

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

<b>Manuscript Number:</b>	GIGA-D-16-00088R1
<b>Full Title:</b>	The draft genome of <i>Megalobrama amblycephala</i> reveals the development of intermuscular bone and adaptation to herbivorous diet
<b>Article Type:</b>	Research
<b>Funding Information:</b>	
<b>Abstract:</b>	<p><b>Background:</b> The blunt snout bream, <i>Megalobrama amblycephala</i>, is the economically most important cyprinid fish species. As an herbivore, it can be grown by eco-friendly and resource-conserving aquaculture. However, the large number of intermuscular bones in the trunk musculature is adverse to fish meat processing and consumption.</p> <p><b>Results:</b> As a first towards optimizing this aquatic livestock, we present a 1.116-Gb draft genome of <i>M. amblycephala</i>, with 779.54 Mb anchored on 24 linkage groups. Integrating spatiotemporal transcriptome analyses we show that intermuscular bone is formed in the more basal teleosts by intramembranous ossification, and may 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><b>Conclusions:</b> As a valuable resource for improvement of <i>M. amblycephala</i> livestock, the draft genome sequence offers new insights into the development of intermuscular bone and herbivorous adaptation.</p>
<b>Corresponding Author:</b>	Weimin Wang  CHINA
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Han Liu
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	Han Liu
	Chunhai Chen
	Zexia Gao
	Jiumeng Min
	Yongming Gu
	Jianbo Jian
	Xiewu Jiang
	Huimin Cai
	Ingo Ebersberger
	Meng Xu
	Xinhui Zhang
	Jianwei Chen

	Wei Luo
	Boxiang Chen
	Junhui Chen
	Hong Liu
	Jiang Li
	Ruifang Lai
	Mingzhou Bai
	Jin Wei
	Shaokui Yi
	Huanling Wang
	Xiaojuan Cao
	Xiaoyun Zhou
	Yuhua Zhao
	Kaijian Wei
	Ruibin Yang
	Bingnan Liu
	Shancen Zhao
	Xiaodong Fang
	Manfred Scharl
	Xueqiao Qian
	Weimin Wang
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>24 November 2016  Dr. Hans Zauner  Journal: GigaScience</p> <p>Dear Dr. Zauner,  Manuscript No.: GIGA-D-16-00088  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 referees' comments which you forwarded to us with your email of 2 November 2016. We would like to express our sincere thanks to the reviewers for the constructive and positive comments. We have addressed all their suggestions, and the manuscript has been edited accordingly. Please find below our responses to the comments of the referees. The major amendments are highlighted in red in the revised manuscript. Responses to each of the reviewers' comments are detailed below in this letter. Because of the amendments, the page and the line numbers referred to by the referees 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 and embedded image data (intermuscular bones and histological stainings) were uploaded to the indicated ftp. We hope that with the amendments made in response to the reviewers' comments, the manuscript is now acceptable for</p>

publication in GigaScience.  
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 Reviewers

##### Reviewer Reports

##### Reviewer 1:

It is clear that a lot of work has gone into the creation of a draft genome for *Megalobrama amblycephala*. The creation of a genetic map to help anchor the assembly is significant and I am please to see it. Their results and discussion flow logically from the genome assembly and annotation as compared with other fish species. Although every bioinformatician has their favorite open source program, the software used in the construction of the assembly and annotation are well known and documented.

I recommend acceptance with the following comments and strong recommendations.

1) All raw data for the assembly and annotation be submitted to a public database.

Author response: The raw sequencing data have been deposited at NCBI under the accession number SRP090157 and raw data is available from the SRA under bioproject number PRJNA343584. We also have uploaded the assemblies, annotations, RNA-seq and microbiome data to the ftp (<ftp://user28@climb.genomics.cn>).

2) All raw data and markers for the genetic map be submitted to a public database or made readily downloadable.

Author response: All the raw data and markers for the genetic map have been deposited at NCBI in the SRA under bioproject number PRJNA343584 and biosample number SRS1797758.

3) The genome assembly be checked for contamination. This can be easily done with programs like blobtools and megablast. If contamination is found please remove it prior to submitting to the public databases and update the paper accordingly. In this way we do not contaminate and propagate contamination in our common public databases.

Author response: We have checked the genome assembly for contamination. The results are shown as bellow:

```
Scaffold_IDStartEndCover_LenScaffold_LenCoverage (%)Tax_IDOrganism  
scaffold16634177975695878.911148Synechocystis sp.
```

```
scaffold1939712958345590850.11111780Stanieria cyanosphaera
```

```
scaffold1189010239429352356.021287681Eutypa lata
```

Only one gene *MamblycephalaGene23156.1* is involved in the contamination region.

We have now shielded them and updated the gene and genome data in the indicated database (<ftp://user28@climb.genomics.cn>).

4) Consider swapping out some of the language of "plays a role" with phrases like "are involved in" or "have reported to" or "have a role in" etc.;

Author response: We have replaced the expression of "plays a role" by "be involved in" or "be associated with" throughout the whole manuscript.

5) On Line 241, you have written bittern, I suspect you either meant bitter or bitterness.

Author response: This was a language error on the out side and it should read bitterness. We now have amended it in the revised manuscript.

6) I downloaded and examined the annotation file *Megalobrama amblycephala.gff*. It only contains transcript models and no gene models. You state in the paper there were 23,696 protein coding genes. This number corresponds to the number of mRNA models you have in this gff file. This struck me as odd as there are usually isoforms for gene models. This file also did not contain any functional annotation. Please create a gff3 file that contains genes models and include functional annotation as this will be of great value to researchers.

Author response: We have created a gff3 file that contains genes models and functional annotations. This will be available in the GigaDB repository associated with the paper publication. The format of this file as shown below:

This is an important work for the development of sustainable aquaculture. Thank you.  
-----

Reviewer 2:

The draft genome of *Megalobrama amblycephala* reveals the development of intermuscular bone and adaptation to herbivorous diet by Liu H., et al. The present manuscript by Liu H., et al. can be divided into three main parts. The first part deals with the genome sequencing of one of the economically important cyprinid fish species *Megalobrama amblycephala*. Authors have generated a draft genome of 1.116 GB and managed to link 779.54 Mb to 24 linkage groups. Linkage groups were also constructed in the present work. Authors further investigate comparative genomics and phylogenetic analysis using the generated data. The second part investigates the characteristics of the feeding strategy of *Megalobrama amblycephala* being an herbivore species.

1) This comprised analysis of “expanded gene families” (this expression is rather unfortunate and not clear. It does not show which gene families finally were looked at and supplementary material like figure S11 does not say much) as well as the gut microbial community, but no differential expression analysis (mRNA). In the method section the experimental set up and analysis of the gut microbial study is missing.

Author response: We have now clarified these questions according to the reviewer’s suggestions. The expanded genes in the *M. amblycephala* and *C. idellus* lineage are now listed in the Additional file 2: Data Note1. The gut microbial community of the larvae (LBSB), domestic adult (DBSB), wild adult (BSB) *M. amblycephala* and wild adult *C. idellus* (GC) at phylum level and genus level are shown in Figure 4E and Additional file 1: Table S16, respectively. The sequences information and alpha diversity indexes are shown in Table S15 in the revised Additional file 1. We have added the experimental set up and analysis of the gut microbial study in the method section of the revised manuscript.

2) The third part studies the development of intermuscular bones comprising the analysis of “expanded gene families” (again figure 2B does not explain the “expanded gene families”) and transcriptome analysis of early developmental stages.

Author response: We apologize for the mistake. Reference should have been made to Figure 2C and not to Figure 2B. This has been corrected in the revised manuscript.

3) Authors present a huge amount of data and analysis but finally do describe and discuss only a little part of their analysis mainly in form of a very small set of genes. The manuscript is well, but not straightforward written. The structure has to be revised and the outcome better worked out.

Author response: We have revised our manuscript according to your suggestions and modified and renumbered the figures and tables clearly.

#### INTRODUCTION

1) LINE 64: Today many genomes (draft and nearly complete) are available. It is suggested to categorize them into fresh water, Mediterranean and Atlantic sea important aquacultured species. One species e.g. important for the Mediterranean aquaculture, The European sea bass (*Dicentrarchus labrax*) which was recently sequenced is not listed (European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature communications, 2014 5, 5770.)

Author response: This important species, the European sea bass, has been listed now in the revised manuscript (Line 65).

2) LINE 68-71: Please re-phrase the sentence. It is not true that the focus of aquaculture is focusing on herbivorous species. It is true however, that the usage of alternative feed is pursued but not the culture of herbivorous species. In contrast several projects are working on new species (including carnivores) for aquaculture purposes.

Author response: This sentence has been re-phrased as “Reports on draft genomes of some resource friendly herbivorous and omnivorous species, in particular cyprinid fish are scarce” in the revised manuscript (Line 69-70).

3) LINE 79 and LINE 84-85: Please refer to Wan et al., 2016 “Dynamic mRNA and miRNA expression analysis in response to intermuscular bone development of blunt snout bream (*Megalobrama amblycephala*)”, Scientific Reports.

Author response: We have added this research paper as a reference in the revised manuscript (Line 80-83 and Line 88-89).

#### DATA DESCRIPTION

1) Description of generated SNP linkage map appears here, but does not appear in the method section. Please accomplish the method section.

Author response: We have added such information in detail now in Data Description section (Line 128-133) and also added the description of the genetic map construction in the method section (Line 403-416) in the revised manuscript.

#### ANALYSES SECTION

1) According to GigaScience Authors guide: “This section should provide details of all of the experiments and analyses that are required to support the conclusions of the paper. The authors should make clear the goal of each analysis and state the basic findings”. Information about analyses (except the linkage map analysis) is partially found in the section named “Results” as well as in the “Method” section.

Author response: We have now modified this section according to your suggestions. We stated the goal and main finding of each analysis in the Results section and also removed the reporting of findings from the Method section.

2) LINES147-153: It is not clear why the enrichment analyses shows that the adaptation to herbivory goes “hand in hand” with the coping with plant secondary metabolites. Do authors have a comparable analysis of a carnivore or omnivore teleosts?

Author response: This sentence was ambiguous and we apologize for the confusion that it generated. We have now changed this expression to “These genes encoding proteins involved in biodegradation of xenobiotics would enhance the ability of an

herbivore to detoxify the secondary compounds present in grasses that are adverse or even toxic to the organism.” (Line 161-164).

3) LINE 170: What do authors mean by the expression “comparative transcriptome analysis”?

Author response: This expression was not accurate. We have now modified the expression and write “differential genes expression (DGE) analysis” or “transcriptome analysis” at the appropriate positions of the manuscript.

4) LINE 174: “notable”. This result is not really surprising. Many other transcriptome studies in teleost have also shown that at the later stages, where the larvae is mainly growing, mostly muscle genes are up-regulated when compared to earlier stages.

Author response: We agree. This word has been deleted in the revised manuscript.

5) LINE 183: change the expression “eventually extended”.

Author response: This expression has been changed to “we performed differential genes expression (DGE) analysis of” in the revised manuscript (Line 192).

6) LINE 203. The link to Figure 3D is not clear. Is Figure 3D showing all 35 identified genes?

Author response: Figure 3D shows several developmental signals regulating key steps of osteoblast and osteoclast differentiation in the process of intramembranous ossification. In Figure 3D, only the colored boxes indicate these 35 identified genes. To make this more clear, we have amended this sentence to “35 bone formation regulatory genes were identified (shown in colored boxes in Figure 3D)” (Line 212).

## DISCUSSION

Mainly a repetition of the previous paragraphs.

1) LINE 134 and 271: Authors did not show that Dre10 and Dre22 are ‘ancestral’, just that those two chromosomes fused to one chromosome in blunt snout bream as this species has 24 chromosomes while zebrafish has 25 chromosomes. Please take into account publications like Nakatani, Y., H. Takeda, Y. Kohara, and S. Morishita, 2007 Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17: 1254–1265 and. Hufton, A. L., D. Groth, M. Vingron, H. Lehrach, A. J. Poustka et al., 2008 Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res.* 18: 1582–1591.

Author response: According to the referee’s suggestion and taking into account the two references, we have modified this sentence to “The most prominent event is a chromosomal fusion in *M. amblycephala* LG02 that joined two *D. rerio* chromosomes, Dre10 and Dre22”.

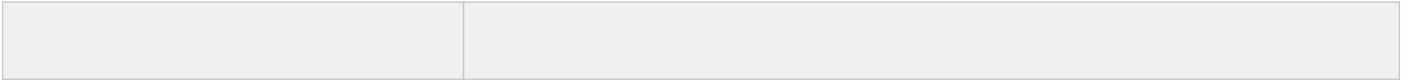
2) LINE 311: ‘comparative transcriptome” This expression leads the reader to the false impression that more than one species was studied. However authors investigated here in differential expression.

Author response: The expression “comparative transcriptome” has been amended to “differential genes expression (DGE) analysis” or “transcriptome analysis” in the appropriate positions of the manuscript.

## METHODS

1) Missing description of microbial community study as well as generation of linkage

	<p>map.—</p> <p>Author response: We have now added more detailed information in the Methods section including genome assembly, construction of gene families, the microbiota analysis, genetic map construction and other related information.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes





1 24 November 2016

2 Dr. Hans Zauner

3  
4 Journal: GigaScience

5  
6  
7  
8 Dear Dr. Zauner,

9  
10 **Manuscript No.: GIGA-D-16-00088**

11  
12 **Title: "The draft genome of *Megalobrama amblycephala* reveals the development of**

13 **intermuscular bone and adaptation to herbivorous diet"**

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

25  
26  
27  
28  
29 We have carefully read the referees' comments which you forwarded to us with your email of 2  
30  
31 November 2016. We would like to express our sincere thanks to the reviewers for the constructive  
32  
33 and positive comments. We have addressed all their suggestions, and the manuscript has been  
34  
35 edited accordingly. Please find below our responses to the comments of the referees. The major  
36  
37 amendments are highlighted in red in the revised manuscript. Responses to each of the reviewers'  
38  
39 comments are detailed below in this letter. Because of the amendments, the page and the line  
40  
41 numbers referred to by the referees have now changed in the edited version of the manuscript.  
42  
43 Please note that all raw reads of genome sequencing and RAD-seq have been deposited at NCBI.  
44  
45 Other data including assemblies, annotations, RNA-seq data, microbiome data and embedded  
46  
47 image data (intermuscular bones and histological stainings) were uploaded to the indicated ftp. We  
48  
49 hope that with the amendments made in response to the reviewers' comments, the manuscript is  
50  
51 now acceptable for publication in GigaScience.

52  
53 I look forward to hearing from you soon.

54  
55  
56  
57  
58 Yours sincerely,

59  
60 Weimin Wang (PhD) (Correspondence author)

61  
62  
63  
64  
65

1 College of Fisheries  
2  
3 Huazhong Agricultural University  
4  
5 Wuhan 430070, P. R. China  
6  
7 E-mail address: [wangwm@mail.hzau.edu.cn](mailto:wangwm@mail.hzau.edu.cn)  
8  
9 Tel: +86-27-8728 4292; Fax: +86-27-8728 4292

## 17 Response to Reviewers

### 18 Reviewer Reports

#### 19 Reviewer 1:

20  
21  
22  
23 It is clear that a lot of work has gone into the creation of a draft genome for *Megalobrama*  
24 *amblycephala*. The creation of a genetic map to help anchor the assembly is significant and I am  
25 please to see it. Their results and discussion flow logically from the genome assembly and  
26 annotation as compared with other fish species. Although every bioinformatician has their favorite  
27 open source program, the software used in the construction of the assembly and annotation are  
28 well known and documented.

29 I recommend acceptance with the following comments and strong recommendations.

30 1) All raw data for the assembly and annotation be submitted to a public database.

31  
32  
33  
34  
35  
36  
37  
38  
39  
40 Author response: The raw sequencing data have been deposited at NCBI under the accession  
41 number SRP090157 and raw data is available from the SRA under bioproject number  
42 PRJNA343584. We also have uploaded the assemblies, annotations, RNA-seq and microbiome  
43 data to the ftp (<ftp://user28@climb.genomics.cn>).  
44  
45  
46

47  
48  
49  
50  
51  
52  
53  
54  
55

Submission Id	Submitter	Updated	State	Status	Comments
Huazhong Agricultural University : RAD-Seq	Han Liu	2016-11-15 12:55	completed	47	<ul style="list-style-type: none"><li>SRS1797758 : blunt snout brea</li><li>20 experiments</li><li>26 runs</li></ul>
Huazhong Agricultural University : Wuchangfish genome	Han Liu	2016-09-23 14:36	completed	38	<ul style="list-style-type: none"><li>SRP090157 : PRJNA343584</li><li>SRS1703996 : Wuchang bream</li><li>11 experiments</li><li>23 runs</li></ul>

56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 2) All raw data and markers for the genetic map be submitted to a public database or made readily  
2  
3 downloadable.

