Revision round #2

Decision for round #2 : *Revision needed*

Revisions needed

---

by **Jitendra Narayan**, 19 May 2024 13:50

Manuscript: **https://doi.org/10.1101/2023.11.30.569369**
version: 4

The authors effectively responded to the ideas made in the initial review, painstakingly implementing the majority of the recommendations to improve the manuscript's reproducibility and clarity.

We thank Dr Narayan for the positive feedback to our revisions and address the remaining points in the following.

Note to reviewers: All referenced lines in this response align with the track changes document, making it easier for reviewers to follow the changes that have been made following their recommendations.

While great progress has been achieved, there are a few areas that may be improved. Specifically, explaining the use of FCS for contaminant screening and removing mitochondrial sequences from genome assembly. In accordance with the reviewers' recommendations, this would considerably improve research transparency.

We added explanatory text for contamination screening in the following sentences to the method section: lines 222-225: “To determine whether any contaminant sequences were present in the genome, we screened for adapter, vector, and foreign sequences using the NCBI Foreign Contamination Screen (FCS) v0.2.1 and FCS-GX v0.3.0) (Astashyn et al. 2024). Resulting hits indicative of contaminants were removed from the assembled sequences”. Concerning the mitochondrial sequence, we added lines 241-245: “A mitochondrial assembly was created using MitoHiFi v2.0 [https://github.com/marcelauliano/MitoHiFi] using the DeepConsensus PacBio Reads and *Phoxinus phoxinus* reference mitochondrial genome NC_020358.1 as input. Any partial mitochondrial contigs remaining in the assembly were removed based on mapping synteny to the fully assembled mitogenome.”

Furthermore, adding documentation to better describe the scripts used is needed.

All scripts used for genome assembly , genome-wide comparison, gene family evolution and demographic history are available from Zenodo (https://doi.org/10.5281/zenodo.12191118) To improve access, we now added explanations on script usage to the README file contained in this repository. Here we include the example of the script used for post-protein-coding gene prediction analyses, named “Busco_Diamond_blast.sh”:

```
### 1. Busco_Diamond_blast.sh

### Description:

This script performs three post-protein-coding gene prediction
analyses, including BUSCO completion assessment, summary of
annotation statistics and finally functional annotation of
predicted protein-coding genes using the following programmes:

- BUSCO (https://busco.ezlab.org/)

- AGAT  (https://github.com/NBISweden/AGAT)

- DIAMOND (https://github.com/bbuchfink/diamond)

Required inputs for these analyses are the genome assembly fasta
file, the selected database for the diamond blast, in this case
TrEMBL, which can be downloaded from https://www.uniprot.org, the
.gff3 file and protein sequences from the BRAKER3 output.
```

Additionally, we added a paragraph to the README file of the annotation pipeline available on Zenodo (https://doi.org/10.5281/zenodo.11925110) to clearly denote where the scripts are coming from and to attribute authorship:

“For repeat masking we use the fasta_split_1.pl script provided by the Sigenae platform and used in the genofish project (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050324/) to split the genome into smaller parts, and Fan Wei's repeat_to_gff.pl script to convert the trf and dustmasker output to gff3”

In addition, a thorough spell-check to correct any leftover typographical problems would improve the paper's overall professional appearance.

We have proofread the manuscript again and hope to have eradicated all remaining errors.

I noticed a discussion about the 11 MB size difference across haplomes. It would be useful to include a summary of the clipped read statistics for both haplomes. Once these changes have been made, I would be happy to write a recommendation.

In order to address this comment, we here provide the clipped base statistics for both haplomes. After excluding secondary and supplementary alignments and using jvarkit v2023.09.07 (https://github.com/lindenb/jvarkit) we obtained the following clipped base proportions. For the primary haplome, for HiFi, we found 9.05% clipped bases and for HiC 11.21%. For the alternate haplome, the HiFi mapping showed 9.54% clipped bases and the HiC mapping 11.32%. These numbers are comparable between the haplomes. We acknowledge that this statistic is commonly done on Illumina read-based assemblies but not on HiFi/HiC based assemblies. Yet, we have screened the literature and think there is a lack of benchmarking for clipped reads statistics for HiFi long reads and HiC short reads and hence would kindly request not to extend the manuscript in this direction.

