Peer Community In Genomics

Exploring evolutionary adaptations through *Phoxinus phoxinus* genomics

Jitendra Narayan *based on peer reviews by* ***Alice Dennis*** *and 2 anonymous reviewers*

Temitope O. Oriowo, Ioannis Chrysostomakis, Sebastian Martin, Sandra Kukowka, Thomas Brown, Sylke Winkler, Eugene W. Myers, Astrid Boehne, Madlen Stange (2024) A chromosome-level, haplotype-resolved genome assembly and annotation for the Eurasian minnow (Leuciscidae: *Phoxinus phoxinus*) provide evidence of haplotype diversity. bioRxiv, ver. 5, peer-reviewed and recommended by Peer Community in Genomics.
https://doi.org/10.1101/2023.11.30.569369

Submitted: 07 December 2023, Recommended: 03 September 2024

Cite this recommendation as:
Narayan, J. (2024) Exploring evolutionary adaptations through *Phoxinus phoxinus* genomics. *Peer Community in Genomics*, 100333. 10.24072/pci.genomics.100333

Published: 03 September 2024

Oriowo et al. (2024) offer a thorough and meticulously conducted study that makes a substantial contribution to our understanding of the Eurasian minnow (*Phoxinus phoxinus*), particularly in terms of its genetic diversity, structural variations, and evolutionary adaptations. The authors have achieved an impressive feat by generating an annotated haplotype-phased, chromosome-level genome assembly (2n = 50). This was accomplished through the integration of high-fidelity long reads with chromosome conformation capture data (Hi-C), resulting in a highly complete and accurate genome assembly. The assembly is characterized by a haploid size of 940 Megabase pairs (Mbp) for haplome one and 929 Mbp for haplome two, with scaffold N50 values of 36.4 Mb and 36.6 Mb, respectively. These metrics, alongside BUSCO scores of 96.9% and 97.2%, highlight the high quality of the genome, making it a robust foundation for further genetic exploration and analyses.

The study's findings are both novel and significant, providing deep insights into the genetic architecture of *P. phoxinus*. The authors report heterozygosity rate of 1.43% and a high repeat content of approximately 54%, primarily consisting of DNA transposons. These transposons play a crucial role in genome rearrangements and variations, contributing to the species' adaptability and evolution (Bourque et al. 2018). The research also identifies substantial structural variations within the genome, including insertions, deletions, inversions, and translocations (Oriowo et al. 2024). Beyond these findings, the genome annotation is exceptionally comprehensive, containing 30,980 mRNAs and 23,497 protein-coding genes. The study's gene family evolution analysis, which compares the *P. phoxinus* proteome to that of ten other teleost species, reveals immune system gene families that favor histone-based disease prevention mechanisms over NLR-based immune responses.

This provides new insight into the evolutionary strategies that have emerged in *P. phoxinus*, enabling its survival in its environment. Moreover, the demographic analysis conducted in the study reveals historical fluctuations in the effective population size of *P. phoxinus*, likely correlated with past climatic changes, offering insights into the species' evolutionary history.

This annotated and phased reference genome not only serves as a crucial resource for resolving taxonomic complexities within the genus *Phoxinus* but also highlights the importance of haplotype-phased assemblies in understanding genetic diversity, particularly in species characterized by high heterozygosity. The authors have delivered a study that is methodologically sound, richly detailed, and highly relevant to the field. The study represents a valuable and impactful contribution to the scientific community, offering resources and knowledge that will likely inform future research in the field.

References:

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C (2018) Ten things you should know about transposable elements. Genome Biology, 19, 199. https://doi.org/10.1186/s13059-018-1577-z

Oriowo TO, Chrysostomakis I, Martin S, Kukowka S, Brown T, Winkler S, Myers EW, Böhne A, Stange M (2024) A chromosome-level, haplotype-resolved genome assembly and annotation for the Eurasian minnow (Leuciscidae: *Phoxinus phoxinus*) provide evidence of haplotype diversity. bioRxiv, ver. 6 peer-reviewed and recommended by PCI Genomics https://doi.org/10.1101/2023.11.30.569369

Reviews

Evaluation round #3

Reviewed by Alice Dennis, 26 July 2024

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). Comparison has been made between the haplomes, and in relation to other Teleosts.