4 **Author response: All the raw data and markers for the genetic map have been deposited at NCBI**  
5 **in the SRA under bioproject number PRJNA343584 and biosample number SRS1797758.**  
6

7  
8 3) The genome assembly be checked for contamination. This can be easily done with programs  
9 like blobtools and megablast. If contamination is found please remove it prior to submitting to the  
10 public databases and update the paper accordingly. In this way we do not contaminate and  
11 propagate contamination in our common public databases.  
12

13 **Author response: We have checked the genome assembly for contamination. The results are shown**  
14 **as bellow:**  
15

Scaffold_ID	Start	End	Cover_Len	Scaffold_Len	Coverage (%)	Tax_ID	Organism
scaffold16634	1	779	756	958	78.91	1148	<i>Synechocystis sp.</i>
scaffold19397	129	583	455	908	50.11	111780	<i>Stanieria cyanosphaera</i>
scaffold11890	102	394	293	523	56.02	1287681	<i>Eutypa lata</i>

16  
17 **Only one gene MamblycephalaGene23156.1 is involved in the contamination region. We have**  
18 **now shielded them and updated the gene and genome data in the indicated database**  
19 **(ftp://user28@climb.genomics.cn).**  
20

21  
22 4) Consider swapping out some of the language of "plays a role" with phrases like "are involved  
23 in" or "have reported to" or "have a role in" etc.;

24 **Author response: We have replaced the expression of "plays a role" by "be involved in" or "be**  
25 **associated with" throughout the whole manuscript.**  
26

27 5) On Line 241, you have written bittern, I suspect you either meant bitter or bitterness.

28 **Author response: This was a language error on the out side and it should read bitterness. We now**  
29 **have amended it in the revised manuscript.**  
30

31  
32 6) I downloaded and examined the annotation file Megalobrama\_amblycephala.gff. It only  
33 contains transcript models and no gene models. You state in the paper there were 23,696 protein  
34 coding genes. This number corresponds to the number of mRNA models you have in this gff file.  
35 This struck me as odd as there are usually isoforms for gene models. This file also did not contain  
36 any functional annotation. Please create a gff3 file that contains genes models and include  
37 functional annotation as this will be of great value to researchers.  
38

39 **Author response: We have created a gff3 file that contains genes models and functional**  
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 information and alpha diversity indexes are shown in Table S15 in the revised Additional file 1.  
2 We have added the experimental set up and analysis of the gut microbial study in the method  
3 section of the revised manuscript.  
4  
5

6 2) The third part studies the development of intermuscular bones comprising the analysis of  
7 “expanded gene families” (again figure 2B does not explain the “expanded gene families”) and  
8 transcriptome analysis of early developmental stages.  
9

10 Author response: We apologize for the mistake. Reference should have been made to Figure 2C  
11 and not to Figure 2B. This has been corrected in the revised manuscript.  
12  
13

14 3) Authors present a huge amount of data and analysis but finally do describe and discuss only a  
15 little part of their analysis mainly in form of a very small set of genes. The manuscript is well, but  
16 not straightforward written. The structure has to be revised and the outcome better worked out.  
17  
18

19 Author response: We have revised our manuscript according to your suggestions and modified and  
20 renumbered the figures and tables clearly.  
21  
22

## 23 INTRODUCTION

24 1) LINE 64: Today many genomes (draft and nearly complete) are available. It is suggested to  
25 categorize them into fresh water, Mediterranean and Atlantic sea important aquacultured species.  
26 One species e.g. important for the Mediterranean aquaculture, The European sea bass  
27 (Dicentrarchus labrax) which was recently sequenced is not listed (European sea bass genome and  
28 its variation provide insights into adaptation to euryhalinity and speciation. Nature  
29 communications, 2014 5, 5770.)  
30  
31

32 Author response: This important species, the European sea bass, has been listed now in the revised  
33 manuscript (Line 65).  
34  
35

36 2) LINE 68-71: Please re-phrase the sentence. It is not true that the focus of aquaculture is  
37 focusing on herbivorous species. It is true however, that the usage of alternative feed is pursued  
38 but not the culture of herbivorous species. In contrast several projects are working on new species  
39 (including carnivores) for aquaculture purposes.  
40  
41

42 Author response: This sentence has been re-phrased as “Reports on draft genomes of some  
43 resource friendly herbivorous and omnivorous species, in particular cyprinid fish are scarce” in the  
44 revised manuscript (Line 69-70).  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
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

3) LINE 79 and LINE 84-85: Please refer to Wan et al., 2016 “Dynamic mRNA and miRNA expression analysis in response to intermuscular bone development of blunt snout bream (*Megalobrama amblycephala*)”, Scientific Reports.

Author response: We have added this research paper as a reference in the revised manuscript (Line 80-83 and Line 88-89).

## DATA DESCRIPTION

1) Description of generated SNP linkage map appears here, but does not appear in the method section. Please accomplish the method section.

Author response: We have added such information in detail now in Data Description section (Line 128-133) and also added the description of the genetic map construction in the method section (Line 403-416) in the revised manuscript.

## ANALYSES SECTION

1) According to GigaScience Authors guide: “This section should provide details of all of the experiments and analyses that are required to support the conclusions of the paper. The authors should make clear the goal of each analysis and state the basic findings”. Information about analyses (except the linkage map analysis) is partially found in the section named “Results” as well as in the “Method” section.

Author response: We have now modified this section according to your suggestions. We stated the goal and main finding of each analysis in the Results section and also removed the reporting of findings from the Method section.

2) LINES147-153: It is not clear why the enrichment analyses shows that the adaptation to herbivory goes “hand in hand” with the coping with plant secondary metabolites. Do authors have a comparable analysis of a carnivore or omnivore teleosts?

Author response: This sentence was ambiguous and we apologize for the confusion that it generated. We have now changed this expression to “These genes encoding proteins involved in biodegradation of xenobiotics would enhance the ability of an herbivore to detoxify the secondary compounds present in grasses that are adverse or even toxic to the organism.” (Line 161-164).

3) LINE 170: What do authors mean by the expression “comparative transcriptome analysis”?

1 Author response: This expression was not accurate. We have now modified the expression and  
2 write “differential genes expression (DGE) analysis” or “transcriptome analysis” at the appropriate  
3 positions of the manuscript.  
4

5  
6 4) LINE 174: “notable”. This result is not really surprising. Many other transcriptome studies in  
7 teleost have also shown that at the later stages, where the larvae is mainly growing, mostly muscle  
8 genes are up-regulated when compared to earlier stages.  
9

10 Author response: We agree. This word has been deleted in the revised manuscript.  
11

12 5) LINE 183: change the expression “eventually extended”.  
13

14 Author response: This expression has been changed to “we performed differential genes  
15 expression (DGE) analysis of” in the revised manuscript (Line 192).  
16

17 6) LINE 203. The link to Figure 3D is not clear. Is Figure 3D showing all 35 identified genes?  
18

19 Author response: Figure 3D shows several developmental signals regulating key steps of  
20 osteoblast and osteoclast differentiation in the process of intramembranous ossification. In Figure  
21 3D, only the colored boxes indicate these 35 identified genes. To make this more clear, we have  
22 amended this sentence to “35 bone formation regulatory genes were identified (shown in colored  
23 boxes in Figure 3D)” (Line 212).  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33

## 34 **DISCUSSION**

35 Mainly a repetition of the previous paragraphs.  
36

37 1) LINE 134 and 271: Authors did not show that Dre10 and Dre22 are ‘ancestral’, just that those  
38 two chromosomes fused to one chromosome in blunt snout bream as this species has 24  
39 chromosomes while zebrafish has 25 chromosomes. Please take into account publications like  
40 Nakatani, Y., H. Takeda, Y. Kohara, and S. Morishita, 2007 Reconstruction of the vertebrate  
41 ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17:  
42 1254–1265 and. Hufton, A. L., D. Groth, M. Vingron, H. Lehrach, A. J. Poustka et al., 2008 Early  
43 vertebrate whole genome duplications were predated by a period of intense genome rearrangement.  
44 *Genome Res.* 18: 1582–1591.  
45  
46  
47  
48  
49  
50  
51  
52  
53

54 Author response: According to the referee’s suggestion and taking into account the two references,  
55 we have modified this sentence to “The most prominent event is a chromosomal fusion in *M.*  
56 *amblycephala* LG02 that joined two *D. rerio* chromosomes, Dre10 and Dre22”.  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 2) LINE 311: ‘comparative transcriptome’ This expression leads the reader to the false impression  
2 that more than one species was studied. However authors investigated here in differential  
3 expression.  
4  
5

6 **Author response:** The expression “comparative transcriptome” has been amended to “differential  
7 genes expression (DGE) analysis” or “transcriptome analysis” in the appropriate positions of the  
8 manuscript.  
9  
10

## 11 **METHODS**

12  
13  
14  
15  
16 1) Missing description of microbial community study as well as generation of linkage map.—

17  
18 **Author response:** We have now added more detailed information in the Methods section including  
19 genome assembly, construction of gene families, the microbiota analysis, genetic map construction  
20 and other related information.  
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<sup>1†</sup>, Chunhai Chen<sup>2†</sup>, Zexia Gao<sup>1†</sup>, Jiumeng Min<sup>2†</sup>, Yongming Gu<sup>3†</sup>, Jianbo Jian<sup>2†</sup>, Xiewu  
7  
8           4           Jiang<sup>3</sup>, Huimin Cai<sup>2</sup>, Ingo Ebersberger<sup>4</sup>, Meng Xu<sup>2</sup>, Xinhui Zhang<sup>1</sup>, Jianwei Chen<sup>2</sup>, Wei Luo<sup>1</sup>,  
9  
10           5           Boxiang Chen<sup>1,3</sup>, Junhui Chen<sup>2</sup>, Hong Liu<sup>1</sup>, Jiang Li<sup>2</sup>, Ruifang Lai<sup>1</sup>, Mingzhou Bai<sup>2</sup>, Jin Wei<sup>1</sup>,  
11  
12           6           Shaokui Yi<sup>1</sup>, Huanling Wang<sup>1</sup>, Xiaojuan Cao<sup>1</sup>, Xiaoyun Zhou<sup>1</sup>, Yuhua Zhao<sup>1</sup>, Kaijian Wei<sup>1</sup>,  
13  
14           7           Ruibin Yang<sup>1</sup>, Bingnan Liu<sup>3</sup>, Shancen Zhao<sup>2</sup>, Xiaodong Fang<sup>2</sup>, Manfred Schartl<sup>5,\*</sup>, Xueqiao  
15  
16           8           Qian<sup>3,\*</sup>, Weimin Wang<sup>1,\*</sup>

17  
18  
19           9  
20  
21           10           \*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           <sup>1</sup>College of Fisheries, Key Lab of Freshwater Animal Breeding, Ministry of Agriculture, Key Lab  
28  
29           14           of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong  
30  
31           15           Agricultural University, Wuhan 430070, China

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

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

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

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

1  
2 59 Fishery and aquaculture play an important role in global alimentation. Over the past decades food  
3  
4 60 fish supply has been increasing with an annual rate of 3.6 percent, about 2 times faster than the  
5  
6 61 human population [1]. This growth of fish production is meanwhile solely accomplished by an  
7  
8 62 extension of aquaculture, as over the past thirty years the total mass of captured fish has remained  
9  
10 63 almost constant [1]. As a consequence of this emphasis on fish breeding, the genomes of various  
11  
12 64 economically important fish species, e.g. Atlantic cod (*Gadus morhua*) [2], rainbow trout  
13  
14 65 (*Oncorhynchus mykiss*) [3], European sea bass (*Dicentrarchus labrax*) [4], yellow croaker  
15  
16 66 (*Larimichthys crocea*) [5], half-smooth tongue sole (*Cynoglossus semilaevis*) [6], tilapia  
17  
18 67 (*Oreochromis niloticus*) [7] and channel catfish (*Ictalurus punctatus*) [8] have been sequenced.  
19  
20 68 Yet, the majority of these species are carnivorous requiring large inputs of protein from wild  
21  
22 69 caught fish or other precious feed. Reports on draft genomes of some resource friendly  
23  
24 70 herbivorous and omnivorous species, in particular cyprinid fish are scarce. It is well known that  
25  
26 71 cyprinids are currently the economically most important group of teleosts for sustainable  
27  
28 72 aquaculture. They grow to large population sizes in the wild and already now account for the  
29  
30 73 majority of freshwater aquaculture production worldwide [1]. Among these, the herbivorous  
31  
32 74 *Megalobrama amblycephala* (Yih, 1955), a particularly eco-friendly and resource-conserving  
33  
34 75 species, is predominant in aquaculture and has been greatly developed in China (Additional file 1:  
35  
36 76 Figure S1) [1]. However, most cyprinids, including *M. amblycephala*, have a large number of  
37  
38 77 intermuscular bones (IBs) in the trunk musculature, which have an adverse effect on fish meat  
39  
40 78 processing and consumption. IBs—a unique form of bone occurring only in the more basal  
41  
42 79 teleosts—are completely embedded within the myosepta and are not connected to the vertebral  
43  
44 80 column or any other bones [9, 10]. Our previous study on IB development of *M. amblycephala*  
45  
46 81 revealed that some miRNA-mRNA interaction pairs may be involved in regulating bone  
47  
48 82 development and differentiation [11]. However, the molecular genetic basis and the evolution of  
49  
50 83 this unique structures remain obscure. Unfortunately, the recent sequencing of two cyprinid  
51  
52 84 genomes common carp (*Cyprinus carpio*) [12] and grass carp (*Ctenopharyngodon idellus*) [13],  
53  
54 85 which provided valuable information for their genetic breeding, contributed little to the  
55  
56 86 understanding of IB formation.  
57  
58  
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

87 In an initial genome survey of *M. amblycephala*, we identified 25,697 single-nucleotide  
88 polymorphism (SNP) [14], 347 conserved miRNAs [15], and many miRNA-mRNA interaction  
89 pairs [11]. However, lack of a whole genome sequence resource limited a thorough investigation  
90 of *M. amblycephala*. Here we report the first high-quality draft genome sequence of *M.*  
91 *amblycephala*. Integrating this novel genome resource with tissue- and developmental  
92 stage-specific gene expression information, as well as with meta-genome data to investigate the  
93 composition of the gut microbiome provides relevant insights into the function and evolution of  
94 two key features characterizing this species: The formation of IB and the adaptation to herbivory.  
95 By that our study lays the foundation for genetically optimizing *M. amblycephala* to further  
96 increase its relevance for securing human food 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 S2). 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  
112 *Cyprinus carpio* and *Danio rerio* (Additional file 1: Figures S3 and S4). Using a comprehensive  
113 annotation strategy combining RNA-seq derived transcript evidence, *de-novo* gene prediction and  
114 sequence similarity to proteins from five further fish species, we annotated a total of 23,696  
115 protein-coding genes (Additional file 1: Table S6). Of the predicted genes, 99.44% (23,563 genes)

116 are annotated by functional database. In addition, we identified 1,796 non-coding RNAs including  
117 474 miRNAs, 220 rRNA, 530 tRNAs, and 572 snRNAs. Transposable elements (TEs) comprise  
118 approximately 34.18% (381.3 Mb) of the *M. amblycephala* genome (Additional file 1: Table S7).  
119 DNA transposons (23.80%) and long terminal repeat retrotransposons (LTRs) (9.89%) are the  
120 most 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 only 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 zebrafish (*Danio rerio*). The unexpected higher number of genes in syntenic regions  
142 shared with the more distantly related *D. rerio* is most likely an effect of the more complete  
143 genome assembly of this species compared to *C. idellus*. The rearrangement events are distributed  
144 across all *M. amblycephala* linkage groups without evidence for a local clustering (Figure 1B).