Regardless, we are confident that the difference between the haplomes is, to our opinion, reasonable and has no indication of technical bias due to the following reasons that we previously established:

1. We are confident of the genome assembly quality based on contiguity and completeness statistics. We calculated low error rates and the satisfactory quality metrics for our assemblies, in line with the EBP/VGP standards (https://www.earthbiogenome.org/report-on-assembly-standards;
   Hap1 QV=58.9, Hap1 N50=36.4 Mbp, Hap1 BUSCO=96.6%,
   Hap2 QV=58.8, Hap2 N50=36.6 Mbp, Hap2 BUSCO=97.2%).
2. For the assembler we used, hifiasm, insufficient coverage or excessive heterozygosity/errors can challenge the accuracy of assemblies. Our data demonstrates that these concerns are minimal here. We only used high-accuracy short and long reads which is evident from the genomescope plots for the HiFi (Figure 1) and from a FastQC analysis (see Figure R1 below) of the HiC which showed no bias in the short reads either. We also obtained high coverage when we backmapped the PacBio subreads, HiFi and HiC to the two haplomes, as shown in the main manuscript (Line 386-393) and in the qualimap plots below (Figures R2 and R3, also shown in the previous review response). For context, for the theoretically less well-assembled, Hap2 scaffolds: the minimum coverage is ~12X and the mean coverage is ~36X. Coverage was only lower than average in the telomeric regions, which is expected, but this is not where the important non-syntenic regions were located (see Figure 5).

Figure R1: MultiQC summary of HiC Illumina paired reads.

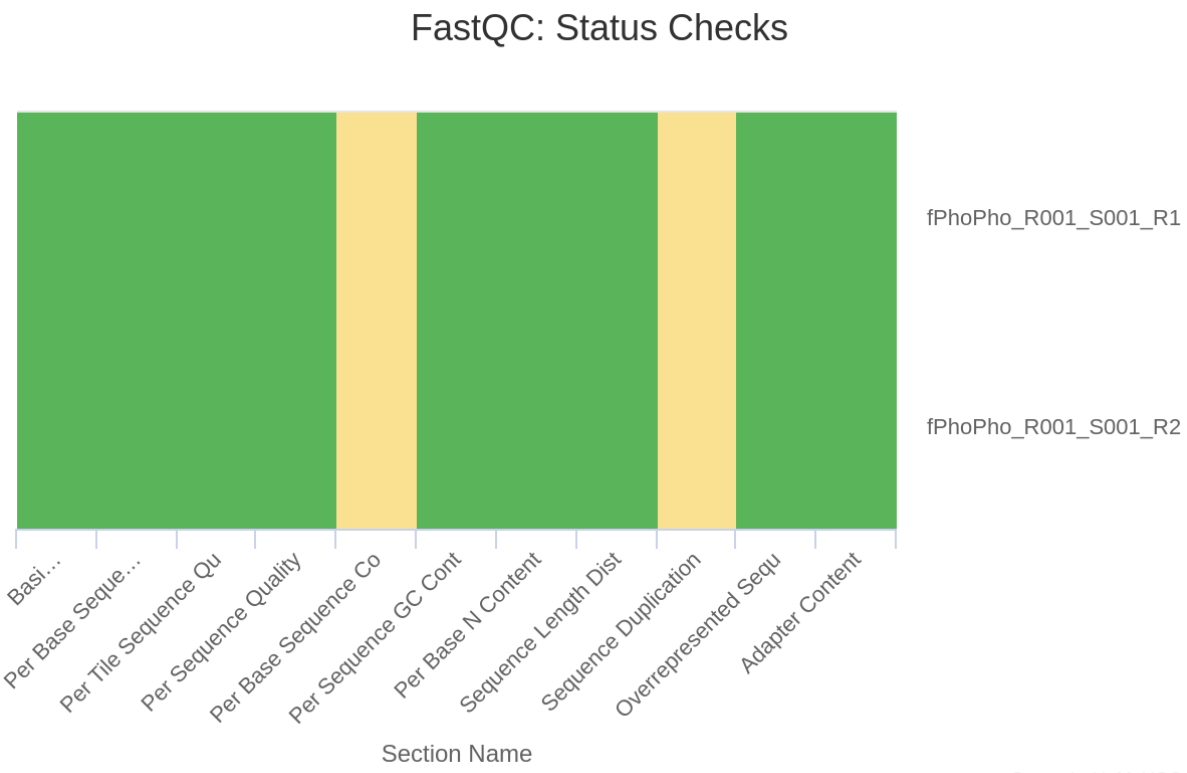


Figure R2: HiFi coverage across alternate haplome.Presented here to demonstrate that even in the traditionally less well-assembled phase coverage is consistent.

