Round #1

Reviews

Reviewed by Gonzalo Riadi, 07 Aug 2023 23:16
PCI Review

The genome sequence of the Montseny horsehair worm, Gordionus montsenyensis sp. nov., a key resource to investigate Ecdysozoa evolution

General questions

Did you read the “guide for reviewers”? (see the Help menu of the thematic PCI or the dedicated blog post)
R: yes.

Is the manuscript well written?
R: yes.

Is the description of the rationale and methods clear and comprehensive?
R: They are enough. Just a couple of points: the methodology of the tree building, and genome size estimation are lacking.

Are there flaws in the design of the research?
R: No.

Are there flaws in the analysis?
R: I missed the genome size estimation, but I wouldn’t qualify it as a major flaw. I suggest to add more information in the methodology and discussion of the manuscript, in the last section of comments of this review.
>Authors: The genome estimation size is included in the first version of the manuscript with GenomeScope, see Fig. 4. Regarding adding more information to the methodology and discussion, we would like to remind the reviewer that this is a genome note, not a regular manuscript, and it is intended to simply report the metrics of the genome assembly. We would like to refer the reviewer to the managing board of PCI Genomics for information on what type of information this type of manuscript is expected to contain.

Are there flaws in the interpretation of results?
R: The Hi-C experiment suggests 5 chromosomes, but the assembly describes 6 main pseudochromosomes. This is not a flaw in the interpretation of results, but calls for a comment in the discussion.
>Authors: We performed additional manual curation to the assembly. We have updated the assembly to a version 2 (Project Accession number: PRJEB63266) and the new metrics of the genome assembly and annotation are being reported in the new version of the manuscript. After

curation, we indeed found that the assembly was formed by 5 pseudochromosomes with robustness (see Fig. 6), despite the first version of the assembly not supporting this with confidence.

Do you have concerns about ethics or scientific misconduct?
R: No.

Did you detect a spin on the results, discussion or abstract? (a spin is a way of twisting the reporting of results such that the true nature and range of the findings are not faithfully represented, https://doi.org/10.1073/pnas.1710755115)
R: No. The genome report, does not precise that the genome reported is partial. That would be desirable (maybe in the discussion as suggested in my comments, last section).
>Authors: We do not consider that the genome is partial. It is indeed chromosome-level. We report that 97% of the assembly is organized in 5 pseudochromosomes, which falls within the requirements of EBP to be considered as a reference, chromosome-level genome. The remaining 3% is assembled in smaller contigs, but it is still included in the fasta file.

Is something critical missing?
R: Apart from the methodology of the tree and the genome size estimation, no. I suggest to add in the methodology and discussion about sequence alignment, in the last section of this review.
>Authors: We have added this information to the Materials and Methods section.

Evaluation of the various components of the article

Title/abstract/introduction
Does the title clearly reflect the content of the article?
R: I think the second phrase in the title, after the comma: “a key resource to investigate Ecdysozoa evolution” while being true, is not too well supported by the text, or this study.
>Authors: We did not conduct any analysis in a comparative context to investigate Ecdysozoa evolution, but this genome will be without doubt a valuable resource for comparative studies of Ecdysozoa. By the day we submitted this manuscript, no genome was published. Now, although there are two more genomes available (Cunha et al., 2023) we still think that Nematomorpha is an understudied phylum and since the interrelationships of Ecdysozoa are unresolved, more genomic information publicly available should definitely be a key resource to investigate Ecdysozoa evolution.

Does the abstract present the supported findings of the study concerned and no other?
R: yes.

Does the introduction clearly explain the motivation for the study?
R: pretty much, yes.

Is the research question/hypothesis/prediction clearly presented?
R: The sequencing of the worm was justified.

Does the introduction build on relevant recent and past research performed in the field?
R: It builds on the necessary information about the biology of the worm in order to understand the need of its genome sequence. However, I would have liked to know a bit about what genomic information is currently available in the databases, if not from the particular species, from its relatives. (and a comparison of the numbers, in the discussion)
>Authors: There was no genomic information available by the date of the submission. Shortly after the submission of this manuscript, two genomes of the phylum of Nematomorpha were published (Cunha et al., 2023), one of them a draft, partial genome, and another one chromosome-level, despite being less contiguous and of a lesser quality than ours. A comparison of the 3 genomes is available in Table 1 of the new version of the manuscript.

Materials and Methods
Are the methods and analysis described in sufficient detail to allow replication by other researchers?
R: Except for the tree, the aligment and the database used for Repeatmasker, yes.
>Authors: The alignment and tree inference methodology are included in the Materials and Methods. The database used for the RepeatMasker is the de novo repeat library constructed with RepeatModeler (already included in the manuscript) and the Dfam curated database v3.6. This information is now added in the corresponding part of the manuscript in the new version.

Is the experimental plan consistent with the questions?
R: Generally, yes.

Are the statistical analyses appropriate?
R: Again, I missed the genome size estimation. The rest of the statistical analyses are standard and well performed by known available programs. I missed technical information on the sequence comparisons and the tree building, too.
>Authors: We have estimated the genome size using kmer analysis with GenomeScope2. No biological method was used for estimation. We do not compare any mitochondrial genomic sequence, since this is a genome note reporting the genome sequence of a new species. The tree building methodology is described in the Materials and Methods.

Have you evaluated the statistical scripts and program codes?
R: No new scripts were developed in this work, so there was no need to evaluate scripts this time.

Results
Have you checked the raw data and their associated description?
R: The data was deposited in the European Nucleotide Archive, ENA, although it is not mentioned explicitly in the text. I also would have liked to see the Genome reports in Supplementary material, which contain more technical information about the sequencing, and genome or transcriptome sample quality checks. No information in the text, on where the genome assembly and annotation files are/will be available for download.

>Authors: This information is available in Table 2. The accession numbers of the version 2 of the genome assembly, with the genome annotation included, and the mitogenome are updated in the new version.

Have you run the data transformations and statistical analyses and checked that you get the same results?
R: No. However, the results are sound respect to the raw data information deposited online, methodology described, and from the tables and figures provided. Discrepancies (like chromosome number, completeness of the assembly, number of protein coding genes, comparisons with other relatives), however, should be discussed in the appropriate section, but are not.
>Authors: We do not understand what the reviewer sees as discrepancies. We would also like to kindly remind them that this manuscript is a genome note, following the instructions of PCI Genomics for such type of submissions, which are in line with the ones from the Wellcome Sanger Institute for the Darwin Tree of Life. Therefore, a discussion in a comparative context is not the goal of this manuscript. In any case, we have provided a table comparing the metrics of the newly generated assembly with the other two nematomorph genome assemblies that were published while this paper was under review (see Table 2).

To the best of your ability, can you detect any obvious manipulation of data (e.g. removal)?
R: No.

Do the statistical results strongly support the conclusion (p< 10-3 or BF>20)?
R: Although no hypothesis testing was explicitly performed, so no p-value of Bayes Factors were calculated, genome data analysis was done correctly.

In the case of negative results, was a statistical power analysis (or an appropriate Bayesian analysis) performed?
R: No “negative” results as in “hypothesis disproved” in a genome report. In the Discussion section, though, a comment on how complete the genome was sequenced and assembled; and a comment on the final number of chromosomes (why they have 6 pseudochromosomes when Hi-C suggests 5) would be desirable.
>Authors: As explained before, the assembly is not partial, it is chromosome-level. After additional manual curation we report in the new version of the manuscript 5 pseudochromosomes.

Did the authors conduct many experiments but retain only some of the results?
R: No.

Discussion
Do the interpretations of the analysis go too far?
R: No.
Are the conclusions adequately supported by the results?
R: Yes.

Does the discussion take into account relevant recent and past research performed in the field?
R: The discussion is centered in the worm's biology, not its genome sequence, assembly or annotation. It would be interesting to read about the comparison of the manuscript genome assembly results, like genome size and number of protein coding genes, with relative species.
>Authors: Again, our main goal is not to provide a comparative study rather than to report one of the first high-quality genomes in the phylum of Nematomorpha by submitting this Genome Note. However, given that 2 genomes were published after the submission of this genome note, we have added a comparison of the metrics of all genomes on Table 1.

Did the authors test many hypotheses but consider only a few in the discussion?
R: No. This is a genome report, describing the sequencing, assembly and annotation of a genome, not a research article testing hypotheses.

References
Are all the references appropriate?
R: Yes.
Are the necessary references present?
R: Yes.
Do the references seem accurate? R: Yes.

Tables and figures
Are the tables and figures clear and comprehensive?
R: Yes.

Are all the tables/figures useful?
R: No.

Are there too many/too few tables and figures?
R: Yes. Figure 5, Figure 7 and Figure 8 could go in supplementary material.

Do the tables and figures have suitable captions such that they can be understood without having to read the main text?
R: No. Figure 2. Please, enrich either the caption or the methodology with more information. For instance, what features were used for generating the tree? What biological sequences? The list of accessions should be available in the supplementary materials. How were they processed? What are the outgroups?
Figure 4 could be further explained. The diploidy peak, and the repeats peak, for example. Is the x-axis k-mer coverage or genome coverage?
>Authors: The methodology of Figure 2 is included in the Materials and Methods. Figure 4 expanded explanation is added.

Comments, questions and suggestions
Abstract suggestions

Change “the most neglected” for “one of the less studied”. Neglected is an active verb, with an emotional load, whereas “less studied” is passive, and probably corresponding to what actually happens.
>Authors: Changed in the new version.

Introduction suggestions
Add a comma after, “As expected” in the first paragraph.
>Authors: Added in the new version.

Figure 1. The electron micrography shows a male. How come it was not described in the materials examined?
>Authors: It was indeed. The images correspond to the holotype. It has been clarified in the caption of Fig. 1.

Genome Sequence Report
“340x coverage of long reads and 80x coverage of small reads”. This is respect to which genome size? One genome previously reported? Maybe one from a cytogenetic study? Or from a genome size estimation? Or from the final assembled genome size? Genome size estimation (and final coverage) is important, since it can be compared with the initial coverage and the assembled size and to estimate how much sequencing was “wasted”, and how much genome is left to be sequenced. Also, the coverage has to be specified, initial (previous analysis, just after sequencing) or final (after analysis, quality control, trimming and genome size estimation, previous assembly).
>Authors: The coverage was estimated against the size of the final genome assembly generated. This comment has been added in the new version.

FASTQC files should go in the supplementary materials.
>Authors: FastQC files are now accessible in the GitHub repository https://github.com/MetazoaPhylogenomicsLab/ChromosomeLevelGenomes/tree/main/Gordionus_montsenyensis.

“Pair-wise (sequence) similarity is 95.04%” (page 3, just before Discussion). Sequence similarity or sequence identity? Please specify. How was the alignment done (global, local, algorithm, parameters)? Were mitochondrial genomes used as pointed out two paragraphs later, in Discussion section? DNA or protein sequences? This, I believe, was not specified in the Methodology.
>Authors: “The mitochondrial genome was aligned using NCBI Blast with the megablast algorithm against the non-redundant nucleotide database (nt) with default parameters.” This information is now added in the new version of the manuscript.

“warty appearance of spines of G. montsenyensis is unique and justifies the description as a new species.” I am not an expert, so I do not feel knowledgeable about the criteria to consider a particular worm as new species. However, as “Nematomorphs are not rich in characters that can

be used for identification", a study using biological sequences (particularly the ITS region known to be important for species determination in worms), both only sequence or phylogenetic could enlighten this question as supporting information. Discussion is lacking in this respect, connecting the phylogenetic analysis with the idea that this is a new species.
>Authors: One of the co-authors, Andreas Schmidt-Rhaesa, is a world authority in nematomorph taxonomy and systematics, and one of the only experts in this phylum. He has published more than 30 papers on taxonomy and systematics of this animal phylum, and has described 25 nematomorph species (see full record of publication on nematomorphs here: https://species.wikimedia.org/wiki/Andreas_Schmidt-Rhaesa). We respectfully disagree in that only sequence or phylogenetics could enlighten the question of whether this is or not a new species, but leaving this debate apart, these sources of information also support its allocation to a newly described species.

"96% of the assembly sequence assigned to 6 pseudochromosomes". However, Figure 6 suggests only 5 chromosomes. Please, comment.
>Authors: Please see comments above.

Figure 2. Please, enrich either the caption or the methodology with more information. For instance, what features were used for the tree building? What biological sequences? DNA or proteins? The list of accessions should be available in the supplementary materials. How were they processed? What are the outgroups? The final log should go in the supplementary materials.
>Authors: We have expanded the information in the Material and Methods section. The accession numbers are indicated in Fig. 2.

Figure 4 could be further explained. The diploidy peak, and the repeats peak, for example. Is the x-axis k-mer coverage or genome coverage?
>Authors: Added in the new version.

Figure 5, Figure 7 and Figure 8 could go in supplementary material.
>Authors: Following the previously published Genome Notes from Wellcome Sanger Institute, and adopted by PCI Genomics for genome notes, these figures are provided in the main text rather than the Supplementary material.

Figure 7 has y-axis title base outward respect to the axis. Please, rotate the y-title 180 degrees so the title reads from bottom up.
>Authors: Changed in both Figure 7 and 8.

How many reads were sequenced for the transcriptome? This is not reported in the methodology.
>Authors: In total ~127 million reads were generated. Added in the new version under the "RNA extraction, library preparation and sequencing" section.

Do the authors think that, in spite of 96% of genome sequence assembled, lack of an estimated 40% of genes by BUSCO in the final genome could be, at least in part, due to the tissue sampling, or not enough depth in transcriptome sequencing? OR could it be due to an assembly problem? A comment about this could enrich the Discussion section.
>Authors: First, we would like to point out that 97% of the genome is assigned to 5 pseudochromosomes. This does not mean that there is 3% not sequenced and assembled. This 3% is assigned in the rest 391 smaller scaffolds. We reported 396 scaffolds in total in the version 2 of the genome assembly, after a second round of manual curation. For the genome we used DNA, so there is no possibility of tissue-based error or depth in transcriptome sequencing. The percentage of missing BUSCO genes is 28.4%, not 40% as indicated by the reviewer, and in our extensive experience with transcriptomics we don't think it is due to a lack of depth in the sequencing (usually 25-30 millions of reads in enough, and in our case we sequenced 127M). Furthermore, the genomes presented in Cunha et al., 2023 are missing similar percentages of BUSCO genes, corroborating our results. Thus, we have compelling evidence that this is not a methodological artifact but rather due to the strange biology of these animals.

Please, specify and reference the database that was used together with Repeatmasker in Repeat Identification section.
>Authors: Specified in the new version.

Latin expressions like "de novo" or "ab initio" should go in italics, in the text.
>Authors: This actually depends on the manual of style chosen or dictated by each journal. Under the Chicago Manual of Style, common Latin phrases like "de novo" and "ab initio" should not be italicized. Chicago Manual of Style typically recommends the use of italics for foreign words or phrases that are not commonly used in English, but for familiar and commonly used Latin phrases, italics are not necessary. So, you can write "de novo" and "ab initio" in regular, non-italicized type in your text when following the Chicago Manual of Style. We will be happy to adhere to the preferences of PCI Genomics in this matter, in the meantime we prefer to leave them not in italics.

No information from were to download the genome assembly and annotation in the text.
>Authors: Updated in the new version in Table 2.

Reviewed by anonymous reviewer, 19 Aug 2023 20:09
In writing this review, I firstly want to flag potential conflicts of interest. I am myself a member of the large team working on the ERGA pilot, I am also collaborating with some of the authors on projects outside of ERGA, and I am host to one of the authors on their MSCA project. I assume that for genome reports as this one, such situations cannot be avoided, simply due to the extremely high number of reports that will be written as part of the EBP.
In their report on the genome of a newly discovered nematomorph species, Gordionus montsenyensis sp. nov., the authors formally describe the species, and describe the genome. The report is very well written and includes all necessary information to be published. I have two suggestions:

1. The authors state, "None of these sequences were filtered, since due to the parasitic life cycle of nematomorphs, these sequences could be horizontally transferred sequences." They could easily include a screen for potential HGT candidates into their report.
>Authors: The host is unknown. Every software we checked required possible hosts as input. This is not possible, since the sampling of the species is quite challenging and was only found in the streams of the river after leaving the body of the host, with many potential hosts around. We decided to leave it like that but of course mention it in the text so the readers are aware in case they want to leverage our datasets.

2. The authors split their report into a section that uses classical morphology to describe the species, and then a second one to describe the genome. I think the morphological description is very valuable, but would still want to suggest for the authors to incorporate the genome into the species description. That is, used basic features of the genome as additional information to describe the new species.
>Authors: We understand the point of the reviewer, however we have a couple of concerns. First, how can we precisely define synapomorphies at the genomic level? Should they be based on the features of the linear sequence, or the genome architecture, or the presence/absence of certain gene families, to name a few attributes that could be coded and explored? Second, even if the previous point was easy to tackle, this was the first genome of the entire phylum, and in order to provide species-specific synapomorphies one should count with abundant information in order to perform a comparative study and understand what features may be diagnostic. This is something that we (as a scientific community) may be able to tackle in a few years, but definitely not now yet.

Additional minor points:
"In this piece of work," -> please consider to use different wording
>Authors: Changed to -> In this study

"this enigmatic animal phyla" -> phylum
>Authors: Changed.

"As expected given their parasitic lifestyle,"" -> this needs a citation if it is really expected
>Authors: A citation has been added to the new version.

Reviewed by anonymous reviewer, 12 Oct 2023 02:48
This study introduces a newly discovered species of Nematomorpha, Gordionus montsenyensis Schmidt-Rhaesa & Fernández sp. nov., along with a chromosome-level genome assembly. The assembly comprises 398 scaffolds, totaling 288 Mb, and boasts an impressive N50 of 52.6 Mb, with 96% of the genome organized into 6 pseudochromosomes. Additionally, the authors have successfully assembled a 15-kilobase circular mitochondrial genome and identified 10,819 protein-coding genes. This valuable genomic resource significantly contributes to the exploration of Ecdysozoa evolution and enhances our understanding of the genetic foundations of parasitic lifestyles.