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

145 The most prominent event is a chromosomal fusion in *M. amblycephala* LG02 that joined two *D.*  
146 *rerio* chromosomes, Dre10 and Dre22. The same fusion is observed in *C. idellus* but not in *C.*  
147 *carpio* suggesting that it probably occurred in a last common ancestor of *M. amblycephala* and *C.*  
148 *idellus*, 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 properties of expanded gene families in the  
156 *M. amblycephala* and *C. idellus* lineage (Additional file 1: Figure S9, Additional file 2: Data  
157 Note1), two typical herbivores mainly feeding on aquatic and terrestrial grasses. Among the  
158 significantly over-represented KEGG pathways (Fisher's exact test,  $P < 0.01$ ), we find olfactory  
159 transduction (ko04740), immune-related pathways (ko04090, ko04672, ko04612 and ko04621),  
160 lipid metabolic related process (ko00590, ko03320, ko00591, ko00565, ko00592 and ko04975), as  
161 well as xenobiotics biodegradation and metabolism (ko00625 and ko00363) (Figure S10). These  
162 genes encoding proteins involved in biodegradation of xenobiotics would enhance the ability of an  
163 herbivore to detoxify the secondary compounds present in grasses that are adverse or even toxic to  
164 the organism. Furthermore, the high-fiber but low-energy grass diet requires a highly effective  
165 intermediate metabolism that accelerates carbohydrate and lipid catabolism and conversion into  
166 energy to maintain physiological functions. Indeed, when tracing positively selected genes (PSG)  
167 in *M. amblycephala* and *C. idellus* (Additional file 3: Date Note2), we identified many candidates  
168 involved in starch and sucrose metabolism (ko00500), citrate cycle (ko00020) and other types of  
169 O-glycan biosynthesis (ko00514). Moreover, 20 genes encoding enzymes involved in lipid and  
170 carbohydrate metabolism appear positively selected in both fish species (Additional file 1: Table  
171 S10).

### 172 Development of Intermuscular Bones

173 To explain the genetic basis of IB, their formation and their function in cyprinids, we first  
174 analyzed the functional annotation of gene families that expanded in this lineage (Figure 2C).  
175 Interestingly, many of these gene families are involved in cell adhesion (GO: 0007155,  
176  $P=5.26E-32$ , 357 genes), myosin complex (GO:0016459,  $P=2.74E-08$ , 100 genes) and cell-matrix  
177 adhesion (GO:0007160,  $P=1.59E-21$ , 69 genes) (Figure 2C), which interact dynamically to  
178 mediate efficient cell motility, migration and muscle construction [19, 20].

179 As a second line of evidence, we performed transcriptome analyses of early developmental  
180 stages (stage1: whole larvae without IBs) and juvenile *M. amblycephala* (stage2: trunk muscle  
181 with partial IBs; stage3: trunk muscle with completed IBs) (Figure 3A). We found 249 genes  
182 significantly up-regulated in stages 2 and 3 (with IB) compared to stage 1 (no IB). Many of these  
183 genes belong to KEGG pathways involved in tight junction (ko04530), regulation of actin  
184 cytoskeleton (ko04810), cardiac muscle contraction (ko04260) and vascular smooth muscle  
185 contraction (ko04270) (Additional file 1: Figure S11). These genes are associated with cell  
186 motility and muscle contraction [19, 21, 22], which resembles the findings from the gene family  
187 expansion analysis. Specifically, some of these genes encoding proteins related to muscle  
188 contraction, including titin, troponin, myosin, actinin, calmodulin and other  $Ca^{2+}$  transporting  
189 ATPases (Figure 3A) point to a strong remodeling of the musculature compartment.

190 To confirm that the observed differences in gene expression are indeed linked to IB formation  
191 and function and are not simply due to the fact that different developmental stages were compared,  
192 we performed differential genes expression (DGE) analysis of muscle tissues, IB, and connective  
193 tissues from the same six months old individual of *M. amblycephala* (Figure 3B, Additional file 1:  
194 Figure S12). Among the genes that are significantly up-regulated in the IB samples many encode  
195 extracellular matrix (ECM) proteins (collagens and integrin-binding protein), Rho GTPase family  
196 (*RhoA*, *Rho GAP*, *Rac*, *Ras*), motor proteins (myosin, dynein, actin), and calcium channel  
197 regulation protein (Additional file 1: Figure S13 and Table S11). Interestingly, it has been  
198 demonstrated that ECM proteins bound to integrins influence cell migration by  
199 actomyosin-generated contractile forces [20, 23]. Rho GTPases, acting as molecular switches, also  
200 is involved in regulating the actin cytoskeleton and cell migration, which in turn initiates  
201 intracellular signaling and contributes to tissue repair and regeneration [24-26].

202 During development of *M. amblycephala*, the first IB appears in muscles of caudal vertebrae  
203 as early as 28 days post fertilization (dpf) when body length is 12.95 mm (Additional file 1: Figure  
204 S14). The system then develops and ossifies predominantly from posterior to anterior (Additional  
205 file 1: Figure S15). IBs are present throughout the body within two months (Additional file 1:  
206 Figure S16) and develop into multiple morphological types in adults (Additional file 1: Figure  
207 S17). The bone is formed directly without an intermediate cartilaginous stage (Additional file 1:  
208 Figures S18 and S19). We also found a large number of mature osteoblasts distributed at the edge  
209 of the bone matrix while some osteocytes were apparent in the center of the mineralized bone  
210 matrix (Additional file 1: Figures S20 and S21). These primary bone-forming cells predominantly  
211 regulate bone formation and function throughout life. Notably, among the genes up-regulated in  
212 IB, 35 bone formation regulatory genes were identified (shown in colored boxes in Figure 3D). In  
213 particular, genes involved in bone morphogenetic protein (BMP) signaling including *Bmp3*,  
214 *Smad8*, *Smad9*, and *Id2*, in fibroblast growth factor (FGF) signaling including *Fgf2*, *Fgfr1a*,  
215 *Fgfbp2*, *Col6a3*, and *Col4a5*, and in Ca<sup>2+</sup> channels including *Cacna1c*, *CaM*, *Creb5*, and *Nfatc*  
216 were highly expressed (>2-fold change) in IB (Additional file 1: Figure S22). It has been  
217 demonstrated that *Bmp*, *Fgf2*, and *Fgfr1* are involved in intramembranous bone development and  
218 affect the expression and activity of other osteogenesis related transcription factors [27, 28]. The  
219 calcium-sensitive transcription factor *NFATc1* together with *CREB* induces the expression of  
220 osteoclast-specific genes [29].

## 221 **Adaptation to Herbivorous Diet**

222 Next to the presence of IB, the herbivory of *M. amblycephala* is the second key feature in  
223 connection to the use of this species as aquatic livestock. Olfaction, the sense of smell, is crucial  
224 for animals to find food. The perception of smell is mediated by a large gene family of olfactory  
225 receptor (OR) genes. The ORs of teleosts are predominantly expressed in the main olfactory  
226 epithelium of the nasal cavity [30, 31] and can discriminate, like those of other vertebrates,  
227 different kinds of odor molecules. However, compared to mammals, e.g. humans having around  
228 400 ORs [32] the OR repertoires in teleosts are considerably small. They range from only about  
229 48 in *Fugu rubripes* up to 161 in *D. rerio* (Figure 4A). In the *M. amblycephala* genome, we



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

230 identified 179 functional olfactory receptor (OR) genes (Figure 4A), and based on the  
231 classification of Niimura [33], 158, 117 and 153 receptors for water-borne odorants were  
232 identified in *M. amblycephala*, *C. idellus* and *D. rerio*, respectively (Additional file 1: [Table S12](#)).  
233 Overall, these receptor repertoires are substantially larger than those of other and carnivorous  
234 teleosts (*G. morhua*, *C. semilaevis*, *O. latipes*, *X. maculatus*) (Additional file 1: [Figures S23 and](#)  
235 [S24](#), [Table S12](#)). This suggests that olfaction—probably for food choice—has a particularly  
236 important role in the cyprinid species. Previous studies have demonstrated that the beta type OR  
237 genes are present in both aquatic and terrestrial vertebrates, indicating that the corresponding  
238 receptors detect both water-soluble and airborne odorants [31, 33]. Intriguingly, we found a  
239 massive expansion of beta-type OR genes in the genomes of the herbivorous *M. amblycephala* and  
240 *C. idellus*, while very few exist in other teleosts (Figure 4B, Additional file 1: [Table S12](#)).

241 Taste is also an important factor in the development of dietary habits. Most animals can  
242 perceive five basic tastes, namely sourness, sweetness, bitterness, saltiness and umami [34].  
243 Interestingly, *TIR1*, the receptor gene necessary for sensing umami, has been lost in herbivorous  
244 *M. amblycephala* but is duplicated in carnivorous *G. morhua*, *C. semilaevis* and omnivorous *O.*  
245 *latipes* and *X. maculatus* ([Figures 4C and 4D](#), Additional file 1: [Figures S25-26](#) and [Table S13](#)). In  
246 contrast, *TIR2*, the receptor gene for sensing sweet, has been duplicated in herbivorous *M.*  
247 *amblycephala* and *C. idellus*, omnivorous *C. carpio* and *D. rerio*, while it has been lost in  
248 carnivorous *G. morhua* and *C. semilaevis* (Additional file 1: [Figure S27](#) and [Table S13](#)). Bitterness  
249 sensed by the *T2R* is particularly crucial for animals to protect them from poisonous compounds  
250 [35]. Probably in the course switching to a diet that contains a larger fraction of [bitterness](#)  
251 containing food, also the *T2R* gene family in *M. amblycephala*, *C. idellus*, *C. carpio* and *D. rerio*  
252 has been expanded (Additional file 1: [Figure S28](#)).

253 To obtain further insights into the genetic adaptation to herbivorous diet, we focused on  
254 further genes that might be associated with digestion. Genes that encode proteases (including  
255 pepsin, trypsin, cathepsin and chymotrypsin) and amylases (including alpha-amylase and  
256 glucoamylase) were identified in the genomes of *M. amblycephala*, carnivorous *C. semilaevis*, *G.*  
257 *morhua* and omnivorous *D. rerio*, *O. latipes* and *X. maculatus*, indicating that herbivorous *M.*  
258 *amblycephala* has a protease repertoire that is not substantially different from those of carnivorous

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

259 and omnivorous fishes (Additional file 1: [Table S14](#)). We did not identify any genes encoding  
260 potentially cellulose-degrading enzymes including endoglucanase, exoglucanase and  
261 beta-glucosidase in the genome of *M. amblycephala*, suggesting that utilization of the herbivorous  
262 diet may largely depend on the gut microbiome. To elucidate this further, we determined the  
263 composition of the gut microbial communities of juvenile, domestic, wild adult *M. amblycephala*  
264 and wild adult *C. idellus* using bacterial 16S rRNA sequencing. A total of 549,020 filtered high  
265 quality sequence reads from 12 samples were clustered at a similarity level of 97%. The resulting  
266 8,558 operational taxonomic units (OTUs) are dominated at phylum level by Proteobacteria,  
267 Fusobacteria, Bacteroidetes, Firmicutes and Actinobacteria (Additional file 1: [Table S15](#), Figure  
268 4E). Increasing the resolution to the genus level, the composition and relative abundance of the  
269 gut microbiota of wild adult *M. amblycephala* and *C. idellus* are still very similar (Additional file  
270 1: [Table S16](#)) and we could identify more than 7% cellulose-degrading bacteria (Additional file 1:  
271 [Table S17](#)). This indicates that indeed the gut microbiome **is crucial** in the digestion of plant  
272 material, and thus in the adaptation to herbivory.

## 273 **Discussion**

274 *M. amblycephala* is the economically most important species for freshwater aquaculture. In  
275 addition to its various superior properties, especially its herbivorous diet, it is also an excellent  
276 model to study IB formation. Here we make available draft genome of *M. amblycephala* with  
277 more than 70% of genome data anchored on 24 linkage groups. Comparative analyses of genome  
278 structure revealed high synteny with three other cyprinid fish and uncovered a chromosomal  
279 fusion event in *M. amblycephala* that joined **two *D. rerio* chromosomes** (Figure 1B), which  
280 supports the previous results in *C. idellus* [13] and also provides novel scientific insights into the  
281 evolution of chromosome fusion events in cyprinids.

282 The evolutionary trajectory analysis of *M. amblycephala* and other teleosts revealed that *M.*  
283 *amblycephala* has the closest relationship to *C. idellus* (Figure 2A). Both the species are  
284 herbivorous fish but which endogenous and exogenous factors affected their feeding habits and  
285 how they adapted to their herbivorous diet is not known. Olfaction and taste are crucial for  
286 animals to find food and to distinguish whether potential food is edible or harmful [31, 35]. 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  
287 search for genes encoding OR showed that herbivorous *M. amblycephala* and *C. idellus* have a  
288 large number of beta-type OR, while other omnivorous and carnivorous fish only have one or two.  
289 This might be attributed to their particular herbivorous diet consisting not only of aquatic grasses  
290 but also the duckweed and terrestrial grasses, which they ingest from the water surface. Previous  
291 studies have demonstrated that the receptor for umami is formed by the T1R1/T1R3 heterodimer,  
292 while T1R2/T1R3 senses sweet taste [36]. We found that the umami gene *T1R1* was lost in  
293 herbivorous *M. amblycephala* but duplicated in the carnivorous *G. morhua* and *C. semilaevis*  
294 (Figure 4C). The loss of the *T1R1* gene in *M. amblycephala* might exclude the expression of a  
295 functional umami taste receptor. Such situations in other organism, e.g. the Chinese panda, have  
296 previously been related to feeding specialization [37]. Interestingly, the sweetness receptor *T1R2*  
297 and bitter receptor *T2R* genes are expanded in the herbivorous fish but few or no copy was found  
298 in carnivorous fish. Collectively, these results not only indicate the genetic adaptation to  
299 herbivorous diet of *M. amblycephala*, but also provided a clear and comprehensive picture of  
300 adaptive evolutionary mechanisms of sensory systems in other fish species with different trophic  
301 specializations.

31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
302 Some insects such as *Tenebrio molitor* [38] and *Neotermes koshunensis* [39], and the mollusc  
303 *Corbicula japonica* [40] have genes encoding endogenous cellulose degradation-related enzymes.  
304 However, all so far analyzed herbivorous vertebrates lack these genes and always rely on their gut  
305 microbiome to digest food [37, 41]. In herbivorous *M. amblycephala* and *C. idellus*, we also did  
306 not find any homologues of digestive cellulase genes. Interestingly, our work on the composition  
307 of gut microbiota of the two fish species identifies more than 7% cellulose-degrading bacteria,  
308 suggesting that the cellulose degradation of herbivorous fish largely depend on their gut  
309 microbiome.

48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
310 IB has evolved several times during teleost evolution [9, 42]. The developmental mechanisms  
311 and ossification processes forming IBs are dramatically distinct from other bones such as ribs,  
312 skeleton, vertebrae or spines. These usually develop from cartilaginous bone and are derived from  
313 the mesenchymal cell population by endochondral ossification [27, 43]. However, IBs form  
314 directly by intramembranous ossification and differentiate from osteoblasts within connective  
315 tissue, forming segmental, serially homologous ossifications in the myosepta. Although various

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 methods of ossification of IB have been proposed, few experiments have been conducted to  
317 confirm the ossification process and little is known about the potential role of IB in teleosts. Based  
318 on our findings of over-represented functional properties of expanded gene families in cyprinid  
319 lineage (Figure 2C) and evidence from DGE of early developmental stages of IB formation  
320 (Figure 3A), we provide molecular evidence that IB might play significant roles not only in  
321 regulating muscle contraction but also in active remodeling at the bone-muscle interface and  
322 coordination of cellular events.

323 It has previously been found that some major developmental signals including BMP, FGF,  
324 WNT, together with calcium/calmodulin signaling [27, 44-46], are essential for regulating the  
325 differentiation and function of osteoblasts and osteocytes and for regulating the RANKL signaling  
326 pathway for osteoclasts [47] in intramembranous bone development. In agreement with this  
327 concept, our DGE analysis of muscle, IB and connective tissues uncovered that 35 bone formation  
328 regulatory genes involved in these signals were highly up-regulated in IB. Taken together, these  
329 results suggest that IB indeed undergoes an intramembranous ossification process, is regulated by  
330 bone-specific signaling pathways, and underlies a homeostasis of maintenance, repair and  
331 remodeling.

## 332 **Conclusions**

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

## 340 **Methods**

### 341 **Sampling and DNA Extraction**

342 DNA for genome sequencing was derived from a double haploid fish from the *M. amblycephala*  
343 genetic breeding center at Huazhong Agricultural University (Wuhan, Hubei, China). Fish blood

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 was collected from adult female fish caudal vein using sterile injectors with pre-added  
345 anticoagulant solutions following anesthetized with MS-222 and sterilization with 75% alcohol.  
346 Genomic DNA was extracted from the whole blood.

