Review by anonymous reviewer 1, 30 Dec 2023 10:43

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.

Thank you very much for your kind evaluation and we appreciate your comments.

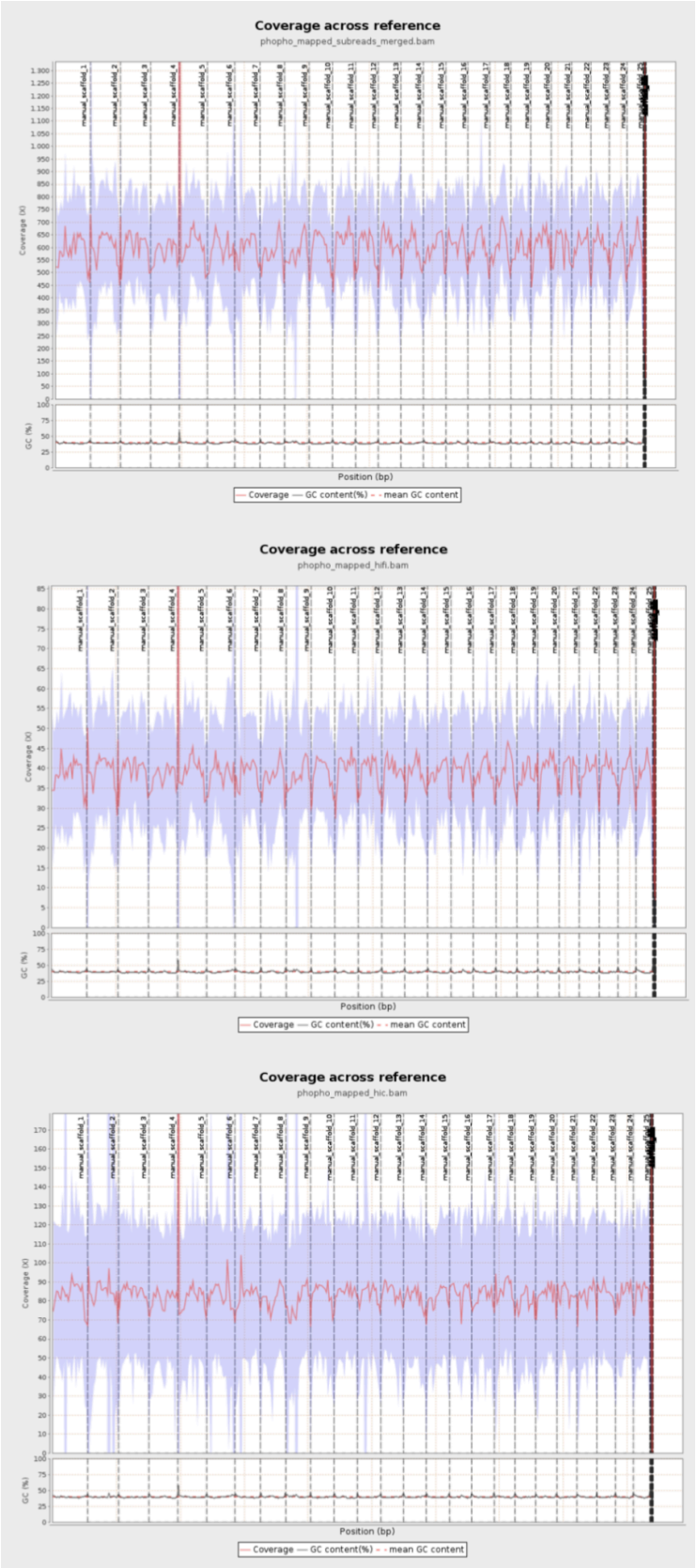
Minor comments:

- 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?

