

## The draft genome of *Megalobrama amblycephala* reveals the development of intermuscular bone and adaptation to herbivorous diet --Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-16-00088
<b>Full Title:</b>	The draft genome of <i>Megalobrama amblycephala</i> reveals the development of intermuscular bone and adaptation to herbivorous diet
<b>Article Type:</b>	Research
<b>Funding Information:</b>	
<b>Abstract:</b>	<p><b>Background:</b> The blunt snout bream, <i>Megalobrama amblycephala</i>, 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.</p> <p><b>Results:</b> As a first towards optimizing this aquatic livestock, we present a 1.116-Gb draft genome of <i>M. amblycephala</i>, 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 play a role 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 <i>M. amblycephala</i>. The composition of gut microflora, which contributed to the herbivorous adaptation of <i>M. amblycephala</i>, was found to be similar to that of other herbivores.</p> <p><b>Conclusions:</b> As a valuable resource for improvement of <i>M. amblycephala</i> livestock, the draft genome sequence offers new insights into the development of intermuscular bone and herbivorous adaptation.</p>
<b>Corresponding Author:</b>	Weimin Wang  CHINA
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Han Liu
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	Han Liu
	Chunhai Chen
	Zexia Gao
	Jiumeng Min
	Yongming Gu
	Jianbo Jian
	Xiewu Jiang
	Huimin Cai
	Ingo Ebersberger
	Meng Xu
	Xinhui Zhang
	Jianwei Chen

	Wei Luo
	Boxiang Chen
	Junhui Chen
	Hong Liu
	Jiang Li
	Ruifang Lai
	Mingzhou Bai
	Jin Wei
	Shaokui Yi
	Huanling Wang
	Xiaojuan Cao
	Xiaoyun Zhou
	Yuhua Zhao
	Kaijian Wei
	Ruibin Yang
	Bingnan Liu
	Shancen Zhao
	Xiaodong Fang
	Manfred Scharf
	Xueqiao Qian
	Weimin Wang
<b>Order of Authors Secondary Information:</b>	
<b>Opposed Reviewers:</b>	Byrappa Venkatesh, Dr. Prof. mcbbv@imcb.a-star.edu.sg
	Yaping Wang, Dr. Prof. wangyp@ihb.ac.cn
	Shaojun Liu, Dr. Prof. lsj@hunnu.edu.cn
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<b>Experimental design and statistics</b>	Yes
Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a> . Information essential to interpreting the data presented should be made available in the figure legends.	

<p>Have you included all the information requested in your manuscript?</p>	
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>

[Click here to view linked References](#)

1           1           **The draft genome of *Megalobrama amblycephala* reveals the development of**  
2  
3           2           **intermuscular bone and adaptation to herbivorous diet**

4  
5  
6           3           Han Liu<sup>1†</sup>, Chunhai Chen<sup>2†</sup>, Zexia Gao<sup>1†</sup>, Jiumeng Min<sup>2†</sup>, Yongming Gu<sup>3†</sup>, Jianbo Jian<sup>2†</sup>, Xiewu  
7  
8           4           Jiang<sup>3</sup>, Huimin Cai<sup>2</sup>, Ingo Ebersberger<sup>4</sup>, Meng Xu<sup>2</sup>, Xinhui Zhang<sup>1</sup>, Jianwei Chen<sup>2</sup>, Wei Luo<sup>1</sup>,  
9  
10          5           Boxiang Chen<sup>1,3</sup>, Junhui Chen<sup>2</sup>, Hong Liu<sup>1</sup>, Jiang Li<sup>2</sup>, Ruifang Lai<sup>1</sup>, Mingzhou Bai<sup>2</sup>, Jin Wei<sup>1</sup>,  
11  
12          6           Shaokui Yi<sup>1</sup>, Huanling Wang<sup>1</sup>, Xiaojuan Cao<sup>1</sup>, Xiaoyun Zhou<sup>1</sup>, Yuhua Zhao<sup>1</sup>, Kaijian Wei<sup>1</sup>,  
13  
14          7           Ruibin Yang<sup>1</sup>, Bingnan Liu<sup>3</sup>, Shancen Zhao<sup>2</sup>, Xiaodong Fang<sup>2</sup>, Manfred Schartl<sup>5,\*</sup>, Xueqiao  
15  
16          8           Qian<sup>3,\*</sup>, Weimin Wang<sup>1,\*</sup>

17  
18  
19          9  
20  
21          10          \*Correspondence: wangwm@mail.hzau.edu.cn; xueqiaoqian@263.net;  
22  
23          11          phch1@biozentrum.uni-wuerzburg.de

24  
25          12          †Equal contributors

26  
27          13          <sup>1</sup>College of Fisheries, Key Lab of Freshwater Animal Breeding, Ministry of Agriculture, Key Lab  
28  
29          14          of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong  
30  
31          15          Agricultural University, Wuhan 430070, China

32  
33          16          <sup>2</sup>Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen 518083, China

34  
35          17          <sup>3</sup>Guangdong Haid Group Co., Ltd., Guangzhou 511400, China

36  
37          18          <sup>4</sup>Dept. for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University,  
38  
39          19          Frankfurt D-60438, Germany

40  
41          20          <sup>5</sup>Physiological Chemistry, University of Würzburg, Biozentrum, Am Hubland, and Comprehensive  
42  
43          21          Cancer Center Mainfranken, University Clinic Würzburg, Würzburg 97070, Germany

29 **Abstract**

30 **Background:** The blunt snout bream, *Megalobrama amblycephala*, is the economically most  
31 important cyprinid fish species. As an herbivore, it can be grown by eco-friendly and  
32 resource-conserving aquaculture. However, the large number of intermuscular bones in the trunk  
33 musculature is adverse to fish meat processing and consumption.

34 **Results:** As a first towards optimizing this aquatic livestock, we present a 1.116-Gb draft genome  
35 of *M. amblycephala*, with 779.54 Mb anchored on 24 linkage groups. Integrating spatiotemporal  
36 transcriptome analyses we show that intermuscular bone is formed in the more basal teleosts by  
37 intramembranous ossification, and may play a role in muscle contractibility and coordinating  
38 cellular events. Comparative analysis revealed that olfactory receptor genes, especially of the beta  
39 type, underwent an extensive expansion in herbivorous cyprinids, whereas the gene for the umami  
40 receptor *TIR1* was specifically lost in *M. amblycephala*. The composition of gut microflora, which  
41 contributed to the herbivorous adaptation of *M. amblycephala*, was found to be similar to that of  
42 other herbivores.

43 **Conclusions:** As a valuable resource for improvement of *M. amblycephala* livestock, the draft  
44 genome sequence offers new insights into the development of intermuscular bone and herbivorous  
45 adaptation.

46  
47 **Keywords:** *Megalobrama amblycephala*, whole genome, herbivorous diet, intermuscular bone,  
48 transcriptome, gut microflora

## 58 **Background**

59 Fishery and aquaculture play an important role in global alimentation. Over the past decades food  
60 fish supply has been increasing with an annual rate of 3.6 percent, about 2 times faster than the  
61 human population [1]. This growth of fish production is meanwhile solely accomplished by an  
62 extension of aquaculture, as over the past thirty years the total mass of captured fish has remained  
63 almost constant [1]. As a consequence of this emphasis on fish breeding, the genomes of various  
64 economically important fish species, e.g. cod (*Gadus morhua*) [2], rainbow trout (*Oncorhynchus*  
65 *mykiss*) [3], yellow croaker (*Larimichthys crocea*) [4], tilapia (*Oreochromis niloticus*) [5] and  
66 half-smooth tongue sole (*Cynoglossus semilaevis*) [6], have been sequenced. Yet, the majority of  
67 these species are carnivorous requiring large inputs of protein from wild caught fish or other  
68 precious feed. The focus of aquacultures is, however, gradually shifting towards more resource  
69 friendly herbivorous and omnivorous species, and in particular cyprinid fish. Consequently,  
70 cyprinids are currently the economically most important group of teleosts for sustainable  
71 aquaculture. They grow to large population sizes in the wild and already now account for the  
72 majority of freshwater aquaculture production worldwide [1]. Among these, the herbivorous  
73 *Megalobrama amblycephala*, a particularly eco-friendly and resource-conserving species, is  
74 predominant in aquaculture and has been greatly developed in China. However, most cyprinids,  
75 including *M. amblycephala*, have a large number of intermuscular bones (IBs) in the trunk  
76 musculature, which have an adverse effect on fish meat processing and consumption. IBs—a  
77 unique form of bone occurring only in the more basal teleosts—are completely embedded within  
78 the myosepta and are not connected to the vertebral column or any other bones [7, 8]. To date little  
79 is known about the molecular genetic basis of IB and its formation during development. Similarly  
80 the evolution of this unique structure remains obscure. Unfortunately, the recent sequencing of  
81 two cyprinid genomes common carp (*Cyprinus carpio*) [9] and grass carp (*Ctenopharyngodon*  
82 *idellus*) [10], which provided valuable information for their genetic breeding, contributed little to  
83 the understanding of IB formation.

