Author's response to reviews

Title: The genome of the largest bony fish, ocean sunfish (Mola mola), provides insights into its fast growth rate

Authors:

Hailin Pan (panhailin@genomics.cn)

Hao Yu (yuhao@genomics.cn)

Vydianathan Ravi (raviv@imcb.a-star.edu.sg)

Cai Li (licaigd@gmail.com)

Alison Lee (alison_lee@bti.a-star.edu.sg)

Michelle Lian (michelle.lianml@gmail.com)

Boon-Hui Tay (mcblab46@imcb.a-star.edu.sg)

Sydney Brenner (sydney.brenner@hotmail.com)

Jian Wang (wangjian@genomics.cn)

Huanming Yang (yanghm@genomics.cn)

Guojie Zhang (zhanggj@genomics.cn)

Byrappa Venkatesh (mcbbv@imcb.a-star.edu.sg)

Version: 1 **Date:** 21 Jul 2016

Author's response to reviews:

Response to reviewers comments:

Reviewer #1: This manuscript characterizes the genomic property of the ocean sunfish and provides insights into its phenotypic specialization. The primary product of the study, the genome assembly, is well prepared and exhibits very high completeness and continuity, largely thanks to the relatively small genome size and the relatively low frequency of repetitive elements

in the genome. Overall, this study, supported by the high-quality genome assembly, should contribute to an advancement of the research field for actinopterygian fish genomics, and I recommend publication of this manuscript in GigaScience, provided that the points below are reconsidered for improving the manuscript.

1. Some morphological characteristics including the loss of caudal fin are mentioned, in search of possible genomic causes. But, little information in included in the manuscript, regarding the developmental process of the unique body plan. Is this because of the lack of records of embryological development?

Response: Unfortunately, there is very little information in the literature about early developmental stages of sunfish. The only published information is in a text book by Fraser-Brunner [The Ocean Sunfishes (Family Molidae); 1951], which reports that the sunfish passes through two larval phases during its metamorphosis to the adult stage one being a Tetraodon-like stage resembling a miniature pufferfish with large pectoral fins, a tail fin and body spines; and the other being a highly transformed stage where the tail is completely absorbed. We have now provided this information in the Background section of the manuscript.

2. Page 6 / line 101, 'show similarity to sequences in public databases' : similarity to nucleotide or protein sequences?

Response: We have now clarified in the text that similarity refers to protein sequences in the public databases.

3. Some parts of the description in 'Analyses' include highly speculative expressions such as 'may have led to the extremely fast growth rate .' in page 9 line 169. It is recommended to move such speculative expressions to 'Discussion'.

Response: As suggested by the reviewer, speculative statements such as the one mentioned above have been removed from the Analyses section.

4. Page 11 / line 207, 'complete loss fa93e10 and scpp7': insert 'of' between 'loss' and 'fa93e10', if I understand correctly.

Response: Suggested change incorporated.

5. Page 11 / line 211, 'an exact orthology' : Is 'exact' ortholog opposed to 'non-exact' orthology? It can't be. Thus, remove 'exact' from this sentence.

Response: Suggested change incorporated.

6. Has any phenotypic evolution been clearly shown to be attributed to a Hox gene loss? If that has not been shown before, it may not be justified to analyze Hox gene repertoire for identifying causes of the sunfish's unique morphology.

Response: There are no clear cut examples relating Hox gene loss to phenotypic evolution of vertebrates. However, loss of Hox gene function has been shown to affect morphological changes. For example, loss of Hox11 paralogous genes results in malformation of limb zeugopod skeletal elements, radius/ulna and tibia/fibula in mice. In addition, it is also associated with the transformation of the sacral region to a lumbar phenotype [Swinehart et al. (2013) Hox11 genes are required for regional patterning and integration of muscle, tendon and bone. Development 140: 4574-4582]. Similarly, targeted disruption of Hox13 paralogs, Hoxa13 and Hoxd13, both individually and in combination, result in altered morphogenesis of the limb autopod in mice [Fromental-Ramain et al. (1996) Hoxa-13 and Hoxd-13 play a crucial role in the patterning of the limb autopod. Development 122: 2997-3011]. Therefore, we hypothesized that the loss of Hox genes has the potential to result in phenotypic evolution. Based on this hypothesis, we had searched for Hox genes in the sunfish genome. We have now included our rationale for this analysis in the main text (Hox genes section).

7. Page 15 / line 280, 'identified 98.4% of CEGs': I wonder if this is a figure for 'Complete' or 'Partial' gene detection in the CEGMA result?

Response: These are complete genes. We have indicated this in the manuscript now. However, we would like to point out that the previous value of 98.4% was a typographical error. The actual value is 99.6%. The error has been corrected.

Table S1: For mate pair libraries, the figures for the column 'insert size' should not be 'insert size' but something like 'mate distance'.

Response: insert size has been changed to mate distance, as suggested.

Reviewer #2: Overall this is a fascinating paper that outlines the insights from genomic analyses of one of the most enigmatic and charismatic fishes in the ocean, the Ocean sunfish (Mola mola). We have a number of suggestions for the authors that will lead to a more comprehensive manuscript.

Please find our comments below.

As a referee we ask that you assess the paper on its own merits. The following list of potential issues may be helpful.

