

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-16-00088R2
Full Title:	The draft genome of <i>Megalobrama amblycephala</i> reveals the development of intermuscular bone and adaptation to herbivorous diet
Article Type:	Research
Funding Information:	
Abstract:	<p>Background: 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>Results: 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 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 <i>M. amblycephala</i>. The composition of gut microflora, which contributes to the herbivorous adaptation of <i>M. amblycephala</i>, was found to be similar to that of other herbivores.</p> <p>Conclusions: 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>
Corresponding Author:	Weimin Wang CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Han Liu
First Author Secondary Information:	
Order of Authors:	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 Scharl
	Xueqiao Qian
	Weimin Wang
Order of Authors Secondary Information:	
Response to Reviewers:	<p>16 March 2017 Dr. Hans Zauner Journal: GigaScience</p> <p>Dear Dr. Zauner,</p> <p>Manuscript No.: GIGA-D-16-00088R1 Title: "The draft genome of <i>Megalobrama amblycephala</i> reveals the development of intermuscular bone and adaptation to herbivorous diet" Author(s): 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 Scharl, Xueqiao Qian, Weimin Wang</p> <p>We have carefully read the referee's comments which you forwarded to us with your email of 20 January 2017. We would like to express our sincere thanks to the reviewer for the constructive comments. We have now addressed all the suggestions, and the manuscript has been edited accordingly. The major amendments are highlighted in red in the revised manuscript. Responses to the reviewer's comments are detailed below in this letter. Because of the amendments, the page and the line numbers referred to by the referee have now changed in the edited version of the manuscript. Please, note that all raw reads of genome sequencing and RAD-seq have been deposited at NCBI. Other data including assemblies, annotations, RNA-seq data, microbiome data, embedded image data and SNP markers were uploaded to the indicated ftp. We hope that with the amendments made in response to the reviewer's comments, the manuscript is now acceptable for publication in GigaScience.</p>

I look forward to hearing from you soon.

Yours sincerely,
Weimin Wang (PhD) (Correspondence author)
College of Fisheries
Huazhong Agricultural University
Wuhan 430070, P. R. China
E-mail address: wangwm@mail.hzau.edu.cn
Tel: +86-27-8728 4292; Fax: +86-27-8728 4292

Response to Reviewer

Reviewer Report

Reviewer #2: It is quite demanding to properly review a paper with this amount of data and therefore the structure of the manuscript is very important. In addition, a figure showing the overview of the workflow would have been helpful.

Author response: We have now included a schematic figure to show the workflow that should be helpful for a better understanding of the workflow behind the data presented in the supporting information as additional file 1: Figure S2. (Line 92)

Concerns of Reviewer 1 and 2 are partially answered and altered accordingly in the manuscript. However some important aspects have to be addressed prior to acceptance.

1. Database submission: All raw data and molecular markers have to be available. The submission to the NCBI SRA database is not sufficient as this database is for sequencing data. There are specific databases for molecular markers such as SNPs.

Author response: We have uploaded the assemblies, annotations, RNA-seq, microbiome data and molecular markers (SNPs) to the ftp (<ftp://user28@climb.genomics.cn>). The URL is included in the revised manuscript.

2. Still the structure has to be revised and the outcome better worked out. The result section comprises parts of the discussion and the discussion contains again descriptions of the results. Line 227 -228 is just one example.

Author response: We have now carefully revised the manuscript according to your suggestions. The parts of the discussion in the results section have been removed and we also refined the discussion section.

3. The reading of the manuscript is still tedious as no structure /concept of the work is given.

Author response: According to your suggestions, we have now carefully revised the manuscript. We hope it is now clear and structured for reading and understanding with these changes and the additional file 1: Figure S2 showing the workflow.

4. Description of the differential expression analysis remains poor. No numbers of significant expressed transcripts are given nor was any further meta-analysis performed. In addition the threshold for significance has not been provided. The sentence at line 502 "DEGs were detected using DESeq" is not informative enough.
Author response: We are sorry for the simple description about differential expression analysis. In the first submitted Supplementary Note, we had detailed the transcriptome analyses. Because there is no Supplementary Note section in the Journal, we deleted it and selectively added some necessary information in the Methods section of the manuscript. We have now clarified these questions in the new revised manuscript according to your suggestions. (Line 175-177, 184-193, 492-496)

5. Line 167 the word "many" in the results section is not an appropriate expression here.

Author response: This sentence has been modified. (Line 163-166)

6. Please provide revised Figure files.

Author response: The revised Figure files are now attached with this new revised manuscript.

7. Still authors do not explain what is meant by "expanded gene families". Figure 2C does not show "gene families" but GO categories. E.g myosin complex belongs to

	<p>GO:0016459. A gene family comprises similar genes, formed by duplication of a single original gene. Those genes normally have also similar biochemical functions. The most known example is the family comprising the human hemoglobin subunits. Author response: We apologize for the confusion. Indeed, Figure 2C does not show expanded gene families but the over-represented GO annotations of cyprinid-specific expansion genes. This has been corrected now in the manuscript and the figure legends. (Line 169-170, 787)</p> <p>8. What is meant by "resource friendly"? It is still not clear what the authors want to pinpoint to. Author response: This sentence was ambiguous and we apologize for the confusion that it generated. This sentence has been re-phrased and reads now: "Reports on draft genomes of herbivorous and omnivorous species....." in the revised manuscript. (Line 69)</p> <p>9. Line 83 change expression: "remain obscure" Author response: This sentence has been re-phrased and reads now: "the molecular genetic basis and the evolution of this unique structure are still unclear". (Line 82)</p> <p>10. Line 88: "many miRNA-mRNA interactions" not really informative. Author response: This sentence has been re-phrased and reads now: "1,136 miRNA-mRNA interaction pairs". (Line 87)</p> <p>11. Line 129 "psuedo-chromosomes". Most probably "pseudo-chromosomes" are meant. Please carefully proof read the manuscript for typos and inadequate expressions. Author response: We apologize for the spelling mistake. It should be "pseudo-chromosomes". This has been corrected in the revised manuscript. (Line 129)</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist . Information essential to interpreting the data presented should be made available in the figure legends.	
Have you included all the information requested in your manuscript?	
Resources	Yes
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 Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.	

<p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (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 Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

1 16 March 2017

2 Dr. Hans Zauner

3
4 Journal: GigaScience

5
6
7 Dear Dr. Zauner,

8
9 **Manuscript No.: GIGA-D-16-00088R1**

10 Title: **"The draft genome of *Megalobrama amblycephala* reveals the development of**
11 **intermuscular bone and adaptation to herbivorous diet"**

12
13 Author(s): Han Liu, Chunhai Chen, Zexia Gao, Jiumeng Min, Yongming Gu, Jianbo Jian, Xiewu
14 Jiang, Huimin Cai, Ingo Ebersberger, Meng Xu, Xinhui Zhang, Jianwei Chen, Wei Luo, Boxiang
15 Chen, Junhui Chen, Hong Liu, Jiang Li, Ruifang Lai, Mingzhou Bai, Jin Wei, Shaokui Yi,
16 Huanling Wang, Xiaojuan Cao, Xiaoyun Zhou, Yuhua Zhao, Kaijian Wei, Ruibin Yang, Bingnan
17 Liu, Shancen Zhao, Xiaodong Fang, Manfred Schartl, Xueqiao Qian, Weimin Wang
18
19
20
21
22
23

24 We have carefully read the referee's comments which you forwarded to us with your email of 20
25 January 2017. We would like to express our sincere thanks to the reviewer for the constructive
26 comments. We have now addressed all the suggestions, and the manuscript has been edited
27 accordingly. The major amendments are highlighted in red in the revised manuscript. Responses to
28 the reviewer's comments are detailed below in this letter. Because of the amendments, the page
29 and the line numbers referred to by the referee have now changed in the edited version of the
30 manuscript. Please, note that all raw reads of genome sequencing and RAD-seq have been
31 deposited at NCBI. Other data including assemblies, annotations, RNA-seq data, microbiome data,
32 embedded image data and SNP markers were uploaded to the indicated ftp. We hope that with the
33 amendments made in response to the reviewer's comments, the manuscript is now acceptable for
34 publication in GigaScience.
35
36
37
38
39
40
41
42
43

44 I look forward to hearing from you soon.

45
46 Yours sincerely,

47 Weimin Wang (PhD) (Correspondence author)

48 College of Fisheries

49 Huazhong Agricultural University

50 Wuhan 430070, P. R. China

51 E-mail address: wangwm@mail.hzau.edu.cn

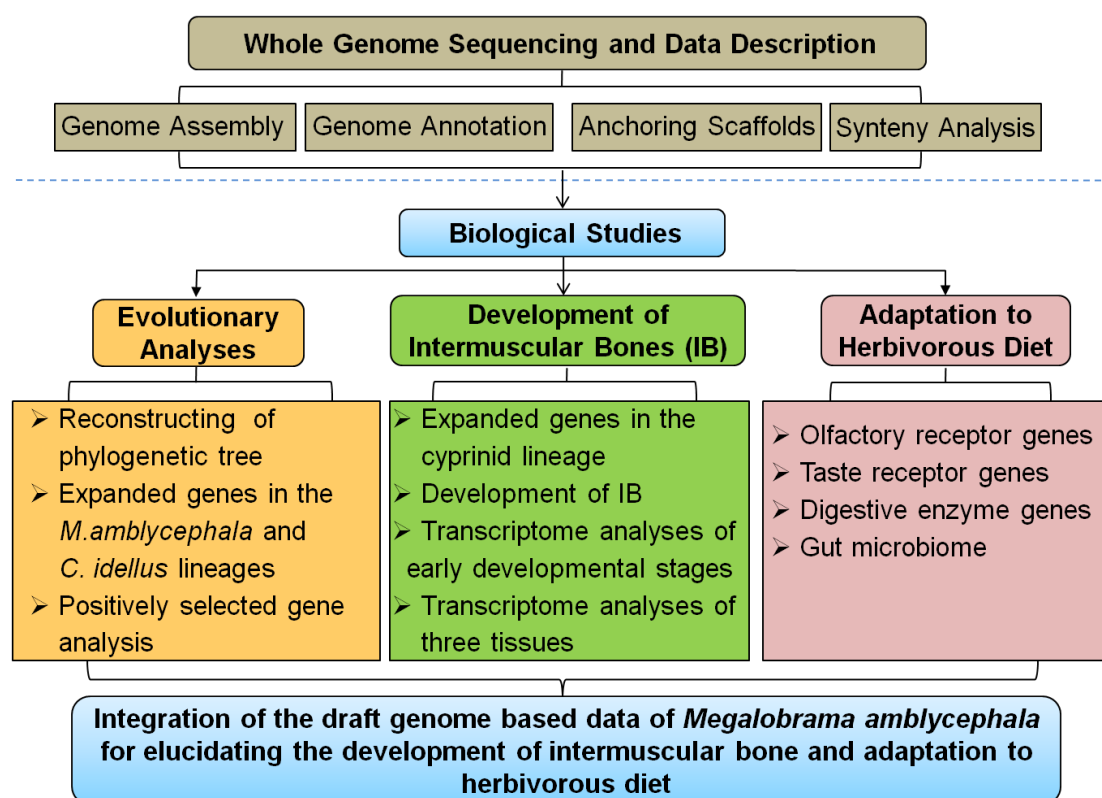
52 Tel: +86-27-8728 4292; Fax: +86-27-8728 4292
53
54
55
56
57
58
59
60
61
62
63
64
65

Response to Reviewer

Reviewer Report

Reviewer #2: It is quite demanding to properly review a paper with this amount of data and therefore the structure of the manuscript is very important. In addition, a figure showing the overview of the workflow would have been helpful.

Author response: We have now included a schematic figure to show the workflow that should be helpful for a better understanding of the workflow behind the data presented in the supporting information as additional file 1: Figure S2. (Line 92)



Additional file 1: Figure S2 Schematic diagram illustrating the workflow of *M. Amblycephala* genome.

Concerns of Reviewer 1 and 2 are partially answered and altered accordingly in the manuscript. However some important aspects have to be addressed prior to acceptance.

1. Database submission: All raw data and molecular markers have to be available. The submission to the NCBI SRA database is not sufficient as this database is for sequencing data. There are specific databases for molecular markers such as SNPs.

Author response: We have uploaded the assemblies, annotations, RNA-seq, microbiome data and

1 molecular markers (SNPs) to the ftp (ftp://user28@climb.genomics.cn). The URL is included in
2 the revised manuscript.
3

4
5 2. Still the structure has to be revised and the outcome better worked out. The result section
6 comprises parts of the discussion and the discussion contains again descriptions of the results.
7
8 Line 227 -228 is just one example.
9

10 **Author response:** We have now carefully revised the manuscript according to your suggestions.
11
12 The parts of the discussion in the results section have been removed and we also refined the
13 discussion section.
14
15
16

17
18 3. The reading of the manuscript is still tedious as no structure /concept of the work is given.
19

20 **Author response:** According to your suggestions, we have now carefully revised the manuscript.
21
22 We hope it is now clear and structured for reading and understanding with these changes and the
23 additional file 1: Figure S2 showing the workflow.
24
25
26

27 4. Description of the differential expression analysis remains poor. No numbers of significant
28 expressed transcripts are given nor was any further meta-analysis performed. In addition the
29 threshold for significance has not been provided. The sentence at line 502 "DEGs were detected
30 using DESeq" is not informative enough.
31
32
33

34 **Author response:** We are sorry for the simple description about differential expression analysis.
35
36 In the first submitted Supplementary Note, we had detailed the transcriptome analyses. Because
37 there is no Supplementary Note section in the Journal, we deleted it and selectively added some
38 necessary information in the Methods section of the manuscript. We have now clarified these
39 questions in the new revised manuscript according to your suggestions. (Line 175-177, 184-193,
40 492-496)
41
42
43
44
45
46

47
48 5. Line 167 the word "many" in the results section is not an appropriate expression here.
49

50 **Author response:** This sentence has been modified. (Line 163-166)
51
52

53 6. Please provide revised Figure files.
54

55 **Author response:** The revised Figure files are now attached with this new revised manuscript.
56
57

58 7. Still authors do not explain what is meant by "expanded gene families". Figure 2C does not
59 show "gene families" but GO categories. E.g myosin complex belongs to GO:0016459. A gene
60
61
62
63
64
65

1 family comprises similar genes, formed by duplication of a single original gene. Those genes
2 normally have also similar biochemical functions. The most known example is the family
3 comprising the human hemoglobin subunits.
4
5

6 **Author response:** We apologize for the confusion. Indeed, Figure 2C does not show expanded
7 gene families but the over-represented GO annotations of cyprinid-specific expansion genes. This
8 has been corrected now in the manuscript and the figure legends. (Line 169-170, 787)
9

10
11
12
13 8. What is meant by "resource friendly"? It is still not clear what the authors want to pinpoint to.
14

15 **Author response:** This sentence was ambiguous and we apologize for the confusion that it
16 generated. This sentence has been re-phrased and reads now: "Reports on draft genomes of
17 herbivorous and omnivorous species....." in the revised manuscript. (Line 69)
18
19
20
21

22 9. Line 83 change expression: "remain obscure"
23

24 **Author response:** This sentence has been re-phrased and reads now: "the molecular genetic basis
25 and the evolution of this unique structure are still unclear". (Line 82)
26
27
28

29 10. Line 88: "many miRNA-mRNA interactions" not really informative.
30

31 **Author response:** This sentence has been re-phrased and reads now: "1,136 miRNA-mRNA
32 interaction pairs". (Line 87)
33
34
35

36 11. Line 129 "psuedo-chromosomes". Most probably "pseudo-chromosomes" are meant. Please
37 carefully proof read the manuscript for typos and inadequate expressions.
38
39

40 **Author response:** We apologize for the spelling mistake. It should be "pseudo-chromosomes".
41 This has been corrected in the revised manuscript. (Line 129)
42
43
44

45 --
46

47 Please also take a moment to check our website at for any additional comments that were saved as
48 attachments. Please note that as GigaScience has a policy of open peer review, you will be able to
49 see the names of the reviewers.
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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^{1†}, Chunhai Chen^{2†}, Zexia Gao^{1†}, Jiumeng Min^{2†}, Yongming Gu^{3†}, Jianbo Jian^{2†}, Xiewu
7
8 4 Jiang³, Huimin Cai², Ingo Ebersberger⁴, Meng Xu², Xinhui Zhang¹, Jianwei Chen², Wei Luo¹,
9
10 5 Boxiang Chen^{1,3}, Junhui Chen², Hong Liu¹, Jiang Li², Ruifang Lai¹, Mingzhou Bai², Jin Wei¹,
11
12 6 Shaokui Yi¹, Huanling Wang¹, Xiaojuan Cao¹, Xiaoyun Zhou¹, Yuhua Zhao¹, Kaijian Wei¹,
13
14 7 Ruibin Yang¹, Bingnan Liu³, Shancen Zhao², Xiaodong Fang², Manfred Schartl^{5,*}, Xueqiao
15
16 8 Qian^{3,*}, Weimin Wang^{1,*}

