

## Genomic and biochemical evidence of dietary adaptation in a marine herbivorous fish

Joseph Heras, Mahul Chakraborty, J. J. Emerson and Donovan P. German

### Article citation details

*Proc. R. Soc. B* **287**: 20192327.

<http://dx.doi.org/10.1098/rspb.2019.2327>

### Review timeline

Original submission: 10 April 2019  
1st revised submission: 3 October 2019  
2nd revised submission: 7 January 2020  
Final acceptance: 27 January 2020

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

## Review History

### RSPB-2019-0808.R0 (Original submission)

#### Review form: Reviewer 1

##### Recommendation

Major revision is needed (please make suggestions in comments)

**Scientific importance: Is the manuscript an original and important contribution to its field?**

Good

**General interest: Is the paper of sufficient general interest?**

Excellent

**Quality of the paper: Is the overall quality of the paper suitable?**

Good

**Is the length of the paper justified?**

Yes

**Should the paper be seen by a specialist statistical reviewer?**

No

**Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.**

No

**It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.**

**Is it accessible?**

Yes

**Is it clear?**

Yes

**Is it adequate?**

Yes

**Do you have any ethical concerns with this paper?**

No

**Comments to the Author**

See attachment. (See Appendix A)

## Review form: Reviewer 2

**Recommendation**

Accept with minor revision (please list in comments)

**Scientific importance: Is the manuscript an original and important contribution to its field?**

Good

**General interest: Is the paper of sufficient general interest?**

Good

**Quality of the paper: Is the overall quality of the paper suitable?**

Excellent

**Is the length of the paper justified?**

Yes

**Should the paper be seen by a specialist statistical reviewer?**

No

**Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.**

No

**It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.**

**Is it accessible?**

Yes

**Is it clear?**

Yes

**Is it adequate?**

Yes

**Do you have any ethical concerns with this paper?**

No

### **Comments to the Author**

This is a nicely constructed paper that provides insight on the molecular mechanisms underlying the adaptive establishment of the digestive machinery, specifically amylase and bile salt activated lipase, for the herbivorous monkeyface prickleback. I have only a few brief comments for this well written manuscript.

This is a fairly methodology-laden paper given the detailed description of assembly and sequencing. I wonder if a hunk of this could be moved to supplemental text.

In your methods on transcript assembly it is noted that differentially expressed genes are viewed via heat maps (3rd line of paragraph). It wasn't clear until I looked at the heat maps that the comparison to assess differences in gene expression was among organs. I suggest making this a little more clear here in the text.

In Physiological genomics of digestive enzymes - Line 12 of that first paragraph - this is a little unclear. You mean to say "...which tend to have one copy or two identical copies of amy2..."

Line 15 of that first paragraph - Spell out TE - transposone element.

Line 16 of this paragraph - do you mean "amy2b" rather than "amy2B"?

Line 16 of this paragraph - spell out LINE - long interspersed nuclear element

Second paragraph of that section, second line. I might be missing something here, but it is noted that "yet Xiphister taxa only have two copies of amy2a..." However, so does *C. violaceus*? Is the difference between *Xiphister* and *C. violaceus* just that *C. violaceus* possesses amy2b and *Xiphister* does not?

In Transcriptomics, it is mentioned (line 8) "Because SCFAs are ketones". I am not sure what definition you are using for ketones, but I have never encountered the labeling of common SCFAs (acetate, propionate, butyrate) as ketones in reference to ketone bodies (acetoacetate, beta-hydroxybutyrate). You may want to clarify or revise.

In the second paragraph of this section, line 6, it is noted that "the mid intestine is rarely recognized as also having pancreatic function". However other statements note the presence of acinar distributed along the small intestine. This just needs some clarification. For fish that lack pyloric ceca, the intestine may be the dominate source of pancreatic secretion.

Fig. 3. The heading of this figure notes gene expression profiles of nine tissues, though there are heat maps for just five tissues.

## Review form: Reviewer 3

### Recommendation

Major revision is needed (please make suggestions in comments)

**Scientific importance: Is the manuscript an original and important contribution to its field?**

Good

**General interest: Is the paper of sufficient general interest?**

Good

**Quality of the paper: Is the overall quality of the paper suitable?**

Good

**Is the length of the paper justified?**

Yes

**Should the paper be seen by a specialist statistical reviewer?**

No

**Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.**

No

**It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.**

**Is it accessible?**

Yes

**Is it clear?**

Yes

**Is it adequate?**

Yes

**Do you have any ethical concerns with this paper?**

No

### Comments to the Author

The manuscript: "Physiological genomics of dietary adaptation in a marine herbivorous fish" by Heras et al, presents a high-quality genome of the herbivorous fish pricklyback. It also presents transcriptomic data from a few tissues, presented as the "digestive transcriptome". Presentations of well assembled genomes has a value by itself, perusing the idea to identify genomic adaptations of the digestive system to new alimentary niche makes the manuscript quite appealing. The ms is also well written.

This is an interesting topic and my major criticism is that I would wish a more holistic approach to the adaptation to an herbivorous (or algivorous) life. Expressions of digestive enzymes are quite plastic and is to a certain degree modulated by the quantity of a nutrient in the food stuff. However, the organism's ability to transport and metabolize different nutrients may be less plastic. For example, the ability to regulate sugars in the blood. Anyway, this ms is about digestion, so let's focus on that.

## Introduction

Digestion and nutrition are pretty large fields to cover. Even though the genomics of digestion in fishes is a fairly new area, there are a few studies out there. E.g. Wang et al., 2010 (intestinal transcriptome in zebrafish), De Santis et al., 2015 on Atlantic salmon Lie et al., genome and intestinal transcriptome on ballan wrasse are a few. Please present state of knowledge on fish in the intro.

## Material and methods

Could the authors provide information on the feeding status of the individuals used for the rna seq experiments? Was there ingesta in the stomach and or intestine?

Please provide RIN values for the rna samples, or the range will do. I'm not a bioinformatician but to my knowledge the method seems sound. I'm impressed with the long N50 achieved.

## Results and discussion

According to HUGO Gene Nomenclature Committee the correct annotation for BSAL is CEL (carboxyl ester lipase) and small letters is usually used for non-human species. Please check the gene nomenclature.

The section "Physiological genomics of digestive enzymes" interestingly point out the difficulties with descriptive science and hypothesis we come up with to fit our findings. The genomic presence of amylases may be explained with or fits the hypothesis of "Adaptive Modulation", when it comes to lipases the auteurs conveniently turn to "Nutrient Balancing". Admittedly this is briefly addressed at the end of the section, but I would wish a deeper reflection on the subject. Another possibility for the additional CEL (or BSAL) gene is that it might be more efficient to hydrolyze galactolipids or betaine lipids that are found in the cell membranes of algae? I am also curious to why there is no focus on phospholipid digestion. Fig 3b includes the membrane bound PLB1 but what about sPLA2 IB ? The pancreatic secretory phospholipase hydrolysing dietary PL into lysoPL has been shown to be important in lipid digestion in fishes.

I am a bit puzzled that the authors have only addressed the digestion of two out of three macronutrients. And now I am generous since digestion of structural lipids have not been addressed, as I pointed out before. What happened to the proteases? After all the dietary protein in such a diet will probably be the limiting factor for growth.

## Transcriptomics

The selection of tissues sequenced seem incomplete considering that digestion was the focus. There is heart, spleen, gill gonad and brain but no stomach or pancreas. I'm not familiar with the anatomy of this particular species, but if it resembles other teleosts I guess the pyloric caeca sample will comprise some pancreas. Actually, it is mentioned in page 16. The authors should provide a better overview of the digestive anatomy than is presented in Fig. 2. If you want to say something about digestion, why look at pepsinogen expression in liver, intestine or spleen, when its digestive role is in the stomach. Please explain why the stomach and therefore gastric digestion has not been analysed and discussed.

It is also important to separate between enzymes that has a role in digestion and enzymes involved in cellular metabolism. E.g. only a few of the lipases listed in Fig. 3 are involved in digestion. Pnpla2 hydrolyses TAG in intracellular lipid droplets and hl is mainly involved in hydrolysis of plasma lipoproteins. Please provide an overview that specifies genes coding for enzymes that actually take part in digestion of ingesta. There is also good reason to conclude that only bsal2 is involved in digestion, considering the expression pattern of the different bsal genes.