84 In an initial genome survey of *M. amblycephala*, we identified 25,697 SNPs [11], 347  
85 conserved miRNAs and 22 novel miRNAs [12]. However, lack of a whole genome sequence  
86 resource limited a thorough investigation of *M. amblycephala*. Here we report the first

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

87 high-quality draft genome sequence of *M. amblycephala*. Integrating this novel genome resource  
88 with tissue- and developmental stage-specific gene expression information, as well as with  
89 meta-genome data to investigate the composition of the gut microbiome provides relevant insights  
90 into the function and evolution of two key features characterizing this species: The formation of  
91 IB and the adaptation to herbivory. By that our study lays the foundation for genetically  
92 optimizing *M. amblycephala* to further increase its relevance for securing human food supply.

## 93 **Data description**

### 94 **Genome Assembly and Annotation**

95 The *M. amblycephala* genome was sequenced and assembled by a whole-genome shotgun strategy  
96 using genomic DNA from a double-haploid line (Additional file 1: Table S1). We assembled a  
97 1.116 Gb reference genome sequence from 142.55 Gb (approximately 130-fold coverage) of clean  
98 data [13] (Additional file 1: Tables S1 and S2, Figure S1). The contig and scaffold N50 lengths  
99 reached 49 Kb and 839 Kb, respectively (Table 1). The largest scaffold spans 8,951 Kb and the  
100 4,034 largest scaffolds cover 90% of the assembly (Additional file 1: Tables S3 and S4). The  
101 mapping of paired end sequence data from the short-insert size WGS libraries (Additional file 1:  
102 Table S5), as well as of published ESTs [11] (Additional file 1: Tables S6 and S7) against the  
103 genome assembly indicated that number and extent of misassemblies is low and comparable to  
104 those of other sequenced fish species (Additional file 1: Table S8).

105 Using a comprehensive annotation strategy combining RNA-seq derived transcript evidence,  
106 *de-novo* gene prediction and sequence similarity to proteins from five further fish species, we  
107 annotated a total of 23,696 protein-coding genes (Additional file 1: Table S9). Of the predicted  
108 genes, 99.94% (23,681 genes) are supported by transcript data and/or by the existence of  
109 homologs in other species (Additional file 1: Table S10). In addition, we identified 1,796  
110 non-coding RNAs including 474 miRNAs, 220 rRNA, 530 tRNAs, and 572 snRNAs (Additional  
111 file 1: Table S11). Transposable elements (TEs) comprise approximately 34% (381.3 Mb) of the *M.*  
112 *amblycephala* genome (Additional file 1: Tables S12 and S13). DNA transposons (23.80%) and  
113 long terminal repeat retrotransposons (LTRs) (9.89%) are the most abundant TEs in *M.*  
114 *amblycephala*. The proportion of LTRs in *M. amblycephala* is highest in comparison with other  
115 teleosts: *G. morhua* (4.88%) [2], *C. semilaevis* (0.08%) [6] and *L. crocea* (2.2%) [4], *C. carpio*

116 (2.28%) [9], *C. idellus* (2.58%) [10], stickleback (*Gasterosteus aculeatus*) (1.9%) [14] (Additional  
117 file 1: Tables S13 and S14, Figures S4 and S5). Notably, the distribution of divergence between  
118 the TEs in *M. amblycephala* peaks at only 7% (Additional file 1: Figure S6), indicating a more  
119 recent activity of these TEs when compared with *O. mykiss* (13%) [3] and *C. semilaevis* (9%) [6].

## 120 **Anchoring Scaffolds and Shared Synteny Analysis**

121 We constructed a high-resolution genetic map based on 5,317 single-nucleotide polymorphism  
122 (SNP) markers extracted from 198 individuals. The map spans 1,701 cM with a mean marker  
123 distance of 0.33 cM and facilitated an anchoring of 1,434 scaffolds comprising 70% (779.54 Mb)  
124 of the *M. amblycephala* genome assembly to form 24 linkage groups (Additional file 1: Table  
125 S15). Of the anchored scaffolds, 598 could additionally be oriented (678.27 Mb, 87.01% of the  
126 total anchored sequences) (Figure 1A, Additional file 1: Table S15). A subsequent comparison of  
127 the gene order between *M. amblycephala* and its close relative *C. idellus* revealed 607 large shared  
128 syntenic blocks encompassing 11,259 genes, and 190 chromosomal rearrangements. The values  
129 change to 1,062 regions, 13,152 genes and 279 rearrangements when considering zebrafish (*Danio*  
130 *rerio*) (Additional file 1: Table S16). The unexpected higher number of genes in syntenic regions  
131 shared with the more distantly related *D. rerio* is most likely an effect of the more complete  
132 genome assembly of this species compared to *C. idellus*. The rearrangement events are distributed  
133 across all *M. amblycephala* linkage groups without evidence for a local clustering (Figure 1B).  
134 The most prominent event is a chromosomal fusion in *M. amblycephala* that joined two ancestral  
135 chromosomes represented by *D. rerio* chromosomes Dre10 and Dre22. The same fusion is  
136 observed in *C. idellus* but not in *C. carpio* suggesting that it probably occurred in a last common  
137 ancestor of *M. amblycephala* and *C. idellus*, approximately 13.1 million years ago (Additional file  
138 1: Figures S7 and S8).

## 139 **Results**

### 140 **Evolutionary Analysis**

141 A phylogenetic analysis of 316 single-copy genes with one to one orthologs in the genomes of 10  
142 other fish species, and coelacanth (*Latimeria chalumnae*) and elephant shark (*Callorhynchus milii*),  
143 as out group served as a basis for investigating the evolutionary trajectory of *M. amblycephala*



144 (Figure 2A, Additional file 1: Figures S9 and S10). To illuminate the evolutionary process  
145 resulting in the adaptation to a grass diet, we analyzed the functional properties of expanded gene  
146 families in the *M. amblycephala* and *C. idellus* lineages (Additional file 1: Figure S11), two  
147 typical herbivores mainly feeding on aquatic and terrestrial grasses. Among the significantly  
148 over-represented KEGG pathways (Fisher's exact test,  $P < 0.01$ ), we find olfactory transduction  
149 (ko04740), immune-related pathways (ko04090, ko04672, ko04612 and ko04621), lipid metabolic  
150 related process (ko00590, ko03320, ko00591, ko00565, ko00592 and ko04975), as well as  
151 xenobiotics biodegradation and metabolism (ko00625 and ko00363) (Additional file 1: Tables S17  
152 and S18, Figure S12). This indicates that an adaptation to herbivory goes hand in hand with  
153 coping with plant secondary metabolites that are adverse or even toxic to the organism.  
154 Furthermore, the high-fiber but low-energy grass diet requires a highly effective intermediate  
155 metabolism that accelerates carbohydrate and lipid catabolism and conversion into energy to  
156 maintain physiological functions. Indeed, when tracing positively selected genes in *M.*  
157 *amblycephala* and *C. idellus* (Additional file 1: Table S19), we identified many candidates  
158 involved in starch and sucrose metabolism (ko00500), citrate cycle (ko00020) and other types of  
159 O-glycan biosynthesis (ko00514). Moreover, 20 genes encoding enzymes involved in lipid and  
160 carbohydrate metabolism appear positively selected in both fish species (Additional file 1: Table  
161 S20).

## 162 **Development of Intermuscular Bones**

163 To explain the genetic basis of intermuscular bones (IBs), their formation and their function in  
164 cyprinids, we first analyzed the functional annotation of gene families that expanded in this  
165 lineage (Figure 2B). Interestingly, many of these gene families are involved in cell adhesion (GO:  
166 0007155,  $P = 5.26E-32$ , 357 genes), myosin complex (GO:0016459,  $P = 2.74E-08$ , 100 genes) and  
167 cell-matrix adhesion (GO:0007160,  $P = 1.59E-21$ , 69 genes), which interact dynamically to  
168 mediate efficient cell motility, migration and muscle construction [15-17] (Figure 2C, Additional  
169 file 1: Tables S21 and S22).

170 As a second line of evidence, we performed comparative transcriptome analyses of early  
171 developmental stages (stage1: whole larvae without IBs) and juvenile *M. amblycephala* (stage2:

172 trunk muscle with partial IBs; stage3: trunk muscle with completed IBs) (Figure 3A). We found  
173 249 genes significantly up-regulated in stages 2 and 3 (with IB) compared to stage 1 (no IB).  
174 Notably, many of these genes belong to KEGG pathways involved in tight junction (ko04530),  
175 regulation of actin cytoskeleton (ko04810), cardiac muscle contraction (ko04260) and vascular  
176 smooth muscle contraction (ko04270) (Additional file 1: Figure S14). These genes play important  
177 roles in cell motility and muscle contraction [16, 18, 19], which resembles the findings from the  
178 gene family expansion analysis. Specifically, some of these genes encoding proteins related to  
179 muscle contraction, including titin, troponin, myosin, actinin, calmodulin and other Ca<sup>2+</sup>  
180 transporting ATPases (Figure 3A) point to a strong remodeling of the musculature compartment.