Thank you for the modifications in response to the reviewer comments. I am satisfied with the manuscript as it is now!

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.1101/2023.11.30.569369
Version of the preprint: 4

Authors' reply, 03 July 2024

We attached a PDF with our reply to the handling editor and the twi reviewers. We thank you for your time.

Download author's reply
Download tracked changes file

Decision by Jitendra Narayan, posted 19 May 2024, validated 20 May 2024

Revisions needed

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. 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. Furthermore, adding documentation to better describe the scripts used is needed. In addition, a thorough spell-check to correct any leftover typographical problems would improve the paper's overall professional appearance.

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.

Reviewed by Alice Dennis, 12 May 2024

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.

I noticed two typos:
Line 462: "the" not "he"
Line 475: "Regions" is missing the R, I think.

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

Reviewed by anonymous reviewer 2, 05 April 2024

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!

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.

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.

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?

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.

Typos:

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

Page 0: Change (2n=25) to (n=25 or 2n=50)
Page 2: Change ”Eurasion” to ”Eurasian”
Page 3: Phoxinus community s (typo/unclear)
Page 4: Change ”fromflash-frozen” to ”from flash-frozen”
Page 5: Change ”ran in genome mode” to ”run in genome mode”
Page 6: Change ”let to misassemblies” to ”led to misassemblies”
Page 7: Change ”wastrained” to ”was trained”
Page 8: Change ”aboveafter” to ”above after”
Page 8: Change ”2017).This estimate is” to ”2017. This estimate is” (a space needs to be added after the parenthesis)
Page 9: Change ”of805.8 Mbp” to ”of 805.8 Mbp” (a space needs to be added)
Page 9: Change ”supportedby” to ”supported by” (a space needs to be added)
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”
Page 14: Change ”egions of reduced” to ”Regions of reduced”
Page 14: Change ”withcentromeres” to ”with centromeres”
Page 16: Change ”(Table 4).The largest” to ”(Table 4). The largest”

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/2023.11.30.569369
Version of the preprint: 1

Authors’ reply, 03 April 2024

Download author’s reply
Download tracked changes file

Decision by Jitendra Narayan, posted 09 January 2024, validated 09 January 2024

Refinements requested

The two reviewers offered constructive criticisms and provided insightful comments - please revise your manuscript accordingly and provide a point-by-point reply to the reviews outlining the changes you made and elaborating on any additional data or analyses you performed. I look forward to reading the next version of your preprint.

Reviewed by anonymous reviewer 1, 30 December 2023

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*). In addition to assembling the two haplomes and annotating their features, they have performed comparative analyses between the two, revealing substantial variation, from indels to inversions. This genome has relatively high heterozygocity, making this comparison especially interesting. They have used gene enrichment tests to explore enriched functions in the genes occupying regions that vary between the haplomes and further explored species-specific genes using a gene family analysis relative to 10 additional species. Using a PSMC analysis, they have inferred historical population dynamics.

This is a very well written paper. The methods are clear as are the purposes of the assembly and analyses. This genus is in great need of taxonomic sorting, and this resource will help achieve this. The results of the gene enrichment analysis presented here also provide a very nice starting point for further study of the adaptive differences among closely related species in the genus. I think it is suitable for publication as is, and the issues below are raised to improve the manuscript.

Minor comments:

- In the comparison between the two haplotypes, there is variation in their size, interpreted as indels. Did you manually inspect any of these areas after the polishing? For example, have you mapped the raw data back to get an idea of what could be missing data or assembly errors?

-The discussion of the PSMC analysis refers to periods/times that are not labeled on the graph. This may be a matter of preference, but this discussion would be easier to follow if there were a few more labels, including: approx. LGM period, 800kya, and 20,000 on the y-axis.