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

24
25 12 †Equal contributors

26
27 13 ¹College 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 ²Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen 518083, China

34
35 17 ³Guangdong Haid Group Co., Ltd., Guangzhou 511400, China

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

40
41 20 ⁵Physiological Chemistry, University of Würzburg, Biozentrum, Am Hubland, and Comprehensive
42
43 21 Cancer Center Mainfranken, University Clinic Würzburg, Würzburg 97070, Germany and Texas
44
45 22 A&M Institute for Advanced Study and Department of Biology, Texas A&M University, College
46
47 23 Station, TX 77843, USA

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 be involved 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 contributes 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. Atlantic cod (*Gadus morhua*) [2], rainbow trout
65 (*Oncorhynchus mykiss*) [3], European sea bass (*Dicentrarchus labrax*) [4], yellow croaker
66 (*Larimichthys crocea*) [5], half-smooth tongue sole (*Cynoglossus semilaevis*) [6], tilapia
67 (*Oreochromis niloticus*) [7] and channel catfish (*Ictalurus punctatus*) [8] have been sequenced.
68 Yet, the majority of these species are carnivorous requiring large inputs of protein from wild
69 caught fish or other precious feed. **Reports on draft genomes of herbivorous and omnivorous**
70 species, in particular cyprinid fish are scarce. It is well known that cyprinids are currently the
71 economically most important group of teleosts for sustainable aquaculture. They grow to large
72 population sizes in the wild and already now account for the majority of freshwater aquaculture
73 production worldwide [1]. Among these, the herbivorous *Megalobrama amblycephala* (Yih, 1955),
74 a particularly eco-friendly and resource-conserving species, is predominant in aquaculture and has
75 been greatly developed in China (Additional file 1: Figure S1) [1]. However, most cyprinids,
76 including *M. amblycephala*, have a large number of intermuscular bones (IBs) in the trunk
77 musculature, which have an adverse effect on fish meat processing and consumption. IBs—a
78 unique form of bone occurring only in the more basal teleosts—are completely embedded within
79 the myosepta and are not connected to the vertebral column or any other bones [9, 10]. Our
80 previous study on IB development of *M. amblycephala* revealed that some miRNA-mRNA
81 interaction pairs may be involved in regulating bone development and differentiation [11].
82 **However, the molecular genetic basis and the evolution of this unique structure are still unclear.**
83 Unfortunately, the recent sequencing of two cyprinid genomes common carp (*Cyprinus carpio*)
84 [12] and grass carp (*Ctenopharyngodon idellus*) [13], which provided valuable information for
85 their genetic breeding, contributed little to the understanding of IB formation.

86 In an initial genome survey of *M. amblycephala*, we identified 25,697 single-nucleotide

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 polymorphism (SNP) [14], 347 conserved miRNAs [15], and 1,136 miRNA-mRNA interaction
88 pairs [11]. However, lack of a whole genome sequence resource limited a thorough investigation
89 of *M. amblycephala*. Here we report the first high-quality draft genome sequence of *M.*
90 *amblycephala*. Integrating this novel genome resource with tissue- and developmental
91 stage-specific gene expression information, as well as with meta-genome data to investigate the
92 composition of the gut microbiome (Workflow shown in additional file 1: Figure S2) provides
93 relevant insights into the function and evolution of two key features characterizing this species:
94 The formation of IB and the adaptation to herbivory. By that our study lays the foundation for
95 genetically optimizing *M. amblycephala* to further increase its relevance for securing human food
96 supply.

97 **Data description**

98 **Genome Assembly and Annotation**

99 The *M. amblycephala* genome was sequenced and assembled by a whole-genome shotgun strategy
100 using genomic DNA from a double-haploid fish (Additional file 1: Table S1). We assembled a
101 1.116 Gb reference genome sequence from 142.55 Gb (approximately 130-fold coverage) of clean
102 data [16] (Additional file 1: Tables S1 and S2, Figure S3). The contig and scaffold N50 lengths
103 reached 49 Kb and 839 Kb, respectively (Table 1). The largest scaffold spans 8,951 Kb and the
104 4,034 largest scaffolds cover 90% of the assembly. To assess the genome assembly quality, the
105 mapping of paired end sequence data from the short-insert size WGS libraries, as well as of
106 published ESTs [14] (Additional file 1: Tables S3 and S4) against the genome assembly indicated
107 that the number and extent of misassemblies is low. To further estimate the completeness of the
108 assembly and gene prediction, the benchmarking universal single-copy orthologs (BUSCO) [17]
109 analysis was used and the results showed that the assembly contains 81.4% complete and 9.1%
110 partial vertebrate BUSCO orthologues (Additional file 1: Table S5).

111 The *M. amblycephala* genome has an average GC content of 37.3%, similar to cyprinid *C.*
112 *carpio* and *Danio rerio* (Additional file 1: Figure S4). Using a comprehensive annotation strategy
113 combining RNA-seq derived transcript evidence, *de-novo* gene prediction and sequence similarity
114 to proteins from five further fish species, we annotated a total of 23,696 protein-coding genes
115 (Additional file 1: Table S6). Of the predicted genes, 99.44% (23,563 genes) are annotated by

116 functional database. In addition, we identified 1,796 non-coding RNAs including 474 miRNAs,
117 220 rRNA, 530 tRNAs, and 572 snRNAs. Transposable elements (TEs) comprise approximately
118 34.18% (381.3 Mb) of the *M. amblycephala* genome (Additional file 1: Table S7). DNA
119 transposons (23.80%) and long terminal repeat retrotransposons (LTRs) (9.89%) are the most
120 abundant TEs in *M. amblycephala*. The proportion of LTRs in *M. amblycephala* is highest in
121 comparison with other teleosts: *G. morhua* (4.88%) [2], *L. crocea* (2.2%) [5], *C. semilaevis*
122 (0.08%) [6], *C. carpio* (2.28%) [12], *C. idellus* (2.58%) [13] and stickleback (*Gasterosteus*
123 *aculeatus*) (1.9%) [18] (Additional file 1: Tables S7 and S8, Figure S5). The distribution of
124 divergence between the TEs in *M. amblycephala* peaks at 7% (Additional file 1: Figure S6),
125 indicating a more recent activity of these TEs when compared with *O. mykiss* (13%) [3] and *C.*
126 *semilaevis* (9%) [6].

127 **Anchoring Scaffolds and Shared Synteny Analysis**

128 Sequencing data from 198 F1 specimens, including the parents, were used as the mapping
129 population to anchor the scaffolds on to 24 **pseudo-chromosomes** of the *M. amblycephala* genome.
130 Following RAD-Seq and sequencing protocol, 1883.5 Mb of 125-bp reads (on average 30.6 Mb
131 and 9.3 Mb of read data for each parent and progeny, respectively) were generated on the HiSeq
132 2500 next-generation sequencing platform. Based on the SOAP bioinformatic pipeline, we
133 generated 5,317 SNP markers for constructing a high-resolution genetic map. The map spans
134 1,701 cM with a mean marker distance of 0.33 cM and facilitated an anchoring of 1,434 scaffolds
135 comprising 70% (779.54 Mb) of the *M. amblycephala* genome assembly to form 24 linkage
136 groups (LG) (Additional file 1: Table S9). Of the anchored scaffolds, 598 could additionally be
137 oriented (678.27 Mb, 87.01% of the total anchored sequences) (Figure 1A). A subsequent
138 comparison of the gene order between *M. amblycephala* and its close relative *C. idellus* revealed
139 607 large shared syntenic blocks encompassing 11,259 genes, and 190 chromosomal
140 rearrangements. The values change to 1,062 regions, 13,152 genes and 279 rearrangements when
141 considering *D. rerio*. The unexpected higher number of genes in syntenic regions shared with the
142 more distantly related *D. rerio* is most likely an effect of the more complete genome assembly of
143 this species compared to *C. idellus*. The rearrangement events are distributed across all *M.*
144 *amblycephala* linkage groups without evidence for a local clustering (Figure 1B). The most

145 prominent event is a chromosomal fusion in *M. amblycephala* LG02 that joined two *D. rerio*
146 chromosomes, Dre10 and Dre22. The same fusion is observed in *C. idellus* but not in *C. carpio*
147 suggesting that it probably occurred in a last common ancestor of *M. amblycephala* and *C. idellus*,
148 approximately 13.1 million years ago (Additional file 1: Figure S7).

149 **Results**

150 **Evolutionary Analysis**

151 A phylogenetic analysis of 316 single-copy genes with one to one orthologs in the genomes of 10
152 other fish species, and coelacanth (*Latimeria chalumnae*) and elephant shark (*Callorhynchus milii*),
153 as out group served as a basis for investigating the evolutionary trajectory of *M. amblycephala*
154 (Figure 2A, Additional file 1: Figure S8). To illuminate the evolutionary process resulting in the
155 adaptation to a grass diet, we analyzed the **functional categories of expanded genes** in the *M.*
156 *amblycephala* and *C. idellus* lineage (Additional file 1: Figure S9, Additional file 2: Data Note1),
157 two typical herbivores mainly feeding on aquatic and terrestrial grasses. Among the significantly
158 over-represented KEGG pathways (Fisher's exact test, $P < 0.01$), we find olfactory transduction
159 (ko04740), immune-related pathways (ko04090, ko04672, ko04612 and ko04621), lipid metabolic
160 related process (ko00590, ko03320, ko00591, ko00565, ko00592 and ko04975), as well as
161 xenobiotics biodegradation and metabolism (ko00625 and ko00363) (Figure S10). Indeed, when
162 tracing positively selected genes (PSG) in *M. amblycephala* and *C. idellus* (Additional file 3: Date
163 Note2), **we identified 10 candidates involved in starch and sucrose metabolism (ko00500), in**
164 **citrate cycle (ko00020) and in other types of O-glycan biosynthesis (ko00514). Moreover, 10**
165 **genes encoding enzymes involved in lipid metabolism appear positively selected in both fish**
166 **species (Additional file 1: Table S10).**

167 **Development of Intermuscular Bones**

168 To explain the genetic basis of IB, their formation and their function in cyprinids, we first
169 analyzed the functional annotation of **genes that expanded** in this lineage (Figure 2C). Many of
170 **these genes are involved in cell adhesion (GO: 0007155, $P=5.26E-32$, 357 genes), myosin**
171 **complex (GO:0016459, $P=2.74E-08$, 100 genes) and cell-matrix adhesion (GO:0007160,**
172 **$P=1.59E-21$, 69 genes) (Figure 2C). As a second line of evidence, we performed transcriptome**

173 analyses of early developmental stages (stage1: whole larvae without IBs) and juvenile *M.*
174 *amblycephala* (stage2: trunk muscle with partial IBs; stage3: trunk muscle with completed IBs)
175 (Figure 3A). Compared with stage1, 388 and 651 differentially expressed genes (DEGs) are
176 up-regulated in stage2 and stage3, respectively. And 249 of them are significantly up-regulated
177 both in stage2 and stage3. KEGG analyses indicate many of these genes involved in tight junction
178 (ko04530), regulation of actin cytoskeleton (ko04810), cardiac muscle contraction (ko04260) and
179 vascular smooth muscle contraction (ko04270) (Additional file 1: Figure S11). Specifically, 26
180 genes encoding proteins related to muscle contraction, including titin, troponin, myosin, actinin,
181 calmodulin and other Ca²⁺ transporting ATPases (Figure 3A) point to a strong remodeling of the
182 musculature compartment. To confirm that the observed differences in gene expression are indeed
183 linked to IB formation and function and are not simply due to the fact that different developmental
184 stages were compared, we performed differential expression analysis of muscle tissues, IB, and
185 connective tissues from the same six months old individual of *M. amblycephala* (Figure 3B,
186 Additional file 1: Figure S12). 1,290 DEGs and 5,231 DEGs are significantly up-regulated in IB
187 compared with connective tissues and muscle, respectively. 24 of these DEGs encode extracellular
188 matrix (ECM) proteins (collagens and integrin-binding protein), Rho GTPase family (*RhoA*, *Rho*
189 *GAP*, *Rac*, *Ras*), motor proteins (myosin, dynein, actin), and calcium channel regulation proteins
190 (Additional file 1: Figure S13 and Table S11). In addition, GO annotations of 963 IB-specific
191 genes indicative of abundance in protein binding (GO:0005515), calcium ion binding
192 (GO:0005509), GTP binding (GO:0005525) and iron ion binding (GO:0005506) were found
193 (Figure 3C).

194 During development of *M. amblycephala*, the first IB appears in muscles of caudal vertebrae
195 as early as 28 days post fertilization (dpf) when body length is 12.95 mm (Additional file 1: Figure
196 S14). The system then develops and ossifies predominantly from posterior to anterior (Additional
197 file 1: Figure S15). IBs are present throughout the body within two months (Additional file 1:
198 Figure S16) and develop into multiple morphological types in adults (Additional file 1: Figure
199 S17). The bone is formed directly without an intermediate cartilaginous stage (Additional file 1:
200 Figures S18 and S19). We also found a large number of mature osteoblasts distributed at the edge
201 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

202 matrix (Additional file 1: Figures S20 and S21). These primary bone-forming cells predominantly
203 regulate bone formation and function throughout life. Notably, among the genes up-regulated in
204 IB, 35 bone formation regulatory genes were identified (shown in colored boxes in Figure 3D). In
205 particular, genes involved in bone morphogenetic protein (BMP) signaling including *Bmp3*,
206 *Smad8*, *Smad9*, and *Id2*, in fibroblast growth factor (FGF) signaling including *Fgf2*, *Fgfr1a*,
207 *Fgfbp2*, *Col6a3*, and *Col4a5*, and in Ca²⁺ channels including *Cacna1c*, *CaM*, *Creb5* and *Nfatc*
208 were highly expressed (>2-fold change) in IB (Additional file 1: Figure S22).

209 **Adaptation to Herbivorous Diet**

210 Next to the presence of IB, the herbivory of *M. amblycephala* is the second key feature in
211 connection to the use of this species as aquatic livestock. Olfaction, the sense of smell, is crucial
212 for animals to find food. The perception of smell is mediated by a large gene family of olfactory
213 receptor (OR) genes. In the *M. amblycephala* genome, we identified 179 functional olfactory
214 receptor (OR) genes (Figure 4A), and based on the classification of Niimura [19], 158, 117 and
215 153 receptors for water-borne odorants were identified in *M. amblycephala*, *C. idellus* and *D.*
216 *rerio*, respectively (Additional file 1: Table S12). Overall, these receptor repertoires are
217 substantially larger than those of other and carnivorous teleosts (*G. morhua*, *C. semilaevis*, *O.*
218 *latipes*, *X. maculatus*) (Additional file 1: Figures S23 and S24, Table S12). In addition, we found
219 a massive expansion of beta-type OR genes in the genomes of the herbivorous *M. amblycephala*
220 and *C. idellus*, while very few exist in other teleosts (Figure 4B, Additional file 1: Table S12).

221 Taste is also an important factor in the development of dietary habits. Most animals can
222 perceive five basic tastes, namely sourness, sweetness, bitterness, saltiness and umami [20]. *T1R1*,
223 the receptor gene necessary for sensing umami, has been lost in herbivorous *M. amblycephala* but
224 is duplicated in carnivorous *G. morhua*, *C. semilaevis* and omnivorous *O. latipes* and *X. maculatus*
225 (Figures 4C and 4D, Additional file 1: Figures S25-26 and Table S13). In contrast, *T1R2*, the
226 receptor gene for sensing sweet, has been duplicated in herbivorous *M. amblycephala* and *C.*
227 *idellus*, omnivorous *C. carpio* and *D. rerio*, while it has been lost in carnivorous *G. morhua* and *C.*
228 *semilaevis* (Additional file 1: Figure S27 and Table S13). Also the *T2R* gene family, most likely
229 important in the course switching to a diet that contains a larger fraction of bitterness containing

230 food, has been expanded in *M. amblycephala*, *C. idellus*, *C. carpio* and *D. rerio* (Additional file 1:
231 Figure S28).

232 To obtain further insights into the genetic adaptation to herbivorous diet, we focused on
233 further genes that might be associated with digestion. Genes that encode proteases (including
234 pepsin, trypsin, cathepsin and chymotrypsin) and amylases (including alpha-amylase and
235 glucoamylase) were identified in the genomes of *M. amblycephala*, carnivorous *C. semilaevis*, *G.*
236 *morhua* and omnivorous *D. rerio*, *O. latipes* and *X. maculatus*, indicating that herbivorous *M.*
237 *amblycephala* has a protease repertoire that is not substantially different from those of carnivorous
238 and omnivorous fishes (Additional file 1: Table S14). We did not identify any genes encoding
239 potentially cellulose-degrading enzymes including endoglucanase, exoglucanase and
240 beta-glucosidase in the genome of *M. amblycephala*, suggesting that utilization of the herbivorous
241 diet may largely depend on the gut microbiome. To elucidate this further, we determined the
242 composition of the gut microbial communities of juvenile, domestic, wild adult *M. amblycephala*
243 and wild adult *C. idellus* using bacterial 16S rRNA sequencing. A total of 549,020 filtered high
244 quality sequence reads from 12 samples were clustered at a similarity level of 97%. The resulting
245 8,558 operational taxonomic units (OTUs) are dominated at phylum level by Proteobacteria,
246 Fusobacteria, Bacteroidetes, Firmicutes and Actinobacteria (Additional file 1: Table S15, Figure
247 4E). Increasing the resolution to the genus level, the composition and relative abundance of the
248 gut microbiota of wild adult *M. amblycephala* and *C. idellus* are still very similar (Additional file
249 1: Table S16) and we could identify more than 7% cellulose-degrading bacteria (Additional file 1:
250 Table S17).

