence.

Total genome size (Mb)	1116
N90 length of scaffold (bp)	20 422
N50 length of scaffold (bp)	838 704
N50 length of contig (bp)	49 400
Total GC content (%)	37.30
Protein-coding genes number	23 696
Average gene length (bp)	15 797
Content of transposable elements (%)	34.18
Number of chromosomes	24
Number of makers in genetic map	5317
Scaffolds anchored on linkage groups (LGs)	1434
Length of scaffolds anchored on LGs (Mb)	779.54 (70%)

220 rRNA, 530 tRNAs, and 572 snRNAs. Transposable elements (TEs) comprise approximately 34.18% (381.3 Mb) of the *M. amblycephala* genome (Additional file 1: Table S7). DNA transposons (23.80%) and long terminal repeat retrotransposons (LTRs; 9.89%) are the most abundant TEs in *M. amblycephala*. The proportion of LTRs in *M. amblycephala* is the highest in comparison with other teleosts: *G. morhua* (4.88%) [2], *L. crocea* (2.2%) [5], *C. semilaevis* (0.08%) [6], *C. carpio* (2.28%) [12], *C. idellus* (2.58%) [13], and stickleback (*Gasterosteus aculeatus*; 1.9%) [18] (Additional file 1: Tables S7 and S8, Fig. S5). The distribution of divergence between the TEs in *M. amblycephala* peaks at 7% (Additional file 1: Fig. S6), indicating a more recent activity of these TEs when compared with *O. mykiss* (13%) [3] and *C. semilaevis* (9%) [6].

Anchoring scaffolds and shared synteny analysis

Sequencing data from 198 F1 specimens, including the parents, were used as the mapping population to anchor the scaffolds onto 24 pseudo-chromosomes of the *M. amblycephala* genome. Following RAD-seq and sequencing protocol, 1883.5 Mb of 125-bp reads (on average 30.6 Mb and 9.3 Mb of read data for each parent and progeny, respectively) were generated on the HiSeq 2500 next-generation sequencing platform. Based on the SOAP bioinformatic pipeline (SOAPdenovo2, RRID:SCR_014986), we generated 5317 SNP markers for constructing a high-resolution genetic map. The map spans 1701 cM with a mean marker distance of 0.33 cM and facilitated an anchoring of 1434 scaffolds comprising 70% (779.54 Mb) of the *M. amblycephala* genome assembly to form 24 linkage groups (LGs) (Additional file 1: Table S9). Of the anchored scaffolds, 598 could additionally be oriented (678.27 Mb, 87.01% of the total anchored sequences) (Fig. 2A). A subsequent comparison of the gene order between *M. amblycephala* and its close relative *C. idellus* revealed 607 large shared syntenic blocks, encompassing 11 259 genes, and 190 chromosomal rearrangements. The values change to 1062 regions, 13 152 genes, and 279 rearrangements when considering *D. rerio*. The unexpected higher number of genes in syntenic regions shared with the more distantly related *D. rerio* is most likely an effect of the more complete genome assembly of this species compared to *C. idellus*. The rearrangement events are distributed across all *M. amblycephala* linkage groups without evidence for a local clustering (Fig. 2B). The most prominent event is the chromosomal fusion in *M. amblycephala* LG02 that joined 2 *D. rerio* chromosomes, Dre10 and Dre22. The same fusion is observed in *C. idellus* but not in *C. carpio*, suggesting that it probably occurred in a last common ancestor of *M. amblycephala* and *C. idellus* approximately 13.1 million years ago (Additional file 1: Fig. S7).

Results

Evolutionary analysis

A phylogenetic analysis of 316 single-copy orthologous genes in the genomes of 10 other fish species, coelacanth (*Latimeria chalumnae*), and elephant shark (*Callorhinchus milii*) as outgroup served as a basis for investigating the evolutionary trajectory of *M. amblycephala* (Fig. 3A; Additional file 1: Fig. S8). We found 9349 orthologous gene families shared among 5 fish species. A total of 246 are specific to the *M. Amblycephala* (Fig. 3B). To illuminate the evolutionary process resulting in the adaptation to a grass diet, we analyzed the functional categories of expanded genes in the *M. amblycephala* and *C. idellus* lineage (Additional file 1: Fig. S9, Additional file 2: Data Note1), 2 typical herbivores mainly feeding on aquatic and terrestrial grasses. Among the significantly over-represented KEGG pathways (KEGG, RRID:SCR_012773; Fisher's exact test, $P < 0.01$), we find olfactory transduction (ko04740), immune-related pathways (ko04090, ko04672, ko04612, and ko04621), lipid metabolic-related process (ko00590, ko03320, ko00591, ko00565, ko00592, and ko04975), and xenobiotics biodegradation and metabolism (ko00625 and ko00363) (Fig. S10). Indeed, when tracing positively selected genes (PSG) in *M. amblycephala* and *C. idellus* (Additional file 3: Date Note2), we identified 10 candidates involved in starch and sucrose metabolism (ko00500), in citrate cycle (ko00020), and in other types of O-glycan biosynthesis (ko00514). Moreover, 10 genes encoding enzymes involved in lipid metabolism appear positively selected in both fish species (Additional file 1: Table S10).

Development of intermuscular bones

To explain the genetic basis of IBs, their formation, and their function in cyprinids, we first analyzed the functional annotation of genes that expanded in this lineage (Fig. 3C). Many of these genes are involved in cell adhesion (GO:0007155, $P = 5.26E-32$, 357 genes), myosin complex (GO:0016459, $P = 2.74E-08$, 100 genes), and cell-matrix adhesion (GO:0007160, $P = 1.59E-21$, 69 genes) (Fig. 3C). As a second line of evidence, we performed transcriptome analyses of early developmental stages (stage 1: whole larvae without IBs) and juvenile *M. amblycephala* (stage 2: trunk muscle with partial IBs; stage 3: trunk muscle with completed IBs) (Fig. 4A). Compared with stage 1, 388 and 651 differentially expressed genes (DEGs) are up-regulated in stage 2 and stage 3, respectively. And 249 of them are significantly up-regulated both in stage 2 and stage 3. KEGG analyses indicate that many of these genes are involved in tight junction (ko04530), regulation of actin cytoskeleton (ko04810), cardiac muscle contraction (ko04260), and vascular smooth muscle contraction (ko04270) (Additional file 1: Fig. S11). Specifically, 26 genes encoding proteins related to muscle contraction, including titin, troponin, myosin, actinin, calmodulin, and other