1. Is the rationale for collecting and analyzing the data well defined?

Is the work carried out on a dataset that can be described as "large-scale" within the context of its field? Does it clearly describe the dataset and provide sufficient context for the reader to understand its potential uses? Does it properly describe previous work?

The motivation for conducting this research was well described in a solidly written introduction. The raw data appears to be of high quality and coverage relative to other work in the field and certainly this data set will be of significant interest to the research community. The types of analyses carried out are creative and will be informative but require some changes to ensure their accuracy and replicability.

2. Is it clear how data was collected and curated?

Credit should be given for transparency and provision of all supporting information.

It would be helpful if the sex of the individual sequenced was specified. Because the sample was obtained in 1998 it would also be useful to know the storage conditions of the blood prior to DNA extraction or, if the DNA was extracted at that time, the storage conditions of the DNA between extraction and sequencing. The method of DNA extraction should also be specified.

Response: Unfortunately, the sex of the fish is not known. The DNA used for genome sequencing was extracted in 1998, suspended in TE buffer and stored at 4°C. We still have some high-molecular weight DNA sitting in our fridge which is in good condition. We had briefly mentioned about sunfish DNA extraction in our paper Venkatesh et al. PNAS (1999) 96: 10267-10271. We have now included this information in the Methods section (under Genome sequencing and assembly).

3. Is it clear - and was a statement provided - on how data and analyses tools used in the study can be accessed?

While we make every effort to make sure this information is available, we appreciate reviewers providing an extra eye to make absolutely certain that this information is clearly stated and properly available. Data availability and access to tools are essential for reproducibility and provide the best means for reuse.

There are a few instances where it is not clear what tools were used to perform certain analyses. Please see detailed comments below.

4. Are accession numbers given or links provided for data that, as a standard, should be submitted to a community approved public repository?

Following community standards for data sharing is a requirement of the journal. Additionally, data sharing in the broadest possible manner expands the ways in which data and tools can be accessed and used.

At the moment there are only tentative NCBI accession numbers given for the assembled genome (XXXXX). I can see that on NCBI a BioProject (Accession: PRJNA305960 ID: 305960) and BioSample (SAMN04335856) have already been registered which is good. It would be helpful if all of the different insert size libraries (additional file 1: table S1) were named in the table and when the raw reads are submitted to the SRA for easy cross-referencing.

Response: The sequences of different insert-size libraries have been uploaded separately to the NCBI under the SRA number SRA319445. They will be released once this paper is accepted for publication.

5. Is the data and software available in the public domain under a Creative Commons license?

Note, that unless otherwise stated, data hosted in our database (GigaDB) is available under a CC0 waiver. Additionally, did the authors indicate where the software tools and relevant source code are available, under an appropriate Open Source Initiative compliant license? If the source code is currently not in a hosted repository, we can help authors copy it over to a GigaScience GitHub repository.

6. Are the data sound and well controlled?

If you feel that inappropriate controls have been used please say so, indicating the reasons for your concerns, and suggesting alternative controls where appropriate. If you feel that further experimental/clinical evidence is required for obtaining solid biological conclusions and substantiating the results, please provide details.

Portions of the analysis, especially the definition of the gene family clusters, require more careful definition and clarification. The text regularly refers to 'single-copy' genes but often does not clearly define their criteria for classifying genes as 'single-copy' and perhaps as a consequence the text reads as internally inconsistent, with different analyses using different datasets of 'single-copy' genes ranging from 1,690 to 3,738 to 10,660. Rather than calling each of these datasets 'single-copy genes' a more specific descriptor should be used for the phylogenetic level at which homology was assessed and whether or not multiple paralogs are present in each case. For example the 1,690 gene set could be called 'single-copy ray-finned fish homologs' as this dataset should comprise only cases where gar and all teleost genomes contain only a single homolog, while the 10,660 gene set could be called simply 'teleost homologs' as this dataset is restricted to teleosts but includes cases where multiple paralogs (e.g. igfr1a, igfr1b) are present and therefore are not 'single-copy'. It is not clear to me if or where the 3,738 'single-copy orthologous' gene set was used or whether this gene set includes paralogs or not. I would additionally urge caution in describing genes as 'orthologous' where simple phenetic (i.e. BLAST-based) methods are used to classify them. Orthology has a very specific phylogenetic meaning and in the context of teleost genomes especially it is important to distinguish between orthologous and paralogous sequences. Where orthology and paralogy have not been assessed using phylogenetic methods the general term 'homology' should be used instead.

Response: We agree with the reviewer that the two different sets of orthologues could be confusing to the readers. We have now named them as ray-finned fish orthologues (1,690) and teleost homologues (10,660). We thank the reviewer for this suggestion.

The 1,690 ray-finned fish orthologues were identified based on a combination of the Ensembl Biomart database which uses phylogenetic analysis and InParanoid analysis (Remm et al. 2001. Automatic clustering of orthologues and in-paralogs from pairwise species comparisons. J Mol Biol. 314: 1041-1052). Thus, this set represents orthologues and not homologues. However, the 10,660 set was identified by Reciprocal Best Hit (RBH) method which we agree is not the most stringent method for identifying orthologues. Therefore, we have named this set as teleost homologues. We have now indicated the methods used for identifying these sets of orthologues in the manuscript.

