Round #1

Author's Reply:

---

by Philipp Schiffer, 31 May 2024 13:32
Manuscript: https://doi.org/10.1101/2024.02.21.581367
version: 1

MATEdb2, a collection of high-quality metazoan proteomes across the Animal Tree of Life to speed up phylogenomic studies

Dear Dr Martínez-Redondo, dear Dr Fernández,

two reviewers have now concluded their assessment of your manuscript "MATEdb2, a collection of high-quality metazoan proteomes across the Animal Tree of Life to speed up phylogenomic studies". Please do excuse that this process has taken some time, I had been waiting for the review of a third reviewer, which eventually was not delivered.

As you can see, both expert reviewers have provided favourable reviews for your work. However, both reviewers also do suggest some improvements to be made, before the manuscript can be recommended. I would kindly ask you to look at the valuable comments made by both reviewers and enact the changes and improvements they suggest, before re-submitting the manuscript for a second round of reviews.

Best regards

Dr Philipp Schiffer

> Authors: We thank Dr. Schiffer, Dr. Natasha Glover, and the anonymous reviewer for their comments and are pleased they liked the resources presented here. Please find below our responses to the reviewers' comments point by point.

Reviews

Reviewer 1

Review by Natasha Glover, 07 May 2024 17:08

The work by Marinez-Redondo et al provides a valuable resource for the animal genomics community. Due to the heterogeneity of data sources (genomes, transcriptomes) and their subsequent structural and functional gene annotations, it is important to be as consistent as possible to obtain the gene repertoire. Additionally, it is important to balance obtaining high-quality gene repertoires with obtaining a taxonomically diverse set of gene repertoires. The authors did a good job of reconciling these two points and provide a nice resource for other researchers. The paper is easy to understand and well-written.

- Title and abstract

○ Does the title clearly reflect the content of the article? [X] Yes, [ ] No (please explain), [ ] I don't know

○ Does the abstract present the main findings of the study? [ ] Yes, [X] No (please explain), [] I don’t know

The main problem I’ve seen in the abstract is that the sentence “Here, we present the newest version of MATEdb MATEdb2) that overcomes some of the previous limitations of our database… (2) we provide gene annotations from genomes obtained using the same pipeline.” This is misleading to me because it gives the impression that the MATEdb2 pipeline performs the structural annotation (i.e. ab initio combined with homology and transcriptomic data). But from what I understand, the pipeline uses the GFF file and assembly provided by the genome sequence’s main research group. Then it has a pipeline to use the GFF coordinates to extract the gene sequences from the assembly. While I appreciate that transcriptomic gene annotation is the main benefit of the MATEdb2 pipeline, rather than genomic annotation, it’s just a bit misleading by the wording.

> Authors: We have corrected the wording of this sentence to make it clearer.

Introduction

○ Are the research questions/hypotheses/predictions clearly presented? [X] Yes, [ ] No (please explain), [ ] I don’t know

○ Does the introduction build on relevant research in the field? [X] Yes, [ ] No (please explain), [ ] I don’t know

I appreciate the motivation for creating MATEdb2, as it is quite cumbersome to process genomes and transcriptomes for comparative genomics studies. The variability in data quality can and does affect the downstream analyses.

It would be good to quantify some of your anecdotal evidence: For example, if you compare a set of transcriptomes with different versions of Trinity– in how many species does the number of genes change significantly?

> Authors: We have used a subset of publicly available transcriptomes to compare the difference in the number of genes obtained by us and the transcriptome used in the original paper. Results can be found in Supp. figure 1.

Also, regarding the paragraph: “However, a closer inspection of both files together with their corresponding genome sequence and annotation revealed incongruences between them that needed to be manually curated. This is caused by the lack of consensus in the annotation and publication of genome files, with some authors uploading modified versions of the protein sequences that do not map directly with the reported GFF and FASTA file, hindering the utili of those files for additional analyses.” How many times exactly in a given set of genomes does the GFF not match the provided proteome fasta file? Is it off by just a few genes, or many genes? This would provide better evidence for the motivation and need for MATEdb2.

> Authors: We added the example of a discarded genome and a comparison of genomes shared between the newest and oldest versions of MATEdb.

● Materials and methods

○ Are the methods and analyses sufficiently detailed to allow replication by other researchers?

