Reply to the recommender
Round #1
PCI Genomics
29 August, 2022

EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes
Daniel J. Richter, Cédric Berney, Jürgen F. H. Strassert, Yu-Ping Poh, Emily K. Herman, Sergio A. Muñoz-Gómez, Jeremy G. Wideman, Fabien Burki, Colomban de Vargas
https://doi.org/10.1101/2020.06.30.180687

Our responses are indented and in and in blue text.

Response to the recommender:

Two reviewers have completed their assessments and have determined that the manuscript is sound and describes a useful resource for the field. They have requested only minor revisions.

I share their enthusiasm for this resource and look forward to seeing the next draft of this manuscript.

> We thank the recommender and the reviewers for their time in assessing our paper, and for their constructive and positive comments on our work.

I concur with reviewer #2 that more detailed discussion regarding how their resource differs from PhyloDB is needed. I also did not find Table 1 very informative, and I would think a supplementary table listing the actual URLs used would be more relevant to interested readers. But since the reviewers did not take issue with the table, I will not require it to be changed, and leave the decision to the authors' preference.

> We added a more detailed discussion of how our resource differs from PhyloDB to our manuscript (please see our response to Reviewer #2 below).
>
> We included Table 1 because we believe it allows the reader to perceive easily that most of the datasets are not available in GenBank, and secondarily that these datasets are scattered across a number of websites, some of which are now inactive/inaccessible. In order to convey this in a more informative way, we have converted the table to a new figure, Figure 2. Also, we would like to note that the actual URLs for each dataset are listed in the metadata files in FigShare (we have added this information to the legend).

Given the small changes requested, the authors should aim to format their manuscript for according to PCI Genomics guidelines (see: https://genomics.peercommunityin.org/help/guide_for_authors#h_32731137856716197052348 47)

Formatting issues that I noticed are

- Table and Figure should be embedded within the text.

- An email for the corresponding author should be indicated

- Rather than a database availability statement, this should be moved to the end of the abstract (for the link to the database webserver), and also mentioned in the “Data, script and code availability” section at the end of the manuscript.

- I believe moving all of the descriptions of the database to a “Results and Discussion” heading would be most appropriate for this article type (and the current headings, such as “The EukProt Database” changed to sub-headings). Based on the formatting guidelines, PCI Genomics strongly recommends separate Results and Discussion headers, but I think a combined section would be acceptable in this case, as the manuscript is very clear as it is.

- The methods section should be moved before the Results section

- Please re-format the acknowledgements section to match the recommended format. Also, is the lower-case “i” in Núria Ros i Rocher a typo? I think this is supposed to be a hyphen.

- Please add a Data, script, and code availability section at the end. Note that the authors’ custom code must also be made available in this section.

- Make in-text citations square brackets when they are within parenthetic phrases (e.g., “Eisen, 2003” at L48.

> Thank you for bringing these formatting issues to our attention. We made all of the changes suggested above.
>
> We also increased the font sizes in Figure 1, to make them more readable in portrait orientation (the figure was originally in landscape orientation).

Response to anonymous reviewer, 11 Jul 2022 19:50:

The present manuscript presents the protein based database EukProt that has been build on reference data from genome, single cells and transcriptomes. This update aligns with the FAIR principles and introduces a new high quality reference dataset that was explicitly setup for comparative genomics and that tries to meet a high taxonomic standard that aligns with UniEuk.

The manuscript is well justified and clear in its description and outlines. Thus, the only critique that I have that it missed to point the limitations of EukProt in a specific manner. For future

users, however, the limitations are as important as the strenghts, in particular, when used as reference for the whole scientific community.
Therefore, I'd like to recommend to add a small paragraph that points out the limitation of the database. This could for instance highlight cases, in which the database will be of only limited use (e.g. a list of lineages that are not well covered (to balance the statement of the lineages that had more than 100 taxa) could be pointed out here and that still require joined sequencing efforts; similarily this could be pointed out for the comparative genomics), or limitations of the current gene prediction models, taxonomic paths, ... (not all maybe neccessary though, but I'd least mention/discuss the most important ones for the users)

Thank you

> We thank the reviewer for their positive comments on our work. We agree that a summary of the limitations of the database would be very helpful to the reader. We added a new section to the Results and Discussion that describes what we believe to be the current limitations of the database, as follows:
>
> 1. “As can be observed in Figure 1, representation of species is highly uneven among taxonomic groups. This is largely due to the difficulty in identifying, cultivating and sequencing species from underrepresented groups; however, some representatives of key taxa have been published but are not publicly accessible (a list of these species can be found in the “not included” metadata table of our FigShare distribution).
> 2. We currently lack an automated, consistent method to identify and remove contaminated data sets or individual contaminant sequences. Currently, contaminated data sets are identified by end users and removed in subsequent versions.
> 3. Most species are not represented by genomes, which are the gold standard for providing gene catalogs. Instead, due to technical limitations in cultivation and/or sequencing methods, they are represented by transcriptomes from laboratory cultures or by single-cell genomes or transcriptomes.
> 4. Proteins smaller than 50 amino acids are not predicted with the settings we used for protein prediction (but may be included in data sets for which we did not perform protein prediction). We used the default settings for protein prediction in TransDecoder (see Methods), in order to compromise between minimum protein length and total number of predicted proteins (as proteomes predicted with protein sizes smaller than 50 amino acids are generally much larger, which may significantly slow many downstream analyses). EukProt users interested in proteins shorter than 50 amino acids (for the species on which we performed protein predictions) would instead have to repeat protein predictions using their desired settings.”

