Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Mironov Aleksei

PhD Program: Life sciences

Title of Thesis: Tissue-specificity and regulation of aberrant alternative splicing

Supervisor: Prof. Dr. Dmitri D. Pervouchine

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

Prof. Dr. Oleg Gusev

At the same time, I believe the study's results should be more open to "an end-user". The applicant may consider introducing an original online resource so that data on alternative splicing will be easily accessed by the research community and used in future projects.

Response: It is, indeed, a good suggestion. I agree that the data should be easily accessible not only by bioinformaticians but also experimental biologists working on alternative splicing. However, building a separate web interface to specifically host TASS and unproductive splicing data would be redundant with respect to the visualization tool for the TASS database through a track hub for the genome browser, which has become a golden standard in the field nowadays. For the purpose of access to the tables related to unproductive splicing, I provided links to the Zenodo open access repository. These external links are sufficient for an end user to access all the information obtained in the Thesis.

Another point is that while the applicant used ENCODE data, there was no attempt to link the activity of specific regulatory elements (enhancer usage, alternative promotes) with aberrant splicing events. Why?

Response: Perhaps the Referee made his comment in regard to transcriptional rather than splicing regulatory elements. The impact of transcriptional enhancers on splicing is not well-studied and clearly falls beyond the scope of this study. In regard to alternative promoters, the only way to tackle these links is through attribution of TASS to 5'- and 3'-UTRs. To address this, I added the characterization of TASS positions within 5'- and 3'-UTRs. Specifically, I added Fig 5-17 and a paragraph about it in the text on p. 76-77. In case if the Referee also pointed to cis-regulating splicing enhancers and splicing silencers that are regulated by RNA-binding proteins (RBPs), I must note that only a handful of RBPs have an established sequence motif [PMID: 29883606]. Generally, such motifs seem to be very degenerate, and the determinants of their binding specificity are a subject of a separate research project that goes far beyond the scope of this work.

The quality of publications matches the standards of an international PhD degree. The only question is why a part of the dissertation (unproductive splicing) is not published or (at least) submitted to a journal.

Response: The paper related to unproductive splicing is now under review in the Nucleic Acids Research journal. The manuscript can be assessed in the biorxiv via the following link: https://www.biorxiv.org/content/10.1101/2022.07.03.498634v1

Prof. Dr. Philipp Khaitovich

The small points what could be considered by the author (not obligatory):

- page 20: “Approximately 95% of mammalian genes are susceptible to AS [20], which strongly influences transcriptome and proteome diversity [21, 22] and pro- vides additional layers of gene expression regulation [13].” – while this statement is certainly correct for the transcriptome, the effect of AS on the proteome diversity is not as extensive, as many splice variants do not affect the protein coding region.

Response: I agree. This statement is generally not correct, as the effects of AS on the proteome have been subject to debates in recent studies and there is no consensus yet [PMID: 27712956, PMID: 28483376]. I corrected the text on p. 20 accordingly.

- tissue-specific TASS: while definition of these events is clearly described in the methods and the examples are also clear, some of the patterns could be considered as tissue-dependent, rather than tissue-specific. The same seems to be the case for unproductive splicing events. This point does not affect the importance and correctness of results, on the opposite – it shows that splicing events with more complex patterns than single-tissue specificity are also included in the analysis.

Response: I agree that “tissue-specificity” in this project implies the variability of splicing over tissues rather than the exquisite presence in just one tissue. A similar approach to the definition of tissue-specificity was used in a previous analysis of NAGNAGs, a subclass of TASS [PMID: 22235189].

- More parallels of obtained results with the tissue specificity of other types of splicing events, for instance their prevalence in brain, skeletal and heart muscle, as well as their regulation specificity across tissues could have been mentioned in the Discussion section.

Response: I agree with this remark and add a paragraph related to the observed tissue-specific patterns of TASS on p. 112. As for USEs, the general tissue-specific patterns are specifically described and discussed in section 6.5. With these additions, the aspect of tissue specificity must be well-covered.

Prof. Dr. Ekaterina Khrameeva

The thesis is clearly written and I have few comments regarding its content and presentation of the results.

The description of used methods is provided with enough detail in general. Yet, I would appreciate it if the author could clarify several issues. First, regarding the catalog of annotated splice sites. It is written that they “were extracted from the comprehensive annotation of the GENCODE database v19 [174] and from UCSC RefSeq database [175]”. But what if these two databases were conflicting with each other? Or simply a union of annotated splice sites was taken here? I suggest adding more details on this procedure, to enable reproducibility.

Response: This is a good point. Both GENCODE and RefSeq are semi-manually curated databases which are prone to human errors. For some reason, which is not obvious to me, some RefSeq transcripts are missing from GENCODE and vice versa. Because of this, I took the **union** of the exon borders from the GENCODE v19 and UCSC RefSeq databases. A comment about it is added to the text on p. 39. In general, I find this is a rather technical issue, which is not worth discussing in much detail.

Second, regarding the read mapping procedure applied to the GTEx data. Was it a de novo mapping, not using information about annotated splice sites? I think this detail is important for the presented analysis and should be specified here.

Response: I agree that this particular detail is very important. To address it, I added a comment on p. 39 explaining that short reads from GTEx were mapped to the human genome using STAR aligner v2.4.2a in two-pass mode by the data providers, and this mapping uses the information about the annotated splice sites. STAR mapper not only uses the information on annotated splice sites but also identifies *de novo* splice

junctions. However, I emphasize that this procedure was done by the GTEx Consortium, and therefore refer the reader to the GTEx portal (https://gtexportal.org/home/methods), which describes the procedure in detail.