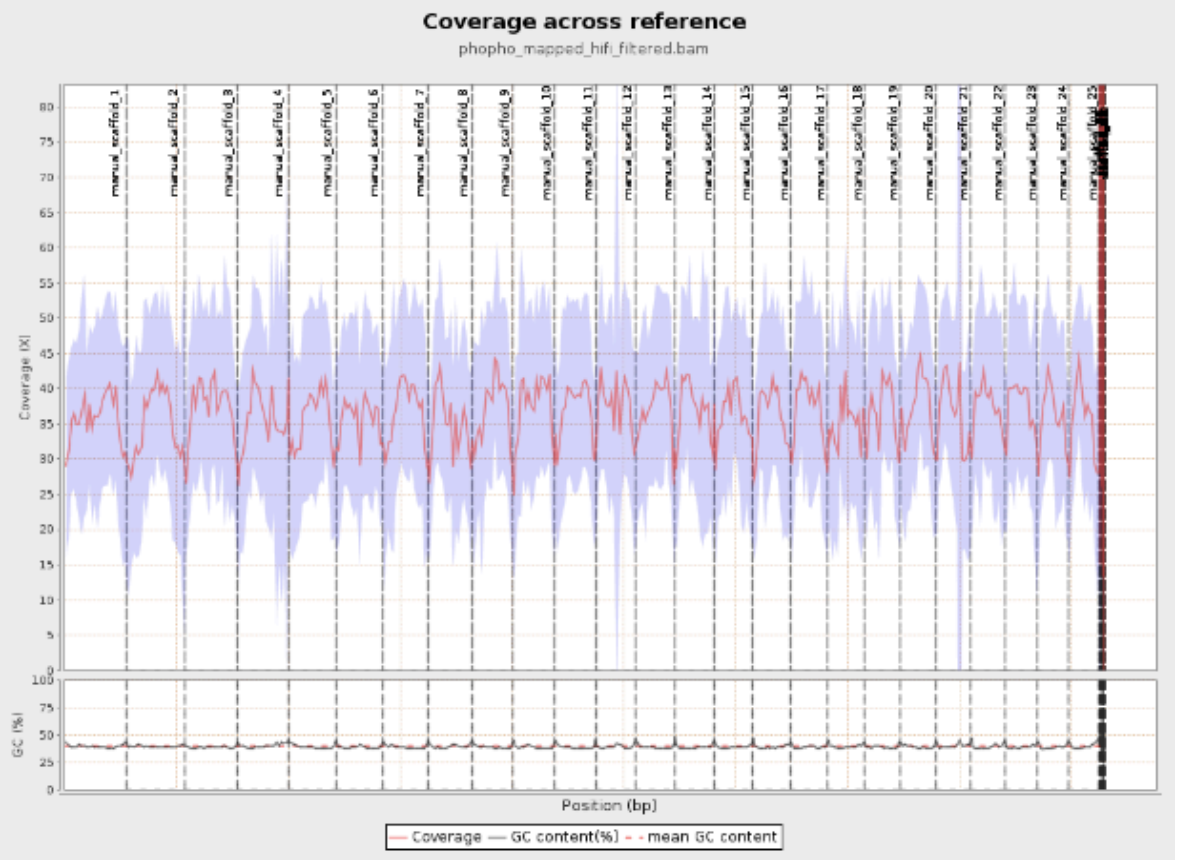
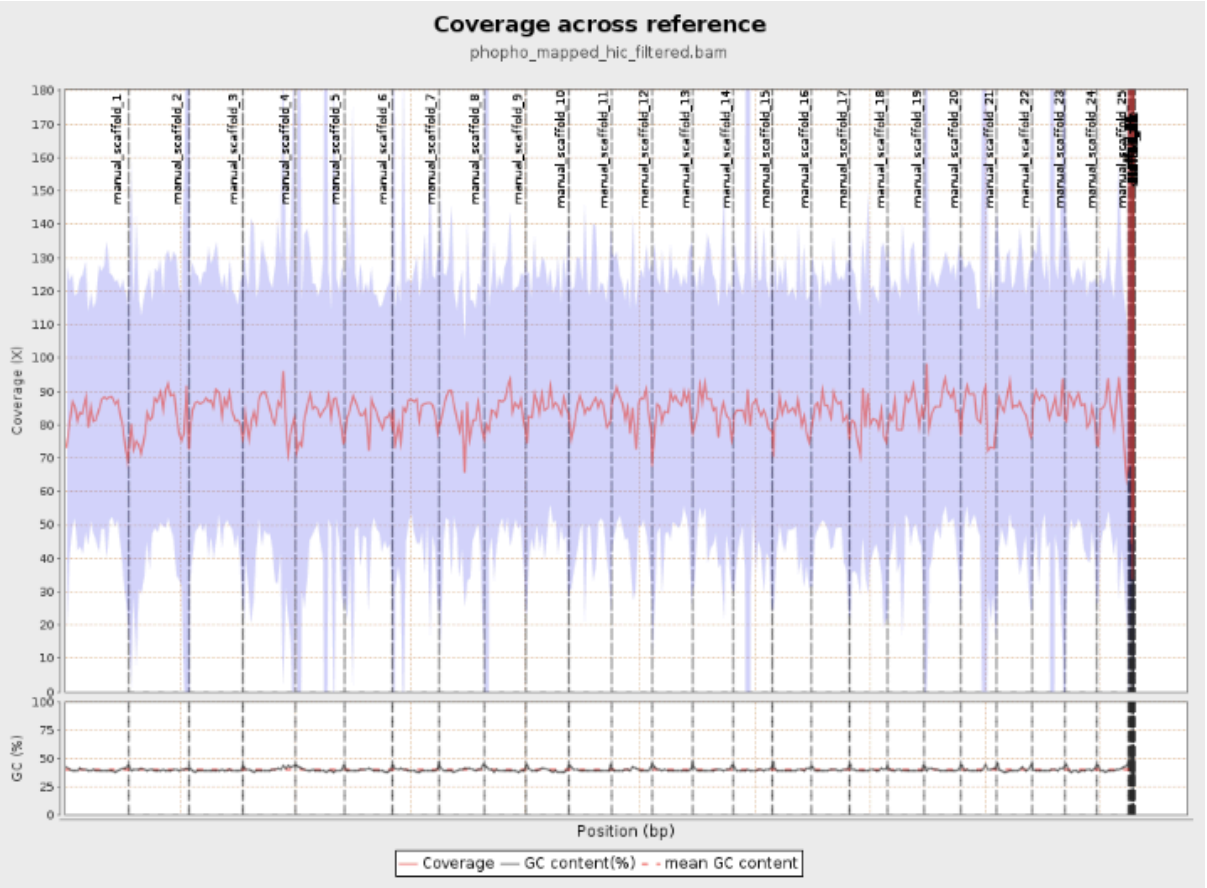