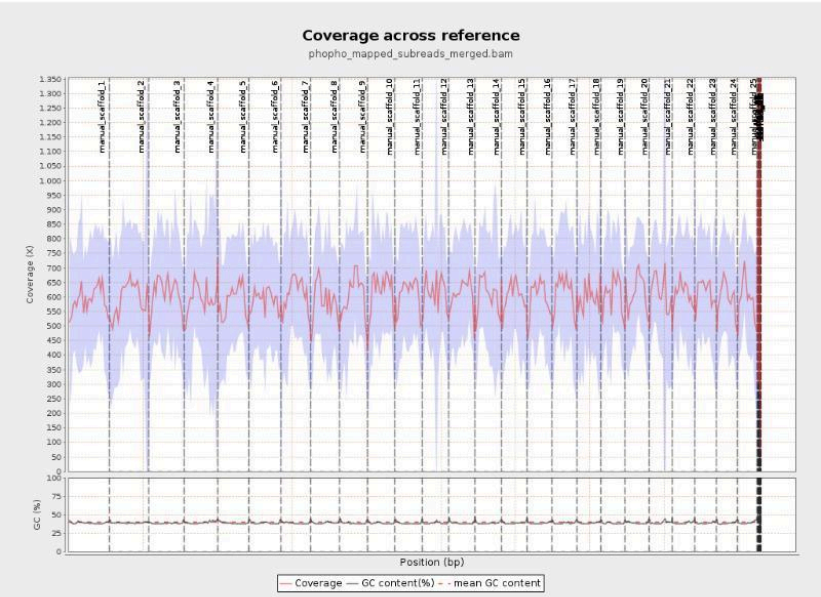
  You’re right that aside from BUSCO and other statistics, like N50, presented in this paper, it is also common to map the input reads back to the output assembly. Based on your suggestion, we performed both analyses in order to verify the quality of the two haplomes.
  So, we mapped the PACBIO subreads, HiFi and Illumina HiC reads back to the assembly and we present the mapping coverage below and some more statistics in Supplementary S1 as well as in the methods under paragraph “De novo genome Assembly and Scaffolding".
  From table S1 we can see that the PACBIO raw subreads, HiFi and illumina HiC for haplome 1 and 2 showed consistent and comparable coverage (600±378X, 39.03±28X 83.6268±132X and 39±25X, 84±82X, 605±353X), mapping rate (99.63, 99.26, 100 and 99.61, 99.18, 100), mapping error (0.0725, 0.0133, 0.1518 and 0.0725, 0.0135, 0.1519) and mapping quality (50, 52, 42 and 50, 52, 43) for the two haplomes. This gives us no hint at a read-based bias in the data.

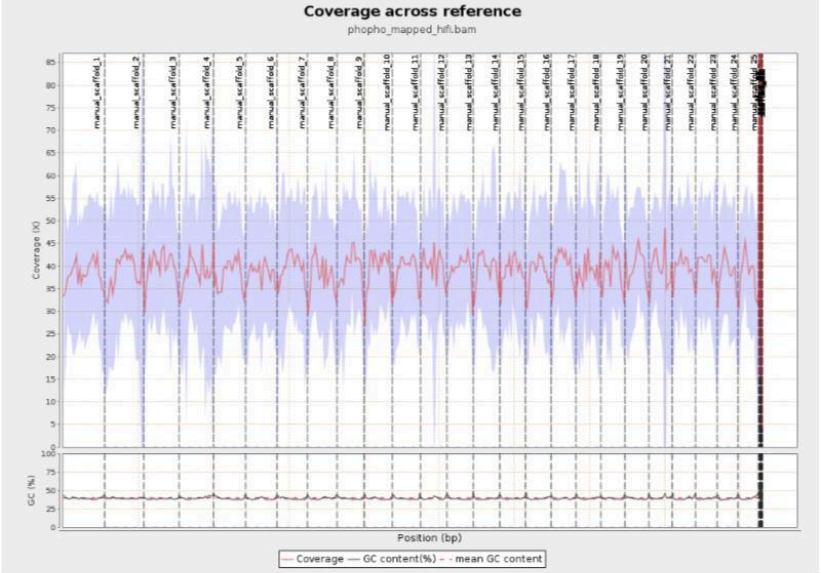
Indeed this new and old information do not give us a good reason to consider the small discrepancy in the size of the two assemblies to be due to large-scale assembly errors.
We observe a reduction in mapping quality within regions where we have detected structural variants between our two haplomes, such as in chromosome 4, the extremities of most chromosomes, and other regions displaying inversions.
This aligns with our expectations, that in structurally variable regions, like ones with inversions, approximately half of the reads would originate from a chromatid with an inversion, while the remainder would come from one without, which should account for the decrease in mapping quality. For a more general view of the mapping coverage and GC% see Figure S1 and the plots below.