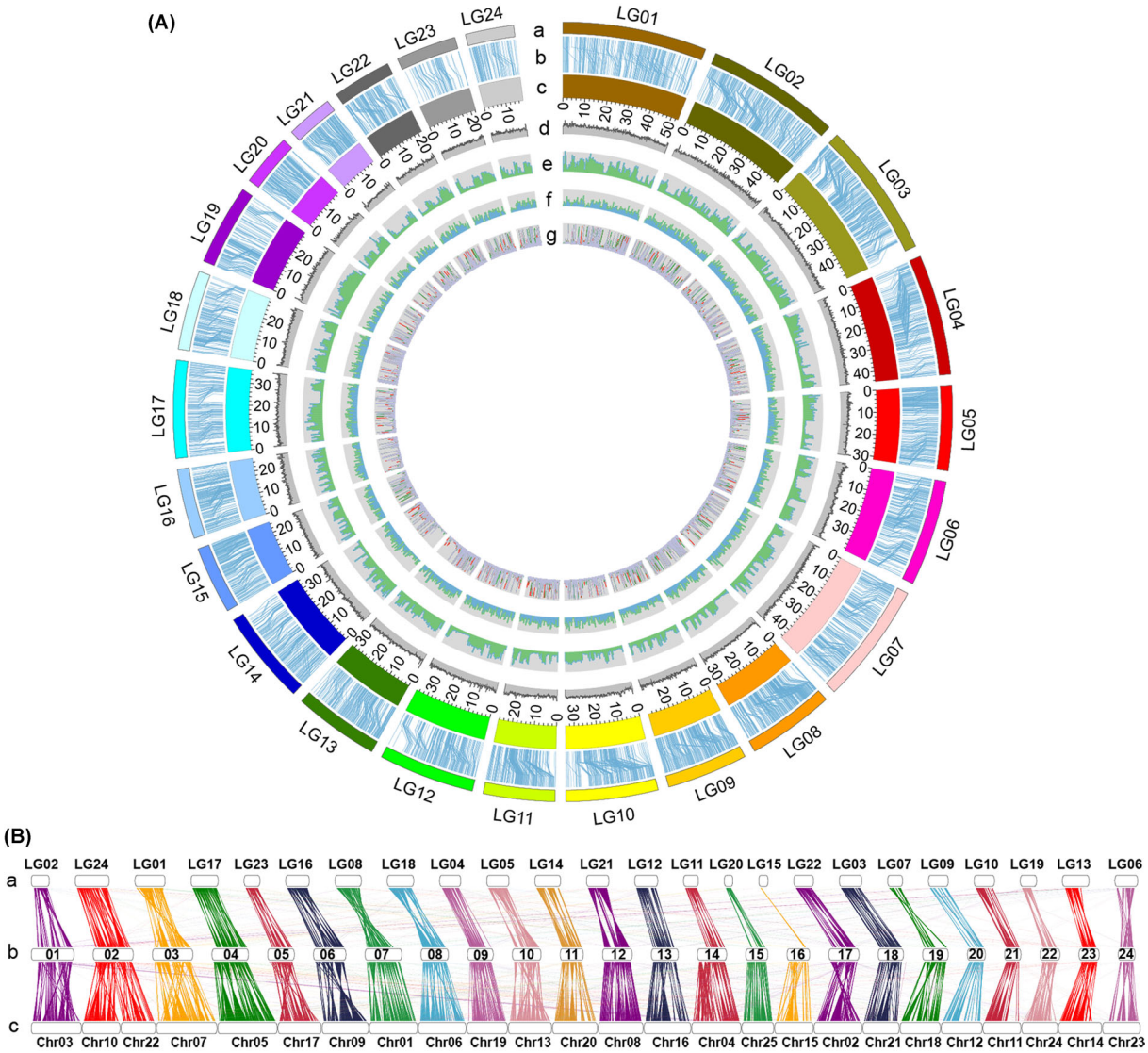

Figure 2: Global view of the *M. amblycephala* genome and syntenic relationship between *C. idellus*, *M. amblycephala*, and *D. rerio*. (**A**) Global view of the *M. amblycephala* genome. From outside to inside, the genetic linkage map (**a**); anchors between the genetic markers and the assembled scaffolds (**b**); assembled chromosomes (**c**); GC content within a 50-kb sliding window (**d**); repeat content within a 500-kb sliding window (**e**); gene distribution on each chromosome (**f**); and different gene expression of 3 transcriptomes (**g**). (**B**) Syntenic relationship between the *C. idellus* (**a**), *M. amblycephala* (**b**), and *D. rerio* (**c**) chromosomes.

Ca^{2+}-transporting ATPases (Fig. 4A) point to a strong remodeling of the musculature compartment. To confirm that the observed differences in gene expression are indeed linked to IB formation and function and are not simply due to the fact that different developmental stages were compared, we performed differential expression analysis of muscle tissues, IBs, and connective tissues from the same 6-month-old individual of *M. amblycephala* (Fig. 4B; Additional file 1: Fig. S12); 1290 DEGs and 5231 DEGs are significantly up-regulated in IBs compared with connective tissues and muscle, respectively. Twenty-four of these DEGs encode extracellular matrix (ECM) proteins (collagens and intergrin-binding protein), Rho GTPase family (*RhoA*, *Rho GAP*, *Rac*, *Ras*), motor proteins (myosin, dynein, actin), and calcium channel regulation proteins (Additional file 1: Fig. S13 and Table S11). In addition, GO annotations of 963 IB-specific genes indicative of abundance in protein binding (GO:0005515), calcium ion binding (GO:0005509), GTP binding (GO:0005525), and iron ion binding (GO:0005506) were found (Fig. 4C).

During development of *M. amblycephala*, the first IB appears in muscles of caudal vertebrae as early as 28 days post fertilization (dpf), when body length is 12.95 mm (Additional file 1: Fig. S14). The system then develops and ossifies predominantly from posterior to anterior (Additional file 1: Fig. S15). IBs are present throughout the body within 2 months (Additional file 1: Fig. S16) and develop into multiple morphological types in adults (Additional file 1: Fig. S17). The bone is formed directly without an intermediate cartilaginous stage (Additional file 1: Figs S18 and S19). We also found a large number of mature osteoblasts distributed at the edge of the bone matrix while some osteocytes were apparent in the center of the mineralized bone matrix (Additional file 1: Figs S20 and S21). These primary bone-forming cells predominantly regulate bone formation and function throughout life. Notably, among the genes up-regulated in IB, 35 bone formation regulatory genes were identified (shown in colored boxes in Fig. 4D). In particular, genes involved in bone morphogenetic protein (BMP) signaling including *Bmp3*, *Smad8*, *Smad9*, and *Id2*, in fibroblast growth factor (FGF) signaling

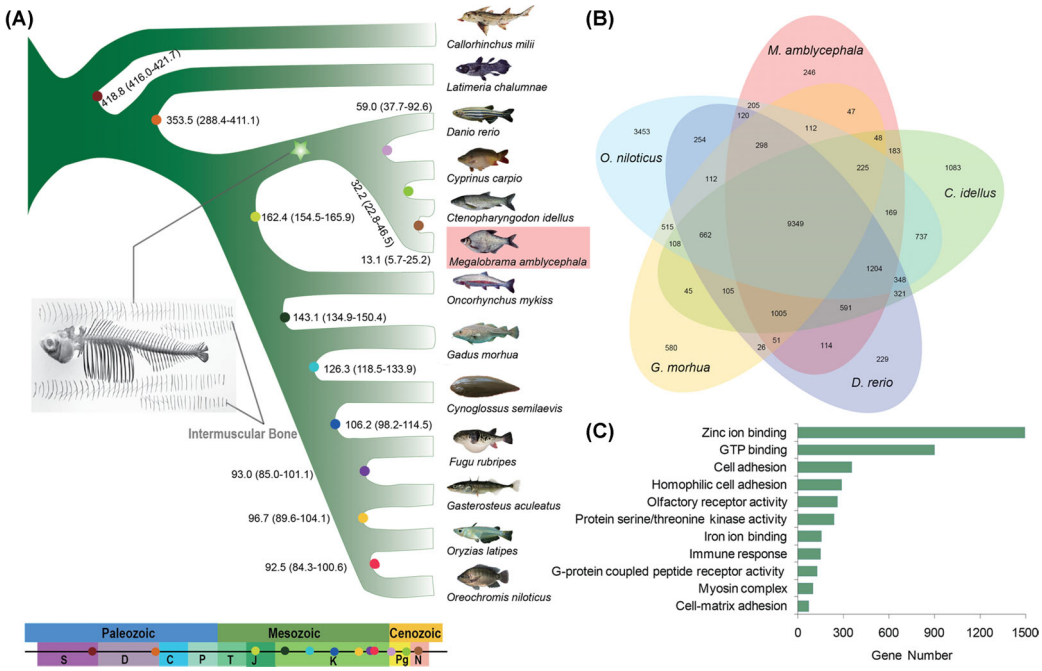

Figure 3: Phylogenetic tree and comparison of orthologous genes in *M. amblycephala* and other fish species. (**A**) Phylogenetic tree of teleosts using 316 single copy orthologous genes. The color circles at the nodes show the estimated divergence times using *O. latipes*–*F. rubripes* (96.9∼150.9 Mya), *F. rubripes*–*D. rerio* (149.85∼165.2 Mya), *F. rubripes*–*C. milii* (416∼421.75 Mya; http://www.timetree.org/) as the calibration time. The pentagram represents 4 cyprinid fish with intermuscular bones. S: Silurian period; D: Devonian period; C: Carboniferous period; P: Permian period in Paleozoic; T: Triassic period; J: Jurassic; K: cretaceous period in Mesozoic; Pg: Paleogene in Cenozoic Era; N: Neogene. (**B**) Venn diagram of shared and unique orthologous gene families in *M. amblycephala* and 4 other teleosts. (**C**) Over-represented GO annotations of cyprinid-specific expansion genes.

including *Fgf2*, *Fgfr1a*, *Fgfbp2*, *Col6a3*, and *Col4a5*, and in Ca^{2+} channels including *Cacna1c*, *CaM*, *Creb5*, and *Nfatc* were highly expressed (>2-fold change) in IB (Additional file 1: Fig. S22).