Based on the expression pattern of the cel (bsal) genes, I'm not convinced that all of them are involved in digestion. Based on the evidence presented it is not possible to conclude that the cel genes described support the Nutrient Balancing Hypothesis.

## Decision letter (RSPB-2019-0808.R0)

31-May-2019

Dear Dr Heras:

I am writing to inform you that your manuscript RSPB-2019-0808 entitled "Physiological genomics of dietary adaptation in a marine herbivorous fish" has, in its current form, been rejected for publication in Proceedings B.

This action has been taken on the advice of referees, who have recommended that substantial revisions are necessary. With this in mind we would be happy to consider a resubmission, provided the comments of the referees are fully addressed. However please note that this is not a provisional acceptance.

The resubmission will be treated as a new manuscript. However, we will approach the same reviewers if they are available and it is deemed appropriate to do so by the Editor. Please note that resubmissions must be submitted within six months of the date of this email. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office. Manuscripts submitted after this date will be automatically rejected.

Please find below the comments made by the referees, not including confidential reports to the Editor, which I hope you will find useful. If you do choose to resubmit your manuscript, please upload the following:

- 1) A 'response to referees' document including details of how you have responded to the comments, and the adjustments you have made.
- 2) A clean copy of the manuscript and one with 'tracked changes' indicating your 'response to referees' comments document.
- 3) Line numbers in your main document.

To upload a resubmitted manuscript, log into <http://mc.manuscriptcentral.com/prsb> and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Resubmission." Please be sure to indicate in your cover letter that it is a resubmission, and supply the previous reference number.

Sincerely,  
Dr Daniel Costa  
<mailto:proceedingsb@royalsociety.org>

Associate Editor  
Board Member: 1  
Comments to Author:

This is an interesting paper, which was liked by the reviewers. However, they identify several points with need for improvement and clarification. A re-submission should address all these concerns.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)  
See attachment

Referee: 2

Comments to the Author(s)

This is a nicely constructed paper that provides insight on the molecular mechanisms underlying the adaptive establishment of the digestive machinery, specifically amylase and bile salt activated lipase, for the herbivorous monkeyface prickleback. I have only a few brief comments for this well written manuscript.

This is a fairly methodology-laden paper given the detailed description of assembly and sequencing. I wonder if a hunk of this could be moved to supplemental text.

In your methods on transcript assembly it is noted that differentially expressed genes are viewed via heat maps (3rd line of paragraph). It wasn't clear until I looked at the heat maps that the comparison to assess differences in gene expression was among organs. I suggest making this a little more clear here in the text.

In Physiological genomics of digestive enzymes - Line 12 of that first paragraph - this is a little unclear. You mean to say "...which tend to have one copy or two identical copies of amy2.."

Line 15 of that first paragraph - Spell out TE - transposone element.

Line 16 of this paragraph - do you mean "amy2b" rather than "amy2B"?

Line 16 of this paragraph - spell out LINE - long interspersed nuclear element

Second paragraph of that section, second line. I might be missing something here, but it is noted that "yet Xiphister taxa only have two copies of amy2a..." However, so does *C. violaceus*? Is the difference between Xiphister and *C. violaceus* just that *C. violaceus* possesses amy2b and Xiphister does not?

In Transcriptomics, it is mentioned (line 8) "Because SCFAs are ketones". I am not sure what definition you are using for ketones, but I have never encountered the labeling of common SCFAs (acetate, propionate, butyrate) as ketones in reference to ketone bodies (acetoacetate, beta-hydroxybutyrate). You may want to clarify or revise.

In the second paragraph of this section, line 6, it is noted that "the mid intestine is rarely recognized as also having pancreatic function". However other statements note the presence of acinar distributed along the small intestine. This just needs some clarification. For fish that lack pyloric ceca, the intestine may be the dominate source of pancreatic secretion.

Fig. 3. The heading of this figure notes gene expression profiles of nine tissues, though there are heat maps for just five tissues.

Referee: 3

Comments to the Author(s)

The manuscript: "Physiological genomics of dietary adaptation in a marine herbivorous fish" by Heras et al, presents a high-quality genome of the herbivorous fish prickleback. It also presents transcriptomic data from a few tissues, presented as the "digestive transcriptome". Presentations of well assembled genomes has a value by itself, perusing the idea to identify genomic adaptations of the digestive system to new alimentary niche makes the manuscript quite appealing. The ms is also well written.

This is an interesting topic and my major criticism is that I would wish a more holistic approach to the adaptation to an herbivorous (or algivorous) life. Expressions of digestive enzymes are

quite plastic and is to a certain degree modulated by the quantity of a nutrient in the food stuff. However, the organism's ability to transport and metabolize different nutrients may be less plastic. For example, the ability to regulate sugars in the blood. Anyway, this ms is about digestion, so let's focus on that.

### Introduction

Digestion and nutrition are pretty large fields to cover. Even though the genomics of digestion in fishes is a fairly new area, there are a few studies out there. E.g. Wang et al., 2010 (intestinal transcriptome in zebrafish), De Santis et al., 2015 on Atlantic salmon Lie et al., genome and intestinal transcriptome on ballan wrasse are a few. Please present state of knowledge on fish in the intro.

### Material and methods

Could the authors provide information on the feeding status of the individuals used for the rna seq experiments? Was there ingesta in the stomach and or intestine?

Please provide RIN values for the rna samples, or the range will do. I'm not a bioinformatician but to my knowledge the method seems sound. I'm impressed with the long N50 achieved.

### Results and discussion

According to HUGO Gene Nomenclature Committee the correct annotation for BSAL is CEL (carboxyl ester lipase) and small letters is usually used for non-human species. Please check the gene nomenclature.

The section "Physiological genomics of digestive enzymes" interestingly point out the difficulties with descriptive science and hypothesis we come up with to fit our findings. The genomic presence of amylases may be explained with or fits the hypothesis of "Adaptive Modulation", when it comes to lipases the auteurs conveniently turn to "Nutrient Balancing". Admittedly this is briefly addressed at the end of the section, but I would wish a deeper reflection on the subject. Another possibility for the additional CEL (or BSAL) gene is that it might be more efficient to hydrolyze galactolipids or betaine lipids that are found in the cell membranes of algae? I am also curious to why there is no focus on phospholipid digestion. Fig 3b includes the membrane bound PLB1 but what about sPLA2 IB ? The pancreatic secretory phospholipase hydrolysing dietary PL into lysoPL has been shown to be important in lipid digestion in fishes.

I am a bit puzzled that the authors have only addressed the digestion of two out of three macronutrients. And now I am generous since digestion of structural lipids have not been addressed, as I pointed out before. What happened to the proteases? After all the dietary protein in such a diet will probably be the limiting factor for growth.

### Transcriptomics

The selection of tissues sequenced seem incomplete considering that digestion was the focus. There is heart, spleen, gill gonad and brain but no stomach or pancreas. I'm not familiar with the anatomy of this particular species, but if it resembles other teleosts I guess the pyloric caeca sample will comprise some pancreas. Actually, it is mentioned in page 16. The authors should provide a better overview of the digestive anatomy than is presented in Fig. 2. If you want to say something about digestion, why look at pepsinogen expression in liver, intestine or spleen, when its digestive role is in the stomach. Please explain why the stomach and therefore gastric digestion has not been analysed and discussed.

It is also important to separate between enzymes that has a role in digestion and enzymes involved in cellular metabolism. E.g. only a few of the lipases listed in Fig. 3 are involved in digestion. Pnpla2 hydrolyses TAG in intracellular lipid droplets and hl is mainly involved in hydrolysis of plasma lipoproteins. Please provide an overview that specifies genes coding for enzymes that actually take part in digestion of ingesta. There is also good reason to conclude that only bsal2 is involved in digestion, considering the expression pattern of the different bsal genes.



Based on the expression pattern of the cel (bsal) genes, I'm not convinced that all of them are involved in digestion. Based on the evidence presented it is not possible to conclude that the cel genes described support the Nutrient Balancing Hypothesis.

