Peer Community In Genomics

MATEdb: a new phylogenomic-driven database for Metazoa

Samuel Abalde *based on reviews by 2 anonymous reviewers*

A recommendation of:

Open Access

MATEdb, a data repository of high-quality metazoan transcriptome assemblies to accelerate phylogenomic studies

Rosa Fernandez, Vanina Tonzo, Carolina Simon Guerrero, Jesus Lozano-Fernandez, Gemma I Martinez-Redondo, Pau Balart-Garcia, Leandro Aristide, Klara Eleftheriadi, Carlos Vargas-Chavez *(2022), bioRxiv, 2022.07.18.500182, ver. 4 peer-reviewed and recommended by Peer Community in Genomics*

https://doi.org/10.1101/2022.07.18.500182

Published: 23 September 2022

Submitted: 20 July 2022, Recommended: 16 September 2022

Cite this recommendation as:
Samuel Abalde (2022) MATEdb: a new phylogenomic-driven database for Metazoa. *Peer Community in Genomics, 100022.* ***https://doi.org/10.24072/pci.genomics.100022***

Recommendation

The development (and standardization) of high-throughput sequencing techniques has revolutionized evolutionary biology, to the point that we almost see as normal fine-detail studies of genome architecture evolution (Robert et al., 2022), adaptation to new habitats (Rahi et al., 2019), or the development of key evolutionary novelties (Hilgers et al., 2018), to name three examples. One of the fields that has benefited the most is phylogenomics, i.e. the use of genome-wide data for inferring the evolutionary relationships among organisms. Dealing with such amount of data, however, has come with important analytical and computational challenges. Likewise, although the steady generation of genomic data from virtually any organism opens exciting opportunities for comparative analyses, it also creates a sort of "information fog", where it is hard to find the most appropriate and/or the higher quality data. I have personally experienced this not so long ago, when I had to spend several weeks selecting the most complete transcriptomes from several phyla, moving back and forth between the NCBI SRA repository and the relevant literature.

In an attempt to deal with this issue, some research labs have committed their time and resources to the generation of taxa- and topic-specific databases (Lathe et al., 2008), such as MolluscDB (Liu et al., 2021), focused on mollusk genomics, or EukProt (Richter et al., 2022), a protein repository representing the diversity of eukaryotes. A new database that promises to become an important resource in the near future is MATEdb (Fernández et al., 2022), a repository of high-quality genomic data from Metazoa. MATEdb has been developed from publicly available and newly generated transcriptomes and genomes, prioritizing quality over quantity. Upon download, the user has access to both raw data and the related datasets: assemblies, several

quality metrics, the set of inferred protein-coding genes, and their annotation. Although it is clear to me that this repository has been created with phylogenomic analyses in mind, I see how it could be generalized to other related problems such as analyses of gene content or evolution of specific gene families. In my opinion, the main strengths of MATEdb are threefold:

1. Rosa Fernández and her team have carefully scrutinized the genomic data available in several repositories to retrieve only the most complete transcriptomes and genomes, saving a lot of time in data mining to the user.
2. These data have been analyzed to provide both the assembly and the set of protein-coding genes, easing the computational burden that usually accompanies these pipelines. Interestingly, all the data have been analyzed with the same software and parameters, facilitating comparisons among taxa.
3. Genomic analysis can be intimidating, and even more for inexperienced users. That is particularly important when it comes to transcriptome and genome assembly because it has an effect in all downstream analyses. I believe that having access to already analyzed data softens this transition. The users can move forward on their research while they learn how to generate and analyze their data at their own pace.

On a negative note, I see two main drawbacks. First, as of today (September 16th, 2022) this database is in an early stage and it still needs to incorporate a lot of animal groups. This has been discussed during the revision process and the authors are already working on it, so it is only a matter of time until all major taxa are represented. Second, there is a scalability issue. In its current format it is not possible to select the taxa of interest and the full database has to be downloaded, which will become more and more difficult as it grows. Nonetheless, with the appropriate resources it would be easy to find a better solution. There are plenty of examples that could serve as inspiration, so I hope this does not become a big problem in the future.