[X] Yes, [ ] No (please explain), [ ] I don't know

○ Are the methods and statistical analyses appropriate and well described? [X] Yes, [ ] No

(please explain), [ ] I don't know

A good summary of the methods is included in the paper and more details on the GitHub. I did not try to replicate but singularity containers are provided.

Minor point: There is a mistake in the formatting for the AGAT citation.

> Authors: The citation format has been corrected.

Results

○ In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [ ] Yes, [ ] No (please explain), [X] I don't know

○ Are the results described and interpreted correctly? [ ] Yes, [X] No (please explain), [ ] I don't know

There is not a Results section per se, as this is a paper to describe a new tool/database. However, I had a look at the supplementary Table S1. It looks mostly good, but I spotted an anomaly that could be a mistake: For *Panulirus ornatus*, there are a reported 252,598 genes annotated. However, this is an unusually high number of genes, and when I checked in the reference paper (https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07636-9), they reported 99,127 genes.

This also seems like a high number, and I imagine due to false positives in the ab initio gene prediction. Please double check these numbers.

> Authors: We double-checked the numbers and the ones reported by us are the ones you get when executing AGAT on the publicly available GFF file the authors uploaded to Zenodo (https://doi.org/10.5281/zenodo.4589425). We agree that this is likely a problem with their prediction and the real number of genes might be much lower. In the original paper, there is also a supplementary table that says that the ab initio gene prediction in the genome of this species gave 163,364 genes and not the 99,127 genes they reported on the main text. However, there is no clear explanation of these discrepancies in the text.

In order to spot these potential outliers/errors, it would be nice to report in the paper a plot with the number of genes inferred for each species or the distribution of gene numbers across all species.

> Authors: We have included a plot with the distribution of gene numbers across phyla.

I appreciated the functional annotation of the genes using both orthology-based methods and protein language models. I did not look into it in detail, so I cannot comment on the suitability of the technique. However, NLP for protein function prediction seems promising, especially for those proteins of unknown function.

Another suggestion is to make it more clear in the paper where the actual data can be found. MATEdb(2) is called and treated as a repository/ database, but in the Data Availability section, it only talks about the necessary scripts and information to obtain all the transcriptomes and proteomes. I had to dig to find out where the actual fasta sequences are found (https://github.com/MetazoaPhylogenomicsLab/MATEdb2/blob/main/linksforMATEdb2.txt). Please make it more clear in the manuscript that the cds data itself is available for download.

> Authors: This has been made clearer in the manuscript.

Discussion

○ Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [ ] Yes, [X] No (please explain), [ ] I don't know

○ Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [X] Yes, [ ] No (please explain), [ ] I don't know

As mentioned above, the main limitation of this database is that for the genome annotations, it still relies on heterogeneous structural gene annotations performed by various research groups. There still may be biases in the different techniques, as illustrated by the variable gene numbers in Table S1. However, I think the transcriptome annotations are more trustworthy since they are done all with the same method, and only the SRA raw data is what is variable.

> Authors: We have included a discussion paragraph at the end of the manuscript emphasizing the aforementioned limitations of the method.

Reviewer 2

Review by anonymous reviewer 1, 22 May 2024 09:52

This manuscript by Martínez-Redondo et al., entitled “MATEdb2, a collection of high-quality metazoan proteomes across the Animal Tree of Life to speed up phylogenomic studies” presents MATEdb2, an updated version of the Metazoan Assemblies from Transcriptomic Ensembles database. This database includes high-quality proteomic data from nearly 1000 animal species across various phyla. MATEdb2 aims to address previous limitations by expanding taxonomic coverage, standardizing gene annotation processes, and utilizing advanced protein language models for functional annotation. The database was generated with the purpose to facilitates comparative genomics and phylogenomic research by providing standardized and easily comparable datasets. A dedicated GitHub repository accompanies the database which provides impressive documentation and computationally reproducible scripts.

Overall, the manuscript is very well written and easy to follow. The bioinformatics analysis pipelines presented by the authors are sound and use community standard software. The integration of protein language models for functional annotation represents a cutting-edge approach, offering deeper insights into protein functions. However, quality control could be improved to further reduce analysis artefacts introduced by meta-analyses.

Together, I can support an editorial decision to recommend this publication for peer review.