## Author's Response to Decision Letter for (RSPB-2019-0808.R0)

See Appendix B.

## RSPB-2019-2327.R0

### Review form: Reviewer 4

#### **Recommendation**

Major revision is needed (please make suggestions in comments)

#### **Scientific importance: Is the manuscript an original and important contribution to its field?**

Acceptable

#### **General interest: Is the paper of sufficient general interest?**

Marginal

#### **Quality of the paper: Is the overall quality of the paper suitable?**

Acceptable

#### **Is the length of the paper justified?**

Yes

#### **Should the paper be seen by a specialist statistical reviewer?**

No

#### **Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.**

No

**It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.**

#### **Is it accessible?**

Yes

#### **Is it clear?**

Yes

#### **Is it adequate?**

Yes

## Do you have any ethical concerns with this paper?

No

### Comments to the Author

This manuscript presents the genome and transcriptome sequences of a herbivorous fish species, *Cebidichthys violaceus*. The authors focus on the adaptation to a herbivorous diet, and thus present evolutionary analyses for specific digestive enzymes, involved in the metabolism of carbohydrates and lipids.

The manuscript provides a valuable genome and transcriptome resource. The genome was sequenced using both long-read (PacBio) and short-read (Illumina) techniques, which resulted in good assembly statistics, comparable with or better than those of other teleost fish species. As a single individual was analyzed, the transcriptome resource is less rich, and does not allow statistical analyses of differential gene expression.

The evolutionary analyses presented in the manuscript are restricted to the candidate genes that encode digestive enzymes. In my opinion, the manuscript needs genome-wide evolutionary analyses, if only to provide an appropriate context for the digestive enzymes data. The authors have made a considerable effort to identify orthologous genes between *C. violaceus* and other fish species. These gene families and the corresponding sequence alignments could be used to identify positive selection events on a genome-wide scale. One could thus ask whether the observations made for amylase and carboxyl ester lipase are exceptional. Related to this, given that all biological analyses focus on a small set of genes, the title of the manuscript is misleading – I would not call this study “physiological genomics”.

As for the dN/dS analyses, gene duplication analyses are also only performed for specific digestive enzymes. A broader evolutionary context is needed here. In particular, the whole-genome duplication that is believed to have occurred in the ancestor of teleost fishes should be addressed. What kind of genes have been preserved in two copies or lost in the *C. violaceus* genome? Are these results consistent with what was previously observed for other teleost species?

This manuscript could also benefit from including genome-wide analyses of transposable element content, non-coding RNA repertoires, conserved non-coding sequences – all of which are possible given the genome and transcriptome sequence data that was generated, and the publicly available resources for other teleost fishes.

## Decision letter (RSPB-2019-2327.R0)

06-Nov-2019

Dear Dr Heras:

Your manuscript has now been peer reviewed again by a different referee from the initial round. This referee points out some important issues that need to be addressed. Modifying the paper according to the issues suggested by the referee will make the paper stronger as well as increasing its interest to the general reader. The reviewers' comments (not including confidential comments to the Editor) are included at the end of this email for your reference. We normally do not allow more than one round of review. However, I think your paper has sufficient merit to allow one final revision.

This will be your final opportunity to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into <http://mc.manuscriptcentral.com/prsb> and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions", click on "Create a Revision". Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" in the "File Upload" section. This should document, point by point, how you have responded to the reviewers' and Editors' comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (<https://royalsociety.org/journals/ethics-policies/>). You should pay particular attention to the following:

#### Research ethics:

If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

#### Use of animals and field studies:

If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

#### Data accessibility and data citation:

It is a condition of publication that you make available the data and research materials supporting the results in the article. Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (<https://royalsociety.org/journals/ethics-policies/data-sharing-mining/>). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (<http://datadryad.org/>) and have not already done so you can submit your data via this link [http://datadryad.org/submit?journalID=RSPB&manu=\(Document not available\)](http://datadryad.org/submit?journalID=RSPB&manu=(Document not available)), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy <http://royalsocietypublishing.org/data-sharing>.

#### Electronic supplementary material:

All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,  
Dr Daniel Costa  
mailto:proceedingsb@royalsociety.org

#### Reviewer(s)' Comments to Author:

Referee: 4

#### Comments to the Author(s).

This manuscript presents the genome and transcriptome sequences of a herbivorous fish species, *Cebidichthys violaceus*. The authors focus on the adaptation to a herbivorous diet, and thus present evolutionary analyses for specific digestive enzymes, involved in the metabolism of carbohydrates and lipids.

The manuscript provides a valuable genome and transcriptome resource. The genome was sequenced using both long-read (PacBio) and short-read (Illumina) techniques, which resulted in good assembly statistics, comparable with or better than those of other teleost fish species. As a single individual was analyzed, the transcriptome resource is less rich, and does not allow statistical analyses of differential gene expression.

The evolutionary analyses presented in the manuscript are restricted to the candidate genes that encode digestive enzymes. In my opinion, the manuscript needs genome-wide evolutionary analyses, if only to provide an appropriate context for the digestive enzymes data. The authors have made a considerable effort to identify orthologous genes between *C. violaceus* and other fish species. These gene families and the corresponding sequence alignments could be used to identify positive selection events on a genome-wide scale. One could thus ask whether the observations made for amylase and carboxyl ester lipase are exceptional. Related to this, given that all biological analyses focus on a small set of genes, the title of the manuscript is misleading – I would not call this study “physiological genomics”.

As for the dN/dS analyses, gene duplication analyses are also only performed for specific digestive enzymes. A broader evolutionary context is needed here. In particular, the whole-genome duplication that is believed to have occurred in the ancestor of teleost fishes should be addressed. What kind of genes have been preserved in two copies or lost in the *C. violaceus* genome? Are these results consistent with what was previously observed for other teleost species?

This manuscript could also benefit from including genome-wide analyses of transposable element content, non-coding RNA repertoires, conserved non-coding sequences – all of which are possible given the genome and transcriptome sequence data that was generated, and the publicly available resources for other teleost fishes.

## Author's Response to Decision Letter for (RSPB-2019-2327.R0)

See Appendix C.

## Decision letter (RSPB-2019-2327.R1)

27-Jan-2020

Dear Dr Heras

I am pleased to inform you that your manuscript entitled "Genomic and biochemical evidence of dietary adaptation in a marine herbivorous fish" has been accepted for publication in *Proceedings B*.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact [procb\\_proofs@royalsociety.org](mailto:procb_proofs@royalsociety.org)

### Open Access

You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700.

Corresponding authors from member institutions

(<http://royalsocietypublishing.org/site/librarians/allmembers.xhtml>) receive a 25% discount to these charges. For more information please visit <http://royalsocietypublishing.org/open-access>.

Your article has been estimated as being 10 pages long. Our Production Office will be able to confirm the exact length at proof stage.

#### Paper charges

An e-mail request for payment of any related charges will be sent out after proof stage (within approximately 2-6 weeks). The preferred payment method is by credit card; however, other payment options are available

#### Electronic supplementary material:

All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,

Dr Daniel Costa

Editor, Proceedings B

mailto:proceedingsb@royalsociety.org

## **A**

### **Review for Proceedings B: ‘Physiological genomics of dietary adaptation in a marine herbivorous fish.’**

Heras et al. provide a genomic perspective on the physiological adaptations to an herbivorous diet. Apart from generating a high quality genome assembly for their study species, they also explore the molecular underpinnings of its adaptation to herbivory. This is a very interesting study with clear objectives and hypotheses. The analyses have been performed thoroughly and the conclusions are sound. Moreover, I think that this topic will appeal to a wide readership and it thus suitable for publication in Proceedings B. However, I do have a few concerns about the structure of the text and some questions that need clarification.

#### *Introduction*

The introduction could be structured better. As it reads now, it jumps between explaining the (molecular) adaptations to herbivory and the objectives/outcomes of this study. Moreover, in the Results and Discussion section, the authors introduce two general hypotheses (Adaptive Modulation Hypothesis and Nutrient Balancing Hypothesis) to explain their findings. I would suggest restructuring the introduction and adding a paragraph on these two hypotheses. Also, I would move the section about the genome assembly to the end and thus have one or two paragraphs that describe the objectives/outcomes of this study. This will provide the reader with the necessary background to understand the remainder of the paper.

