Review of the bachelor thesis

"**Characterization of oocyte- and embryo-specific 5' untranslated regions (UTRs) of mouse mRNAs**"

by Alexandra Austenová

The present thesis is focussed on identification and analyses of oocyte- and embryo-specific 5' untranslated regions (5' UTR) from available mouse RNAseq datasets using a pipeline, which was already established in the Laboratory of early mammalian developmental biology. Genes expressed in oocytes and preimplantation embryos with 5' UTR longer than the one annotated based on expression data from somatic tissues were identified in several filtering steps. Their GC content was determined, and they were searched for upstream open reading frames (uORFs) and regulatory sequence motifs. The thesis is written in good English, consists of 75 pages and is organized into Abstract, Introduction, Background, Aims, Material and methods, Results, Discussion, Conclusion, References, and Appendices. The text is supplemented with workflow overview, 11 figures, and 9 tables. Appendices contain custom python scripts used for analyses written by other member of the lab and R scripts for plotting of data.

The Background chapter introduces 5' UTRs and their various features affecting initiation of translation. I came across following issues:

(i) Sections 3.2.6.3 and 3.2.8 dedicated to G-quadruplexes could have been merged;
(ii) I am not sure how to read the Kozak sequence (p. 5), should it be GCCAUG or GCCAUGG;
(iii) how to reconcile claims from pp. 3–4 that "vertebral transcripts of transcription factors, protooncogenes, growth factors, and receptors tend to all have a 5' UTR that is longer", "transcripts coding for regulatory proteins have 5' UTRs with high GC content", and observed "inverse relationship between the length of 5' UTR with respect to its GC content". I find these contradictory.

The Aims are clearly defined. The Material and Methods were, however, difficult to follow, although it could be because I am not so familiar with the methodology used. I encountered several issues:

(i) crucial information such as number of libraries, from which tissues they were prepared, and whether these were replicated was missing; also some graphical scheme, other than the workflow overview which should be part of methods, explaining logic behind the filtering process would be helpful;
(ii) transcriptomes were not assembled de novo, i.e. without reference, their assembly was genome guided;
(iii) files "overlapping_xxx.txt" and "upstream_xxx.txt" (p. 20) were not mentioned in the text. The author probably meant "overlapping_(plus|minus)_xxx.txt" and "upstream_(plus|minus)_xxx.txt".

I have also one, probably, naïve question. The strand information (pp. 17–18) for genes was obtained in a rather complicated way. Cannot it be parsed from genome files?

The Results to great extent recapitulate the Methods and do so more clearly as they are supplemented with tables and figures I missed in the previous section. Dataset comprising before filtering 35579 genes was reduced to 3080, 6126, and 4101 genes in oocyte, embryo, and somatic

tissues, respectively, and the genes were matched to annotated genes. Oocyte and embryonic genes with 5' UTR longer than annotation based on somatic tissues were selected for further analyses and categorised based on their length to (i) ≤ 500 bp, (ii) ≤ 1000 bp, and (iii) ≤ 3000 bp. However, these categories were not used consistently as descriptive statistics were counted from all genes (given the "Greatest elongation size" and differences between mean and median listed in Table 6), while genes ≤ 5000 bp are plotted in Figure 8 and non-overlapping ranges are plotted in Figure 11.

The Discussion is detailed and critical. It reads well and the author even proposed possible follow-up experiments. I have following questions:

(i) There are clear differences between 5' UTRs of oocytes and embryos and somatic tissues. Interestingly, on several occasions it is hypothesized that novel upstream 5' UTRs discovered in early embryos were retained from oocytes. What is the overlap between genes with upstream 5' UTRs expressed in oocytes and embryos? It seems oocytes and embryos together were compared against soma (see Table 5) but not against each other, which is pity as it could allow to test the hypothesis right away.

(ii) It is argued that some 5' UTRs might have been misannotated due to a low coverage. What was the coverage threshold used in the Cufflinks assembly?

(iii) Have you considered some functional annotation of the genes with longer 5' UTRs e.g. using Gene Ontology?

(iv) How do you reconcile presence of miRNA binding sites in oocyte-specific 5' UTRs with supposed inactivity of the miRNA pathway in mouse oocytes?

(v) The work assumes that longer 5' UTRs provide better control of translation. But what about oocyte and embryonic genes with shorter 5' UTRs compared to genes expressed in soma. Have you checked those? Are there any and how would you interpret their presence?

As a final remark, I would like to note that tools with graphical user interface were mostly used for the analyses and their outputs were parsed in MS Excel, which, as I well remember, is rather cumbersome for large files. As a bioinformatician, the author should switch to stand alone tools in future, which would allow to write and share a well commented scripts able to rerun the analyses.

Conclusion

Despite issues listed above, I can conclude that the author produced original results, met the aims of the thesis, and acquired hands-on experience with processing NGS data and manipulating big datasets. The present thesis thus in my opinion fulfils all requirements and I recommend it for successful defence.

On July 1, 2020, in České Budějovice

Petr Nguyen