Adaptation to herbivorous diet

Next to the presence of IB, the herbivory of *M. amblycephala* is the second key feature in connection to the use of this species as aquatic livestock. Olfaction, the sense of smell, is crucial for animals to find food. The perception of smell is mediated by a large gene family of olfactory receptor (OR) genes. In the *M. amblycephala* genome, we identified 179 functional olfactory receptor genes (Fig. 5A), and based on the classification of Niimura [19], 158, 117, and 153 receptors for water-borne odorants were identified in *M. amblycephala*, *C. idellus*, and *D. rerio*, respectively (Additional file 1: Table S12). Overall, these receptor repertoires are substantially larger than those of other and carnivorous teleosts (*G. morhua*, *C. semilaevis*, *O. latipes*, *X. maculatus*) (Additional file 1: Figs S23 and S24, Table S12). In addition, we found a massive expansion of beta-type OR genes in the genomes of the herbivorous *M. amblycephala* and *C. idellus*, while very few exist in other teleosts (Fig. 5B; Additional file 1: Table S12).

Taste is also an important factor in the development of dietary habits. Most animals can perceive 5 basic tastes, namely sourness, sweetness, bitterness, saltiness, and umami [20]. *T1R1*, the receptor gene necessary for sensing umami, has been lost in herbivorous *M. amblycephala* but is duplicated in carnivorous *G. morhua* and *C. semilaevis* and omnivorous *O. latipes* and *X. maculatus* (Figs 5C and D; Additional file 1: Figs S25–S6 and Table S13). In contrast, *T1R2*, the receptor gene for sensing sweet, has been duplicated in herbivorous *M. amblycephala* and *C. idellus* and omnivorous *C. carpio* and *D. rerio*, while it has been lost in carnivorous *G. morhua* and *C. semilaevis* (Additional file 1: Fig. S27 and Table S13). Also the *T2R* gene family, most likely important in the course switching to a diet that contains a larger fraction of bitterness-containing food, has been expanded in *M. amblycephala*, *C. idellus*, *C. carpio*, and *D. rerio* (Additional file 1: Fig. S28).

To obtain further insights into the genetic adaptation to herbivorous diet, we focused on further genes that might be associated with digestion. Genes that encode proteases (including pepsin, trypsin, cathepsin, and chymotrypsin) and amylases (including alpha-amylase and glucoamylase) were identified in the genomes of *M. amblycephala*, carnivorous *C. semilaevis and G. morhua* and omnivorous *D. rerio, O. latipes*, and *X. maculatus*, indicating that herbivorous *M. amblycephala* has a protease repertoire that is not substantially different from those of carnivorous and omnivorous fishes (Additional file 1: Table S14). We did not identify any genes encoding potentially cellulose-degrading enzymes including endoglucanase, exoglucanase, and beta-glucosidase in the genome of *M. amblycephala*, suggesting that utilization of the herbivorous diet may largely depend on the gut microbiome. To elucidate this further, we determined the composition of the gut microbial communities of juvenile, domestic, and wild adult *M. amblycephala* and wild adult *C. idellus* using bacterial 16S rRNA sequencing. A total of 549 020 filtered high-quality sequence reads from 12 samples were clustered at a similarity level of 97%. The resulting 8558 operational taxonomic units (OTUs) are dominated at phylum level by Proteobacteria, Fusobacteria, Bacteroidetes, Firmicutes, and Actinobacteria (Fig. 5E; Additional file 1: Table S15). Increasing the resolution to the genus level, the composition and relative abundance of the gut microbiota of wild adult *M. amblycephala* and *C. idellus* are still very similar (Additional file 1: Table S16), and we could identify more than 7% cellulose-degrading bacteria (Additional file 1: Table S17).

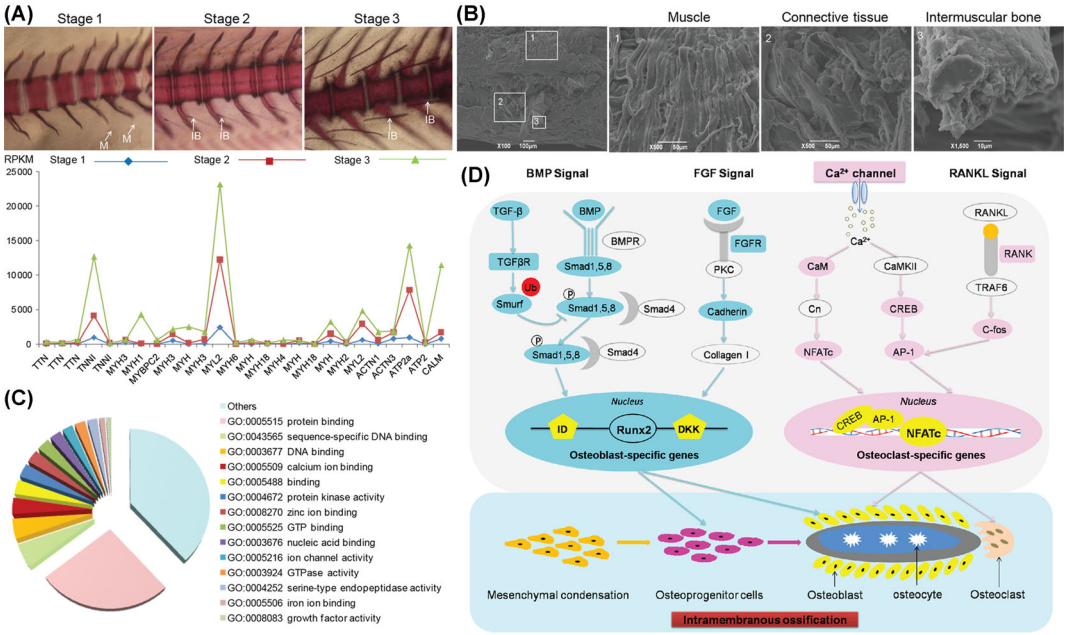

Figure 4: Regulation of genes related to intermuscular bone formation and function identified from developmental stages and adult tissues transcriptome data. (**A**) The gene expression pattern involved in muscle contraction–regulated genes in early developmental stages corresponds to the intermuscular bone formation of *M. amblycephala* (alizarin red staining). M: myosepta. (**B**) Scanning electron microscope photos of muscle tissues, connective tissues, and intermuscular bone. (**C**) Distribution of intermuscular bone–specific genes in GO annotations indicative of abundance in protein binding, calcium ion binding, and GTP binding functions. (**D**) Several developmental signals regulating key steps of osteoblast and osteoclast differentiation in the process of intramembranous ossification. Colored boxes indicate that significantly up-regulated genes in these signals specifically occurred in intermuscular bone.

