Review of the Bachelor thesis of Lirik Behluli

Is DNMT3L a key factor for de novo DNA methylation in mammalian oocytes?

As the name suggests, the thesis investigates factors affecting the DNA methylation in mammalian oocytes but using the *in silico* analysis of relevant transcriptomic data. Since given the bloom of NGS, we have much more genomic/transcriptomic data than we'll ever be able to properly analyze, I think the topic/approach is very well chosen. I am by no means an expert on the topic and was chosen as a reviewer based on methodological rather than thematic overlap. Therefore, I will focus mostly on conceptual and formal things, but I hope that my distance and layman's view could be eventually beneficial for the author and supervisor.

As usual, the thesis has a conventional structure. Lengthwise, although on the lower margin, it falls within the tolerable range. The first thing I noticed is that the Contents lacks number of pages and was to complain about it, only to find out a couple of pages later that the whole thesis lacks page numbering. Is there some specific reason behind it? It really makes reading and reviewing more difficult.

The Introduction is very dense with abbreviations and complicated names and is targeted at a specialist (not me 😊). Although it is written in good English, it doesn't read well. Some broader introduction to the mammalian ontogeny as well as figures illustrating the main points would definitely help.

The Aims are maybe too concise. The author should explain the broader concept of Aims by some generalizing sentences at the beginning and also pay more attention to their formulation. For example, the difference between the first and third aim is negligible. Compared to the Introduction, the language of M&Ms chapter gets clumsier. There are several examples of unnecessarily repeated information: „*SeqMonk software was utilized to visualize and quantify RNA-seq datasets. SeqMonk is a bioinformatical program (software) for visualizing and analyzing mapped next-generation sequencing (NGS) data.*" The text suggests that both paired-end and single-end libraries were used. If so, I would like to see the information about the type of utilized libraries in Table 1.

In Results, the level of English further deteriorates, but the text is mostly comprehensible. There are several clumsy wordings, sentences are unnecessarily long, and I found some typos (shorten/shorter…). However, given the degree of the author, I still find it acceptable and mention it for the sake of future improvement. However, I have a problem with the graphic presentation of most of the results (see my detailed comments below). Looks like there was either a lack of time, a lack of attention to the detail, or indifference. I hope the first possibility is true. Individual subsections of Results are also poorly interconnected. I miss some red twine that would guide me through it, so I understand why a particular analysis was done and what was the question at the beginning, why does this matter?

Compared to the Results, the Discussion/Conclusions are well written and as in the Introduction author shows his good work with literature. Suddenly, some poorly explained analyses make sense and fall into place. Although a bit too short, it is the best part of the thesis. And it scores some extra points for the suggestions for future research at the end.

In general, despite my critique, I think this is a decent thesis. Sure, there is some room for improvement, but on the other hand, I read some much worse. I am inclined to evaluate it as very good but will reserve my final judgment after the defense. Below please find my detailed comments. I'd like to see/hear those highlighted in bold font addressed during the defense.

Comments:

I think the chapter 4.2 would be better placed at the end of M&Ms. Moreover, the author should have paid more attention to explaining why it is so important to know that "*Dnmt3l is annotated at the correct position in the naked mole rat genome, and in case it is not; to identify where it is in the genome based on the sequence similarity with human Dnmt3l sequence.*"

Chapter 4.3. I strongly disagree with the statement that "*Clustal program performs multiple sequence alignment, while other programs (generally) will do mostly just alignments of 2 sequences (pairwise sequence alignment)*". In general, there are several aligning software that vastly outperforms ClustalW. I understand the reasons for using ClustalW in the context of this study, however, I'd recommend confirming resulting alignments using some iterative aligning SW, like MAFFT.

Chapter 4.4 What is the 'normal' level of the conservancy of introns? How was this 'normality' estimated?

Chapter 4.5 doesn't say much and could be probably substantially reduced (if not omitted from the thesis). BTW, isn't the grep a better and more efficient tool to do the job instead of Notepad++?

Chapter 5.1 The boxplots as presented provide a good overview of expression in individual species. However, given the picture size and number of plots, it is hard to compare it across the species. Maybe it would be better to squeeze all the boxplots into one large figure (in a 3x3 panel)? Alternatively, a table summarizing the expression levels of studied genes across all examined species or the heatmap would be really helpful!

Chapter 5.2 Autor mentions the isoform of Dnmt3 genes in naked mole rat. Are these isoforms confirmed experimentally, or were they found in the transcriptome assembly, and therefore there is a high chance they are artificial?

Figure 10 is not very intuitive, I think there are better ways how to visualize the results of the Blast search.

Chapter 5.3 **Why study oocyte promoter sequence in mouse, rat, and hamster?** This comment applied to most of the results though. I can come up with several explanations, but the motivation behind each specific aim has to be clear from the text, so the reader doesn't have to wonder.

Figures 12. and 13. Why on Earth author did not include species names in the alignment?! Also, it looks like the Figures 12 and 13 could be presented only once merging the sampling of both datasets into a single alignment? What was the purpose of splitting it in two?

Chapter 5.4 Figure 14 is very hard to get the actual meaning. I think these data would be much better visualized by value matrices instead of bar plots.

Chapter 5.5 The evaluation of results presented in the text is purely subjective and I would rather see some more exact comparison of alignment conservancy.

The legend on the x-axis of Figure 19 is not readable. Again, to get better overview Figures 15-19 could have probably been merged into one

Chapter 5.6 **I wonder, what is the rationale behind repeating the comparison for mouse, naked mole rat, and human only when there are no apparent differences when analysis of a more general**

dataset reveals there that the presence/absence of oocyte Dnmtl3l has no apparent effect on expression profiles of epigenetic factors. What do the results really mean, and what is their biological value? Also, what do you mean by "*difference at least 3*"? I'd definitely go for a different style of presentation. Maybe the points would be better illustrated by a heatmap in the main text rather than a lengthy and confusing table in the Appendix?

Ceske Budejovice, May 26th Ales Horak