### 347 **Genomic Sequencing and Assembly**

348 Libraries with different insert sized inserts of 170 bp, 500 bp, 800 bp, 2 Kb, 5 Kb, 10 Kb and 20  
349 Kb were constructed from the genomic DNA at BGI-Shenzhen. The libraries were sequenced  
350 using a HiSeq2000 instrument. In total, 11 libraries, sequenced in 23 lanes were constructed. To  
351 obtain high quality data, we applied filtering criteria for the raw reads. As a result, 142.55 Gb of  
352 filtered data were used to complete the genome assembly using **SOAPdenovo\_V2.04** [16]. Only  
353 filtered data were used in the genome assembly. First, the short insert size library data were used  
354 to construct a de Bruijn graph. The tips, merged bubbles and connections with low coverage were  
355 removed before resolving the small repeats. Second, all high-quality reads were realigned with the  
356 contig sequences. The number of shared paired-end relationships between pairs of contigs was  
357 calculated and weighted with the rate of consistent and conflicting paired ends before constructing  
358 the scaffolds in a stepwise manner from the short-insert size paired ends to the long-insert size  
359 paired ends. Third, the gaps between the constructed scaffolds were composed mainly of repeats,  
360 which were masked during scaffold construction. These gaps were closed using the paired-end  
361 information to retrieve read pairs in which one end mapped to a unique contig and the other was  
362 located in the gap region. Subsequently, local assembly was conducted for these collected reads.

363 **To assess the genome assembly quality, approximately 42.82 Gb Illumina reads generated from**  
364 **short-insert size libraries were mapped onto the genome. Bwa0.5.9-r16 software [48] with default**  
365 **parameters was used to assess the mapping ratio and Soap coverage 2.27 was used to calculate the**  
366 **sequencing depth. We also assessed the accuracy of the genome assembly by Trinity [49],**  
367 **including number of ESTs and new mRNA reads from early stages of embryos and multiple**  
368 **tissues, by aligning the scaffolds to the assembled transcriptome sequences.**

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

373 **Genome Annotation**

374 The genome was searched for repetitive elements using Tandem Repeats Finder (version 4.04)  
375 [50]. TEs were identified using homology-based approaches. The Repbase (version 16.10) [51]  
376 database of known repeats and a *de novo* repeat library generated by RepeatModeler were used.  
377 This database was mapped using the software of RepeatMasker (version 3.3.0). Four types of  
378 non-coding RNAs (microRNAs, transfer RNAs, ribosomal RNAs and small nuclear RNAs) were  
379 also annotated using tRNAscan-SE (version 1.23) and the Rfam database45 (Release 9.1) [52].

380 For gene prediction, *de novo* gene prediction, homology-based methods and RNA-seq data  
381 were used to perform gene prediction. For the sequence similarity based prediction, *D. rerio*, *G.*  
382 *aculeatus*, *O. niloticus*, *O. latipes* and *G. morhua* protein sequences were downloaded from  
383 Ensembl (release 73) and were aligned to the *M. amblycephala* genome using TBLASTN. Then  
384 homologous genome sequences were aligned against the matching proteins using GeneWise [53]  
385 to define gene models. Augustus was employed to predict coding genes using appropriate  
386 parameters in *de novo* prediction. For the RNA-seq based prediction, we mapped transcriptome  
387 reads to the genome assembly using TopHat [54]. Then, we combined TopHat mapping results  
388 together and applied Cufflinks [55] to predict transcript structures. All predicted gene structures  
389 were integrated by GLEAN [56] (<http://sourceforge.net/projects/glean-gene/>) to obtain a  
390 consensus gene set. Gene functions were assigned to the translated protein-coding genes using  
391 Blastp tool, based on their highest match to proteins in the SwissProt and TrEMBL [57] databases  
392 (Uniprot release 2011-01). Motifs and domains in the protein-coding genes were determined by  
393 InterProScan (version 4.7) searches against six different protein databases: ProDom, PRINTS,  
394 Pfam, SMART, PANTHER and PROSITE. Gene Ontology [58] IDs for each gene were obtained  
395 from the corresponding InterPro entries. All genes were aligned against KEGG [59] (Release 58)  
396 database, and the pathway in which the gene might be involved was derived from the matched  
397 genes in KEGG. tRNA genes were *de novo* predicted by tRNAscan-SE software [60], with  
398 eukaryote parameters on the repeat pre-masked genome. The rRNA fragments were identified by  
399 aligning the rRNA sequences using BlastN at E-value 1e-5. The snRNA and miRNA were  
400 searched by the method of aligning and searching with INFERNAL (version 0.81) [61] against  
401 Rfam database (release 9.1).

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 **Genetic map construction**

403 To anchor the scaffolds into pseudo-chromosomes, 198 F1 population individuals were used to  
404 obtain the genetic map. Each of the individual genomic DNA was digested with the restriction  
405 endonuclease EcoR I, following the RAD-Seq protocol [62]. The SNP calling process was carried  
406 out using the SOAP bioinformatic pipeline. The RAD-based SNP calling was done by SOAPsnp  
407 software [63] after each individual's paired-end RAD reads was mapped onto the assembled  
408 reference genome with the alignment software SOAP2 [64]. The potential SNP markers were used  
409 for the linkage analysis if the following criteria were satisfied: for parents - sequencing depth  $\geq 8$   
410 and  $\leq 100$ , base quality  $\geq 25$ , copy number  $\leq 1.5$ ; for progeny - sequencing depth  $\geq 5$ , base quality  
411  $\geq 20$ , copy number  $\leq 1.5$ . If the markers were showing significantly distorted segregation ( $P$ -value  
412  $< 0.01$ ), they were excluded from the map construction. Linkage analysis was performed only for  
413 markers present in at least 80% of the genomes, using JoinMap 4.0 software with CP population  
414 type codes and applying the double pseudo-test cross strategy [65]. The linkage groups were  
415 formed at a logarithm of odds threshold of 6.0 and ordered using the regression mapping  
416 algorithm.

## 417 **Construction of gene families**

418 We identified gene families using TreeFam software [66] as follows: Blast was used to compare  
419 all the protein sequences from 13 species: *M. amblycephala*, *C. idellus*, *C. semilaevis*, *C. carpio*,  
420 *D. rerio*, *Callorhinchus milii*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *Oncorhynchus*  
421 *mykiss*, *O. niloticus*, *O. latipes*, *Fugu rubripes*, with the E-value threshold set as  $1e-7$ . In the next  
422 step, HSP segments of each protein pair were concatenated by Solar software. H-scores were  
423 computed based on Bit-scores and these were taken to evaluate the similarity among genes.  
424 Finally, gene families were obtained by clustering of homologous gene sequences using  
425 Hcluster\_sg (Version 0.5.0). Specific genes of *M. amblycephala* were those that did not cluster  
426 with other vertebrates that were chosen for gene family construction, and those that did not have  
427 homologs in the predicted gene repertoire of the compared genomes. If these genes had functional  
428 motifs, they were annotated by GO.

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

430 The coding sequences of single-copy gene families conserved among *M. amblycephala*, *C. idellus*,

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 *C. carpio*, *D. rerio*, *C. semilaevis*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *O. mykiss*, *O.*  
432 *niloticus*, *O. latipes*, *C. milii* and *Fugu rubripes* (Ensembl Gene v.77) were extracted and aligned  
433 with guidance from amino-acid alignments created by the MUSCLE program [67]. The individual  
434 sequence alignments were then concatenated to form one supermatrix. PhyML [68, 69] was  
435 applied to construct the phylogenetic tree under an HKY85+gamma model for nucleotide  
436 sequences. ALRT values were taken to assess the branch reliability in PhyML. The same set of  
437 codon sequences at position 2 was used for phylogenetic tree construction and estimation of the  
438 divergence time. The PAML mcmctree program (PAML version 4.5) [70, 71] was used to  
439 determine divergence times with the approximate likelihood calculation method and the  
440 'correlated molecular clock' and 'REV' substitution model.

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

442 Protein sequences of *M. amblycephala* and 11 other related species (Ensembl Gene v.77)) were  
443 used in BLAST searches to identify homologs. We identified gene families using CAFÉ [72],  
444 which employs a random birth and death model to study gene gains and losses in gene families  
445 across a user-specified phylogeny. The global parameter  $\lambda$ , which describes both the gene birth ( $\lambda$ )  
446 and death ( $\mu = -\lambda$ ) rate across all branches in the tree for all gene families, was estimated using  
447 maximum likelihood. A conditional *P*-value was calculated for each gene family, and families  
448 with conditional *P*-values less than the threshold (0.05) were considered as having notable gain or  
449 loss. We identified branches responsible for low overall *P*-values of significant families.

#### 450 **Detection of Positively Selected Genes**

451 We calculated Ka/Ks ratios for all single copy orthologs of *M. amblycephala* and *C. semilaevis*, *D.*  
452 *rerio*, *G. morhua*, *O. niloticus* and *C. carpio*. Alignment quality was essential for estimating  
453 positive selection. Thus, orthologous genes were first aligned by PRANK [73], which is  
454 considerably conservative for inferring positive selection. We used Gblocks [74] to remove  
455 ambiguously aligned blocks within PRANK alignments and employed 'codeml' in the PAML  
456 package with the free-ratio model to estimate Ka, Ks, and Ka/Ks ratios on different branches. The  
457 differences in mean Ka/Ks ratios for single-copy genes between *M. amblycephala* and each of the  
458 other species were compared using paired Wilcoxon rank sum tests. Genes that showed values of  
459 Ka/Ks higher than 1 along the branch leading to *M. amblycephala* were reanalyzed using 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

460 codon based branch-site tests implemented in PAML. The branch-site model allowed  $\omega$  to vary  
461 both among sites in the protein and across branches, and was used to detect episodic positive  
462 selection.

### 463 **Developmental process of intermuscular bone in *M. amblycephala***

464 To better understand the number and morphological types of IBs in adult *M. amblycephala*,  
465 specimens with a body length ranging from 15.5 to 20.5 cm were collected and each individual  
466 was wrapped in gauze and boiled. The fish body was divided into two sections: anterior (snout to  
467 cloaca) and posterior (cloaca to the base of caudal fin), and the length of each section was  
468 measured. The IBs were retrieved, counted, arranged in order and photographed with a digital  
469 camera. Fertilized *M. amblycephala* eggs were brought from hatching facilities at Freshwater Fish  
470 Genetics Breeding Center of Huazhong Agricultural University (Wuhan, Hubei, China) to our  
471 laboratory. *M. amblycephala* larvae were maintained in a re-circulating aquaculture system at  $23 \pm$   
472  $1^\circ\text{C}$  with a 14-hr photoperiod. To explore the early development of IBs, larvae at different stages  
473 from 15 to 40 dpf were collected and fixed in 4% paraformaldehyde and transferred to 70%  
474 ethanol for storage. Specimens were stained with alizarin red for bone following the method  
475 described by Dawson [75]. The appearance of red color was recorded as the appearance of IB  
476 because bone ossification is accompanied by the uptake of alizarin red, resulting in red staining of  
477 the mineralized bone matrix. Myosepta, either not yet ossified, or poorly ossified, are not visible  
478 with alizarin red staining. For histologic analysis, specimens were paraffin-embedded and  
479 sectioned following standard protocols. Sections were stained with hematoxylin and eosin (HE)  
480 and Masson trichrome [76] and photographed using a Nikon microscope (Nikon, Tokyo, Japan)  
481 with a DP70 digital camera (Olympus, Japan). Scanning electron microscopy (SEM) and  
482 transmission electron microscopy (TEM) were also conducted to analyze the ultrastructure of IB.  
483 The specimens were fixed with 2.5% (v/v) glutaraldehyde in a solution of 0.1 M sodium  
484 cacodylate buffer (pH 7.3) for 2 h at room temperature. The SEM and TEM samples were  
485 prepared according to a standard protocol described by Ott [77]. The samples were then visualized  
486 with a JSM-6390LV scanning electron microscope (SEM, Japan) and the stained ultrathin sections  
487 with a H-7650 transmission electron microscope (Hitachi, Japan).

### 488 **RNA Sequencing Analysis**

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 *M. amblycephala* specimens belonging to three different developmental stages of IBs (stage 1:  
490 whole larvae without distribution of IB; stage 2: muscle tissues with partial distribution of IBs; stage 3:  
491 muscle tissues with completed distribution of IBs were identified under microscope and immediately  
492 frozen in liquid nitrogen. In addition, dorsal white muscle, IBs and connective tissue surrounding the  
493 IBs from six months old fish were also collected. RNA was extracted from total fish samples at  
494 different stages and from individual muscle, connective tissue, and intermuscular bone samples of  
495 *M. amblycephala* using RNAisoPlus Reagent (TaKaRa, China) according to the manufacturer's  
496 protocol. The integrity and purity of the RNA was determined by gel electrophoresis and Agilent  
497 2100 BioAnalyzer (Agilent, USA) before preparing the libraries for sequencing. Paired-end RNA  
498 sequencing was performed using the Illumina HiSeq 2000 platform. Low quality score reads were  
499 filtered and the clean data were aligned to the reference genome using Bowtie [78]. Genes and  
500 isoforms expression level were quantified by a software package: RSEM (RNASeq by  
501 Expectation Maximization) [79]. Gene expression levels were calculated by using the RPKM  
502 method (Reads per kilobase transcriptome per million mapped reads) [80] and adjusted by a  
503 scaling normalization method [81]. DEGs were detected using DESeq [82]. Annotation of DEGs  
504 were mapped to GO categories in the database (<http://www.geneontology.org/>) and the number of  
505 genes for every term were calculated to identify GO terms that were significantly enriched in the  
506 input list of DEGs. The calculated *P*-value was adjusted by the Bonferroni Correction, taking  
507 corrected *P*-value  $\leq 0.05$  as a threshold. KEGG automatic annotation was used to perform  
508 pathway enrichment analysis of DEGs.

### 509 **Prediction of Olfactory Receptor Genes**

510 Olfactory receptor genes were identified by previously described methods [83], with the exception  
511 of a first-round TBLASTN [84] search, in which 1,417 functional olfactory receptor genes from *H.*  
512 *sapiens*, *D. rerio*, *L. chalumnae*, *Lepisosteus oculatus*, *L. vexillifer*, *O. niloticus*, *O. latipes*, *F.*  
513 *rubripes* and *Xenopus tropicalis* were used as queries. We then predicted the structure of  
514 sequenced genes using the blast-hit sequence with the software GeneWise extending in both 3' and  
515 5' directions along the genome sequences. The results were further confirmed by NR annotation.  
516 Then the coding sequences from the start (ATG) to stop codons were extracted, while sequences  
517 that contained interrupting stop codons or frame-shifts were regarded as pseudogenes. To

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 construct phylogenetic trees, the amino-acid sequences encoded by olfactory receptor genes were  
519 first aligned using the program MUSCLE nested in MEGA 5.10 [85]. We then constructed the  
520 phylogenetic tree using the neighbor-joining method with Poisson correction distances using the  
521 program MEGA 5.10. We also identified and compared the genes for five basic tastes (sour, sweet,  
522 bitter, umami and salty) using a similar method as in OR gene identification.

### 523 **Gut microbiota analysis**

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

### 537 **Additional files**

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

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

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

### 542 **Abbreviations**

543 IB, intermuscular bone; SNP, single-nucleotide polymorphism; BUSCO, benchmarking universal  
544 single-copy orthologs; TE, transposable element; LTR, long terminal repeat retrotransposon; LG,  
545 linkage group; PSG, positively selected gene; ECM, extracellular matrix; dpf, days post  
546 fertilization; BMP, bone morphogenetic protein; FGF, fibroblast growth factor; OR, olfactory

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

547 receptor; OTU, operational taxonomic unit; DGE, differential genes expression; HE, hematoxylin  
548 and eosin; SEM, scanning electron microscopy; TEM, transmission electron microscopy

## 549 **Acknowledgements**

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

## 555 **Availability of data and materials**

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

## 559 **Authors' contributions**

560 W.W. initiated and conceived the project and provided scientific input. X.Q. organized financial  
561 support and designed the project. M.S. discussed the data, wrote and modified the paper. H.L. and  
562 C.C. conducted the biological experiments, analyzed the data and wrote the paper with input from  
563 other authors. I.E. wrote and modified the manuscript and discussed the data. The RAD-Seq data  
564 analyses and the genetic map construction were performed by Z.G., Y.G., J.J. and X.J. Genome  
565 assembly and annotation were performed by J.M., H.C., M.X. and J.C. X.Z., W.L., R.L., B.C.,  
566 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  
567 and data collection. J.L. and J.C. identified the gene families and analyzed the RNA-seq data. M.B.  
568 coordinated the project. S.Z. and X.F. modified the manuscript and discussed the data. All authors  
569 read the manuscript and provided comments and suggestions for improvements. The authors  
570 declare no competing financial interests.