251 Discussion

252 The evolutionary trajectory analysis of *M. amblycephala* and other teleosts revealed that *M.*
253 *amblycephala* has the closest relationship to *C. idellus*. Both the species are herbivorous fish but
254 which endogenous and exogenous factors affected their feeding habits and how they adapted to
255 their herbivorous diet is not known. Our results from the expanded genes and PSG in the lineage
256 of the two herbivores uncovered a number of genes that are involved in glucose, lipid and
257 xenobiotics metabolism, which would enhance the ability of an herbivore to detoxify the

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 secondary compounds present in grasses that are adverse or even toxic to the organism.
259 Furthermore, the high-fiber but low-energy grass diet requires a highly effective intermediate
260 metabolism that accelerates carbohydrate and lipid catabolism and conversion into energy to
261 maintain physiological functions.

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

272 It is known that the receptor for umami is formed by the T1R1/T1R3 heterodimer, while
273 T1R2/T1R3 senses sweet taste [24]. We found that the umami gene *TIR1* was lost in herbivorous
274 *M. amblycephala* but duplicated in the carnivorous *G. morhua* and *C. Semilaevis*. The loss of the
275 *TIR1* gene in *M. amblycephala* might exclude the expression of a functional umami taste receptor.
276 Such situations in other organism, e.g. the Chinese panda, have previously been related to feeding
277 specialization [25]. Bitterness sensed by the *T2R* is particularly crucial for animals to protect them
278 from poisonous compounds [22]. Interestingly, the bitter receptor *T2R* genes are expanded in the
279 herbivorous fish but few or no copy was found in carnivorous fish. These results not only indicate
280 the genetic adaptation to herbivorous diet of *M. amblycephala*, but also provided a clear and
281 comprehensive picture of adaptive evolutionary mechanisms of sensory systems in other fish
282 species with different trophic specializations.

283 It has been reported that some insects such as *Tenebrio molitor* [26] and *Neotermes*
284 *koshunensis* [27], and the mollusc *Corbicula japonica* [28] have genes encoding endogenous
285 cellulose degradation-related enzymes. However, all so far analyzed herbivorous vertebrates lack
286 these genes and always rely on their gut microbiome to digest food [25, 29]. In herbivorous *M.*

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

287 *amblycephala* and *C. idellus*, we also did not find any homologues of digestive cellulase genes.

288 Interestingly, our work on the composition of gut microbiota of the two fish species identifies
289 more than 7% cellulose-degrading bacteria, suggesting that the cellulose degradation of
290 herbivorous fish largely depend on their gut microbiome.

291 IB has evolved several times during teleost evolution [9, 30]. The developmental mechanisms
292 and ossification processes forming IB are dramatically distinct from other bones such as ribs,
293 skeleton, vertebrae or spines. These usually develop from cartilaginous bone and are derived from
294 the mesenchymal cell population by endochondral ossification [31, 32]. However, IB form directly
295 by intramembranous ossification and differentiate from osteoblasts within connective tissue,
296 forming segmental, serially homologous ossifications in the myosepta. Although various methods
297 of ossification of IB have been proposed, few experiments have been conducted to confirm the
298 ossification process and little is known about the potential role of IB in teleosts. **Based on our
299 findings of expanded genes in cyprinid lineage and evidence from transcriptome of developmental
300 stages of IB formation, a number of genes were found to interact dynamically to mediate efficient
301 cell motility, migration and muscle construction [33-36]. In addition, transcriptome analyses of
302 three tissues indicated that ECM, Rho GTPase, motor and calcium channel regulation protein
303 displayed high expression in IB. It is known that ECM proteins bound to integrins influence cell
304 migration by actomyosin-generated contractile forces [34, 37]. Rho GTPases, acting as molecular
305 switches, are also involved in regulating the actin cytoskeleton and cell migration, which in turn
306 initiates intracellular signaling and contributes to tissue repair and regeneration [38-40]. Thus, our
307 results provide molecular evidence that IB might play significant roles not only in regulating
308 muscle contraction but also in active remodeling at the bone-muscle interface and coordination of
309 cellular events.**

310 Some major developmental signals including BMP, FGF, WNT, together with
311 calcium/calmodulin signaling [31, 41-43], are essential for regulating the differentiation and
312 function of osteoblasts and osteocytes and for regulating the RANKL signaling pathway for
313 osteoclasts [44]. In agreement with this concept, we found 35 bone formation regulatory genes
314 involved in these signals were highly up-regulated in IB. **Among these signaling pathways, in
315 particular, *Bmp*, *Fgf2*, and *Fgfr1* are closely related to intramembranous bone development and**

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

316 affect the expression and activity of other osteogenesis related transcription factors [31, 45]. The
317 calcium-sensitive transcription factor *NFATc1* together with *CREB* induces the expression of
318 osteoclast-specific genes [46]. Taken together, these results suggest that IB indeed undergoes an
319 intramembranous ossification process, is regulated by bone-specific signaling pathways, and
320 underlies a homeostasis of maintenance, repair and remodeling.

321 **Conclusions**

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

329 **Methods**

330 **Sampling and DNA Extraction**

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

336 **Genomic Sequencing and Assembly**

337 Libraries with different insert sized inserts of 170 bp, 500 bp, 800 bp, 2 Kb, 5 Kb, 10 Kb and 20
338 Kb were constructed from the genomic DNA at BGI-Shenzhen. The libraries were sequenced
339 using a HiSeq2000 instrument. In total, 11 libraries, sequenced in 23 lanes were constructed. To
340 obtain high quality data, we applied filtering criteria for the raw reads. As a result, 142.55 Gb of
341 filtered data were used to complete the genome assembly using SOAPdenovo_V2.04 [16]. Only
342 filtered data were used in the genome assembly. First, the short insert size library data were used
343 to construct a de Bruijn graph. The tips, merged bubbles and connections with low coverage were

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

344 removed before resolving the small repeats. Second, all high-quality reads were realigned with the
345 contig sequences. The number of shared paired-end relationships between pairs of contigs was
346 calculated and weighted with the rate of consistent and conflicting paired ends before constructing
347 the scaffolds in a stepwise manner from the short-insert size paired ends to the long-insert size
348 paired ends. Third, the gaps between the constructed scaffolds were composed mainly of repeats,
349 which were masked during scaffold construction. These gaps were closed using the paired-end
350 information to retrieve read pairs in which one end mapped to a unique contig and the other was
351 located in the gap region. Subsequently, local assembly was conducted for these collected reads.
352 To assess the genome assembly quality, approximately 42.82 Gb Illumina reads generated from
353 short-insert size libraries were mapped onto the genome. Bwa0.5.9-r16 software [47] with default
354 parameters was used to assess the mapping ratio and Soap coverage 2.27 was used to calculate the
355 sequencing depth. We also assessed the accuracy of the genome assembly by Trinity [48],
356 including number of ESTs and new mRNA reads from early stages of embryos and multiple
357 tissues, by aligning the scaffolds to the assembled transcriptome sequences.

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

362 **Genome Annotation**

363 The genome was searched for repetitive elements using Tandem Repeats Finder (version 4.04)
364 [49]. TEs were identified using homology-based approaches. The Repbase (version 16.10) [50]
365 database of known repeats and a *de novo* repeat library generated by RepeatModeler were used.
366 This database was mapped using the software of RepeatMasker (version 3.3.0). Four types of
367 non-coding RNAs (microRNAs, transfer RNAs, ribosomal RNAs and small nuclear RNAs) were
368 also annotated using tRNAscan-SE (version 1.23) and the Rfam database45 (Release 9.1) [51].

369 For gene prediction, *de novo* gene prediction, homology-based methods and RNA-seq data
370 were used to perform gene prediction. For the sequence similarity based prediction, *D. rerio*, *G.*
371 *aculeatus*, *O. niloticus*, *O. latipes* and *G. morhua* protein sequences were downloaded from
372 Ensembl (release 73) and were aligned to the *M. amblycephala* genome using TBLASTN. Then

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

373 homologous genome sequences were aligned against the matching proteins using GeneWise [52]
374 to define gene models. Augustus was employed to predict coding genes using appropriate
375 parameters in *de novo* prediction. For the RNA-seq based prediction, we mapped transcriptome
376 reads to the genome assembly using TopHat [53]. Then, we combined TopHat mapping results
377 together and applied Cufflinks [54] to predict transcript structures. All predicted gene structures
378 were integrated by GLEAN [55] (<http://sourceforge.net/projects/glean-gene/>) to obtain a
379 consensus gene set. Gene functions were assigned to the translated protein-coding genes using
380 Blastp tool, based on their highest match to proteins in the SwissProt and TrEMBL [56] databases
381 (Uniprot release 2011-01). Motifs and domains in the protein-coding genes were determined by
382 InterProScan (version 4.7) searches against six different protein databases: ProDom, PRINTS,
383 Pfam, SMART, PANTHER and PROSITE. Gene Ontology [57] IDs for each gene were obtained
384 from the corresponding InterPro entries. All genes were aligned against KEGG [58] (Release 58)
385 database, and the pathway in which the gene might be involved was derived from the matched
386 genes in KEGG. tRNA genes were *de novo* predicted by tRNAscan-SE software [59], with
387 eukaryote parameters on the repeat pre-masked genome. The rRNA fragments were identified by
388 aligning the rRNA sequences using BlastN at E-value 1e-5. The snRNA and miRNA were
389 searched by the method of aligning and searching with INFERNAL (version 0.81) [60] against
390 Rfam database (release 9.1).

391 **Genetic Map Construction**

392 To anchor the scaffolds into pseudo-chromosomes, 198 F1 population individuals were used to
393 obtain the genetic map. Each of the individual genomic DNA was digested with the restriction
394 endonuclease EcoR I, following the RAD-Seq protocol [61]. The SNP calling process was carried
395 out using the SOAP bioinformatic pipeline. The RAD-based SNP calling was done by SOAPSnp
396 software [62] after each individual's paired-end RAD reads was mapped onto the assembled
397 reference genome with the alignment software SOAP2 [63]. The potential SNP markers were used
398 for the linkage analysis if the following criteria were satisfied: for parents - sequencing depth ≥ 8
399 and ≤ 100 , base quality ≥ 25 , copy number ≤ 1.5 ; for progeny - sequencing depth ≥ 5 , base quality
400 ≥ 20 , copy number ≤ 1.5 . If the markers were showing significantly distorted segregation (*P*-value
401 < 0.01), they were excluded from the map construction. Linkage analysis was performed only for

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

402 markers present in at least 80% of the genomes, using JoinMap 4.0 software with CP population
403 type codes and applying the double pseudo-test cross strategy [64]. The linkage groups were
404 formed at a logarithm of odds threshold of 6.0 and ordered using the regression mapping
405 algorithm.

406 **Construction of Gene Families**

407 We identified gene families using TreeFam software [65] as follows: Blast was used to compare
408 all the protein sequences from 13 species: *M. amblycephala*, *C. idellus*, *C. semilaevis*, *C. carpio*,
409 *D. rerio*, *Callorhinchus milii*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *Oncorhynchus*
410 *mykiss*, *O. niloticus*, *O. latipes*, *Fugu rubripes*, with the E-value threshold set as 1e-7. In the next
411 step, HSP segments of each protein pair were concatenated by Solar software. H-scores were
412 computed based on Bit-scores and these were taken to evaluate the similarity among genes.
413 Finally, gene families were obtained by clustering of homologous gene sequences using
414 Hcluster_sg (Version 0.5.0). Specific genes of *M. amblycephala* were those that did not cluster
415 with other vertebrates that were chosen for gene family construction, and those that did not have
416 homologs in the predicted gene repertoire of the compared genomes. If these genes had functional
417 motifs, they were annotated by GO.

418 **Phylogenetic Tree Reconstruction and Divergence Time Estimation**

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

430 **Gene Family Expansion and Contraction Analyses**

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

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

439 **Detection of Positively Selected Genes**

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

452 **Developmental Process of Intermuscular Bone in *M. amblycephala***

453 To better understand the number and morphological types of IBs in adult *M. amblycephala*,
454 specimens with a body length ranging from 15.5 to 20.5 cm were collected and each individual
455 was wrapped in gauze and boiled. The fish body was divided into two sections: anterior (snout to
456 cloaca) and posterior (cloaca to the base of caudal fin), and the length of each section was
457 measured. The IBs were retrieved, counted, arranged in order and photographed with a digital
458 camera. Fertilized *M. amblycephala* eggs were brought from hatching facilities at Freshwater Fish
459 Genetics Breeding Center of Huazhong Agricultural University (Wuhan, Hubei, China) to our

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
460 laboratory. *M. amblycephala* larvae were maintained in a re-circulating aquaculture system at 23 ±
461 1°C with a 14-hr photoperiod. To explore the early development of IBs, larvae at different stages
462 from 15 to 40 dpf were collected and fixed in 4% paraformaldehyde and transferred to 70%
463 ethanol for storage. Specimens were stained with alizarin red for bone following the method
464 described by Dawson [74]. The appearance of red color was recorded as the appearance of IB
465 because bone ossification is accompanied by the uptake of alizarin red, resulting in red staining of
466 the mineralized bone matrix. Myosepta, either not yet ossified, or poorly ossified, are not visible
467 with alizarin red staining. For histologic analysis, specimens were paraffin-embedded and
468 sectioned following standard protocols. Sections were stained with hematoxylin and eosin (HE)
469 and Masson trichrome [75] and photographed using a Nikon microscope (Nikon, Tokyo, Japan)
470 with a DP70 digital camera (Olympus, Japan). Scanning electron microscopy (SEM) and
471 transmission electron microscopy (TEM) were also conducted to analyze the ultrastructure of IB.
472 The specimens were fixed with 2.5% (v/v) glutaraldehyde in a solution of 0.1 M sodium
473 cacodylate buffer (pH 7.3) for 2 h at room temperature. The SEM and TEM samples were
474 prepared according to a standard protocol described by Ott [76]. The samples were then visualized
475 with a JSM-6390LV scanning electron microscope (SEM, Japan) and the stained ultrathin sections
476 with a H-7650 transmission electron microscope (Hitachi, Japan).

477 **RNA Sequencing Analysis**

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
478 *M. amblycephala* specimens belonging to three different developmental stages of IBs (stage 1: whole
479 larvae without distribution of IB; stage 2: muscle tissues with partial distribution of IBs; stage 3:
480 muscle tissues with completed distribution of IBs) were identified under microscope and immediately
481 frozen in liquid nitrogen. In addition, dorsal white muscle, IBs and connective tissue surrounding the
482 IBs from six months old fish were also collected. RNA was extracted from total fish samples at
483 different stages and from individual muscle, connective tissue, and intermuscular bone samples of
484 *M. amblycephala* using RNAisoPlus Reagent (TaKaRa, China) according to the manufacturer's
485 protocol. The integrity and purity of the RNA was determined by gel electrophoresis and Agilent
486 2100 BioAnalyzer (Agilent, USA) before preparing the libraries for sequencing. Paired-end RNA
487 sequencing was performed using the Illumina HiSeq 2000 platform. Low quality score reads were
488 filtered and the clean data were aligned to the reference genome using Bowtie [77]. Genes and

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

489 isoforms expression level were quantified by a software package: RSEM (RNASeq by
490 Expectation Maximization) [78]. Gene expression levels were calculated by using the RPKM
491 method (Reads per kilobase transcriptome per million mapped reads) [79] and adjusted by a
492 scaling normalization method [80]. We detected DEGs from three stages of IBs with software
493 NOIseq and three different tissues with PossionDis as requested. NOIseq is based on noisy
494 distribution model, performed as described by Tarazona [81]. The parameters were set as: fold
495 change ≥ 2.00 and probability ≥ 0.7 . PossionDis is based on the Poisson distribution, performed
496 as described by Audic [82]. The parameters were set as: fold change ≥ 2.00 and FDR ≤ 0.001 .
497 Annotation of DEGs were mapped to GO categories in the database
498 (<http://www.geneontology.org/>) and the number of genes for every term were calculated to
499 identify GO terms that were significantly enriched in the input list of DEGs. The calculated
500 *P*-value was adjusted by the Bonferroni Correction, taking corrected *P*-value ≤ 0.05 as a threshold.
501 KEGG automatic annotation was used to perform pathway enrichment analysis of DEGs.