181 To confirm that the observed differences in gene expression are indeed linked to IB formation  
182 and function and are not simply due to the fact that different developmental stages were compared,  
183 we eventually extended the comparative transcriptome analysis to muscle tissues, IB, and  
184 connective tissues from the same six months old individual of *M. amblycephala* (Figure 3B,  
185 Additional file 1: Figures S15-17). Among the genes that are significantly up-regulated in the IB  
186 samples many encode extracellular matrix (ECM) proteins (collagens and intergrin-binding  
187 protein), Rho GTPase family (*RhoA*, *Rho GAP*, *Rac*, *Ras*), motor proteins (myosin, dynein, actin),  
188 and calcium channel regulation protein (Additional file 1: Figure S18 and Table S23). Interestingly,  
189 it has been demonstrated that ECM proteins bound to integrins influence cell migration by  
190 actomyosin-generated contractile forces [17, 20]. Rho GTPases, acting as molecular switches, also  
191 play a pivotal role in regulating the actin cytoskeleton and cell migration, which in turn initiates  
192 intracellular signaling and contributes to tissue repair and regeneration [21-23].

193 During development of *M. amblycephala*, the first IB appears in muscles of caudal vertebrae  
194 as early as 28 days post fertilization (dpf) when body length is 12.95 mm (Additional file 1: Figure  
195 S19). The system then develops and ossifies predominantly from posterior to anterior (Additional  
196 file 1: Figure S20). IBs are present throughout the body within two months (Additional file 1:  
197 Figure S21) and develop into multiple morphological types in adults (Additional file 1: Figure  
198 S22). The bone is formed directly without an intermediate cartilaginous stage (Additional file 1:  
199 Figures S23 and S24). We also found a large number of mature osteoblasts distributed at the edge  
200 of the bone matrix while some osteocytes were apparent in the center of the mineralized bone

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

201 matrix (Additional file 1: Figures S25 and S26). These primary bone-forming cells predominantly  
202 regulate bone formation and function throughout life. Notably, among the genes up-regulated in  
203 IB, 35 bone formation regulatory genes were identified (Figure 3D). In particular, genes involved  
204 in Bmp signaling including *Bmp3*, *Smad8*, *Smad9*, and *Id2*, in FGF signaling including *Fgf2*,  
205 *Fgfr1a*, *Fgfbp2*, *Col6a3*, and *Col4a5*, and in Ca<sup>2+</sup> channels including *Cacna1c*, *CaM*, *Creb5*, and  
206 *Nfatc* were highly expressed (>2-fold change) in IB (Additional file 1: Figure S27). It has been  
207 demonstrated that *Bmp*, *Fgf2*, and *Fgfr1* play significant roles in intramembranous bone  
208 development and affect the expression and activity of other osteogenesis related transcription  
209 factors [24, 25]. The calcium-sensitive transcription factor *NFATc1* together with *CREB* induces  
210 the expression of osteoclast-specific genes [26].

### 211 **Adaptation to Herbivorous Diet**

212 Next to the presence of IB, the strict herbivory of *M. amblycephala* is the second key feature in  
213 connection to the use of this species as aquatic livestock. Olfaction, the sense of smell, is crucial  
214 for animals to find food. The perception of smell is mediated by a large gene family of olfactory  
215 receptor (OR) genes. The ORs of teleosts are predominantly expressed in the main olfactory  
216 epithelium of the nasal cavity [27, 28] and can discriminate, like those of other vertebrates,  
217 different kinds of odor molecules. However, compared to mammals, e.g. humans having around  
218 400 ORs [29] the OR repertoires in teleosts are considerably small. They range from only about  
219 48 in *Fugu rubripes* up to 161 in *D. rerio* (Figure 4A). In the *M. amblycephala* genome, we  
220 identified 179 functional olfactory receptor (OR) genes (Figure 4A), and based on the  
221 classification of Niimura [30], 158, 117 and 153 receptors for water-borne odorants were  
222 identified in *M. amblycephala*, *C. idellus* and *D. rerio*, respectively (Additional file 1: Table S24).  
223 Overall, these receptor repertoires are substantially larger than those of other and carnivorous  
224 teleosts (*G. morhua*, *C. semilaevis*, *O. latipes*, *X. maculatus*) (Additional file 1: Figures S28 and  
225 S29, Table S24). This suggests that olfaction—probably for food choice—plays a particularly  
226 important role in the cyprinid species. Previous studies have demonstrated that the beta type OR  
227 genes are present in both aquatic and terrestrial vertebrates, indicating that the corresponding  
228 receptors detect both water-soluble and airborne odorants [28, 30]. Intriguingly, we found a

229 massive expansion of beta-type OR genes in the genomes of the herbivorous *M. amblycephala* and  
230 *C. idellus*, while very few exist in other teleosts (Figure 4B, Additional file 1: Tables S24 and  
231 S25).

232 Taste is also an important factor in the development of dietary habits. Most animals can  
233 perceive five basic tastes, namely sourness, sweetness, bitterness, saltiness and umami [31].  
234 Interestingly, *TIR1*, the receptor gene necessary for sensing umami, has been lost in herbivorous  
235 *M. amblycephala* but is duplicated in carnivorous *G. morhua*, *C. semilaevis* and omnivorous *O.*  
236 *latipes* and *X. maculatus* (Figures 4C and 4D, Additional file 1: Figures S30-32 and Table S26). In  
237 contrast, *TIR2*, the receptor gene for sensing sweet, has been duplicated in herbivorous *M.*  
238 *amblycephala* and *C. idellus*, omnivorous *C. carpio* and *D. rerio*, while it has been lost in  
239 carnivorous *G. morhua* and *C. semilaevis* (Additional file 1: Figure S33 and Table S26). Bitterness  
240 sensed by the *T2R* is particularly crucial for animals to protect them from poisonous compounds  
241 [32]. Probably in the course switching to a diet that contains a larger fraction of bitter containing  
242 food, also the *T2R* gene family in *M. amblycephala*, *C. idellus*, *C. carpio* and *D. rerio* has been  
243 expanded (Additional file 1: Figure S34).

244 To obtain further insights into the genetic adaptation to herbivorous diet, we focused on  
245 further genes that might be associated with digestion. Genes that encode proteases (including  
246 pepsin, trypsin, cathepsin and chymotrypsin) and amylases (including alpha-amylase and  
247 glucoamylase) were identified in the genomes of *M. amblycephala*, carnivorous *C. semilaevis*, *G.*  
248 *morhua* and omnivorous *D. rerio*, *O. latipes* and *X. maculatus*, indicating that herbivorous *M.*  
249 *amblycephala* has a protease repertoire that is not substantially different from those of carnivorous  
250 and omnivorous fishes (Additional file 1: Table S27). We did not identify any genes encoding  
251 potentially cellulose-degrading enzymes including endoglucanase, exoglucanase and  
252 beta-glucosidase in the genome of *M. amblycephala*, suggesting that utilization of the herbivorous  
253 diet may largely depend on the gut microbiome. To elucidate this further, we determined the  
254 composition of the gut microbial communities of juvenile, domestic, wild adult *M. amblycephala*  
255 and wild adult *C. idellus* using bacterial 16S rRNA sequencing (Additional file 1: Figure S35). A  
256 total of 549,020 filtered high quality sequence reads from 12 samples were clustered at a similarity  
257 level of 97%. The resulting 8,558 operational taxonomic units (OTUs) are dominated at phylum

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

258 level by Proteobacteria, Fusobacteria, Bacteroidetes, Firmicutes and Actinobacteria (Figure 4E,  
259 Additional file 1: Table S28). Increasing the resolution to the genus level, the composition and  
260 relative abundance of the gut microbiota of wild adult *M. amblycephala* and *C. idellus* are still  
261 very similar (Additional file 1: Table S29) and we could identify more than 7%  
262 cellulose-degrading bacteria (Additional file 1: Table S30). This indicates that indeed the gut  
263 microbiome plays a prominent role in the digestion of plant material, and thus in the adaptation to  
264 herbivory.

## 265 **Discussion**

266 *M. amblycephala* is the economically most important species for freshwater aquaculture. In  
267 addition to its various superior properties, especially its herbivorous diet, it is also an excellent  
268 model to study intermuscular bones (IB) formation. Here we make available draft genome of *M.*  
269 *amblycephala* with more than 70% of genome data anchored on 24 linkage groups. Comparative  
270 analyses of genome structure revealed high synteny with three other cyprinid fish and uncovered a  
271 chromosomal fusion event in *M. amblycephala* that joined two ancestral chromosomes (Figure  
272 1B), which supports the previous results in *C. idellus* [10] and also provides novel scientific  
273 insights into the evolution of chromosome fusion events in cyprinids.

