General comments

I thank all the reviewers for their comments. My responses will be written in red. Apart from changes in response to the Reviewers, these are the additional enhancements made in this version of the article and the database:

- SequenceServer was updated to the latest version (3.0.1) from 2.0.0.
- Some typos were fixed.
- Minor metadata corrections were made.

Anonymous reviewer 1:

LukProt is a eukaryote-wide protein database that combines much of the data from the previously assembled databases EukProt and AniProt but is enriched in datasets especially from early diverging animal lineages such as ctenophores, sponges and cnidarians. The author provides both web-portal and command line versions of the database so those with varying degrees of bioinformatic proficiency should be able to take advantage of LukProt. The curation effort is well-done, and I believe the comparative genomics community, especially those interested in animal origins, will find LukProt to be a useful resource. I have only minor suggestions for improvement below.

For the local version of database in the file LukProt_metadata/README.txt it would be nice to have the tested versions of the dependencies necessary for the use of the associated companion scripts so the user does not have to refer back to the manuscript. It might also be helpful to include instructions on how a user might create a conda environment and install all correct versions of the necessary dependencies in case they are operating on community resources such as university clusters that are maintained by administrators and lack the necessary permissions to install software system-wide themselves. This also will guarantee performance if system wide versions of dependencies change.

Additionally, some more detail regrading the usage of these associated scripts would be useful to add to the documentation. Provide example commands, detail the exact input and expected output, when should they be run etc. No instructions can be too clear and the easier the tools are to use the more users will adopt them in their regular workflows.

Thank you for the suggestions. All of the used software should in principle be forwards compatible and one of the tools (TrimAl) is not being developed any more and will not change. However it is indeed important to provide the exact versions used. The documentation was modified accordingly, namely:

- The README file was improved and instructions to install all software were added.
- Examples of script usage are now provided, together with with the expected output.

Reviewer: Giacomo Mutti

Summary

The author expands a previously published eukaryotic proteomes database (EukProt), greatly increasing the taxon sampling in holozoans. This effort will definitely be useful for the field investigating this clade and its origin and in general for comparative genomics. However, I have some criticism especially regarding the longevity and reproducibility of the database itself and the lack of any quality control/statistics on the newly added genomes. I consider the article suitable for publication after some revisions.

Major

- I see the point the author is trying to make in line 123 but I would have expected at least a minimum analysis of the database content (or at least for the newly added proteomes), for example: number of proteins, average protein length, N50/L50 for genomes (when available). Moreover, a “comparative set” and BUSCO completeness score are listed as limitations but I consider them to be quickly and easily solvable. First, EukProt already has a “comparative set” so it would only be necessary to choose the best and most representative proteomes from the 216 new ones. Secondly, BUSCO (or other softwares such as the newly published OMArk, which also tries to assess contamination) is very quick and I do not really see why not running it. Overall, it’s very difficult this way to assess how trustable this database is, even if you assume that a single contaminated or low-quality dataset will not influence much.

Following the Reviewer’s suggestion, BUSCO was run on all the sequences and additionally on all of the included EukProt sequences, so that the results are comparable (database: eukaryotes_odb10, BUSCO version: 5.7.1, hmmsearch: 3.1). I appreciate the mention of OMArk, which I was not aware of and 2 Reviewers mentioned. For a proper analysis, OMArk needs a single protein for each gene, which I cannot generate in a feasible amount of time for all of the EukProt and LukProt datasets. Still, the OMArk analysis was performed, with taxids included, but the results should be treated with caution, as not all the proteomes have 1 protein per gene. As a proxy measure of proteome quality without assigning isoforms to genes I added an OMArk “Single + Duplicated” measure to OMArk results, which should be roughly equivalent to BUSCO Complete score. All the statistics from both programs are now included in the metadata spreadsheet and the output of these programs is included in the Zenodo repository. In addition, more dataset statistics were calculated using SeqKit and are available in the metadata: number of sequences, sum of sequence lengths, minimum sequence length, average sequence length and maximum sequence length.