The 3,738 set was not used in any of the analyses reported in the manuscript and therefore has been deleted.

7. Is the interpretation (Analysis and Discussion) well balanced and supported by the data?

The interpretation should discuss the relevance of all the results in an unbiased manner. Are the interpretations overly positive or negative? Note that the authors may include opinions and speculations in an optional 'Potential Implications' section of the manuscript; thus, if there is material in other parts of the manuscript that you feel would be better suited in such a section, please state that. Conclusions drawn from the study should be valid and result directly from the data shown, with reference to other relevant work as applicable. Have the authors provided references wherever necessary?

The authors are appropriately careful in drawing biological conclusions from their data and throughout the analysis and discussion always imply potential roles rather than implying direct causality.

8. Are the methods appropriate, well described, and include sufficient details and supporting information to allow others to evaluate and replicate the work?

Please remark on the suitability of the methods for the study.

If statistical analyses have been carried out, please indicate if you feel they need to be assessed specifically by an additional reviewer with statistical expertise.

In some cases more detailed descriptions of the methodology including parameters are needed. See details below.

Response: We have provided more detailed descriptions. See details below.

9. What are the strengths and weaknesses of the methods?

Please comment on any improvements that could be made to the study design to enhance the quality of the results. If any additional experiments are required, please give details. If novel experimental techniques were used please pay special attention to their reliability and validity.

In some instances methodological improvements during analysis seem to be necessary to meet minimum requirements for publication. Please see below for details.

Response: Necessary improvements have been made. See details below.

10. Have the authors followed best-practices in reporting standards?

This is an essential component as ease of reproducibility and usability are key criteria for manuscript publication. Please note, the methodology sections should never contain "protocol available upon request" or "e-mail author for detailed protocol". Have the authors followed and used reporting checklists recommended by the Biosharing network and if the methods are amenable, have the authors used workflow management systems such as Galaxy, Taverna or one of the many related systems listed on MyExperiment? We can also host these in our Giga-Galaxy server if they currently do not have a home. We also encourage use of virtual machines and containers such as Docker. And the use and deposition of both wet-lab and computational protocols in a protocols repository like protocols.io.

In some cases additional details are required, particularly for methodology during the annotation and homologous gene cluster building.

Response: Yes, additional details have been provided.

11. Can the writing, organization, tables and figures be improved?

Although the editorial team may also assess the quality of the written English, please do comment if you consider the standard is below that expected for a scientific publication.

If the manuscript is organized in such a manner that it is illogical or not easily accessible to the reader please suggest improvements. Please provide feedback on whether the data are presented in the most appropriate manner; for example, is a table being used where a graph would give increased clarity? Do the figures appear to be genuine, i.e. without evidence of manipulation, and of a high enough quality to be published in their present form?

The manuscript is clearly written. I have suggested moving analysis of bone-forming genes to the main text as it warrants attention. Some minor changes to figures have been recommended. Please see below for details.

Response: These changes have been made.

12. When revisions are requested.

Reviewers may recommend revisions for any or all of the following reasons: the data require additional testing to ensure their quality, additional data are required to support the authors' conclusions; better justification is needed for the arguments based on existing data; or the clarity and/or coherence of the paper needs to be improved.

Several changes and/or clarifications are necessary prior to being published. Please see below for details.

Response: Necessary changes have been made.

13. Are there any ethical or competing interests issues you would like to raise?

The study should adhere to ethical standards of scientific/medical research and the authors should declare that they have received ethics approval and/or patient consent for the study, where appropriate.

Whilst we do not expect reviewers to delve into authors' competing interests, if you are aware of any issues that you do not think have been adequately addressed, please inform the Editorial office.

No issues.

Detailed Revision Requests

Introduction

63 "other tetraodontid fishes such as pufferfish, boxfish and triggerfish"

KM1: This should be changed to tetraodontiform fishes to refer to the whole order and to avoid confusion with the family tetraodontidae (pufferfishes only).

Response: Suggested change incorporated.

Genome assembly and annotation

KM2: The number and sizes of the different libraries should be mentioned in the main text along with (an abbreviated version perhaps reporting only N50, contig number, scaffold number, and total size) of the assembly metrics - possibly in a brief concatenated figure combining S1, S2, and S3.

Response: The number and sizes of the different libraries are given in tables S1 and S2; this is also indicated in the Methods section with a brief mention in the main text. These tables are informative on their own and hence we do not think it is necessary to show these numbers in a separate figure.

KM3: It should be made clear in the text that the estimate of 134X coverage is based on a (later described) k-mer counting method of genome size estimation. Table S1 also says 131X coverage rather than 134X so whichever is correct should be used.

Response: We have now mentioned that the genome size was estimated by k-mer method. 134× coverage is the correct statement. 131× is a typo. We regret this oversight.

KM4: My preference would also be that the coverage statistics reported in the main text should refer to the reads actually used to produce the assembly and not the discarded data i.e. table S2 "statistics of clean reads" as this is a more accurate reflection of what was used for producing the assembly used in all downstream analyses, so 96X coverage rather than 131X.

Response: We agree that the actual reads used for the assembly should be used for estimating the genome coverage. We have now consistently mentioned 96× coverage in the revised manuscript.