502 **Prediction of Olfactory Receptor Genes**

503 Olfactory receptor genes were identified by previously described methods [83], with the exception
504 of a first-round TBLASTN [84] search, in which 1,417 functional olfactory receptor genes from *H.*
505 *sapiens*, *D. rerio*, *L. chalumnae*, *Lepisosteus oculatus*, *L. vexillifer*, *O. niloticus*, *O. latipes*, *F.*
506 *rubripes* and *Xenopus tropicalis* were used as queries. We then predicted the structure of
507 sequenced genes using the blast-hit sequence with the software GeneWise extending in both 3' and
508 5' directions along the genome sequences. The results were further confirmed by NR annotation.
509 Then the coding sequences from the start (ATG) to stop codons were extracted, while sequences
510 that contained interrupting stop codons or frame-shifts were regarded as pseudogenes. To
511 construct phylogenetic trees, the amino-acid sequences encoded by olfactory receptor genes were
512 first aligned using the program MUSCLE nested in MEGA 5.10 [85]. We then constructed the
513 phylogenetic tree using the neighbor-joining method with Poisson correction distances using the
514 program MEGA 5.10. We also identified and compared the genes for five basic tastes (sour, sweet,
515 bitter, umami and salty) using a similar method as in OR gene identification.

516 **Gut Microbiota Analysis**

517 To characterize the microbial diversity of herbivorous *M. amblycephala*, a total of 12 juvenile

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

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

530 **Additional files**

531 Additional file 1: Tables S1 to S17 and Figures S1 to S28.

532 Additional file 2: Data Note1 Expanded genes in the *M. amblycephala* and *C. idellus* lineage.

533 Additional file 3: Data Note2 Positively selected genes in the *M. amblycephala* and *C. idellus*
534 genomes.

535 **Abbreviations**

536 IB, intermuscular bone; SNP, single-nucleotide polymorphism; BUSCO, benchmarking universal
537 single-copy orthologs; TE, transposable element; LTR, long terminal repeat retrotransposon; LG,
538 linkage group; PSG, positively selected gene; ECM, extracellular matrix; dpf, days post
539 fertilization; BMP, bone morphogenetic protein; FGF, fibroblast growth factor; OR, olfactory
540 receptor; OTU, operational taxonomic unit; DEGs, differentially expressed genes; HE,
541 hematoxylin and eosin; SEM, scanning electron microscopy; TEM, transmission electron
542 microscopy

543 **Acknowledgements**

544 This work was supported by the Fundament Research Funds for the Central Universities
545 (2662015PY019), the Modern Agriculture Industry Technology System Construction Projects of

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

546 China titled as—Staple Freshwater Fishes Industry Technology System (No. CARS-46-05),
547 Guangdong Haid Group Co., Ltd and the International Scientific and Technology Cooperation
548 Program of Wuhan City (2015030809020365).

549 **Availability of data and materials**

550 Datasets supporting the results of this article are available in the GigaDB repository associated
551 with this publication [88]. Raw whole genome sequencing, transcriptome and RAD-Seq data have
552 been deposited at NCBI in the SRA under [BioProject](#) number PRJNA343584.

553 **Authors' contributions**

554 W.W. initiated and conceived the project and provided scientific input. X.Q. organized financial
555 support and designed the project. M.S. discussed the data, wrote and modified the paper. H.L. and
556 C.C. conducted the biological experiments, analyzed the data and wrote the paper with input from
557 other authors. I.E. wrote and modified the manuscript and discussed the data. The RAD-Seq data
558 analyses and the genetic map construction were performed by Z.G., Y.G., J.J. and X.J. Genome
559 assembly and annotation were performed by J.M., H.C., M.X. and J.C. X.Z., W.L., R.L., B.C.,
560 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 preparation
561 and data collection. J.L. and J.C. identified the gene families and analyzed the RNA-seq data. M.B.
562 coordinated the project. S.Z. and X.F. modified the manuscript and discussed the data. All authors
563 read the manuscript and provided comments and suggestions for improvements. The authors
564 declare no competing financial interests.

565 **Competing interests**

566 The authors declare that they have no competing interests.

567 **Ethics approval and consent to participate**

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

571 **References**

- 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
- 572 1. FAO Fisheries and Aquaculture Department. FAO yearbook Fishery and Aquaculture Statistics
573 2014 (Food and Agriculture Organization of the United Nations, Rome, 2016).
 - 574 2. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, et al. The genome
575 sequence of Atlantic cod reveals a unique immune system. *Nature*. 2011; 477:207–10.
 - 576 3. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al. The rainbow trout
577 genome provides novel insights into evolution after whole-genome duplication in vertebrates.
578 *Nat. Commun.* 2014; 5:3657.
 - 579 4. Tine M., Kuhl H., Gagnaire PA, Louro B, Desmarais E, Martins RS, et al. European sea bass
580 genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat.*
581 *Commun.* 2014; 5:5770.
 - 582 5. Wu C, Zhang D, Kan M, Lv Z, Zhu A, Su Y, et al. The draft genome of the large yellow
583 croaker reveals well-developed innate immunity. *Nat. Commun.* 2014; 5:5227.
 - 584 6. Chen S, Zhang G, Shao C, Huang Q, Liu G, Zhang P, et al. Whole-genome sequence of a
585 flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic
586 lifestyle. *Nat. Genet.* 2014; 46:253–60.
 - 587 7. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for
588 adaptive radiation in African cichlid fish. *Nature*. 2014; 513:375–82.
 - 589 8. Chen X, Zhong L, Bian C, Xu P, Qiu Y, You X, et al. High-quality genome assembly of
590 channel catfish, *Ictalurus punctatus*. *GigaScience*. 2016; 5:39.
 - 591 9. Gemballa S, Britz R. Homology of intermuscular bones in acanthomorph fishes. *Am. Mus.*
592 *Novit.* 1998; 3241:1–25.
 - 593 10. Danos N, Ward AB. The homology and origins of intermuscular bones in fishes: Phylogenetic
594 or biomechanical determinants? *Biol. J. Linn. Soc.* 2012; 106:607–22.
 - 595 11. Wan SM, Yi SK, Zhong J, Nie CH, Guan NN, Zhang WZ, et al. Dynamic mRNA and miRNA
596 expression analysis in response to intermuscular bone development of blunt snout bream
597 (*Megalobrama amblycephala*). *Sci. Rep.* 2016; 6:31050.
 - 598 12. Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, et al. Genome sequence and genetic diversity
599 of the common carp, *Cyprinus carpio*. *Nat. Genet.* 2014; 46:1212–9.
 - 600 13. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, et al. The draft genome of the grass carp
601 (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation.
602 *Nat. Genet.* 2015; 47:625–31.
 - 603 14. Gao Z, Luo W, Liu H, Zeng C, Liu X, Yi S, et al. Transcriptome analysis and SSR/SNP
604 markers information of the blunt snout bream (*Megalobrama amblycephala*). *PLoS One.*
605 2012; 7:e42637.

- 606 15. Yi S, Gao ZX, Zhao H, Zeng C, Luo W, Chen B, et al. Identification and characterization of
607 microRNAs involved in growth of blunt snout bream (*Megalobrama amblycephala*) by
608 Solexa sequencing. BMC Genomics. 2013; 14:754.
- 609 16. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved
610 memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.
- 611 17. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing
612 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.
613 2015;31:3210–2.
- 614 18. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis
615 of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61.
- 616 19. Niimura Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative
617 genome analysis among 23 chordate species. Genome Biol. Evol. 2009; 1:34–44.
- 618 20. Lindemann B. Receptors and transduction in taste. Nature. 2001; 413:219–25.
- 619 21. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene
620 repertoires: roles of chance and necessity. Nat. Rev. Genet. 2008; 9:951–63.
- 621 22. Chandrashekar J, Mueller KL, Hoon MA, Adler E, Feng L, Guo W, et al. T2Rs function as
622 bitter taste receptors. Cell. 2000; 100:703–11.
- 623 23. Niimura Y, Nei M. Evolutionary dynamics of olfactory and other chemosensory receptor
624 genes in vertebrates. J. Hum. Genet. 2006; 51:505–17.
- 625 24. Nelson G, Chandrashekar J, Hoon MA, Feng L, Zhao G, Ryba NJ, et al. An amino-acid taste
626 receptor. Nature. 2002; 416:199–202.
- 627 25. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and *de novo* assembly of the
628 giant panda genome. Nature. 2010; 463:311–7.
- 629 26. Ferreira AHP, Marana SR, Terra WR, Ferreira C. Purification, molecular cloning, and
630 properties of a β -glycosidase isolated from midgut lumen of *Tenebrio molitor* (Coleoptera)
631 larvae. Insect Biochem. Mol. Biol. 2001; 31:1065–76.
- 632 27. Tokuda G, Saito H, Watanabe H. A digestive β -glucosidase from the salivary glands of the
633 termite, *Neotermes koshunensis* (Shiraki): Distribution, characterization and isolation of its
634 precursor cDNA by 5'- and 3'-RACE amplifications with degenerate primers. Insect
635 Biochem. Mol. Biol. 2002; 32:1681–9.
- 636 28. Sakamoto K, Uji S, Kurokawa T, Toyohara H. Molecular cloning of endogenous
637 β -glucosidase from common Japanese brackish water clam *Corbicula japonica*. Gene. 2009;
638 435:72–9.
- 639 29. Zhu L, Wu Q, Dai J, Zhang S, Wei F. Evidence of cellulose metabolism by the giant panda gut

- 640 microbiome. Proc. Natl. Acad. Sci. USA. 2011; 108:17714–9.
- 641 30. Bird NC, Mabee PM. Developmental morphology of the axial skeleton of the zebrafish, *Danio*
642 *rerio* (Ostariophysi: Cyprinidae). Dev. Dyn. 2003; 228:337–57.
- 643 31. Ornitz D, Marie P. FGF signaling pathways in endochondral and intramembranous bone
644 development and human genetic disease. Genes Dev. 2002; 16:1446–65.
- 645 32. Ortega N, Behonick DJ, Werb Z. Matrix remodeling during endochondral ossification. Trends
646 Cell Biol. 2004; 14:86–93.
- 647 33. Even-Ram S, Doyle AD, Conti MA, Matsumoto K, Adelstein RS, Yamada KM. Myosin IIA
648 regulates cell motility and actomyosin-microtubule crosstalk. Nat. Cell Biol. 2007;
649 9:299–309.
- 650 34. Sheetz MP, Felsenfeld DP, Galbraith CG. Cell migration: Regulation of force on
651 extracellular-complexes. Trends Cell Biol. 1998; 8:51–4.
- 652 35. Gunst SJ, Zhang W. Actin cytoskeletal dynamics in smooth muscle: a new paradigm for the
653 regulation of smooth muscle contraction. Am. J. Physiol. Cell Physiol. 2008; 295:C576–87.
- 654 36. Webb RC. Smooth muscle contraction and relaxation. Adv. Physiol. Educ. 2003; 27:201–6.
- 655 37. Ulrich TA, De Juan Pardo EM, Kumar S. The mechanical rigidity of the extracellular matrix
656 regulates the structure, motility, and proliferation of glioma cells. Cancer Res. 2009;
657 69:4167–74.
- 658 38. Ridley AJ. Rho GTPases and cell migration. J. Cell Sci. 2001; 114:2713–22.
- 659 39. Etienne-Manneville S, Hall A. Rho GTPases in Cell Biology. Nature. 2002; 420:629–35.
- 660 40. Ridley AJ. Cell Migration: Integrating Signals from Front to Back. Science. 2003;
661 302:1704–9.
- 662 41. Chen G, Deng C, Li YP. TGF- β and BMP signaling in osteoblast differentiation and bone
663 formation. Int. J. Biol. Sci. 2012; 8:272–88.
- 664 42. Harada S, Rodan GA. Control of osteoblast function and regulation of bone mass. Nature.
665 2003; 423:349–55.
- 666 43. Long F. Building strong bones: molecular regulation of the osteoblast lineage. Nat. Rev. Mol.
667 Cell Biol. 2011; 13:27–38.
- 668 44. Boyle WJ, Simonet WS, Lacey DL. Osteoclast differentiation and activation. Nature. 2003;
669 423:337–42.
- 670 45. Fakhry A, Ratisoontorn C, Vedhachalam C, Salhab I, Koyama E, Leboy P, et al. Effects of
671 FGF-2/-9 in calvarial bone cell cultures: Differentiation stage-dependent mitogenic effect,
672 inverse regulation of BMP-2 and noggin, and enhancement of osteogenic potential. Bone.

673 2005; 36:254–66.

674 46. Sato K, Suematsu A, Nakashima T, Takemoto-Kimura S, Aoki K, Morishita Y, et al.
675 Regulation of osteoclast differentiation and function by the CaMK-CREB pathway. *Nat. Med.*
676 2006; 12:1410–6.

677 47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
678 *Bioinformatics.* 2009; 25:1754–60.

679 48. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity:
680 reconstructing a full-length transcriptome without a genome from RNA-Seq data . *Nat.*
681 *Biotechnol.* 2011; 29:644–52.

682 49. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.*
683 1999; 27:573–80.

684 50. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase
685 Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 2005;
686 110:462–7.

687 51. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: Annotating
688 non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005; 33:121–4.

689 52. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res.* 2004; 14:988–95.

690 53. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq.
691 *Bioinformatics.* 2009; 25:1105–11.

692 54. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript
693 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform
694 switching during cell differentiation. *Nat. Biotechnol.* 2010; 28:511–5.

695 55. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey
696 bee consensus gene set. *Genome Biol.* 2007; 8:R13.

697 56. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement
698 TrEMBL in 2000. *Nucleic Acids Res.* 2000; 28:45–8.

699 57. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool
700 for the unification of biology. *Nat. Genet.* 2000; 25:25–9.

701 58. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*
702 2000; 28:27-30.

703 59. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes
704 in genomic sequence. *Nucleic Acids Res.* 1997; 25:955–64.

- 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
- 705 60. Nawrocki EP, Kolbe DL, Eddy SR. *Infernal 1.0: Inference of RNA alignments*. *Bioinformatics*.
706 2009; 25:1335–7.
- 707 61. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP
708 discovery and genetic mapping using sequenced RAD markers. *PloS One*. 2008; 3:e3376.
- 709 62. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively
710 parallel whole-genome resequencing. *Genome Res*. 2009; 19:1124–32.
- 711 63. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: An improved ultrafast tool
712 for short read alignment. *Bioinformatics*. 2009; 25:1966–7.
- 713 64. Grattapaglia D, Sederoff R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus*
714 *urophylla* using a pseudo-testcross: Mapping strategy and RAPD markers. *Genetics*. 1994;
715 137:1121–37.
- 716 65. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, et al. TreeFam: a curated
717 database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 2006;
718 34:D572–80.
- 719 66. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.
720 *Nucleic Acids Res*. 2004; 32:1792–7.
- 721 67. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and
722 methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML
723 3.0. *Syst. Biol*. 2010; 59:307–21.
- 724 68. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by
725 maximum likelihood. *Syst. Biol*. 2003; 52:696–704.
- 726 69. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. 2007;
727 24:1586–91.
- 728 70. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock
729 using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol*. 2006; 23:212–26.
- 730 71. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics*.
731 2007; 177:1941–9.
- 732 72. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with
733 insertions. *Proc. Natl. Acad. Sci. USA*. 2005; 102:10557–62.
- 734 73. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and
735 ambiguously aligned blocks from protein sequence alignments. *Syst. Biol*. 2007; 56:564–77.
- 736 74. Dawson AB. A note on the staining of the skeleton of cleared specimens with alizarin red S.
737 *Biotech. Histochem*. 1926; 1:123-4.

- 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
- 738 75. Gruber HE. Adaptations of Goldner's Masson trichrome stain for the study of undecalcified
739 plastic embedded bone. *Biotech. Histochem.* 1992; 67:30–4.
- 740 76. Ott HC, Matthiesen TS, Goh SK, Black LD, Kren SM, Netoff TI, et al.
741 Perfusion-decellularized matrix: using nature's platform to engineer a bioartificial heart. *Nat.*
742 *Med.* 2008; 14:213–21.
- 743 77. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 2012;
744 9:357–9.
- 745 78. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or
746 without a reference genome. *BMC Bioinformatics.* 2011; 12:323.
- 747 79. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying
748 mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 2008; 5:621–8.
- 749 80. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis
750 of RNA-seq data. *Genome Biol.* 2010; 11:R25.
- 751 81. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in
752 RNA-seq: a matter of depth. *Genome Res.* 2011; 21:2213–23.
- 753 82. Audic S, Claverie JM. The significance of digital gene expression profiles. *Genome Res.* 1997;
754 7:986–95.
- 755 83. Niimura Y. Evolutionary dynamics of olfactory receptor genes in chordates: interaction
756 between environments and genomic contents. *Hum. Genomics.* 2009; 4:107–18.
- 757 84. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
758 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;
759 25:3389–402.
- 760 85. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary
761 analysis of DNA and protein sequences. *Brief. Bioinform.* 2008; 9:299–306.
- 762 86. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al.
763 Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc.*
764 *Natl. Acad. Sci. USA.* 2011; 108:4516–22.
- 765 87. Edgar, RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat.*
766 *Methods.* 2013; 10:996–8.
- 767 88. Liu H, Chen CH, Gao ZX, Min JM, Gu YM, Jian JB, et al. The draft genome of *Megalobrama*
768 *amblycephala* reveals the development of intermuscular bone and adaptation to herbivorous
769 diet. 2016. GigaScience Database.