#### *Methods*

The methods are very clearly described, including all parameter settings (which enables replication of the results). However, I do have a few questions that need further explanation.

- Which perl-script did you use to trim the contig?
- What settings did you use in MUSCLE to align the orthologous sequences?
- Provide some more information about the way aBSREL and MEME work. This is relevant to understand some of the figures.
- Why only sequences with 60 amino acids or more? I understand you need sequences that are sufficiently long, but is there a rationale behind the threshold of 60? Or is it just an arbitrary choice?
- Which model did you use to generate the phylogenetic tree? Did you run a model testing analysis?

- You compare syntenic regions between certain species. I assume these have been selected because they have been assembled on a chromosome level. Please mention this for clarity, some readers might wonder why you used these particular species.

### *Results and Discussion*

The Results and Discussion section has a clear structure, although I would put the section on transcriptomics second (it is now the last one). This will result in a better flow, going from genome assembly over main transcriptomic patterns to candidate genes. Moreover, you refer to some of the transcriptomic data when discussing the candidate genes.

You find that *C. violaceus* has two copies of *amy2a* and one copy of *amy2b*. Next, you state that other prickleback species have one or two identical copies of *amy2*. How reliable are the assemblies of these species? You specifically focused on these genes and found copy number variation.

In each of the candidate genes (amylase and lipase), you find positions under positive selection. Several positions represent non-synonymous substitutions. These amino acid changes could be interesting from a functional point of view. The observation that AMY2A and AMY2B have different isoelectric points suggests changes in protein structure. Can you elaborate on this in the text or add the amino acid changes in a supplementary text? The same goes for BSA lipase.

Given the long discussion on the BSA lipase, I suggest moving Figure S11 to main text.

In the transcriptomic section, you mention that *C. violaceus* uses microbial fermentation. How did you discriminate between microbial and fish proteins in the analyses? This is not clear from the text. It might be useful to add a few lines to the method section to address this issue.

I miss the relevance of the paragraph on the pyloric cecal tissue at the end of the Results and Discussion section. This topic had not been properly introduced and seems unrelated to the rest of the manuscript. Moreover, you only refer to supplementary figures here. I suggest linking this paragraph better with the rest of the text or perhaps deleting it.



## *Figures*

In general, the figures are nicely constructed and easy to follow. Just a few minor comments.

In figure 3, you can highlight the candidate genes (amylase and BSA lipase) so the reader can immediately see where they are expressed (as discussed in the main text).

In figure 4a, some words have red lines. I assume it has been copied from Microsoft Word or Powerpoint. Please correct this.

## **Minor comments**

### *Introduction*

Why is the term 'herbivorous' controversial amongst ichthyologists? This is a bit confusing for non-ichthyologists and might require more background. Perhaps better to remove this statement.

You mention *X. mucosus* without providing the complete genus-name.

### *Methods*

No need to mention the adaptor indexes used in the RNA sequencing. Referring to Supplementary Table S3 should be sufficient.

Mention that CUMMERBUND is an R-library. Do you have a reference for it?

### *Results and Discussion*

The use of *amy2b* is not consistent in the text. You also mention *amy2B*.

I noticed that the numbering of supplementary figures is not correct.

## **B**

### **Heras et al. Response to Reviewers**

Associate Editor Board Member: 1 Comments to Author: This is an interesting paper, which was liked by the reviewers. However, they identify several points with need for improvement and clarification. A re-submission should address all these concerns.

Reviewer(s)' Comments  
to Author:

#### **Referee: 1**

Comments to the Author(s)

Heras et al. provide a genomic perspective on the physiological adaptations to an herbivorous diet. Apart from generating a high quality genome assembly for their study species, they also explore the molecular underpinnings of its adaptation to herbivory. This is a very interesting study with clear objectives and hypotheses. The analyses have been performed thoroughly and the conclusions are sound. Moreover, I think that this topic will appeal to a wide readership and it thus suitable for publication in Proceedings B. However, I do have a few concerns about the structure of the text and some questions that need clarification.

#### *Introduction*

The introduction could be structured better. As it reads now, it jumps between explaining the (molecular) adaptations to herbivory and the objectives/outcomes of this study. Moreover, in the Results and Discussion section, the authors introduce two general hypotheses (Adaptive Modulation Hypothesis and Nutrient Balancing Hypothesis) to explain their findings. I would suggest restructuring the introduction and adding a paragraph on these two hypotheses.

We added the Adaptive Modulation Hypothesis and Nutrient Balancing Hypothesis to the introduction and explained their relevance to amylase and lipase activities, respectively (lines: 42-55).

*“The Adaptive Modulation Hypothesis (17) suggests a positive correlation between digestive enzyme activities and ingested quantity of the substrate for those enzymes. Based on economic principles, this hypothesis is well supported in the literature for carbohydrases, as herbivorous and omnivorous animals tend to have elevated amylase activities in their guts (11, 18-23), and achieve these activities largely via gene duplications (22). The Nutrient Balancing Hypothesis (24) suggests that there can be elevated expression of enzymes towards limiting dietary resources to ensure acquisition of essential nutrients, like essential fatty acids. Indeed, elevated carboxyl ester lipase activities are observed in fishes consuming low-lipid, high-fiber*

*foods (11, 25-26). In this investigation, we sequenced the genome of the herbivorous fish Cebidichthys violaceus, which revealed extensive genetic variation and adaptive amino acid variation for amylase and carboxyl ester lipase, suggesting multiple mechanisms underlying the novel derived dietary physiology in C. violaceus, and simultaneously supporting the Adaptive Modulation and Nutrient Balancing Hypotheses in the same organism.”*

We also added a supplementary figure (Supplementary Figure S1) to help explain these non-mutually exclusive hypotheses.

Also, I would move the section about the genome assembly to the end and thus have one or two paragraphs that describe the objectives/outcomes of this study. This will provide the reader with the necessary background to understand the remainder of the paper.

We have moved our description of the genome assembly to the bottom of our introduction section (lines: 75-85). In addition, we have outlined our objectives in the last paragraph of our introduction. This includes assembling a draft genome and transcriptomic datasets. We also outline the use of our transcriptomic datasets for the annotation of the genome and as a resource for gene expression of multiple tissue types. We also include a statement about estimating episodic diversifying selection for candidate genes and identifying gene copy number for each candidate gene within our draft genome. Lastly, we included a statement about our comparative genomic analysis of syntenic regions to better understand the conservation order of chromosome information of multiple fish genomes relative to the *C. violaceus* genome.

## *Methods*

The methods are very clearly described, including all parameter settings (which enables replication of the results). However, I do have a few questions that need further explanation.

Which perl-script did you use to trim the contig?

We used a perl-script called FASCUT, which is part of the FAST Analysis of Sequences Toolbox (Lawrence et al. 2015), which is composed of multiple perl scripts used for sequence analysis (lines: 136-138).

*“We then used FASCUT, a perl script that is part of the FAST Analysis of Sequences Toolbox (38) to trim the contig that contained amylase loci and neighboring genes.”*

What settings did you use in MUSCLE to align the orthologous sequences?

We used default settings for the alignment and we reviewed the alignments by eye and used only alignments with the lowest average gaps present in each alignment (lines: 174-175).

Provide some more information about the way aBSREL and MEME work. This is relevant to

understand some of the figures.

We have added a couple of statements about aBSREL and MEME which is described in the selection analysis that we conducted (lines: 146-153).

*“Selection was estimated using branch-site models and using adaptive Branch Site REL (aBSREL), a branch-site model that infers the optimal number of  $\omega$  (nonsynonymous/synonymous rate ratio) classes for each branch, testing whether a proportion of sites have evolved under positive selection. Next, a Mixed Effects Model of Evolution (MEME) was used to test the individual sites subject to episodic positive or diversifying selection, and a signatures of recombination Genetic Algorithm for Recombination Detection (GARD) was used as part of the datamonkey v2.0 web application (42).”*

Why only sequences with 60 amino acids or more? I understand you need sequences that are sufficiently long, but is there a rationale behind the threshold of 60? Or is it just an arbitrary choice?

