Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

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 D. Pervouchine

The thesis document includes the following changes in answer to the external review process.

I would like to thank all the jury members for their reviews and comments. The changes made to the text and the answers to the comments are reported below.

Professor Dmitry Ivankov

1. P. 22: "... protein primary sequence." This should be corrected either to "... protein primary structure" or to "... protein amino acid sequence".

Thank you for pointing out the mistake. The correction was introduced to the text on p. 23.

2. Sometimes terms are linked to its position in the Glossary, sometimes not. I did not understand the pattern, so maybe to consider making the terms in the same format throughout the text?

Thank you for your observation. In sections where the use of acronyms, such as "PAS", is particularly frequent, highlighting each occurrence made the text visually overwhelming. My general approach was to highlight the first appearance of an acronym in each section. Following your comment, I have reviewed and adjusted the links throughout the text to adhere more consistently to this rule.

Professor Oleg Gusev

One part I really missed in the dissertation is an idea about data represented in Table 5.1. The number of PAS in protein-coding sequences greatly exceeds the number of genes, and there are hundreds of thousands of cases of PAS in intergenic regions. What are they? Is it the polyadenylation "noise", like one we see in transcription starting sites modules, or something functional? Maybe it an evidence of new overlooked mini-genes? I hope to be able to discuss it during the defense Q&A session.

Thank you for your interest and observations regarding the data in Table 5.1. The high number of PAS in protein-coding sequences, exceeding the number of genes, aligns with the fact that about 70% of genes have multiple polyadenylation sites. As for the intergenic PAS, while their absolute number is high, their density (number of PAS per kilobase) is very low. Several studies in plants, yeast, and mammals have also reported tens of thousands of intergenic PAS and found that a significant fraction of them are located within several kilobases of known transcripts. These PAS could be associated with antisense transcripts or long 3'-extensions of known genes, as noted by Ozsolak et al. (Cell 143, 2010) and Miura et al. (Genome Res. 23, 2013). Indeed, some intergenic PAS may represent the ends of novel functional transcripts, as indicated by Wu et al. (2015). Alternatively, these PAS might originate from non-functional transcripts near enhancers or transposable elements, as suggested by Ozsolak et al. (Cell 143, 2010) and Wang et al. (Genome Res., 2018), respectively. They could also arise from background or pervasive transcription. Furthermore, it's possible that some intergenic PAS are artefacts due to analysis issues like misalignment. I look forward to discussing these possibilities further.

Professor Mikhail Gelfand
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).
I have included a list of attended conferences on page 4 and corrected the misprint.
(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.
Thank you for highlighting the discrepancy. I have added a paragraph in Background section 2.1.1 (page 20) that comprehensively describes various sequence elements around functional polyadenylation sites. This addition specifically links the motifs displayed in the figures to their corresponding discussions in the text.
(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".
The correction was introduced to the text.
(3) The signal in the legend to Fig. 5-6 is used without definition (defined much later in the text)?
The term is indeed defined earlier in the text on page 60 (section 5.1.2), prior to Figure 5-6.
(4) Why calculating local GU-content is relevant to the identification of a "U-rich region" (p. 67)?
Indeed, the mentioned statement is confusing. To clarify, the downstream sequence element is a GU‑rich sequence containing GU‑ or U‑repeats. Both motifs were shown to be recognised by the CSTF complex (Zarudnaya,M. et.al. *NAR* 31 (2003), Tian,B. et.al. *NAR* 33 (2005)). A correction was introduced to the text: the "U-rich" was substituted with "GU/U-rich" on p. 65.
(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).
Skewed to the right is a less confusing term. A correction was introduced to the text on p. 79: *"skewed towards positive values"* was replaced with *"skewed to the right"*
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).
Thank you for the insightful suggestions. The concept in question was originally proposed by A. Gruber and M. Zavolan in their 2019 review paper (Nat. Rev. Genet., 2019). This idea primarily stems from observations by Sheets et al. (NAR, 1990) and other groups, which indicated a correlation between the frequency of polyadenylation signal hexamers in the genome and the cleavage and polyadenylation efficiency at the corresponding PAS.
In response to your suggestion, I explored research testing this hypothesis and discovered a relevant study by Kainov and Aushev et al. (*GBE*, 2016). Their analysis of human polymorphisms and interspecies divergence showed that selection acts on both consensus and non-consensus PAS hexamers, albeit is weaker on the latter. The authors also examined SNPs that disrupt, impair, or improve cleavage efficiency. Notably, substantial negative selection was observed against disrupting SNPs. The results for impairing SNPs were less conclusive, with no evident negative selection against SNPs that improve CPA efficiency. The latter

finding challenges the assumption of strong conservation of the non-consensus nucleotides in AAUAAA-like sequences.