274 The evolutionary trajectory analysis of *M. amblycephala* and other teleosts revealed that *M.*  
275 *amblycephala* has the closest relationship to *C. idellus* (Figure 2A). Both the species are  
276 herbivorous fish but which endogenous and exogenous factors affected their feeding habits and  
277 how they adapted to their herbivorous diet is not known. Olfaction and taste are crucial for  
278 animals to find food and to distinguish whether potential food is edible or harmful [28, 32]. The  
279 search for genes encoding OR showed that herbivorous *M. amblycephala* and *C. idellus* have a  
280 large number of beta-type OR, while other omnivorous and carnivorous fish only have one or two.  
281 This might be attributed to their particular herbivorous diet consisting not only of aquatic grasses  
282 but also the duckweed and terrestrial grasses, which they ingest from the water surface. Previous  
283 studies have demonstrated that the receptor for umami is formed by the T1R1/T1R3 heterodimer,  
284 while T1R2/T1R3 senses sweet taste [33]. We found that the umami gene *TIR1* was lost in  
285 herbivorous *M. amblycephala* but duplicated in the carnivorous *G. morhua* and *C. semilaevis*

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
286 (Figure 4C). The loss of the *TIR1* gene in *M. amblycephala* excludes the expression of a  
287 functional umami taste receptor. Such situations in other organism, e.g. the Chinese panda, have  
288 previously been related to feeding specialization [13, 34]. Interestingly, the sweetness receptor  
289 *TIR2* and bitter receptor *T2R* genes are expanded in the herbivorous fish but few or no copy was  
290 found in carnivorous fish. Collectively, these results not only indicate the genetic adaptation to  
291 herbivorous diet of *M. amblycephala*, but also provided a clear and comprehensive picture of  
292 adaptive evolutionary mechanisms of sensory systems in other fish species with different trophic  
293 specializations.

294       Some insects such as *Tenebrio molitor* [35] and *Neotermes koshunensis* [36], and the mollusc  
295 *Corbicula japonica* [37] have genes encoding endogenous cellulose degradation-related enzymes.  
296 However, all so far analyzed herbivorous vertebrates lack these genes and always rely on their gut  
297 microbiome to digest food [13, 38]. In herbivorous *M. amblycephala* and *C. idellus*, we also did  
298 not find any homologues of digestive cellulase genes. Interestingly, our work on the composition  
299 of gut microbiota of the two fish species identifies more than 7% cellulose-degrading bacteria,  
300 suggesting that the cellulose degradation of herbivorous fish largely depend on their gut  
301 microbiome.

302       Intermuscular bones (IBs) have evolved several times during teleost evolution [7, 39]. The  
303 developmental mechanisms and ossification processes forming IBs are dramatically distinct from  
304 other bones such as ribs, skeleton, vertebrae or spines. These usually develop from cartilaginous  
305 bone and are derived from the mesenchymal cell population by endochondral ossification [24, 40].  
306 However, IBs form directly by intramembranous ossification and differentiate from osteoblasts  
307 within connective tissue, forming segmental, serially homologous ossifications in the myosepta.  
308 Although various methods of ossification of IB have been proposed, few experiments have been  
309 conducted to confirm the ossification process and little is known about the potential role of IB in  
310 teleosts. Based on our findings of over-represented functional properties of expanded gene  
311 families in cyprinid lineage (Figure 2C) and evidence from the comparative transcriptome  
312 analyses of early developmental stages of IB formation (Figure 3A), we provide molecular  
313 evidence that IB might play significant roles not only in regulating muscle contraction but also in  
314 active remodeling at the bone-muscle interface and coordination of cellular events.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

315 It has previously been found that some major developmental signals including bone  
316 morphogenetic proteins (BMPs), fibroblast growth factors (FGFs), and WNT, together with  
317 calcium/calmodulin signaling [24, 41-43], are essential for regulating the differentiation and  
318 function of osteoblasts and osteocytes and for regulating the RANKL signaling pathway for  
319 osteoclasts [44] in intramembranous bone development. In agreement with this concept, our  
320 comparative transcriptome analysis of muscle, IB and connective tissues uncovered that 35 bone  
321 formation regulatory genes involved in these signals were highly up-regulated in IB. Taken  
322 together, these results suggest that IB indeed undergoes an intramembranous ossification process,  
323 is regulated by bone-specific signaling pathways, and underlies a homeostasis of maintenance,  
324 repair and remodeling.

## 325 **Conclusions**

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

## 333 **Methods**

### 334 **Sampling and DNA Extraction**

335 DNA for genome sequencing was derived from a double haploid line from the *M. amblycephala*  
336 genetic breeding center at Huazhong Agricultural University (Wuhan, Hubei, China). Fish blood  
337 was collected from adult female fish caudal vein using sterile injectors with pre-added  
338 anticoagulant solutions following anesthetized with MS-222 and sterilization with 75% alcohol.  
339 Genomic DNA was extracted from the whole blood. All experimental procedures involving fish  
340 were performed in accordance with the guidelines and regulations of the National Institute of  
341 Health Guide for the Care and Use of Laboratory Animals. The experiments were also approved  
342 by the Animal Care and Use Committee of Huazhong Agricultural University.

## 343 **Genomic Sequencing and Assembly**

1  
2 344 Libraries with different insert sized inserts of 170 bp, 500 bp, 800 bp, 2 Kb, 5 Kb, 10 Kb and 20  
3  
4 345 Kb were constructed from the genomic DNA at BGI-Shenzhen. The libraries were sequenced  
5  
6 346 using a HiSeq2000 instrument. In total, 11 libraries, sequenced in 23 lanes were constructed. To  
7  
8 347 obtain high quality data, we applied filtering criteria for the raw reads. As a result, 142.55 Gb  
9  
10 348 (129.59X) of filtered data were used to complete the genome assembly using SOAPdenovo-1.05  
11  
12 349 [13]. Only filtered data were used in the genome assembly. First, the short insert size library data  
13  
14 350 were used to construct a de Bruijn graph. The tips, merged bubbles and connections with low  
15  
16 351 coverage were removed before resolving the small repeats. Second, all high-quality reads were  
17  
18 352 realigned with the contig sequences. The number of shared paired-end relationships between pairs  
19  
20 353 of contigs was calculated and weighted with the rate of consistent and conflicting paired ends  
21  
22 354 before constructing the scaffolds in a stepwise manner from the short-insert size paired ends to the  
23  
24 355 long-insert size paired ends. Third, the gaps between the constructed scaffolds were composed  
25  
26 356 mainly of repeats, which were masked during scaffold construction. These gaps were closed using  
27  
28 357 the paired-end information to retrieve read pairs in which one end mapped to a unique contig and  
29  
30 358 the other was located in the gap region. Subsequently, local assembly was conducted for these  
31  
32 359 collected reads.  
33

## 360 **Genome Annotation**

361 The genome was searched for repetitive elements using Tandem Repeats Finder (version 4.04)  
362 [45]. TEs were identified using homology-based approaches. The Repbase (version 16.10) [46]  
363 database of known repeats and a *de novo* repeat library generated by RepeatModeler were used.  
364 This database was mapped using the software of RepeatMasker (version 3.3.0). Four types of  
365 non-coding RNAs (microRNAs, transfer RNAs, ribosomal RNAs and small nuclear RNAs) were  
366 also annotated using tRNAscan-SE (version 1.23) and the Rfam database45 (Release 9.1) [47].  
367 For gene prediction, *de novo* gene prediction, homology-based methods and RNA-seq data were  
368 used to perform gene prediction. For the sequence similarity based prediction, *D. rerio*, *G.*  
369 *aculeatus*, *O. niloticus*, *O. latipes* and *G. morhua* protein sequences were downloaded from  
370 Ensembl (release 73) and were aligned to the *M. amblycephala* genome using TBLASTN. Then  
371 homologous genome sequences were aligned against the matching proteins using GeneWise [48]



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

372 to define gene models. Augustus was employed to predict coding genes using appropriate  
373 parameters in *de novo* prediction. For the RNA-seq based prediction, we mapped transcriptome  
374 reads to the genome assembly using TopHat [49]. Then, we combined TopHat mapping results  
375 together and applied Cufflinks [50] to predict transcript structures. All predicted gene structures  
376 were integrated by GLEAN [51] (<http://sourceforge.net/projects/glean-gene/>) to obtain a  
377 consensus gene set.

### 378 **Phylogenetic Tree Reconstruction and Divergence Time Estimation**

379 The coding sequences of single-copy gene families conserved among *M. amblycephala*, *C. idellus*,  
380 *C. carpio*, *D. rerio*, *C. semilaevis*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *O. mykiss*, *O.*  
381 *niloticus*, *O. latipes*, *C. milii* and *Fugu rubripes* (Ensembl Gene v.77) were extracted and aligned  
382 with guidance from amino-acid alignments created by the MUSCLE program [52]. The individual  
383 sequence alignments were then concatenated to form one supermatrix. PhyML [53, 54] was  
384 applied to construct the phylogenetic tree under an HKY85+gamma model for nucleotide  
385 sequences. ALRT values were taken to assess the branch reliability in PhyML. The same set of  
386 codon sequences at position 2 was used for phylogenetic tree construction and estimation of the  
387 divergence time. The PAML mcmctree program (PAML version 4.5) [55, 56] was used to  
388 determine divergence times with the approximate likelihood calculation method and the  
389 'correlated molecular clock' and 'REV' substitution model.