Discussion

The evolutionary trajectory analysis of *M. amblycephala* and other teleosts revealed that *M. amblycephala* has the closest relationship to *C. idellus*. Both the species are herbivorous fish, but which endogenous and exogenous factors affected their feeding habits and how they adapted to their herbivorous diet are not known. Our results from the expanded genes and PSG in the lineage of the 2 herbivores uncovered a number of genes that are involved in glucose, lipid, and xenobiotics metabolism, which would enhance the ability of an herbivore to detoxify the secondary compounds present in grasses that are adverse or even toxic to the organism. Furthermore, the high-fiber but low-energy grass diet requires a highly effective intermediate metabolism that accelerates carbohydrate and lipid catabolism and conversion into energy to maintain physiological functions.

Olfaction and taste are also crucial for animals to find food and to distinguish whether potential food is edible or harmful [21, 22]. The ORs of teleosts are predominantly expressed in the main olfactory epithelium of the nasal cavity [21, 23] and can discriminate, like those of other vertebrates, different kinds of odor molecules. Previous studies have demonstrated that the beta type OR genes are present in both aquatic and terrestrial vertebrates, indicating that the corresponding receptors detect both water-soluble and airborne odorants [19, 21]. In the present study, the search for genes encoding OR showed that herbivorous *M. amblycephala* and *C. idellus* have a large number of beta-type ORs, while other omnivorous and carnivorous fish only have 1 or 2. This might be attributed to their particular herbivorous diet consisting not only of aquatic grasses but also the duckweed and terrestrial grasses, which they ingest from the water surface.

It is known that the receptor for umami is formed by the T1R1/T1R3 heterodimer, while T1R2/T1R3 senses sweet taste [24]. We found that the umami gene *T1R1* was lost in herbivorous *M. amblycephala* but duplicated in the carnivorous *G. morhua* and *C. Semilaevis*. The loss of the *T1R1* gene in *M. amblycephala* might exclude the expression of a functional umami taste receptor. Such situations in other organisms, e.g., the Chinese panda, have previously been related to feeding specialization [25]. Bitterness sensed by the *T2R* is particularly crucial for animals to protect them from poisonous compounds [22]. Interestingly, the bitter receptor *T2R* genes are expanded in herbivorous fish, but few or no copies were found in carnivorous fish. These results not only indicate the genetic adaptation to a herbivorous diet of *M. amblycephala*, but also provide a clear and comprehensive picture of adaptive evolutionary mechanisms of sensory systems in other fish species with different trophic specializations.

It has been reported that some insects such as *Tenebrio molitor* [26] and *Neotermes koshunensis* [27] and the mollusc *Corbicula japonica* [28] have genes encoding endogenous cellulose degradation–related enzymes. However, so far all analyzed herbivorous vertebrates lack these genes and always rely on their gut microbiome to digest food [25, 29]. In herbivorous *M. amblycephala* and *C. idellus*, we also did not find any homologs of digestive cellulase genes. Interestingly, our work on the composition of gut microbiota of the 2 fish species identifies more than 7% cellulose-degrading bacteria, suggesting that the cellulose degradation of herbivorous fish largely depends on their gut microbiome.

IB has evolved several times during teleost evolution [9, 30]. The developmental mechanisms and ossification processes forming IB are dramatically distinct from other bones such as ribs, skeleton, vertebrae, or spines. These usually develop from cartilaginous bone and are derived from the mesenchymal cell population by endochondral ossification [31, 32]. However, IBs form directly by intramembranous ossification and

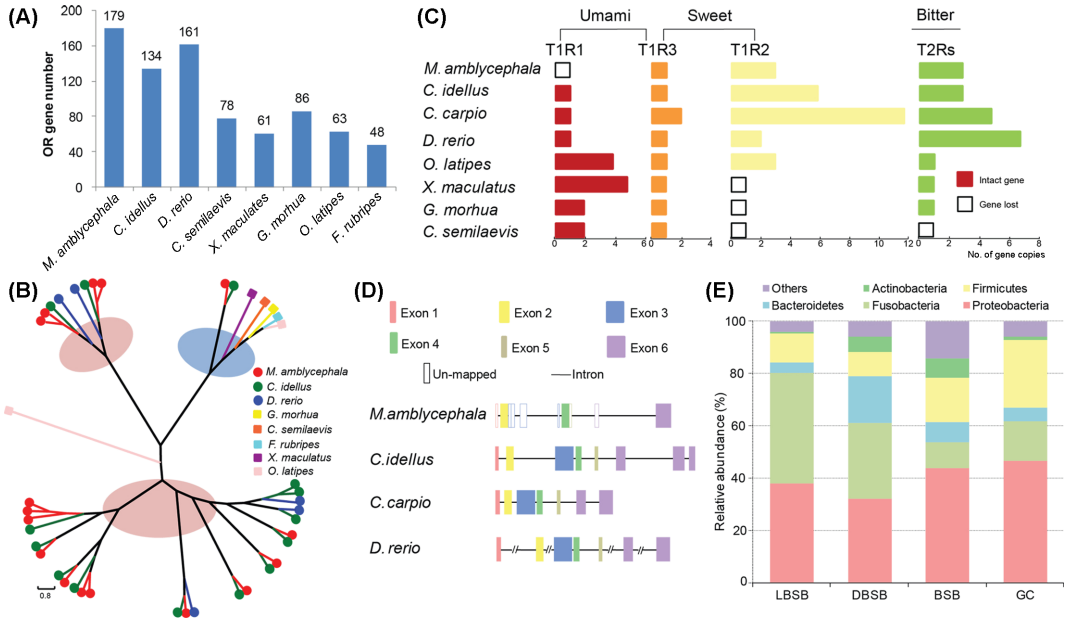

Figure 5: Molecular characteristics of sensory systems and the composition of gut microbiota in *M. amblycephala*. (**A**) Extensive expansion of olfactory receptor genes (ORs) in *M. amblycephala* compared with other teleosts. (**B**) Phylogeny of "beta" type ORs in 8 representative teleost species showing the significant expansion of "beta" ORs in *M. amblycephala* and *C. idellus*. The pink background shows cyprinid-specific "beta" types of ORs. (**C**) Umami, sweet, and bitter taste–related gene families in teleosts with different feeding habits. (**D**) Structure of the umami receptor encoding T1R1 gene in cyprinid fish. (**E**) Relative abundance of microbial flora and taxonomic assignments in juvenile (LBSB), domestic adult (DBSB), and wild adult (BSB) *M. amblycephala* and wild adult *C. idellus* (GC) samples at the phylum level.

differentiate from osteoblasts within connective tissue, forming segmental, serially homologous ossifications in the myosepta. Although various methods of ossification of IB have been proposed, few experiments have been conducted to confirm the ossification process, and little is known about the potential role of IB in teleosts. Based on our findings of expanded genes in cyprinid lineage and evidence from the transcriptome of the developmental stages of IB formation, a number of genes were found to interact dynamically to mediate efficient cell motility, migration, and muscle construction [33–36]. In addition, transcriptome analyses of 3 tissues indicated that ECM, Rho GTPase, and motor and calcium channel regulation protein displayed high expression in IB. It is known that ECM proteins bound to integrins influence cell migration by actomyosin-generated contractile forces [34, 37]. Rho GTPases, acting as molecular switches, are also involved in regulating the actin cytoskeleton and cell migration, which in turn initiates intracellular signaling and contributes to tissue repair and regeneration [38–40]. Thus, our results provide molecular evidence that IBs might play significant roles not only in regulating muscle contraction but also in active remodeling at the bone-muscle interface and coordination of cellular events.

