Skoltech

Skolkovo Institute of Science and Technology

# Jury Member Report – Doctor of Philosophy thesis.

**Name of Candidate:** Mariia Vlasenok

**PhD Program:** Life Sciences

**Title of Thesis:** Transcriptomic analysis of the interaction between pre-mRNA splicing and intronic polyadenylation

**Supervisor:** Associate Professor Dmitri Pervouchine

**Name of the Reviewer:** Mikhail Gelfand

| | |
|---|---|
| I confirm the absence of any conflict of interest | **Date:** 11-11-2023 |

**Reviewer's Report**

As suggested by the title, the dissertation is about an interesting problem, the interplay between (alternative) splicing and (alternative) polyadenylation. The thesis is logically structured and generally well-written. The review, while not very long, is sufficiently complete and detailed.

While the use of reads containing non-templated adenines had been applied before, the candidate was a pioneer in doing it large-scale (the methods were also improved). It allowed her to discover new, interesting and important biology. In particular, the developed methods allowed the candidate to quantify the level of intronic alternative polyadenylation without bias caused by low coverage, and she observed that it was more frequent than exonic alternative polyadenylation. This unexpected observation was interpreted as a sign of on-going competition between splicing and polyadenylations: spliced polyadenylated introns (a term introduced by the candidate) are generated when the intron is simultaneously polyadenylated and spliced. This is an important hypothesis running against current wisdom and, if confirmed, quite deserving a paragraph in the textbooks. I should also specifically mention the detailed and explicit description of fine-tuning the methods, motivation and preliminary analysis for selecting particular values of parameters, saturation and robustness checks etc.

The level and quality of publications is quite high, as they include a first-author publication in *NAR – Genomics and Bioinformatics* and a paper in *Nucleic Acids Research* sensu stricto. Nothing is said about conferences, but as a member of the candidate's IDC I know that the formal requirements in this respect are met (still, it might be a good idea to mention the conferences explicitly).

I have no major comments; some misprints still remain, e.g. *Xenopus t.* on page 37 (although the quality of the manuscript is very high). Some more minor points: (1) It is not clear what is the relationship between various sequence motifs in Fig. 2-1 and Fig. 2-2: only the hexanucleotide AAUAAA and its variants are mentioned in the text, while some others are seen in the figures (GUGU, UGUA, CA); a U-rich downstream region is mentioned in the Results chapter (p. 67) but not in the Review chapter. (2) *From the 565,387 PASs with $H \geq 2$, 331,563 contained a sequence motif similar to the canonical consensus CPA signal* – should be "out of". (3) *The signal* in the legend to Fig. 5-6 is used without definition (defined much later in the text)? (4) Why calculating local GU-content is relevant to the identification of a "U-rich region" (p. 67)? (5) *the ratio $wi_1/wi_2$ was skewed towards positive values* – the ratio is positive by definition; the author probably means "higher than 1" (or positive logarithm).

However, I have some suggestions for discussion and maybe further analysis.

(I) The author posits that deviations of the AAUAAA-like sites from the consensus are needed to maintain a proper level of (alternative) polyadenylation. If this is true, one would expect conservation of non-consensus nucleotides (as demonstrated by Stepan Denisov for splicing sites and by Eugenia Belousova for binding sites of bacterial transcription factors; the latter study not published yet, but reported at ITaS(b) seminars) – given hundreds of available genomes, it might be interesting to look at. In the same vein, one may directly compare site strength (measured by a positional weight matrix) and polyadenylation efficiency (although that might be difficult as it would require non-trivial normalization to account to the competition by the parallel splicing process).

(II) Would not restricting the analysis to *PAS clusters located in genes containing at least one RNA-seq-derived PAS and at least one 3'seq-derived PAS* overestimate the precision and recall?

(III) The author mentions that PASs of highly expressed genes tend to be missed by the 3'-seq: why?

(IV) The author considers PASs in intergenic regions to be a consequence of pervasive transcription? If so, are these transcripts spliced as well?

(V) It might be interesting to speculate how the analyses could be improved once long-read data e.g. from nanopore sequencing become available.

**Provisional Recommendation**

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

□ ~~*I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*~~

□ ~~*The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*~~