Review of: "CRSP: Comparative RNA-seq pipeline for species lacking both of sequenced genomes and reference transcripts"

Pedro Prudêncio[1], Pedro Barbosa

1 University of Lisbon

Potential competing interests: The author(s) declared that no potential competing interests exist.

Authors present CRSP, a pipeline to facilitate the analysis of transcriptomes on species that lack reference genome or annotated transcripts. Despite the tremendous progress witnessed in the last couple of years, we are still quite far from having a proper catalog of genetic information representing the full spectrum of biological diversity we have on earth. Particularly, there are genomes that are too big, repetitive, or highly polyploid, making the research cost-prohibitive for many labs. The alternative is to sequence the RNA, which can be sufficient to get biological insights on some properties that one may be interested in. Hence, the development of a pipeline such as the one presented here, is of sufficient scientific interest to be published.

However, we do have some concerns on the pipeline itself, as well as on the results.

1 - The pipeline requires the user to manually configure the environment (e.g. install dependencies) and run each step separately. Wrapping all of this in a single executable (or a nextflow/snakemake pipeline) would be of much greater value, and it is something that is not hard to do.

2- What are the additional contributions that CRSP adds in comparison to the analysis performed in a previous paper (Jiang P, 2017) mentioned in the manuscript ? Couldn't we just replicate the steps described in section 2.3 of that paper?

3 - The user needs to run the assembly (e.g. trinity) before running CRSP. Would it be feasible to allow CRSP to run from fastq data and actually do the assemblies for the sample schemes the user defines? It would then be an end-to-end analysis pipeline, thus increasing the value of the tool.

4 - Were additional abundance estimators, such as Salmon or Kallisto, tested? It could be that for different species (or even assembly qualities), read assignment from these tools could be more effective. Although It seems RSEM results are consistent with Kallisto and Salmon, these last two are significantly faster (https://www.nature.com/articles/s41598-019-44499-3)

5 - It is not clear in the paper how the redundancy after merging different assemblies is conducted. Given that we expect it

to be an important step of CRSP, we would like to know more on that. we may have missed, but doing a quick search on the source code we did not find anything.

6 - Evaluations were only performed on H1 cell line. It is indeed useful to compare between different data types as it was done (single-end vs paired-end), but in this paper we would find more important to benchmark different species. We understand that CRSP is highly dependent on the assembly quality, and there are a lot of challenges to overcome, but we would like to see results in species with different genome architecture, such as high vs low GC content. The same way authors picked from standard RSEM counts to human genes, gene expression matrixes for other species are available to do so.

Some minor details:

- Formatting issues in figure 1 (Very small x/y axis values in 1B and 1C; Different x axis label size in 1B and 1C)
- Typo in section 3 (results): “we also generated we 80 million”
- Very confusing last phrase in the manuscript: “Together, our study suggest CRSP is a useful tool making RNA-seq as a general tool to interrogate transcriptomic dynamics but not only limited to species with sequenced genomes or known transcripts.”

Given the comments above, we think additional work is necessary to make CRSP a “sellable” tool. Better benchmarks are required as well.