Přírodovědecká fakulta
Faculty of Science

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

OPPONENT'S REVIEW ON BACHELOR/~~DIPLOMA~~[1]* THESIS

Name of the student: Lara Marie Höscheler

Thesis title: Microbial community composition of Arctic cryoconite holes
Supervisor: Mgr. Marie Šabacká, Ph.D.

Referee: RNDr. Pavlína Věchtová, Ph.D.
Referee`s affiliation: Faculty of Science, University of South Bohemia

	Point scale[2]	Points
(1) FORMAL REQUIREMENTS		
Extent of the thesis (for bachelor theses min. 18 pages, for masters theses min. 25 pages)**, balanced length of the thesis parts** (recommended length of the theoretical part is max. 1/3 of the total length)**, logical structure of the thesis**	0-3	2.5
Quality of the theoretical part (review) (number and relevancy of the references, recency of the references)	0-3	2
Accuracy in citing of the references (presence of uncited sources, uniform style of the references, use of correct journal titles and abbreviations)	0-3	3
Graphic layout of the text and of the figures/tables	0-3	2
Quality of the annotation	0-3	2.5
Language and stylistics, complying with the valid terminology	0-3	3
Accuracy and completeness of figures/tables legends (clarity without reading the rest of the text, explanation of the symbols and labeling, indication of the units)	0-3	2.5
Formal requirements – points in total		**17.5**
(2) PRACTICAL REQUIREMENTS		
Clarity and fulfillment of the aims	0-3	1
Ability to understand the results, their interpretation, and clarity of the results, discussion, and conclusions	0-3	1.5
Discussion quality – interpretation of the results and their discussion with the literature (absence of discussion with the literature is not acceptable)	0-3	1.5
Logic in the course of the experimental work	0-3	1.5
Completeness of the description of the used techniques	0-3	1
Experimental difficulty of the thesis, independence in experimental work	0-3	0

1 *Choose one

2 Mark as: 0-unsatisfactory, 1-satisfactory, 2-average, 3-excellent.

Quality of experimental data presentation	0-3	1.5
The use of up-to-date techniques	0-3	1
Contribution of the thesis to the knowledge in the field and possibility to publish the results (after eventual supplementary experiments)	0-3	0
Practical requirements – points in total		**9**
POINTS IN TOTAL (MAX/AWARDED)	**48**	**(26.5)[3]**

Comments of the reviewer on the student and the thesis:

In the following Comments section both comments and questions are formulated. The questions that require answering are restated and reformulated in the Suggestions and questions section of this opponent review.

Introduction

What does "highly concentrated environment" on page 3 mean?

Materials and Methods

Due to the methods used in this thesis, I find it important to introduce 16S rRNA sequencing, which represents a crucial method for the entire thesis.

A more detailed sampling procedure is missing in The Materials and Methods chapter. What tools were used for sample collection? Were triplicates collected from a single cryoconite hole or perhaps from different cryoconite holes located close to each other? If so, were cryoconite holes of similar size sampled to receive a biological replication? With respect to a large size range from several millimetres to tens of centimetres, I would expect differences in abiotic factors influencing the composition of microbial communities in cryoconite holes of different sizes and stating the size of sampled cryoconite hole sounds as an important variable especially when sampling biological replicates.

Why were samples from different sites collected at different times of the year? With respect to altering seasons and associated alterations in weather conditions, I would expect abiotic conditions and subsequently microbial community composition to be altered. If microbial community composition is to be compared, it is crucial to avoid the inclusion of important variables such as seasonal weather changes. I believe that sampling at different times of the year makes sample sites with respect to microbial composition incomparable.

It is very unfortunate that both sets of samples, i.e. those used for isolation, PCR and sequencing (listed in Tab 1.) and those provided for data analysis (listed in Tab 3.), were not introduced in one place in the Materials and Methods chapter. Instead, the sample set used for data analysis is introduced in chapter Results and what is more, sampling dates are missing in the set of samples used for data analysis. This structure makes the result chapter very difficult to read and the reader is forced to make long searches through the thesis to find details about sampling sites.

Why is the entire DNA isolation procedure, using Quick-DNA Fecal/Soil Miniprep Kit by Zymo Research, rewritten in the Materials and Methods section? The isolation protocol is usually present

[3] Enter the number of points awarded.

and clearly written within the manufacturer's manual of each isolation kit. It does not make any sense to rewrite the entire procedure when it can be easily referred to in the manufacturer's instructions. Instead, it is a common practice to only highlight potential changes made to the manufacturer protocol to help other researchers reproduce the methodology and results.

It is also a common practice to evaluate the purity of isolated nucleic acid samples using simple spectrophotometric measurements. Spectrophotometric evaluation of nucleic acid samples is an important value required by any downstream analyses due to potential contamination of the sample with guanidine salts or phenol and with proteins. The presence of these impurities could cause the downstream enzymatic reactions, such as PCR or sequencing library prep, to perform suboptimally and provide low-quality and finally incomparable results.