Third, regarding differential expression analysis procedure (page 46). It relies on each tissue against all other tissues comparisons.
Potentially, this procedure can be biased if tissues are not uniformly presented in the dataset. For example, if the dataset contains too many samples representing similar brain tissues (e.g., many cortical areas) with similar expression profiles, gene expression differences attributed to these brain tissues might be under-represented by this analysis strategy. I think this potential issue should be discussed in the thesis.

Response: This is, indeed, a valid concern deserved to be discussed. I added a paragraph discussing general patterns of tissue-specificity of TASS and the bias associated with it on p. 112. In the current methodology that is based on linear regression modelling over GTEx samples, I can not control for such a bias per se. As suggested by the reviewer, the abundance of brain-specific events can be underestimated due to the bias, because brain samples constitute the largest cohort in the GTEx panel. However, most identified tissue-specific TASS are brain-specific (Table A.9), in agreement with previous studies. Therefore, I suppose that the bias does not qualitatively affect the results.

Fourth, why marmoset and galago genomes were chosen in the evolutionary analysis (page 51)? Was their genome assembly and annotation quality sufficient? I think it should also be discussed in the thesis.

Response: I thank the Referee for this question. The quality of marmoset and galago genomes was more than sufficient, at least in what relates to protein-coding genes, which are analyzed in this Thesis. I refer the reader to the original paper by Denisov et al, which describes in detail the procedure of evolutionary selection analysis [PMID:24966225] and demonstrates its feasibility. No changes were made to the Thesis in this regard.

All results presented in the thesis seem to be solid. Yet, a few of them could be a bit improved. At page 85, the author writes that “significantly expressed miSS preferentially affects disordered protein regions, and tissue-specific miSS are found in disordered regions even more frequently (Fig 5-28).” Is there a way to estimate the statistical significance of the observed increase?

Response: I agree that the significance of this particular comparison is hard to see in Fig 5-28 (Fig 5-29 in the revised version). To address this comment, I add the result of the Fisher exact test on p. 86.

At page 89, it is written that “a significant positive selection was detected in maSS and constitutive splice sites (Fig 5-33, A, right).” I suggest marking this significant observation with a star (or stars) at the Fig. 5-33. It would simplify perceiving this result.

Response: I disagree with this comment. On p. 60 I stated that error bars in all figures and the numbers after the ± sign represent 95% confidence intervals. In the legend of Fig 5-33 (Fig 5-34 in the revised version) I additionally stated that the error bars denote confidence intervals for the ratio of two binomial proportions based on likelihood scores. In (Fig 5-34, A, right) the reader can clearly see that error bars for constitutive splice sites and maSS are above the horizontal line O/E=1 indicating a significant positive selection. Adding stars indicating significance over confidence intervals is generally not recommended, as it is redundant. Statistical inference at the 5% significance level can be made directly from confidence intervals.

At page 91, it is written that “miSS are considerably weaker (Fig 5-35, A).” Is this decrease significant? If yes, it should be also marked with a star in Fig. 5-35.

Response: I agree with this comment and add the result of a statistical test on p. 92 of the manuscript. However, I would prefer to avoid adding stars on the Fig 5-35 (Fig 5-36 in the revised version), because the

difference in strength is statistically significant (Mann-Whitney U test, p-value<10^-15) in every comparison: constitutive vs. maSS, constitutive vs. miSS, and maSS vs. miSS, but the effect size is remarkably different.

Prof. Dr. Maria Poptsova

Overall the text is clearly written and easy to follow, however here is I list my comments and suggestions to the presented text:

To me the subject of tissue-specificity should be discussed in more detail. For TASS sites this discussion is almost lacking. How the author explains relative abundance of AS events in one tissue compared to the other. Which tissues are more prone to aberrant splicing? The reader would benefit from obtaining the comprehensive view.

Response: I agree with this remark and add a paragraph related to the observed tissue-specific patterns of TASS on p. 112. As for USEs, the general tissue-specific patterns are specifically described and discussed in section 6.5. With these additions, the aspect of tissue specificity must be well-covered.

RBP analysis revealed some significant players associated with AS. It would also be beneficial to the reader if the author would summarize classes/types of RBP participating in all studied types of AS and highlight the corresponding pathways.

Response: It is a great comment, however I believe that this information is already present in the thesis. For TASS, we found only a handful of tissue-specific events with strong evidence of targeting by RBPs, and the most prominent cases represent the regulation by PTBP1. PTBP1-mediated regulation of TASS was further investigated on p. 80-83. The regulatory network of RBPs that tissue-specifically regulate USEs is presented in Fig 6-7, described on p. 104-107. Again, the most prominent regulator is PTBP1, and PTBP1-mediated regulation is extensively described in the text.

It would be also beneficial to present GO-enrichment analysis for genes detected as having AS events for TASS and USEs, both overall non-tissue-specific and tissue-specific.

Response: It is, indeed, a great suggestion. I performed a GO analysis of target genes and found a strong enrichment in the nuclear localization and chromatin organization. However, I found that target genes systematically contain more exons and often harbor several miSS. I therefore decided to match the set of target genes with background genes that have the same number of exons. GO enrichment analysis in GOrilla [PMID: 19192299] and DAVID [PMID: 35325185] did not return any significant association with ontologies. This is now described on p. 72-74 and 109-110 and Fig. A-4.