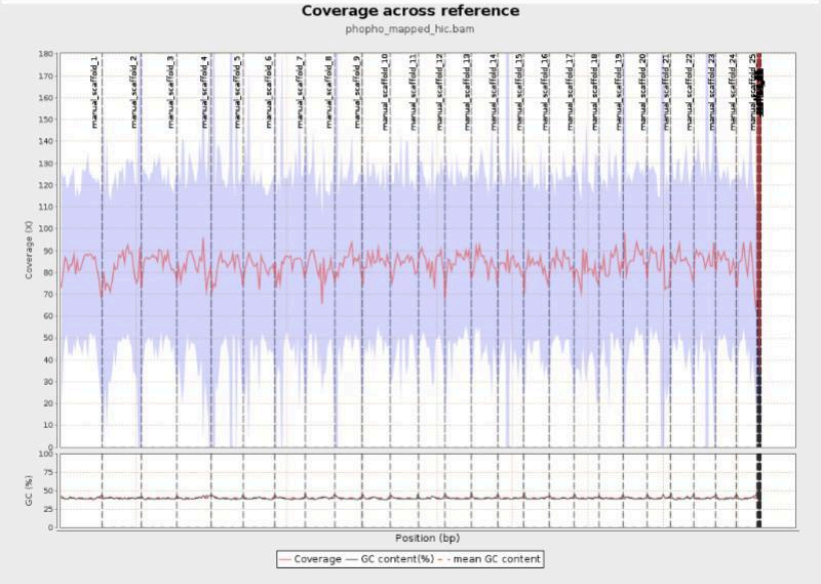
Mapping of raw reads to haplome 1:



Mapping of input reads to haplome 2:







- 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.

  To increase readability of the text and figure, we added labels for dates that we specifically mention in the main text.

- 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?
  Thank you for pointing this out. We now explicitly state that we refer to the preceding section

- 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

  The way the window-based heterozygosity was estimated involved mapping the HiFi reads to the assembly and using ANGSD to call genotypes (with the GATK approach) and then calculate the heterozygosity from the SFS. ANGSD only calls biallelic SNPs and the called sites were filtered for depth and mapping quality. The k-mer based approach, on the other hand, uses the raw reads to count all 19-mers in the dataset. Not only does this use a lot more data than ANGSD, because of a lack of filtering, but also takes structural variants into account, which we show might be abundant between the two haplomes. We believe that all these factors together can account for the discrepancy. Some internal testing we have done (little-to-no filtering ANGSD-based heterozygosity estimates) find much higher heterozygosity than presented in the main text of the paper. This behaviour was also reported in the Genescope paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5870704/#:~:text=The%20results%20are%20generally%20concordant,the%20lowest%20quality%20draft%20genomes.

- 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.
  Other roles of histone genes outside immunity have now been included: “Histones also play important roles in gene expression, DNA replication, and DNA damage repair (Best et al., 2018; Seal et al., 2022), in addition to their involvement in immune activity. It is important to note that histones are not restricted to immune activity

alone. In fish, they have also been linked to spermatogenesis and can be an indicator of sperm quality (Herráez et al., 2017)."

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

"Review by anonymous reviewer 2, 22 Dec 2023 11:12

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.
We were under the impression that the editor had shared all the raw data (RNA-seq and haplomes and annotations) with the reviewers. At the time of submission, our data was still being processed by NCBI, but both haplomes and annotations are now public. GCA_037504875.1 (haplome 1), GCA_037504845.1 (haplome 2) and the genome project can be found under the umbrella of the EuroFish project PRJNA768423, and the HiFi , RNA-seq data and the annotations are now directly accessible under the umbrella of PRJNA1040855.

