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

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.

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? 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.

- **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.

- **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.

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.

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.

- **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.