Some major developmental signals including BMP, FGF, and WNT, together with calcium/calmodulin signaling [31, 41–43], are essential for regulating the differentiation and function of osteoblasts and osteocytes and for regulating the RANKL signaling pathway for osteoclasts [44]. In agreement with this concept, we found that 35 bone formation regulatory genes involved in these signals were highly up-regulated in IBs. Among these signaling pathways, in particular, *Bmp*, *Fgf2*, and *Fgfr1* are closely related to intramembranous bone development and affect the expression and activity of other osteogenesis-related transcription factors [31, 45]. The calcium-sensitive transcription factor *NFATc1*, together with *CREB*, induces the expression of osteoclast-specific genes [46]. Taken together, these results suggest that IB indeed undergoes an intramembranous ossification process, is regulated by bone-specific signaling pathways, and underlies a homeostasis of maintenance, repair, and remodeling.

Conclusions

Our results provide novel functional insights into the evolution of cyprinids. Importantly, the *M. amblycephala* genome data come up with novel insights, shedding light on the adaptation to herbivorous nutrition and the evolution and formation of IB. Our results on the evolution of gene families, as well as the digestive and sensory system, and our microbiome meta-analysis and transcriptome data provide powerful evidence and a key database for future investigations to increase the understanding of the specific characteristics of *M. amblycephala* and other fish species.

Methods

Sampling and DNA extraction

DNA for genome sequencing was derived from a double haploid fish from the *M. amblycephala* genetic breeding center at Huazhong Agricultural University (Wuhan, Hubei, China). Fish blood was collected from adult female fish caudal vein using sterile injectors with pre-added anticoagulant solutions following anesthetized with MS-222 and sterilization with 75% alcohol. Genomic DNA was extracted from the whole blood.

Genomic sequencing and assembly

Libraries with different insert sizes of 170 bp, 500 bp, 800 bp, 2 Kb, 5 Kb, 10 Kb, and 20 Kb were constructed from the genomic DNA at BGI-Shenzhen. The libraries were sequenced

using a HiSeq2000 instrument. In total, 11 libraries, sequenced in 23 lanes, were constructed. To obtain high-quality data, we applied filtering criteria for the raw reads. As a result, 142.55 Gb of filtered data were used to complete the genome assembly using SOAPdenovo_V2.04 (SOAPdenovo2, RRID:SCR_014986) [16]. Only filtered data were used in the genome assembly. First, the short–insert size library data were used to construct a de Bruijn graph. The tips, merged bubbles, and connections with low coverage were removed before resolving the small repeats. Second, all high-quality reads were realigned with the contig sequences. The number of shared paired-end relationships between pairs of contigs was calculated and weighted with the rate of consistent and conflicting paired ends before constructing the scaffolds in a stepwise manner from the short–insert size paired ends to the long–insert size paired ends. Third, the gaps between the constructed scaffolds were composed mainly of repeats, which were masked during scaffold construction. These gaps were closed using the paired-end information to retrieve read pairs in which 1 end mapped to a unique contig and the other was located in the gap region. Subsequently, local assembly was conducted for these collected reads. To assess the genome assembly quality, approximately 42.82 Gb of Illumina reads generated from short–insert size libraries were mapped onto the genome. Bwa0.5.9-r16 software (BWA, RRID:SCR_010910) [47] with default parameters was used to assess the mapping ratio, and SOAP coverage 2.27 was used to calculate the sequencing depth. We also assessed the accuracy of the genome assembly by Trinity (Trinity, RRID:SCR_013048) [48], including number of ESTs and new mRNA reads from early stages of embryos and multiple tissues, by aligning the scaffolds to the assembled transcriptome sequences.

After obtaining K-mers from the short–insert size (<1 Kb) reads with just 1 bp slide, frequencies of each K-mer were calculated. The K-mer frequency fits the Poisson distribution when a sufficient amount of data is present. The total genome size was deduced from these data in the following way: genome size = K–mer num/Peak_depth.

Genome annotation

The genome was searched for repetitive elements using Tandem Repeats Finder (version 4.04) [49]. TEs were identified using homology-based approaches. The Repbase (version 16.10) [50] database of known repeats and a *de novo* repeat library generated by RepeatModeler (RepeatModeler, RRID:SCR_015027) were used. This database was mapped using the software of RepeatMasker (RepeatMasker, RRID:SCR_012954; version 3.3.0). Four types of non-coding RNAs (microRNAs, transfer RNAs, ribosomal RNAs, and small nuclear RNAs) were also annotated using tRNAscan-SE (version 1.23) and the Rfam database45 (Rfam, RRID:SCR_007891; release 9.1) [51].

For gene prediction, *de novo* gene prediction, homology-based methods, and RNA-seq data were used to perform gene prediction. For the sequence similarity–based prediction, *D. rerio, G. aculeatus, O. niloticus, O. latipes,* and *G. morhua* protein sequences were downloaded from Ensembl (Ensembl, RRID:SCR_002344; release 73) and were aligned to the *M. amblycephala* genome using TBLASTN (TBLASTN, RRID:SCR_011822). Then homologous genome sequences were aligned against the matching proteins using GeneWise [52] to define gene models. Augustus was employed to predict coding genes using appropriate parameters in *de novo* prediction. For the RNA-seq-based prediction, we mapped transcriptome reads to the genome assembly using TopHat (TopHat, RRID:SCR_013035) [53]. Then, we combined TopHat mapping results together and applied Cufflinks (Cufflinks, RRID:SCR_014597) [54] to predict transcript structures. All predicted gene structures were integrated by GLEAN [55] to obtain a consensus gene set. Gene functions were assigned to the translated protein-coding genes using the Blastp tool (BLASTP, RRID:SCR_001010), based on their highest match to proteins in the SwissProt and TrEMBL [56] databases (UniProt, RRID:SCR_002380; Uniprot release 2011–01). Motifs and domains in the protein-coding genes were determined by InterProScan (InterProScan, RRID:SCR_005829; version 4.7) searches against 6 different protein databases: ProDom, PRINTS, Pfam, SMART, PANTHER, and PROSITE. Gene ontology (GO: RRID:SCR_002811) [57] IDs for each gene were obtained from the corresponding InterPro entries. All genes were aligned against the KEGG (KEGG, RRID:SCR_012773; release 58) [58] database, and the pathway in which the gene might be involved was derived from the matched genes in KEGG. tRNA genes were *de novo* predicted by tRNAscan-SE software (tRNAscan-SE, RRID:SCR_010835) [59], with eukaryote parameters on the repeat pre-masked genome. The rRNA fragments were identified by aligning the rRNA sequences using BlastN at E-value 1e-5 (BLASTN, RRID:SCR_001598). The snRNA and miRNA were searched by the method of aligning and searching with INFERNAL (Infernal, RRID:SCR_011809; version 0.81) [60] against the Rfam database (Rfam, RRID:SCR_007891; release 9.1).

Genetic map construction

To anchor the scaffolds into pseudo-chromosomes, 198 F1 population individuals were used to obtain the genetic map. Each of the individual genomic DNA was digested with the restriction endonuclease EcoR I, following the RAD-seq protocol [61]. The SNP calling process was carried out using the SOAP bioinformatic pipeline. The RAD-based SNP calling was done by SOAPsnp software (SOAPsnp, RRID:SCR_010602) [62] after each individual's paired-end RAD read was mapped onto the assembled reference genome with the alignment software SOAP2 (SOAPaligner/soap2, RRID:SCR_005503) [63]. The potential SNP markers were used for the linkage analysis if the following criteria were satisfied: for parents—sequencing depth $\geq$ 8 and $\leq$ 100, base quality $\geq$ 25, copy number $\leq$ 1.5; for progeny—sequencing depth $\geq$5, base quality $\geq$20, copy number $\leq$1.5. If the markers were showing significantly distorted segregation ($P < 0.01$), they were excluded from the map construction. Linkage analysis was performed only for markers present in at least 80% of the genomes, using JoinMap 4.0 software (JOINMAP, RRID:SCR_009248) with cross pollination (CP) population-type codes and applying the double pseudo-test cross strategy [64]. The linkage groups were formed at a logarithm of odds threshold of 6.0 and ordered using the regression mapping algorithm.