## 571 **Competing interests**

572 The authors declare that they have no competing interests.

## 573 **Ethics approval and consent to participate**

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

## 577 **References**

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

- 606 13. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, et al. The draft genome of the grass carp  
607 (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation.  
608 Nat. Genet. 2015; 47:625–31.
- 609 14. Gao Z, Luo W, Liu H, Zeng C, Liu X, Yi S, et al. Transcriptome analysis and SSR/SNP  
610 markers information of the blunt snout bream (*Megalobrama amblycephala*). PLoS One.  
611 2012; 7:e42637.
- 612 15. Yi S, Gao ZX, Zhao H, Zeng C, Luo W, Chen B, et al. Identification and characterization of  
613 microRNAs involved in growth of blunt snout bream (*Megalobrama amblycephala*) by  
614 Solexa sequencing. BMC Genomics. 2013; 14:754.
- 615 16. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved  
616 memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.
- 617 17. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing  
618 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.  
619 2015;31:3210–2.
- 620 18. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis  
621 of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61.
- 622 19. Even-Ram S, Doyle AD, Conti MA, Matsumoto K, Adelstein RS, Yamada KM. Myosin IIA  
623 regulates cell motility and actomyosin-microtubule crosstalk. Nat. Cell Biol. 2007;  
624 9:299–309.
- 625 20. Sheetz MP, Felsenfeld DP, Galbraith CG. Cell migration: Regulation of force on  
626 extracellular-complexes. Trends Cell Biol. 1998; 8:51–4.
- 627 21. Gunst SJ, Zhang W. Actin cytoskeletal dynamics in smooth muscle: a new paradigm for the  
628 regulation of smooth muscle contraction. Am. J. Physiol. Cell Physiol. 2008; 295:C576–87.
- 629 22. Webb RC. Smooth muscle contraction and relaxation. Adv. Physiol. Educ. 2003; 27:201–6.
- 630 23. Ulrich TA, De Juan Pardo EM, Kumar S. The mechanical rigidity of the extracellular matrix  
631 regulates the structure, motility, and proliferation of glioma cells. Cancer Res. 2009;  
632 69:4167–74.
- 633 24. Ridley AJ. Rho GTPases and cell migration. J. Cell Sci. 2001; 114:2713–22.
- 634 25. Etienne-Manneville S, Hall A. Rho GTPases in Cell Biology. Nature. 2002; 420:629–35.
- 635 26. Ridley AJ. Cell Migration: Integrating Signals from Front to Back. Science. 2003;  
636 302:1704–9.
- 637 27. Ornitz D, Marie P. FGF signaling pathways in endochondral and intramembranous bone  
638 development and human genetic disease. Genes Dev. 2002; 16:1446–65.

- 639 28. Fakhry A, Ratisoontorn C, Vedhachalam C, Salhab I, Koyama E, Leboy P, et al. Effects of  
640 FGF-2/-9 in calvarial bone cell cultures: Differentiation stage-dependent mitogenic effect,  
641 inverse regulation of BMP-2 and noggin, and enhancement of osteogenic potential. *Bone*.  
642 2005; 36:254–66.
- 643 29. Sato K, Suematsu A, Nakashima T, Takemoto-Kimura S, Aoki K, Morishita Y, et al.  
644 Regulation of osteoclast differentiation and function by the CaMK-CREB pathway. *Nat. Med.*  
645 2006; 12:1410–6.
- 646 30. Niimura Y, Nei M. Evolutionary dynamics of olfactory and other chemosensory receptor  
647 genes in vertebrates. *J. Hum. Genet.* 2006; 51:505–17.
- 648 31. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene  
649 repertoires: roles of chance and necessity. *Nat. Rev. Genet.* 2008; 9:951–63.
- 650 32. Lopez C, Raper J. Cloning and functional characterization of odorant receptors expressed in  
651 the zebrafish olfactory system. *FASEB J.* 2015; 29:727–37.
- 652 33. Niimura Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative  
653 genome analysis among 23 chordate species. *Genome Biol. Evol.* 2009; 1:34–44.
- 654 34. Lindemann B. Receptors and transduction in taste. *Nature.* 2001; 413:219–25.
- 655 35. Chandrashekar J, Mueller KL, Hoon MA, Adler E, Feng L, Guo W, et al. T2Rs function as  
656 bitter taste receptors. *Cell.* 2000; 100:703–11.
- 657 36. Nelson G, Chandrashekar J, Hoon MA, Feng L, Zhao G, Ryba NJ, et al. An amino-acid taste  
658 receptor. *Nature.* 2002; 416:199–202.
- 659 37. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and *de novo* assembly of the  
660 giant panda genome. *Nature.* 2010; 463:311–7.
- 661 38. Ferreira AHP, Marana SR, Terra WR, Ferreira C. Purification, molecular cloning, and  
662 properties of a  $\beta$ -glycosidase isolated from midgut lumen of *Tenebrio molitor* (Coleoptera)  
663 larvae. *Insect Biochem. Mol. Biol.* 2001; 31:1065–76.
- 664 39. Tokuda G, Saito H, Watanabe H. A digestive  $\beta$ -glucosidase from the salivary glands of the  
665 termite, *Neotermes koshunensis* (Shiraki): Distribution, characterization and isolation of its  
666 precursor cDNA by 5'- and 3'-RACE amplifications with degenerate primers. *Insect*  
667 *Biochem. Mol. Biol.* 2002; 32:1681–9.
- 668 40. Sakamoto K, Uji S, Kurokawa T, Toyohara H. Molecular cloning of endogenous  
669  $\beta$ -glucosidase from common Japanese brackish water clam *Corbicula japonica*. *Gene.* 2009;  
670 435:72–9.
- 671 41. Zhu L, Wu Q, Dai J, Zhang S, Wei F. Evidence of cellulose metabolism by the giant panda gut  
672 microbiome. *Proc. Natl. Acad. Sci. USA.* 2011; 108:17714–9.

- 673 42. Bird NC, Mabee PM. Developmental morphology of the axial skeleton of the zebrafish, *Danio*  
674 *rerio* (Ostariophysi: Cyprinidae). *Dev. Dyn.* 2003; 228:337–57.
- 675 43. Ortega N, Behonick DJ, Werb Z. Matrix remodeling during endochondral ossification. *Trends*  
676 *Cell Biol.* 2004; 14:86–93.
- 677 44. Chen G, Deng C, Li YP. TGF- $\beta$  and BMP signaling in osteoblast differentiation and bone  
678 formation. *Int. J. Biol. Sci.* 2012; 8:272–88.
- 679 45. Harada S, Rodan GA. Control of osteoblast function and regulation of bone mass. *Nature.*  
680 2003; 423:349–55.
- 681 46. Long F. Building strong bones: molecular regulation of the osteoblast lineage. *Nat. Rev. Mol.*  
682 *Cell Biol.* 2011; 13:27–38.
- 683 47. Boyle WJ, Simonet WS, Lacey DL. Osteoclast differentiation and activation. *Nature.* 2003;  
684 423:337–42.
- 685 48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.  
686 *Bioinformatics.* 2009; 25:1754–60.
- 687 49. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity:  
688 reconstructing a full-length transcriptome without a genome from RNA-Seq data . *Nat.*  
689 *Biotechnol.* 2011; 29:644–52.
- 690 50. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.*  
691 1999; 27:573–80.
- 692 51. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase  
693 Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 2005;  
694 110:462–7.
- 695 52. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: Annotating  
696 non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005; 33:121–4.
- 697 53. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res.* 2004; 14:988–95.
- 698 54. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq.  
699 *Bioinformatics.* 2009; 25:1105–11.
- 700 55. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript  
701 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform  
702 switching during cell differentiation. *Nat. Biotechnol.* 2010; 28:511–5.
- 703 56. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey  
704 bee consensus gene set. *Genome Biol.* 2007; 8:R13.



- 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 57. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement  
706 TrEMBL in 2000. *Nucleic Acids Res.* 2000; 28:45–8.
- 707 58. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool  
708 for the unification of biology. *Nat. Genet.* 2000; 25:25–9.
- 709 59. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*  
710 2000; 28:27-30.
- 711 60. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes  
712 in genomic sequence. *Nucleic Acids Res.* 1997; 25:955–64.
- 713 61. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: Inference of RNA alignments. *Bioinformatics.*  
714 2009; 25:1335–7.
- 715 62. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP  
716 discovery and genetic mapping using sequenced RAD markers. *PloS One.* 2008; 3:e3376.
- 717 63. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively  
718 parallel whole-genome resequencing. *Genome Res.* 2009; 19:1124–32.
- 719 64. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: An improved ultrafast tool  
720 for short read alignment. *Bioinformatics.* 2009; 25:1966–7.
- 721 65. Grattapaglia D, Sederoff R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus*  
722 *urophylla* using a pseudo-testcross: Mapping strategy and RAPD markers. *Genetics.* 1994;  
723 137:1121–37.
- 724 66. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, et al. TreeFam: a curated  
725 database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 2006;  
726 34:D572–80.
- 727 67. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.  
728 *Nucleic Acids Res.* 2004; 32:1792–7.
- 729 68. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and  
730 methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML  
731 3.0. *Syst. Biol.* 2010; 59:307–21.
- 732 69. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by  
733 maximum likelihood. *Syst. Biol.* 2003; 52:696–704.
- 734 70. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;  
735 24:1586–91.
- 736 71. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock  
737 using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* 2006; 23:212–26.

- 738 72. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics*.  
739 2007; 177:1941–9.
- 740 73. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with  
741 insertions. *Proc. Natl. Acad. Sci. USA*. 2005; 102:10557–62.
- 742 74. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and  
743 ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 2007; 56:564–77.
- 744 75. Dawson AB. A note on the staining of the skeleton of cleared specimens with alizarin red S.  
745 *Biotech. Histochem.* 1926; 1:123-4.
- 746 76. Gruber HE. Adaptations of Goldner’s Masson trichrome stain for the study of undecalcified  
747 plastic embedded bone. *Biotech. Histochem.* 1992; 67:30–4.
- 748 77. Ott HC, Matthiesen TS, Goh SK, Black LD, Kren SM, Netoff TI, et al.  
749 Perfusion-decellularized matrix: using nature’s platform to engineer a bioartificial heart. *Nat.*  
750 *Med.* 2008; 14:213–21.
- 751 78. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 2012;  
752 9:357–9.
- 753 79. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or  
754 without a reference genome. *BMC Bioinformatics.* 2011; 12:323.
- 755 80. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying  
756 mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 2008; 5:621–8.
- 757 81. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis  
758 of RNA-seq data. *Genome Biol.* 2010; 11:R25.
- 759 82. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.*  
760 2010; 11:R106.
- 761 83. Niimura Y. Evolutionary dynamics of olfactory receptor genes in chordates: interaction  
762 between environments and genomic contents. *Hum. Genomics.* 2009; 4:107–18.
- 763 84. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and  
764 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;  
765 25:3389–402.
- 766 85. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary  
767 analysis of DNA and protein sequences. *Brief. Bioinform.* 2008; 9:299–306.
- 768 86. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al.  
769 Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc.*  
770 *Natl. Acad. Sci. USA.* 2011; 108:4516–22.

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

771 87. Edgar, RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat.  
772 Methods. 2013; 10:996–8.  
773 88. Liu H, Chen CH, Gao ZX, Min JM, Gu YM, Jian JB, et al. The draft genome of *Megalobrama*  
774 *amblycephala* reveals the development of intermuscular bone and adaptation to herbivorous  
775 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

776 **Figure Legends**

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

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

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

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

829 **Table**

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

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

831

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

4  
5  
6  
7           3    Han Liu<sup>1†</sup>, Chunhai Chen<sup>2†</sup>, Zexia Gao<sup>1†</sup>, Jiumeng Min<sup>2†</sup>, Yongming Gu<sup>3†</sup>, Jianbo Jian<sup>2†</sup>, Xiewu  
8  
9           4    Jiang<sup>3</sup>, Huimin Cai<sup>2</sup>, Ingo Ebersberger<sup>4</sup>, Meng Xu<sup>2</sup>, Xinhui Zhang<sup>1</sup>, Jianwei Chen<sup>2</sup>, Wei Luo<sup>1</sup>,  
10  
11          5    Boxiang Chen<sup>1,3</sup>, Junhui Chen<sup>2</sup>, Hong Liu<sup>1</sup>, Jiang Li<sup>2</sup>, Ruifang Lai<sup>1</sup>, Mingzhou Bai<sup>2</sup>, Jin Wei<sup>1</sup>,  
12  
13          6    Shaokui Yi<sup>1</sup>, Huanling Wang<sup>1</sup>, Xiaojuan Cao<sup>1</sup>, Xiaoyun Zhou<sup>1</sup>, Yuhua Zhao<sup>1</sup>, Kaijian Wei<sup>1</sup>,  
14  
15          7    Ruibin Yang<sup>1</sup>, Bingnan Liu<sup>3</sup>, Shancen Zhao<sup>2</sup>, Xiaodong Fang<sup>2</sup>, Manfred Schartl<sup>5,\*</sup>, Xueqiao  
16  
17          8    Qian<sup>3,\*</sup>, Weimin Wang<sup>1,\*</sup>

19  
20  
21  
22          10   \*Equally contributing corresponding authors: wangwm@mail.hzau.edu.cn; xueqiaoqian@263.net;  
23  
24          11   phch1@biozentrum.uni-wuerzburg.de

25  
26          12   †Equal contributors

27  
28          13   <sup>1</sup>College of Fisheries, Key Lab of Freshwater Animal Breeding, Ministry of Agriculture, Key Lab  
29  
30          14   of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong  
31  
32          15   Agricultural University, Wuhan 430070, China

33  
34          16   <sup>2</sup>Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen 518083, China

35  
36          17   <sup>3</sup>Guangdong Haid Group Co., Ltd., Guangzhou 511400, China

37  
38          18   <sup>4</sup>Dept. for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University,  
39  
40          19   Frankfurt D-60438, Germany

41  
42          20   <sup>5</sup>Physiological Chemistry, University of Würzburg, Biozentrum, Am Hubland, and Comprehensive  
43  
44          21   Cancer Center Mainfranken, University Clinic Würzburg, Würzburg 97070, Germany and Texas  
45  
46          22   A&M Institute for Advanced Study and Department of Biology, Texas A&M University, College  
47  
48          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 some resource friendly  
70 herbivorous and omnivorous species, in particular cyprinid fish are scarce. It is well known that  
71 cyprinids are currently the economically most important group of teleosts for sustainable  
72 aquaculture. They grow to large population sizes in the wild and already now account for the  
73 majority of freshwater aquaculture production worldwide [1]. Among these, the herbivorous  
74 *Megalobrama amblycephala* (Yih, 1955), a particularly eco-friendly and resource-conserving  
75 species, is predominant in aquaculture and has been greatly developed in China (Additional file 1:  
76 Figure S1) [1]. However, most cyprinids, including *M. amblycephala*, have a large number of  
77 intermuscular bones (IBs) in the trunk musculature, which have an adverse effect on fish meat  
78 processing and consumption. IBs—a unique form of bone occurring only in the more basal  
79 teleosts—are completely embedded within the myosepta and are not connected to the vertebral  
80 column or any other bones [9, 10]. Our previous study on IB development of *M. amblycephala*  
81 revealed that some miRNA-mRNA interaction pairs may be involved in regulating bone  
82 development and differentiation [11]. However, the molecular genetic basis and the evolution of  
83 this unique structures remain obscure. Unfortunately, the recent sequencing of two cyprinid  
84 genomes common carp (*Cyprinus carpio*) [12] and grass carp (*Ctenopharyngodon idellus*) [13],  
85 which provided valuable information for their genetic breeding, contributed little to the  
86 understanding of IB formation.