Major comments:

1.) **Quality Threshold Adjustment**: Lowering the quality threshold for inclusion (from 85% to 70% BUSCO scores) might introduce less reliable datasets, potentially affecting downstream analyses. While BUSCO is indeed used broadly for genome quality assessments, this approach is still heavily debated within the bioinformatics community and even BUSCO scores >90% can be misleading for individual cases. Any meta-analysis should be aware of this. Have the authors thought about adding additional quality controls that were particularly designed for meta-analyses?

> Authors: As mentioned, it is a current area of research and debate in the scientific community. We decided to go with the standard on the field and leave any additional controls that may be more specific to the research objectives up to the user. Whenever there were many options within a lineage, we selected transcriptomes or genomes of higher quality in terms of read or assembly quality, from which it is more likely to get a more complete gene repertoire.

2.) **Dependence on Public Data**: While species sampling has vastly increased thanks to their extended database. The database continues to exclude certain taxa/species or introduce biases based on the availability of high-quality data for certain species/domains/groups. It would be be very useful if the authors would provide a meta annotation table where an accumulation of poorer quality species within a domain/group/lineage are highlighted, so that users are more careful when interpreting downstream results from this particular domain/group/lineage.

> Authors: Low-quality data from certain taxa/species will affect any downstream analysis, which is why selecting the part that has a high quality (and is more reliable) is important even though doing that will create some taxonomic bias, especially for some research questions. Huge efforts are currently being made to generate high-quality genomes of eukaryotic species that will reduce in the following years the taxonomic bias from public databases. In the meantime, to deal with this taxonomic bias, our database will be expanded as we incorporate new datasets from underrepresented lineages, such as nematodes, or as requested to be incorporated by the scientific community if resources allow it. Still, for the current version of the database, some publicly available datasets of underrepresented lineages had to be discarded due to low quality. We have included a supplementary table of

data that we discarded when creating the database so that anyone interested could see which public datasets have low quality.

3.) **Manual Curation Needs**: Did the authors provide any manual curation or manual quality checks to confirm that quality metrics applied in the meta-analysis are indeed meaningful when randomly sampling species for manual inspection?

> Authors: We have not performed any manual check of the quality metrics, as we are using standard methods for a general quality check and further analyses go beyond the scope of the paper.

4.) **Annotation Consistency**: While the standardized pipeline improves consistency, differences in annotation quality and completeness across datasets might still pose challenges for comparative studies. Have the authors taken this annotation bias into account (see e.g. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000862)?

> Authors: We are aware of the putative challenges of heterogeneity across datasets for further studies, but most published studies completely ignore these differences, which is why we tried to reduce this as much as possible at the bioinformatic level. A discussion paragraph has been added discussing this heterogeneity in the genome datasets. Nonetheless, there is still a limitation associated with the lack of resources for some lineages, as most benchmarks or statistics are usually estimated on a small set of organisms. For instance, the current standards for genome assemblies proposed by the Earth Biogenome Project are based on vertebrates. For some lineages, it is very difficult to reach these standards due to biological limitations (such as the amount of sample needed from one individual). The same applies to any type of data. Even if we increased the quality of the data or reduced the heterogeneity of the datasets, which would require an extra effort not only at the bioinformatic level, the problems would only be alleviated to some extent. Using the example from the paper above, there would still be a considerable amount of putative lineage-specific genes due to the limitations of the current methods.

5.) **Data Retrieval**: Currently data retrieval is not automated. It would be very useful if the authors would provide a download script or detailed tutorial on how users can efficiently retrieve the full database. Also a database management scheme (how was the data organised and standardised) would be useful to further facilitated automated downstream analysis.

> Authors: The previous MATEdb contained the full database on a single compressed file heavy in size. In this newer version, we decided to create download links for each of the files as some users may only need a part of the datasets and downloading the full database would take a long time due to its large size. Currently, an automated retrieval of data is not possible due to hosting and resource limitations.

6.) **Referencing Software Dependencies**: Although the authors list the software their workflow is depending on, they don't cite the corresponding papers to the software. I strongly recommend citing the relevant papers for the software and software version they employ.

> Authors: We have included citations of the used software.

7.) **Long-term database management**: It was not clear to me what the long-term plan for database hosting is. Will the database be hosted for XY years and further extended? More details about the "hosting service" aspect would be useful for users to decide whether or not they wish to invest in relying on this database infrastructure.

> Authors: The database is hosted on our own server and will be there indefinitely. The database will be expanded as we incorporate new datasets from underrepresented lineages, such as nematodes, or as requested to be incorporated by the scientific community if resources allow it.