Construction of gene families

We identified gene families using TreeFam software (Tree families database, RRID:SCR_013401) [65] as follows: Blast was used to compare all the protein sequences from 13 species: *M. amblycephala*, *C. idellus, C. semilaevis*, *C. carpio*, *D. rerio*, *Callorhinchus milii*, *G. morhua, G. aculeatus, Latimeria chalumnae*, *Oncorhynchus mykiss, O. niloticus, O. latipes,* and *Fugu rubripes*, with the E-value threshold set as 1e-7. In the next step, HSP segments of each protein pair were concatenated by Solar software. H-scores were computed based on Bit-scores, and these were taken to evaluate the similarity among genes. Finally, gene families were obtained by clustering of homologous gene sequences using Hcluster_sg

(version 0.5.0). Specific genes of *M. amblycephala* were those that did not cluster with other vertebrates that were chosen for gene family construction and those that did not have homologs in the predicted gene repertoire of the compared genomes. If these genes had functional motifs, they were annotated by GO.

Phylogenetic tree reconstruction and divergence time estimation

The coding sequences of single-copy gene families conserved among *M. amblycephala*, *C. idellus*, *C. carpio*, *D. rerio*, *C. semilaevis*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *O. mykiss*, *O. niloticus*, *O. latipes*, *C. milii*, and *Fugu rubripes* (Ensembl Gene version 77) were extracted and aligned with guidance from amino-acid alignments created by the MUSCLE program (MUSCLE, RRID:SCR_011812) [66]. The individual sequence alignments were then concatenated to form 1 supermatrix. PhyML (PhyML, RRID:SCR_014629) [67, 68] was applied to construct the phylogenetic tree under an HKY85+ gamma model for nucleotide sequences. Approximate likelihood ratio test (aLRT) values were taken to assess the branch reliability in PhyML. The same set of codon sequences at position 2 was used for phylogenetic tree construction and estimation of the divergence time. The PAML mcmctree program (PAML, RRID:SCR_014932; PAML version 4.5) [69, 70] was used to determine divergence times with the approximate likelihood calculation method and the correlated molecular clock and REV substitution model.

Gene family expansion and contraction analyses

Protein sequences of *M. amblycephala* and 11 other related species (Ensembl Gene version 77) were used in BLAST searches to identify homologs. We identified gene families using CAFÉ [71], which employs a random birth and death model to study gene gains and losses in gene families across a user-specified phylogeny. The global parameter λ, which describes both the gene birth (λ) and death ($\mu = -\lambda$) rate across all branches in the tree for all gene families, was estimated using maximum likelihood. A conditional *P*-value was calculated for each gene family, and families with conditional *P*-values of less than the threshold (0.05) were considered to have a notable gain or loss. We identified branches responsible for low overall *P*-values of significant families.

Detection of positively selected genes

We calculated Ka/Ks ratios for all single copy orthologs of *M. amblycephala* and *C. semilaevis*, *D. rerio*, *G. morhua*, *O. niloticus*, and *C. carpio*. Alignment quality was essential for estimating positive selection. Thus, orthologous genes were first aligned by PRANK [72], which is considerably conservative for inferring positive selection. We used Gblocks [73] to remove ambiguously aligned blocks within PRANK alignments and employed "codeml" in the PAML package with the free-ratio model to estimate Ka, Ks, and Ka/Ks ratios on different branches. The differences in mean Ka/Ks ratios for single-copy genes between *M. amblycephala* and each of the other species were compared using paired Wilcoxon rank sum tests. Genes that showed values of Ka/Ks higher than 1 along the branch leading to *M. amblycephala* were reanalyzed using the codon-based branch site tests implemented in PAML (PAML, RRID:SCR_014932). The branch site model allowed ω to vary both among sites in the protein and across branches, and it was used to detect episodic positive selection.

Developmental process of intermuscular bone in *M. amblycephala*

To better understand the number and morphological types of IBs in adult *M. amblycephala*, specimens with a body length ranging from 15.5 to 20.5 cm were collected, and each individual was wrapped in gauze and boiled. The fish body was divided into 2 sections: anterior (snout to cloaca) and posterior (cloaca to base of caudal fin), and the length of each section was measured. The IBs were retrieved, counted, arranged in order, and photographed with a digital camera. Fertilized *M. amblycephala* eggs were brought from hatching facilities at the Freshwater Fish Genetics Breeding Center of Huazhong Agricultural University (Wuhan, Hubei, China) to our laboratory. *M. amblycephala* larvae were maintained in a re-circulating aquaculture system at 23°C ± 1°C with a 14-hour photoperiod. To explore the early development of IBs, larvae at different stages from 15 to 40 dpf were collected and fixed in 4% paraformaldehyde and transferred to 70% ethanol for storage. Specimens were stained with alizarin red for bone following the method described by Dawson [74]. The appearance of the red color was recorded as the appearance of IB because bone ossification is accompanied by the uptake of alizarin red, resulting in red staining of the mineralized bone matrix. Myosepta, either not yet ossified or poorly ossified, are not visible with alizarin red staining. For histologic analysis, specimens were paraffin-embedded and sectioned following standard protocols. Sections were stained with hematoxylin and eosin and Masson trichrome [75] and photographed using a Nikon microscope (Nikon, Tokyo, Japan) with a DP70 digital camera (Olympus, Japan). Scanning electron microscopy (SEM) and transmission electron microscopy (TEM) were also conducted to analyze the ultrastructure of IB. The specimens were fixed with 2.5% (v/v) glutaraldehyde in a solution of 0.1 M sodium cacodylate buffer (pH 7.3) for 2 hours at room temperature. The SEM and TEM samples were prepared according to a standard protocol described by Ott [76]. The samples were then visualized with a JSM-6390LV scanning electron microscope (SEM, Japan), and the stained ultrathin sections with an H-7650 transmission electron microscope (Hitachi, Japan).

RNA sequencing analysis

M. amblycephala specimens belonging to 3 different developmental stages of IBs (stage 1: whole larvae without distribution of IB; stage 2: muscle tissues with partial distribution of IBs; stage 3: muscle tissues with completed distribution of IBs that were identified under microscope and immediately frozen in liquid nitrogen). In addition, dorsal white muscle, IBs, and connective tissue surrounding the IBs from 6-month-old fish were also collected. RNA was extracted from total fish samples at different stages and from individual muscle, connective tissue, and intermuscular bone samples of *M. amblycephala* using RNAisoPlus Reagent (TaKaRa, China) according to the manufacturer's protocol. The integrity and purity of the RNA was determined by gel electrophoresis and Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA, USA) before preparing the libraries for sequencing. Paired-end RNA sequencing was performed using the Illumina HiSeq 2000 platform. Low-quality score reads were filtered, and the clean data were aligned to the reference genome using Bowtie [77]. Gene and isoform expression levels were quantified by a software package: RSEM (RNASeq by Expectation Maximization; RSEM, RRID:SCR_013027) [78]. Gene expression levels were calculated by using the reads per kilobase transcriptome per

million mapped reads (RPKM) method [79] and adjusted by a scaling normalization method [80]. We detected DEGs from 3 stages of IBs with software NOIseq and 3 different tissues with PossionDis as requested. NOIseq is based on a noisy distribution model, performed as described by Tarazona [81]. The parameters were set as: fold change $\geq$ 2.00 and probability $\geq$ 0.7. PossionDis is based on the Poisson distribution, performed as described by Audic [82]. The parameters were set as: fold change $\geq$ 2.00 and FDR $\leq$ 0.001. Annotation of DEGs was mapped to GO categories in the database, and the numbers of genes for every term were calculated to identify GO terms that were significantly enriched in the input list of DEGs. The calculated *P*-value was adjusted by the Bonferroni Correction, taking corrected *P*-value $\leq$ 0.05 as a threshold. KEGG automatic annotation was used to perform pathway enrichment analysis of DEGs.