Lines: 170-173

We used 60 amino acids or more because we wanted to ensure a robust phylogenetic analysis with ample sequence information. Also, by providing sequences with 60 amino acids or more is intended to remove putative sequence fragments. This method has been used in other studies which use transcriptomic data for identifying orthologous sequences (Heras et al. 2015; Liu et al. 2018).

References:

Heras J, McClintock K, Sunagawa S, Aguilar A. 2015 Gonadal transcriptomics elucidate patterns of adaptive evolution within marine rockfishes (Sebastes). BMC Genomics 16. (doi:10.1186/s12864-015-1870-0)

Liu H et al. 2018 Draft genome of Glyptosternon maculatum, an endemic fish from Tibet Plateau. GigaScience 7. (doi:10.1093/gigascience/giy104)

Which model did you use to generate the phylogenetic tree? Did you run a model testing analysis?

We used jmodeltest v2.1.0 for model testing and GTR+I+G was selected under the AICc criteria. We used this model for phylogenetic analysis. All this information has been included in our methods section (lines: 175-178).

*“GTR+I+G was selected as the best model of sequence evolution (jmodeltest v2.1.0; 43). Sequences were concatenated with sequencematrix 1.8 (47). Maximum likelihood (ML) phylogenetic trees were constructed with PhyML v3.1 (44) with the GTR+I+G model and*

*1,000 bootstrap replicates.”*

You compare syntenic regions between certain species. I assume these have been selected because they have been assembled on a chromosome level. Please mention this for clarity, some readers might wonder why you used these particular species.

*We selected these four species based on their quality draft genomes (e.g. assembled on a linkage group or chromosome level) and their broad evolutionary distance in respect to the *C. violaceus* genome (lines: 178-181).*

*“Syntenic regions of the *C. violaceus* genome was compared to *Gasterosteus aculeatus*, *Oryzias latipes*, *Danio rerio*, and *Lepisosteus oculatus* (Supplemental Material) based on their quality draft genomes and broad evolutionary distance with respect to *C. violaceus*.”*

### *Results and Discussion*

The Results and Discussion section has a clear structure, although I would put the section on transcriptomics second (it is now the last one). This will result in a better flow, going from genome assembly over main transcriptomic patterns to candidate genes. Moreover, you refer to some of the transcriptomic data when discussing the candidate genes.

*The transcriptomics section has been moved to the second section of the Results and Discussion section (lines: 207-237).*

You find that *C. violaceus* has two copies of *amy2a* and one copy of *amy2b*. Next, you state that other prickleback species have one or two identical copies of *amy2*. How reliable are the assemblies of these species? You specifically focused on these genes and found copy number variation.

In each of the candidate genes (amylase and lipase), you find positions under positive selection.

*In German, D. P., D. M. Foti, et al. (2016 Physiol Biochem Zool 89: 277-293), the estimate of gene copy number was conducted by real-time quantitative polymerase chain reaction and not through genomic assemblies. Moreover, German et al. (2016) used Native PAGE to detect each AMY protein, and no AMY2B was present in any prickleback species other than *C. violaceus*. From this present study, nucleotide sequences of *amy2* were generated with RNA extractions from pyloric cecal tissues for transcriptomic analyses. Thus, the lack of *amy2B* is inferred from previous investigations of amylase genes in these fishes (German et al. 2016), and by the lack of any *amy2a* transcripts in any species other than *C. violaceus*.*

Several positions represent non-synonymous substitutions. These amino acid changes could be interesting from a functional point of view. The observation that AMY2A and AMY2B have

different isoelectric points suggests changes in protein structure. Can you elaborate on this in the text or add the amino acid changes in a supplementary text?

We detail the comparisons of the AMY2A and AMY2B proteins in German et al. (2016). Thus, we do not detail them again in this manuscript.

The same goes for BSA lipase.

We have significantly expanded our discussion of carboxyl ester lipase (*cel*), which is what other reviewers have requested that we call this gene and protein. The main *cel* figure (Fig. 4) is now in the main article with additional *cel* figures and data (Supplementary figure S29-S31; table S13) in the supplements. We removed the prickleback phylogeny (what was Fig. 2) and put it in the supplements (Supplementary figure S2) to make room for the *cel* figure.

In the transcriptomic section, you mention that *C. violaceus* uses microbial fermentation. How did you discriminate between microbial and fish proteins in the analyses? This is not clear from the text. It might be useful to add a few lines to the method section to address this issue.

These methods are detailed in the following articles, which are cited in the current manuscript:

German, D. P. and R. A. Bittong (2009). "Digestive enzyme activities and gastrointestinal fermentation in wood-eating catfishes." *Journal of Comparative Physiology B* 179: 1025-1042.

German, D. P. et al. (2015). "More than one way to be an herbivore: convergent evolution of herbivory using different digestive strategies in prickleback fishes (family Stichaeidae)." *Zoology* 118: 161-170.

The measurements of microbial fermentation were made with gas chromatography, whereas the inference about microbially-produced enzymes comes from the fractions in which the enzymes were measured (tissue vs contents), and which gut region shows the most activity. Enzymes that spike in the distal intestine, are likely of microbial origin, given that this is the gut region with the most microbial fermentation. This is also detailed in German et al. (2015) cited above. We can make this clearer in the manuscript and have adjusted the writing.

I miss the relevance of the paragraph on the pyloric cecal tissue at the end of the Results and Discussion section. This topic had not been properly introduced and seems unrelated to the rest of the manuscript. Moreover, you only refer to supplementary figures here. I suggest linking this paragraph better with the rest of the text or perhaps deleting it.

This topic is important because it emphasizes why we focused on the pyloric cecal region (it has dense pancreatic tissue). We added an image of the *C. violaceus* intestine to Fig. 2 panel c, and have incorporated more of this earlier in the manuscript (lines: 227-237).

*"The pyloric caecal tissue of C. violaceus and other pricklebacks is recognized as*

*“pancreatic” because it is sheathed in acinar cells (9) and shows elevated activity levels of pancreatic enzymes (11, 22). However, this tissue only has two differentially expressed genes in comparison to the mid intestine (Supplemental Figure S15), which is a highly absorptive region of the fish gut (55-56). Although pyloric caeca have been documented as absorptive (i.e., similar function to the mid intestine; 57), the mid intestine is rarely recognized as also having pancreatic function. In fishes, pancreatic tissue can be embedded in the liver (forming a hepatopancreas) or diffuse along the proximal intestine, particularly in fishes with pyloric caeca (58). Our transcriptomic and biochemical data suggest an absorptive function of the pyloric caeca, and that the acinar cells are distributed at least down to the mid intestine and not restricted to the pyloric caecal region in C. violaceus (Supplemental Figures S16-S17).”*

### *Figures*

In general, the figures are nicely constructed and easy to follow. Just a few minor comments.

In figure 3, you can highlight the candidate genes (amylase and BSA lipase) so the reader can immediately see where they are expressed (as discussed in the main text).

Amylase and carboxyl ester lipase genes and other candidate genes were denoted with (\*) for clarity in Fig. 2.

In figure 4a, some words have red lines. I assume it has been copied from Microsoft Word or Powerpoint. Please correct this.

This figure, which is now Fig. 3, has been corrected and updated.

### *Minor comments*

#### *Introduction*

Why is the term ‘herbivorous’ controversial amongst ichthyologists? This is a bit confusing for non-ichthyologists and might require more background. Perhaps better to remove this statement.

We removed this sentence. This statement stems from Clements et al. (2017) Biol J Linn Soc, in which parrotfishes are argued to be microphages on cyanobacteria. Many fish ecologists, especially those working on coral reefs, struggle with the point that herbivory in fishes is a complicated topic and that not all herbivores are the same (e.g., German et al. 2015). But, in the interest of saving space and not opening a can of worms in this concise work for PRSB, we deleted the contested statement.

You mention *X. mucosus* without providing the complete genus-name.