- Line 306 refers to the "above described proteomes". Could you replace this with the specifics? i.e. I think you mean the section at the end of the protein alignment, maybe?

- You found that heterozygocity of the assembled haplomes was substantially less than the kmer-based estimates. Can you speculate on why this happened? Did you run genomescope with the HiC/Illumina reads for comparison? I am curious if this is a systematic bias, something specific to PacBio data, or something else.

- One of the most striking gene-family expansions in P. phoxinus is in histone genes. These are discussed in light of their role in the immune system. However, Histones have many other functions, including in transcriptional regulation, or perhaps (especially in the case of duplications) in tissue-specific activities. An inclusion of these alternative roles would be nice.

- italics missing in a few places (*ab initio* in lines 227, 228)

Reviewed by anonymous reviewer 2, 22 December 2023

This manuscript constitutes a good summary of a in general well planned and performed study. It is in many ways a classic genome paper, and more presents a resource for future studies rather than providing any deep biological insights on its own. I agree with the authors that this haplotype-resolved assembly can facilitate new insights into this quite heterogeneous species. My comments, and in some cases concerns, are more focused on reproducibility and how I think some of the analyses are not documented to a level that is satisfactory.

The biggest strength of the manuscript is the assembled genome itself. The biggest weakness is the information that is missing from how the assembly and annotation was performed. At the moment it seems very likely to me that the genome is correctly assembled and of high quality, but without my comments below being adressed, I cannot be certain.

The references in general seem satisfactory and correctly applied.

I found it most difficult to comment on the orthology analyses, although I can find no obvious errors.

Here follows some general and specific comments:

Data availability:

I can easily find the RNA-seq datasets, but not the HiFi-datasets. This needs to be addressed. Also, I cannot find the annotation anywhere. The assembled haplogenomes are available in Zenodo, but I cannot find them in GenBank. A general recommendation of data management is that data and results should be made available in specific rather than general repositories if possible, and I would thus strongly recommend that the assemblies and annotations are made available in GenBank rather than Zenodo.

I would also greatly prefer to see the scripts made available in GitHub rather than Zenodo, although I would not consider it mandatory that this is changed.

Line 115: Specimen Collection and Sampled Tissues

What has been done to make sure that the identity of the sampled individuals later can be verified? Have any voucher material been preserved in a natural history museum? I understand that due to the size of the species it is difficult to preserve the specimen in a state that allows for morphological identification, but a voucher consisting of a third individual (as this is a schooling species) could have given some help. A photo of the live specimens before dissection would also be helpful. Two identifiers are given (starting with ZFMK...). Do these identifiers represent material preserved in a biobank? If that is the case, I would prefer to see this spelled out and the name of the biobank made clear. I would also like to see a more detailed description of the sampling locality, preferably with coordinates. The information given on lines 116-117 is not quite satisfactory.

I consider the lack of information about the material used to be extra problematic as this is a species, which the authors clearly note, which is very heterogenous and may be considered a species complex.

Line 192: De novo genome Assembly and Scaffolding

The assembly process needs to be described better. Here are some pieces of information that are missing:

Please make clear that the HiC reads were used together with the PacBio HiFI reads in the assembly process. This can be deduced from the parameters, but especially since the parameters are not correctly given (see below), this needs to be made clear.

What was done to assure assembly quality? Looking at the supplied information it seems as if HiFiasm was run once using default parameters and that this assembly was then picked for scaffolding without any effort to verify its quality. Best practices include running several assembly tools, or at least running HiFiasm with different parameters, and then picking the best assembly based on BUSCO scores, contiguity (not in itself a measure of quality though), kmer-content, and more.

Presence of contaminants/symbionts needs to be verified, and they need to be removed if present. Blobtoolkit can be used to investigate the presence of these sequences and will also supply a list of contigs that are identified as coming from other organisms.