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.
When preparing our manuscript for PCI Genomics we followed the recommended guidelines, stating: under "2.3 Repeatability of science and open science" that "**Raw data**, made available directly in the text or through an open repository, such as Zenodo, Dryad or some other institutional repository (see Directory of Open Access Repositories) with a DOI.

Data must be reusable, and the metadata and accompanying text must, therefore, carefully describe the data." and have hence decided to publish code on Zenodo.

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.
Thank you for pointing out that we weren't clear enough. Actually, fin clips from the sampled specimens were deposited in the Museum's Biobank (we are a natural history museum) as voucher material (formerly known as Zoological Research Museum Alexander Koenig, and hence "ZFMK"). We added the information of the Biobank name to the manuscript and clarified that the accession IDs are meant to be voucher material.
After dissection not much was left from the individual. We admit that a publishable picture of the individuals before dissection would have been nice (we include the picture we took below), but species differentiation in *Phoxinus* from outer morphology outside the breeding season (no spawning coloration) is currently not possible. We also added info about a proxy-voucher specimen from the neotype locality that includes a fin-clip and a body to the main text.

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.
We have now further clarified the origin of the specimens in the text. The specimens originate from the museum's (LIB, Museum Koenig, formerly known as Zoological Research Museum Alexander Koenig) live display, which shows the local fish fauna to our visitors. For subsequent publications within the same project (Leibniz J96/2020) we collected fresh specimens of *Phoxinus* spp. (bodies and tissues) from more than 1500 specimens. The added proxy voucher derives from these sampling efforts. Tissues have been deposited in the LIB Biobank. Bodies that remain intact (we are also performing trophic analyses that partially destroy the body) will be deposited in the ZFMK Ichthyological Collection, hence lots more proxy material will soon be available.

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.

Thank you for the suggestion. We have added a table (now Table 1) to the methods with BUSCO and Contiguity scores for the parameters sweeps run when assembling the contigs with hifiasm +/- purge-dups. The aim, as always is to maximise BUSCO completeness and contiguity, while minimising BUSCO duplication and contig number. We found that the combination applied of -l2 and purge-dups to give the best assemblies. Regarding assembly tools, we were keen to assemble the two haplotypes separately using the HiC reads, and at the time HiFiasm was the only robustly tested tool capable of doing this. To my knowledge this capability also only exists in Verkko and can potentially be found by assembling all unitigs with HiCanu and grouping contigs together into quasi haplotype-phased scaffolds using the HiC density plots.

We hope the reviewer agrees that the manuscript has been strengthened by the addition of Table 1 to the methods section.

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.

Decontaminants were removed using the NCBI tool fcs (adapters and vectors) and fcs-gx (foreign contaminants). As this is the current gold standard, we would prefer to maintain this over older methods, such as the blast/diamond search inside blobtoolkit.

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.

Indeed, the mitochondrial assembly was assembled using MitoHifi, but not analysed as part of this manuscript. The sequence has been submitted to INSDC alongside the nuclear assembly, but is still being processed.