The genus name *Xiphister* is now spelled out at the first occurrence of this word, and referred to *X. mucosus* throughout the manuscript.

## Methods

No need to mention the adaptor indexes used in the RNA sequencing. Referring to Supplementary Table S3 should be sufficient.

We have removed the text about adaptor indexes IDs in the manuscript and refer to the IDs in the Supplementary Table S8, which provides the nucleotide sequence for each of the adaptors.

Mention that CUMMERBUND is an R-library. Do you have a reference for it?

We have updated this section with the appropriate citation and appropriate term, which can be found in the supplemental materials (line 1217-1219)

*“Differentially expressed genes for all tissue types were viewed with a heatmap that was generated with the cummerbund R package (<http://comphio.mit.edu/cummeRbund/>; Supplemental Figure S13).”*

## Results and Discussion

The use of *amy2b* is not consistent in the text. You also mention *amy2B*.

We have made edits where *amy2a* and *amy2b* is consistent throughout the entire manuscript. We are sticking with the convention of lowercase names for genes, all CAPS names for proteins.

## Referee: 2

Comments to the Author(s) This is a nicely constructed paper that provides insight on the molecular mechanisms underlying the adaptive establishment of the digestive machinery, specifically amylase and bile salt activated lipase, for the herbivorous monkeyface pricklyback. I have only a few brief comments for this well written manuscript.

This is a fairly methodology-laden paper given the detailed description of assembly and sequencing. I wonder if a hunk of this could be moved to supplemental text.

In your methods on transcript assembly it is noted that differentially expressed genes are viewed



via heat maps (3rd line of paragraph). It wasn't clear until I looked at the heat maps that the comparison to assess differences in gene expression was among organs. I suggest making this a little more clear here in the text.

This has been clarified in the text (lines: 1217-1219).

*“Differentially expressed genes for all tissue types were viewed with a heatmap that was generated with the cummerbund R package (<http://compbio.mit.edu/cummeRbund/>; Supplemental Figure S13).”*

In Physiological genomics of digestive enzymes - Line 12 of that first paragraph – this is a little unclear. You mean to say “...which tend to have one copy or two identical copies of amy2..”

We have clarified this statement (lines: 247-252).

*“C. violaceus has three tandem pancreatic amylase genes: two copies of amy2a and one copy of amy2b (Fig. 3). The three C. violaceus amy genes in tandem differs from other pricklebacks (and most other fishes for which genetic data are available), which tend to have one or two identical copies of amy2 (Fig. 3; 22). The two amy2a copies in C. violaceus are supported by three spanning reads, emphasizing the correct assembly of the amy2a tandem duplicates (Supplemental Figure S25).”*

Line 15 of that first paragraph – Spell out TE – transposone element.

At the first occurrence, we have changed TE to transposable element (TE) and refer to TE throughout the remainder of the manuscript (line: 254).

Line 16 of this paragraph – do you mean “amy2b” rather than “amy2B”?

We have updated and consistently use *amy2a* and *amy2b* throughout our manuscript.

Line 16 of this paragraph – spell out LINE – long interspersed nuclear element

We have updated our manuscript and spell out LINE at the first occurrence (line: 255-256).

Second paragraph of that section, second line. I might be missing something here, but it is noted that “yet Xiphister taxa only have two copies of amy2a...” However, so does *C. violaceus*? Is the difference between Xiphister and *C. violaceus* just that *C. violaceus* possesses amy2b and Xiphister does not?

Yes, that is correct. The difference is that *C. violaceus* has *amy2b*.

In Transcriptomics, it is mentioned (line 8) “Because SCFAs are ketones”. I am not sure what definition you are using for ketones, but I have never encountered the labeling of common SCFAs (acetate, propionate, butyrate) as ketones in reference to ketone bodies (acetoacetate, beta-hydroxybutyrate). You may want to clarify or revise.

We revised this sentence to be more specific as follows (lines: 214-216): “Because the two most commonly produced SCFAs acetate and propionate are largely metabolized to ketones in the liver, animals reliant on hindgut fermentation tend to have active ketotic pathways in their tissues (8, 52).”

In the second paragraph of this section, line 6, it is noted that “the mid intestine is rarely recognized as also having pancreatic function”. However other statements note the presence of acinar distributed along the small intestine. This just needs some clarification. For fish that lack pyloric ceca, the intestine may be the dominate source of pancreatic secretion.

We clarified this by stating that the acinar cells are largely recognized as being distributed in the proximal intestine in species with pyloric ceca. Our results show that the acinar cells can extend at least down to the mid intestine in *C. violaceus*. We stated that more explicitly in this section.

Lines: 227-229

*“The pyloric caecal tissue of C. violaceus and other pricklebacks is recognized as “pancreatic” because it is sheathed in acinar cells (9) and shows elevated activity levels of pancreatic enzymes (11, 22).”*

Lines: 234-237

*“Our transcriptomic and biochemical data suggest an absorptive function of the pyloric caeca, and that the acinar cells are distributed at least down to the mid intestine and not restricted to the pyloric caecal region in C. violaceus (Supplemental Figures S16-S17).”*

Fig. 3. The heading of this figure notes gene expression profiles of nine tissues, though there are heat maps for just five tissues.

We actually generated transcriptomic datasets for nine tissues, but in Fig. 2, we are only showing five of them (those relevant to digestion, plus the spleen). We have rephrased the description of Fig. 2 to convey that gene expression profiles for liver, pyloric caeca, proximal intestine, middle intestine, and spleen tissues are represented in this figure.

**Referee: 3**

Comments to the Author(s) The manuscript: “Physiological genomics of dietary adaptation in a marine herbivorous fish ” by Heras et al, presents a high-quality genome of the herbivorous fish prickleback. It also presents transcriptomic data from a few tissues, presented as the “digestive transcriptome”. Presentations of well assembled genomes has a value by itself, perusing the idea to identify genomic adaptations of the digestive system to new alimentary niche makes the manuscript quite appealing. The ms is also well written.

This is an interesting topic and my major criticism is that I would wish a more holistic approach to the adaptation to an herbivorous (or algivorous) life. Expressions of digestive enzymes are quite plastic and is to a certain degree modulated by the quantity of a nutrient in the food stuff. However, the organism’s ability to transport and metabolize different nutrients may be less plastic. For example, the ability to regulate sugars in the blood. Anyway, this ms is about digestion, so let’s focus on that.

Introduction Digestion and nutrition are pretty large fields to cover. Even though the genomics of digestion in fishes is a fairly new area, there are a few studies out there. E.g. Wang et al., 2010 (intestinal transcriptome in zebrafish), De Santis et al., 2015 on Atlantic salmon Lie et al., genome and intestinal transcriptome on ballan wrasse are a few. Please present state of knowledge on fish in the intro.

We have added the mentioned citations (and a few more) to better reflect the state of the literature.

(Lines: 33-37)

*“Animal digestion is an ideal model phenotype because it is central to fitness, is understood in many species at genetic, molecular, biochemical, and physiological levels, and is variable throughout animal evolution (8-11). Untangling the genetic basis of digestion is contingent on quality genomic resources (12), which have traditionally been lacking in non-model species (13-14).”*

References 12-14 are Wang et al., 2010, De Santis et al., 2015, Lie et al., 2018

Material and methods Could the authors provide information on the feeding status of the individuals used for the rna seq experiments? Was there ingesta in the stomach and or intestine?

Yes. We added the following line (lines: 1202-1204) to the methods (supplemental materials):

*“All individuals had digesta in their guts during dissection (i.e., they had all eaten), and digesta was removed prior to tissue fixation in RNAlater®.”*

Please provide RIN values for the rna samples, or the range will do.

We have included our RIN criteria to the methods section. We have included the RIN values in Supplementary Table S8 as well.

I'm not a bioinformatician but to my knowledge the method seems sound. I'm impressed with the long N50 achieved.

So are we!

## Results and discussion

According to HUGO Gene Nomenclature Committee the correct annotation for BSAL is CEL (carboxyl ester lipase) and small letters is usually used for non-human species. Please check the gene nomenclature.