As the dataset quality analyses have now been done, selecting a comparative dataset is indeed easier. However, I would prefer to have some feedback before deciding upon specific representatives. The metadata now contain 2 columns (in_TCS, in_LCS) where it is indicated if a dataset is in the EukProt TCS (The Comparative Set) and whether it is proposed for its extension, a LukProt Comparative Set (LCS). The proposed (beta version) of LCS consists of 233 datasets (18% of LukProt), compared to 196 of LCS (19% of EukProt v3). The species were chosen based on the taxonomic breadth and BUSCO/OMArk quality scores. The LCS is also avalable to search using the BLAST server.

- My second main point is on the Huntingtin example analysis. Firstly, no criterias are listed for manual curation of the outliers which can heavily bias the final tree topology. Further, when clustering it is not clear if the representative tip of each cluster was annotated with the taxonomic latest common ancestor of the cluster or not. In my opinion this should be done when clustering sequences. Further, the author tests many different trimming parameters but no discussion on how variable the tree topologies of the fast trees are. Is this parameter that much important? Overall, the analysis seems a bit unnecessarily complicated and I think that they make the example much less powerful than what it could actually be. I consider this database a good effort, especially how the author took care in homogenizing with EukProt and granting this would be kept in newer versions.

I agree with the Reviewer that the associated explanation makes the example less powerful. It was chosen to show that inference of deep phylogeny can be achieved using fast methods, although if one already knows the answer they are looking for, it is indeed less convincing. I have provided a different example, based on concatenated BUSCOs, with a simpler protocol.

The new procedure is now explained in the Methods section. In brief, the number of BUSCOs in each species out of the 255 scanned was assessed. The BUSCOs were ranked according to how many species contained a single given BUSCO. The sequences of top 20 BUSCOs from each species were extracted. The species which had fewer than 10 of the 20 BUSCOs were discarded. BUSCO sequences from the remaining species were concatenated and aligned using FAMSA, trimmed using TrimAl (50% gap threshold) and then phylogenies were made. This avoids the somewhat arbitrary choice of Huntingtin as an example and the manual outlier pruning and might represent a more “raw” view of the data one may expect to find in LukProt.

- However, as too often in bioinformatics, this resource may be quickly discontinued as, currently, it’s a single individual’s effort. Further, there is no code repository to try to reproduce/update what has been done. The scripts in the zenodo folder are just utilities to parse/analyse the database. I think it would be ideal to share a version controlled repository with scripts/documentation to try to solve this.

Lack of updates is of course a valid concern. I am able and plan to maintain the database indefinitely, even without any targeted funding. However, it is prudent to have a backup plan. Once the database is citable and there is some community involvement, I will start actively searching for people interested in taking over, if I am ever unable to continue. I can also consider renaming the database if the current name is ever found to be inappropriate.

Regarding a repository to reproduce what was done: I have documented the analyses step by step with all the terminal commands recorded. To be completely honest, I have no formal training in computer science and I do not feel comfortable enough with GitHub CLI to put together a repository and release them at this point. If the Reviewer allows, I will release a repository to reproduce all analyses for the next major version of LukProt, as this would be a major effort and cause a long delay. I believe that, with the additional BUSCO and OMArk statistics now added, the database can be used as is and most users would consult these statistics. At the moment, if there are any further questions about the analyses done, I can be contacted by email. A line about this was added in the “Data, scripts, code, and supplementary information availability” section.

Minor

Main text

line 37: phrased like this it almost seems that the debate was on Eukaryotes, not Metazoa
Thank you, indeed, the phrasing is awkward here - this has been corrected.
line 39: the much of -> much/most of
Corrected.
line 30: for not only for -> not only for
Corrected.
line 100: An R package -> The R package
Corrected.
line 146: followin -> following
Corrected.
line 150: misspelling in “databate metadata”
Corrected.
Fig2: caption does not explain A and B
Corrected, the figure was replaced as well.

Metadata

- sometimes “MMETSP – be very careful” other times only “MMETSP”

The “MMETSP” caution was present only in the datasets not included in the current version (1.5.1) but the previous one (1.4.1). The caution is now the same.

- In the metadata csv file: Chlorochromonas danica notes: “also EP01039 – why are there 2?” Indeed, why?