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

770 **Figure Legends**

771 **Figure 1** Global view of the *M. amblycephala* genome and syntenic relationship between *C.*
772 *idellus*, *M. amblycephala* and *D. rerio*. (A) Global view of the *M. amblycephala* genome. From
773 outside to inside, the genetic linkage map (a); Anchors between the genetic markers and the
774 assembled scaffolds (b); Assembled chromosomes (c); GC content within a 50-kb sliding window
775 (d); Repeat content within a 500-kb sliding window (e); Gene distribution on each chromosome (f);
776 Different gene expression of three transcriptomes (g). (B) Syntenic relationship between *C. idellus*
777 (a), *M. amblycephala* (b) and *D. rerio* (c) chromosomes.

778 **Figure 2** Phylogenetic tree and comparison of orthologous genes in *M. amblycephala* and other
779 fish species. (A) Phylogenetic tree of teleosts using 316 single copy orthologous genes. The color
780 circles at the nodes shows the estimated divergence times using *O. latipes*–*F. rubripes*
781 [96.9~150.9Mya], *F. rubripes*–*D. rerio* [149.85~165.2Mya], *F. rubripes*–*C. milii* [416~421.75Mya]
782 (<http://www.timetree.org/>) as the calibration time. Pentagram represents four cyprinid fish with
783 intermuscular bones. S, silurian period; D, devonian period; C, carboniferous period; P, permian
784 period in Paleozoic; T, triassic period; J, jurassic and k-cretaceous period in Mesozoic; Pg,
785 paleogene in Cenozoic Era, N, Neogene. (B) Venn diagram of shared and unique orthologous gene
786 families in *M. amblycephala* and four other teleosts. (C) Over-represented GO annotations of
787 cyprinid-specific **expansion genes**.

788 **Figure 3** Regulation of genes related to intermuscular bone formation and function identified from
789 developmental stages and adult tissues transcriptome data. (A) Gene expression pattern involved
790 in muscle contraction regulated genes in early developmental stages corresponds to intermuscular
791 bone formation of *M. amblycephala*, (alizarin red staining). M, myosepta; IB, intermuscular bone.
792 (B) Scanning electron microscope photos of muscle tissues, connective tissues, and intermuscular
793 bone. (C) Distribution of intermuscular bone specific genes in GO annotations indicative of
794 abundance in protein binding, calcium ion binding, GTP binding functions. (D) Several
795 developmental signals regulating key steps of osteoblast and osteoclast differentiation in the
796 process of intramembranous ossification. Colored boxes indicate significantly up-regulated genes
797 in these signals specifically occurred in intermuscular bone.

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

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823 **Table**

824 **Table 1 Features of the *M. 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%)

825

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

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^{1†}, Chunhai Chen^{2†}, Zexia Gao^{1†}, Jiumeng Min^{2†}, Yongming Gu^{3†}, Jianbo Jian^{2†}, Xiewu
7
8 4 Jiang³, Huimin Cai², Ingo Ebersberger⁴, Meng Xu², Xinhui Zhang¹, Jianwei Chen², Wei Luo¹,
9
10 5 Boxiang Chen^{1,3}, Junhui Chen², Hong Liu¹, Jiang Li², Ruifang Lai¹, Mingzhou Bai², Jin Wei¹,
11
12 6 Shaokui Yi¹, Huanling Wang¹, Xiaojuan Cao¹, Xiaoyun Zhou¹, Yuhua Zhao¹, Kaijian Wei¹,
13
14 7 Ruibin Yang¹, Bingnan Liu³, Shancen Zhao², Xiaodong Fang², Manfred Schartl^{5,*}, Xueqiao
15
16 8 Qian^{3,*}, Weimin Wang^{1,*}

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

24
25 12 †Equal contributors

26
27 13 ¹College 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 ²Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen 518083, China

34
35 17 ³Guangdong Haid Group Co., Ltd., Guangzhou 511400, China

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

40
41 20 ⁵Physiological Chemistry, University of Würzburg, Biozentrum, Am Hubland, and Comprehensive
42
43 21 Cancer Center Mainfranken, University Clinic Würzburg, Würzburg 97070, Germany and Texas
44
45 22 A&M Institute for Advanced Study and Department of Biology, Texas A&M University, College
46
47 23 Station, TX 77843, USA

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 be involved 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 contributes 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. Atlantic cod (*Gadus morhua*) [2], rainbow trout
65 (*Oncorhynchus mykiss*) [3], European sea bass (*Dicentrarchus labrax*) [4], yellow croaker
66 (*Larimichthys crocea*) [5], half-smooth tongue sole (*Cynoglossus semilaevis*) [6], tilapia
67 (*Oreochromis niloticus*) [7] and channel catfish (*Ictalurus punctatus*) [8] have been sequenced.
68 Yet, the majority of these species are carnivorous requiring large inputs of protein from wild
69 caught fish or other precious feed. Reports on draft genomes of herbivorous and omnivorous
70 species, in particular cyprinid fish are scarce. It is well known that cyprinids are currently the
71 economically most important group of teleosts for sustainable aquaculture. They grow to large
72 population sizes in the wild and already now account for the majority of freshwater aquaculture
73 production worldwide [1]. Among these, the herbivorous *Megalobrama amblycephala* (Yih, 1955),
74 a particularly eco-friendly and resource-conserving species, is predominant in aquaculture and has
75 been greatly developed in China (Additional file 1: Figure S1) [1]. However, most cyprinids,
76 including *M. amblycephala*, have a large number of intermuscular bones (IBs) in the trunk
77 musculature, which have an adverse effect on fish meat processing and consumption. IBs—a
78 unique form of bone occurring only in the more basal teleosts—are completely embedded within
79 the myosepta and are not connected to the vertebral column or any other bones [9, 10]. Our
80 previous study on IB development of *M. amblycephala* revealed that some miRNA-mRNA
81 interaction pairs may be involved in regulating bone development and differentiation [11].
82 However, the molecular genetic basis and the evolution of this unique structure are still unclear.
83 Unfortunately, the recent sequencing of two cyprinid genomes common carp (*Cyprinus carpio*)
84 [12] and grass carp (*Ctenopharyngodon idellus*) [13], which provided valuable information for
85 their genetic breeding, contributed little to the understanding of IB formation.

86 In an initial genome survey of *M. amblycephala*, we identified 25,697 single-nucleotide

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 polymorphism (SNP) [14], 347 conserved miRNAs [15], and 1,136 miRNA-mRNA interaction
88 pairs [11]. However, lack of a whole genome sequence resource limited a thorough investigation
89 of *M. amblycephala*. Here we report the first high-quality draft genome sequence of *M.*
90 *amblycephala*. Integrating this novel genome resource with tissue- and developmental
91 stage-specific gene expression information, as well as with meta-genome data to investigate the
92 composition of the gut microbiome (Workflow shown in additional file 1: Figure S2) provides
93 relevant insights into the function and evolution of two key features characterizing this species:
94 The formation of IB and the adaptation to herbivory. By that our study lays the foundation for
95 genetically optimizing *M. amblycephala* to further increase its relevance for securing human food
96 supply.

97 **Data description**

98 **Genome Assembly and Annotation**

99 The *M. amblycephala* genome was sequenced and assembled by a whole-genome shotgun strategy
100 using genomic DNA from a double-haploid fish (Additional file 1: Table S1). We assembled a
101 1.116 Gb reference genome sequence from 142.55 Gb (approximately 130-fold coverage) of clean
102 data [16] (Additional file 1: Tables S1 and S2, Figure S3). The contig and scaffold N50 lengths
103 reached 49 Kb and 839 Kb, respectively (Table 1). The largest scaffold spans 8,951 Kb and the
104 4,034 largest scaffolds cover 90% of the assembly. To assess the genome assembly quality, the
105 mapping of paired end sequence data from the short-insert size WGS libraries, as well as of
106 published ESTs [14] (Additional file 1: Tables S3 and S4) against the genome assembly indicated
107 that the number and extent of misassemblies is low. To further estimate the completeness of the
108 assembly and gene prediction, the benchmarking universal single-copy orthologs (BUSCO) [17]
109 analysis was used and the results showed that the assembly contains 81.4% complete and 9.1%
110 partial vertebrate BUSCO orthologues (Additional file 1: Table S5).

111 The *M. amblycephala* genome has an average GC content of 37.3%, similar to cyprinid *C.*
112 *carpio* and *Danio rerio* (Additional file 1: Figure S4). Using a comprehensive annotation strategy
113 combining RNA-seq derived transcript evidence, *de-novo* gene prediction and sequence similarity
114 to proteins from five further fish species, we annotated a total of 23,696 protein-coding genes
115 (Additional file 1: Table S6). Of the predicted genes, 99.44% (23,563 genes) are annotated by

116 functional database. In addition, we identified 1,796 non-coding RNAs including 474 miRNAs,
117 220 rRNA, 530 tRNAs, and 572 snRNAs. Transposable elements (TEs) comprise approximately
118 34.18% (381.3 Mb) of the *M. amblycephala* genome (Additional file 1: Table S7). DNA
119 transposons (23.80%) and long terminal repeat retrotransposons (LTRs) (9.89%) are the most
120 abundant TEs in *M. amblycephala*. The proportion of LTRs in *M. amblycephala* is highest in
121 comparison with other teleosts: *G. morhua* (4.88%) [2], *L. crocea* (2.2%) [5], *C. semilaevis*
122 (0.08%) [6], *C. carpio* (2.28%) [12], *C. idellus* (2.58%) [13] and stickleback (*Gasterosteus*
123 *aculeatus*) (1.9%) [18] (Additional file 1: Tables S7 and S8, Figure S5). The distribution of
124 divergence between the TEs in *M. amblycephala* peaks at 7% (Additional file 1: Figure S6),
125 indicating a more recent activity of these TEs when compared with *O. mykiss* (13%) [3] and *C.*
126 *semilaevis* (9%) [6].

127 **Anchoring Scaffolds and Shared Synteny Analysis**

128 Sequencing data from 198 F1 specimens, including the parents, were used as the mapping
129 population to anchor the scaffolds on to 24 pseudo-chromosomes of the *M. amblycephala* genome.
130 Following RAD-Seq and sequencing protocol, 1883.5 Mb of 125-bp reads (on average 30.6 Mb
131 and 9.3 Mb of read data for each parent and progeny, respectively) were generated on the HiSeq
132 2500 next-generation sequencing platform. Based on the SOAP bioinformatic pipeline, we
133 generated 5,317 SNP markers for constructing a high-resolution genetic map. The map spans
134 1,701 cM with a mean marker distance of 0.33 cM and facilitated an anchoring of 1,434 scaffolds
135 comprising 70% (779.54 Mb) of the *M. amblycephala* genome assembly to form 24 linkage
136 groups (LG) (Additional file 1: Table S9). Of the anchored scaffolds, 598 could additionally be
137 oriented (678.27 Mb, 87.01% of the total anchored sequences) (Figure 1A). A subsequent
138 comparison of the gene order between *M. amblycephala* and its close relative *C. idellus* revealed
139 607 large shared syntenic blocks encompassing 11,259 genes, and 190 chromosomal
140 rearrangements. The values change to 1,062 regions, 13,152 genes and 279 rearrangements when
141 considering *D. rerio*. The unexpected higher number of genes in syntenic regions shared with the
142 more distantly related *D. rerio* is most likely an effect of the more complete genome assembly of
143 this species compared to *C. idellus*. The rearrangement events are distributed across all *M.*
144 *amblycephala* linkage groups without evidence for a local clustering (Figure 1B). The most

145 prominent event is a chromosomal fusion in *M. amblycephala* LG02 that joined two *D. rerio*
146 chromosomes, Dre10 and Dre22. The same fusion is observed in *C. idellus* but not in *C. carpio*
147 suggesting that it probably occurred in a last common ancestor of *M. amblycephala* and *C. idellus*,
148 approximately 13.1 million years ago (Additional file 1: Figure S7).

149 **Results**

150 **Evolutionary Analysis**

151 A phylogenetic analysis of 316 single-copy genes with one to one orthologs in the genomes of 10
152 other fish species, and coelacanth (*Latimeria chalumnae*) and elephant shark (*Callorhynchus milii*),
153 as out group served as a basis for investigating the evolutionary trajectory of *M. amblycephala*
154 (Figure 2A, Additional file 1: Figure S8). To illuminate the evolutionary process resulting in the
155 adaptation to a grass diet, we analyzed the functional categories of expanded genes in the *M.*
156 *amblycephala* and *C. idellus* lineage (Additional file 1: Figure S9, Additional file 2: Data Note1),
157 two typical herbivores mainly feeding on aquatic and terrestrial grasses. Among the significantly
158 over-represented KEGG pathways (Fisher's exact test, $P < 0.01$), we find olfactory transduction
159 (ko04740), immune-related pathways (ko04090, ko04672, ko04612 and ko04621), lipid metabolic
160 related process (ko00590, ko03320, ko00591, ko00565, ko00592 and ko04975), as well as
161 xenobiotics biodegradation and metabolism (ko00625 and ko00363) (Figure S10). Indeed, when
162 tracing positively selected genes (PSG) in *M. amblycephala* and *C. idellus* (Additional file 3: Date
163 Note2), we identified 10 candidates involved in starch and sucrose metabolism (ko00500), in
164 citrate cycle (ko00020) and in other types of O-glycan biosynthesis (ko00514). Moreover, 10
165 genes encoding enzymes involved in lipid metabolism appear positively selected in both fish
166 species (Additional file 1: Table S10).

167 **Development of Intermuscular Bones**

168 To explain the genetic basis of IB, their formation and their function in cyprinids, we first
169 analyzed the functional annotation of genes that expanded in this lineage (Figure 2C). Many of
170 these genes are involved in cell adhesion (GO: 0007155, $P = 5.26E-32$, 357 genes), myosin
171 complex (GO:0016459, $P = 2.74E-08$, 100 genes) and cell-matrix adhesion (GO:0007160,
172 $P = 1.59E-21$, 69 genes) (Figure 2C). As a second line of evidence, we performed transcriptome

173 analyses of early developmental stages (stage1: whole larvae without IBs) and juvenile *M.*
174 *amblycephala* (stage2: trunk muscle with partial IBs; stage3: trunk muscle with completed IBs)
175 (Figure 3A). Compared with stage1, 388 and 651 differentially expressed genes (DEGs) are
176 up-regulated in stage2 and stage3, respectively. And 249 of them are significantly up-regulated
177 both in stage2 and stage3. KEGG analyses indicate many of these genes involved in tight junction
178 (ko04530), regulation of actin cytoskeleton (ko04810), cardiac muscle contraction (ko04260) and
179 vascular smooth muscle contraction (ko04270) (Additional file 1: Figure S11). Specifically, 26
180 genes encoding proteins related to muscle contraction, including titin, troponin, myosin, actinin,
181 calmodulin and other Ca²⁺ transporting ATPases (Figure 3A) point to a strong remodeling of the
182 musculature compartment. To confirm that the observed differences in gene expression are indeed
183 linked to IB formation and function and are not simply due to the fact that different developmental
184 stages were compared, we performed differential expression analysis of muscle tissues, IB, and
185 connective tissues from the same six months old individual of *M. amblycephala* (Figure 3B,
186 Additional file 1: Figure S12). 1,290 DEGs and 5,231 DEGs are significantly up-regulated in IB
187 compared with connective tissues and muscle, respectively. 24 of these DEGs encode extracellular
188 matrix (ECM) proteins (collagens and integrin-binding protein), Rho GTPase family (*RhoA*, *Rho*
189 *GAP*, *Rac*, *Ras*), motor proteins (myosin, dynein, actin), and calcium channel regulation proteins
190 (Additional file 1: Figure S13 and Table S11). In addition, GO annotations of 963 IB-specific
191 genes indicative of abundance in protein binding (GO:0005515), calcium ion binding
192 (GO:0005509), GTP binding (GO:0005525) and iron ion binding (GO:0005506) were found
193 (Figure 3C).