KM5: The number of reads from each library actually used to produce the assembly should be reported so it is clear how much of the "clean" 68.87Gb was used by the assembler and how much was discarded. If this isn't available as a direct output from SOAPdenovo, all of the clean reads should be realigned to the genome assembly (e.g. using bwa or bowtie) and it should be reported what proportion of the clean reads align uniquely and concordantly to the genome assembly. This will also give a good idea of the completeness of the assembly.

Response: We do not believe that it is important to identify and report how many clean reads were used for the assembly. What matters is the overall quality of the assembly. As indicated by our QC based on CEGMA (and now the additional BUSCO analysis), the genome assembly is of high quality.

KM6: The parameters used for the SOAPdenovo assembly need to be stated. Justification for why these parameters and not others were used should be given, even if you only decided on these parameters a posteriori after comparing assemblies. Did you try a range of parameters and compare assembly metrics? Did you try a range of assembly programs? If yes this should be stated and summarized as a supplementary table. If not then it needs to be made clear that you only produced one assembly and did not compare, but the parameters you used still need to be stated.

Response: SOAPdenovo is one of the best assembly programs currently available (see its performance reported by Earl et al. 2011, Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res. 21: 2224-2241) and has been widely used to generate high-quality assemblies of several complex vertebrate genomes (e.g. Ruiqiang Li et al. 2010, The sequence and de novo assembly of the giant panda genome. Nature 463: 311-317; Qiu et al. 2012, The yak genome and adaptation to life at high altitude. Nature Genetics 44: 946949; Zhan et al. 2013, Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. Nature Genetics 45: 563566; Wang et al. 2013, The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nature Genetics 45: 701706; Yim et al. 2014, Minke whale genome and aquatic adaptation in cetaceans. Nature Genetics 46: 8892). The SOAPdenovo parameters used in our assembly are similar to those previously used for generating high quality assembly of

some other genomes (e.g. Xinxin You et al. 2014, Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. Nat. Commun. 5:5594). Thus, we do not see the necessity of trying a range of parameters and comparing the metrics which normally one would perform for a newly-developed program.

For the information of the reviewer, we used the following parameters: K=23, map_len=32 for libraries with insert lengths shorter than 1 kb; and map_len=35 for libraries with inserts longer than 1 kb. These parameters are now mentioned in the Methods section (under Genome sequencing and assembly).

KM7: The programs used for filtering, trimming and/or correcting the raw reads need to be stated along with the thresholds for calling a read or a base "low quality" and discarding it.

Response: Reads with more than 40 bases with quality scores less than 8 (Phred+64), or containing more than 10 Ns were defined as low quality and removed. We have mentioned this in the manuscript (Methods: Genome sequencing and assembly section).

KM8: More detailed results from the CEGMA analysis should be provided. Did you identify 98.4% predicted 'full-length' proteins, or only partial proteins? Please report both values. Although I think CEGMA is still a useful tool, the authors should note that CEGMA is no longer maintained by the creators and they have released an alternative (BUSCO): http://www.acgt.me/blog/2015/5/18/goodbye-cegma-hello-busco

Response: CEGMA actually reported 99.6% of genes as complete. The figure 98.4% indicated in the manuscript is incorrect. We have corrected this error. We have now indicated in the manuscript that 99.6% of CEGMA genes are complete in the assembly. In addition, as suggested by the reviewer, we have carried out BUSCO analysis as well. BUSCO analysis revealed the presence of 74% complete and 18% partial vertebrate BUSCO orthologues in the assembly. We have included the BUSCO results as well in the manuscript.

KM9: Given that the assembly comprised 642Mb (88%) of an estimated 730Mb genome estimated by the authors using a kmer counting method, it would be useful to have some discussion of sunfish genome size estimated by other methods e.g. flow cytometry, see (Rainerd,

E.L.L.B. et al., 2001. Patterns of Genome Size Evolution in Tetraodontiform Fishes.55(11), pp.2363-2368) and some personal communications by the authors themselves communicated in T. Ryan Gregory's genome size database (http://www.genomesize.com/) which both suggest even larger genome sizes for sunfish. A stringent realignment of the clean reads to the genome assembly should also give an idea of what proportion of the read data has been used by the assembly and what proportion has been discarded.

Response: We had estimated the genome size of sunfish using flow cytometry as 0.97 pg [Venkatesh et al. (2000) FEBS Lett. 476: 3-7] whereas Brainerd et al. (2001), who also used flow cytometry, estimated it as 0.85 pg. Flow cytometry gives only an approximate estimation of the genome size in terms of pg and the result varies depending on the state of the blood used, duration of storage, etc. We believe that it is irrelevant to discuss these results in our current manuscript.

The genome size estimated by k-mer distribution of millions of NGS sequences is a better approximation of the actual genome size. Thus the 730 Mb genome size estimated by the k-mer method in this manuscript is likely to be closer to the actual genome size. An approximate relationship between pg of DNA in a haploid cell and the genome length is 1.2 pg = 1 Gb (Flow cytometry estimation of the chicken genome size is 1.2 pg and the assembled chicken genome is 1.07 Gb). Using this approximation, 0.97 pg translates to 860 Mb and 0.85 pg translates to 758 Mb. As you can see, the latter flow cytometric estimation (0.85 pg) is not far from the genome size estimated by the k-mer method.

