Peer Community In
Genomics

MATEdb2 is a valuable phylogenomics resource across Metazoa

Philipp Schiffer *based on peer reviews by* ***Natasha Glover*** *and 1 anonymous reviewer*

Gemma I. Martínez-Redondo, Carlos Vargas-Chávez, Klara Eleftheriadi, Lisandra Benítez-Álvarez, Marçal Vázquez-Valls, Rosa Fernández (2024) MATEdb2, a collection of high-quality metazoan proteomes across the Animal Tree of Life to speed up phylogenomic studies. bioRxiv, ver. 2, peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2024.02.21.581367

Submitted: 07 March 2024, Recommended: 21 August 2024

Cite this recommendation as:
Schiffer, P. (2024) MATEdb2 is a valuable phylogenomics resource across Metazoa. *Peer Community in Genomics*, 100371. 10.24072/pci.genomics.100371

Published: 21 August 2024

Martínez-Redondo and colleagues (2024) present MATEdb2, which provides the scientific community with Metazoa proteomes that have been predicted and annotated in a standardised way. The authors improved the taxon representation from the earlier MATEdb and their current database has a strong focus on Arthropoda, Annelida, and Mollusca. In particular, for the latter two groups not many high-quality reference genomes are available. Standardisation of the prediction and annotation process in a reproducible pipeline, as integrated in MATEdb2, is of great value, in particular to infer phylogenies as correctly as possible. Thus, I am sure that MATEdb2 will be an excellent go-to resource for phylogenomic studies, as well as for probing the biology of new, obscure species, especially marine ones.

The manuscript was evaluated by two experts in the field of orthology search and orthology databases, and computational biology. The authors diligently implemented the modifications suggested by both referees and I am gladly recommending the manuscript at this point.

References

Martínez-Redondo GI, Vargas-Chávez C, Eleftheriadi K, Benítez-Álvarez L, Vázquez-Valls M, Fernández R (2024) MATEdb2, a collection of high-quality metazoan proteomes across the Animal Tree of Life to speed up phylogenomic studies. bioRxiv, ver. 2 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2024.02.21.581367

Reviews

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/2024.02.21.581367
Version of the preprint: 1

Authors' reply, 25 June 2024

Download author's reply
Download tracked changes file

Decision by Philipp Schiffer, posted 31 May 2024, validated 31 May 2024

Revision and additional work needed for your MATEdb2 manuscript

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

Reviewed by Natasha Glover, 07 May 2024

Download the review

Reviewed by anonymous reviewer 1, 22 May 2024

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

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

Results

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

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

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

Please see review.

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

Review:

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?

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.

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?

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)?

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.

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.

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.