Accordingly, I have updated the relevant paragraph in my manuscript (p. 19) to include these findings.

(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?

Indeed, restricting the analysis to PAS located in genes with at least one RNA-seq-derived PAS and at least one 3'-seq-derived PAS could inflate the precision and recall metrics when comparing these two datasets. Conversely, including all annotated genes in the comparison of a PAS set against transcript ends would underestimate precision and recall. This is because many genes are not expressed in B cells, and hence their transcript ends could not be represented in our dataset.

I performed three pairwise comparisons: 3'seq vs GENCODE, RNA-seq vs GENCODE, and 3'-seq vs RNA-seq. To ensure a uniform basis for comparison across all three datasets, I chose to use the intersection of gene sets containing RNA-seq-derived PAS and 3'-seq-derived PASs.

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

The observation that many highly expressed genes do not appear to contain PAS captured by the 3'-seq protocol, as illustrated in Figure 5-2B, can be attributed to several factors. The primary one arises from the analysis of the 3'-seq dataset. Due to the specific limitations of this dataset, it was challenging to reliably identify the genomic strand of the candidate PAS. Consequently, as described in Methods 4.3.1, I inferred strand information based on the gene harbouring the PAS. This approach resulted in the exclusion of genes overlapping with genes from a different strand, impacting about 5,000 out of 20,089 protein-coding genes. Therefore, these excluded genes could not contain any 3'-seq-derived PAS. Additionally, patient variability could play a role, as cells from only 9 out of 15 patients were used for both RNA-seq and 3'-seq libraries, leading to potential differences in gene expression patterns. Protocol differences, with samples processed under varying conditions, might also contribute to this discrepancy.

To provide a clearer explanation, I have revised the discussion of these reasons on page 98 of the manuscript.

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

The PASs in intergenic regions could be associated with antisense transcripts or long 3'-extensions of known genes, as noted by Ozsolak et al. (Cell 143, 2010) and Miura et al. (Genome Res. 23, 2013). A small fraction may represent the ends of novel functional transcripts, as indicated by Wu et al. (2015). Alternatively, these PAS might originate from non-functional transcripts near enhancers or transposable elements, as suggested by Ozsolak et al. (Cell 143, 2010) and Wang et al. (Genome Res., 2018), respectively. As suggested, they could also arise from background or pervasive transcription. Finally, some intergenic PAS are probably artifacts due to analysis issues like misalignment.

In this study, I did not specifically examine the obtained intergenic PAS or splicing patterns around them.

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

The ability of long-read libraries to show entire isoforms makes these methods valuable for the analysis of interactions between RNA processing events. For instance, data from long-read sequencing of nascent transcripts have demonstrated a link between the splicing of the terminal intron and the cleavage and polyadenylation of pre-mRNA (Reimer,K.A. et.al. *Mol. Cell* 81 (2021)) Additionally, other researchers have employed targeted long-read sequencing to illustrate the coordination between cassette exon splicing within the gene body and alternative polyadenylation in the 3'-UTR (Zhang,Z. et.al. *Nat. Comm.* 14 (2023)).

In the context of this work, long-read data would be very helpful in the identification of the type of alternative terminal exons associated with the intronic PASs (composite or skipped). However, such data would not help to capture spliced polyadenylated introns, since long-read protocols tend to be less sensitive and can miss the RNAs with low concentrations. Moreover, there are no long-read RNA-seq datasets with total coverage depth of such magnitude as GTEx.

Professor Ivan V. Kulakovskiy

(1) The text is overflown by acronyms and their usage could have been more careful, e.g. the title of section 5.2.4. is a certain pinnacle: "Abundance and tissue-specificity of STE, CTE and SPI iPASCs".

In response to the comment, I have revised sections 5.2.3 and 5.2.4 and reduced the number of acronyms. Furthermore, I have changed the title of section 5.2.4 from "Abundance and tissue-specificity of STE, CTE, and SPI iPASCs" to "Abundance and tissue-specificity of alternative terminal exons and spliced polyadenylated introns."

(2) The list of publications (page 4) misses certain metadata of the applicant's papers, such as page numbers and DOI identifiers. Also, most of the thesis text is written from a first-person
point of view. I expect that all the research described in the text was indeed fully performed by
the applicant, but it would be beneficial to explicitly state the author's contribution to each of the publications.