The description of any PCR requires a proper introduction of all PCR components and PCR reaction conditions. For example, the amplicon length of the PCR reaction is very important information, which also helps the reader to evaluate the validity of the results. In addition, primers for 16S rRNA amplicon sequencing possess a more complex structure and it would be useful to introduce it in the thesis to support the introduction part and for readers to fully comprehend the principle of 16S rRNA library prep and sequencing.

In chapter 3.5. Illumina sequencing, it is stated that "All samples….were sent to SEQme for DNA sequencing with Illumina". It is important to properly introduce all commercial chemicals and services used to produce the data in a thesis. Thus, the SEQme should be introduced as "SEQme s.r.o. (Dobris, Czech Republic) a company providing sequencing services" or similar.

The sample list in Table 3 in Chapter 3.6. provided for Data analysis, shows a different number of samples per sampling site. Why were not all the sampling sites examined three times to ensure the statistical significance of the microbial species identification? What is more, the sampling date in the case of this list of samples is absent. This information is crucial for the evaluation of data reliability, comparability, and statistical evaluation.

What exactly do "values" and "variables" refer to in the sentence "It removes two types of variables with extremely small values and variables with constant values (detected by robust estimate interquantile range (IQR))" in chapter 3.6.2. Downstream processing? Similarly, what does "value" mean in the sentence "For data normalization, row-wise normalization (sample against sample) by sum was chosen in order to ensure that values within sample are proportionally distributed based on their relative contribution to the total sum in order to be able to compare each sample with different sizes to each other." also in chapter 3.6.2.? This also applies to the rest of the chapter 3.6.2. where term "variable" appears a few more times. Explanation of this term is essential to understand all the results in this chapter. In addition, I believe that "IRQ" normalization stands for "Interquartile range" and not "Interquantile range".

Results and Evaluation

In Table 5, there are lots of DNA concentration values that say >120.00, which indicates that the sample concentration value is above the dynamic range of the device, in this case of the Qubit 3 fluorimeter. If this happens, the sample should be diluted and remeasured to receive a concentration that falls within the dynamic range of concentration measurement of the device to receive an exact concentration value. Ideally, the sample concentration can first be estimated using less precise but more simple and less expensive spectrophotometric measurement and the obtained value can be used to dilute the sample to fall within the dynamic range of the more precise and

specific fluorimetric device.

The information about 16S rRNA sequencing is entirely missing in the thesis. For example, the information about the length and number of reads per sample, including single-end or pair-end mode, is important to prove that the presented results were produced using a representative amount of data that is informative enough and reliable in order to allow the scientist to make firm and statistically supported conclusions about the presented research.

In chapter 4.2. Gel electrophoresis the author claims that the absence of DNA bands in Fig. 5 was caused by insufficient electrophoretic separation of the samples. In my opinion, if DNA separation was insufficient, we would be able to see visible bands of non-separated DNA samples close under the agarose gel wells, similarly to the DNA marker. In addition, if the DNA separation is insufficient, it can be easily placed back into the electrophoretic device and electrophoresis can be rerun for a longer period to allow for sufficient separation. If the author concluded that this unsuccessful result was caused by insufficient electrophoretic separation, why did not she prolong the separation time?

For future work, I recommend that the author provide sample labels right above each well in photographs of agarose gels. In addition, the sample identity is absent in both the photo and the figure legend in Fig. 5 and is only present in the legend of Fig. 6.

Why is the agarose gel concentration 1.3% in Fig. 5. and 1.5% in Fig 6.?

The Chapter 4.3. Data analysis, which ranges 7 pages, is for the most part composed of a somewhat tiresome description of what the reader can easily find in the presented pie charts. Instead, it would be much more useful to only highlight interesting or unusual findings.

The data in Table 7 in Chapter 4.3.8 show the relative abundances of the presented genera in all glaciers. Were the relative abundances calculated from the averaged relative abundance values of each sampling site? If so, why is not data variance, such as standard deviation, also presented for each abundance value? It is essential information for the evaluation of data reliability and significance.

There is no explanation of the results of PCA analysis presented in the 2D score plot in Fig. 21. Why is the plot presented in the thesis and what significance does it pose for the respective data? Although some description of it is given in the Discussion chapter, it does not provide any information about the meaning of the analysis and thus the information about the significance of this kind of result for the presented data is unclear. Additionally, the main description of the result should be given in the place where the result is first presented.

Similarly, there is no explanation of what the “Chao” and “Shannon” indexes are in a chart of Alpha Diversity in Fig. 23 and actually, there is no explanation of what alpha diversity is and why it is important for this kind of data.

