Round #1

Author's Reply:

by Samuel Abalde, 22 Aug 2022 11:02
Manuscript: **https://www.biorxiv.org/content/10.1101/2022.07.18.500182v1**

MATEdb: a new phylogenomic-driven database for Metazoa

Dear members of the Metazoa Phylogenomics Lab,

Thank you for your submission and apologies for the delay in our response; I am sure you understand this is a hard time of the year to get the review of a manuscript done. You will see two reviewers have agreed to review your manuscript. I was waiting for a third one, but because I did not want to delay my decision any longer I have decided that my own review would be enough.

We all agree on the merit of this work. Not only MATEdb has the potential to become an important resource for the community, but the manuscript clearly describes the importance of making such database available. Thus far, this project is a great example of open science and I would like to congratulate you for that.

> Authors: We thank Dr. Abalde and the two anonymous reviewers for their comments and are pleased that they are excited about the resources presented here. Please find below our responses to the reviewers' comments point by point.

Importantly, we all have comments that should be addressed before the final acceptance or, in PCI terms, recommendation of this manuscript. You can see the reviewers comments in the revision panel, but I will detail my own comments here below:

My major concern regarding this manuscript and the database is that it feels misleading. The database name and repeating sentences along the manuscript such as "*we present here Metazoan Assemblies from Transcriptomic Ensembles (MATEdb v1), a continuously updated and curated database of hundreds of high-quality transcriptome assemblies from different animal phyla*," make you think of a sound metazoan database with many phyla represented. However, up to this moment only two phyla are included: Arthropoda and Mollusca. I presume your idea is to expand this database, but this manuscript should present the database as it is now. This (temporary) incompleteness should be addressed, either by toning down this kind of sentences or by explicitly saying something along the lines "*we include these phyla now but we are working on incorporating these other at the moment*". Related to this, a paragraph summarizing your road map of future work would make a good addition to the manuscript.

> Authors: Indeed, we are thinking of expanding the database very soon (ie, in a few months) with genomes and transcriptomes of multiple animal phyla, this is why we decided to coin the database as MATEdb. We intend this manuscript to be the core of the database and the general description, and aim at updating it with new versions as soon as we accumulate a reasonable number of new genomes and transcriptomes. The updates will be explained in detail in the new versions of MATEdb, so we are confident that the readers will find all the information they need to fully understand the content of each of the different versions. We are clarifying this in the revised version of the manuscript, as suggested.

Another major concern -that could also be considered a suggestion for future work- is that no references have been made about how you made sure the transcriptome of a species actually belongs to that species. I am thinking of contamination, mislabeled transcriptomes in the databases and that kind of things. This is not a frequent issue but it is potentially problematic, and its consequences in downstream analyses can be important. Since you already have the set of proteins for all species, it could be a good idea to make tree inferences at several phylogenetic levels to confirm the recovered topologies and branch lengths make sense. This could also help you pinpoint contaminants (or contaminated transcriptomes) from metazoan sources now that non-metazoan contaminants have been removed with BlobTools.

> Authors: The genomes and transcriptomes of MATEdb are the by-product of extensive comparative genomics studies that we are currently undertaking in the lab. All of them include orthology inference and phylogenetic inference, so we are confident that the species correspond to their given names in the public databases - at least, they fall in the trees where they belong based on our current understanding of the phylogenetic relationships of each group. We are currently preparing the publications of these studies, which will be available in the following months, and will make the link to MATEdb as the source of the assemblies used for each study.

Finally, I have to agree with one of the reviewers about the necessity of differentiating this new database from other available resources such as MolluscDB. The goal of a database is to make data easily available for the community, but the duplication of resources can actually become an obstacle and make more harm than good. What does MATEdb bring to the table that other databases do not?

