Přírodovědecká fakulta
Faculty of Science
Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Review of the Bachelor thesis

"Multispecies analysis of the role of epigenetic factors in aging"

by Miriam Müller

The present bachelor thesis aims to investigate the expression patterns of genes encoding epigenetic factors in representatives of short- and long-lived mammals and studies how they change with aging. The thesis is divided into two main parts, namely (i) analysis of differential expression of these factors in young and aged tissues of species with different lifespans (human, naked mole rat vs. mouse, rat) and (ii) identification of their orthologs across an extended dataset (human, naked mole rat, guinea pig, rat, blind mole rat, mouse, marmoset, bowhead whale).

The thesis has 50 pages and is organized into classical chapters (Abstract, Introduction, Aims, Methods, Results, Discussion, References and Appendices). It contains 11 figures, 5 tables, 7 Appendix tables and the text is supplemented with two dataset files and a custom python script used for data processing. The thesis is written in good English and the quality of the text is high, with minimum mistakes.

The Introduction is 2.5 pages long and briefly describes the nature of epigenetic changes and their linkage to aging. This chapter would benefit from a more in-depth literature review, focused for example on gene groups later evaluated in the study (Polycomb, PIWI, Trithorax…). Some information would deserve to be supported by citations (**"**Around one percent of the genome contains sequences encoding proteins.**"**) and citations in the References mostly list only the first author of the study, omitting **"**et al.", which may be caused by use of citation software.

The Aims of the study are clearly stated. Materials and Methods provide a description of the input data and their processing, as well as the python functions written by the author. Again, a few citations are missing (SeqMonk, Python) and it is not always clear whether the tools were run in default mode or with some parameters (Trim Galore). The python script itself is well explained and commented. I also appreciate the Tables 1-5 showing examples of the data structures and working files.

In Results, the custom-made python functions, their workings and outputs are further described in great detail. Expression values of gene classes and individual genes are compared between species with short and long lifespans and plotted in a series of graphs. However, the Figures would deserve more attention (Figure 2 – wrong label, Figure 4 – the top label is too large and the figures seem to be partially overlapping, Figures 6-9 - contain tables, also an indication of a,b,c is missing). Despite the lack of technical/biological replicas and the necessity of statistics to test for

significant differences, the author was able to identify potentially interesting genes, representing candidates for future research.

The Discussion has four pages, the results are critically evaluated with added context. Aware of the weaknesses of the input datasets, the author offers a possible explanation of the observed phenomena and elaborates on the up/downregulation of genes that showed differences between short- and long-lived species. Nevertheless, the usage of the phrase **"**significant differences" should be avoided, as the significance was not tested.

I have the following questions:

(1) 2 - **"**Eukaryotes, including the human, store their genetic information within the DNA in the nucleus. Around one percent of the genome contains sequences encoding proteins." Is that true for all Eukaryotes?

(2) 4.2 - Can you please elaborate on read trimming with TrimGalore? Did you use any parameters and how exactly were the reads treated by the program?

(3) 4.4 – How was the information on strand specificity obtained?

(4) 4.5 – **"**Then the ratio of aged vs young was calculated by subtracting the value "young" from "aged"." The result of subtraction is a difference, but have you considered using ratio?

(5) 5.2.1 – Based on Table 3, the "average_log_ratio" values of mouse kidney datasets GSE181797 and GSE163821 are quite different. Have you tried comparing expression patterns (gene classes and individual genes) between these two datasets?

(6) 5.3 – Can you please hypothesize what caused the different numbers of successfully obtained orthologs per species?

As a final remark, I would like to note that an analysis of differential expression, in general, is a rather complex and possibly tricky topic and it certainly required courage and hard work to tackle the project.

Despite the listed issues, I can conclude that the author gained the ability to handle and process bioinformatical data, obtained original results and produced scripts that can be used in future studies. In my opinion, the thesis fulfills the requirements and I recommend it for defence.

May 25th, 2022, České Budějovice

Anna Voleníková