They are different strains of the same species but their name is not differentiated in EukProt. I have contacted a EukProt maintainer and he told me this is an error on their part and EP00981 will be removed. This will be reflected in LukProt after EukProt v4 is released. I have noticed some data regarding strains was not up to date with EukProt v3, for example names of these 2 strains. This has been corrected.

- In general, and I consider this is something that also EukProt is lacking, it would be ideal to also link to downloadable genomes and gffs when the proteome comes from a genome. If the author is interested he may feel free to contact me as I’ve been collecting these URLs for EukProt.

I am interested in this, if it is indeed appropriate, I can include these data, but I am not sure if Reviewers are allowed to provide data for the article they are reviewing. It would probably be better if the Reviewer could contact EukProt maintainers to include it. It would be then added automatically to the next versions of LukProt.
Following this suggestion, links to genome sequence and/or annotation files were added to the LPXXXXX datasets that were not downloaded from NCBI, UniProt or AniProt. In case of these, the relevant files can be easily found by users using the sequence IDs.

Finally, just a little note that MateDB v2 was recently released (https://doi.org/10.1101/2024.02.21.581367) greatly expanding the number of proteomes which might be useful to the author for future releases of LukProt.
I am aware of this and considered including some of the datasets. MateDB seems to be focused on Protostomia, of which there are plenty species in the database already. If the Reviewer can suggest some important taxa that were omitted, they can be added to LukProt. They would then have the prefix “MPXXXXX”. I want to be careful about including many more species from Protostomia, as they could make the database too skewed towards them, not least due to of the number of known/sequenced species from this clade (especially Arthropoda).

Anonymous reviewer 2

> The manuscript of Sobala provides a new pervasive and curated eukaryotic database. It gathers information from EukProt and many other resources to increase the Metazoa sampling in released protein databases. I consider the paper interesting for the phylogenomics and comparative genomics communities, and I thank the author for their work. The analysis and data management are suitable. However, I have found some weaknesses that would make the database user distort the interpretation of the given protein sets. I would recommend this paper for publication after some corrections.
>
> **Overview and general comments**
> The database construction is rigorous as it considers different sources of information, homogenises the IDs, and distributes it using standard formats. The metadata incorporated to the database is easy to read and parse, as well as it is completely integrated with EukProt, maximising the compatibility between both databases. The server that the author provides is

> accessible, and the taxonomic structure implemented in the server helps the user to perform clade-specific analyses.

> I miss an analysis of the contamination and quality of each genome. Providing this information to the users would let them better choose the genomes for downstream analyses. Moreover, this point is essential for lateral gene transfer analyses.

Following suggestions from two Reviewers, BUSCO and OMArk analyses are now included. I agree that they should have been there in the first place.

> The taxonomy accounts for up-to-date literature, and they performed a readable table, which I personally think is a sound synthesis work and thank. It enforces the deep branches of the events the author considers important for the Metazoan researchers. The author may improve the documentation for the taxonomic tree in the supplementary file. The structure of the taxonomic groups and how they cluster is not clear to me. An indentation table structure with the literature would be more understandable as you would easily identify shallow and deep groups.

I agree that the supplementary information was not very clear. I have rearranged this section and added indentations, which are necessarily shallow (one space character) due to the many levels of nesting.

> Regarding the example analysis, I found the first paragraph in the results more suitable to be written in methods, as the huntingtin example method's section seems incomplete and unclear. I could not understand why the author included two inference software and which criteria they used for removing branches from the preliminary trees. I emphasise this in the detailed comments of my review.

Because there were reservations about the example from 2 reviewers, it was exchanged for a more rigorous one, based on concatenated, abundant single BUSCOs. All the details about it can be found in the Methods sections, some information can be found in reply to Reviewer Giacomo Mutti.

> I finally consider that the assessment of limitations is fair, although I think that the author should add a contamination assessment to the database. BUSCO completeness and contamination values (for instance calculated with OMArk) would be valuable and would make the database even more complete.

I do agree. Both BUSCO and OMArk results for all datasets were added (see above).

