Review of:

TransPi – a comprehensive TRanscriptome ANalysiS PIpeline for de novo transcriptome assembly

The authos presents a new standardized pipeline / wrapper for automating transcriptome assembly of any organism regardless of the availability of a reference genome. It is aimed particularly at non-model organisms for which there may not be reliable gene models or a reference genome.

They do not claim to develop any new approach/algorithm for improving denovo transcriptome assembly, but rather to reduce human input and thus bias in the assembly outcome by automating the best combination of tools and parameters to produce optimal assemblies.

Introduction

It seems appropriate with acknowledgement to previous transcriptome assembly algorithms and software available. They do not go into details about the algorithm differences between the tools, but rather to the fact that no assembly is optimal for all cases and that, rather their combined output is preferable. It then tackles the issue of combining different assembly outputs into a single coherent assembly. The authors clearly specify their preference for EvidentialGene for the production of a non-redundant consensus assembly with higher BUSCO scores and other reference-free metrics.

Methods

Clear explanation of the pipeline steps.

1) 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.
2) 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.
3) rnaSPADES uses different kmer sizes for a single assembly. Did the authors use a single kmer size with this assembler as they did with other traditional deBruijn assemblers like Trinity, Oases or transAbyss?

Results

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.

Discussion

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

The use of the advantages and disadvantages of using different kmer sizes was studied and published by Peng et al. (2012) and the use of different kmer 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.

‘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.

‘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.

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).

Conclusions

The authors present a novel pipeline that includes several important preprocessing steps, a wide combination of assemblers and kmer sizes for the preliminary assemblies and the addition of a non-optional EvidentialGene step to produce a final transcriptome assembly.

The authors show that the pipeline is at least as good as a Trinity if not better when measuring the BUSCO scores, particularly for non-model organisms.

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.

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.

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