Thank you for pointing out the missing metadata. I have updated the publication list on page 4 to include page numbers and DOI identifiers for each paper. Concerning the use of the first-person narrative and my contributions to the cited publications, here is a detailed account:

- "Transcriptome sequencing suggests that pre-mRNA splicing counteracts pre-mature intronic polyadenylation": My role encompassed the entire data analysis, including identification of PAS from GTEx RNA-seq data, quantification of tissue-specific PAS usage, processing of IPSA pipeline output for tissue-specific splicing quantification, and correlating tissue-specific patterns of alternative splicing with intronic polyadenylation.
- "Tissue-specific regulation of gene expression via unproductive splicing": I was actively involved in the development and implementation of the IPSA pipeline utilized in this study.
- "RNAcontacts: A Pipeline for Predicting Contacts from RNA Proximity Ligation Assays": It is important to note that the results from this paper are not presented in the thesis.

(3) The Introduction lacks literature references for key claims to better substantiate the setting of
the study.

The intention was to keep the Introduction concise and easily digestible. Thus, I focused on a broad overview without the detailed support of specific references. The detailed references were included in the Background section to support all the claims that were made. However, acknowledging the importance of substantiating key claims from the start, I have now added essential literature references to the Introduction as well.

(4) Page 31: it is unclear how particular precision and recall metrics relate to 'the best performance'. Were they selected based on e.g. an optimal F1?

The description of the APAeval benchmark study in the thesis is brief. To clarify, in the described paper, several tools for PAS identification were examined, every tool was applied to an RNA-seq dataset, and the obtained PAS predictions were compared against the annotated transcript ends. The best tool, TAPAS, was selected based on the precision and recall metrics. To address the comment, I have extended the description of the benchmark study on p.32.

(5) A review of PAS identification methods in Chapter 2 would benefit from a brief discussion of the applications of modern long-read and whole-transcript sequencing (e.g. Nanopore).

Thank you for the suggestion. In response, I have added a paragraph on this topic on page 34, enhancing the chapter with current methodologies in PAS identification.

In short, long-read and whole-transcript sequencing data, while useful for transcriptome assembly and identifying transcript ends, are not typically used for de novo polyadenylation site identification. Recent studies, however, have started to explore their application in this area. For example, a tool for PAS identification from long-read RNA-seq data was introduced (Celik,M.H,Mortazavi,A. bioRxiv (2022)). However, its authors showed that 3'-seq is still more effective at capturing PAS, especially in genes with low expression levels, despite long reads having lower mapping error rates. Another study found that 3'-seq identified more PAS in introns compared to PacBio Iso-Seq (Shah,A. et.al. Genome Bio. 22(2021)). Additionally, long-read sequencing is valuable for analyzing interactions between RNA processing events and illustrating coordination between splicing and alternative polyadenylation (Reimer,K. et.al. Mol. Cell 81 (2021), Zhang,Z. et.al. Nat. Com. 14 (2023)).

(6) Throughout the text, the authors repeatedly refer to the 'expression' of polyadenylation sites, which might be a common jargon but to me seems strongly confusing and incorrect. I would suggest referring to PAS usage instead, as it seems widely accepted in the field.

Thank you for pointing out the use of 'expression of polyadenylation sites' in the text. While this term is sometimes used in the field (e.g., in the widely cited “quantitative atlas of polyadenylation in five mammals” paper by Derti et. al.), I agree that 'PAS usage' is clearer and more accurate. I have revised the manuscript and replaced 'PAS expression' with 'PAS usage' to ensure more precise and standard terminology.

(7) Page 43: it is unclear why different read mapping tools were used for RNA-Seq and 3'-seq data, as both STAR and HISAT can perform splicing-aware mapping.

Initially, both the 3'-seq and RNA-seq datasets were processed using HISAT2. However, to align with the GTEx-processing pipeline standards, the RNA-seq data was subsequently reanalyzed with STAR. This step was taken to ensure consistency with lab protocols. It is important to note that this reanalysis did not significantly alter the identified PAS set.

(8) Page 44, 1st paragraph: it would be useful to explicitly show the fraction of gene annotations unused due to overlaps. What do you mean by "all bookended" intervals (same paragraph)?

Thank you for the valuable suggestion. To clarify, I have added specific information on page 45 detailing the number of genes that do not overlap with genes from another strand: “Practically, I selected all annotated genes that did not intersect with a gene from another strand (14,922 out of 20,089 protein-coding genes).”