Discussion

I appreciate the explanation for using the gene for 16S rRNA for the taxonomic identification of microbial communities in this study. However, I believe that the main reason for using this gene as a molecular marker is its ideal sequence evolutionary rate which allows for the identification and taxonomic classification of microbial composition on the species level.

Do you have any explanation for why replicated samples of one sampling site cluster with samples from different sampling sites as presented in PCA analysis in Fig. 21 and Pearson's clustering algorithm in Fig. 22 as presented in the discussion? How could you make any conclusion about the microbial composition in each sampling site if replicated sampling gives absolutely different results and clusters randomly with other sampling sites?

Conclusion

Why is this sampling environment prone to rapid DNA degradation as written in the conclusion? Does this statement refer to local climate or perhaps limited tools for sample collection and/or storage or other reasons?

Instead of making any conclusions about sampling site compositions and their comparison among each other, I would focus on the reasons why the data from different sampling sites are incomparable and what reasons caused the data irreproducibility such as different sampling site dates, different sampling methods or conditions or differences in quality of isolated DNA samples causing suboptimal performance of downstream enzymatic reactions within library prep and sequencing pipeline.

Suggestions and questions, to which the student has to answer during the defense. Mistakes, which the students should avoid in the future:

1) What does "highly concentrated environment" on page 3 mean?

2) What tools were used for sample collection? Were replicates collected from a single cryoconite hole or perhaps from different cryoconite holes located close to each other? If so, were cryoconite holes of similar size sampled for each biological replication?

3) Why were samples from different sites collected at different times of the year?

4) Why was not sample quality measured spectrophotometrically?

5) Could you, please, explain the principle of the amplicon sequencing method? Could you also clarify the structure of primers used for 16S rRNA amplicon library preparation? What was the amplicon length and why was the V3-V5 16S rRNA region used for this project?

6) Why were not all the sampling sites examined three times to ensure the statistical significance of the microbial species identification?

7) What exactly do "values" and "variables" refer to in the sentence "It removes two types of variables with extremely small values and variables with constant values (detected by robust estimate interquantile range (IQR))" in chapter 3.6.2. Downstream processing? Similarly, what does "value" mean in the sentence "For data normalization, row-wise normalization (sample against sample) by sum was chosen in order to ensure that values within the sample are proportionally distributed based on their relative contribution to the total sum in order to be able to compare each sample with different sizes to each other." also in chapter 3.6.2.?

8) Could you, please, provide the information about 16S rRNA sequencing raw data used for

data analysis in this thesis? In particular, provide the information about the sequencing mode used for library sequencing (i.e. SE or PE), the length of the reads and the number of reads per sample. In addition, how many reads were discarded after low-quality read filtering?

9) Why was not electrophoretic separation prolonged in the case of the agarose gel presented in Fig. 5?

10) Why is the agarose gel concentration 1,3% in Fig. 5. and 1,5% in Fig 6.?

11) Were the relative abundances calculated from averaged relative abundance values of each sampling site in the case of samples listed in Table 7? If so, why is not data variance, such as standard deviation, also presented for each abundance value?

12) Could you clarify the meaning of PCA analysis in Fig. 21?

13) What are the "Chao" and "Shannon" indexes in the Alpha diversity chart in Fig 23?

14) Do you have any hypothesis about why replicated samples from the same sampling site do not cluster together? What implications does this phenomenon have for your data interpretation?

15) Why is this sampling environment prone to rapid DNA degradation as written in the conclusion?

Conclusion: The thesis complies with all the formal requirements given by the Faculty of Science JU. It is written in nice English with a minimum of mistakes and typos. The thesis structure requires some rearrangements to make it more organized and easier to keep track of the flow of the ideas. There are more topics that should be introduced and are important for understanding the methods and results of the thesis. It is very unfortunate that different samples were used for processing in a laboratory and data from different samples were used for data analysis. It makes the entire thesis very difficult to read and lowers the value of the entire study. In addition, the Materials and Methods section lacks lots of important information about sample collection and identity and in general there seem to be many issues in overall sampling methodology. The major issue lies in sample replication, the lack of statistical evaluation and subsequent problems with results interpretation, which makes the entire project questionable and unpublishable. This study may provide general insight into cryoconite hole microbial communities in Svalbard in general allowing an approximate comparison with other glaciers worldwide. However, I find comparisons among different sampling sites in Svalbard highly questionable due to the aforementioned problems with the circumstances of sample collection and the absence of statistical evaluation.

In conclusion, I

r e c o m m e n d / ~~d o n o t r e c o m m e n d~~*

the thesis for the defense and I suggest the grade 3.[4]

[4] You can suggest a grade, which can be modified during the defense based on the presentation. However, if the reviewer is not present at the defense, the grade will not be counted. Grades: excellent (1). Very good (2), Good (3), Unsatisfactory/failed (4).

In České Budějovice date 11. 9. 2023

……

signature