Figure R3: HiC coverage across alternate haplome.



Review by Alice Dennis, 12 May 2024 08:59

Review of “A chromosome-level, haplotype-resolved genome assembly and annotation for the Eurasian minnow (Leuciscidae – Phoxinus phoxinus) provide evidence of haplotype diversity”

The authors of this study have used long-read (PacBio Hifi) sequencing and HiC scaffolding to assemble a phased genome of the Eurasian minnow (Phoxinux phoxinus).

This is a very well written paper. As written in my first review of this paper, I think it is suitable for publication as is, and the modifications in this second version have strengthened this case. For example, I think the PMSC graph is much nicer with the dates you have added. I also appreciate the small addition to the section discussing histones.

We thank Dr Dennis for her kind recommendation.

I noticed two typos:

Line 462: “the” not “he”

changed accordingly, now line 492

Line 475: “Regions” is missing the R, I think.

changed accordingly, now line 507

(P.S. I uploaded this last week and it does not seem to have been completed. Apologies if this came through twice!).

Review by anonymous reviewer 2, 05 Apr 2024 14:23

Dear authors,

the comments in my first review have to a large extent been answered satisfactorily, and I have very few remaining comments. I would only ask to include a few more sections on the assembly process (see below) to increase reproducibility, and that some typos are corrected, otherwise I find the manuscript fit for publication. Congratulations on a strong contribution to the understanding of this interesting species!

We thank the reviewer 2 for their detailed comments and suggestions.

Comments on assembly methods and reproducibility:

Despite careful reading of the manuscript and going through the scripts multiple times, I cannot find that FCS was used to check for contaminants. The only mention I find of this is in the authors' answer to my original comment. Please include a section in the manuscript that FCS was used to screen for contamination.

We follow the reviewers suggestion and, instead of just detailing this section in the reply to reviewers only, added a sentence to “De novo genome assembly and scaffolding” section between lines 222 and 225: “To determine whether any contaminant sequences were present in the genome, we screened for adapter, vector, and foreign sequences using NCBI Foreign Contamination Screen (FCS) v0.2.1 and FCS-GX v0.3.0 (Astashyn et al. 2024). Resulting hits indicative of contaminants were removed from the assembled sequences.”

I would also ask the authors to include a sentence stating that the mitochondrion has been identified and removed from the genome assembly. I cannot see that this has been done, and it needs to be detailed.