KM10: For the estimation of genome size using k-mer analysis, please state the tools used to make the calculation. How was the depth of 17mers counted? Is this an output of SOAPdenovo or another program like jellyfish?

Response: We used the C/C++ program Genomic Character Estimator (GCE) (Liu, B. et al. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. http://arxiv.org/abs/1308.2012v1) for k-mer analysis. GCE counts the depth of 17-mers in 17 bp sliding windows with 1 bp step length throughout each read. It estimates genome size based on Poisson theory as well as the depth of each 17-mer. We have now cited this reference in the manuscript.

97 "The sunfish genome comprises approximately 11% repetitive sequences,

98 which is comparable to the repeat content of the fugu genome (Figure 1)."

KM11: It could be made clearer in the main text if the figure of 11% refers to interspersed repeats only or is a combination including transposable elements, tandem repeats, and simple-sequence repeats. A breakdown of transposable element composition by type should be accessible from the RepeatMasker runs already carried out and would enhance this analysis and should be included in the supplementary data.

Response: 11% refers to the combined set of transposable elements, tandem repeats, and simple-sequence repeats. This is now mentioned in the manuscript. We have also added a new table (Additional file 1_Table S6) which gives the breakup of different types of repetitive elements in various teleosts.

99 "Using homology-based and de novo annotation methods, we predicted 19,605 protein-coding genes

100 in the sunfish assembly"

KM12: The type of homology-based and de novo annotation methods should be mentioned in the main text (i.e. tBLASTn against protein predictions from 5 genomes and AUGUSTUS). In the methods it should be described what the cut-off thresholds for tBLASTn alignments were and what criteria for annotating the sunfish homolog were used (i.e. where more than one protein aligned did you choose the one with the greatest length, %ID, E-Value?) Because the final gene set merged with GLEAN also contains AUGUSTUS please also report the sensitivity and specificity of the AUGUSTUS parameters chosen during the training.

Response: Details of the parameters used at different steps of annotation are now given in the Methods section. In brief: 1e-5 cut-off for tBLASTn searches; 0.25 cut-off for Genewise alignment; for AUGUSTUS training 1000 high quality sunfish gene predictions were used.

101 "Using a genome-wide set of 1,690 one-to-one

102 orthologs in sunfish and seven other ray-finned fishes (fugu, Tetraodon, stickleback, medaka,

103 tilapia, zebrafish and spotted gar), we reconstructed a phylogenetic tree and estimated the

104 divergence times of various fish lineages using MCMCtree [8]."

KM13: It needs to be clearly stated how this set of 1,690 one-to-one orthologs was chosen and verified. Ensembl is a large database with many types of export tools. Please specify the tools used and the thresholds/criteria used for defining one-to-one orthologs. Please also report the genome assembly and annotation version for each genome separately rather than the Ensembl release version. A supplementary file containing the gene names and accession numbers for each of the additional ray-finned fish genes and the corresponding sunfish gene model numbers used to form each cluster would be necessary to make this analysis reproducible.

Response: We used the Ensembl Biomart (release 76) to extract one-to-one orthologues for fugu, Tetraodon, stickleback, medaka, tilapia, zebrafish and spotted gar. InParanoid was used to identify sunfish-fugu one-to-one orthologues. We next compared the Biomart (seven fishes) one-to-one orthologue set to the sunfish-fugu InParanoid orthologues to get the 1,690 one-to-one orthologues for the eight fishes. InParanoid was run at default settings, details of which have now been mentioned in the Methods (Phylogenetic reconstruction and divergence time estimation section). Additionally, we have also included the assembly version for each of the fishes in the same section. Further, as suggested by the reviewer, an additional file containing the gene IDs of the 1,690 one-to-one orthologues from the eight fishes (Additional file 2_Table S7) has been included.

Figure 1

KM14: The bootstrap support for each of the nodes in the tree should be reported on the figure. The figure (preferably) or at least the legend needs to specify which assembly and annotation version of each of the genomes reported are being used to source the values for genome size, repeat content, and number of genes. If the repeat content comes from your own analysis rather than the published genomes this should be made clear as well. The value of 1.3% for the spotted gar repeat content is very different from the reported value of 20% from the gar genome paper (Braasch, I. et al., 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics, 48(4)) and this should be double-checked.

Response: We have now indicated the bootstrap support values at the nodes in the tree. The versions of the assembly and annotation used are indicated in the Methods section.

Thanks for pointing out the incorrect repeat content of the gar genome. As per the gar genome paper (Braasch et al. 2016), it is 26%. We have corrected this error.

Population size history

KM15: Having never carried out such analyses my expertise is limited here but I would appreciate a very brief explanation of the core methodology of PSMC in the text or methods and a brief justification of its use highlighting its potential strengths and weaknesses. Preferably cite one or two examples that show that PSMC analysis is appropriate for comparing genomes which diverged >50mya rather than 250 thousand years (over two orders of magnitude difference) as this seems like it might be problematic.