Prediction of olfactory receptor genes

Olfactory receptor genes were identified by previously described methods, with the exception of a first-round TBLASTN (TBLASTN, RRID:SCR_011822) [83] search, in which 1417 functional olfactory receptor genes from *H. sapiens*, *D. rerio*, *L. chalumnae*, *Lepisosteus oculatus*, *L. vexillifer*, *O. niloticus*, *O. latipes*, *F. rubripes*, and *Xenopus tropicalis* were used as queries. We then predicted the structure of sequenced genes using the blast-hit sequence with the software GeneWise [52], extending in both 3′ and 5′ directions along the genome sequences. The results were further confirmed by non-redundant (NR) annotation. Then the coding sequences from the start (ATG) to stop codons were extracted, while sequences that contained interrupting stop codons or frame shifts were regarded as pseudogenes. To construct phylogenetic trees, the amino acid sequences encoded by olfactory receptor genes were first aligned using the program MUSCLE nested in MEGA 5.10 (MEGA Software, RRID:SCR_000667) [84]. We then constructed the phylogenetic tree using the neighbor-joining method with Poisson correction distances using the program MEGA 5.10. We also identified and compared the genes for 5 basic tastes (sour, sweet, bitter, umami, and salty) using a similar method as in OR gene identification.

Gut microbiota analysis

To characterize the microbial diversity of herbivorous *M. amblycephala*, 12 juvenile (LBSB), domestic adult (DBSB), and wild adult *M. amblycephala* (BSB) and wild adult *C. idellus* (GC) intestinal fecal samples were collected. Bacterial genomic DNA was extracted from the 200-mg gut content of each sample using a QIAamp DNA Stool Mini Kit (Qiagen, Valencia, USA). Quality and integrity of each DNA sample were determined by 1% agarose gel electrophoresis in Tris-acetate-EDTA buffer. DNA concentration was quantified using a NanoDrop ND-2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA). To determine the diversity and composition of the bacterial communities of each sample, 20 μg of genomic DNA were sequenced using the Illumina MiSeq sequencing platform. Polymerase chain reaction amplifications were conducted from each sample to produce the V4 hypervariable region (515F and 806 R) of the 16S rRNA gene according to the previously described method [86]. We used the UPARSE pipeline [87] to pick OTUs at an identity threshold of 97% and picked representative sequences for each OTU and used the Ribosomal Database Project (RDP) classifier to assign taxonomic data to each representative sequence.

Additional files

Additional file 1: Tables S1 to S17 and Figs S1 to S28.
Additional file 2: Data Note 1: Expanded genes in the *M. amblycephala* and *C. idellus* lineage.
Additional file 3: Data Note 2: Positively selected genes in the *M. amblycephala* and *C. idellus* genomes.

Abbreviations

BMP: bone morphogenetic protein; BUSCO: benchmarking universal single-copy orthologs; DEGs: differentially expressed genes; dpf: days post fertilization; ECM: extracellular matrix; FGF: fibroblast growth factor; IB: intermuscular bone; LG: linkage group; LTR: long terminal repeat retrotransposon; PSG: positively selected gene; OR: olfactory receptor; SEM: scanning electron microscopy; SNP: single-nucleotide polymorphism; TEM: transmission electron microscopy; TE: transposable element.

Acknowledgements

This work was supported by the Fundament Research Funds for the Central Universities (2662015PY019), the Modern Agriculture Industry Technology System Construction Projects of China, titled as “Staple Freshwater Fishes Industry Technology System” (No. CARS-46–05), Guangdong Haid Group Co., Ltd., and the International Scientific and Technology Cooperation Program of Wuhan City (2015030809020365).

Availability of data and materials

Datasets and source images supporting the results of this article are available in the GigaDB repository associated with this publication [88]. Raw whole-genome sequencing and RAD-seq data have been deposited at NCBI in the SRA under accession number SRP090157 (BioProject Number: PRJNA343584).

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

All experimental procedures involving fish were performed in accordance with the guidelines and regulations of the National Institutes of Health Guide for the Care and Use of Laboratory Animals and the Institutional Review Board on Bioethics and Biosafety of BGI (BGI-IRB).

Author contributions

W.W. initiated and conceived the project and provided scientific input. X.Q. organized financial support and designed the project. M.S. discussed the data and wrote and modified the paper. H.L. and C.C. conducted the biological experiments, analyzed the data, and wrote the paper with input from other authors. I.E. wrote and modified the manuscript and discussed the data. The RAD-seq data analyses and the genetic map construction were performed by Z.G., Y.G., J.J., and X.J. Genome assembly and annotation were performed by J.M., H.C., M.X., and J.C. X.Z., W.L., R.L., B.C., J.W., H.L., S.Y., H.W., X.C., X.Z., Y.Z., K.W., R.Y., and B.L. carried out the sample preparation and data collection. J.L. and J.C. identified the gene families and analyzed the RNA-seq data. M.B. coordinated the project. S.Z. and X.F. modified the manuscript

and discussed the data. All authors read the manuscript and provided comments and suggestions for improvements. The authors declare no competing financial interests.