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 In an initial genome survey of *M. amblycephala*, we identified 25,697 single-nucleotide  
88 polymorphism (SNP) [14], 347 conserved miRNAs [15], and many miRNA-mRNA interaction  
89 pairs [11]. However, lack of a whole genome sequence resource limited a thorough investigation  
90 of *M. amblycephala*. Here we report the first high-quality draft genome sequence of *M.*  
91 *amblycephala*. Integrating this novel genome resource with tissue- and developmental  
92 stage-specific gene expression information, as well as with meta-genome data to investigate the  
93 composition of the gut microbiome provides relevant insights into the function and evolution of  
94 two key features characterizing this species: The formation of IB and the adaptation to herbivory.  
95 By that our study lays the foundation for genetically optimizing *M. amblycephala* to further  
96 increase its relevance for securing human food 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 S2). 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  
112 *Cyprinus carpio* and *Danio rerio* (Additional file 1: Figures S3 and S4). Using a comprehensive  
113 annotation strategy combining RNA-seq derived transcript evidence, *de-novo* gene prediction and  
114 sequence similarity to proteins from five further fish species, we annotated a total of 23,696  
115 protein-coding genes (Additional file 1: Table S6). Of the predicted genes, 99.44% (23,563 genes)

116 are annotated by functional database. In addition, we identified 1,796 non-coding RNAs including  
117 474 miRNAs, 220 rRNA, 530 tRNAs, and 572 snRNAs. Transposable elements (TEs) comprise  
118 approximately 34.18% (381.3 Mb) of the *M. amblycephala* genome (Additional file 1: Table S7).  
119 DNA transposons (23.80%) and long terminal repeat retrotransposons (LTRs) (9.89%) are the  
120 most 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 only 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 zebrafish (*Danio rerio*). The unexpected higher number of genes in syntenic regions  
142 shared with the more distantly related *D. rerio* is most likely an effect of the more complete  
143 genome assembly of this species compared to *C. idellus*. The rearrangement events are distributed  
144 across all *M. amblycephala* linkage groups without evidence for a local clustering (Figure 1B).

145 The most prominent event is a chromosomal fusion in *M. amblycephala* LG02 that joined two *D.*  
146 *rerio* chromosomes, Dre10 and Dre22. The same fusion is observed in *C. idellus* but not in *C.*  
147 *carpio* suggesting that it probably occurred in a last common ancestor of *M. amblycephala* and *C.*  
148 *idellus*, 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 (*Callorhinchus 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 properties of expanded gene families in the  
156 *M. amblycephala* and *C. idellus* lineage (Additional file 1: Figure S9, Additional file 2: Data  
157 Note1), two typical herbivores mainly feeding on aquatic and terrestrial grasses. Among the  
158 significantly over-represented KEGG pathways (Fisher's exact test,  $P < 0.01$ ), we find olfactory  
159 transduction (ko04740), immune-related pathways (ko04090, ko04672, ko04612 and ko04621),  
160 lipid metabolic related process (ko00590, ko03320, ko00591, ko00565, ko00592 and ko04975), as  
161 well as xenobiotics biodegradation and metabolism (ko00625 and ko00363) (Figure S10). These  
162 genes encoding proteins involved in biodegradation of xenobiotics would enhance the ability of an  
163 herbivore to detoxify the secondary compounds present in grasses that are adverse or even toxic to  
164 the organism. Furthermore, the high-fiber but low-energy grass diet requires a highly effective  
165 intermediate metabolism that accelerates carbohydrate and lipid catabolism and conversion into  
166 energy to maintain physiological functions. Indeed, when tracing positively selected genes (PSG)  
167 in *M. amblycephala* and *C. idellus* (Additional file 3: Date Note2), we identified many candidates  
168 involved in starch and sucrose metabolism (ko00500), citrate cycle (ko00020) and other types of  
169 O-glycan biosynthesis (ko00514). Moreover, 20 genes encoding enzymes involved in lipid and  
170 carbohydrate metabolism appear positively selected in both fish species (Additional file 1: Table  
171 S10).

### 172 **Development of Intermuscular Bones**

173 To explain the genetic basis of IB, their formation and their function in cyprinids, we first  
174 analyzed the functional annotation of gene families that expanded in this lineage (Figure 2C).  
175 Interestingly, many of these gene families are involved in cell adhesion (GO: 0007155,  
176  $P=5.26E-32$ , 357 genes), myosin complex (GO:0016459,  $P=2.74E-08$ , 100 genes) and cell-matrix  
177 adhesion (GO:0007160,  $P=1.59E-21$ , 69 genes) (Figure 2C), which interact dynamically to  
178 mediate efficient cell motility, migration and muscle construction [19, 20].

179 As a second line of evidence, we performed transcriptome analyses of early developmental  
180 stages (stage1: whole larvae without IBs) and juvenile *M. amblycephala* (stage2: trunk muscle  
181 with partial IBs; stage3: trunk muscle with completed IBs) (Figure 3A). We found 249 genes  
182 significantly up-regulated in stages 2 and 3 (with IB) compared to stage 1 (no IB). Many of these  
183 genes belong to KEGG pathways involved in tight junction (ko04530), regulation of actin  
184 cytoskeleton (ko04810), cardiac muscle contraction (ko04260) and vascular smooth muscle  
185 contraction (ko04270) (Additional file 1: Figure S11). These genes are associated with cell  
186 motility and muscle contraction [19, 21, 22], which resembles the findings from the gene family  
187 expansion analysis. Specifically, some of these genes encoding proteins related to muscle  
188 contraction, including titin, troponin, myosin, actinin, calmodulin and other  $Ca^{2+}$  transporting  
189 ATPases (Figure 3A) point to a strong remodeling of the musculature compartment.

190 To confirm that the observed differences in gene expression are indeed linked to IB formation  
191 and function and are not simply due to the fact that different developmental stages were compared,  
192 we performed differential genes expression (DGE) analysis of muscle tissues, IB, and connective  
193 tissues from the same six months old individual of *M. amblycephala* (Figure 3B, Additional file 1:  
194 Figure S12). Among the genes that are significantly up-regulated in the IB samples many encode  
195 extracellular matrix (ECM) proteins (collagens and integrin-binding protein), Rho GTPase family  
196 (*RhoA*, *Rho GAP*, *Rac*, *Ras*), motor proteins (myosin, dynein, actin), and calcium channel  
197 regulation protein (Additional file 1: Figure S13 and Table S11). Interestingly, it has been  
198 demonstrated that ECM proteins bound to integrins influence cell migration by  
199 actomyosin-generated contractile forces [20, 23]. Rho GTPases, acting as molecular switches, also  
200 is involved in regulating the actin cytoskeleton and cell migration, which in turn initiates  
201 intracellular signaling and contributes to tissue repair and regeneration [24-26].

202 During development of *M. amblycephala*, the first IB appears in muscles of caudal vertebrae  
203 as early as 28 days post fertilization (dpf) when body length is 12.95 mm (Additional file 1: Figure  
204 S14). The system then develops and ossifies predominantly from posterior to anterior (Additional  
205 file 1: Figure S15). IBs are present throughout the body within two months (Additional file 1:  
206 Figure S16) and develop into multiple morphological types in adults (Additional file 1: Figure  
207 S17). The bone is formed directly without an intermediate cartilaginous stage (Additional file 1:  
208 Figures S18 and S19). We also found a large number of mature osteoblasts distributed at the edge  
209 of the bone matrix while some osteocytes were apparent in the center of the mineralized bone  
210 matrix (Additional file 1: Figures S20 and S21). These primary bone-forming cells predominantly  
211 regulate bone formation and function throughout life. Notably, among the genes up-regulated in  
212 IB, 35 bone formation regulatory genes were identified (shown in colored boxes in Figure 3D). In  
213 particular, genes involved in bone morphogenetic protein (BMP) signaling including *Bmp3*,  
214 *Smad8*, *Smad9*, and *Id2*, in fibroblast growth factor (FGF) signaling including *Fgf2*, *Fgfr1a*,  
215 *Fgfbp2*, *Col6a3*, and *Col4a5*, and in Ca<sup>2+</sup> channels including *Cacna1c*, *CaM*, *Creb5*, and *Nfatc*  
216 were highly expressed (>2-fold change) in IB (Additional file 1: Figure S22). It has been  
217 demonstrated that *Bmp*, *Fgf2*, and *Fgfr1* are involved in intramembranous bone development and  
218 affect the expression and activity of other osteogenesis related transcription factors [27, 28]. The  
219 calcium-sensitive transcription factor *NFATc1* together with *CREB* induces the expression of  
220 osteoclast-specific genes [29].

## 221 **Adaptation to Herbivorous Diet**

222 Next to the presence of IB, the herbivory of *M. amblycephala* is the second key feature in  
223 connection to the use of this species as aquatic livestock. Olfaction, the sense of smell, is crucial  
224 for animals to find food. The perception of smell is mediated by a large gene family of olfactory  
225 receptor (OR) genes. The ORs of teleosts are predominantly expressed in the main olfactory  
226 epithelium of the nasal cavity [30, 31] and can discriminate, like those of other vertebrates,  
227 different kinds of odor molecules. However, compared to mammals, e.g. humans having around  
228 400 ORs [32] the OR repertoires in teleosts are considerably small. They range from only about  
229 48 in *Fugu rubripes* up to 161 in *D. rerio* (Figure 4A). In the *M. amblycephala* genome, we

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

230 identified 179 functional olfactory receptor (OR) genes (Figure 4A), and based on the  
231 classification of Niimura [33], 158, 117 and 153 receptors for water-borne odorants were  
232 identified in *M. amblycephala*, *C. idellus* and *D. rerio*, respectively (Additional file 1: Table S12).  
233 Overall, these receptor repertoires are substantially larger than those of other and carnivorous  
234 teleosts (*G. morhua*, *C. semilaevis*, *O. latipes*, *X. maculatus*) (Additional file 1: Figures S23 and  
235 S24, Table S12). This suggests that olfaction—probably for food choice—has a particularly  
236 important role in the cyprinid species. Previous studies have demonstrated that the beta type OR  
237 genes are present in both aquatic and terrestrial vertebrates, indicating that the corresponding  
238 receptors detect both water-soluble and airborne odorants [31, 33]. Intriguingly, we found a  
239 massive expansion of beta-type OR genes in the genomes of the herbivorous *M. amblycephala* and  
240 *C. idellus*, while very few exist in other teleosts (Figure 4B, Additional file 1: Table S12).

241 Taste is also an important factor in the development of dietary habits. Most animals can  
242 perceive five basic tastes, namely sourness, sweetness, bitterness, saltiness and umami [34].  
243 Interestingly, *TIR1*, the receptor gene necessary for sensing umami, has been lost in herbivorous  
244 *M. amblycephala* but is duplicated in carnivorous *G. morhua*, *C. semilaevis* and omnivorous *O.*  
245 *latipes* and *X. maculatus* (Figures 4C and 4D, Additional file 1: Figures S25-26 and Table S13). In  
246 contrast, *TIR2*, the receptor gene for sensing sweet, has been duplicated in herbivorous *M.*  
247 *amblycephala* and *C. idellus*, omnivorous *C. carpio* and *D. rerio*, while it has been lost in  
248 carnivorous *G. morhua* and *C. semilaevis* (Additional file 1: Figure S27 and Table S13). Bitterness  
249 sensed by the *T2R* is particularly crucial for animals to protect them from poisonous compounds  
250 [35]. Probably in the course switching to a diet that contains a larger fraction of bitterness  
251 containing food, also the *T2R* gene family in *M. amblycephala*, *C. idellus*, *C. carpio* and *D. rerio*  
252 has been expanded (Additional file 1: Figure S28).

253 To obtain further insights into the genetic adaptation to herbivorous diet, we focused on  
254 further genes that might be associated with digestion. Genes that encode proteases (including  
255 pepsin, trypsin, cathepsin and chymotrypsin) and amylases (including alpha-amylase and  
256 glucoamylase) were identified in the genomes of *M. amblycephala*, carnivorous *C. semilaevis*, *G.*  
257 *morhua* and omnivorous *D. rerio*, *O. latipes* and *X. maculatus*, indicating that herbivorous *M.*  
258 *amblycephala* has a protease repertoire that is not substantially different from those of carnivorous

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

259 and omnivorous fishes (Additional file 1: Table S14). We did not identify any genes encoding  
260 potentially cellulose-degrading enzymes including endoglucanase, exoglucanase and  
261 beta-glucosidase in the genome of *M. amblycephala*, suggesting that utilization of the herbivorous  
262 diet may largely depend on the gut microbiome. To elucidate this further, we determined the  
263 composition of the gut microbial communities of juvenile, domestic, wild adult *M. amblycephala*  
264 and wild adult *C. idellus* using bacterial 16S rRNA sequencing. A total of 549,020 filtered high  
265 quality sequence reads from 12 samples were clustered at a similarity level of 97%. The resulting  
266 8,558 operational taxonomic units (OTUs) are dominated at phylum level by Proteobacteria,  
267 Fusobacteria, Bacteroidetes, Firmicutes and Actinobacteria (Additional file 1: Table S15, Figure  
268 4E). Increasing the resolution to the genus level, the composition and relative abundance of the  
269 gut microbiota of wild adult *M. amblycephala* and *C. idellus* are still very similar (Additional file  
270 1: Table S16) and we could identify more than 7% cellulose-degrading bacteria (Additional file 1:  
271 Table S17). This indicates that indeed the gut microbiome is crucial in the digestion of plant  
272 material, and thus in the adaptation to herbivory.

## 273 Discussion

274 *M. amblycephala* is the economically most important species for freshwater aquaculture. In  
275 addition to its various superior properties, especially its herbivorous diet, it is also an excellent  
276 model to study IB formation. Here we make available draft genome of *M. amblycephala* with  
277 more than 70% of genome data anchored on 24 linkage groups. Comparative analyses of genome  
278 structure revealed high synteny with three other cyprinid fish and uncovered a chromosomal  
279 fusion event in *M. amblycephala* that joined two *D. rerio* chromosomes (Figure 1B), which  
280 supports the previous results in *C. idellus* [13] and also provides novel scientific insights into the  
281 evolution of chromosome fusion events in cyprinids.

282 The evolutionary trajectory analysis of *M. amblycephala* and other teleosts revealed that *M.*  
283 *amblycephala* has the closest relationship to *C. idellus* (Figure 2A). Both the species are  
284 herbivorous fish but which endogenous and exogenous factors affected their feeding habits and  
285 how they adapted to their herbivorous diet is not known. Olfaction and taste are crucial for  
286 animals to find food and to distinguish whether potential food is edible or harmful [31, 35]. 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  
287 search for genes encoding OR showed that herbivorous *M. amblycephala* and *C. idellus* have a  
288 large number of beta-type OR, while other omnivorous and carnivorous fish only have one or two.  
289 This might be attributed to their particular herbivorous diet consisting not only of aquatic grasses  
290 but also the duckweed and terrestrial grasses, which they ingest from the water surface. Previous  
291 studies have demonstrated that the receptor for umami is formed by the T1R1/T1R3 heterodimer,  
292 while T1R2/T1R3 senses sweet taste [36]. We found that the umami gene *T1R1* was lost in  
293 herbivorous *M. amblycephala* but duplicated in the carnivorous *G. morhua* and *C. semilaevis*  
294 (Figure 4C). The loss of the *T1R1* gene in *M. amblycephala* might exclude the expression of a  
295 functional umami taste receptor. Such situations in other organism, e.g. the Chinese panda, have  
296 previously been related to feeding specialization [37]. Interestingly, the sweetness receptor *T1R2*  
297 and bitter receptor *T2R* genes are expanded in the herbivorous fish but few or no copy was found  
298 in carnivorous fish. Collectively, these results not only indicate the genetic adaptation to  
299 herbivorous diet of *M. amblycephala*, but also provided a clear and comprehensive picture of  
300 adaptive evolutionary mechanisms of sensory systems in other fish species with different trophic  
301 specializations.

31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
302 Some insects such as *Tenebrio molitor* [38] and *Neotermes koshunensis* [39], and the mollusc  
303 *Corbicula japonica* [40] have genes encoding endogenous cellulose degradation-related enzymes.  
304 However, all so far analyzed herbivorous vertebrates lack these genes and always rely on their gut  
305 microbiome to digest food [37, 41]. In herbivorous *M. amblycephala* and *C. idellus*, we also did  
306 not find any homologues of digestive cellulase genes. Interestingly, our work on the composition  
307 of gut microbiota of the two fish species identifies more than 7% cellulose-degrading bacteria,  
308 suggesting that the cellulose degradation of herbivorous fish largely depend on their gut  
309 microbiome.