### 390 **Gene Family Expansion and Contraction Analysis**

391 Protein sequences of *M. amblycephala* and 11 other related species (Ensembl Gene v.77)) were  
392 used in BLAST searches to identify homologs. We identified gene families using CAFÉ [57],  
393 which employs a random birth and death model to study gene gains and losses in gene families  
394 across a user-specified phylogeny. The global parameter  $\lambda$ , which describes both the gene birth ( $\lambda$ )  
395 and death ( $\mu = -\lambda$ ) rate across all branches in the tree for all gene families, was estimated using  
396 maximum likelihood [53]. A conditional *P*-value was calculated for each gene family, and families  
397 with conditional *P*-values less than the threshold (0.05) were considered as having notable gain or  
398 loss. We identified branches responsible for low overall *P*-values of significant families.

### 399 **Detection of Positively Selected Genes**

400 We calculated Ka/Ks ratios for all single copy orthologs of *M. amblycephala* and *C. semilaevis*, *D.*

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
401 *rerio*, *G. morhua*, *O. niloticus* and *C. carpio*. Alignment quality was essential for estimating  
402 positive selection. Thus, orthologous genes were first aligned by PRANK [58], which is  
403 considerably conservative for inferring positive selection. We used Gblocks [59] to remove  
404 ambiguously aligned blocks within PRANK alignments and employed 'codeml' in the PAML  
405 package with the free-ratio model to estimate Ka, Ks, and Ka/Ks ratios on different branches. The  
406 differences in mean Ka/Ks ratios for single-copy genes between *M. amblycephala* and each of the  
407 other species were compared using paired Wilcoxon rank sum tests. Genes that showed values of  
408 Ka/Ks higher than 1 along the branch leading to *M. amblycephala* were reanalyzed using the  
409 codon based branch-site tests implemented in PAML. The branch-site model allowed  $\omega$  to vary  
410 both among sites in the protein and across branches, and was used to detect episodic positive  
411 selection.

#### 412 **Prediction of Olfactory Receptor Genes**

413 Olfactory receptor genes were identified by previously described methods [60], with the exception  
414 of a first-round TBLASTN [61] search, in which 1,417 functional olfactory receptor genes from *H.*  
415 *sapiens*, *D. rerio*, *L. chalumnae*, *Lepisosteus oculatus*, *L. vexillifer*, *O. niloticus*, *O. latipes*, *F.*  
416 *rubripes* and *Xenopus tropicalis* were used as queries. We then predicted the structure of  
417 sequenced genes using the blast-hit sequence with the software GeneWise extending in both 3' and  
418 5' directions along the genome sequences. To construct phylogenetic trees, the amino-acid  
419 sequences encoded by olfactory receptor genes were first aligned using the program MUSCLE  
420 nested in MEGA 5.10 [62]. We then constructed the phylogenetic tree using the neighbor-joining  
421 method [63] with Poisson correction distances using the program MEGA 5.10.

#### 422 **RNA Sequencing Analysis**

423 RNA was extracted from total fish samples at different stages and from individual muscle,  
424 connective tissue, and intermuscular bone samples of adult *M. amblycephala*. Paired-end RNA  
425 sequencing was performed using the Illumina HiSeq 2000 platform. Low quality score reads were  
426 filtered and the clean data were aligned to the reference genome using Bowtie [64]. Genes and  
427 isoforms expression level were quantified by a software package: RSEM (RNASeq by  
428 Expectation Maximization) [65]. Gene expression levels were calculated by using the RPKM  
429 method (Reads per kilobase transcriptome per million mapped reads) [66] and adjusted by a

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

430 scaling normalization method [67]. Differentially expressed genes (DEGs) were detected using  
431 DESeq [68]. Annotation of DEGs were mapped to GO categories in the database  
432 (<http://www.geneontology.org/>) and the number of genes for every term were calculated to  
433 identify GO terms that were significantly enriched in the input list of DEGs. The calculated  
434 *P*-value was adjusted by the Bonferroni Correction, taking corrected *P*-value  $\leq 0.05$  as a threshold.  
435 KEGG [69] automatic annotation was used to perform pathway enrichment analysis of DEGs.

### 436 **Additional file**

437 Additional file 1: Supplementary Material contains Supplementary Figs. S1–S35, Supplementary  
438 Tables S1–S30, Supplementary Note, and Supplementary References.

### 439 **Competing interests**

440 The authors declare that they have no competing interests.

### 441 **Authors' Contributions**

442 W.W. initiated and conceived the project and provided scientific input. X.Q. organized financial  
443 support and designed the project. M.S. discussed the data, wrote and modified the paper. H.L. and  
444 C.C. conducted the biological experiments, analyzed the data and wrote the paper with input from  
445 other authors. I.E. wrote and modified the manuscript and discussed the data. The RAD sequence  
446 data analyses and the genetic map construction were performed by Z.G., Y.G., J.J. and X.J.  
447 Genome assembly and annotation were performed by J.M., H.C., M.X. and J.C. X.Z., W.L., R.L.,  
448 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 samples  
449 preparation and data collection. J.L. and J.C. identified the gene families and analyzed the  
450 RNA-seq data. M.B. coordinated the project. S.Z. and X.F. modified the manuscript and discussed  
451 the data. All authors read the manuscript and provided comments and suggestions for  
452 improvements. The authors declare no competing financial interests.

### 453 **Acknowledgements**

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

458 of Wuhan City (2015030809020365).

459 **Data access**

460 The *M. amblycephala* genome sequence and related data have been deposited at the *M.*  
461 *amblycephala* genome database and can be downloaded from  
462 <http://bream.hzau.edu.cn/page/species/download.html>.