> **Detailed comments on the methods**
> The methods are proper and suitable for the kind of data that Sobala used. However, I see some weaknesses that they should address:
>
> - **Dataset naming**: the renaming sounds good. Although the author incorporates the sequence identifier of the original source after the protein ID (L81), a file connecting the protein IDs of LukProt, EukProt, AniProtDB, would be helpful for the community in the cases they change or just for comparison purposes. Despite this, the files are accessible and well-documented in the Zenodo repository.

A tabulated file connecting the differing IDs between AniProtDB, EukProt and LukProt is now provided in the repository.

- **Distribution of the database**: it may be helpful to separate the BLAST databases by the taxonomic depth of the database in separate compressed files, as Zenodo allows a folder structure. I propose the author to share the folder structure with the compressed version of the database rather than the complete set of databases in a single compressed folder if they consider it appropriate. It would be easier to use and download.

I have tried to apply this but Zenodo is not permitting what the Reviewer suggests here. Zenodo does not allow to upload folders (see https://support.zenodo.org/help/en-gb/1-upload-deposit/74-can-i-upload-folders-directories). When LukProt is uploaded as a .zip file to allow Zenodo to show its structure, the viewer does not show the full structure anyway because of the number of files in the archive. Individual files from the zip cannot be downloaded either. I do not think therefore that the format needs any changes – only if the full structure could be visible, I would change it to .zip.

- **Data processing**: Sobala uses two different software (Trinity and TransAbyss) and multiple versions of the Trinity software for assembling transcriptomes. Moreover, they also use different versions of the software for protein prediction (TransDecoder). However, they do not explain why the author chooses one or the other and the criteria substantiating the software election. The author should include this information in the manuscript. Regarding the clustering, I see the same as previously commented. The author does not specify why and when proteomes are clustered (“in most cases”, L96). Moreover, for the strain “pangenome”, they do not determine whether the CD-HIT parameters are or are not maintained.

TransAbyss was used in a single case where Trinity was failing due to a bug in the software; Trinity was always preferred. This is now explained. As for the TransDecoder, the reason for using different versions is that it was updated after performing part of the analyses – the database was prepared over multiple years. The genome assembly “state of the art” – transXpress (https://github.com/transXpress/transXpress-nextflow) – was tried but did not work properly and the second best option, Trinity, was used.
I believe that for reproducibility it is enough to list the version that was used, but if the Reviewer considers it necessary, older analyses can be re-run for a new LukProt version. For now, a line explaining software versions was added in the Methods section and the metadata indicate the versions used, if they were changing.

- **Huntingtin example**: as I previously said, the methods section of this example is incomplete. The author should describe better how they obtained the phylogeny with a clear description of each step. They should move the first paragraph of the results section to the methods section with a few changes and clarifications: 1) the procedural scheme is diffuse, I had to read twice to understand the steps they followed; 2) the changes in the CD-HIT clustering identity parameter are not justified; 3) I miss a definition for “outlier” (L192), as some sequences have been removed, I consider necessary to explain which criterion has been used to remove them.

I agree that the example could be better. Using the performed BUSCO analyses I have switched it for another one. The procedure is detailed in the Methods but in brief it is thus: the number of BUSCOs in each species out of the 255 sought was assessed. The BUSCOs were ranked according to how many species contained a single given BUSCO. The sequences of top 20 BUSCOs from each species were extracted. The species which had 9 or fewer of the 20 BUSCOs were discarded. BUSCO sequences of the remaining species were concatenated and aligned using FAMSA, then trimmed using TrimAl and then phylogenies were made. This avoids the somewhat arbitrary choice of Huntingtin as an example and the manual outlier pruning.

> **General detailed comments**
> L58: the author should define AniProtDB, Animal Proteome DataBase (AniProtDB).

It is now defined.

> L58,64,134,138: AniProtDB appears written differently. The author should homogenise them to the database name in the reference "AniProtDB".

Thank you for pointing this out. The correct name should be "AniProtDB" (corrected everywhere).

> L127-128: HGT analyses are sensitive to contamination, as they consider that a given protein originated through HGT when it is more similar to a distant taxon than to a close one. For this reason, although I agree with the author that phylogeny will help "drawing conclusions", I consider that this argument does not apply to HGT. In my opinion, they should at least release a bona fide list of proteins for each organism and a bona fide sister database.

