Review of the Master Thesis of Tereza Flegrová

Analysis of the dicyemid diversity using amplicon sequencing

Despite its several intrinsic problems, metabarcoding recently revolutionized our understanding of the diversity of (especially) microscopic biota. Since our lab used this concept in several projects dealing with the diversity of plankton, I was more than happy to inspect, how it works also for multicellular 'parasites'. Right off the bat, I have to admit I like the background and design of the project. It nicely extends the effort of the lab to uncover more about the neglected-and-therefore-mysterious dicyemids. I also like how the author accepted the challenge of methodically complex and technically demanding work.

The thesis is written in English. It is mostly comprehensible, there are some cumbersome phrases, but the level is generally good. It has a standard and logical structure. The Introduction is rather brief, and I would expect/welcome a more thorough Introduction of the metabarcoding concept and its advantages and shortcomings to the readers. Aims are well defined, although not in the logical order (I think aim 4. had to be completed first so the other question could be answered).

On the other hand, M&Ms are described in great (but not excessive) detail. Still, the motivation behind some procedures is missing (see my comments below for P15-22) and could be explained even more (IMHO). Results are again concise and, in my eyes, also the most problematic part of the thesis. Maybe the author lacked the time or the patience to explain everything properly, but the way the results of her hard and demanding work are presented is not entirely satisfying (to say the least). See again my comments below for details. It is by no means an easy task, but I find execution lacking in clarity and attention to detail.

On the other hand, the Discussion is comprehensible, and witty, and shows a good work with the literature as well as the authors' ability of critical scientific thinking. Although some aspects could have been addressed more (such as the impact of oligonucleotide choice on the metabarcoding results), I think it is the best chapter of the thesis. In general, apart from the somewhat lacking presentation of the Results, I like the thesis quite a bit. The author managed to learn an impressive number of techniques, and gather a lot of usable data, which I have no doubts will build a solid foundation of nice publication in a (hopefully not far) future. To my best knowledge, It fulfills all the criteria for a Master's thesis and therefore author deserves her degree, I will suggest my grade after the defense though. Bellow, please find my specific comments/questions. Bold font highlights those I would like to see/hear addressed during the defense.

P15 First paragraph describing the preparation of amplicon libraries refers to Fig.5. vs Fig.7

P16 I miss some details on the oligonucleotides used for dicyemid metabarcoding (16S bacterial/18S dicyemid?). Seems like the author used oligonucleotides designed to amplify

prokaryotic SSU rRNA (Parada et al. 2015/16)? If so why and why she did not attempt to design primers specifically amplifying dicyemids? I think the impact of this step of the pipeline should have been critically evaluated in the Discussion.

P19 Please explain how running the analysis in qiime2 using the forward reads only helps you to validate the accuracy of your approach.

P22 What is denoising and why do you use it in your pipeline?

P22 Chapters 3.6.8 – 3.6.11 comprise a single sentence. Such over-structuring doesn't help the flow of the text and merging several chapters into one would likely help.

P22 **What is the distribution of variability in V4 in dicyemids? By excluding the central part of the region from your analyses, did you lose some substantial part of the information? If so, how does it affect your estimates?**

P22 Have you tried also PR2 db for ASV taxonomy mapping? In our experience (at least for protists), it is vastly better than Silva.

P25 What is a GenBank accession? Why should they be eliminated from the results of species delimitation?

P25 What is a 20% filtering approach?

P25 Why there is a phylogenetic tree (missing the scale bar) with bGMYC species delimitation approach shown in Fig. 11 and on P26 there is the same tree with all three methods of species delimitation (including bGMYC) shown? Did the author try to also map the geography or taxonomy on the tree to see if there are some interesting patterns?

P26 In several places in the text, the author mentions chromidinids. Since this obscure lineage of ciliates with just a few genera described is not generally known, I think it would deserve at least one or two explanatory sentences, so the reader doesn't have to look the info him/herself up.

P27 In Fig.13, I assume the numbers above bars of individual octopus/sepia statistics are achieved P-values. I think the information is not that useful since just looking at the bars tells us the distribution of values of alpha-diversity indices among both hosts is similar and the notion mentioning this in the legend suffices. Nevertheless, an explanation of a number's meaning must be mentioned somewhere and not rely on the interpretation of readers.

P28-30 Fig 14. I understand the need to present all the results in one place, but not sure this is as comprehensible as the author intended. It spans over three pages, but the coding is shown on page 28. So, if you want to understand Fig. 14C, you have to switch between pages 28 and 30. Moreover, the color-coding of individual ASVs is chosen very poorly and doesn't help to differentiate the geographic origin of (some) ASVs at all. For example, the color difference between two ASVs from Tyrrhenian Sea Tyrh01 and Tyrh13 is much bigger than the difference between Tyrh13 and Viet01 Also, what are the black-to-greyish lines interrupting individual bars/samples. I'd rather see this Figure in Suppl. material

P31 Tables 7-10 show the distribution of individual dicyemid ASVs among the host samples using double coding for abundance categories, using both numbers and colors. In my opinion, this is again confusing, one has to spend unnecessary time to understand this duality and color-coding would IMO suffice.

P35 Fig. 15 is much more comprehensible than Fig. 14, however, it shares the same problem of poorly chosen color-coding.

P37 Fig. 16 The idea behind the figure is great. However, I am afraid it again lacks comprehensibility by trying to tackle both host and geographic patterns of ASV distribution into one picture with the same poor color coding. **Also, I am missing an analysis of the impact of individual variables on the distribution. Is geography more important than host specificity?**

P37 and P38 I miss the more detailed evaluation of the results of network analysis (Fig. 17) and the comparison with the results of ordination plots (Fig. 16).

České Budějovice, May 18th Aleš Horák