

Supervisor's statement – Laura Kroll-Nadine

The bachelor's thesis titled "Testing of different quality filtering algorithms on contig quality in plant transcriptome" aims to test different tools for quality filtering in RNA sequences and discuss the influence of assembly algorithms and pipelines after quality filter. The thesis analyses quality filtering tools (PRINSEQ, Trimmomatic and Filter by Tile) commonly used in DNA and RNA sequencing researcher. The first aim has a relative importance in bioinformatics science since few published studies compare the different tools created for this purpose. Second aim is also essential to evaluate the influence of that tools in the assembly processes, which generate the results used in several published papers. Therefore, the integration of the two aims in the thesis was important for a good research with solid results.

There are some considerations that I believe can make the thesis clearer and more complete. First of all, it is important to have a correctly format of the text (justified, standard spacing and paragraphs) and references for a good reading. Abstract may have more summarized information such as: tested tools, more relevant results and main conclusion of the thesis. Inclusion of more details on how the sequences were obtained can make the context of the thesis clearer. The description of the tools used in the methodology must contain the parameters used for each one. For example, what was the `min_In`, `trim_left` and `trim_qual_right` in PRINSEQ tool? A detailed methodology can help many researchers in the future who want to study topics similar to those presented here. Caption should be placed at the top of the tables and at the bottom of the figures. Captions should to have all the information needed to understand the contents of the tables and the figures (for example, table 2 was not clear if the files correspond to merged or not sequences). Graphical figures must contain the meaning of the x and y axes. Results about the influence of quality filtering tools in assembly process could be merged into a one table with all the values obtained (including all files and all parameters analysed, such as contigs total and larger contigs) instead of many graphics. Table 4 could be more complete with the results for all analysed files using N50 statistics. I strongly recommend that the results of all files analysed by the different quality filtering tools be included in the results section, even if the data are similar. This is one of the objectives of the thesis and it is essential to have this information in the results.

The comments above are recommendations for improving the text of the thesis that shows good results. However, here are some issues that should be addressed for the author to clarify in the state final examination:

- Why were just five transcriptome files chosen for the thesis analysis? What were the criteria for choosing the files?

- What are the differences between figures 4A and 4B? What do these differences mean?
- What are the differences between figures 7A, 7B and 7C? What does the largest contigs in figure 7C mean?
- After assembly, some graphics show less contigs in unfiltered files than in files after quality filtering. How is this possible?
- Are there any researches published in the scientific literature, using the same quality filtering tools, that present similar or different results for other kinds of transcriptomes? If yes, I recommend including it in the discussion of the thesis.
- For this data set from *Utricularia vulgaris* transcriptome what was the best quality filtering tool tested, considering the results obtained after assembly? I recommend that the conclusion has this statement more clearly and more conclusively.

I hope you find these recommendations and questions helpful to further improve the thesis developed during the bachelor's degree. After making the changes you feel are appropriate, I believe that the thesis is ready to be deposited in the IS STAG database coordinated by the University of South Bohemia.

I really appreciate the invitation as thesis opponent and for sharing with me this work resulting from your bachelor's degree.

Best regards,

In České Budějovice on 30th June 2020,

Dr Caio César Pires de Paula