48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
310 IB has evolved several times during teleost evolution [9, 42]. The developmental mechanisms  
311 and ossification processes forming IBs are dramatically distinct from other bones such as ribs,  
312 skeleton, vertebrae or spines. These usually develop from cartilaginous bone and are derived from  
313 the mesenchymal cell population by endochondral ossification [27, 43]. However, IBs form  
314 directly by intramembranous ossification and differentiate from osteoblasts within connective  
315 tissue, forming segmental, serially homologous ossifications in the myosepta. Although various

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 methods of ossification of IB have been proposed, few experiments have been conducted to  
317 confirm the ossification process and little is known about the potential role of IB in teleosts. Based  
318 on our findings of over-represented functional properties of expanded gene families in cyprinid  
319 lineage (Figure 2C) and evidence from DGE of early developmental stages of IB formation  
320 (Figure 3A), we provide molecular evidence that IB might play significant roles not only in  
321 regulating muscle contraction but also in active remodeling at the bone-muscle interface and  
322 coordination of cellular events.

323 It has previously been found that some major developmental signals including BMP, FGF,  
324 WNT, together with calcium/calmodulin signaling [27, 44-46], are essential for regulating the  
325 differentiation and function of osteoblasts and osteocytes and for regulating the RANKL signaling  
326 pathway for osteoclasts [47] in intramembranous bone development. In agreement with this  
327 concept, our DGE analysis of muscle, IB and connective tissues uncovered that 35 bone formation  
328 regulatory genes involved in these signals were highly up-regulated in IB. Taken together, these  
329 results suggest that IB indeed undergoes an intramembranous ossification process, is regulated by  
330 bone-specific signaling pathways, and underlies a homeostasis of maintenance, repair and  
331 remodeling.

## 332 **Conclusions**

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

## 340 **Methods**

### 341 **Sampling and DNA Extraction**

342 DNA for genome sequencing was derived from a double haploid fish from the *M. amblycephala*  
343 genetic breeding center at Huazhong Agricultural University (Wuhan, Hubei, China). Fish blood

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 was collected from adult female fish caudal vein using sterile injectors with pre-added  
345 anticoagulant solutions following anesthetized with MS-222 and sterilization with 75% alcohol.  
346 Genomic DNA was extracted from the whole blood.

### 347 **Genomic Sequencing and Assembly**

348 Libraries with different insert sized inserts of 170 bp, 500 bp, 800 bp, 2 Kb, 5 Kb, 10 Kb and 20  
349 Kb were constructed from the genomic DNA at BGI-Shenzhen. The libraries were sequenced  
350 using a HiSeq2000 instrument. In total, 11 libraries, sequenced in 23 lanes were constructed. To  
351 obtain high quality data, we applied filtering criteria for the raw reads. As a result, 142.55 Gb of  
352 filtered data were used to complete the genome assembly using SOAPdenovo\_V2.04 [16]. Only  
353 filtered data were used in the genome assembly. First, the short insert size library data were used  
354 to construct a de Bruijn graph. The tips, merged bubbles and connections with low coverage were  
355 removed before resolving the small repeats. Second, all high-quality reads were realigned with the  
356 contig sequences. The number of shared paired-end relationships between pairs of contigs was  
357 calculated and weighted with the rate of consistent and conflicting paired ends before constructing  
358 the scaffolds in a stepwise manner from the short-insert size paired ends to the long-insert size  
359 paired ends. Third, the gaps between the constructed scaffolds were composed mainly of repeats,  
360 which were masked during scaffold construction. These gaps were closed using the paired-end  
361 information to retrieve read pairs in which one end mapped to a unique contig and the other was  
362 located in the gap region. Subsequently, local assembly was conducted for these collected reads.  
363 To assess the genome assembly quality, approximately 42.82 Gb Illumina reads generated from  
364 short-insert size libraries were mapped onto the genome. Bwa0.5.9-r16 software [48] with default  
365 parameters was used to assess the mapping ratio and Soap coverage 2.27 was used to calculate the  
366 sequencing depth. We also assessed the accuracy of the genome assembly by Trinity [49],  
367 including number of ESTs and new mRNA reads from early stages of embryos and multiple  
368 tissues, by aligning the scaffolds to the assembled transcriptome sequences.

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

373 **Genome Annotation**

1  
2 374 The genome was searched for repetitive elements using Tandem Repeats Finder (version 4.04)  
3  
4 375 [50]. TEs were identified using homology-based approaches. The Repbase (version 16.10) [51]  
5  
6 376 database of known repeats and a *de novo* repeat library generated by RepeatModeler were used.  
7  
8 377 This database was mapped using the software of RepeatMasker (version 3.3.0). Four types of  
9  
10 378 non-coding RNAs (microRNAs, transfer RNAs, ribosomal RNAs and small nuclear RNAs) were  
11  
12 379 also annotated using tRNAscan-SE (version 1.23) and the Rfam database45 (Release 9.1) [52].

14 380 For gene prediction, *de novo* gene prediction, homology-based methods and RNA-seq data  
15  
16 381 were used to perform gene prediction. For the sequence similarity based prediction, *D. rerio*, *G.*  
17  
18 382 *aculeatus*, *O. niloticus*, *O. latipes* and *G. morhua* protein sequences were downloaded from  
19  
20 383 Ensembl (release 73) and were aligned to the *M. amblycephala* genome using TBLASTN. Then  
21  
22 384 homologous genome sequences were aligned against the matching proteins using GeneWise [53]  
23  
24 385 to define gene models. Augustus was employed to predict coding genes using appropriate  
25  
26 386 parameters in *de novo* prediction. For the RNA-seq based prediction, we mapped transcriptome  
27  
28 387 reads to the genome assembly using TopHat [54]. Then, we combined TopHat mapping results  
29  
30 388 together and applied Cufflinks [55] to predict transcript structures. All predicted gene structures  
31  
32 389 were integrated by GLEAN [56] (<http://sourceforge.net/projects/glean-gene/>) to obtain a  
33  
34 390 consensus gene set. Gene functions were assigned to the translated protein-coding genes using  
35  
36 391 Blastp tool, based on their highest match to proteins in the SwissProt and TrEMBL [57] databases  
37  
38 392 (Uniprot release 2011-01). Motifs and domains in the protein-coding genes were determined by  
39  
40 393 InterProScan (version 4.7) searches against six different protein databases: ProDom, PRINTS,  
41  
42 394 Pfam, SMART, PANTHER and PROSITE. Gene Ontology [58] IDs for each gene were obtained  
43  
44 395 from the corresponding InterPro entries. All genes were aligned against KEGG [59] (Release 58)  
45  
46 396 database, and the pathway in which the gene might be involved was derived from the matched  
47  
48 397 genes in KEGG. tRNA genes were *de novo* predicted by tRNAscan-SE software [60], with  
49  
50 398 eukaryote parameters on the repeat pre-masked genome. The rRNA fragments were identified by  
51  
52 399 aligning the rRNA sequences using BlastN at E-value 1e-5. The snRNA and miRNA were  
53  
54 400 searched by the method of aligning and searching with INFERNAL (version 0.81) [61] against  
55  
56 401 Rfam database (release 9.1).  
57  
58  
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

## 402 Genetic map construction

403 To anchor the scaffolds into pseudo-chromosomes, 198 F1 population individuals were used to  
404 obtain the genetic map. Each of the individual genomic DNA was digested with the restriction  
405 endonuclease EcoR I, following the RAD-Seq protocol [62]. The SNP calling process was carried  
406 out using the SOAP bioinformatic pipeline. The RAD-based SNP calling was done by SOAPsnp  
407 software [63] after each individual's paired-end RAD reads was mapped onto the assembled  
408 reference genome with the alignment software SOAP2 [64]. The potential SNP markers were used  
409 for the linkage analysis if the following criteria were satisfied: for parents - sequencing depth  $\geq 8$   
410 and  $\leq 100$ , base quality  $\geq 25$ , copy number  $\leq 1.5$ ; for progeny - sequencing depth  $\geq 5$ , base quality  
411  $\geq 20$ , copy number  $\leq 1.5$ . If the markers were showing significantly distorted segregation ( $P$ -value  
412  $< 0.01$ ), they were excluded from the map construction. Linkage analysis was performed only for  
413 markers present in at least 80% of the genomes, using JoinMap 4.0 software with CP population  
414 type codes and applying the double pseudo-test cross strategy [65]. The linkage groups were  
415 formed at a logarithm of odds threshold of 6.0 and ordered using the regression mapping  
416 algorithm.

## 417 Construction of gene families

418 We identified gene families using TreeFam software [66] as follows: Blast was used to compare  
419 all the protein sequences from 13 species: *M. amblycephala*, *C. idellus*, *C. semilaevis*, *C. carpio*,  
420 *D. rerio*, *Callorhinchus milii*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *Oncorhynchus*  
421 *mykiss*, *O. niloticus*, *O. latipes*, *Fugu rubripes*, with the E-value threshold set as  $1e-7$ . In the next  
422 step, HSP segments of each protein pair were concatenated by Solar software. H-scores were  
423 computed based on Bit-scores and these were taken to evaluate the similarity among genes.  
424 Finally, gene families were obtained by clustering of homologous gene sequences using  
425 Hcluster\_sg (Version 0.5.0). Specific genes of *M. amblycephala* were those that did not cluster  
426 with other vertebrates that were chosen for gene family construction, and those that did not have  
427 homologs in the predicted gene repertoire of the compared genomes. If these genes had functional  
428 motifs, they were annotated by GO.

## 429 Phylogenetic Tree Reconstruction and Divergence Time Estimation

430 The coding sequences of single-copy gene families conserved among *M. amblycephala*, *C. idellus*,

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 *C. carpio*, *D. rerio*, *C. semilaevis*, *G. morhua*, *G. aculeatus*, *Latimeria chalumnae*, *O. mykiss*, *O.*  
432 *niloticus*, *O. latipes*, *C. milii* and *Fugu rubripes* (Ensembl Gene v.77) were extracted and aligned  
433 with guidance from amino-acid alignments created by the MUSCLE program [67]. The individual  
434 sequence alignments were then concatenated to form one supermatrix. PhyML [68, 69] was  
435 applied to construct the phylogenetic tree under an HKY85+gamma model for nucleotide  
436 sequences. ALRT values were taken to assess the branch reliability in PhyML. The same set of  
437 codon sequences at position 2 was used for phylogenetic tree construction and estimation of the  
438 divergence time. The PAML mcmctree program (PAML version 4.5) [70, 71] was used to  
439 determine divergence times with the approximate likelihood calculation method and the  
440 'correlated molecular clock' and 'REV' substitution model.

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

442 Protein sequences of *M. amblycephala* and 11 other related species (Ensembl Gene v.77)) were  
443 used in BLAST searches to identify homologs. We identified gene families using CAFÉ [72],  
444 which employs a random birth and death model to study gene gains and losses in gene families  
445 across a user-specified phylogeny. The global parameter  $\lambda$ , which describes both the gene birth ( $\lambda$ )  
446 and death ( $\mu = -\lambda$ ) rate across all branches in the tree for all gene families, was estimated using  
447 maximum likelihood. A conditional *P*-value was calculated for each gene family, and families  
448 with conditional *P*-values less than the threshold (0.05) were considered as having notable gain or  
449 loss. We identified branches responsible for low overall *P*-values of significant families.

#### 450 **Detection of Positively Selected Genes**

451 We calculated Ka/Ks ratios for all single copy orthologs of *M. amblycephala* and *C. semilaevis*, *D.*  
452 *rerio*, *G. morhua*, *O. niloticus* and *C. carpio*. Alignment quality was essential for estimating  
453 positive selection. Thus, orthologous genes were first aligned by PRANK [73], which is  
454 considerably conservative for inferring positive selection. We used Gblocks [74] to remove  
455 ambiguously aligned blocks within PRANK alignments and employed 'codeml' in the PAML  
456 package with the free-ratio model to estimate Ka, Ks, and Ka/Ks ratios on different branches. The  
457 differences in mean Ka/Ks ratios for single-copy genes between *M. amblycephala* and each of the  
458 other species were compared using paired Wilcoxon rank sum tests. Genes that showed values of  
459 Ka/Ks higher than 1 along the branch leading to *M. amblycephala* were reanalyzed using 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

460 codon based branch-site tests implemented in PAML. The branch-site model allowed  $\omega$  to vary  
461 both among sites in the protein and across branches, and was used to detect episodic positive  
462 selection.

### 463 **Developmental process of intermuscular bone in *M. amblycephala***

464 To better understand the number and morphological types of IBs in adult *M. amblycephala*,  
465 specimens with a body length ranging from 15.5 to 20.5 cm were collected and each individual  
466 was wrapped in gauze and boiled. The fish body was divided into two sections: anterior (snout to  
467 cloaca) and posterior (cloaca to the base of caudal fin), and the length of each section was  
468 measured. The IBs were retrieved, counted, arranged in order and photographed with a digital  
469 camera. Fertilized *M. amblycephala* eggs were brought from hatching facilities at Freshwater Fish  
470 Genetics Breeding Center of Huazhong Agricultural University (Wuhan, Hubei, China) to our  
471 laboratory. *M. amblycephala* larvae were maintained in a re-circulating aquaculture system at  $23 \pm$   
472  $1^\circ\text{C}$  with a 14-hr photoperiod. To explore the early development of IBs, larvae at different stages  
473 from 15 to 40 dpf were collected and fixed in 4% paraformaldehyde and transferred to 70%  
474 ethanol for storage. Specimens were stained with alizarin red for bone following the method  
475 described by Dawson [75]. The appearance of red color was recorded as the appearance of IB  
476 because bone ossification is accompanied by the uptake of alizarin red, resulting in red staining of  
477 the mineralized bone matrix. Myosepta, either not yet ossified, or poorly ossified, are not visible  
478 with alizarin red staining. For histologic analysis, specimens were paraffin-embedded and  
479 sectioned following standard protocols. Sections were stained with hematoxylin and eosin (HE)  
480 and Masson trichrome [76] and photographed using a Nikon microscope (Nikon, Tokyo, Japan)  
481 with a DP70 digital camera (Olympus, Japan). Scanning electron microscopy (SEM) and  
482 transmission electron microscopy (TEM) were also conducted to analyze the ultrastructure of IB.  
483 The specimens were fixed with 2.5% (v/v) glutaraldehyde in a solution of 0.1 M sodium  
484 cacodylate buffer (pH 7.3) for 2 h at room temperature. The SEM and TEM samples were  
485 prepared according to a standard protocol described by Ott [77]. The samples were then visualized  
486 with a JSM-6390LV scanning electron microscope (SEM, Japan) and the stained ultrathin sections  
487 with a H-7650 transmission electron microscope (Hitachi, Japan).

### 488 **RNA Sequencing Analysis**

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 *M. amblycephala* specimens belonging to three different developmental stages of IBs (stage 1:  
490 whole larvae without distribution of IB; stage 2: muscle tissues with partial distribution of IBs; stage 3:  
491 muscle tissues with completed distribution of IBs were identified under microscope and immediately  
492 frozen in liquid nitrogen. In addition, dorsal white muscle, IBs and connective tissue surrounding the  
493 IBs from six months old fish were also collected. RNA was extracted from total fish samples at  
494 different stages and from individual muscle, connective tissue, and intermuscular bone samples of  
495 *M. amblycephala* using RNAisoPlus Reagent (TaKaRa, China) according to the manufacturer's  
496 protocol. The integrity and purity of the RNA was determined by gel electrophoresis and Agilent  
497 2100 BioAnalyzer (Agilent, USA) before preparing the libraries for sequencing. Paired-end RNA  
498 sequencing was performed using the Illumina HiSeq 2000 platform. Low quality score reads were  
499 filtered and the clean data were aligned to the reference genome using Bowtie [78]. Genes and  
500 isoforms expression level were quantified by a software package: RSEM (RNASeq by  
501 Expectation Maximization) [79]. Gene expression levels were calculated by using the RPKM  
502 method (Reads per kilobase transcriptome per million mapped reads) [80] and adjusted by a  
503 scaling normalization method [81]. DEGs were detected using DESeq [82]. Annotation of DEGs  
504 were mapped to GO categories in the database (<http://www.geneontology.org/>) and the number of  
505 genes for every term were calculated to identify GO terms that were significantly enriched in the  
506 input list of DEGs. The calculated *P*-value was adjusted by the Bonferroni Correction, taking  
507 corrected *P*-value  $\leq 0.05$  as a threshold. KEGG automatic annotation was used to perform  
508 pathway enrichment analysis of DEGs.