194 During development of *M. amblycephala*, the first IB appears in muscles of caudal vertebrae
195 as early as 28 days post fertilization (dpf) when body length is 12.95 mm (Additional file 1: Figure
196 S14). The system then develops and ossifies predominantly from posterior to anterior (Additional
197 file 1: Figure S15). IBs are present throughout the body within two months (Additional file 1:
198 Figure S16) and develop into multiple morphological types in adults (Additional file 1: Figure
199 S17). The bone is formed directly without an intermediate cartilaginous stage (Additional file 1:
200 Figures S18 and S19). We also found a large number of mature osteoblasts distributed at the edge
201 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
202 matrix (Additional file 1: Figures S20 and S21). These primary bone-forming cells predominantly
203 regulate bone formation and function throughout life. Notably, among the genes up-regulated in
204 IB, 35 bone formation regulatory genes were identified (shown in colored boxes in Figure 3D). In
205 particular, genes involved in bone morphogenetic protein (BMP) signaling including *Bmp3*,
206 *Smad8*, *Smad9*, and *Id2*, in fibroblast growth factor (FGF) signaling including *Fgf2*, *Fgfr1a*,
207 *Fgfbp2*, *Col6a3*, and *Col4a5*, and in Ca²⁺ channels including *Cacna1c*, *CaM*, *Creb5* and *Nfatc*
208 were highly expressed (>2-fold change) in IB (Additional file 1: Figure S22).

209 **Adaptation to Herbivorous Diet**

17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
210 Next to the presence of IB, the herbivory of *M. amblycephala* is the second key feature in
211 connection to the use of this species as aquatic livestock. Olfaction, the sense of smell, is crucial
212 for animals to find food. The perception of smell is mediated by a large gene family of olfactory
213 receptor (OR) genes. In the *M. amblycephala* genome, we identified 179 functional olfactory
214 receptor (OR) genes (Figure 4A), and based on the classification of Niimura [19], 158, 117 and
215 153 receptors for water-borne odorants were identified in *M. amblycephala*, *C. idellus* and *D.*
216 *rerio*, respectively (Additional file 1: Table S12). Overall, these receptor repertoires are
217 substantially larger than those of other and carnivorous teleosts (*G. morhua*, *C. semilaevis*, *O.*
218 *latipes*, *X. maculatus*) (Additional file 1: Figures S23 and S24, Table S12). In addition, we found
219 a massive expansion of beta-type OR genes in the genomes of the herbivorous *M. amblycephala*
220 and *C. idellus*, while very few exist in other teleosts (Figure 4B, Additional file 1: Table S12).

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
221 Taste is also an important factor in the development of dietary habits. Most animals can
222 perceive five basic tastes, namely sourness, sweetness, bitterness, saltiness and umami [20]. *T1R1*,
223 the receptor gene necessary for sensing umami, has been lost in herbivorous *M. amblycephala* but
224 is duplicated in carnivorous *G. morhua*, *C. semilaevis* and omnivorous *O. latipes* and *X. maculatus*
225 (Figures 4C and 4D, Additional file 1: Figures S25-26 and Table S13). In contrast, *T1R2*, the
226 receptor gene for sensing sweet, has been duplicated in herbivorous *M. amblycephala* and *C.*
227 *idellus*, omnivorous *C. carpio* and *D. rerio*, while it has been lost in carnivorous *G. morhua* and *C.*
228 *semilaevis* (Additional file 1: Figure S27 and Table S13). Also the *T2R* gene family, most likely
229 important in the course switching to a diet that contains a larger fraction of bitterness containing

230 food, has been expanded in *M. amblycephala*, *C. idellus*, *C. carpio* and *D. rerio* (Additional file 1:
231 Figure S28).

232 To obtain further insights into the genetic adaptation to herbivorous diet, we focused on
233 further genes that might be associated with digestion. Genes that encode proteases (including
234 pepsin, trypsin, cathepsin and chymotrypsin) and amylases (including alpha-amylase and
235 glucoamylase) were identified in the genomes of *M. amblycephala*, carnivorous *C. semilaevis*, *G.*
236 *morhua* and omnivorous *D. rerio*, *O. latipes* and *X. maculatus*, indicating that herbivorous *M.*
237 *amblycephala* has a protease repertoire that is not substantially different from those of carnivorous
238 and omnivorous fishes (Additional file 1: Table S14). We did not identify any genes encoding
239 potentially cellulose-degrading enzymes including endoglucanase, exoglucanase and
240 beta-glucosidase in the genome of *M. amblycephala*, suggesting that utilization of the herbivorous
241 diet may largely depend on the gut microbiome. To elucidate this further, we determined the
242 composition of the gut microbial communities of juvenile, domestic, wild adult *M. amblycephala*
243 and wild adult *C. idellus* using bacterial 16S rRNA sequencing. A total of 549,020 filtered high
244 quality sequence reads from 12 samples were clustered at a similarity level of 97%. The resulting
245 8,558 operational taxonomic units (OTUs) are dominated at phylum level by Proteobacteria,
246 Fusobacteria, Bacteroidetes, Firmicutes and Actinobacteria (Additional file 1: Table S15, Figure
247 4E). Increasing the resolution to the genus level, the composition and relative abundance of the
248 gut microbiota of wild adult *M. amblycephala* and *C. idellus* are still very similar (Additional file
249 1: Table S16) and we could identify more than 7% cellulose-degrading bacteria (Additional file 1:
250 Table S17).

251 Discussion

252 The evolutionary trajectory analysis of *M. amblycephala* and other teleosts revealed that *M.*
253 *amblycephala* has the closest relationship to *C. idellus*. Both the species are herbivorous fish but
254 which endogenous and exogenous factors affected their feeding habits and how they adapted to
255 their herbivorous diet is not known. Our results from the expanded genes and PSG in the lineage
256 of the two herbivores uncovered a number of genes that are involved in glucose, lipid and
257 xenobiotics metabolism, which would enhance the ability of an herbivore to detoxify the

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 secondary compounds present in grasses that are adverse or even toxic to the organism.
259 Furthermore, the high-fiber but low-energy grass diet requires a highly effective intermediate
260 metabolism that accelerates carbohydrate and lipid catabolism and conversion into energy to
261 maintain physiological functions.

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

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

283 It has been reported that some insects such as *Tenebrio molitor* [26] and *Neotermes*
284 *koshunensis* [27], and the mollusc *Corbicula japonica* [28] have genes encoding endogenous
285 cellulose degradation-related enzymes. However, all so far analyzed herbivorous vertebrates lack
286 these genes and always rely on their gut microbiome to digest food [25, 29]. In herbivorous *M.*

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

287 *amblycephala* and *C. idellus*, we also did not find any homologues of digestive cellulase genes.
288 Interestingly, our work on the composition of gut microbiota of the two fish species identifies
289 more than 7% cellulose-degrading bacteria, suggesting that the cellulose degradation of
290 herbivorous fish largely depend on their gut microbiome.

291 IB has evolved several times during teleost evolution [9, 30]. The developmental mechanisms
292 and ossification processes forming IB are dramatically distinct from other bones such as ribs,
293 skeleton, vertebrae or spines. These usually develop from cartilaginous bone and are derived from
294 the mesenchymal cell population by endochondral ossification [31, 32]. However, IB form directly
295 by intramembranous ossification and differentiate from osteoblasts within connective tissue,
296 forming segmental, serially homologous ossifications in the myosepta. Although various methods
297 of ossification of IB have been proposed, few experiments have been conducted to confirm the
298 ossification process and little is known about the potential role of IB in teleosts. Based on our
299 findings of expanded genes in cyprinid lineage and evidence from transcriptome of developmental
300 stages of IB formation, a number of genes were found to interact dynamically to mediate efficient
301 cell motility, migration and muscle construction [33-36]. In addition, transcriptome analyses of
302 three tissues indicated that ECM, Rho GTPase, motor and calcium channel regulation protein
303 displayed high expression in IB. It is known that ECM proteins bound to integrins influence cell
304 migration by actomyosin-generated contractile forces [34, 37]. Rho GTPases, acting as molecular
305 switches, are also involved in regulating the actin cytoskeleton and cell migration, which in turn
306 initiates intracellular signaling and contributes to tissue repair and regeneration [38-40]. Thus, our
307 results provide molecular evidence that IB might play significant roles not only in regulating
308 muscle contraction but also in active remodeling at the bone-muscle interface and coordination of
309 cellular events.

310 Some major developmental signals including BMP, FGF, WNT, together with
311 calcium/calmodulin signaling [31, 41-43], are essential for regulating the differentiation and
312 function of osteoblasts and osteocytes and for regulating the RANKL signaling pathway for
313 osteoclasts [44]. In agreement with this concept, we found 35 bone formation regulatory genes
314 involved in these signals were highly up-regulated in IB. Among these signaling pathways, in
315 particular, *Bmp*, *Fgf2*, and *Fgfr1* are closely related to intramembranous bone development and

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

316 affect the expression and activity of other osteogenesis related transcription factors [31, 45]. The
317 calcium-sensitive transcription factor *NFATc1* together with *CREB* induces the expression of
318 osteoclast-specific genes [46]. Taken together, these results suggest that IB indeed undergoes an
319 intramembranous ossification process, is regulated by bone-specific signaling pathways, and
320 underlies a homeostasis of maintenance, repair and remodeling.

321 **Conclusions**

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

329 **Methods**

330 **Sampling and DNA Extraction**

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

336 **Genomic Sequencing and Assembly**

337 Libraries with different insert sized inserts of 170 bp, 500 bp, 800 bp, 2 Kb, 5 Kb, 10 Kb and 20
338 Kb were constructed from the genomic DNA at BGI-Shenzhen. The libraries were sequenced
339 using a HiSeq2000 instrument. In total, 11 libraries, sequenced in 23 lanes were constructed. To
340 obtain high quality data, we applied filtering criteria for the raw reads. As a result, 142.55 Gb of
341 filtered data were used to complete the genome assembly using SOAPdenovo_V2.04 [16]. Only
342 filtered data were used in the genome assembly. First, the short insert size library data were used
343 to construct a de Bruijn graph. The tips, merged bubbles and connections with low coverage were

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

344 removed before resolving the small repeats. Second, all high-quality reads were realigned with the
345 contig sequences. The number of shared paired-end relationships between pairs of contigs was
346 calculated and weighted with the rate of consistent and conflicting paired ends before constructing
347 the scaffolds in a stepwise manner from the short-insert size paired ends to the long-insert size
348 paired ends. Third, the gaps between the constructed scaffolds were composed mainly of repeats,
349 which were masked during scaffold construction. These gaps were closed using the paired-end
350 information to retrieve read pairs in which one end mapped to a unique contig and the other was
351 located in the gap region. Subsequently, local assembly was conducted for these collected reads.
352 To assess the genome assembly quality, approximately 42.82 Gb Illumina reads generated from
353 short-insert size libraries were mapped onto the genome. Bwa0.5.9-r16 software [47] with default
354 parameters was used to assess the mapping ratio and Soap coverage 2.27 was used to calculate the
355 sequencing depth. We also assessed the accuracy of the genome assembly by Trinity [48],
356 including number of ESTs and new mRNA reads from early stages of embryos and multiple
357 tissues, by aligning the scaffolds to the assembled transcriptome sequences.

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

362 **Genome Annotation**

363 The genome was searched for repetitive elements using Tandem Repeats Finder (version 4.04)
364 [49]. TEs were identified using homology-based approaches. The Repbase (version 16.10) [50]
365 database of known repeats and a *de novo* repeat library generated by RepeatModeler were used.
366 This database was mapped using the software of RepeatMasker (version 3.3.0). Four types of
367 non-coding RNAs (microRNAs, transfer RNAs, ribosomal RNAs and small nuclear RNAs) were
368 also annotated using tRNAscan-SE (version 1.23) and the Rfam database45 (Release 9.1) [51].

369 For gene prediction, *de novo* gene prediction, homology-based methods and RNA-seq data
370 were used to perform gene prediction. For the sequence similarity based prediction, *D. rerio*, *G.*
371 *aculeatus*, *O. niloticus*, *O. latipes* and *G. morhua* protein sequences were downloaded from
372 Ensembl (release 73) and were aligned to the *M. amblycephala* genome using TBLASTN. Then

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

373 homologous genome sequences were aligned against the matching proteins using GeneWise [52]
374 to define gene models. Augustus was employed to predict coding genes using appropriate
375 parameters in *de novo* prediction. For the RNA-seq based prediction, we mapped transcriptome
376 reads to the genome assembly using TopHat [53]. Then, we combined TopHat mapping results
377 together and applied Cufflinks [54] to predict transcript structures. All predicted gene structures
378 were integrated by GLEAN [55] (<http://sourceforge.net/projects/glean-gene/>) to obtain a
379 consensus gene set. Gene functions were assigned to the translated protein-coding genes using
380 Blastp tool, based on their highest match to proteins in the SwissProt and TrEMBL [56] databases
381 (Uniprot release 2011-01). Motifs and domains in the protein-coding genes were determined by
382 InterProScan (version 4.7) searches against six different protein databases: ProDom, PRINTS,
383 Pfam, SMART, PANTHER and PROSITE. Gene Ontology [57] IDs for each gene were obtained
384 from the corresponding InterPro entries. All genes were aligned against KEGG [58] (Release 58)
385 database, and the pathway in which the gene might be involved was derived from the matched
386 genes in KEGG. tRNA genes were *de novo* predicted by tRNAscan-SE software [59], with
387 eukaryote parameters on the repeat pre-masked genome. The rRNA fragments were identified by
388 aligning the rRNA sequences using BlastN at E-value 1e-5. The snRNA and miRNA were
389 searched by the method of aligning and searching with INFERNAL (version 0.81) [60] against
390 Rfam database (release 9.1).

391 **Genetic Map Construction**

392 To anchor the scaffolds into pseudo-chromosomes, 198 F1 population individuals were used to
393 obtain the genetic map. Each of the individual genomic DNA was digested with the restriction
394 endonuclease EcoR I, following the RAD-Seq protocol [61]. The SNP calling process was carried
395 out using the SOAP bioinformatic pipeline. The RAD-based SNP calling was done by SOAPSnp
396 software [62] after each individual's paired-end RAD reads was mapped onto the assembled
397 reference genome with the alignment software SOAP2 [63]. The potential SNP markers were used
398 for the linkage analysis if the following criteria were satisfied: for parents - sequencing depth ≥ 8
399 and ≤ 100 , base quality ≥ 25 , copy number ≤ 1.5 ; for progeny - sequencing depth ≥ 5 , base quality
400 ≥ 20 , copy number ≤ 1.5 . If the markers were showing significantly distorted segregation (P -value
401 < 0.01), they were excluded from the map construction. Linkage analysis was performed only for

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

402 markers present in at least 80% of the genomes, using JoinMap 4.0 software with CP population
403 type codes and applying the double pseudo-test cross strategy [64]. The linkage groups were
404 formed at a logarithm of odds threshold of 6.0 and ordered using the regression mapping
405 algorithm.

406 **Construction of Gene Families**

407 We identified gene families using TreeFam software [65] as follows: Blast was used to compare
408 all the protein sequences from 13 species: *M. amblycephala*, *C. idellus*, *C. semilaevis*, *C. carpio*,
409 *D. rerio*, *Callorhinchus milii*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *Oncorhynchus*
410 *mykiss*, *O. niloticus*, *O. latipes*, *Fugu rubripes*, with the E-value threshold set as 1e-7. In the next
411 step, HSP segments of each protein pair were concatenated by Solar software. H-scores were
412 computed based on Bit-scores and these were taken to evaluate the similarity among genes.
413 Finally, gene families were obtained by clustering of homologous gene sequences using
414 Hcluster_sg (Version 0.5.0). Specific genes of *M. amblycephala* were those that did not cluster
415 with other vertebrates that were chosen for gene family construction, and those that did not have
416 homologs in the predicted gene repertoire of the compared genomes. If these genes had functional
417 motifs, they were annotated by GO.

418 **Phylogenetic Tree Reconstruction and Divergence Time Estimation**

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

430 **Gene Family Expansion and Contraction Analyses**

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

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

439 **Detection of Positively Selected Genes**

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

452 **Developmental Process of Intermuscular Bone in *M. amblycephala***

453 To better understand the number and morphological types of IBs in adult *M. amblycephala*,
454 specimens with a body length ranging from 15.5 to 20.5 cm were collected and each individual
455 was wrapped in gauze and boiled. The fish body was divided into two sections: anterior (snout to
456 cloaca) and posterior (cloaca to the base of caudal fin), and the length of each section was
457 measured. The IBs were retrieved, counted, arranged in order and photographed with a digital
458 camera. Fertilized *M. amblycephala* eggs were brought from hatching facilities at Freshwater Fish
459 Genetics Breeding Center of Huazhong Agricultural University (Wuhan, Hubei, China) to our

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
460 laboratory. *M. amblycephala* larvae were maintained in a re-circulating aquaculture system at 23 ±
461 1°C with a 14-hr photoperiod. To explore the early development of IBs, larvae at different stages
462 from 15 to 40 dpf were collected and fixed in 4% paraformaldehyde and transferred to 70%
463 ethanol for storage. Specimens were stained with alizarin red for bone following the method
464 described by Dawson [74]. The appearance of red color was recorded as the appearance of IB
465 because bone ossification is accompanied by the uptake of alizarin red, resulting in red staining of
466 the mineralized bone matrix. Myosepta, either not yet ossified, or poorly ossified, are not visible
467 with alizarin red staining. For histologic analysis, specimens were paraffin-embedded and
468 sectioned following standard protocols. Sections were stained with hematoxylin and eosin (HE)
469 and Masson trichrome [75] and photographed using a Nikon microscope (Nikon, Tokyo, Japan)
470 with a DP70 digital camera (Olympus, Japan). Scanning electron microscopy (SEM) and
471 transmission electron microscopy (TEM) were also conducted to analyze the ultrastructure of IB.
472 The specimens were fixed with 2.5% (v/v) glutaraldehyde in a solution of 0.1 M sodium
473 cacodylate buffer (pH 7.3) for 2 h at room temperature. The SEM and TEM samples were
474 prepared according to a standard protocol described by Ott [76]. The samples were then visualized
475 with a JSM-6390LV scanning electron microscope (SEM, Japan) and the stained ultrathin sections
476 with a H-7650 transmission electron microscope (Hitachi, Japan).