> FURTHER COMMENT CLARIFICATION:

> A bona fide sister database is the database you use to detect the origin of the HGT.

> Imagine that we are interested in detecting HGT in a worm from bacteria. The worm would be the organism of interest, and a broad database with many bacteria would be the bona fide sister database. You would search the worm proteins in the bacterial database, then reconstruct the trees and, finally, assess whether the sister sequences to our worm sequence (the sequences next to the worm sequence in the phylogeny) are bacterial. In such a case, you would be in front of a case of HGT from bacteria to the worm. If the sister database is incomplete in terms of diversity, you will not find the donor. Moreover, imagine that the worm's genome is contaminated with bacterial sequences of the gut. Without a contamination assessment of the proteins in the genome, you would be overestimating the number of HGT from bacteria to the worm. My concern was more about the latter, which is why I asked the author to assess contamination and change this sentence, as the genome needs to be as clean as possible to determine bona fide HGT analyses.

> The sister database is a requirement for the HGT analysis, but here, it is not that important because this is not a paper about HGT in a clade. I am sorry that this secondary database has generated some confusion.

Thank you for the comment and clarifications. I have now provided the BUSCO and OMArk analyses, which should be indicative of contamination, at least partially. Regarding the bona fide sister database, the downstream users should provide one if they would like to do the HGT analyses. I would argue, however, that even without it, some phylogenies can be suggestive of gene transfers. An example of this is Supplementary Figure 1 to the article "Evolution and phylogenetic distribution of endo-α-mannosidase", Sobala, Ł. F. *Glycobiology* 33, 687–699 (2023), where a previous version of LukProt was used. There, some GH99 sequences from Rhizaria cluster together with other rhizarians and some with chlorophytes. The latter are from the photosynthetic endosymbionts of Rhizaria. The Reviewer might also take a look at the sequences from Chatonellales within Medusozoa, coming from multiple species and suggesting transfer from an ancestral “placnidian” to an ancestral species of Chatonellales. A LukProt subset “Placnidia” could serve as a bona fide sister database here.

> L133: I suppose that a dataset is the whole protein set for a species, but it is not defined. It would be helpful.

It is now explained.

> L134: a brief comment on the sources of the newly added datasets would be necessary, at least commenting that they have been collected from repositories of numerous studies and the source is available in the metadata table.

I have added a few sequences explaining the sources more clearly, thank you for the suggestion.

> L152: the figure 1 colours need to be more different to be easily distinguished. Most of them are similar colours (Apusomonadida, Streptophyta, Breviatea…).

As much as I would like to comply with this suggestion, the number of colors distinguishable by eye is quite limited and further limited by the necessity of choosing colors that are dark enough to be visible when printed out. This means that some colors will be very similar. The chosen scheme takes into account the fact that proteins from less closely related groups are not likely to be found clustering together. Indeed, no color RGB hex code is duplicated in the scheme.

Regarding Apusomonadida and Breviatea, when printed out, these colors do look quite different; this also depends on the monitor used. There is a trade-off when choosing colors for neighboring (on the tree) or closely related groups. On one hand, using colors that are too similar makes them difficult to distinguish. On the other hand, using colors that are too different “takes them away” from other groups. The chosen color scheme traverses the color space just twice. Some colors were intentionally chosen for intuitive understanding, for example shades of green for Archaeplastida and shades of red/yellow for vertebrates (from blood). Streptophyta are unlikely to cluster with Breviatea, so this similarity should not interfere with intuitive phylogeny inspections.

> L152: the table 1 title starts with an uppercase letter, while the figure 1 caption starts in lowercase; it should be uppercase.

Corrected.

> L202: the figure 2 caption lacks the caption for the panel B, it should be added. Moreover, I suggest to increase the size of the panel B and put the label “A” on top. There are some groups which do not match the topology (Ambulacraria placement, Chordata…), I would understand the size when both topologies match, but I do not see the point here.

Figure 2 was reworked with a different example.

> L214-216: sequence similarity networks gain insights into deeper relationships and higher detection of far homologs. However, the whole analysis is trying to remove homologs by clustering and manual curation, this sentence may confuse the reader.

I agree that the sentence was confusing. A new paragraph was added to explain what was meant.