A paragraph was added to the end of the “De novo genome assembly and scaffolding” section, lines 241-244: “A mitochondrial assembly was created using MitoHiFi v2.0 [https://github.com/marcelauliano/MitoHiFi (Uliano-Silva et al., 2023) using the DeepConsensus PacBio Reads and *Phoxinus phoxinus* reference mitochondrial genome NC_020358.1 as input. Any partial mitochondrial contigs remaining in the assembly were removed based on mapping synteny to the fully assembled mitogenome”.

The scripts deposited in Zenodo includes scripts developed by the authors and scripts developed by other groups. Only by manually inspecting the scripts can I identify if the scripts are new or a copy of something that is already published elsewhere. Is it possible to make this more clear?

We added a paragraph to the README file of the annotation pipeline (https://doi.org/10.5281/zenodo.11925110) to clearly denote where the scripts are coming from and to attribute authorship:

“For repeat masking we use the fasta_split_1.pl script provided by the Sigenae platform and used in the genofish project (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050324/) to split the genome into smaller parts, and Fan Wei's repeat_to_gff.pl script to convert the trf and dustmasker output to gff3.

Additionally, we use custom perl code provided by the genofish pipeline ((https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050324/) together with the rmOutToGFF3.pl script from the RepeatMasker suite (http://www.girinst.org/) to transform the RepeatMasker gtf to gff3 format.”

Additionally, the sources of the scripts were added as references at the bottom of the README file together with the other programs used in the pipeline with consultation with the

authors of the genofish pipeline which was the inspiration for ours. We hope this is satisfactory and sufficient.

Page 7: "Using our assembled transcripts as input" is still not correct. If, as the authors say in their answer to my previous comment, BRAKER3 assembles the transcripts internally using Stringtie2, then it is the bam-files that are used as input for BRAKER3, not the assembled transcripts.

Rephrased to: “With the RNA mapping files (.bam) as input, BRAKER3 uses StringTie2 to create a draft transcriptome, which then serves as the basis to predict protein-coding genes with GeneMarkS-T (Tang et al., 2015). The genes with the best similarity scores and quality of *ab initio* predictions are then selected with GeneMark-ETP (Brůna et al., 2020).” between lines 273 and 277.

Typos:

The manuscript would benefit from a spell-check/read-through. Below I indicate some typos I have found, but there might be more.

We have checked the submitted documents and found that many typos affecting joined words originated from the track changes version of the manuscript. These typos are not present in the PDF version on biorxiv, which was generated from the clean version. We didn't expect the track changes version of the manuscript to be used for review apart from visualising the made changes. We will be more careful this time to work cleanly on the track changes version as well

Page 0: Change (2n=25) to (n=25 or 2n=50)

changed to 2n=50 in line 21

Page 2: Change "Eurasion" to "Eurasian"

changed as recommended in line 69

Page 3: Phoxinus community s (typo/unclear)

rephrased to: “Firstly, it should provide the *Phoxinus* research community with the necessary basis for phylogenetic unravelling of this complex genus” between lines 108 and 109”

Page 4: Change "fromflash-frozen" to "from flash-frozen"

typo not present

Page 5: Change "ran in genome mode" to "run in genome mode"

changed as recommended in line 205

Page 6: Change "let to misassemblies" to "led to misassemblies"

changed as recommended in line 233

Page 7: Change "wastrained" to "was trained"

typo not present

Page 8: Change "aboveafter" to "above after"

typo not present

Page 8: Change "2017).This estimate is" to "2017. This estimate is" (a space needs to be added after the parenthesis)

typo not present

Page 9: Change "of805.8 Mbp" to "of 805.8 Mbp" (a space needs to be added)

typo not present

Page 9: Change "supportedby" to "supported by" (a space needs to be added)

typo not present

Page 13: This sentence does not feel complete, please correct: "confidently mapped and the SNPs, he k-mer-based approach however additionally incorporates structural variants and is"

Rephrased: “In the present case, the difference between the two approaches used is likely the cause: the genome-wide approach generates heterozygosity estimates from direct observation of the confidently mapped loci as well as the thereof derived SNPs; the k-mer-based approach, however, additionally incorporates structural variants and is more sensitive to low coverage and error-prone regions and hence results in higher heterozygosity estimates.” between lines 490 and 494

Page 14: Change "egions of reduced" to "Regions of reduced"

changed accordingly in line 507

Page 14: Change "withcentromeres" to "with centromeres"

typo not present

Page 16: Change "(Table 4).The largest" to "(Table 4). The largest"

typo not present