36 477 **RNA Sequencing Analysis**

37
38 478 *M. amblycephala* specimens belonging to three different developmental stages of IBs (stage 1: whole
39
40 479 larvae without distribution of IB; stage 2: muscle tissues with partial distribution of IBs; stage 3:
41
42 480 muscle tissues with completed distribution of IBs were identified under microscope and immediately
43
44 481 frozen in liquid nitrogen. In addition, dorsal white muscle, IBs and connective tissue surrounding the
45
46 482 IBs from six months old fish were also collected. RNA was extracted from total fish samples at
47
48 483 different stages and from individual muscle, connective tissue, and intermuscular bone samples of
49
50 484 *M. amblycephala* using RNAisoPlus Reagent (TaKaRa, China) according to the manufacturer's
51
52 485 protocol. The integrity and purity of the RNA was determined by gel electrophoresis and Agilent
53
54 486 2100 BioAnalyzer (Agilent, USA) before preparing the libraries for sequencing. Paired-end RNA
55
56 487 sequencing was performed using the Illumina HiSeq 2000 platform. Low quality score reads were
57
58 488 filtered and the clean data were aligned to the reference genome using Bowtie [77]. Genes and
59
60
61
62
63
64
65

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

489 isoforms expression level were quantified by a software package: RSEM (RNASeq by
490 Expectation Maximization) [78]. Gene expression levels were calculated by using the RPKM
491 method (Reads per kilobase transcriptome per million mapped reads) [79] and adjusted by a
492 scaling normalization method [80]. We detected DEGs from three stages of IBs with software
493 NOIseq and three different tissues with PossionDis as requested. NOIseq is based on noisy
494 distribution model, performed as described by Tarazona [81]. The parameters were set as: fold
495 change ≥ 2.00 and probability ≥ 0.7 . PossionDis is based on the Poisson distribution, performed
496 as described by Audic [82]. The parameters were set as: fold change ≥ 2.00 and FDR ≤ 0.001 .
497 Annotation of DEGs were mapped to GO categories in the database
498 (<http://www.geneontology.org/>) and the number of genes for every term were calculated to
499 identify GO terms that were significantly enriched in the input list of DEGs. The calculated
500 *P*-value was adjusted by the Bonferroni Correction, taking corrected *P*-value ≤ 0.05 as a threshold.
501 KEGG automatic annotation was used to perform pathway enrichment analysis of DEGs.

502 **Prediction of Olfactory Receptor Genes**

503 Olfactory receptor genes were identified by previously described methods [83], with the exception
504 of a first-round TBLASTN [84] search, in which 1,417 functional olfactory receptor genes from *H.*
505 *sapiens*, *D. rerio*, *L. chalumnae*, *Lepisosteus oculatus*, *L. vexillifer*, *O. niloticus*, *O. latipes*, *F.*
506 *rubripes* and *Xenopus tropicalis* were used as queries. We then predicted the structure of
507 sequenced genes using the blast-hit sequence with the software GeneWise extending in both 3' and
508 5' directions along the genome sequences. The results were further confirmed by NR annotation.
509 Then the coding sequences from the start (ATG) to stop codons were extracted, while sequences
510 that contained interrupting stop codons or frame-shifts were regarded as pseudogenes. To
511 construct phylogenetic trees, the amino-acid sequences encoded by olfactory receptor genes were
512 first aligned using the program MUSCLE nested in MEGA 5.10 [85]. We then constructed the
513 phylogenetic tree using the neighbor-joining method with Poisson correction distances using the
514 program MEGA 5.10. We also identified and compared the genes for five basic tastes (sour, sweet,
515 bitter, umami and salty) using a similar method as in OR gene identification.

516 **Gut Microbiota Analysis**

517 To characterize the microbial diversity of herbivorous *M. amblycephala*, a total of 12 juvenile

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

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

530 **Additional files**

531 Additional file 1: Tables S1 to S17 and Figures S1 to S28.

532 Additional file 2: Data Note1 Expanded genes in the *M. amblycephala* and *C. idellus* lineage.

533 Additional file 3: Data Note2 Positively selected genes in the *M. amblycephala* and *C. idellus*
534 genomes.

535 **Abbreviations**

536 IB, intermuscular bone; SNP, single-nucleotide polymorphism; BUSCO, benchmarking universal
537 single-copy orthologs; TE, transposable element; LTR, long terminal repeat retrotransposon; LG,
538 linkage group; PSG, positively selected gene; ECM, extracellular matrix; dpf, days post
539 fertilization; BMP, bone morphogenetic protein; FGF, fibroblast growth factor; OR, olfactory
540 receptor; OTU, operational taxonomic unit; DEGs, differentially expressed genes; HE,
541 hematoxylin and eosin; SEM, scanning electron microscopy; TEM, transmission electron
542 microscopy

543 **Acknowledgements**

544 This work was supported by the Fundament Research Funds for the Central Universities
545 (2662015PY019), the Modern Agriculture Industry Technology System Construction Projects of

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

546 China titled as—Staple Freshwater Fishes Industry Technology System (No. CARS-46-05),
547 Guangdong Haid Group Co., Ltd and the International Scientific and Technology Cooperation
548 Program of Wuhan City (2015030809020365).

549 **Availability of data and materials**

550 Datasets supporting the results of this article are available in the GigaDB repository associated
551 with this publication [88]. Raw whole genome sequencing, transcriptome and RAD-Seq data have
552 been deposited at NCBI in the SRA under BioProject number PRJNA343584.

553 **Authors' contributions**

554 W.W. initiated and conceived the project and provided scientific input. X.Q. organized financial
555 support and designed the project. M.S. discussed the data, wrote and modified the paper. H.L. and
556 C.C. conducted the biological experiments, analyzed the data and wrote the paper with input from
557 other authors. I.E. wrote and modified the manuscript and discussed the data. The RAD-Seq data
558 analyses and the genetic map construction were performed by Z.G., Y.G., J.J. and X.J. Genome
559 assembly and annotation were performed by J.M., H.C., M.X. and J.C. X.Z., W.L., R.L., B.C.,
560 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 preparation
561 and data collection. J.L. and J.C. identified the gene families and analyzed the RNA-seq data. M.B.
562 coordinated the project. S.Z. and X.F. modified the manuscript and discussed the data. All authors
563 read the manuscript and provided comments and suggestions for improvements. The authors
564 declare no competing financial interests.

565 **Competing interests**

566 The authors declare that they have no competing interests.

567 **Ethics approval and consent to participate**

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

571 **References**

- 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
- 572 1. FAO Fisheries and Aquaculture Department. FAO yearbook Fishery and Aquaculture Statistics
573 2014 (Food and Agriculture Organization of the United Nations, Rome, 2016).
 - 574 2. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, et al. The genome
575 sequence of Atlantic cod reveals a unique immune system. *Nature*. 2011; 477:207–10.
 - 576 3. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al. The rainbow trout
577 genome provides novel insights into evolution after whole-genome duplication in vertebrates.
578 *Nat. Commun.* 2014; 5:3657.
 - 579 4. Tine M., Kuhl H., Gagnaire PA, Louro B, Desmarais E, Martins RS, et al. European sea bass
580 genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat.*
581 *Commun.* 2014; 5:5770.
 - 582 5. Wu C, Zhang D, Kan M, Lv Z, Zhu A, Su Y, et al. The draft genome of the large yellow
583 croaker reveals well-developed innate immunity. *Nat. Commun.* 2014; 5:5227.
 - 584 6. Chen S, Zhang G, Shao C, Huang Q, Liu G, Zhang P, et al. Whole-genome sequence of a
585 flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic
586 lifestyle. *Nat. Genet.* 2014; 46:253–60.
 - 587 7. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for
588 adaptive radiation in African cichlid fish. *Nature*. 2014; 513:375–82.
 - 589 8. Chen X, Zhong L, Bian C, Xu P, Qiu Y, You X, et al. High-quality genome assembly of
590 channel catfish, *Ictalurus punctatus*. *GigaScience*. 2016; 5:39.
 - 591 9. Gemballa S, Britz R. Homology of intermuscular bones in acanthomorph fishes. *Am. Mus.*
592 *Novit.* 1998; 3241:1–25.
 - 593 10. Danos N, Ward AB. The homology and origins of intermuscular bones in fishes: Phylogenetic
594 or biomechanical determinants? *Biol. J. Linn. Soc.* 2012; 106:607–22.
 - 595 11. Wan SM, Yi SK, Zhong J, Nie CH, Guan NN, Zhang WZ, et al. Dynamic mRNA and miRNA
596 expression analysis in response to intermuscular bone development of blunt snout bream
597 (*Megalobrama amblycephala*). *Sci. Rep.* 2016; 6:31050.
 - 598 12. Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, et al. Genome sequence and genetic diversity
599 of the common carp, *Cyprinus carpio*. *Nat. Genet.* 2014; 46:1212–9.
 - 600 13. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, et al. The draft genome of the grass carp
601 (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation.
602 *Nat. Genet.* 2015; 47:625–31.
 - 603 14. Gao Z, Luo W, Liu H, Zeng C, Liu X, Yi S, et al. Transcriptome analysis and SSR/SNP
604 markers information of the blunt snout bream (*Megalobrama amblycephala*). *PLoS One.*
605 2012; 7:e42637.

- 606 15. Yi S, Gao ZX, Zhao H, Zeng C, Luo W, Chen B, et al. Identification and characterization of
607 microRNAs involved in growth of blunt snout bream (*Megalobrama amblycephala*) by
608 Solexa sequencing. BMC Genomics. 2013; 14:754.
- 609 16. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved
610 memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.
- 611 17. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing
612 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.
613 2015;31:3210–2.
- 614 18. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis
615 of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61.
- 616 19. Niimura Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative
617 genome analysis among 23 chordate species. Genome Biol. Evol. 2009; 1:34–44.
- 618 20. Lindemann B. Receptors and transduction in taste. Nature. 2001; 413:219–25.
- 619 21. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene
620 repertoires: roles of chance and necessity. Nat. Rev. Genet. 2008; 9:951–63.
- 621 22. Chandrashekar J, Mueller KL, Hoon MA, Adler E, Feng L, Guo W, et al. T2Rs function as
622 bitter taste receptors. Cell. 2000; 100:703–11.
- 623 23. Niimura Y, Nei M. Evolutionary dynamics of olfactory and other chemosensory receptor
624 genes in vertebrates. J. Hum. Genet. 2006; 51:505–17.
- 625 24. Nelson G, Chandrashekar J, Hoon MA, Feng L, Zhao G, Ryba NJ, et al. An amino-acid taste
626 receptor. Nature. 2002; 416:199–202.
- 627 25. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and *de novo* assembly of the
628 giant panda genome. Nature. 2010; 463:311–7.
- 629 26. Ferreira AHP, Marana SR, Terra WR, Ferreira C. Purification, molecular cloning, and
630 properties of a β -glycosidase isolated from midgut lumen of *Tenebrio molitor* (Coleoptera)
631 larvae. Insect Biochem. Mol. Biol. 2001; 31:1065–76.
- 632 27. Tokuda G, Saito H, Watanabe H. A digestive β -glucosidase from the salivary glands of the
633 termite, *Neotermes koshunensis* (Shiraki): Distribution, characterization and isolation of its
634 precursor cDNA by 5'- and 3'-RACE amplifications with degenerate primers. Insect
635 Biochem. Mol. Biol. 2002; 32:1681–9.
- 636 28. Sakamoto K, Uji S, Kurokawa T, Toyohara H. Molecular cloning of endogenous
637 β -glucosidase from common Japanese brackish water clam *Corbicula japonica*. Gene. 2009;
638 435:72–9.
- 639 29. Zhu L, Wu Q, Dai J, Zhang S, Wei F. Evidence of cellulose metabolism by the giant panda gut

- 640 microbiome. Proc. Natl. Acad. Sci. USA. 2011; 108:17714–9.
- 641 30. Bird NC, Mabee PM. Developmental morphology of the axial skeleton of the zebrafish, *Danio*
642 *rerio* (Ostariophysi: Cyprinidae). Dev. Dyn. 2003; 228:337–57.
- 643 31. Ornitz D, Marie P. FGF signaling pathways in endochondral and intramembranous bone
644 development and human genetic disease. Genes Dev. 2002; 16:1446–65.
- 645 32. Ortega N, Behonick DJ, Werb Z. Matrix remodeling during endochondral ossification. Trends
646 Cell Biol. 2004; 14:86–93.
- 647 33. Even-Ram S, Doyle AD, Conti MA, Matsumoto K, Adelstein RS, Yamada KM. Myosin IIA
648 regulates cell motility and actomyosin-microtubule crosstalk. Nat. Cell Biol. 2007;
649 9:299–309.
- 650 34. Sheetz MP, Felsenfeld DP, Galbraith CG. Cell migration: Regulation of force on
651 extracellular-complexes. Trends Cell Biol. 1998; 8:51–4.
- 652 35. Gunst SJ, Zhang W. Actin cytoskeletal dynamics in smooth muscle: a new paradigm for the
653 regulation of smooth muscle contraction. Am. J. Physiol. Cell Physiol. 2008; 295:C576–87.
- 654 36. Webb RC. Smooth muscle contraction and relaxation. Adv. Physiol. Educ. 2003; 27:201–6.
- 655 37. Ulrich TA, De Juan Pardo EM, Kumar S. The mechanical rigidity of the extracellular matrix
656 regulates the structure, motility, and proliferation of glioma cells. Cancer Res. 2009;
657 69:4167–74.
- 658 38. Ridley AJ. Rho GTPases and cell migration. J. Cell Sci. 2001; 114:2713–22.
- 659 39. Etienne-Manneville S, Hall A. Rho GTPases in Cell Biology. Nature. 2002; 420:629–35.
- 660 40. Ridley AJ. Cell Migration: Integrating Signals from Front to Back. Science. 2003;
661 302:1704–9.
- 662 41. Chen G, Deng C, Li YP. TGF- β and BMP signaling in osteoblast differentiation and bone
663 formation. Int. J. Biol. Sci. 2012; 8:272–88.
- 664 42. Harada S, Rodan GA. Control of osteoblast function and regulation of bone mass. Nature.
665 2003; 423:349–55.
- 666 43. Long F. Building strong bones: molecular regulation of the osteoblast lineage. Nat. Rev. Mol.
667 Cell Biol. 2011; 13:27–38.
- 668 44. Boyle WJ, Simonet WS, Lacey DL. Osteoclast differentiation and activation. Nature. 2003;
669 423:337–42.
- 670 45. Fakhry A, Ratisoontorn C, Vedhachalam C, Salhab I, Koyama E, Leboy P, et al. Effects of
671 FGF-2/-9 in calvarial bone cell cultures: Differentiation stage-dependent mitogenic effect,
672 inverse regulation of BMP-2 and noggin, and enhancement of osteogenic potential. Bone.

673 2005; 36:254–66.

674 46. Sato K, Suematsu A, Nakashima T, Takemoto-Kimura S, Aoki K, Morishita Y, et al.
675 Regulation of osteoclast differentiation and function by the CaMK-CREB pathway. *Nat. Med.*
676 2006; 12:1410–6.

677 47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
678 *Bioinformatics.* 2009; 25:1754–60.

679 48. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity:
680 reconstructing a full-length transcriptome without a genome from RNA-Seq data . *Nat.*
681 *Biotechnol.* 2011; 29:644–52.

682 49. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.*
683 1999; 27:573–80.

684 50. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase
685 Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 2005;
686 110:462–7.

687 51. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: Annotating
688 non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005; 33:121–4.

689 52. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res.* 2004; 14:988–95.

690 53. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq.
691 *Bioinformatics.* 2009; 25:1105–11.

692 54. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript
693 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform
694 switching during cell differentiation. *Nat. Biotechnol.* 2010; 28:511–5.

