Answers to Reviewers

We thank the reviewers for their comments and suggestions. Below we divided the document by each section of the manuscript, including the comments by the reviewer and our answers.

Abstract

Reviewer:
“This is distinctly true for non-model organisms, where no genome information is available; yet, studies of differential gene expression, DNA enrichment baits design, and phylogenetics can all be accomplished with the data gathered at the transcrip- tomic level.”

Please split this sentence or improve it. It is a bit too large.

> Sentence was modified

Introduction

Reviewers:
“Moreover, factors such as reads length and number have to be taken into account for the assembly of a reference transcriptome (Grabherr et al., 2011; Schulz et al., 2012; Francis et al., 2013).”

This is a bit out of context. Maybe improve the sentence or move the sentence before?

> Sentence was modified

Methods

Reviewer:
“Next, TransPi uses **this** non-redundant reference transcriptome to run several

downstream analyses commonly applied to de novo transcriptomes projects:”

I think "this" is not correct here cause it belongs to another paragraph and left the word out of context

Fixed

Reviewer:
Have to wonder about read normalization before assembly. The algorithm of some assemblers is not well suited for normalized libraries since they rely on difference in edge depth to differentiate between isoforms and/or paralogous genes.

TransPi does a read normalization step. By default this is done when running the tool, however, we provide the option to skip the normalization step. Also, the user can select the values for minimum and maximum coverage per read for this step. Default values for these are 1 and 100, respectively. It has been pointed out that 30 is more than enough for the maximum coverage for reads during the normalization step (Haas et al., 2014). Since an assembly can vary widely depending on organism and datasets we let the users decide if they want to do normalization or not.

A case example: we were working with a coral sample and we wanted to create a reference transcriptome. After concatenation of the multiple datasets from the coral (i.e. same experiment, treatment, etc) it produced files over 400M of reads. Thus normalizing before assembly reduced drastically the files, thus reducing time and resources used for the assembly but without losing information.

Reference

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, *8*(8), 1494-1512.

Reviewer:
Benchmarking was performed on *C. elegans*, *D. melanogaster* and *M. musculus* based solely on BUSCO scores. No mention of assembly accuracy is done, despite the availability of well curated gene models and isoforms for all three model organisms.

We relied on the model species *C. elegans*, *D. melanogaster* and *M. musculus* for the different tests we did such as k-mer sizes, read length and read quantity. This was due to the fact that multiple libraries per each species are easily findable for performing such experiments and with multiple samples per experiment. A test against the gene models will, in fact, be very pertinent here. See Results section below for a test including the gene models.

Reviewer:
rnaSPADES uses different k-mer sizes for a single assembly. Did the authors use a single k-mer size with this assembler as they did with other traditional deBruijn assemblers like Trinity, Oases or transAbyss?

The k-mer list was used to run rnaSPADES with each k-mer individually to then concatenate all the assemblies of different k-mers together.

Results

Reviewer:
The results are essentially centered around BUSCO scores and comparisons between the different datasets and to Trinity as the gold standard of RNAseq assembly

Missed opportunity: Measure the percentage of chimeric transcripts in Trinity and TransPi by comparing the model organisms assemblies to their gold standard annotations. Not only is it important to understand how a new tool outperforms older tools in terms of BUSCO completeness, but also if the transcript accuracy is higher or lower using the novel method.

We agree, this test could be a good indication on the quality of the final assembly. To investigate the number of chimerics transcripts between TransPi and Trinity we used an approach similar to Kerkvliet et al., 2019. First, we downloaded the following reference genes for each model species:

- Dmel-all-transcript-r6.39.fasta (*D. melanogaster* - flybase)
- GCF_000001635.27_GRCm39_rna_from_genomic.fna (*M. musculus* NCBI)
- c_elegans.PRJNA13758.WS279.mRNA_transcripts.fa (*C. elegans* - Wormbase)

We then performed a BLASTN search using the transcriptomes from TransPi and Trinity against each corresponding gene set. Parameters used were as specified

by Kerkvliet et al., 2019 (-perc_identity .90 -evalue .001). BLASTN output was filtered using a minimum length of 300bp for each match. Transcripts with one match per gene were identified as non-chimera. Transcripts with two or more matches were classified as chimeras. Table 2 from the manuscript presents the results of the chimera test for model species *C. elegans*, *D. melanogaster*, and *M. musculus*. The TransPi transcriptome has a higher percentage of unique (i.e. non chimeric) transcripts when compared to Trinity alone. Only in one sample (i.e. *M. musculus* SRR8329326) the percentage of non-chimeric transcripts of Trinity was higher than TransPi. However, this Trinity sample had over 200,000 more transcripts in the assembly.Nevertheless, the percentage difference was only 0.59%. BUSCO scores followed the same pattern as explained above in the Results section.

