Reviewer's report

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

Version: 0 **Date:** 01 Jun 2016

Reviewer: Gareth Fraser

Reviewer's report:

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.

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.

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.

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.

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.

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.

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.

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.

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).

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.

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.

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.

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.

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.

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.

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

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.

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?

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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).

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.

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.

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.

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.

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?

"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?

"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.

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.

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.

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.

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.

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.

Are the methods appropriate to the aims of the study, are they well described, and are necessary controls included?

If not, please specify what is required in your comments to the authors.

No

Are the conclusions adequately supported by the data shown?

If not, please explain in your comments to the authors.

Yes

Does the manuscript adhere to the journal's guidelines on <a href='http://resource-cms.springer.com/springer-cms/rest/v1/content/7117202/data/v1/Minimum+standards+of+reporting+checklist'target=' new'>minimum standards of reporting?</a>

If not, please specify what is required in your comments to the authors.

Yes

Are you able to assess all statistics in the manuscript, including the appropriateness of statistical tests used?

(If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.)

Yes, and I have assessed the statistics in my report.

Quality of written English

Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests

Please complete a declaration of competing interests, consider the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
5. Do you have any other financial competing interests?
6. Do you have any non-financial competing interests in relation to this manuscript?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal