DR. PHILIPP SCHIFFER, GROUP LEADER, INSTITUTE OF ZOOLOGY, UNIVERSITY OF COLOGNE

p.schiffer@uni-koeln.de

Institut für Zoologie
Biocenter Köln
Zülpicher Str. 47b
50674 Köln
Deutschland

Tuesday, 30 May 2023

Ref: Re-submission PCI Genomics #235

Dear Dr Fernández,

With this letter we are are re-submitting a revised version of our manuscript PCI Genomics #235. We have worked to incorporate the extensive comments, in particular by Dr Laumer, made on the first version of our manuscript. We would like to thank both reviewers, for the comments they made and are convinced that they were very helpful to improve our manuscript. You will find the revised manuscript on bioRxiv, as required by the PCI system.

We provide point-by-point replies and comments to and on the reviewers' suggestions in a document attached to this letter.

We also fixed some issues pertaining to the data availability on the NCBI databank.

We hope that this revised version of our manuscript will be acceptable for a recommendation by you and are looking forward to receiving this.

With kind regards, on behalf of all authors

Dr. Philipp Schiffer

Dr Philipp Schiffer

Round #1

Author's Reply:

by Rosa Fernández, 10 Jan 2023 11:48
Manuscript: **https://doi.org/10.1101/2022.06.24.497508** version 2

Invitation to revise your manuscript

Dear Dr. Schiffer,

Two reviewers have now assessed your manuscript. While they are both quite enthusiastic about the findings of this piece of work, they both raised major concerns that should be addressed before recommendation. In particular, they are concerned about the robustness of your results supporting the monophyly of Xenacoelomorpha, and the lack of sufficient detail ensuring transparency and reproducibility, among other minor issues. I encourage you to carefully address these concerns in a revised version of your manuscript.

Yours sincerely,

Rosa Fernández

Reviews

Reviewed by anonymous reviewer, 15 Dec 2022 11:55

The paper «The slow evolving genome of the xenacoelomorph worm Xenoturbella bocki" by Schiffer et al. presents the first high-quality genome for a species of the genus Xenoturbella, which is the same as Xenoturbellida. The paper is very well written and relative easy to follow. The analytical steps are sound and state-of-the-art. This genome is large step forward towards our understanding of genome evolution in animals and will be a very useful tool for different kinds of research in the future. Most conclusions are well funded by the results, but I disagree with one conclusion the authors draw.

None of the results the authors present support the monophyly of Xenacoelomorpha (Acoelomorpha and Xenoturbella) and the phylogenetic analyses actually supports the non-monophyly. While the authors follow their phylogenetic results in concluding that Xenoturbella is placed within Deuterostomia, they disregard the non-monophyly of Xenacoelomorpha stating that Acoelomorpha are so fast evolving that this would mislead the phylogenetic reconstruction they conducted. However, their own results showed that the loss of genes in Xenoturbella and Acoelomorpha is not that different between the two groups. Moreover, there is no strong support yet that evolutionary rate at the sequence level is correlated with the rate of evolution at the synteny level. Finally, the tip-to-root distance for Hofstenia is not that different from the one of Xenoturbella. Hence, discarding the finding of non-monophyly of Xenacoelomorpha as an artefact appears like an ad hoc assumption in its strict meaning. As a consequence of this decision, the authors often treat findings related to the genome of Xenoturbella as relevant for the whole of Xenacoelomorpha, while the situation found in the genomes of Acoelomorpha are regarded as derived or data-insufficient given that they are of poorer quality.

My suggestion would be that the authors more clearly state the decision of favoring monophyly of Xenoacoelomorpha is a subjective decision by the authors in contrast to their own results.

- We have amended the introduction to clearly state the evidence for monophyletic Xenacoelomorpha and also state that we are following this line of argument in the present manuscript: *"Analyses of phylogenomic data sets have shown that Xenoturbellida and Acoelomorpha constitute their own phylum, the Xenacoelomorpha[3,4]. The monophyly of Xenacoelomorpha is also convincingly supported by their sharing unique amino acid signatures in their Caudal genes[3] and Hox4/5/6 gene[5]. In the present work we analyse our data in this phylogenetic framework of a monophyletic taxon Xenacoelomorpha."*

Moreover, anytime they represent the findings for Xenoturbella as representative of Xenacoelomorpha, it should be indicated each time that this is done given a supposed monophyly of Xenacoelomorpha. In this way, it is might clear that this conclusion is not necessarily in agreement with their own phylogenetic results and might apply in the end only to Xenoturbella and not Xenacoelomorpha. In this way, the opinion of the authors is still present in the paper, but the presentation is a little bit more balanced.

- There have been no published studies in the last 10 to 15 years not agreeing with Xenacoelomorpha. Nevertheless, we have now clearly stated that we are following our own line of argument here (see above).

In Figure 1A, please explain what the abbreviations mean.

- We have done so.

The section at line 185, the first paragraph should not be part of this section as it deals with a different question. The paragraph deals with phylogenetic position of Xenacoelomorpha and does not contribute anything to the molecular toolkit. The remainder of the section is on the toolkit and that it is not different from other bilaterians. However, this does not add anything to the phylogenetic position. Hence, this should be separated to avoid any implicit conclusions that one has to add something to the topic of the other, which they can't.

- We have split this section into two paragraphs with independent headings, as the reviewer suggests.

On lines 208-210, the authors state "Using our phylogenomic matrix of gene presence/absence (see above) we identified all orthologs present in any bilaterian and any non-bilaterian; these must have existed in the bilaterian ancestor." The conclusion is not a given. There are other options such hybridization, introgression, horizontal gene transfer and convergent evolution under similar selection pressures that could also result in this.

- The evolutionary scenarios suggested might well have happened, but are rather unlikely to affect a large number of genes. In particular, we would be very hesitant to suggest that the bilaterian ancestor arose through hybridisation, or that introgression played a huge role in its evolution. We are not excluding this possibility, but we do not wish to suggest this without any evidence in our data.

Finally, wrong gene family detection due to poorly supported gene trees can be another reason. This should be mentioned.

- We have added a caution in regard to the reviewer's sentence and a reference to the discussion of our presence/absence phylogeny.

On lines 255-258, the authors conclude that "Xenoturbella is, however, not significantly less complete when compared to other bilaterians considered to have low morphological complexity and which have been shown to have reduced gene content, such as C. elegans, the annelid parasite Intoshia linei, or the acoel Hofstenia miamia." The taxon sampling is very limited for this conclusion and many more complete bilaterian genomes are available. These should be mentioned here. Moreover, it should be explained how the author measure morphological complexity and what characters are the basis for this conclusion.

- We have changed the text accordingly, to now say that a large amount of evolutionary change happened in these lineages. We also refer to this in the introduction (see response to comment by reviewer 2). In regard to the taxon sampling, this has always to be incomplete. We selected some species, for which there are high-quality genomes available.

At line 266, the order of Hox genes in the brackets should reflect the order on the genome. This would highlight that typical colinearity of Hox genes seems not to be given.

- We have modified the listing as suggested.

In Figure 4b, it is not self-explanatory what the inset in bottom right corner explains? Please provide this explanation in the legend.

- We have inserted such an explanation into the figure description.

At line 356, it should be mentioned here that the ALG R is also present in the sea scallop.

- We mention the sea scallop now.

In Figure 5b, it is not clear, which scaffold is the aberrant scaffold. I assume that it is c1896 as it is highlighted in red, but it should be explicitly stated to avoid any confusion.

- We have made this clear.

Another small point is that the authors write concerning the phylogenetic reconstruction: "We calculated phylogenetic trees on these matrices using RevBayes (…), as described in ref 74,". This should be described in detail here as 74 runs two models, a reversible one and a Dollo one. It is not clear here, which one was used. This is especially relevant, if the Dollo analysis was applied. The Dollo criterion is a very strong assumption and hence it should be clearly stated if it had been used. If Dollo had been applied, I would advise to also conduct an analysis with the reversible model of 74 to test how the reconstruction performs without applying the very strong Dollo assumption.

- We have clarified this in the respective Methods section.

Reviewed by Christopher Laumer, 21 Dec 2022 19:47

Synopsis

Schiffer et al present a genome assembly, annotation, and comparative analysis of a representative of Xenoturbellida, perhaps the most evolutionarily interesting (and controversy-hounded) lineage of Bilateria, owing to its relatively simple gross morphology and uncertain phylogenetic position. They demonstrate a reasonably gene-space complete primary assembly and annotation of this small genome, and using HiC libraries scaffold roughly 3/5ths of the assembly into 18 linkage groups showing high levels of macrosynteny conservation with non-bilaterians and a representative deuterostome. A comprehensive orthology analysis shows, perhaps surprisingly regardless of the phylogenetic position of this lineage, a relatively Bilaterian-like gene content from the perspective of orthologue occupancy and signaling/transcription factor machinery, albeit showing a slightly higher than average loss of ancestral bilaterian orthologs (en par, I was surprised to see, with Hofstenia miamia, a representative of the acoelomorph sister lineage). A gene presense/absence phylogeny made with a different orthology declaration method and reduced taxon set shows strong support for a Xenambulacraria topology, even while splitting Xenacoelomorpha. Such an analysis is only possible with the whole-genome data presented here, and represents a refreshingly different and still somewhat novel approach to tackle this difficult phylogenetic problem to the familiar sequence alignment based inference methods, about which much has been published elsewhere. Numerous other "small" analyses (which I'm sure represent months of work in many cases), e.g. of miRNA content, neuropeptide complement, homeobox gene organization, phylostratigraphy, and symbiont genomics are presented, which shed light on many aspects of xenoturbellidan biology - doubtless this manuscript will help solidify our understanding of this enigmatic lineage, and stimulate deeper study in some unexpected areas. The phylostratigraphically anomalous & sparsely methylated chromosome is particularly interesting.

There are a few apparent weaknesses of the manuscript. It's evident that these data were generated some time ago, and that the technologies used to generate the primary assembly are now basically obsolete - I'm sure that 1/3 of a Sequel II flow cell or a single MinION flow cell could generate a much more contiguous (and probably somewhat more complete) assembly of this genome with much less bioinformatic acrobatics these days. This said, I think the authors demonstrate convincingly that for the specific analyses shown in this paper, focusing on coding gene content and a birds' eye view of macrosynteny conservation, this assembly is adequate to the task at hand, and a reviewer shouldn't ask for more than that. This said, I would not present this as a "high quality genome" by today's standards - it is fundamentally a highly scaffolded Illumina genome which was just about contiguous enough to further scaffold to a pseudochromosomal stage with HiC data.

- We do agree with the reviewer that sequencing this genome with long read-technologies would improve our assembly and have rephrased our wording to say “highly scaffolded”. Unfortunately, as the reviewer will appreciate, sampling *X. bocki* is not simple and usually only very view individuals are retrieved in a collection trip. We are not in a position to conduct more sequencing now.

Obviously, one of the major uses of the genome will be in providing new evidence for the phylogenetic position of this lineage. I think the strength of the gene presence/absence phylogeny based on whole genomes assigned to OMA orthogroups speaks for itself, and I have no particular qualm with the authors' methods or interpretation of these results. However I did find it strange that this mode of phylogeny-building was not explored for the taxonomically much larger orthogroup assignment done using Orthofinder. True there can be failures to detect true gene presence in transcriptomes, and the acoel transcriptomes that exist vary quite a bit in quality, but the cynic in me did wonder whether such analyses were conducted and not presented because it yielded results incompatible with the authors' previous body of work on this phylogenetic problem.

- We feel rather uncomfortable with the insinuation that we have been selective with the results presented. We represented the work we have done in its totality.

Some further justification of this decision, in any case, seems appropriate.

- We have not conducted Orthofinder based analysis of the phylogenetic position of *X. bocki*. Here we used OMA as the method as it is generally seen to be more stringent and we intended to reduce noise. We also focussed on a limited set of high quality genomes (our Orthofinder run does, for example, include transcriptomes and genomes with low BUSCO completeness).

By far the most glaring problems with the manuscript are in its method section and overall transparency/reproducibility. Almost all of the primary data used to generate these results was not made available during review so that even basic sanity checks e.g. through a k-mer analysis of genome size & heterozygosity were not possible. Numerous basic reports e.g. on library quality and assembly statistics in various stages of the assembly pipeline were not presented. Important analyses are alluded to but not shown (e.g. blobplots, de novo transcriptome assembly statistics/completeness). Several clear factual errors are apparent (e.g. in the instrument used to generate the core assembly), and where both lab and bioinformatic protocols are remarked on, they are often presented with such a low level of detail to as to forbid reproducibility. Indeed, many data types which were used for various small analyses (e.g. bisulfite sequencing, ONT sequencing) are not mentioned at all in the methods or supplement. I've given a fairly detailed account of where I see the absences in the notes below. For the most part, I have confidence in the quality of the datasets used to underpin this work, which was doubtless a lot of labor over many years, done the lab of a well established research leader in this field. I also do realize that these are *lots* of different experiments, and some of the data types are now no longer even on the market (e.g. TSLR). However, all published scientific literature should hold itself to a basic standard of transparency and reproducibility, which I would say this manuscript in its current form does not meet.

- We have taken the reviewers comments into account and addressed the detailed notes he makes below.

Signed,

Christopher Laumer

Detailed notes on introduction:

It seems the authors have made a choice not to cite any of the early molecular work plagued by contamination with molluscan gut contents. Follow up note: have the authors themselves done any screening for molluscan DNA?

- Yes, we did this and could not find evidence for this. As this problem was resolved years ago, we decided not to explicitly report on it in the main text. Along with the blobplot the reviewer requests, we are now alluding to this in the Methods and Supplementary.

From "line 70" - the authors refer to "a majority of studies" but cite only one (Cannon et al) - perhaps other citations are needed here?

- We have added 3 further citations at this point in the text.

From line 87 "The loss of...": A bit of a strange review - I'm not sure a barnacle or urochordate or neodermatan morphologist would characterize their study systems as morphologically simple. And is neoteny really a "new mode of living"? - I think of it as a hypothetical model of evolutionary change. I would have less issue with a statement to the effect that major ecological transitions are often accompanied by major morphological shifts, including loss of "bodyplan" level features and organ systems.

- We have rephrased this section, including to mention “remodelling” instead of simplification alone.

Line 103 "The only Xenacoelomorpha genomes available...": this is now out-of-date, with the preprint on the Symsagittifera roscoffensis genome, which is albeit very closely related to Praesagittifera. https://www.biorxiv.org/content/10.1101/2022.08.27.505549v1.full

- We have included a mention and citation of the *Symsagittifera* genome.

Detailed notes on Results:

The difference in size between the primary assembly (121M) versus the final assembly (117M) suggests that very little sequence was removed by redundans as haplotypic duplication - is this correct? Was the genome relatively homozygous, e.g. as judged by kmer content?

- This is correct. The genome was relatively homozygous and we have now included a GenomeScope profile into the Supplementary to inform on this.

I find the interpretation of false-negative orthology detection due to fast rates of sequence substitution leading to a splitting of Xenacoelomorpha in the p/a phylogeny quite credible, actually. There was an interesting paper recently published that looked at rates of false negative orthology detection and showed this to be a pervasive problem in taxonomic lineages that are poorly sampled and/or fast-evolving:
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000862

- We are glad to read that reviewer approves of our interpretation of the data. As requested by reviewer 1, we have nevertheless added some further discussion on this, including citing the paper the reviewer suggests and one by us.

I would be interested to know how many lineage-specific gene births and losses are recovered for Xenoturbella+Ambulacraria in this presence-absence analysis. Does the taxon-restricted gene set have any particular characteristics – e.g. average alignment length, compositional bias, distribution throughout the genome? If they are “perfectly normal” genes it strengthens the argument that this relationship is unlikely to be an artifact. I would be particularly interested in knowing if any of the “Xenambulacrarian” genes are particularly enriched on c1896, which was a very striking outcome of your analysis.

- The reviewer is very right that these effects could be of great interest. At the present time we do not have made any striking observations made based on our methods. In particular, there is no enrichment of Xenambulacrarian genes on c1896.

Line 207: is Xenoturbella slow-evolving or fast-evolving? The title says one thing, here another. Perhaps a little more clarity on what precisely is meant by this is worthwhile.

- We have clarified that the evolutionary change is seen mostly on the morphological level.

A comment on BUSCO completeness. One of the authors (PHS) was kind enough to share the genome annotation and assembly with me for use in a classroom module, before I was asked to review this manuscript. By my own hand, the BUSCO completeness of the genome annotation against Metazoa odb10 in BUSCO5 (run in peptide mode) was:

C:82.5%[S:77.8%,D:4.7%],F:7.0%,M:10.5%,n:954

Which is substantially lower than the 90% quoted in the manuscript - is this down to a difference in database versions or has something else happened?

Also, in comparison to the figures from the genome annotation (whether 82.5% or 90%), I have separately done some re-analyses of publicly available RNA-seq data from X. bocki generated by other groups. My trinity assemblies give figures like:

SRR2681987 C:93.8%[S:22.6%,D:71.2%],F:2.2%,M:4.0%,n:954

SRR5760181 C:94.8%[S:27.5%,D:67.3%],F:1.7%,M:3.5%,n:954

This to my mind does indicate there's some 40-50 metazoan BUSCOs present in adult RNA-seq data which are not represented in the genome annotation, potentially indicating some, although not a large level of, incompleteness of the genome assembly.

- The peptide set contains ~90% partial BUSCOs and 82,5% of these are complete. We now include this information correctly.

Line 211 "derived bilaterians" - In my view & in the views of many other systematists, no living species is more or less derived than any other - individual characters can be derived, but not whole organisms. I have a similar problem with reference to "early branching lineages" made elsewhere in the manuscript - from the perspective of cnidarians, for instance, mammals are a representative of an early-branching lineage. I am sure that the authors won't disagree with these fundamental points, but the language that we sometimes use to describe these phenomena I think does bias our thinking towards a ladder-like view of evolution, and I would prefer to see these organismic comparisons made with less hypothesis-laden descriptors.

- We have edited the manuscript to contain less of the descriptors the reviewer objects to.

Very interesting that the bilaterian orthogroups have a similar occupancy for both Xenoturbella and the supposedly much faster evolving Hofstenia!

I am missing in Figure 3a data from the acoelomorphs incorporated into this analysis. Obviously you cannot show all 155 species in an easily readable way, but as a comparative point, it does seem essential to compare the Xenoturbella species available to the acoelomorphs, particularly if you are trying to argue that this genome is especially faster or slower evolving than those from representatives its sister lineage.

- We have included two acoels from our analysis into figure 3a as suggested by the reviewer.

For figures 3b/3c, are the trees at the bottom of the heatmaps dendrograms on the heatmap data, or are these schematic cladograms, and if so, from where do they originate?

- These are schematic cladograms drawn by us. We have mentioned this in the figure legend now.

I shall mostly refrain from commenting on the neuropeptide section as this is outside my expertise. I was curious, however - you see the K/R-RFP-K/R motif, which is reasonably compelling if anecdotal as a molecular synapomorphy, in Xenoturbella and in the ambulacrarians, but not in the nemertodermatid transcriptomes - is it also present in the reasonably complete transcriptome from X. profunda? Or any of the acoels?

- The analysis the reviewer is asking for was already performed previously, and we like to refer them to Thiel et al. 2018 (10.1093/molbev/msy160) to answer their question. Out of curiosity, we now conducted some blasts using the X. *bocki* Prokineticin sequence against various database in NCBI (with acoels and *Xenoturbella*) and the only 4 matching sequences are reads from a *Xenoturbella profunda* transcriptome used in Thiel et al. (Bioproject: PRJNA295687 / SRA: SRX1280426 / SRA Study: SRP064117 / SRA Run: SRR2500940). As the analysis of acoels is not an aim of our manuscript we did not elaborate on this further.

On ancestral linkage groups. Indeed the conservation of macrosynteny visible with other metazoans is an impressive feature of this assembly, and a compelling demonstration of the success of this paper. I think it would help improve the readability of this manuscript if you could e.g. put a bold-line box around some of the clade-specific linkage groups you discuss in the oxford plots.

- We have added boxes to the figure.

Also, I don't understand the argumentation around "prebilaterian" linkage groups.

Neither the Nephrozoa hypothesis nor the Xenambulacraria hypothesis posits that Xenacoelomorpha are non-bilaterians - why would the absence of a eumetazoan plesiomorphy say anything decisive about either hypothesis?

- The reviewer is right in his observation. However, at this point we are only reporting our findings in the light of a hypothesis. We have made our argument clearer at the end of the discussion.

On the anomalous small chromosome: super interesting result, and indeed perhaps the start of some really interesting Xenoturbella-specific biology. I wonder if this will be seen to occur in any acoelomorph genomes going forward. I would be very keen to see if gene tree topologies/delta-likelihoods of orthologs occurring on this chromosome are any different on average to those occurring elsewhere in the genome - for instance indicating a potential horizontal origin (albeit you don't see a large signature of this in the global analysis of HGT). One other thing: I couldn't really find anything in the figures corresponding to the synteny with the E. muelleri scaffolds described in the text - could you make this clearer? Indeed I don't see the c1896 labelled on the dot plot with E. muelleri.

- There is no signal for HGT into c1896. Additional tests in regard to gene content do not show striking results (see reply below). In figure 5a we only show scaffolds with clear synteny blocks, this is not the case for c1896. We have clarified this through a change in the main text: “We did observe links between the aberrant scaffold and several scaffolds from the genome of the sponge *E. muelleri* in regard to synteny, but could not detect distinct synteny relationships to a single scaffold in another species.”

Detailed notes on Discussion:

The discussion of "intermediate" genomic traits such as miRNA counts and linkage group organization feels a bit phenetic to me. Surely all of the traits mentioned could be analyzed cladistically, in a search for synapomorphies. Simply intermediate numbers of various character states shouldn't be compelling on their own.

- This is beyond the scope of the current manuscript, but could be a very valuable contribution for the future.

Is the relatively canonical gene content of Xenoturbella informative either way on the Xenambulacraria vs Nephrozoa debate? I'm not sure I agree that it is. I think for instance of the Dimorphilus genome which was recently published, showing many features we would expect a typical annelid genome to have, despite the highly reduced body size and morphological simplification of this lineage. This to me shows a decoupling between a birds' eye view of genome biology and morphology. So indeed, while the relatively bilaterian-gene rich genome of Xenoturbella is consistent with the Xenambulacraria hypothesis, it's not *inconsistent* with the Nephrozoa hypothesis either. I do like your phrasing of the "strong" Nephrozoa hypothesis not being supported - this does imply, however, that a "weak" Nephrozoa hypothesis is possible (presumably meaning a Nephrozoa true tree topology but little obvious genomic "pre-bilaterian-ness" as one might naively think if one interprets xenacoelomorph morphology as primitively simple and gene content as predictive of morphology). Indeed, you say as much in the final paragraph of the discussion - I think this measured reading is appropriate and commendable.

- We are glad the reviewer appreciates our line of argument and our aim to not overstate findings.

Sentence beginning line 461: I had thought that the long branch lengths seen for acoels in your presence/absence trees would indicate a high rate of acoelomorph-specific gene births, rather than a high rate of loss? Is it possible to disentangle these?

- The reviewer is right that this is an interesting question, but we think not a core focus of our manuscript on *Xenoturbella*. Surely, a comparative assay into acoel genome evolution is needed, once an array of high-quality assemblies are available for the taxon.

Another possibility about the anomalous chromosome: could this be a germline restricted chromosome? We do expect that these should have younger genes on average, and these are also usually small. Does it show a different level of average coverage in the raw reads to other chromosomes?

- We prefer not speculate and suggest leaving this for future analysis including more species of the genus.

Detailed notes on Methods:

Which phenol-chloroform protocols and Qiagen kits were used for DNA extraction? In the results it's asserted that HMW gDNA was extracted – how was this ascertained/QC'd?

- We have provided more detailed information on this.

It's concerning to me that the authors state a 2x250 bp read format was used on the HiSeq 4000, as this platform does not offer that read length. Perhaps it was a HiSeq 2500 2x250 rapid run?

- The reviewer is of course correct in his observation and we thank them for catching this typo. We have corrected the Methods section accordingly.

It would be good to see one of the blobplots referred to, to convince the reader that this really is an uncontaminated genome assembly, despite the efforts to starve the specimens before extraction. This is a part of the tree of life for which few close references exist and it can be tricky sometimes to judge the source of contaminants from blobplots on such species - perhaps better to show rather than tell.

- We have included a blobplot into the supplementary information now (see above).

I am missing here some basic statistics (perhaps best shown in a table) on these assemblies during various points in the process, e.g. right after SPAdes, after redundans, after BADGER. The authors cite a scaffold N50 of 60 kb before HiC scaffolding, but what is the contig N50 before any scaffolding?

- We have included this information now in the Supplementary as a table.

Indeed, the authors refer to mate-pair libraries but do not give any details on the protocols used to generate these datasets, the size of the mate pairs, QC statistics...

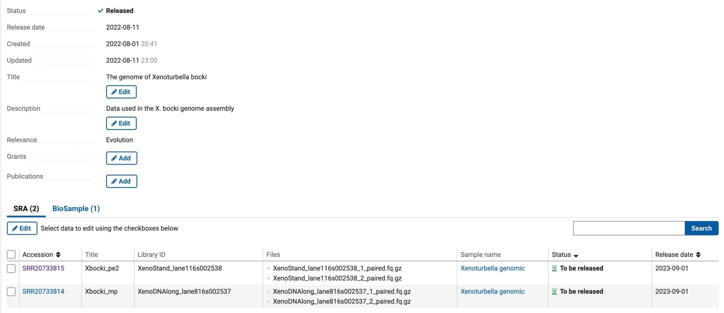
We have included this information now via a reference to a manuscript in which this initial dataset is described in the Methods section.

Redundans should be cited.

- We have included this citation now.

I was unable to find a link to the raw reads (except for two HiC datasets) or assemblies used in this paper. I was hoping to do some basic analyses, e.g. kmer spectra, just to cross-check for instance that the assembled genome size matches the kmer estimated size, and to determine what proportion of kmers in the reads were not represented in the assembly. Without such data (which could have been uploaded to SRA and embargoed for public release) it's difficult to fully review this manuscript. I will note that the genomescope analyses I made of the two HiC datasets were somewhat concerning, with no visible spectra outside an error distribution - perhaps these are low-diversity libraries, or highly contaminated libraries?

- We are very sorry the reviewer could not access the reads on the NCBI portal. However, this seems to be an error on the portal as we had made the data available in accordance with the PCI requirements. The screenshot below shows that the status of the data is “released”, while in fact they are still not. We have been in contact with NCBI and the data was manually released now.



Status	✓ Released
Release date	2022-08-11
Created	2022-08-01 20:41
Updated	2022-08-11 23:00
Title	The genome of Xenoturbella bocki
Description	Data used in the X. bocki genome assembly
Relevance	Evolution
Grants	
Publications	

SRA (2) BioSample (1)

Select data to edit using the checkboxes below

□	Accession	Title	Library ID	Files	Sample name	Status	Release date
□	SRR20733815	Xbocki_pe2	XenoStand_lane116s002538	XenoStand_lane116s002538_1_paired.fq.gz XenoStand_lane116s002538_2_paired.fq.gz	Xenoturbella genomic	To be released	2023-09-01
□	SRR20733814	Xbocki_mp	XenoDNAlong_lane816s002537	XenoDNAlong_lane816s002537_1_paired.fq.gz XenoDNAlong_lane816s002537_2_paired.fq.gz	Xenoturbella genomic	To be released	2023-09-01

Some concerns about the HiC protocol and data presented here. Fixing a whole animal vs fixing cryohomogenized tissue is likely to lead to poor results from autolysis as the fixative penetrates large volumes of tissue.

- The proximity ligation data were generated using a fresh live sample immediately in the aftermath of a collection campaign. At the time (2014) we worked with a custom made variation of a protocol that was developed and applied over a variety of species, and not from current standardized Hi-C kits or more recent optimized versions of the protocol. In our hands, fixation after cryogenization leads to poorer results than fresh fixation when it comes to proximity ligation protocols. Here we have fixed a segment immediately after segmentation from a living X. bocki for 1 hour. Given the ability of FA to propagate quickly into tissues, we consider autolysis had only very limited effect if any on the outcome. Note that the fixation on fresh samples was applied on other species, some of which were larger than the small fragments of *Xenorturbella bocki*, with good results. We recommend also in the methods more recent version of the protocol if the readers is interested.

There's no indication that the DpnII enzyme was heat-killed before proceeding to fill-in.

- The enzyme was indeed inactivated before centrifugation, then the extremities filled-in.

Numerous volumes used (for formaldehyde and SDS, for example), and enzyme details (which "ligase"? what manufacturer?) are missing.

- We now provide more information regarding these details.
  Formaldehyde = 3% final in 30 mL (Sigma - F8775-4x25mL)
  SDS = 0.3% final (7.5 µL of SDS 20% for 0.5mL of reaction)
  Ligase = T4 DNA ligase (New England Biolabs)

It's not clear what protocol was used to prepare the extracted 3C DNA into an Illumina library, or how the biotin selection was performed.

- The Hi-C protocol used here correspond to our custom made approach at the time of the sampling, it does involve a biotin selection but the efficiency at the time remained relatively low compared to current kits. Nevertheless, it provided sufficient contacts to bridge scaffolds. Process for illumina sequencing was done using the illumina kit TrueSeq, we believe that pointing the reader to the manufacturer instruction should be sufficient for such a commonly used approach nowadays.

And most concerning of all, I can't really see any QC data on this library - at very least, the authors should be showing the pair-length distribution and the contact heatmaps which have become standard in the field, so that readers can judge how strong the evidence for the chromosome scaffolding is.

- We have included the requested figures in the supplementary now.

I think, as instaGRAAL is a published method, it's not necessary to explain its algorithm in detail here - just the parameters that were used to run it.

- As we were using a pre-release version of the software, we have chosen to keep the detailed description. We hope the reviewer and readers will appreciate this.

The protocols used for RNA extraction and cDNA library preparation should be specified in enough detail for another lab to reproduce this work.

- We have provided more information on this and the exact protocol can be requested from the corresponding author (stated in the manuscript now).

It would also be good to see some rough statistics on the Trinity assembly, so readers can judge its completeness and contiguity. Again, I could not find any RNA-seq reads used in this study uploaded to the SRA.

- We have included this information and will upload the reads to SRA (SAMN35083895).

The authors refer to additional single-cell transcriptome data - if these were used in the annotation of this genome, surely the experiments used to derive these should also be described in the methods section and deposited into public databases?

- We originally planned to submit a manuscript describing these experiments in parallel with this manuscript. The sc-Seq manuscript has now been submitted elsewhere and is in-review. We have mentioned this now in the Methods section and included a reference.

Question: during setup for orthology inference, for the species for which RNA-seq data only were used as input, were *only* those genes with positive hits against UniProt/Pfam retained in the protein prediction, or was this simply used to improve the sensitivity of the predictions? I am wondering if this pipeline might exclude novel taxon-specific orthologs with no sequence similarity to existing databases.

- We did not remove proteins, the UniProt/Pfam hits were only used for improvement of sensitivity. We thus assume that novel taxon-specific orthologs are present.

I don't see any problem per se in the way that the gene presence/absence phylogenies were generated, but I am curious why the OMA algorithm, and apparently a separate species set, was employed in this while the Orthofinder analysis should also in principle be well-suited to this kind of analysis. Do the results differ with a larger taxon sample?

- The reviewer is correct in principle that the Orthofinder analysis can be used. However, to reduce noise in the analysis we opted to only select a subset of species (which were part of the OF analysis) and used the very stringent OMA algorithm for this.

In the homeobox section, ONT reads are mentioned for the first time. There's no information given in the manuscript about the volume and quality of these data, and how they were generated. I also find it strange that these were used only in the context of homeodomain-containing contig analysis - why not also incorporate them into the primary SPAdes assembly?

- We obtained very limited ONT data amounting to 0.5x genome coverage from the last *X. bocki* individual we were able to source in the framework of this project and after the genome was scaffolded. These data were not useful in the genome construction process. We are now reporting how we obtained these reads and that they were not useful to further improve the genome. A comment on this was added to the Methods section.

Line 751 "We extracted a highly contiguous..." - how was this extraction performed bioinformatically?

- This was done by identifying contigs with blast. We have added this information now.

Line 752 - As I understand it, LINKS is a scaffolder, not a polisher.

- We corrected this.

BUSCO is mentioned - including re-analyses of public data such as the Hofstenia genome - but the parameters/database versions used to run this software seem not to have been reported.

- We have included the information that BUSCO_v5 was used.

Similarly: there are some results on methylation reported in the supplement, but no mention is made of how these results were obtained - was this bisulfite sequencing? If so, how were these libraries generated and these analyses performed?

- We have now included a short description of the methods used for bisulfite sequencing, and the corresponding downstream analysis in the Methods section.