> Authors: The main differences between MATEdb and MolluscaDB (or other lineage-specific databases) are as follows: (i) the datasets included in MATEdb are all high quality (i.e. high BUSCO completeness). This is key to minimize biases during downstream analyses such as orthology inference or gene repertoire evolution studies; (ii) all datasets in MATEdb have been analyzed with the exact same pipeline. This is to our belief one of the most valuable features of our resource, since different versions of the same software (e.g. Trinity) change substantially (e.g. in our experience the number of 'genes' inferred with Trinity may vary up to one order of magnitude depending on the version); (iii) MATEdb pays more attention to lineage representation than to the total number of datasets included, ie we prioritize the inclusion of the main lineages within each phyla over the number of species included; and (iv) the number of species in MATEdb is higher than current databases (e.g., MolluscaDB includes 20 and 22 species in the versions of Liu et al. 2021 and Caucel et al. 2021 , respectively). We believe this

is particularly helpful for phylogenomic studies, particularly for the selection of adequate outgroups. We have now added this information to the main text.

Apart from that, all I have are minor comments:

- You say in the introduction that the database includes new transcriptomes generated by you but there are no references to this work. How did you generate them? I see in Supp. Mat. Table 1 there is only one new transcriptome, is that correct?

> Authors: For the moment we have just generated one new transcriptome, but will add more very soon. We have now uploaded a file to the github repository with the wetlab protocol followed (Suppl. File S1 within the folder 'Protocols'), which has been mentioned in the main text.

- What happened to all the data in Supp. Mat, Table 1 without database information? Where does it come from?

> Authors: We thank the recommender for noticing this. We have now filled in this column with the corresponding information.

- The tables should have a caption. I would like to see them embedded within the tables, but to include a caption in the manuscript would be an acceptable solution. For instance in SM Table 1 I personally understand "C, S, D, F, M" mean "Complete, Single-Copy, Duplicated, Fragmented, Missing" genes in BUSCO, but you need to make the table self-explanatory.

> Authors: We have now included a caption with this information.

- I see no link to MATEdb in the manuscript. I think you could include one in the "Database availability" section, but you could also mention that such link is available through the Github repository.

> Authors: We have now included a link in the manuscript.

All in all, I think we all agree on that this is a great and exciting initiative and look forward to seeing how you address our comments.

> Authors: Thanks a lot, we really welcome all the positive criticism provided and are sure that by addressing these concerns our resource and manuscript will be definitely more solid.

Sincerely,

Samuel Abalde

Reviews

Reviewer 1

Reviewed by anonymous reviewer, 16 Aug 2022 08:23

Authors present a new database in which to house and distribute curated genomic and transcriptomic datasets of metazoans that follow a quality cutoff and strict version-controlled set of scripts for maintaining proper data cleanliness and traceability. The authors demonstrate a clear grasp of the field's current issues and their consequences for downstream data analysis. Authors highlight the drawbacks of acquiring genome annotations and transcriptome assemblies from across multiple databases and repositories. This results in genome assemblies and annotations requiring a large amount of time investment to acquire and format rather than be easily accessible as would be the case with a database such as MateDB. I can report that all the code and scripts available on github work and I am able to produce a 'mateDB ready' transcriptome assembly.

[Major Concerns]

-Authors cite other important databases such as MolluscDB but do not make attempts to provide a roadmap or features that would greatly improve the utility of MateDB. These include features that are a part of MolluscDB such as transposable elements, gene families, interactive blast, etc. While exciting, currently MateDB is a data repository that otherwise could be included with a major publication focusing on metazoan genomes. The utility of large databases is their accessibility and ability to pre-parse data for the user such that the process of doing actual analysis (rather than data mining and wrangling) can occur with ease.

> Authors: While we understand the reviewer's point of view, we consider the inclusion of these features out of scope for our current aims. MATEdb is not a stand-alone project but rather a by-product of the several projects that we are currently undertaking in the lab. As such, it is challenging to go beyond the scope of the data generated for these projects. We still believe that the data presented in MATEdb is a very valuable resource for the scientific community in its current shape. On a more positive note, the upcoming publications from our lab using the resources from MATEdb will include some of the elements suggested by the reviewer (eg, gene families, transposable elements, etc); we will make sure that the make the link clear to the data repository so the scientific community can track these features and retrieve them.