Response to anonymous reviewer, 13 Jul 2022 13:31:

This article presents the release of EukProt: a database of eukaryotic genome-scale predicted proteins. The manuscript nicely outlines the pitfalls in shared genomics data accessibility and

presents EukProt as a solution for several challenges of comparative genomics analyses, which will become even stronger with the exponential increase in genomic data production. It then continues by describing the database utilities, downloadables, generic structure, abidance to FAIR principles, and community-provided update possibilities and finishes with a detailed description of the methodology.

The title and abstract are clear and straight to the point. Overall the article excellently stands out for its clarity, detailed methodology, input database specifications, comprehensiveness, and range of bioinformatics challenges that the authors address with the development of this resource. The amount of considered repositories from which the database is constructed is impressive, and so is the subsequent integration of custom processed raw data (assemblies, annotations). The authors have a clear and deep knowledge of the comparative genomics issues that the scientific community is facing and provide an elegant solution through a genomic analysis framework enriched with some of the most solid and state-of-art comparative genomics tools (examples: the UniEuk taxonomic framework, BUSCO completeness scores). It indicates particular sensitivity and integrity of the authors toward a modern (e.g. foreseeing the ocean metagenomics data integration) and virtuous (e.g. providing various downloadables such as genome annotations) way of approaching bioinformatics resource development. This sensitivity is mostly exemplified by the presentation of The Comparative Set (TCS), a selection of taxonomically fairly-distributed, highly complete predicted-protein sets, which will hopefully serve as a basis for many comparative genomics analyses in future eukaryotic biology studies.

This reviewer will only provide a few minor comments about the clarity of some sentences, as well as mention the fact that they cannot technically evaluate the tools and parameters selection for the de novo transcriptome assembly paragraphs (lines 300-306) and the automated genome annotation (lines 329-338). This reviewer particularly praises the care given to the methods producing The Comparative Set.

> We thank the reviewer for their positive and encouraging comments on our work.
>
> In order to further emphasize the value of The Comparative Set and its potential uses, we have produced a new figure, Figure 3, that displays the taxonomic relationships among the 196 species in the set, in the form of a tree that uses the same color scheme as Figure 1.

This reviewer would be happy to see this resource further expand and recommends it for PCI Genomics validation.

Minor comments:

Lines 68-70: The authors could better explain how EukProt differentiates from PhyloDB.

> We agree. To address this comment, we added a more detailed discussion of how our resource differs from PhyloDB, as follows:

> “We note that an existing database of genome-scale protein data sets, PhyloDB (https://github.com/allenlab/PhyloDB), contains 550 eukaryotic species with at least 500 proteins, as of version 1.076. PhyloDB was most recently updated in 2015; EukProt includes data made available since then, allowing it to provide more species with generally higher completeness and representing a greater phylogenetic breadth. In addition, we placed each species within a universal eukaryotic taxonomic framework, UniEuk (Berney et al., 2017), in order to ensure that the evolutionary relationships among data sets are accurately and consistently described.”

Lines 73-75: This reviewer could not find protein data files comprising protein domains, Interpro, or gene ontologies from the downloadables (genome annotations, protein fasta files). Not clear if they are provided or if they are mentioned as an example of data with difficult accessibility. Either way, it could be better explained. EDIT: found the mention of potential addition in the future at lines 205-208, this reviewer would still advise rephrasing lines 73-75 for immediate clarity.

> We agree with the reviewer’s comment. We rephrased this part of the text, and added an explanation of our proposed strategy that, in order to encourage flexibility, annotations can be added to the GitHub in between EukProt releases (for example, in the case of protein domains, at the time we released v3 of EukProt, the latest release of InterProScan had not yet integrated the most recent version of Pfam). We hope that the community will also contribute annotations to be disseminated via our GitHub. The two modified sections of the text are below.
>
> Introduction:
>
> “In addition, because the large majority of the protein data sets from diverse eukaryotes are not included in major public databases (see Figure 2), researchers cannot easily access them via standard tools such as NCBI BLAST (Sayers et al., 2020) in order to find the homologs of a protein of interest, nor are they automatically integrated into major public databases containing annotations such as protein domains (e.g., Pfam [El-Gebali et al., 2019], Interpro [Mitchell et al., 2019]) or gene ontology (The Gene Ontology Consortium, 2019).”
>
> Growing the EukProt database with community involvement:
>
> “Additionally, we propose to use our GitHub as a flexible repository to disseminate community-contributed annotations on the level of individual protein sequences, such as protein domains from Pfam (El-Gebali et al., 2019)/Interpro (Mitchell et al., 2019), gene ontology (The Gene Ontology Consortium, 2019)/eggNOG (Huerta-Cepas et al., 2019), which may be updated in between EukProt releases.”