Reply to the recommender
Round #2
PCI Genomics
9 September, 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:

I think your changes have addressed most of the reviewer's comments (except for one minor comment – see below) and I think the manuscript is in excellent condition, and requires only a small tweak prior to recommendation.

One important thing to note is that I received a "timed out" error when trying to load http://evocellbio.com/eukprot/ - I'm guessing this was just a transient problem, but should be checked.

> Yes, this appears to be a transient problem that we were experiencing with the Amazon Web Server (AWS) that hosts our web site. In order to avoid this problem in the future, we increased the AWS capacity, and we are currently migrating to a different AWS provider that we expect will provide more reliable hosting in the future.

The minor comment that I think the authors perhaps missed was this partial statement from reviewer 1:
"…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)"

Those line numbers no longer match, but the sections correspond to the paragraphs starting with "'assemble mRNA': de novotranscriptome assembly " and "'predict genes': we used EukMetaSanity", respectively. I think either a little more explanation of why these parameters were chosen (e.g., why stating why using the same parameters as Alexander et al. 2021 makes sense, in the case of the predict genes). If the options are somewhat arbitrary, which might be the case with the assembly and filtering options, then the authors could mention that these options were not evaluated but are similar to what are commonly used, which I believe would address the reviewer's point.

The recommender is correct, we did indeed miss this statement from Reviewer #1. We apologize for any inconvenience this may have caused. To respond to this comment, we updated the corresponding sections to reflect the fact that default parameters were always used (as those are most likely to correspond to the options recommended by the authors of the software), and, in the absence of default parameter values, options were chosen following publications working with data similar to ours, as follows (new text is in bold):

"All software parameter values were default (unless otherwise specified below or in the metadata record for a given data set), **as default parameter values are most likely to correspond to the options recommended for general use by the authors of the software. Due to the large volume of data sets we processed, and variability among them, we were not able to test parameter values beyond those specified below**."

"'assemble mRNA': de novo transcriptome assembly using Trinity v. 2.8.4, http://trinityrnaseq.github.io/ (Haas et al., 2013). We trimmed Illumina input reads for adapters and sequence quality using the built-in '--trimmomatic' option **(whose default trimming settings are based on optimal trimming parameters from [MacManes, 2014])**. We trimmed 454 input reads prior to running Trinity with Trimmomatic v. 0.3.9, http://www.usadellab.org/cms/?page=trimmomatic (Bolger et al., 2014) with the directives 'ILLUMINACLIP:[454 adapters FASTA file]:2:30:10 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25' **(corresponding to the default Trinity trimming parameters, but for single-end reads)**. When the sequence library was described as stranded in the NCBI Sequence Read Archive, we used the corresponding 'SS_lib_type' option."

"'predict genes': we used EukMetaSanity https://github.com/cjneely10/EukMetaSanity (Neely et al., 2021) to perform automated annotation of genome sequences lacking publicly available protein predictions. We used the following parameters, as specified in (Alexander et al., 2021): --min_contig 500 --min_contig_in_predict 500 --max_contig 100000000. **These settings were designed for generalized gene prediction on genomes expected to be sampled from across eukaryotic diversity, which is also the case for the data sets in EukProt.** We used the parameter --min_contig_in_predict 200 (as it matched the default minimum contig length in Trinity). By default, we selected the proteins at Tier 2 (predictions supported by at least 2 sources). If Tier 2 produced fewer than 15,000 predicted proteins, we instead selected Tier 1. All other parameter values were left at their defaults. We did not perform gene prediction on unannotated genomes for which a transcriptome was already available for the same species (under the assumption that the gene predictions of the transcriptome would be of higher quality, due to potential errors in the gene annotation process)."

Last, I recommend two very minor changes:

In your title, I recommend that you change "a database" after the "Eukprot:" to be "A database". I believe that most style guides suggest the latter format, but the former is widespread in the scientific literature so I leave that choice to you.

> We capitalized the A after the title, as suggested.

I do however strongly think that the link to the webserver should be added to the abstract, which I think many readers comes to expect when reading about bioinformatics resource.

> We added a link to the web server to the abstract.

Once these last changes are addressed I would be pleased to recommend your article.

> We appreciate the recommender’s time and constructive criticism, and look forward to the possible recommendation of our article.