-As of current, the mateDB dataset is available on figshare. While convenient for now, once properly established, this will be inefficient and cause major headache. I would suggest moving the database to a place in which users can quickly find the data they need without having to download the entire dataset. For example, if I wanted to get all gastropod mollusc transcriptome assemblies, I would have to download the entire file (which in the future could be composed of 100s or 1000s of transcriptomes) and parse that dataset myself for what I need. Rather, it would be convenient to select my taxonomic level of interest, quickly generate a .csv of information (such that is contained in table_S1 on github) and download the assemblies. The strength of mateDB is that it contains assemblies that are treated equally using field standard methods and are curated such that no troublesome versioning issues between trimming, assembly, and processing potentially influence downstream analysis. However, if these datasets begin to become too large and intractable on the user end, then the purpose will ultimately be defeated. I recognize that without funding, recognition, or proper citation such databases provide little to advance the prospects of scientists. Such is the thankless task of database management and curation.

> Authors: We thank the reviewer for her/his tips about database management. We will take them into consideration for future releases as the dataset grows, but for the moment (and without specific resources to fund this project) this is the best we can do.

[Minor concerns/comments/suggestions]

-All genomic datasets need to have size, contigs, N50, etc. such as from the output of quast, or obtained from their original repositories.

> Authors: We have now included this information in Table S1.

-I personally would love to see an international community develop around such a database however this requires the proper channels for feedback and collaboration. I would suggest authors think on ways to develop a community around mateDB such as a forum or online splash page.

> Authors: This is an excellent suggestion. We will think of developing a channel of communication with the scientific community and will implement it in upcoming versions of MATEdb.

-A streamlined process in which researchers can submit assemblies to the mateDB database would both reduce the workload of the authors and increase the reach of the database to other research groups.

> Authors: This is an excellent idea as well. We will try to implement it in the near future.

-I would suggest that as new orthoDB datasets are produced mateDB updates the information contained in the BUSCO summary tables on the github. This can be done easily using a custom script. Perhaps also including completeness metrics for other clade specific orthoDB datasets would be useful (such as Mollusca specific BUSCO scores).

> Authors: We thank the reviewer for her/his suggestion. We will explore this option as well.

-The manuscript itself should include which orthoDB dataset was used to obtain BUSCO scores in the Transcriptome assembly processing section (I suspect due to the date accessed this is metazoan_odb10).

> Authors: We have now included this information in the manuscript.

-The figures and workflow are readable and of high quality.

[Summary]
Overall, I believe mateDB to be a valuable addition to the metazoan phylogenetics community and will provide researchers with complete datasets requiring little or no pre-processing. I am confident authors will continue to update and modify the mateDB database to provide the best resource for the Community.

> Authors: We thank the reviewer for her/his assessment and for providing very valuable tips to improve it, which we will definitely explore in the near future.

Reviewer 2
Reviewed by anonymous reviewer, 04 Aug 2022 12:13

General

The idea of a transcriptome database for metazoans is a useful initiative. This manuscript describes MATEdb, a repository for high-quality transcriptome assemblies from different animal phyla analyzed following a common analysis pipeline. The motivation and rationale behind such an undertaking, I think, have been well laid out in the manuscript. I agree with the authors, particularly on the potential of such a database enhancing the reproducibility of studies and ensuring that when studies are compared, such comparisons are not skewed by methodological differences. MATEdb is unique in being transparent with the analysis pipeline (including tools used, their versions and their command parameters). Providing a container for the complete suite of tools used is also a good idea, for both reproducibility and portability.

Hopefully, well will see more taxonomic groups represented as well. I think the authors should look at nematodes in a subsequent version. There are many genome and transcriptome data sets for non-model nematode taxa.

> Authors: We thank the reviewer for her/his positive comments. We are actually reproducing the pipeline presented in MATEdb in an extended metazoan dataset including other animal phyla, and these will be included in the near future. This includes nematodes as well.

Specific

Abstract

It does provide an adequate synopsis of the paper.

Introduction

Brief but captures what the goal is with this database as well as why and how it will be useful.

Methodology

Other than the single comment below, I think the methodology gives a detailed description of how the data was analysed.

In Figure 2, does the “Manual downloading” process under transcriptomes connect to the “fastp” process?

> Authors: As currently implemented, it does not. The rationale behind it is to check that the download is complete before proceeding with downstream analyses. However, it could be easily connected in the future.