As for the term "bookended" intervals, it refers to consecutive intervals where the end of one interval coincides with the start of the next. I agree that it is not a widely used term, so I have replaced it with “adjacent”.

(9) Page 49: 1st paragraph, Qi is not defined; 2nd paragraph: how many shuffles were generated per sequence? Can we consider the ddG estimates to be stable in the case only a single shuffled sequence was used as a control?

The weighted interquartile range is initially defined at the start of the paragraph, starting with “For weighted IQR each PAS is weighted by its polyA read support”. Realizing that the subsequent reformulated definition of the weighted IQR (starting with “In other words”) caused confusion, I have removed this redundant explanation from the text on p.49.

As for the number of shuffled sequences used, one shuffled sequence was generated for each transcript end. While we cannot guarantee the stability of the ddG estimate for each individual case, our overall conclusion is based on a large number of independent observations (one for each transcript end). This approach ensures that the overall findings are reliable, even with the use of a single shuffled sequence for each transcript end.

(10) Page 52: there are explicitly provided normalization factors, but without extra information or a dedicated clarification it is impossible to judge whether they are adequate or not.

Thank you for pointing out the lack of information. To clarify, I have revised the paragraph (p.52) to explicitly reference the relevant results section (5.2.1) and provided additional information to justify the used normalization factors.

“In 5.2.1 I evaluated the CPA rate, at which it acts on the nascent pre-mRNA. To account for the bias arising from intron degradation, I normalized the polyA read count to the average read coverage in exons and introns. To estimate the read coverage in constitutive exons, alternative exons, and introns, the total read coverage values per nucleotide in GTEx samples were averaged between windows (wi and we) located in the respective regions, resulting in the normalization factors of $3.3*10^6$, $3.2*10^6$, and $8.0*10^4$, respectively.”

(11) Page 54: if you consider the impact of PCR duplicates, why duplicate filtering (e.g. with Picard) was not performed explicitly?

Thank you for your suggestion regarding duplicate filtering. I filter the PAS set by entropy because our confidence in each PAS correlates with the diversity of the polyA reads supporting it. This diversity can be quantified by the entropy of the overhang lengths distribution. Filtering by entropy inherently removes duplicates and makes duplicate elimination with a specific tool unnecessary. I have realized that this aspect was not sufficiently emphasized in the text, so I have now expanded the rationale behind the entropy threshold on page 54 to enhance clarity.

(12) Page 55: how and when did you filter PAS in adenine-rich regions? How these regions were defined in the first place? These details seem to be missing from Methods.

Thank you for highlighting this issue. To clarify, I have defined adenine-rich regions as genomic loci with a sequence of 10 consecutive adenines or 10 consecutive thymines. All candidate PAS found within these regions were excluded from the analysis for both the 3’-seq-matched and GTEx RNA-seq datasets. To provide a comprehensive understanding, I have now described the process of identifying and filtering out PAS from adenine-rich regions in the Methods subsection about PAS identification from RNA-seq data (Section 4.3.2, page 45).

(13) Page 56: ".. genes .. to be expressed in .. datasets". Genes are probably expressed in cells, not in datasets.

A correction was introduced to the text on page 56: "Since the samples were matched, it was expected that the 3'-seq and RNA-seq datasets would contain reads from the same genes."

(14) Page 64: by IQR of PAS here you mean the IQR of PAS borders or locations, right?

A correction was introduced to the text on p. 63: "... the IQRs of PAS …" was substituted with "… the IQRs of PAS positions …"

(15) When switching from a small-scale 'benchmarking' study of PAS found with RNA-Seq versus 3'-seq to a large-scale study with GTEx data, there were more stringent filters introduced, which, for some reason, were not re-tested on a smaller-scale benchmarking data, although it is of interest to know if it could improve the precision/recall of PAS recognition achieved in benchmarking.

The question regarding the application of more stringent filters used in the large-scale GTEx study to the smaller-scale benchmarking data is indeed valid. Initially, these filters were not applied to the smaller dataset due to the already low number of PAS identified (about 32,742 PAS supported by multiple polyA reads), and the concern that further filtering would reduce this number significantly.

However, following your comment, I reapplied these filters to the RNA-seq samples from the matched dataset. As anticipated, the total number of identified PAS decreased substantially to 114,475, with only 18,237 supported by multiple polyA reads. This adjustment resulted in a slight increase in precision (by approximately 5%) but a more notable decrease in recall (by about 10%). Additionally, the number of genes with both 3'-seq-derived and RNA-seq-derived PAS was reduced from 3398 to 2741.