695 55. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey
696 bee consensus gene set. *Genome Biol.* 2007; 8:R13.

697 56. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement
698 TrEMBL in 2000. *Nucleic Acids Res.* 2000; 28:45–8.

699 57. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool
700 for the unification of biology. *Nat. Genet.* 2000; 25:25–9.

701 58. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*
702 2000; 28:27-30.

703 59. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes
704 in genomic sequence. *Nucleic Acids Res.* 1997; 25:955–64.

- 705 60. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: Inference of RNA alignments. *Bioinformatics*.
706 2009; 25:1335–7.
- 707 61. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP
708 discovery and genetic mapping using sequenced RAD markers. *PloS One*. 2008; 3:e3376.
- 709 62. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively
710 parallel whole-genome resequencing. *Genome Res*. 2009; 19:1124–32.
- 711 63. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: An improved ultrafast tool
712 for short read alignment. *Bioinformatics*. 2009; 25:1966–7.
- 713 64. Grattapaglia D, Sederoff R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus*
714 *urophylla* using a pseudo-testcross: Mapping strategy and RAPD markers. *Genetics*. 1994;
715 137:1121–37.
- 716 65. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, et al. TreeFam: a curated
717 database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 2006;
718 34:D572–80.
- 719 66. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.
720 *Nucleic Acids Res*. 2004; 32:1792–7.
- 721 67. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and
722 methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML
723 3.0. *Syst. Biol*. 2010; 59:307–21.
- 724 68. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by
725 maximum likelihood. *Syst. Biol*. 2003; 52:696–704.
- 726 69. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. 2007;
727 24:1586–91.
- 728 70. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock
729 using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol*. 2006; 23:212–26.
- 730 71. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics*.
731 2007; 177:1941–9.
- 732 72. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with
733 insertions. *Proc. Natl. Acad. Sci. USA*. 2005; 102:10557–62.
- 734 73. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and
735 ambiguously aligned blocks from protein sequence alignments. *Syst. Biol*. 2007; 56:564–77.
- 736 74. Dawson AB. A note on the staining of the skeleton of cleared specimens with alizarin red S.
737 *Biotech. Histochem*. 1926; 1:123-4.

- 738 75. Gruber HE. Adaptations of Goldner's Masson trichrome stain for the study of undecalcified
739 plastic embedded bone. *Biotech. Histochem.* 1992; 67:30–4.
- 740 76. Ott HC, Matthiesen TS, Goh SK, Black LD, Kren SM, Netoff TI, et al.
741 Perfusion-decellularized matrix: using nature's platform to engineer a bioartificial heart. *Nat.*
742 *Med.* 2008; 14:213–21.
- 743 77. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 2012;
744 9:357–9.
- 745 78. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or
746 without a reference genome. *BMC Bioinformatics.* 2011; 12:323.
- 747 79. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying
748 mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 2008; 5:621–8.
- 749 80. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis
750 of RNA-seq data. *Genome Biol.* 2010; 11:R25.
- 751 81. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in
752 RNA-seq: a matter of depth. *Genome Res.* 2011; 21:2213–23.
- 753 82. Audic S, Claverie JM. The significance of digital gene expression profiles. *Genome Res.* 1997;
754 7:986–95.
- 755 83. Niimura Y. Evolutionary dynamics of olfactory receptor genes in chordates: interaction
756 between environments and genomic contents. *Hum. Genomics.* 2009; 4:107–18.
- 757 84. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
758 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;
759 25:3389–402.
- 760 85. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary
761 analysis of DNA and protein sequences. *Brief. Bioinform.* 2008; 9:299–306.
- 762 86. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al.
763 Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc.*
764 *Natl. Acad. Sci. USA.* 2011; 108:4516–22.
- 765 87. Edgar, RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat.*
766 *Methods.* 2013; 10:996–8.
- 767 88. Liu H, Chen CH, Gao ZX, Min JM, Gu YM, Jian JB, et al. The draft genome of *Megalobrama*
768 *amblycephala* reveals the development of intermuscular bone and adaptation to herbivorous
769 diet. 2016. GigaScience Database.

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

770 **Figure Legends**

771 **Figure 1** Global view of the *M. amblycephala* genome and syntenic relationship between *C.*
772 *idellus*, *M. amblycephala* and *D. rerio*. (A) Global view of the *M. amblycephala* genome. From
773 outside to inside, the genetic linkage map (a); Anchors between the genetic markers and the
774 assembled scaffolds (b); Assembled chromosomes (c); GC content within a 50-kb sliding window
775 (d); Repeat content within a 500-kb sliding window (e); Gene distribution on each chromosome (f);
776 Different gene expression of three transcriptomes (g). (B) Syntenic relationship between *C. idellus*
777 (a), *M. amblycephala* (b) and *D. rerio* (c) chromosomes.

778 **Figure 2** Phylogenetic tree and comparison of orthologous genes in *M. amblycephala* and other
779 fish species. (A) Phylogenetic tree of teleosts using 316 single copy orthologous genes. The color
780 circles at the nodes shows the estimated divergence times using *O. latipes*–*F. rubripes*
781 [96.9~150.9Mya], *F. rubripes*–*D. rerio* [149.85~165.2Mya], *F. rubripes*–*C. milii* [416~421.75Mya]
782 (<http://www.timetree.org/>) as the calibration time. Pentagram represents four cyprinid fish with
783 intermuscular bones. S, silurian period; D, devonian period; C, carboniferous period; P, permian
784 period in Paleozoic; T, triassic period; J, jurassic and k-cretaceous period in Mesozoic; Pg,
785 paleogene in Cenozoic Era, N, Neogene. (B) Venn diagram of shared and unique orthologous gene
786 families in *M. amblycephala* and four other teleosts. (C) Over-represented GO annotations of
787 cyprinid-specific expansion genes.

788 **Figure 3** Regulation of genes related to intermuscular bone formation and function identified from
789 developmental stages and adult tissues transcriptome data. (A) Gene expression pattern involved
790 in muscle contraction regulated genes in early developmental stages corresponds to intermuscular
791 bone formation of *M. amblycephala*, (alizarin red staining). M, myosepta; IB, intermuscular bone.
792 (B) Scanning electron microscope photos of muscle tissues, connective tissues, and intermuscular
793 bone. (C) Distribution of intermuscular bone specific genes in GO annotations indicative of
794 abundance in protein binding, calcium ion binding, GTP binding functions. (D) Several
795 developmental signals regulating key steps of osteoblast and osteoclast differentiation in the
796 process of intramembranous ossification. Colored boxes indicate significantly up-regulated genes
797 in these signals specifically occurred in intermuscular bone.

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

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823 **Table**

824 **Table 1 Features of the *M. 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%)

825

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

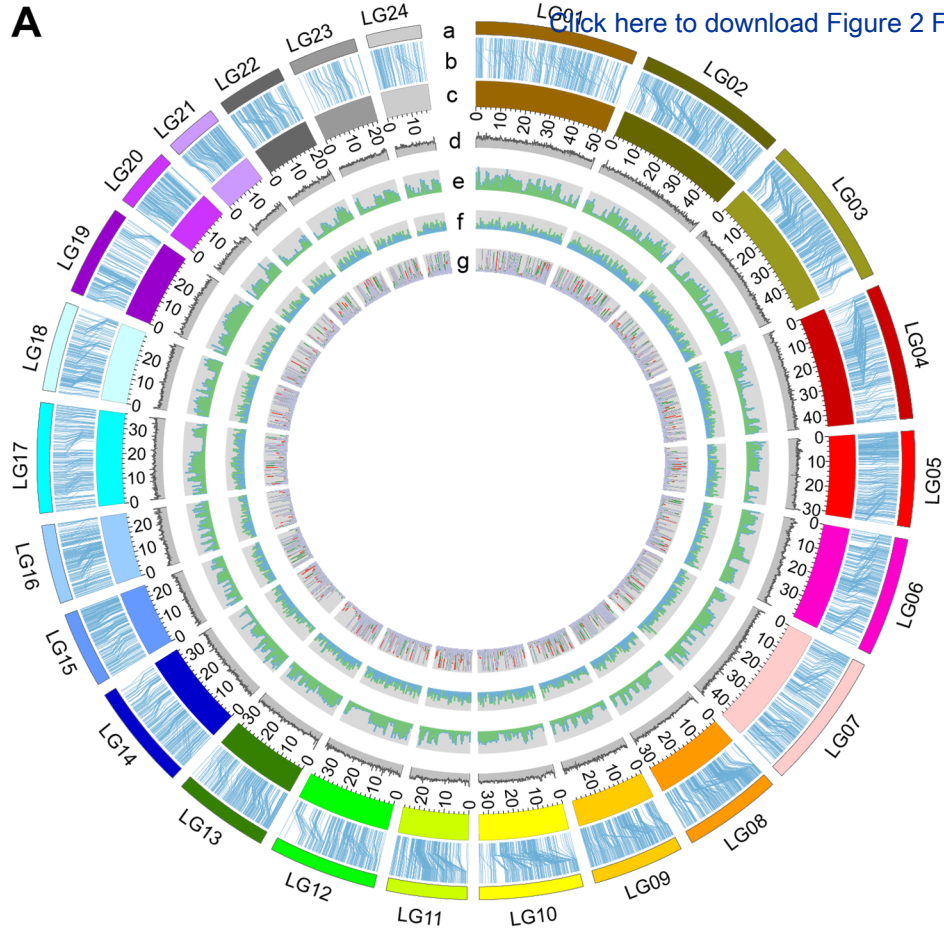
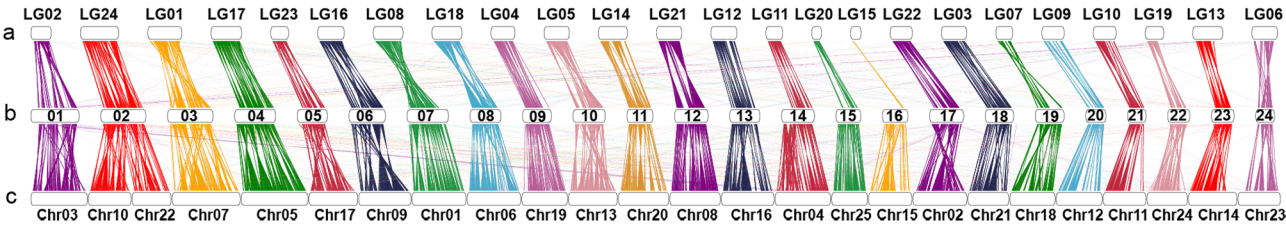
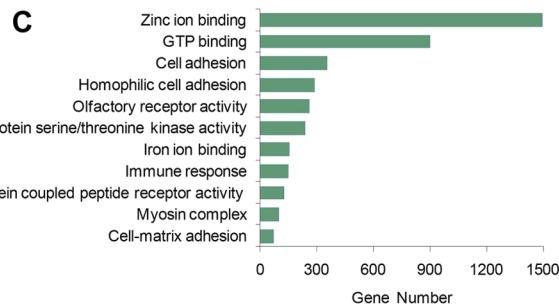
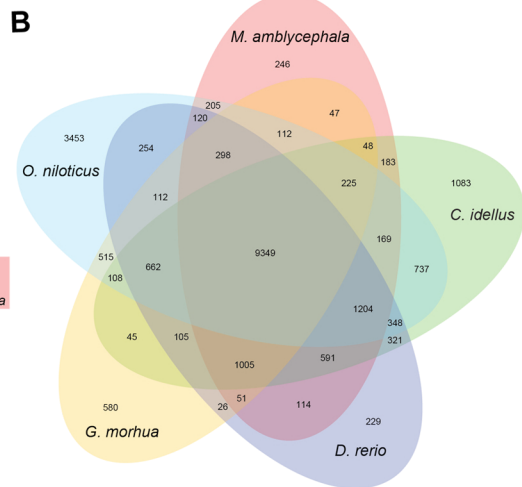
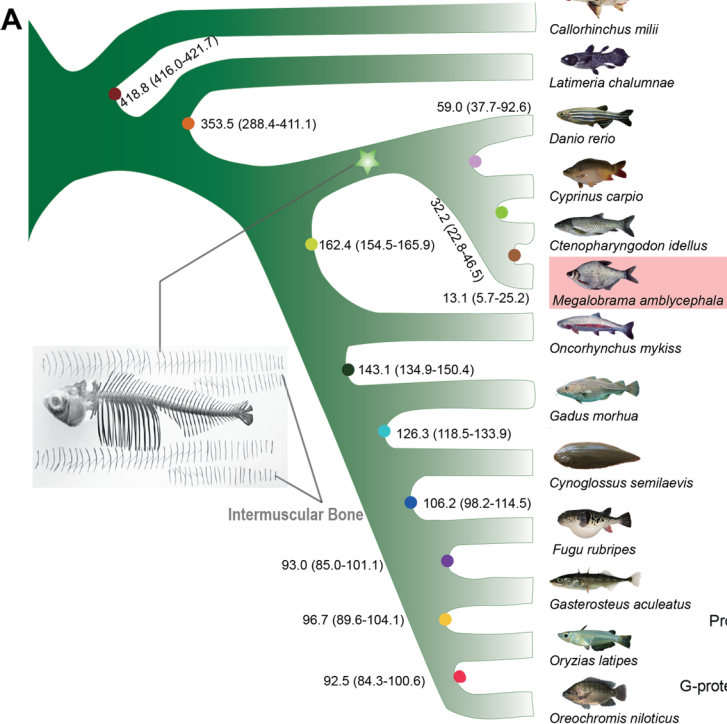
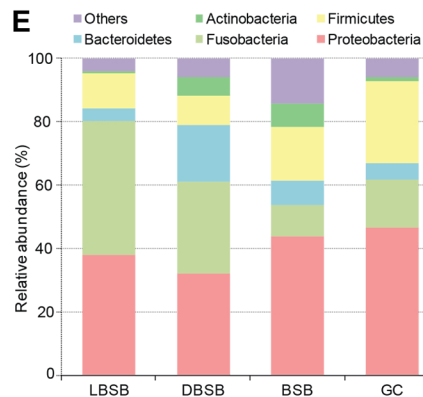
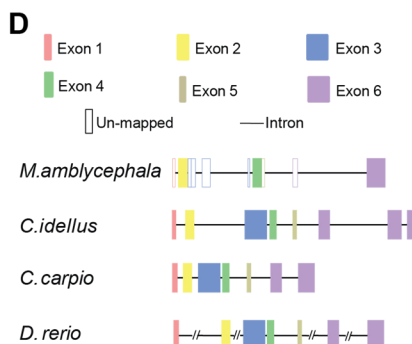
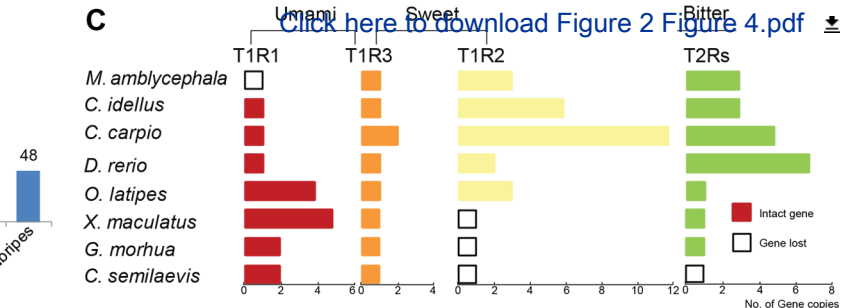
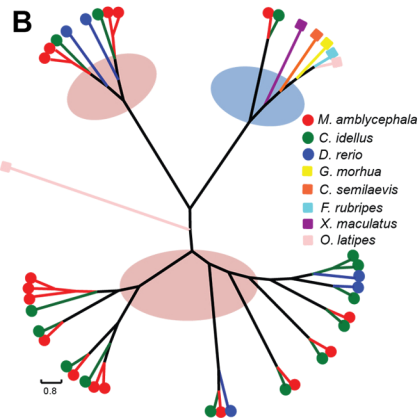
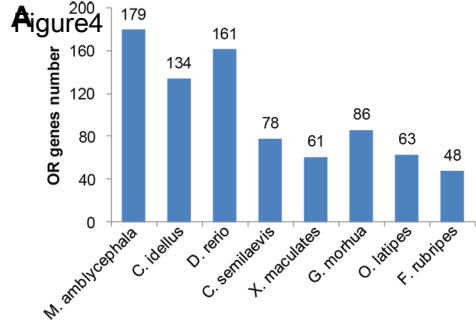

A**B**

Figure2

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





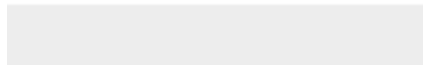
Click here to access/download

Supplementary Material

3 Additional file 1 Tables S1 to S17 and Figures S1 to
S28.pdf



Click here to access/download
Supplementary Material
3 Additional file 2 Data note1.xlsx





Click here to access/download
Supplementary Material
3 Additional file 3 Data note2.xlsx

