De novo genome assembly of *Arthrobacter sp.* isolated from arctic permafrost soil

Opponent's review

Author of bachelor thesis: Mariana Šatrová
Opponent: Mgr. Vojtěch Tláskal
Date: May 25, 2017

The bachelor thesis of Mariana Šatrová summarizes bioinformatic processing of data obtained by genome sequencing of the permafrost isolate *Arthrobacter*. The workflow goes from raw data up to annotation of assembled genome and includes several assembly strategies such as optimization of raw reads, use of different assembly algorithms in two software environments. English language of the work is at high level and very well understandable without any bigger mistakes or typos. I would like to appreciate the structure and stylistic aspect of the work. Mariana wrote the text which can be followed easily even without deeper knowledge of otherwise complicated topic. Moreover, annotated genome can provide valuable insight into physiology of microorganisms in permafrost as habitat with possible large environmental impact in a near future.

The work is clearly separated into introduction, literature review, etc. The text in individual parts fits well. Following bibliography contains correctly-formatted references to 27 publications and 19 internet articles and technical sheets. It is clear from the text that Mariana understands to the principles of sequence analysis like preprocessing, assembly itself and evaluation of the results. Mariana is able to see biological meaning behind outputs from analysis steps (like quality check of sequences). Further, she is able to troubleshoot basic bioinformatic issues (like lack of computational memory and unsufficient documentation) and can provide description of the workflow which will be very useful during future publishing of the results. Notably, thesis also contains correctly working bash script thus proving Mariana's experience with UNIX environment in which the tools are used.

There are a lot of possibilities how to analyze genomic data, moreover the tools are still evolving. Thus it is a question of previous personal experience which analysis should be preferred. The workflow in the thesis is, however, chosen appropriately. Comparison of several assemblers and following integration of contigs by CISA is a logical workflow. With used sequencing method the resulting genome assembly is satisfying. After suggested additional sequencing by up-to-date technology of long reads, the dataset has promising potential to provide even more detailed informations.

I would appreciate more ecological interpretations from genome annotation in the text. But I am aware of the bioinformatical focus of the work and so I think that the amount of tasks needed for successful genome annotation is sufficient for bachelor thesis. The work can serve as basis for reporting the draft genome and possibly for publication of ecological model of the *Arthrobacter*. I recommend the thesis for approving with excellent mark.

Notes:

1) In introduction, I missed references to original articles in which mentioned assembly algorithms were published (pages 9-11). Although SPAdes tool is later used with satisfying results, it is not listed in introduction among algorithms based on k-mer approach (page 11).

2) There is the paragraph about pyrosequencing method in introduction (page 3) but any note that this technology is no longer available due to finished support from manufacturer. When talking about pyrosequencing, I would consider to mention it.

3) Difference between types of the reads which are produced by second generation of sequencers is comprehensively explained, however, when mentioning Illumina MiSeq technology it is quite important that it offers also sequencing with paired-end read lengths of 300 bp which is not listed in offered lengths on the page 4.

4) In materials and methods there is typo when referring to original FASTQ files (page 13), both have same name although sequencer is producing forward and reverse files called differently. Fragment length of original genome library is 180 bp but I missed statement about used length for sequencing itself. It appears later in the results that chemistry for 150 bp reads was used but this needs to be mentioned in methods.

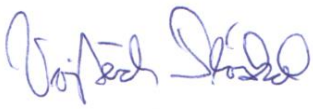
5) It is better to refer version numbers of assembly algorithms included in MyPro software since some of them are under rapid development (e.g. SPAdes, page 14).

Questions:

1) If there is possible presence of plasmid in sequenced DNA, how would you identify it in the data?

2) Can you explain annotation by RAST subsystems in more details? How does it predict genes? And what is the actual content of "subsystems"?

3) What is the phylogenetical relatedness of isolated *Arthrobacter* to other members of this genus (e.g. based on 16S rRNA gene). Would it be possible to align assembled genome to its closest relative? Which tools would you suggest for such alignment?



Mgr. Vojtěch Tláskal

Laboratory of Environmental Microbiology

Institute of Microbiology of the CAS

Prague, May 12, 2017