463 **References**

- 464 1. FAO Fisheries and Aquaculture Department. The State of World Fisheries and Aquaculture  
465 2012 (Food and Agriculture Organization of the United Nations, Rome, 2014).
- 466 2. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, et al. The genome  
467 sequence of Atlantic cod reveals a unique immune system. *Nature*. 2011; 477:207–10.
- 468 3. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al. The rainbow trout  
469 genome provides novel insights into evolution after whole-genome duplication in vertebrates.  
470 *Nat. Commun.* 2014; 5, 3657.
- 471 4. Wu C, Zhang D, Kan M, Lv Z, Zhu A, Su Y, et al. The draft genome of the large yellow  
472 croaker reveals well-developed innate immunity. *Nat. Commun.* 2014; 5, 5227.
- 473 5. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for  
474 adaptive radiation in African cichlid fish. *Nature*. 2014; 513:375–82.
- 475 6. Chen S, Zhang G, Shao C, Huang Q, Liu G, Zhang P, et al. Whole-genome sequence of a  
476 flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic  
477 lifestyle. *Nat. Genet.* 2014; 46:253–60.
- 478 7. Gemballa S, Britz R. Homology of intermuscular bones in acanthomorph fishes. *Am. Mus.*  
479 *Novit.* 1998; 3241:1–25.
- 480 8. Danos N, Ward AB. The homology and origins of intermuscular bones in fishes: Phylogenetic  
481 or biomechanical determinants? *Biol. J. Linn. Soc.* 2012; 106:607–22.
- 482 9. Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, et al. Genome sequence and genetic diversity  
483 of the common carp, *Cyprinus carpio*. *Nat. Genet.* 2014; 46:1212–9.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- 484 10. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, et al. The draft genome of the grass carp  
485 (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation.  
486 Nat. Genet. 2015; 47:625–31.
  - 487 11. Gao Z, Luo W, Liu H, Zeng C, Liu X, Yi S, et al. Transcriptome analysis and SSR/SNP  
488 markers information of the blunt snout bream (*Megalobrama amblycephala*). PLoS One.  
489 2012; 7, e42637.
  - 490 12. Yi S, Gao ZX, Zhao H, Zeng C, Luo W, Chen B, et al. Identification and characterization of  
491 microRNAs involved in growth of blunt snout bream (*Megalobrama amblycephala*) by  
492 Solexa sequencing. BMC Genomics. 2013; 14, 754.
  - 493 13. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and *de novo* assembly of the  
494 giant panda genome. Nature. 2010; 463:311–7.
  - 495 14. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis  
496 of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61.
  - 497 15. Etienne-Manneville S. Actin and microtubules in cell motility: Which one is in control?  
498 Traffic. 2004; 5:470–7.
  - 499 16. Even-Ram S, Doyle AD, Conti MA, Matsumoto K, Adelstein RS, Yamada KM. Myosin IIA  
500 regulates cell motility and actomyosin-microtubule crosstalk. Nat. Cell Biol. 2007;  
501 9:299–309.
  - 502 17. Sheetz MP, Felsenfeld DP, Galbraith CG. Cell migration: Regulation of force on  
503 extracellular- complexes. Trends Cell Biol. 1998; 8:51–4.
  - 504 18. Gunst SJ, Zhang W. Actin cytoskeletal dynamics in smooth muscle: a new paradigm for the  
505 regulation of smooth muscle contraction. Am. J. Physiol. Cell Physiol. 2008; 295:C576–87.
  - 506 19. Webb RC. Smooth muscle contraction and relaxation. Adv. Physiol. Educ. 2003; 27:201–6.
  - 507 20. Ulrich TA, De Juan Pardo EM, Kumar S. The mechanical rigidity of the extracellular matrix  
508 regulates the structure, motility, and proliferation of glioma cells. Cancer Res. 2009;  
509 69:4167–74.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- 510 21. Ridley AJ. Rho GTPases and cell migration. *J. Cell Sci.* 2001; 114:2713–22.
- 511 22. Etienne-Manneville S, Hall A. Rho GTPases in Cell Biology. *Nature.* 2002; 420:629–35.
- 512 23. Ridley AJ. Cell Migration: Integrating Signals from Front to Back. *Science.* 2003;  
513 302:1704–9.
- 514 24. Ornitz D, Marie P. FGF signaling pathways in endochondral and intramembranous bone  
515 development and human genetic disease. *Genes Dev.* 2002; 16:1446–65.
- 516 25. Fakhry A, Ratisoontorn C, Vedhachalam C, Salhab I, Koyama E, Leboy P, et al. Effects of  
517 FGF-2/-9 in calvarial bone cell cultures: Differentiation stage-dependent mitogenic effect,  
518 inverse regulation of BMP-2 and noggin, and enhancement of osteogenic potential. *Bone.*  
519 2005; 36:254–66.
- 520 26. Sato K, Suematsu A, Nakashima T, Takemoto-Kimura S, Aoki K, Morishita Y, et al.  
521 Regulation of osteoclast differentiation and function by the CaMK-CREB pathway. *Nat. Med.*  
522 2006; 12:1410–6.
- 523 27. Niimura Y, Nei M. Evolutionary dynamics of olfactory and other chemosensory receptor  
524 genes in vertebrates. *J. Hum. Genet.* 2006; 51:505–17.
- 525 28. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene  
526 repertoires: roles of chance and necessity. *Nat. Rev. Genet.* 2008; 9:951–63.
- 527 29. Lopez C, Raper J. Cloning and functional characterization of odorant receptors expressed in  
528 the zebrafish olfactory system. *FASEB J.* 2015; 29:727–37.
- 529 30. Niimura Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative  
530 genome analysis among 23 chordate species. *Genome Biol. Evol.* 2009; 1:34–44.
- 531 31. Lindemann B. Receptors and transduction in taste. *Nature.* 2001; 413:219–25.
- 532 32. Chandrashekar J, Mueller KL, Hoon MA, Adler E, Feng L, Guo W, et al. T2Rs function as  
533 bitter taste receptors. *Cell.* 2000; 100:703–11.
- 534 33. Nelson G, Chandrashekar J, Hoon MA, Feng L, Zhao G, Ryba NJ, et al. An amino-acid taste  
535 receptor. *Nature.* 2002; 416:199–202.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- 536 34. Jiang P, Josue J, Li X, Glaser D, Li W, Brand JG, et al. From the Cover: Major taste loss in  
537       carnivorous mammals. *Proc. Natl. Acad. Sci. USA.* 2012; 109:4956–61.
- 538 35. Ferreira AHP, Marana SR, Terra WR, Ferreira C. Purification, molecular cloning, and  
539       properties of a  $\beta$ -glycosidase isolated from midgut lumen of *Tenebrio molitor* (Coleoptera)  
540       larvae. *Insect Biochem. Mol. Biol.* 2001; 31:1065–76.
- 541 36. Tokuda G, Saito H, Watanabe H. A digestive  $\beta$ -glucosidase from the salivary glands of the  
542       termite, *Neotermes koshunensis* (Shiraki): Distribution, characterization and isolation of its  
543       precursor cDNA by 5'- and 3'-RACE amplifications with degenerate primers. *Insect*  
544       *Biochem. Mol. Biol.* 2002; 32:1681–9.
- 545 37. Sakamoto K, Uji S, Kurokawa T, Toyohara H. Molecular cloning of endogenous  
546        $\beta$ -glucosidase from common Japanese brackish water clam *Corbicula japonica*. *Gene.* 2009;  
547       435:72–9.
- 548 38. Zhu L, Wu Q, Dai J, Zhang S, Wei F. Evidence of cellulose metabolism by the giant panda gut  
549       microbiome. *Proc. Natl. Acad. Sci. USA.* 2011; 108:17714–9.
- 550 39. Bird NC, Mabee PM. Developmental morphology of the axial skeleton of the zebrafish, *Danio*  
551       *rerio* (Ostariophysi: Cyprinidae). *Dev. Dyn.* 2003; 228:337–57.
- 552 40. Ortega N, Behonick DJ, Werb Z. Matrix remodeling during endochondral ossification. *Trends*  
553       *Cell Biol.* 2004; 14:86–93.
- 554 41. Chen G, Deng C, Li YP. TGF- $\beta$  and BMP signaling in osteoblast differentiation and bone  
555       formation. *Int. J. Biol. Sci.* 2012; 8:272–88.
- 556 42. Harada S, Rodan GA. Control of osteoblast function and regulation of bone mass. *Nature.*  
557       2003; 423:349–55.
- 558 43. Long F. Building strong bones: molecular regulation of the osteoblast lineage. *Nat. Rev. Mol.*  
559       *Cell Biol.* 2011; 13:27–38.
- 560 44. Boyle WJ, Simonet WS, Lacey DL. Osteoclast differentiation and activation. *Nature.* 2003;  
561       423:337–42.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- 562 45. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.*  
563 1999; 27:573–80.
- 564 46. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase  
565 Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 2005;  
566 110:462–7.
- 567 47. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: Annotating  
568 non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005; 33:121–4.
- 569 48. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res.* 2004; 14:988–95.
- 570 49. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq.  
571 *Bioinformatics.* 2009; 25:1105–11.
- 572 50. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript  
573 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform  
574 switching during cell differentiation. *Nat. Biotechnol.* 2010; 28:511–5.
- 575 51. Elisk CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey  
576 bee consensus gene set. *Genome Biol.* 2007; 8, R13.
- 577 52. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.  
578 *Nucleic Acids Res.* 2004; 32:1792–7.
- 579 53. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and  
580 methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML  
581 3.0. *Syst. Biol.* 2010; 59:307–21.
- 582 54. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by  
583 maximum likelihood. *Syst. Biol.* 2003; 52:696–704.
- 584 55. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;  
585 24:1586–91.
- 586 56. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock  
587 using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* 2006; 23:212–26.



- 588 57. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics*.  
589 2007; 177:1941–9.
- 590 58. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with  
591 insertions. *Proc. Natl. Acad. Sci. USA*. 2005; 102:10557–62.
- 592 59. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and  
593 ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 2007; 56:564–77.
- 594 60. Niimura Y. Evolutionary dynamics of olfactory receptor genes in chordates: interaction  
595 between environments and genomic contents. *Hum. Genomics*. 2009; 4:107–18.
- 596 61. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and  
597 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;  
598 25:3389–402.
- 599 62. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary  
600 analysis of DNA and protein sequences. *Brief. Bioinform.* 2008; 9:299–306.
- 601 63. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic  
602 trees. *Mol. Biol. Evol.* 1987; 4:406–25.
- 603 64. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 2012;  
604 9:357–9.
- 605 65. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or  
606 without a reference genome. *BMC Bioinformatics*. 2011; 12, 323.
- 607 66. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying  
608 mammalian transcriptomes by RNA-Seq. *Nat. Methods*. 2008; 5:621–8.
- 609 67. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis  
610 of RNA-seq data. *Genome Biol.* 2010; 11, R25.
- 611 68. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.*  
612 2010; 11, R106.
- 613 69. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking

1	614	genomes to life and the environment. <i>Nucleic Acids Res.</i> 2008; 36:480–4.
2		
3	615	
4		
5	616	
6		
7		
8	617	
9		
10	618	
11		
12		
13	619	
14		
15		
16	620	
17		
18		
19	621	
20		
21	622	
22		
23		
24	623	
25		
26	624	
27		
28		
29	625	
30		
31		
32	626	
33		
34		
35	627	
36		
37	628	
38		
39		
40	629	
41		
42	630	
43		
44		
45	631	
46		
47		
48	632	
49		
50		
51	633	
52		
53	634	
54		
55		
56	635	
57		
58	636	
59		
60		
61		
62		
63		
64		
65		

637 **Figure Legends**

1  
2  
3 638 **Figure 1** Global view of the *M. amblycephala* genome and syntenic relationship between  
4  
5 639 *Ctenopharyngodon idellus*, *M. amblycephala* and *Danio rerio*. (A) Global view of the *M.*  
6  
7 640 *amblycephala* genome. From outside to inside, the genetic linkage map (a); Anchors between the  
8  
9 641 genetic markers and the assembled scaffolds (b); Assembled chromosomes (c); GC content within  
10  
11 642 a 50-kb sliding window (d); Repeat content within a 500-kb sliding window (e); Gene distribution  
12  
13 643 on each chromosome (f); Different gene expression of three transcriptomes (g). (B) Syntenic  
14  
15 644 relationship between *C. idellus* (a), *M. amblycephala* (b) and *D. rerio* (c) chromosomes.