Response: PSMC (pairwise sequentially Markovian coalescent) model is a coalescent-based hidden Markov model which can be used to estimate the history of effective population sizes based on genome-wide diploid sequence data and was proposed by Li and Durbin (2011). This method has been used to understand the population history of vertebrates beyond 1 million years (Nadachowska-Brzyska et al. 2015. Temporal Dynamics of Avian Populations during Pleistocene Revealed by Whole-Genome Sequences. Curr. Biol. 25: 13751380; You et al. 2014. Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. Nat. Commun. 5: 5594; Liu et al. 2016. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nat. Commun. 7: 11757). One drawback of this method is that since only two sequences are modelled, the coalescent event joining the sequences at the most recent common ancestor is almost always older than 20 thousand years ago (see Schiffels and Durbin, 2014 Nat. Gen. 46: 919-925), so PSMC can only infer population size estimates beyond 20 kya. However, since we are looking at the population history of sunfish over a much longer period, this limitation is not an issue for our analysis. We have given a brief introduction to this method in the Methods section and have cited the above-mentioned publications in which this method was used for inferring demographic history beyond 25 kya.

Positively-selected and fast-evolving genes

127 "Using a set of 10,660 one-to-one orthologues from five teleost species (sunfish, fugu,

128 Tetraodon, medaka and zebrafish) we conducted positive selection analyses"

KM16: Calling this 10,660 gene set 'one-to-one orthologues' is confusing as it contains multiple paralogs present in different quantities in different teleost genomes. It should be described how many sunfish paralogs are found in each case, and whether the subsequent selection analyses

used the teleost 'a' or 'b' paralogy groups as the sunfish genes do not seem to be classified within the teleost 'a' or 'b' paralogy groups. For example, insulin growth factor 1 receptor (igf1r) is present as 2 paralogs in fugu, Tetraodon, medaka and zebrafish (igf1ra, igf1rb) but only one sunfish homolog (Sunfish09150) is reported in the selection analyses (Table S6, S7). Is this the ortholog of igf1ra or igf1rb? Table S8 suggests 2 copies of igf1r are found in sunfish and reports 2 dN/dS values and LRT p-values but doesn't distinguish which is which. Furthermore, the LRT p-values reported in table S6 and S7 don't correspond with those reported in table S8 (5.78x10-4 for the one igf1r

paralog presented in S6, and 3.64x10-7, 2.3 x10-3 for the two igf1r paralogs presented in S8). It would help if the sunfish gene models were annotated with 'a' or 'b' if this has been assessed - and if orthology hasn't been assessed calling them (1 of 2) and (2 of 2) would be more appropriate. If different paralogs, rather than orthologs, were used in any alignments the dN/dS estimations and inferences of evolutionary rates are meaningless so it is crucial that the methods used to assess orthology are careful and clearly described.

Response: As stated above, we used Reciprocal Best Hits (RBH) strategy for identifying this set of orthologues. However, as rightly suggested by the reviewer, we have named this set as teleost homologues. RBH does identify orthologues fairly accurately (but still we will call this set homologues) for single copy (a or b) as well as duplicate copies (a and b) of genes in the genome. In the case of genes of interest, particularly those with duplicate copies (e.g. igf1ra and igf1rb), we generated phylogenetic trees to assign orthology (a and b) to the two copies of the gene. Accordingly, we have updated Tables S8 (now Additional file 1_Table S10) and S9 (now Additional file 1_Table S11) and indicated adjusted LRT p-values separately for a and b copies.

395 "We picked

396 up genes whose likelihood values of H1 are significantly larger (LRT p-value of <0.05) than

397 H0 and likelihood values of H2 are not significantly larger than H1."

KM17: During the hypothesis testing it would also be more appropriate to select genes whose likelihood values of H1 (sunfish evolving independently from rest of the tree) are significantly greater than both H0 (all branches evolving at the same rate) and H2 (all branches evolving independently) before then sorting from this set which sunfish genes have a larger . It would also be interesting to report which sunfish genes have a lower as this might imply a greater amount of constraint.

Response: We used a likelihood ratio test (LRT) to compare the goodness of fit of two hypotheses. Null hypothesis is a special case of alternative hypothesis which means freedom of null hypothesis is smaller than alternative hypothesis. In general, the model with larger freedom results in a larger likelihood. As freedom of H2 is larger than H1, we could only test if likelihood value of H2 is significantly greater than H1 but not vice versa.

As suggested by the reviewer, we have shown the list of genes with lower than the rest of the tree in a new sheet of Table S8 (Additional file 3_Table S8b).

144 "Using the branch models in PAML [20], we found multiple genes in the

145 GH/IGF1 axis (ghr1, igf1r, grb2, irs1, irs2, jak2, stat5, akt3) with significantly higher dN/dS

146 values compared to other lineages, suggesting that these genes are evolving rapidly in the

147 sunfish lineage"

KM18: Contrary to the above statement, the authors are not reporting sunfish genes with significantly higher dN/dS than other lineages but rather sunfish genes for which hypothesis H1 (sunfish genes evolving at a different rate from the rest of the tree) is a significantly better hypothesis than H0 (all branches evolving equally). There are also multiple examples (both paralogs of irs1, one of the paralogs of irs2, one of the paralogs of jak2, and stat5) where the dN/dS value in sunfish is actually lower than the background dN/dS implying the sunfish genes are actually evolving slower than the background.