Altogether, I and the researchers that participated in the revision process believe that MATEdb has the potential to become an important and valuable addition to the metazoan phylogenomics community. Personally, I wish it was available just a few months ago, it would have saved me so much time.

References

Fernández R, Tonzo V, Guerrero CS, Lozano-Fernandez J, Martínez-Redondo GI, Balart-García P, Aristide L, Eleftheriadi K, Vargas-Chávez C (2022) MATEdb, a data repository of high-quality metazoan transcriptome assemblies to accelerate phylogenomic studies. bioRxiv, 2022.07.18.500182, ver. 4 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2022.07.18.500182

Hilgers L, Hartmann S, Hofreiter M, von Rintelen T (2018) Novel Genes, Ancient Genes, and Gene Co-Option Contributed to the Genetic Basis of the Radula, a Molluscan Innovation. Molecular Biology and Evolution, 35, 1638–1652. https://doi.org/10.1093/molbev/msy052

Lathe W, Williams J, Mangan M, Karolchik, D (2008). Genomic data resources: challenges and promises. Nature Education, 1(3), 2.

Liu F, Li Y, Yu H, Zhang L, Hu J, Bao Z, Wang S (2021) MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. Nucleic Acids Research, 49, D988–D997. https://doi.org/10.1093/nar/gkaa918

Rahi ML, Mather PB, Ezaz T, Hurwood DA (2019) The Molecular Basis of Freshwater Adaptation in Prawns: Insights from Comparative Transcriptomics of Three Macrobrachium Species. Genome Biology and Evolution, 11, 1002–1018. https://doi.org/10.1093/gbe/evz045

Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, Vargas C de (2022) EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. bioRxiv, 2020.06.30.180687, ver. 5 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2020.06.30.180687

Robert NSM, Sarigol F, Zimmermann B, Meyer A, Voolstra CR, Simakov O (2022) Emergence of distinct syntenic density regimes is associated with early metazoan genomic transitions. BMC Genomics, 23, 143. https://doi.org/10.1186/s12864-022-08304-2

Reviews

Toggle reviews

Evaluation round #2

DOI or URL of the preprint: *https://doi.org/10.1101/2022.07.18.500182*

Version of the preprint: 2

Author's Reply, 14 Sep 2022

Download tracked changes file

Dear Dr. Abalde,

Thanks for your prompt review of our revised manuscript. We have corrected a couple of typos and have uploaded the final version to bioRxiv (it should be online within the next few hours). Please do let us know if you spot something else that we may have missed.

Yours sincerely, on behalf of the Metazoa Phylogenomics team,

Rosa Fernández

Decision by Samuel Abalde, 14 Sep 2022

Dear member of the Metazoa Phylogenomics Lab,

I would like to thank you for your prompt reply to mine and the reviewers' comments. I think all the comments have been correctly addressed and I find the new version of the manuscript to be of a higher quality. Nonetheless, I would proofread the manuscript to correct some typos I found. I would tell you exactly where they are, but I have seen different typos in the "track changes" and the new bioRxiv versions, so I am not sure we are reading the same manuscript version. I am talking of double dots ending a paragraph, "eg" or "e.g." used interchangeably in the text and, in the "track changes" version, mid-paragraph format changes. In any case, these are just nitpicking details that do not hamper at all the read.

Therefore, I will be happy to write a recommendation for this manuscript. Could you, please, proofread it and share the latest version with us?

Congratulations for your good work.

Sincerely,

Samuel Abalde

Evaluation round #1

DOI or URL of the preprint: *https://www.biorxiv.org/content/10.1101/2022.07.18.500182v1*

Version of the preprint: 1

Author's Reply, 01 Sep 2022

Download author's replyDownload tracked changes file

Decision by Samuel Abalde, 22 Aug 2022

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.

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.

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.

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?

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?

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

- 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.

- 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.

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.

Sincerely,

Samuel Abalde

Reviewed by anonymous reviewer, 16 Aug 2022

Download the review

Reviewed by anonymous reviewer, 04 Aug 2022

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.

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?