### 509 **Prediction of Olfactory Receptor Genes**

510 Olfactory receptor genes were identified by previously described methods [83], with the exception  
511 of a first-round TBLASTN [84] search, in which 1,417 functional olfactory receptor genes from *H.*  
512 *sapiens*, *D. rerio*, *L. chalumnae*, *Lepisosteus oculatus*, *L. vexillifer*, *O. niloticus*, *O. latipes*, *F.*  
513 *rubripes* and *Xenopus tropicalis* were used as queries. We then predicted the structure of  
514 sequenced genes using the blast-hit sequence with the software GeneWise extending in both 3' and  
515 5' directions along the genome sequences. The results were further confirmed by NR annotation.  
516 Then the coding sequences from the start (ATG) to stop codons were extracted, while sequences  
517 that contained interrupting stop codons or frame-shifts were regarded as pseudogenes. To



1  
2  
3  
4  
5  
6  
7  
8  
9  
518 construct phylogenetic trees, the amino-acid sequences encoded by olfactory receptor genes were  
519 first aligned using the program MUSCLE nested in MEGA 5.10 [85]. We then constructed the  
520 phylogenetic tree using the neighbor-joining method with Poisson correction distances using the  
521 program MEGA 5.10. We also identified and compared the genes for five basic tastes (sour, sweet,  
522 bitter, umami and salty) using a similar method as in OR gene identification.

### 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 65 523 **Gut microbiota analysis**

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

### 537 **Additional files**

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

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

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

### 542 **Abbreviations**

543 IB, intermuscular bone; SNP, single-nucleotide polymorphism; BUSCO, benchmarking universal  
544 single-copy orthologs; TE, transposable element; LTR, long terminal repeat retrotransposon; LG,  
545 linkage group; PSG, positively selected gene; ECM, extracellular matrix; dpf, days post  
546 fertilization; BMP, bone morphogenetic protein; FGF, fibroblast growth factor; OR, olfactory

1  
2 547 receptor; OTU, operational taxonomic unit; DGE, differential genes expression; HE, hematoxylin  
3  
4 548 and eosin; SEM, scanning electron microscopy; TEM, transmission electron microscopy  
5

## 6 549 **Acknowledgements**

7 550 This work was supported by the Modern Agriculture Industry Technology System Construction  
8  
9 551 Projects of China titled as—Staple Freshwater Fishes Industry Technology System (No.  
10  
11 552 CARS-46-05), Guangdong Haid Group Co., Ltd, the Fundament Research Funds for the Central  
12  
13 553 Universities (2662015PY019), the International Scientific and Technology Cooperation Program  
14  
15 554 of Wuhan City (2015030809020365).  
16  
17

## 18 555 **Availability of data and materials**

19  
20  
21 556 Datasets supporting the results of this article are available in the GigaDB repository associated  
22  
23 557 with this publication [88]. Raw whole genome sequencing, transcriptome and RAD-Seq data have  
24  
25 558 been deposited at NCBI in the SRA under bioproject number PRJNA343584.  
26  
27

## 28 559 **Authors' contributions**

29  
30 560 W.W. initiated and conceived the project and provided scientific input. X.Q. organized financial  
31  
32 561 support and designed the project. M.S. discussed the data, wrote and modified the paper. H.L. and  
33  
34 562 C.C. conducted the biological experiments, analyzed the data and wrote the paper with input from  
35  
36 563 other authors. I.E. wrote and modified the manuscript and discussed the data. The RAD-Seq data  
37  
38 564 analyses and the genetic map construction were performed by Z.G., Y.G., J.J. and X.J. Genome  
39  
40 565 assembly and annotation were performed by J.M., H.C., M.X. and J.C. X.Z., W.L., R.L., B.C.,  
41  
42 566 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  
43  
44 567 and data collection. J.L. and J.C. identified the gene families and analyzed the RNA-seq data. M.B.  
45  
46 568 coordinated the project. S.Z. and X.F. modified the manuscript and discussed the data. All authors  
47  
48 569 read the manuscript and provided comments and suggestions for improvements. The authors  
49  
50 570 declare no competing financial interests.  
51  
52  
53

## 54 571 **Competing interests**

55  
56  
57 572 The authors declare that they have no competing interests.  
58  
59

## 60 573 **Ethics approval and consent to participate**

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

## 577 **References**

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

- 606 13. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, et al. The draft genome of the grass carp  
607 (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation.  
608 Nat. Genet. 2015; 47:625–31.
- 609 14. Gao Z, Luo W, Liu H, Zeng C, Liu X, Yi S, et al. Transcriptome analysis and SSR/SNP  
610 markers information of the blunt snout bream (*Megalobrama amblycephala*). PLoS One.  
611 2012; 7:e42637.
- 612 15. Yi S, Gao ZX, Zhao H, Zeng C, Luo W, Chen B, et al. Identification and characterization of  
613 microRNAs involved in growth of blunt snout bream (*Megalobrama amblycephala*) by  
614 Solexa sequencing. BMC Genomics. 2013; 14:754.
- 615 16. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved  
616 memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.
- 617 17. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing  
618 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.  
619 2015;31:3210–2.
- 620 18. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis  
621 of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61.
- 622 19. Even-Ram S, Doyle AD, Conti MA, Matsumoto K, Adelstein RS, Yamada KM. Myosin IIA  
623 regulates cell motility and actomyosin-microtubule crosstalk. Nat. Cell Biol. 2007;  
624 9:299–309.
- 625 20. Sheetz MP, Felsenfeld DP, Galbraith CG. Cell migration: Regulation of force on  
626 extracellular-complexes. Trends Cell Biol. 1998; 8:51–4.
- 627 21. Gunst SJ, Zhang W. Actin cytoskeletal dynamics in smooth muscle: a new paradigm for the  
628 regulation of smooth muscle contraction. Am. J. Physiol. Cell Physiol. 2008; 295:C576–87.
- 629 22. Webb RC. Smooth muscle contraction and relaxation. Adv. Physiol. Educ. 2003; 27:201–6.
- 630 23. Ulrich TA, De Juan Pardo EM, Kumar S. The mechanical rigidity of the extracellular matrix  
631 regulates the structure, motility, and proliferation of glioma cells. Cancer Res. 2009;  
632 69:4167–74.
- 633 24. Ridley AJ. Rho GTPases and cell migration. J. Cell Sci. 2001; 114:2713–22.
- 634 25. Etienne-Manneville S, Hall A. Rho GTPases in Cell Biology. Nature. 2002; 420:629–35.
- 635 26. Ridley AJ. Cell Migration: Integrating Signals from Front to Back. Science. 2003;  
636 302:1704–9.
- 637 27. Ornitz D, Marie P. FGF signaling pathways in endochondral and intramembranous bone  
638 development and human genetic disease. Genes Dev. 2002; 16:1446–65.

- 639 28. Fakhry A, Ratisoontorn C, Vedhachalam C, Salhab I, Koyama E, Leboy P, et al. Effects of  
640 FGF-2/-9 in calvarial bone cell cultures: Differentiation stage-dependent mitogenic effect,  
641 inverse regulation of BMP-2 and noggin, and enhancement of osteogenic potential. *Bone*.  
642 2005; 36:254–66.
- 643 29. Sato K, Suematsu A, Nakashima T, Takemoto-Kimura S, Aoki K, Morishita Y, et al.  
644 Regulation of osteoclast differentiation and function by the CaMK-CREB pathway. *Nat. Med.*  
645 2006; 12:1410–6.
- 646 30. Niimura Y, Nei M. Evolutionary dynamics of olfactory and other chemosensory receptor  
647 genes in vertebrates. *J. Hum. Genet.* 2006; 51:505–17.
- 648 31. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene  
649 repertoires: roles of chance and necessity. *Nat. Rev. Genet.* 2008; 9:951–63.
- 650 32. Lopez C, Raper J. Cloning and functional characterization of odorant receptors expressed in  
651 the zebrafish olfactory system. *FASEB J.* 2015; 29:727–37.
- 652 33. Niimura Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative  
653 genome analysis among 23 chordate species. *Genome Biol. Evol.* 2009; 1:34–44.
- 654 34. Lindemann B. Receptors and transduction in taste. *Nature.* 2001; 413:219–25.
- 655 35. Chandrashekar J, Mueller KL, Hoon MA, Adler E, Feng L, Guo W, et al. T2Rs function as  
656 bitter taste receptors. *Cell.* 2000; 100:703–11.
- 657 36. Nelson G, Chandrashekar J, Hoon MA, Feng L, Zhao G, Ryba NJ, et al. An amino-acid taste  
658 receptor. *Nature.* 2002; 416:199–202.
- 659 37. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and *de novo* assembly of the  
660 giant panda genome. *Nature.* 2010; 463:311–7.
- 661 38. Ferreira AHP, Marana SR, Terra WR, Ferreira C. Purification, molecular cloning, and  
662 properties of a  $\beta$ -glycosidase isolated from midgut lumen of *Tenebrio molitor* (Coleoptera)  
663 larvae. *Insect Biochem. Mol. Biol.* 2001; 31:1065–76.
- 664 39. Tokuda G, Saito H, Watanabe H. A digestive  $\beta$ -glucosidase from the salivary glands of the  
665 termite, *Neotermes koshunensis* (Shiraki): Distribution, characterization and isolation of its  
666 precursor cDNA by 5'- and 3'-RACE amplifications with degenerate primers. *Insect*  
667 *Biochem. Mol. Biol.* 2002; 32:1681–9.
- 668 40. Sakamoto K, Uji S, Kurokawa T, Toyohara H. Molecular cloning of endogenous  
669  $\beta$ -glucosidase from common Japanese brackish water clam *Corbicula japonica*. *Gene.* 2009;  
670 435:72–9.
- 671 41. Zhu L, Wu Q, Dai J, Zhang S, Wei F. Evidence of cellulose metabolism by the giant panda gut  
672 microbiome. *Proc. Natl. Acad. Sci. USA.* 2011; 108:17714–9.

- 673 42. Bird NC, Mabee PM. Developmental morphology of the axial skeleton of the zebrafish, *Danio*  
674 *rerio* (Ostariophysi: Cyprinidae). *Dev. Dyn.* 2003; 228:337–57.
- 675 43. Ortega N, Behonick DJ, Werb Z. Matrix remodeling during endochondral ossification. *Trends*  
676 *Cell Biol.* 2004; 14:86–93.
- 677 44. Chen G, Deng C, Li YP. TGF- $\beta$  and BMP signaling in osteoblast differentiation and bone  
678 formation. *Int. J. Biol. Sci.* 2012; 8:272–88.
- 679 45. Harada S, Rodan GA. Control of osteoblast function and regulation of bone mass. *Nature.*  
680 2003; 423:349–55.
- 681 46. Long F. Building strong bones: molecular regulation of the osteoblast lineage. *Nat. Rev. Mol.*  
682 *Cell Biol.* 2011; 13:27–38.
- 683 47. Boyle WJ, Simonet WS, Lacey DL. Osteoclast differentiation and activation. *Nature.* 2003;  
684 423:337–42.
- 685 48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.  
686 *Bioinformatics.* 2009; 25:1754–60.
- 687 49. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity:  
688 reconstructing a full-length transcriptome without a genome from RNA-Seq data . *Nat.*  
689 *Biotechnol.* 2011; 29:644–52.
- 690 50. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.*  
691 1999; 27:573–80.
- 692 51. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase  
693 Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 2005;  
694 110:462–7.
- 695 52. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: Annotating  
696 non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005; 33:121–4.
- 697 53. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res.* 2004; 14:988–95.
- 698 54. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq.  
699 *Bioinformatics.* 2009; 25:1105–11.
- 700 55. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript  
701 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform  
702 switching during cell differentiation. *Nat. Biotechnol.* 2010; 28:511–5.
- 703 56. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey  
704 bee consensus gene set. *Genome Biol.* 2007; 8:R13.

- 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 57. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement  
706 TrEMBL in 2000. *Nucleic Acids Res.* 2000; 28:45–8.
- 707 58. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool  
708 for the unification of biology. *Nat. Genet.* 2000; 25:25–9.
- 709 59. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*  
710 2000; 28:27-30.
- 711 60. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes  
712 in genomic sequence. *Nucleic Acids Res.* 1997; 25:955–64.
- 713 61. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: Inference of RNA alignments. *Bioinformatics.*  
714 2009; 25:1335–7.
- 715 62. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP  
716 discovery and genetic mapping using sequenced RAD markers. *PloS One.* 2008; 3:e3376.
- 717 63. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively  
718 parallel whole-genome resequencing. *Genome Res.* 2009; 19:1124–32.
- 719 64. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: An improved ultrafast tool  
720 for short read alignment. *Bioinformatics.* 2009; 25:1966–7.
- 721 65. Grattapaglia D, Sederoff R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus*  
722 *urophylla* using a pseudo-testcross: Mapping strategy and RAPD markers. *Genetics.* 1994;  
723 137:1121–37.
- 724 66. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, et al. TreeFam: a curated  
725 database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 2006;  
726 34:D572–80.
- 727 67. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.  
728 *Nucleic Acids Res.* 2004; 32:1792–7.
- 729 68. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and  
730 methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML  
731 3.0. *Syst. Biol.* 2010; 59:307–21.
- 732 69. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by  
733 maximum likelihood. *Syst. Biol.* 2003; 52:696–704.
- 734 70. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;  
735 24:1586–91.
- 736 71. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock  
737 using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* 2006; 23:212–26.

- 738 72. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics*.  
739 2007; 177:1941–9.
- 740 73. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with  
741 insertions. *Proc. Natl. Acad. Sci. USA*. 2005; 102:10557–62.
- 742 74. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and  
743 ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 2007; 56:564–77.
- 744 75. Dawson AB. A note on the staining of the skeleton of cleared specimens with alizarin red S.  
745 *Biotech. Histochem.* 1926; 1:123-4.
- 746 76. Gruber HE. Adaptations of Goldner’s Masson trichrome stain for the study of undecalcified  
747 plastic embedded bone. *Biotech. Histochem.* 1992; 67:30–4.
- 748 77. Ott HC, Matthiesen TS, Goh SK, Black LD, Kren SM, Netoff TI, et al.  
749 Perfusion-decellularized matrix: using nature’s platform to engineer a bioartificial heart. *Nat.*  
750 *Med.* 2008; 14:213–21.
- 751 78. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 2012;  
752 9:357–9.
- 753 79. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or  
754 without a reference genome. *BMC Bioinformatics.* 2011; 12:323.
- 755 80. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying  
756 mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 2008; 5:621–8.
- 757 81. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis  
758 of RNA-seq data. *Genome Biol.* 2010; 11:R25.
- 759 82. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.*  
760 2010; 11:R106.
- 761 83. Niimura Y. Evolutionary dynamics of olfactory receptor genes in chordates: interaction  
762 between environments and genomic contents. *Hum. Genomics.* 2009; 4:107–18.
- 763 84. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and  
764 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;  
765 25:3389–402.
- 766 85. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary  
767 analysis of DNA and protein sequences. *Brief. Bioinform.* 2008; 9:299–306.
- 768 86. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al.  
769 Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc.*  
770 *Natl. Acad. Sci. USA.* 2011; 108:4516–22.



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

771 87. Edgar, RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat.  
772 Methods. 2013; 10:996–8.

773 88. Liu H, Chen CH, Gao ZX, Min JM, Gu YM, Jian JB, et al. The draft genome of *Megalobrama*  
774 *amblycephala* reveals the development of intermuscular bone and adaptation to herbivorous  
775 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

776 **Figure Legends**

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

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

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

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

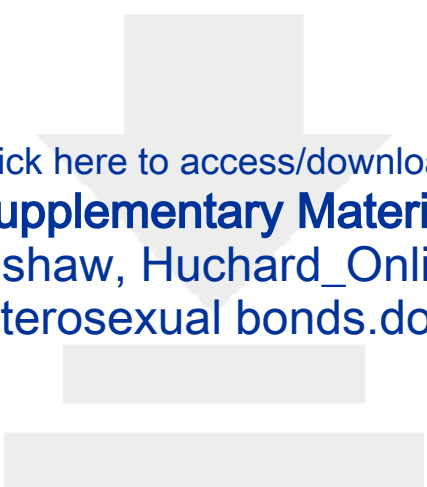
829 **Table**

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

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

831

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



Click here to access/download

**Supplementary Material**

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



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





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