Point of concern:
1. The author focuses on N50, it is a metric that provides insight into the 'average' sizes of long scaffolds within an assembly. This measurement is particularly effective when the genome assembly is "predominantly error-free".
I suggest the author assess assembly quality by examining the k-mer distribution in the assembly and compare it to the expected k-mer distribution derived from the sequencing reads. By utilizing this k-mer correctness metric in conjunction with N50, we can gain insight into whether a high N50 value is the result of mostly accurate junctions or if it is influenced by numerous incorrect junctions.
>Authors: If the reviewer is referring to the error rate that Merqury outputs, it is 0.000183556. In any case, we consider the combination of all the metrics we provided is adequate to justify a high-quality chromosome-level genome assembly. We do not focus simply on the N50 value.

2.The English in some sections of the manuscript could benefit from minor improvements. For instance, the statement "The nematomorph fauna from Spain is only fragmentarily known" could be rephrased as "Our understanding of the nematomorph fauna in Spain is limited and fragmented" to enhance clarity.
>Authors: Changed in the new version.

3. Maximum likelihood phylogenetic tree “Figure 2” illustrating the positioning of Gordionus montsenyensis sp., with support values of 83/85/85 for standard nonparametric bootstrap, SH-aLRT and UFBoot.
Generally, we tend to place more confidence in a clade when its SH-aLRT exceeds 80% and UFBoot exceeds 95%. I would recommend providing the Average Nucleotide Identity (ANI) values to support this relationship.
>Authors: The tree was inferred only with the sequence of the cytochrome C oxidase subunit I. We believe this measurement is adequate for full genomes and we won’t be able to calculate it in our case.

4. Page number 14 “The size of each circle is proportional to the scaffold length and each color represents taxonomic assignment by a blast search against the nt database”
Did you take the second-best match into account during your assignment? It's recommended to employ the Kraken k-mer-based approach and conduct a cross-comparison of the results for enhanced accuracy.
>Authors: Indeed we took only the first best match. Although, we are quite confident in the accuracy since we have scaffolded the genome using Hi-C data and contaminants don’t have contacts with the main chromosomes.

5. On page 17, could you elaborate on your rationale for selecting 'with k=27' for the analysis? Please provide an explanation for this choice.
>Authors: The coverage of the reads was so good that we increased the k number compared to the default value (k=21). We do not think that this value is pivotal in the output, which is in any case of very high quality.

6. On page 18, “The mitochondrial genome was assembled with MITGARD (Nachtigall, Grazziotin, and Junqueira-de-Azevedo 2021) using the WGS Illumina paired-end reads and the mitochondria of Gordionus alpestris (NC_044095.1) as a reference and selecting the clade Arthropoda. The mitochondrial genome was annotated using MitoZ with parameters annotate--genetic_code auto--clade Arthropoda (Meng et al. 2019)”
MITGARD is specifically crafted to extract the mitochondrial genome from "RNA-seq" data of various Eukaryotic species. To clarify the type of data involved in mitochondrial genome assembly, it's important to note that MITGARD focuses on RNA-seq data, which differs from approaches like MitoFinder, which exclusively rely on whole genome sequencing reads for assembly. It is also designed to find and annotate mitochondrial sequences in existing genomic assemblies. To further illustrate the contrast, it would be valuable to include comparative results between MITGARD and MitoFinder, and explain the methods in detail for clarity.
>Authors: We tried both MitoFinder with WGS and MITGARD with RNAseq data. However, MitoFinder somehow could not retrieve the mitochondrial genome in one fragment. However, MITGARD worked just fine with WGS Illumina data. We believe it may be due to the extremely high AT content in the genome of this species. In any case, we are not able to provide a comparison since the first method did not work out.

7. Table 1 highlights the completeness of BUSCO*completeness C:59.5%[S:59.3%,D:0.2%],F:12.1%,M:28.4:%,n:954. It suggests a relatively low completeness, and a lots of missing fragments which may indirectly imply potential issues with the assembly. It would be highly valuable to include a KAT kmer completeness score for a more comprehensive evaluation of the data quality.
>Authors: We are including a KAT k-mer comparison plot (Figure 5). However, we strongly think that this is not an error in the assembly. Please note that shortly after the submission of our manuscript, two more genomes of Nematomorpha were published (Cunha et al., 2023) with very similar percentages of missing genes. Taking into consideration that these are the first genomes for the full phylum and the peculiar parasitic lifestyle of these animals, we strongly believe this is not an artifact and indeed represents a genomic feature of these animals.