16  
17  
18 645 **Figure 2** Phylogenetic tree and comparison of orthologous genes in *M. amblycephala* and other  
19  
20 646 fish species. (A) Phylogenetic tree of teleosts using 316 single copy orthologous genes. The color  
21  
22 647 circles at the nodes shows the estimated divergence times using *O. latipes*–*F. rubripes*  
23  
24 648 [96.9~150.9Mya], *F. rubripes*–*D. rerio* [149.85~165.2Mya], *F. rubripes*–*C. milii* [416~421.75Mya]  
25  
26 649 (<http://www.timetree.org/>) as the calibration time. Pentagon represents four cyprinid fish with  
27  
28 650 intermuscular bones. S, silurian period; D, devonian period; C, carboniferous period; P, permian  
29  
30 651 period in Paleozoic; T, triassic period; J, jurassic and k-cretaceous period in Mesozoic; Pg,  
31  
32 652 paleogene in Cenozoic Era, N, Neogene. (B) Venn diagram of shared and unique orthologous gene  
33  
34 653 families in *M. amblycephala* and four other teleosts. (C) Over-represented GO annotations of  
35  
36 654 cyprinid-specific expansion gene families.

37  
38  
39  
40 655 **Figure 3** Regulation of genes related to intermuscular bone formation and function identified from  
41  
42 656 developmental stages and adult tissues transcriptome data. (A) Gene expression pattern involved  
43  
44 657 in muscle contraction regulated genes in early developmental stages corresponds to intermuscular  
45  
46 658 bone formation of *M. amblycephala*, (alizarin red staining). M, myosepta; IB, intermuscular bone.  
47  
48 659 (B) Scanning electron microscope photos of muscle tissues, connective tissues, and intermuscular  
49  
50 660 bone. (C) Distribution of intermuscular bone specific genes in GO annotations indicative of  
51  
52 661 abundance in protein binding, calcium ion binding, GTP binding functions. (D) Several  
53  
54 662 developmental signals regulating key steps of osteoblast and osteoclast differentiation in the  
55  
56 663 process of intramembranous ossification. Colored boxes indicate significantly up-regulated genes  
57  
58 664 in these signals specifically occurred in intermuscular bone.

665 **Figure 4** Molecular characteristics of sensory systems and the composition of gut microbiota in *M.*  
666 *amblycephala*. (A) Extensive expansion of olfactory receptor genes (ORs) in *M. amblycephala*  
667 compared with other teleosts. (B) Phylogeny of ‘beta’ type ORs in eight representative teleost  
668 species showing the significant expansion of ‘beta’ ORs in *M. amblycephala* and *C. idellus*. The  
669 pink background shows cyprinid-specific ‘beta’ types of ORs. (C) Umami, sweet and bitter tastes  
670 related gene families in teleosts with different feeding habits. (D) Structure of the umami receptor  
671 encoding T1R1 gene in cyprinid fish. (E) Relative abundance of microbial flora and taxonomic  
672 assignments in juvenile (LBSB), domestic adult (DBSB), wild adult (BSB) *M. amblycephala* and  
673 wild adult *C. idellus* (GC) samples at the phylum level.

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690 **Table**

691 **Table 1 Features of the *Megalobrama amblycephala* whole genome sequence**

4	Total genome size (Mb)	1,116
5	N90 length of scaffold (bp)	20,422
6	N50 length of scaffold (bp)	838,704
7	N50 length of contig (bp)	49,400
8	Total GC content (%)	37.30
9	Protein-coding genes number	23,696
10	Average gene length (bp)	15,797
11	Content of transposable elements (%)	34.18
12	Number of chromosomes	24
13	Number of makers in genetic map	5,317
14	Scaffolds anchored on linkage groups (LGs)	1,434
15	Length of scaffolds anchored on LGs (Mb)	779.54 (70%)