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

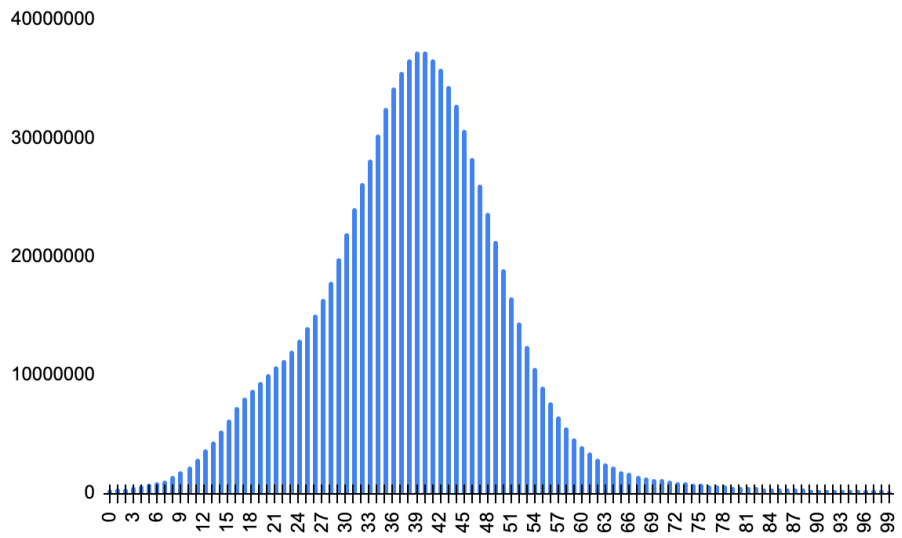
Many thanks for pointing out this error. The parameters have been corrected in the methods section and the script on Zenodo was updated , the new version doi is (https://zenodo.org/doi/10.5281/zenodo.10210240).

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.

As detailed in the assembly scripts, all steps with purge-dups were run as default, or as otherwise recommended (e.g. including -e parameter in get_seqs https://github.com/dfguan/purge_dups?tab=readme-ov-file#step-3-get-purged-primary-and-haplotig-sequences-from-draft-assembly). The cutoff values selected by default were found to be correct, most importantly the haplotypic-threshold parameter (number 4) was selected as midway between the heterozygous (20) and homozygous (40) peak. Default inferred cutoff values for the two haplotpyes were:

5 15 25 30 50 90
5 15 23 30 48 90

, respectively. Below is the coverage profile calculated for hap1, generated via the pbcstat command from purge-dups for comparison.



Line 208: Change "ran" to "run".
Done.

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.

The Dustmasker and TRF outputs were converted using Fan Wei's repeat_to_gff.pl v 1.0 perl script and the RepeatMasker output was converted with rmOutToGFF3.pl from the RepeatMasker container. More details can be found in the annotation pipeline in Zenodo.

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.
Changed accordingly.

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.
The transcript assembly was performed internally by BRAKER3 using Stringtie2 before being used for prediction with GeneMarkS-T, We removed the paragraph break before this sentence and added "[...] from the seven organs" to this sentence to make clear that the previous paragraph and this sentence belong together.

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.
Changed accordingly.

Line 264: Change "blasted" to "mapped".
Changed as suggested.

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.
Thank you very much for pointing this out. We changed "homologous" to "similar" in this instance and re-read and adjusted the text keeping the difference in mind.

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.
Done

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"
Done

Line 328: "We chose the 19-mer length due to a lower error rate...". I do not understand this sentence, please elaborate.
We needed to choose a k-mer length that we could use as a basis for comparison not only with heterozygosity estimates in other species, but also with our heterozygosity estimate from mapping. The error rate reported by GenomeScope generally indicates model fit; the lower it is, the better the model fits the chosen k-mer length.

Line 397: "Protein annotation". See comment for line 226.
Done.

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.
It has been removed as suggested and left as just counts instead, it is an output statistic generated by AGAT.

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

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.
This has been adjusted and made clearer. Former Table 2 is also now Table 3.

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.
The tone of this paragraph has been altered and more references were provided.

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.
The point of this statement is that until haplotype-phasing was possible the genomes and their diversity was collapsed into a single reference sequence, which ultimately led to not-captured diversity lost to all subsequent analyses, like in population genomics and species diversity estimations.

Line 444-448: "Previous studies...". This section feels out of place here and should be moved.
It has been moved to a more suitable section

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

Done.

Line 517: Change "haplomes," to "haplomes.".
Done.

Line 623: Change "...contiguous and complete Eurasian minnow..." to "contiguous and complete genome of the Eurasian minnow...".
Changed as suggested.

Line 639: Change "lead" to "led".
Done.

Line 641: Who is SM? Have the letters for Madlen Stange been switched around?"
SM stands for Sebastian Martin. To avoid confusion MS was changed to MSt.