Mitochondrional sequences need to be identified and removed if present in the assembly. Best is if the assembled and annotated mitochondrion is then submitted under a separate accession number to GenBank, although this can be omitted if considered outside of the scope of the study. Most important is that the mitochondrial sequences are not submitted as part of the nuclear genome assembly.

Line 194: ”...parameters –hic and -l2...”. This is not correct. Looking at the script genome_assembly.sh, the syntax is different.

Line 194: I would like the authors to detail how purge-dups was run, especially how cut-off values, were chosen. Purge-dups can significantly change the assembly, and how it was used needs to be detailed.

Line 208: Change ”ran” to ”run”.

Line 223: How were the output-files converted to GFF3? GFF3 is a complex and heterogenous format and would be interesting to see which standard was followed.

Line 226: The term ”protein annotation” is used here and in several other places in the manuscript. I would argue that this is not a suitable term and would change to ”Annotation of protein coding genes”. It is after all genes that are identified in structural annotation, not proteins.

Line 234: ”Using our assembled transcripts...”. I cannot find anywhere how the transcripts were assembled or any stats about them. Not in the manuscript, not in the linked scripts. This needs to be included.

Line 244: ”Structural annotation...”: Is this a typo and should state ”Functional annotation”? That would fit better with the rest of the sentence and the section
in general.

Line 264: Change ”blasted” to ”mapped”.

Line 265: ”..., to identify homologous sequences”: Here, and in other sections, there is a confusion about homology and what protein similarity can be used for. Diamond uses similarity and can only be used to identify the most similar sequences, not to determine homology. Homology implies shared ancestry, either in the way of paralogy (result of a duplication event) or in orthology (result of a speciation event). On line 265 the problem can be avoided by simply changing ”homologous” to ”similar”, but the authors need to be wary of the meaning of homology and what Diamond/BLAST can be used for in the rest of the manuscript as well.

Line 278: Remove ”To estimate heterozygosity...” and start sentence with ”Site allele...”. As the sentence is currently written I was led to believe that the authors were talking about a new process to estimate heteroszygosity and not a follow-up of the previous section.

Line 287: ”Demographic History of P. phoxinus”: I have little experience in the process described and cannot with confidence review the validity of the methods used here.

Line 324: Change ”length” to ”size”

Line 328: ”We chose the 19-mer length due to a lower error rate...”. I do not understand this sentence, please elaborate.

Line 397: ”Protein annotation”. See comment for line 226.

Line 401: ”...covering 49.9% of the genome...”. How is this calculated? Including intronic sequence? I find this statistic rather uninteresting and it could easily be removed, but if included needs to described better.

Line 410: ”Structural annotation...”. Should this also be functional annotation? See comment for line 244.

Line 415: Table2. I find this table mixes terms and is confusing to the reader. Swissprot, TrEMBL and PDBAA are protein databases and the scores supplied simply implies similarity. Egg-Nog uses phylogenetic information and is a much stronger indication of orthology. Gene overlap is a summary of the other four results and looks strange in a column called ”Database”. The results need to be presented in a better way where different types of results are not mixed.

Line 433: ”It is possible that the regions of high heterozygosity are linked to telomeric regions...”. Perhaps not necessary, but there are tools that can be used to identify telomeric regions. ”It is possible...” is a rather weak statement.

Line 440: ”Genomes with high heterozygosity can pose assembly challenges...”. A high heterozygosity is most likely a positive factor when assembling haplotypes as is done in this study. If the haplogenomes are very different, the assembler can more easily pick them apart. It causes most problems when a consensus sequence is assembled.

Line 444-448: ”Previous studies...”. This section feels out of place here and should be moved.

Line 498: Change to ”We investigated what type of genes were enriched in regions of copy number...”.

Line 517: Change ”haplomes,” to ”haplomes.”.

Line 623: Change ”...contiguous and complete Eurasian minnow...” to ”contiguous and complete genome of the Eurasian minnow...”.

Line 639: Change ”lead” to ”led”.

Line 641: Who is SM? Have the letters for Madlen Stange been switched around?