692

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Figure 1

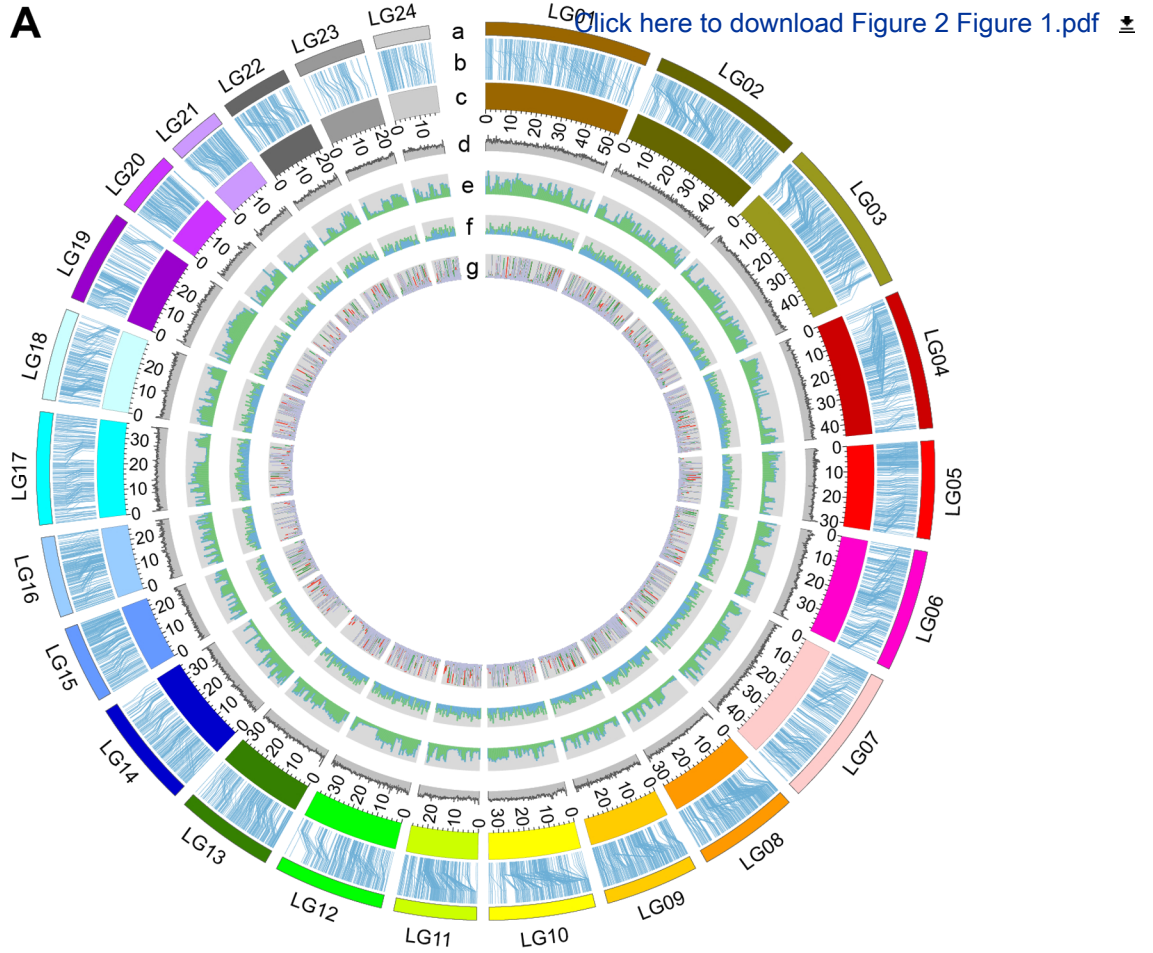
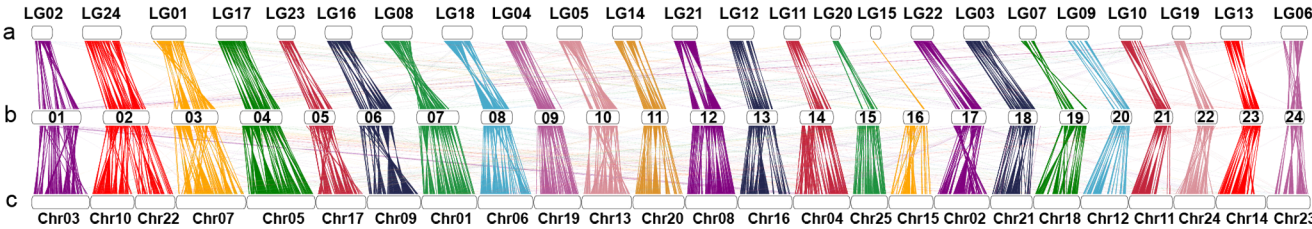
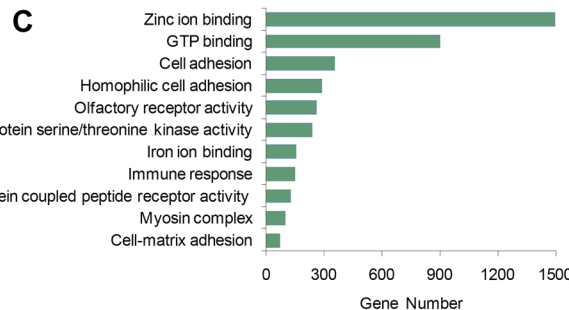
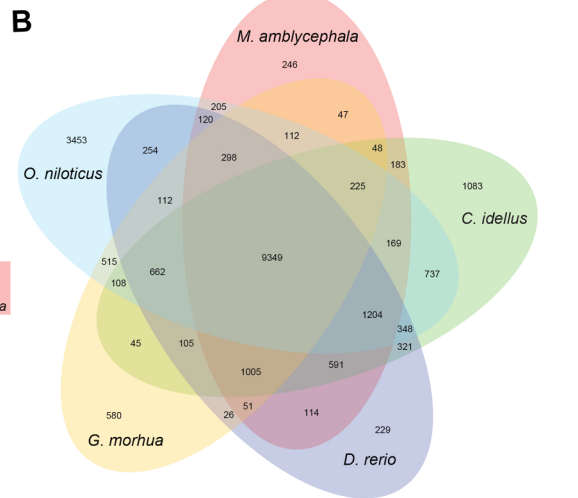
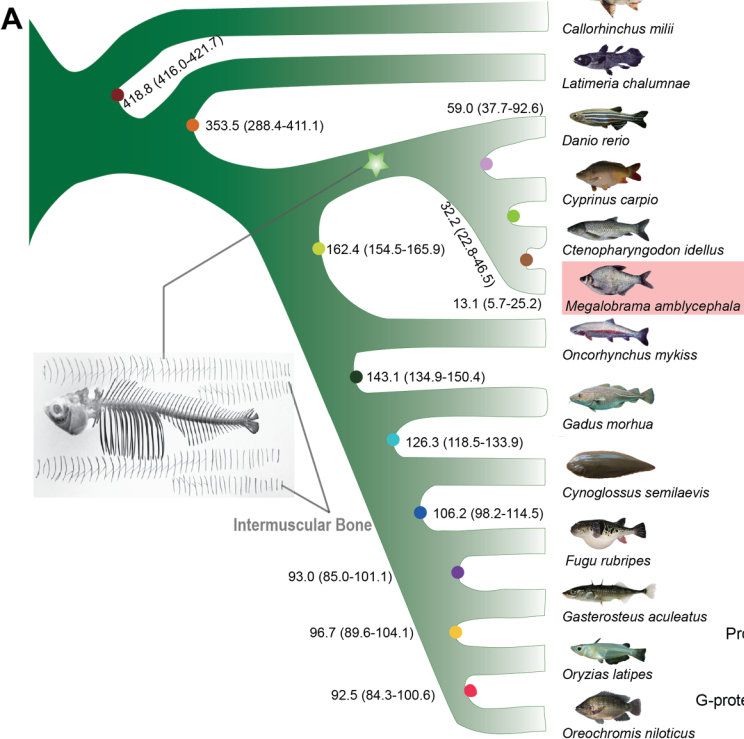
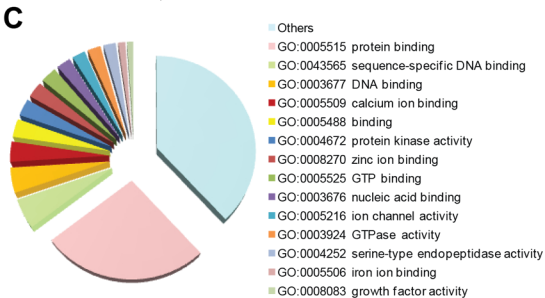
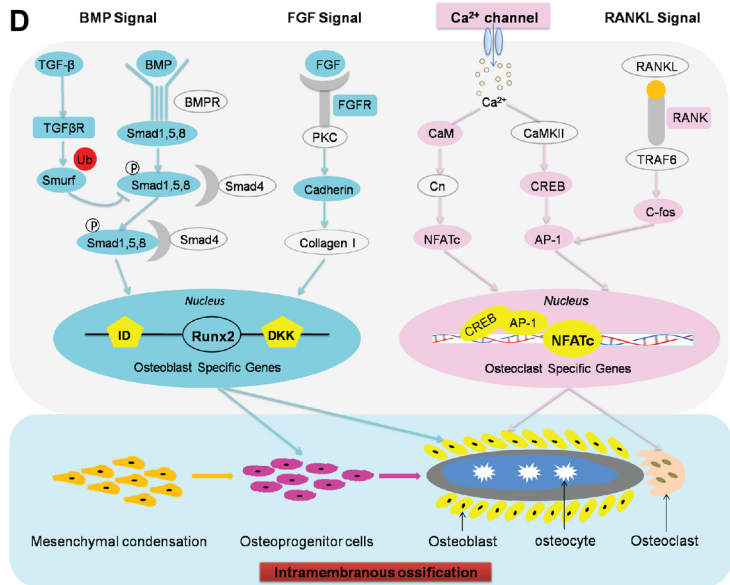
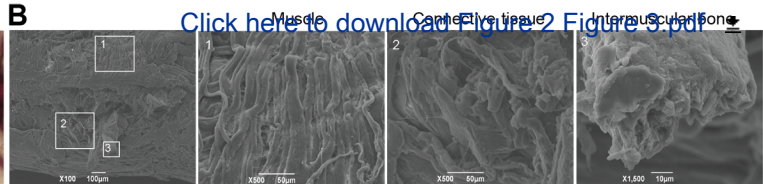
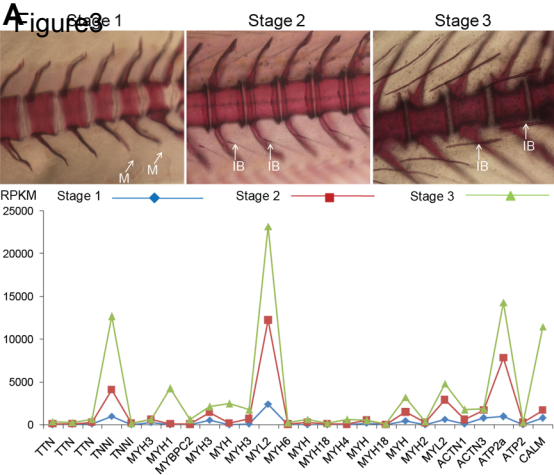
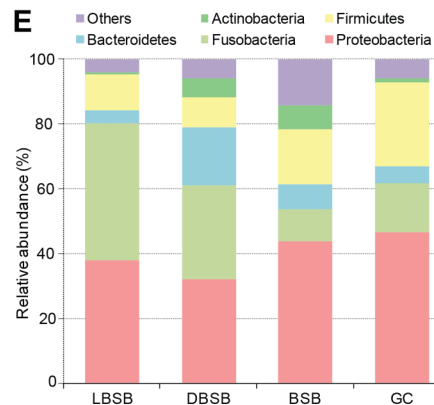
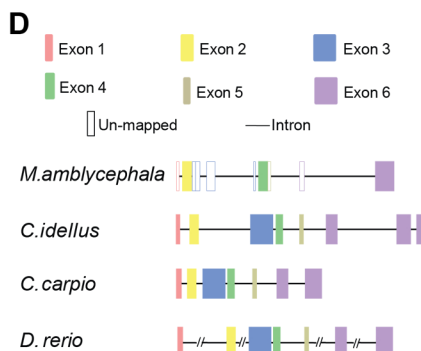
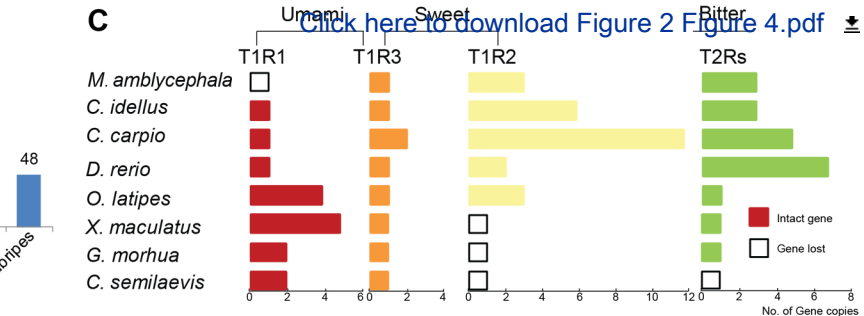
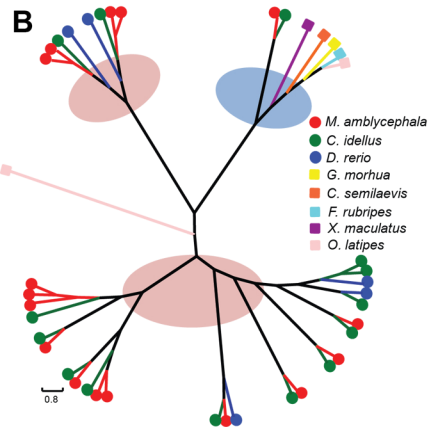
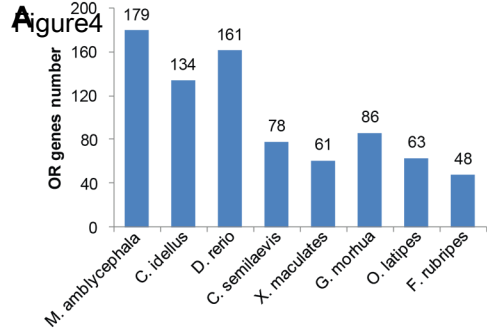
**A****B**

Figure2

[Click here to download Figure 2 Figure 2.pdf](#)











[Click here to access/download](#)

**Supplementary Material**

3 2016.9.8 Supplementary Material.pdf





Click here to access/download  
**Supplementary Material**  
3 Supplementary Note 1.xlsx