We changed BSAL to CEL at the reviewer's request. We chose BSAL originally because these enzymes are known as BSAL in the fish digestive literature. But, we agree that to be consistent across fields, CEL is more correct.

The section "Physiological genomics of digestive enzymes" interestingly point out the difficulties with descriptive science and hypothesis we come up with to fit our findings. The genomic presence of amylases may be explained with or fits the hypothesis of "Adaptive Modulation", when it comes to lipases the auteurs conveniently turn to "Nutrient Balancing". Admittedly this is briefly addressed at the end of the section, but I would wish a deeper reflection on the subject. Another possibility for the additional CEL (or BSAL) gene is that it might be more efficient to hydrolyze galactolipids or betaine lipids that are found in the cell membranes of algae? I am also curious to why there is no focus on phospholipid digestion. Fig 3b includes the membrane bound PLB1 but what about sPLA2 IB ? The pancreatic secretory phospholipase hydrolysing dietary PL into lysoPL has been shown to be important in lipid digestion in fishes.

We agree with the reviewer. We have expanded the lipolytic enzymes up to 15 (from the original eight in the version the reviewer read). We have included phospholipases A2 and B1, as well as pancreatic triacylglycerol lipase (Supplemental Figures S27-S28; Fig. 2b). From all of this, it is clear that the only interesting pattern remains with *cel*. We also agree that whether one invokes Adaptive Modulation Hypothesis or Nutrient Balancing Hypothesis is a matter of scale. On a gross mass level, algae are low in lipid (<10%), and thus, herbivorous fishes with elevated lipolytic activities certainly fit within the Nutrient Balancing Hypothesis. However, if one focuses on the lipids only, and the fact that galactolipids and betaine lipids, both of which would be hydrolyzed by CEL, are the most common lipids in algae, then an increase in *cel* gene copy number in an herbivorous fish actually probably better fits the Adaptive Modulation Hypothesis. This discrepancy is now addressed in the discussion (lines 273-285). We thank the reviewer for bringing this to our attention. But, CEL is still recognized as the most important digestive lipase in fishes, even in carnivores, so we don't feel we should abandon the Nutrient Balancing Hypothesis.

I am a bit puzzled that the authors have only addressed the digestion of two out of three macronutrients. And now I am generous since digestion of structural lipids have not been

addressed, as I pointed out before. What happened to the proteases? After all the dietary protein in such a diet will probably be the limiting factor for growth.

The reviewer raises a good point: what is going on with the proteases? In our original analysis, we included 10 proteases. Our analysis now includes 16. None of the proteolytic enzyme genes show any convincing signs of selection, but chymotrypsin gene copy number (4) is of interest, and that is addressed in the discussion. Our analyses of aminopeptidases took the longest because it is quite interesting. Although there are no obvious signs of selection on aminopeptidases in prickleback fishes, it appears that aminopeptidases in fishes show signs of retention following whole genome duplication events. This is noted on the ohnolog website <http://ohnologs.curie.fr>. *Danio rerio*, for instance, appears to have three different loci for aminopeptidases, as does *C. violaceus* and other fishes. Aminopeptidases are recognized as ohnologs in *D. rerio* and *O. latipes*, but to our knowledge, this has not been investigated in terms of what that means for digestion. Thus, we now provide four new supplemental figures (supplemental figures S21-S24) that address aminopeptidases and their syntenic regions within various fish genomes. Our phylogenetic analyses of these genes suggests that *aminopeptidase a* may be the original chordate aminopeptidase, and that subsequent duplication events led to *aminopeptidase b*, *aminopeptidase N*, and *aminopeptidase EY*, all of which show gut expression. Each of these proteins needs to be examined more closely to understand what functional differences exist amongst them. This is also discussed in the Supplemental Discussion section (section V - *Evaluation of Candidate Genes Associated with Digestion*).

**Transcriptomics** The selection of tissues sequenced seem incomplete considering that digestion was the focus. There is heart, spleen, gill gonad and brain but no stomach or pancreas. I'm not familiar with the anatomy of this particular species, but if it resembles other teleosts I guess the pyloric caeca sample will comprise some pancreas. Actually, it is mentioned in page 16. The authors should provide a better overview of the digestive anatomy than is presented in Fig. 2.

We have now added an image of the *C. violaceus* gut to Fig. 2 so that the different gut regions are more obvious to the reader, and make it clear that pancreatic cells (acinar cells) are sheathed around the pyloric caeca.

If you want to say something about digestion, why look at pepsinogen expression in liver, intestine or spleen, when its digestive role is in the stomach. Please explain why the stomach and therefore gastric digestion has not been analysed and discussed.

We agree that including pepsinogen in the liver and intestine is not a worthwhile investigation. Pepsinogen has been removed from Fig. 2. However, we are quite interested in the stomach (and pepsinogens), but that is the focus of a separate, comparative transcriptomics study that investigates pepsins and chitinases. We have presented this work at a recent Society for Integrative and Comparative Biology conference: German, D.P., M.H. Herrera, and J. Heras (2018) Can you stomach it? Comparative transcriptomics of the stomachs of prickleback fishes

(Stichaeidae) consuming different diets. Integrative and Comparative Biology 58 Supplement 1: E74. Because that is its own dataset that we are preparing for publication elsewhere, we will not add it into this current submission for PRSB. Pepsin has been removed from the heatmap since stomach is not included.

It is also important to separate between enzymes that has a role in digestion and enzymes involved in cellular metabolism. E.g. only a few of the lipases listed in Fig. 3 are involved in digestion. Pnpla2 hydrolyses TAG in intracellular lipid droplets and hl is mainly involved in hydrolysis of plasma lipoproteins. Please provide an overview that specifies genes coding for enzymes that actually take part in digestion of ingesta. There is also good reason to conclude that only bsal2 is involved in digestion, considering the expression pattern of the different bsal genes.

We thank the reviewer for this comment. We have removed lipases (including pnpla2) that are not part of the digestive process. Although we included more lipolytic genes in the analysis, we still focus on *cel*.

Based on the expression pattern of the *cel* (bsal) genes, I'm not convinced that all of them are involved in digestion. Based on the evidence presented it is not possible to conclude that the *cel* genes described support the Nutrient Balancing Hypothesis.

We agree that with this current analysis, we cannot conclusively state which CEL proteins are involved in digestion. We do call for more research in this area involving isolated CEL proteins and more detailed assay methods (e.g., pH stat) that would discern what types of fatty acids are best hydrolyzed by these enzymes. Of the *cel* genes, *cel2* and *cel-like* show clear gut expression, whereas *cel-1a*, *b*, and *c* have more ubiquitous expression patterns. It may be that CEL-1 is a lysosomal protein, but this will require more specific analyses that are beyond the scope of this paper. We did include an amino acid sequence alignment of all the CEL proteins (Supplemental Figure S31) to show the similarities amongst the proteins, including the bile-salt binding domain.

Referee: 4

As you will see below, our reply to referee 4 can be broadly seen as justifying why we view the suggestions as either off topic or beyond that to which we can reply. We don't wish to imply that the reviewer's suggestions aren't valuable (they are) or that we don't appreciate the time that the reviewer invested in reviewing our manuscript (we do). However, as we outline below, most of the suggestions don't pertain to digestive physiology and/or greatly expand the scope in ways we are simply incapable of addressing.

Comments to the Author(s).

This manuscript presents the genome and transcriptome sequences of a herbivorous fish species, *Cebidichthys violaceus*. The authors focus on the adaptation to a herbivorous diet, and thus present evolutionary analyses for specific digestive enzymes, involved in the metabolism of carbohydrates and lipids.

The manuscript provides a valuable genome and transcriptome resource. The genome was sequenced using both long-read (PacBio) and short-read (Illumina) techniques, which resulted in good assembly statistics, comparable with or better than those of other teleost fish species. As a single individual was analyzed, the transcriptome resource is less rich, and does not allow statistical analyses of differential gene expression.

