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

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.

-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.

[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.

-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.

-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.

-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).

-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).

-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.