Name of the student: Ms. Karolína Kraváriková
Thesis: Identification and Characterisation of Novel Transcripts in Mouse Oocytes and Embryos.

Reviewer's statement

In her Bachelor thesis, Karolína identifies novel transcripts in mouse oocytes and preimplantation embryos, analyze their expression profile, protein-coding potential, the activity of transposable elements as promoters of such transcripts, to identify transcripts with potential roles in mammalian development.

Karolína shows a broad knowledge of the topic of her thesis. Information within the literature review is appropriate for the research topics studied. The aims are clear and fulfilled by the presented results. The results are mostly observational and descriptive. Nevertheless, Karolína demonstrates promising expertise in a transcriptomic data analysis, but she didn't present the analysis pipeline clearly in some cases. Her findings contribute to identifying and characterizing novel transcripts in mouse early developmental stages. Karolína highlights the basis for a better understanding of transposable elements and their role as promoters as well as future experimental functional characterization of the candidate novel genes with potential roles during mouse early development. The thesis satisfies the formal criteria. Information is correctly referenced, including the figures. Figures are generally well designed but sometimes without legends or with low quality.

I have comments regarding the text :

The **introduction** without any descriptive information with a few vague sentences while the **abstract** and **background** sections were well written and informative. The **methods** section was not well written with several language errors, which make it hard to understand. The **methods** section needs to improve and could be made shorter without too much detailed information.

The **results** section was well written and the results correctly presented. But, methods statements have been repeated in the results section. A none common terminology was used in both methods and results sections which make the text hard to understand "i.e. using Probe instead of genes/transcripts". Some results could be more understandably with better presentation "i.e. venn diagrams". Finally, the **discussion** section was informative and I appreciate that other possible analysis strategies and future analysis have been discussed as well.

Additionally, I have few comments regarding the formal side of the thesis, pointing out minor issues that could be improved:

1. Combining the de novo assemblies from different assemblers and using different k-mers, should increase the possibility of identifying novel transcripts.

2. Implement a structural annotation for gene prediction using " i.e. AUGUSTUS" may improve the pipeline for identifying true de novo transcripts/genes and functional annotation of the identified transcripts/genes will confirm the potential role of the novel transcripts/genes.

3. I recommend following a different approach to identify lncRNAs and confirming the identifying by searching against one or more of the long non-coding RNA public databases "i.e. NONCODE".

4. Unifying the format is highly appreciated.
   The text is Times New Roman, but the page and section numbers are Calibri.
   Section headlines, tables and figures labels have different format.

Questions for the author:

1. In section **5.1 Processing and mapping of RNA-seq datasets:** the trimmed reads were mapped to the mouse reference genome. why the mapping step has been performed? do you think that should affect the de novo assembly by being biased to the mouse reference genome assembly?

2. In section **5.4 Quantification of gene expression:** I assume that the merged datasets from the same developmental stage were used for the quantification. Which quantification statistical matrix has been used to normalized read counts?

3. In section **5.3 De novo transcriptome assembly:** it was stated that genes without directions has been removed, why these genes have been excluded?

4. In section **5.6 ChIP-seq datasets processing and mapping, peak calling, and selection of high confidence novel transcripts:** How the promoter of the novel transcripts has been identified?

Overall, I appropriate the time and effort of Karolína to produce such much results in a short time. Therefore, I recommend the thesis for defense with grade 1 (excellent).

In České Budějovice, 25.06.2020 Abdoallah Sharaf, Ph.D.