Response: We agree that our analysis identified genes that are evolving at a rate different from the rest of the tree. This set contains several genes (ghr1, igf1ra, ifg1rb, grb2, akt3, irs2a and jak2a) whose dN/dS values are higher than the background and these are the ones evolving at a rapid rate in the sunfish lineage. We have changed the sentence as Using the branch models in PAML (Yang Z, 1997) we found multiple genes that are evolving at a different rate from the rest of the tree. Among these, several genes in the GH/IGF1 axis (ghr1, igf1ra, ifg1rb, grb2, akt3, irs2a and jak2a) are evolving rapidly in the sunfish lineage (dN/dS values higher than the background).

Table S8. Copy number and LRT p-values of sunfish genes in the GH/IGF-1 axis.

KM19: This should be changed to "select genes in the GH/IGF-1 axis" as this is not a comprehensive list of genes involved in this pathway.

Response: We have modified the title of Table S8 (now Additional file 1_Table S10) as suggested.

131 "we identified 1117 genes that contained positively-selected sites

132 specifically in sunfish (Additional file 3: Table S7)."

KM20: The authors should report how many sites (either absolute number or proportion of coding sequence) appear to be under positive selection for each of these cases in their supplementary data. Could the authors please also clarify whether their claim that these 1117 genes contained positively-selected sites specifically in sunfish means that the sites or that the genes show signs of positive selection only in sunfish.

Response: We have now shown the number of sites under positive selection in each gene in Table S7 (now Additional file 4_Table S9).

With regard to the 1117 genes, we meant that these 1117 genes contained positively-selected sites only in sunfish.

132 "Inspection of the fast-evolving and

133 positively-selected gene sets revealed several interesting genes."

KM21: 'Positively-selected genes' should be replaced with 'genes with positively selected sites' as none of the genes showed outright signs of positive selection (dN/dS > 1).

Response: Yes, we have inserted genes containing positively-selected sites in place of positively-selected genes.

KM22: Ideally the authors would perform a type of overrepresentation analysis using for example GO or KEGG pathway terms to determine without bias whether the GH/IGF pathway, ECM components, or bone formation for example turn up more or less frequently than expected at random in their set of 'rapidly-evolving' or 'positively-selected' genes. Otherwise it should be made clear that the authors specifically looked at genes in the GH/IGF pathway and ECM. For example "we examined genes in the GH/IGF pathway" rather than "inspectionrevealed" as this implies that these genes somehow stood out form the rest of the data - which might be the case but without an overrepresentation analysis it is not clear.

Response: We have changed the sentence as follows: We examined genes involved in the growth pathway and found several fast-evolving genes and genes containing positively-selected sites.

144 "we found multiple genes in the

145 GH/IGF1 axis (ghr1, igf1r, grb2, irs1, irs2, jak2, stat5, akt3) with significantly higher dN/dS

146 values compared to other lineages, suggesting that these genes are evolving rapidly in the

147 sunfish lineage"

KM23: Again here as I understand it the analysis tested whether there was a significant difference between H1 and H0, not whether there was a significant difference in dN/dS between sunfish and other lineages. If this is a separate analysis it should be clearly stated. Furthermore several dN/dS values reported for sunfish in table S8 are actually lower than the background reported.

Response: This question is the same as KM18 and has been addressed before.

147 "We found that both copies of igf1r

148 (igf1ra and igf1rb) are under positive selection in the sunfish (Figure 2, Additional file 1: Table

149 S8)"

KM24: Here please also replace "under positive selection" with "contain sites under positive selection". The same applies to ECM analysis. If you have indeed assessed orthology with igf1ra and igf1rb please make this clear in earlier methods sections and report orthology in table S7, S8 and elsewhere.

Response: We have replaced "under positive selection" with "contain sites under positive selection" with regard to igfr genes as well as ECM genes. We have also mentioned that We found that both copies of igf1r (igf1ra and igf1rb) contain positively selected sites in the sunfish (Figure 2, Additional file 1_Table S10)

190 "However, the sunfish

191 possesses intact orthologues for most of these genes except for some SCPP genes (see

192 Supplementary Material)"

KM25: I find it disjointed that this analysis alone is described in supplementary materials. As it is integral to the motivation for conducting the study the analysis of bone forming genes should be included in the main text.

Response: We thank the reviewer for this suggestion. This was initially a single section in the main text. However, for brevity, we had moved portion of this analysis to the Supplementary Information and had retained only Scpp genes in the main text with a brief mention about the other bone-genes analyzed. As suggested by the reviewer, we have now moved back the entire bone-gene analysis section to the main text.

Additional File 1

"We identified orthologues for all the above genes in the ocean sunfish genome on (a) scaffold10.1, (b) scaffold39.1, (c) scaffold20.1, and (d) scaffold77.1, except Optc and Omd."

KM26: Please state how you identified these homologs. Did you perform tblastn, or tblastx genome wide against your assembly and what did you use as your query sequences? What were the similarity thresholds you used?

Response: We carried out reciprocal BLAST searches to identify orthologues of these genes in sunfish. We first ran BLASTP of the human and/or zebrafish proteins against the annotated ocean sunfish proteins. For genes that could not be identified using this method, we proceeded to a TBLASTN of the human and fish proteins against the ocean sunfish genome assembly followed by a BLASTX of the resulting sunfish genomic loci against NCBI NR protein database. We used the default BLAST settings in order to maximize the sensitivity of identifying possibly divergent orthologues. This strategy is now mentioned in the main text (under the section Genes involved in bone formation and subsection Proteoglycan-encoding genes).