References

1. FAO Fisheries and Aquaculture Department. FAO Yearbook Fishery and Aquaculture Statistics 2014. Rome: Food and Agriculture Organization of the United Nations, 2016.
2. Star B, Nederbragt AJ, Jentoft S et al. The genome sequence of Atlantic cod reveals a unique immune system. Nature 2011;**477**:207–10.
3. Berthelot C, Brunet F, Chalopin D et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 2014;**5**:3657.
4. Tine M, Kuhl H, Gagnaire PA et al. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun 2014;**5**:5770.
5. Wu C, Zhang D, Kan M et al. The draft genome of the large yellow croaker reveals well-developed innate immunity. Nat Commun 2014;**5**:5227.
6. Chen S, Zhang G, Shao C et al. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat Genet 2014;**46**:253–60.
7. Brawand D, Wagner CE, Li YI et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature 2014;**513**:375–82.
8. Chen X, Zhong L, Bian C et al. High-quality genome assembly of channel catfish, *Ictalurus punctatus*. Gigascience 2016;**5**:39.
9. Gemballa S, Britz R. Homology of intermuscular bones in acanthomorph fishes. Am Mus Novit 1998;**3241**:1–25.
10. Danos N, Ward AB. The homology and origins of intermuscular bones in fishes: phylogenetic or biomechanical determinants? Biol J Linn Soc 2012;**106**:607–22.
11. Wan SM, Yi SK, Zhong J et al. Dynamic mRNA and miRNA expression analysis in response to intermuscular bone development of blunt snout bream (*Megalobrama amblycephala*). Sci Rep 2016;**6**:31050.
12. Xu P, Zhang X, Wang X et al. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. Nat Genet 2014;**46**:1212–9.
13. Wang Y, Lu Y, Zhang Y et al. The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. Nat Genet 2015;**47**:625–31.
14. Gao Z, Luo W, Liu H et al. Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). PLoS One 2012;**7**:e42637.
15. Yi S, Gao ZX, Zhao H et al. Identification and characterization of microRNAs involved in growth of blunt snout bream (*Megalobrama amblycephala*) by Solexa sequencing. BMC Genomics 2013;**14**:754.
16. Luo R, Liu B, Xie Y et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 2012;**1**:18.
17. Simao FA, Waterhouse RM, Ioannidis P et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**:3210–2.
18. Jones FC, Grabherr MG, Chan YF et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 2012;**484**:55–61.
19. Niimura Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol Evol 2009;**1**:34–44.
20. Lindemann B. Receptors and transduction in taste. Nature 2001;**413**:219–25.
21. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet 2008;**9**:951–63.
22. Chandrashekar J, Mueller KL, Hoon MA et al. T2Rs function as bitter taste receptors. Cell 2000;**100**:703–11.
23. Niimura Y, Nei M. Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. J Hum Genet 2006;**51**:505–17.
24. Nelson G, Chandrashekar J, Hoon MA et al. An amino-acid taste receptor. Nature 2002;**416**:199–202.
25. Li R, Fan W, Tian G et al. The sequence and de novo assembly of the giant panda genome. Nature 2010;**463**:311–7.
26. Ferreira AHP, Marana SR, Terra WR et al. Purification, molecular cloning, and properties of a β-glycosidase isolated from midgut lumen of *Tenebrio molitor* (Coleoptera) larvae. Insect Biochem Mol Biol 2001;**31**:1065–76.
27. Tokuda G, Saito H, Watanabe H. A digestive β-glucosidase from the salivary glands of the termite, *Neotermes koshunensis* (Shiraki): Distribution, characterization and isolation of its precursor cDNA by 5′- and 3′-RACE amplifications with degenerate primers. Insect Biochem Mol Biol 2002;**32**:1681–9.
28. Sakamoto K, Uji S, Kurokawa T et al. Molecular cloning of endogenous β-glucosidase from common Japanese brackish water clam *Corbicula japonica*. Gene 2009;**435**:72–9.
29. Zhu L, Wu Q, Dai J et al. Evidence of cellulose metabolism by the giant panda gut microbiome. Proc Natl Acad Sci USA 2011;**108**:17714–9.
30. Bird NC, Mabee PM. Developmental morphology of the axial skeleton of the zebrafish, *Danio rerio* (Ostariophysi: Cyprinidae). Dev Dyn 2003;**228**:337–57.
31. Ornitz D, Marie P. FGF signaling pathways in endochondral and intramembranous bone development and human genetic disease. Genes Dev 2002;**16**:1446–65.
32. Ortega N, Behonick DJ, Werb Z. Matrix remodeling during endochondral ossification. Trends Cell Biol 2004;**14**:86–93.
33. Even-Ram S, Doyle AD, Conti MA et al. Myosin IIA regulates cell motility and actomyosin-microtubule crosstalk. Nat Cell Biol 2007;**9**:299–309.
34. Sheetz MP, Felsenfeld DP, Galbraith CG. Cell migration: regulation of force on extracellular-complexes. Trends Cell Biol 1998;**8**:51–4.
35. Gunst SJ, Zhang W. Actin cytoskeletal dynamics in smooth muscle: a new paradigm for the regulation of smooth muscle contraction. Am J Physiol Cell Physiol 2008;**295**:C576–87.
36. Webb RC. Smooth muscle contraction and relaxation. Adv Physiol Educ 2003;**27**:201–6.
37. Ulrich TA, De Juan Pardo EM, Kumar S. The mechanical rigidity of the extracellular matrix regulates the structure, motility, and proliferation of glioma cells. Cancer Res 2009;**69**:4167–74.
38. Ridley AJ. Rho GTPases and cell migration. J Cell Sci 2001;**114**:2713–22.
39. Etienne-Manneville S, Hall A. Rho GTPases in cell biology. Nature 2002;**420**:629–35.
40. Ridley AJ. Cell migration: integrating signals from front to back. Science 2003;**302**:1704–9.

41. Chen G, Deng C, Li YP. TGF-β and BMP signaling in osteoblast differentiation and bone formation. Int J Biol Sci 2012;**8**:272–88.
42. Harada S, Rodan GA. Control of osteoblast function and regulation of bone mass. Nature 2003;**423**:349–55.
43. Long F. Building strong bones: molecular regulation of the osteoblast lineage. Nat Rev Mol Cell Biol 2011;**13**:27–38.
44. Boyle WJ, Simonet WS, Lacey DL. Osteoclast differentiation and activation. Nature 2003;**423**:337–42.
45. Fakhry A, Ratisoontorn C, Vedhachalam C et al. Effects of FGF-2/-9 in calvarial bone cell cultures: differentiation stage-dependent mitogenic effect, inverse regulation of BMP-2 and noggin, and enhancement of osteogenic potential. Bone 2005;**36**:254–66.
46. Sato K, Suematsu A, Nakashima T et al. Regulation of osteoclast differentiation and function by the CaMK-CREB pathway. Nat Med 2006;**12**:1410–6.
47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;**25**:1754–60.
48. Grabherr MG, Haas BJ, Yassour M et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. Nat Biotechnol 2011;**29**:644–52.
49. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;**27**:573–80.
50. Jurka J, Kapitonov VV, Pavlicek A et al. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;**110**:462–7.
51. Griffiths-Jones S, Moxon S, Marshall M et al. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 2005;**33**:121–4.
52. Birney E, Clamp M, Durbin R. Gene wise and genomewise. Genome Res 2004;**14**:988–95.
53. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-seq. Bioinformatics 2009;**25**:1105–11.
54. Trapnell C, Williams BA, Pertea G et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;**28**:511–5.
55. Elsik CG, Mackey AJ, Reese JT et al. Creating a honey bee consensus gene set. Genome Biol 2007;**8**:R13.
56. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;**28**:45–8.
57. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;**25**:25–9.
58. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;**28**:27–30.
59. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997;**25**:955–64.
60. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics 2009;**25**:1335–7.
61. Baird NA, Etter PD, Atwood TS et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One 2008;**3**:e3376.
62. Li R, Li Y, Fang X et al. SNP detection for massively parallel whole-genome resequencing. Genome Res 2009;**19**:1124–32.
63. Li R, Yu C, Li Y et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009;**25**:1966–7.
64. Grattapaglia D, Sederoff R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 1994;**137**:1121–37.
65. Li H, Coghlan A, Ruan J et al. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 2006;**34**:D572–80.
66. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;**32**:1792–7.
67. Guindon S, Dufayard JF, Lefort V et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;**59**:307–21.
68. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;**52**:696–704.
69. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;**24**:1586–91.
70. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 2006;**23**:212–26.
71. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. Genetics 2007;**177**:1941–9.
72. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci USA 2005;**102**:10557–62.
73. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;**56**:564–77.
74. Dawson AB. A note on the staining of the skeleton of cleared specimens with alizarin red S. Biotech Histochem 1926;**1**:123–4.
75. Gruber HE. Adaptations of Goldner's Masson trichrome stain for the study of undecalcified plastic embedded bone. Biotech Histochem 1992;**67**:30–4.
76. Ott HC, Matthiesen TS, Goh SK et al. Perfusion-decellularized matrix: using nature's platform to engineer a bioartificial heart. Nat Med 2008;**14**:213–21.
77. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**:357–9.
78. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 2011;**12**:323.
79. Mortazavi A, Williams BA, McCue K et al. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 2008;**5**:621–8.
80. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;**11**:R25.
81. Tarazona S, García-Alcalde F, Dopazo J et al. Differential expression in RNA-seq: a matter of depth. Genome Res 2011;**21**:2213–23.
82. Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 1997;**7**:986–95.
83. Niimura Y. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. Hum Genomics 2009;**4**:107–18.
84. Altschul S, Madden T, Schaffer A et al.t Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;**25**:3389–402.
85. Kumar S, Nei M, Dudley J et al. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 2008;**9**:299–306.

86. Caporaso JG, Lauber CL, Walters WA et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 2011;**108**:4516–22.
87. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 2013;**10**:996–8.
88. Liu H, Chen C, Gao Z et al. Supporting data for "The draft genome of blunt snout bream (Megalobrama amblycephala) reveals the development of intermuscular bone and adaptation to herbivorous diet" GigaScience Database 2017. http://dx.doi.org/10.5524/100305.