The implementation of a procedure like the Bellerophon pipeline (Kerkvliet et al., 2019) in our tool is hindered by multiple factors. First, for calculating the chimeras a reference transcriptome or gene set is needed. TransPi is intended to be used mainly on non-model species where the majority of these species do not have a reference transcriptome. Second, the Bellerophon pipeline makes use of a software (i.e. transrate) which has not been updated in a while. This creates reproducibility problems since the tool relies on old versions for some dependencies. Also, it does not offer a conda installation or container images. However, it should be noted that one of the critical steps in the Bellerophon pipeline is the use of CD-HIT-EST for decreasing redundancy in the assemblies. This step is already incorporated in the EvidentialGene software for the same purpose.

Reference

Kerkvliet, J., de Fouchier, A., van Wijk, M., & Groot, A. T. (2019). The Bellerophon pipeline, improving de novo transcriptomes and removing chimeras. *Ecology and evolution*, *9*(18), 10513-10521.

Discussion

Reviewer:

‘Thus, by combining various k-mer sizes (i.e. short and long k-mers), a more comprehensive representation of the transcriptome can be achieved’

The use of the advantages and disadvantages of using different k-mer sizes was studied and published by Peng et al. (2012) and the use of different k-mer sizes in a single

assembly was exploited in the IDBA-UD assembler. I have not seen this paper cited by the author despite its detailed exploration of the subject.

> Peng citation was incorporated

Reviewer:
'It has been previously shown that using more than 30M read pairs does not significantly improve the quality of the transcriptome assembly '
This largely depends on the organism.

> We agree. This is to show that studies tackling these have been done previously but given our results we see that it depends on various factors (e.g. organism). That is why we mentioned afterwards: "However, in our tests mixed results were observed when comparing reads quantities and BUSCO scores (Supplementary File 5)". Sentence was modified at the end.

Reviewer:
'Another major disadvantage of keeping false isoforms is in phylogenomic analyses'

The presence of alternative isoforms is also beneficial, so it should be up to the user to decide depending on the downstream analysis.

> After the transcriptome reduction we do not perform any other filtering. The user has the option to further clean the transcriptome if necessary. Our point here was that we reduced the number of transcripts that may not bring valuable information like in the cases of false positives. The user has to inspect the output of TransPi and decide if it requires further processing.

Reviewer:
The authors show that BUSCO scores were consistently high in most TransPi assemblies, similar to Trinity assemblies. Despite this, there seems to be a reduction in read mapping which they attribute to the smaller assembly. Although that is true, this also indicates that the EvidentialGene step has removed real transcripts that are present in the reads and in the Trinity assembly, but missing in the TransPi assembly. This is further shown in the following paragraph, where they show that some genes missing in the final TransPi assembly are found in some of the preliminary assemblies that are produced prior to merging (Figure 6).

Yes, the EvidentialGene process seems to eliminate some of the BUSCO as the missing category is a little higher in some cases. However, the paper of BUSCO (Simão et al., 2015) says that a BUSCO score should have complete single genes and the duplicates should be rare. Many duplicated BUSCO genes indicate an erroneous assembly.

As the reviewer pointed out, the mapping is reduced considerably due to the reduction in number of transcripts. We cannot discard the possibility of some real transcripts being filtered by EvidentiaGene. However, given that the differences in the missing category are not significant and the final TransPi assembly will have high BUSCO scores (i.e. complete and single genes). Furthermore, as observed in the test for chimeric transcripts (see Results section above), the TransPi assembly has a higher amount of non-chimeric transcripts than Trinity. Thus, we argue that this should not be a major concern. Nevertheless, the users have to decide if they need to do more filtering and processing of the TransPi output. The interactive report is a useful way to determine this.

Reference

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

Reviewer:

It is not clear if the additional CPU-hours invested result in an equally improved assembly. A comparison of the CPU-hours between Trinity and TransPi may help other users make a decision about which tool to use.

TransPi automatically creates a report on the CPUs and RAM usage per task and in total. This is one of the many advantages of using Nextflow. One example of such file was added to the supplementary files (i.e. Supplementary File 8). Even though it may take longer to do the entire TransPi analysis, we argue that by doing so, you essentially have all the necessary information that you need to assess the libraries, assemblies, and experiments. Also, TransPi automates the entire process making the user save time on installation and setting up scripts to run the software. TransPi is simple, user-friendly, but provides a thorough analysis in a reproducible way. A sentence was added to the manuscript to let users know about these files.

Reviewer:

As noted by the authors in the Introduction, denovo transcriptome assembly tends to

generate many partial and chimeric transcripts. It is important to measure the accuracy of the transcripts assembled and I think the authors missed an opportunity to show that with the model organisms. I suggest they compare the results of the Trinity and TransPi assemblies to the curated annotations of the model organisms and measure their correctness.

Test for chimeric transcripts in the results section (above)

Reviewer:
In some paragraphs the reference to Supplementary tables is incorrect.
Some minor redaction mistakes were found but the general ideas are still understandable.

Fixed