"We BLASTX-searched the ocean sunfish loci of (a) and (b) to identify Optc and Omd respectively, but did not identify these genes."

KM27: Again, please clarify the type of BLAST algorithm you ran and the query and target sequences you used. The above statement implies you used blastx to run the sunfish scaffolds as a query against a database containing Optc and Omd protein sequences. Is this correct? Which species were the Optc and Omd proteins sourced from? What were the cutoff parameters used?

Response: We BLASTX-searched the loci against NCBI NR protein database, and not a database containing only OPTC or OMD sequences. We used the default BLAST settings. We have added this information in the main text (under the section Genes involved in bone formation and subsection Proteoglycan-encoding genes).

"An alignment of Runx2 proteins shows that ocean sunfish Runx2 is highly conserved (e.g. its DNA-binding domain is perfectly conserved; its central and C-terminal domains look intact as well) (data not shown)."

KM28: I have no reason to doubt this but if you are reporting it I suggest you show the data especially as your supplementary data is not restricted.

Response: We have now presented the Runx2 alignment as Additional file 1_Figure S3 and have accordingly cited the figure instead of data not shown.

KM29: The analysis of presence/absence of each of the target bone-formation related genes should be presented in a table (in either the main text or SI). In each case where homologs of bone-formation genes were found in sunfish the exact number of homologs found should be stated. E.g. "For Smad4, we identified up to four copies in ocean sunfish" is confusing and the exact number should be reported.

Response: We have included a supplementary table (Additional file 5_Table S12) that lists all the genes that we searched and the identifiers of the orthologues that we identified in the ocean sunfish genome.

200 "However, it has lost two P/Q-rich SCPP genes (fa93e10 and scpp7) that are conserved in the

201 other two teleosts"

KM30: Before concluding gene loss please make it clear if you have searched the whole genome assembly and not just the identified clusters for these genes, and whether you have also searched the raw genomic reads which may contain unassembled reads corresponding to the missing genes.

Response: Besides searching the genomic vicinity of Spacl1, we did indeed search the entire ocean sunfish genome using fugu and zebrafish SCPP protein sequences by TBLASTN. As suggested by the reviewer, we have now searched raw genome reads of sunfish using zebrafish fa93e10 and scpp7 proteins. While we could not find reads that matched fa93e10, we identified several reads that aligned to scpp7. However, assembling these reads and searching the consensus sequence against NCBI NR protein database showed they show high similarity to fugu scpp5. They also mapped to scpp5 gene predicted by us in the sunfish assembly. These searches provide additional support to the conclusion that fa93e10 and scpp7 genes are missing in the sunfish.

KM31: Because of the complex duplication history of SCPP genes I would consider it essential to carefully assess homology of each of the genes in the P/Q-rich SCPP gene cluster with phylogenetic methods to ensure that scpp7 is indeed lost and that additional sunfish SCPP genes reported as scpp3b1 and scpp3b2 for example are not actually orthologs of scpp4, and that the reported pseudogene of scpp4 is not in fact scpp7.

Response: As suggested by the reviewer, we generated a Maximum Likelihood (ML) tree using protein sequences of the P/Q-rich SCPP genes in ocean sunfish, fugu, medaka and zebrafish. The ML tree (see new Additional file 1_Figure S4) shows clearly that the ocean sunfish genes that we have identified as scpp3b1/b2 are indeed orthologs of scpp3 as they cluster with fugu and medaka scpp3 genes. In addition, the ocean sunfish gene that we identified as scpp4 pseudogene, clusters with fugu scpp4. Lastly, none of the ocean sunfish SCPP genes cluster with zebrafish scpp7 suggesting that this gene is indeed missing in ocean sunfish. We have mentioned the phylogenetic analysis in the main text of the manuscript (Genes involved in bone formation: SCPP section) and included the ML tree as Additional file 1_Figure S4.

KM32: To confirm that scpp4 is indeed a pseudogene and that the insertion of the "T" is not a sequencing/assembly error please report the results of a read re-mapping to this locus to verify that the additional "T" is present in most raw reads which realign to this site.

Response: We do not think that this exercise is necessary as we have confirmed this mutation by Sanger sequencing this region from two other specimens of sunfish. Presence of this mutation in three unrelated individuals of sunfish is strong evidence that this mutation is present in the population and not a sequencing error.

Hox genes

KM33: In figure S3 a more appropriate or additional outgroup for analysis of Hox clusters in teleosts would be the spotted gar, which the authors have also previously used in their own analyses. See (Braasch, I. et al., 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics, 48(4)). The figure would be ameliorated if the authors marked the independent gene losses which occurred on each branch to highlight the differences in sunfish from other teleosts. It should also be reported what scaffold numbers in the sunfish assembly each Hox cluster corresponds to, in a similar fashion as reported for SCPP genes in Figure 4.

Response: As suggested by the reviewer, we have now included the Hox clusters from spotted gar in the figure but would also like to retain coelacanth as it is a representative lobe-finned fish. The inferred gene losses in each of the species have now been presented in the figure and the inferred state of the Hox cluster in the ancestor of the four teleosts has also been shown. The scaffold numbers for the sunfish Hox clusters has been mentioned in the figure legend. We would like to point out that this figure is now Additional file 1_Figure S5.