The evolutionary analyses presented in the manuscript are restricted to the candidate genes that encode digestive enzymes. In my opinion, the manuscript needs genome-wide evolutionary analyses, if only to provide an appropriate context for the digestive enzymes data. The authors have made a considerable effort to identify orthologous genes between *C. violaceus* and other fish species. These gene families and the corresponding sequence alignments could be used to identify positive selection events on a genome-wide scale. One could thus ask whether the observations made for amylase and carboxyl ester lipase are exceptional. Related to this, given that all biological analyses focus on a small set of genes, the title of the manuscript is misleading – I would not call this study “physiological genomics”.

While we agree with the reviewer that genome-wide perspectives have great value, we disagree with the reviewer's conclusion that the manuscript needs such a genome-wide evolutionary analysis. The implication of this request is that a genome-wide analysis is the best way to know whether there has been dietary-related selection on the genomic regions on which we focus. This is not the case. This study is primarily a study on digestive physiology with an emphasis on the genomics and molecular evolutionary aspects of well-known digestive genes. Irrespective of how the rest of the genome is evolving, we are interested in understanding the digestive physiology of the derived herbivore *C. violaceus*. Indeed, one of the strengths of our approach is that our study design was pre-specified to address gene families we knew *a priori* are relevant to nutritional physiology. This focus is justified in many ways. It permits us to target a strictly circumscribed number of genes. This not only allows us to manually curate alignments and troubleshoot problems directly, it also obviates the application of controlling for thousands of statistical tests and the attendant false discoveries inherent in unfocused analyses, such as those requested by the reviewer. Moreover, the statistical tools we applied to interrogate the evolution of the gene families we chose have their own internal logic and long history of application that are not



dependent on placing special value on exceptional cases. It could be a brute fact that many other gene families unrelated to digestion or nutrient metabolism evolve faster than the families we selected. But this still would not invalidate the observations we made or their implications for the evolution of digestive physiology. In short, while we reiterate that genome-wide studies are valuable, such analyses add little to our digestive physiology-focused study.

Another concern we have in addressing the comments of reviewer 4 has to do with the scope of work implied by the reviewer's suggestion. Genome-wide analysis requires not only a well-curated genome in our study taxon, but also in all genomes to which we compare it. The level of curation done manually for our *a priori* digestive and metabolic genes of interest is literally impossible to do in a reasonable time frame on the whole genome in all taxa. Manual curation of thousands of multiple alignments would simply not be an option. Approaches that require less supervision offer viable alternatives, but they would necessitate substantially more functional genomic data in all taxa to arrive at several independent rigorous annotations that would be less dependent on manual curation. For example, the rigorous comparative genome-wide scans contrasting dogs and wolves benefited from well-curated genomes in these well-studied taxa (e.g., Axelsson et al. 2013 *Nature* 495 360-364). To accomplish this in fishes would require considerable time and resources that we don't have. Finally, it is well-known that molecular evolutionary analyses are extremely sensitive to misspecification of homologous nucleotide residues via errors in the multiple alignment (e.g. Markova-Raina and Petrov. 2011. *Genome Research* 21 (6): 863–74). This concern is even more pronounced when the residues aren't even homologous as would certainly happen when performing automated alignment of gene models derived from draft annotations available to us for the taxa we included in our analyses.

Moreover, it isn't clear what we would be looking for even if we were able to perform a genome-wide analysis. The reviewer suggests that such work provides "appropriate context". We disagree. As described above, the internal logic of the comparative methods we employ permits molecular evolutionary inference on the genes directly (e.g. for amino acid evolution: Nei, M., and T. Gojobori. 1986. *Molecular Biology and Evolution* 3 (5): 418–26; Li, W. H. 1993. *Journal of Molecular Evolution* 36 (1): 96–99; Yang, Z. 1998. *Molecular Biology and Evolution* 15 (5): 568–73; Yang, Z., and R. Nielsen. 2000. *Molecular Biology and Evolution* 17 (1): 32–43). This is well-trodden ground, and an ecosystem of comparative approaches appropriate for carrying out robust inference on single genes exists that does not rely on genome-wide analysis to provide context. Given this, it is unclear what other scientific questions in digestive physiology a genome-wide approach would be addressing.

Perhaps most importantly, the requested analysis outlines an entirely separate study that is only tenuously tied to the one on which we embarked and wrote about. The other three reviewers of this manuscript accepted the premise and context of our original submission, which was based on a rich nutritional ecology dataset for *C. violaceus* (spanning back to the 1970's). Neither the editor nor reviewers expressed concerns that we were drilling down to the genetic underpinnings of ecological, physiological, and biochemical abilities known in this species. The reviewer appears to want us to also take an unfocused genome-wide approach to observe where selection has occurred in the genome, neglecting the rich array of nutritionally-relevant hypotheses based on years of physiological data that drives our work. And, while valuable, these additional revisions recommended by the reviewer 4 address fundamentally different questions than those



we seek to address. As such, their review amounts to requesting we complete another study on a topic not of our choosing. The focus on nutritional physiology outlined in our cover letter was not seen as a limitation when the paper was initially sent out for review, nor were analyses similar to those requested by reviewer 4 suggested by three other reviewers in the first round of revision. Hence, while we again reiterate the value in the suggestions given by the reviewer, we submit that they are not crucial to our study and greatly exceed the scope of manuscript revisions, especially coming as they do in the second round of review.

We propose to change the title to “Genomic and biochemical evidence of dietary adaptation in a marine herbivorous fish”, which removes the term “Physiological Genomics”, but still captures the essence of the paper, retains the word “genomic” (which was an important part of our effort), and will still be broadly appealing to readers from diverse scientific backgrounds.

As for the dN/dS analyses, gene duplication analyses are also only performed for specific digestive enzymes. A broader evolutionary context is needed here. In particular, the whole-genome duplication that is believed to have occurred in the ancestor of teleost fishes should be addressed. What kind of genes have been preserved in two copies or lost in the *C. violaceus* genome? Are these results consistent with what was previously observed for other teleost species?

Please see our response above why we believe that genome-wide analysis is not necessary for context. However, we did notice “Ohnolog” (i.e., whole genome duplication) signatures for the aminopeptidase genes, and we wrote about this in the supplemental discussion section (starting on line 1385, Supplemental Figures S21-S24). Regrettably, due to Proc B. page limits, we don’t have space to devote a great deal of attention this in the main text, but it is indeed already part of the supplemental section. We are not aware of any other Ohnolog signatures for other digestive enzyme genes in *C. violaceus* or other fishes. Although a deeper dissection of this topic is of interest for future analyses, it deserves its own study. However, we do insert a pointer to this observation in the main manuscript (Page 12 of the main document, lines 232-235).

This manuscript could also benefit from including genome-wide analyses of transposable element content, non-coding RNA repertoires, conserved non-coding sequences – all of which are possible given the genome and transcriptome sequence data that was generated, and the publicly available resources for other teleost fishes.

We have now added a brief overview of the repeat content. Interestingly, ~92% of the TEs we found were not present in the existing repeat database of the fugu genome (Page 10 of the main document, lines 200-203).

One of the strengths of our manuscript is that, in order to answer physiological questions, we generated a high-quality assembly that will serve as a valuable resource to the fish genomics community. And we expect that our assembly will be used by the community to explore features like TEs, ncRNAs, CNSes, and many more. However, we would like to note again that the reviewer has shifted focus away from digestive physiology. Moreover, while all of these analyses are certainly valuable and of great interest to the community, we also note that reviewer 4 has listed three additional genome-wide annotation efforts that each require domain-specific

knowledge and expertise to perform rigorously. Two of the authors here (Chakraborty and Emerson) study structural variation and genome assembly and have recently completed a comparative analysis of 4 closely related species of *Drosophila*. Compared to the current study, the *Drosophila* study benefits from dealing with simpler genomes, higher quality assemblies, and more comprehensive functional data. And the scope of the reviewer's suggestions for TEs alone in *Drosophila* is vast, to say nothing of performing analogous analyses for ncRNAs and CNSes. To suggest all three in a vertebrate genome is beyond what we are capable of performing in this revision.

As a result, while we are happy to describe the broad features of the TE annotation employed in the manuscript, we must unfortunately point out again that suggesting that we expand the scope of the manuscript to address untargeted analyses of ncRNAs and CNSes is beyond the scope of a digestive physiology manuscript, especially given that such suggestions are just now being raised in the second round of review.