Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Mikhail Moldovan

PhD Program: Life Sciences

Title of Thesis: Heritable modifications of transmitted biological information as possible sources of adaptation

Supervisor: Prof. Mikhail S. Gelfand

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

1. **Prof. Peter Sergiev's review**

   1.1. Have you considered a possibility that massive A deamination might be a way to deal with probable dC to dU deamination in the genomes of coleoids? I don't know whether it is indeed the case. Sometimes nature uses "general" cure to overcome the consequence of massive damage to the genome happening in too many places, which make it impossible to correct via random mutational restoration.

   We did consider a possibility of editing being a general compensatory mechanism for rescuing deleterious dA-to-dG substitutions. This hypothesis is one of the general concepts regarding the role of editing and was formulated and discussed in the work of Jiang and Zhang, 2019, which is cited and discussed in Ch. 3.1 and 3.4.2. of the present thesis. However, we did not explicitly discuss the possible role of abundant dC-to-dU deamination in the generation of deleterious adenine states. The following sentence has been added to the second passage of Ch. 3.4.2: "In particular, deleterious G states may arise due to dC-to-dU deamination, which results in complementary dG-to-dA mutations and is known as one of the major mechanisms of mutagenesis in the human population (Seplyarskiy et al. 2021).".
   Regarding the Jiang and Zhang's concept, we did not find solid proof of its validity. As shown on Fig. 3C and discussed in the text, we observed only a slight tendency of coleoid A-to-I editing sites to originate from guanines more frequently than from cytosines or thymines. Although this result does not completely rule out the possibility of editing being a mechanism behind coping with enhanced mutation rates and specifically extensive dC-to-dU deamination, it shows that it is at least a very non-specific mechanism. Moreover, the elevated mutation rates would result in higher level of polymorphism in coleoid population, which we did not observe. On the contrary, we observed coleoids to be rather low-polymorphic (Ch. 3.4.2.).

2. **Prof. Olga Kalinina's review**

   2.1. ...I think a little more care could have been spent on proper definition of all terms (e.g., the word 'decodings' appears here for the first time, does not seem to be standard, and is never precisely defined).

   In Ch. 1.1., the word "decodings" has been changed to "expressed variants". In Ch. 2.2. the term is defined as "expressed states".

2.2. …one has to guess what 'E' means in the context of 'E-to-G and/or G-to-E transitions' – I imagine that means the edited adenines, but that is only my assumption.

The A/E distinction has been clarified in Ch. 3.2.5. by adding the following sentence: "A and E are the edited and non-edited states of adenines, respectively.".

2.3. …I have to point out a redundancy I spotted in the Methods section: in my opinion, sections 4.2.2 and 4.2.9 can be placed together and merged.

Chapters 4.2.2. and 4.2.9. have been merged into Ch. 4.2.2.

2.4. Also, I think Fig. 12B should be Fig 6B.

This was indeed a misleading figure reference which is now fixed.

2.5. However, from Fig. 13B I can see that mutations to lysine (a positively charged amino acid) are almost as prominent. I wish Mr. Moldovan could comment on this.

This is an excellent topic for discussion, and the following paragraph has been added to the text: "**5.4.4. Mutations to lysine.** As shown on Fig. 13, pS-to-K mutations are overrepresented in mouse and HMR datasets along with mutations to NCA. We see two possible explanations here: 1) This pattern may be due to the dynamics of basic contexts of clustered phosphosites. Indeed, if a phosphosite cluster possesses a basic context, the substitution of one of the phosphosites to a positively charged lysine should reinforce the context of neighbouring sites. This hypothesis is indirectly supported by pS-to-K mutations being overrepresented for clustered, but not for individual sites in the HMR dataset (Fig. 13B). 2) The most overrepresented mutation of phosphoserines relative to non-phosphorylated serines is pS-to-E. Serine is coded by TCN and AGY codon families, and glutamate – by GAR codons. N represents any of the four nucleotides, Y – cytosine or thymine and R – adenine or guanine. Thus, S-to-K a mutation would require at least two transversions: TCR (S) → GCR (A) → GAR (E)/TCR (S) → TAR (Stop) → GAR (E), or, alternatively, at least two transitions and one transversion: AGY (S) → AGR (R) → GGR (G) → GAR (E)/AGY (S) → AGR (R) → AAR (K) → GAR (E)/AGY (S) → AAY (N) → AAR (K) → GAR (E)/AGY (S) → AAY (N) → GAY (D) → GAR (E)/AGY (S) → GGY (G) → GAY (D) → GAR (E)/AGY (S) → GGY (G) → GGR (G) → GAR (E). Note that lysine is present in two of the paths, hence the overrepresentation of pS-to-K mutations may be due to the presence of intermediate stages of pS-to-E substitutions. Both explanations cannot currently be proven due to the lack of data, however this may be a subject for future investigations.".

Actually, we have investigated the validity of the second explanation by analysing sequences of codon substitutions along the phylogeny. However, our results were not statistically significant when a multiple testing correction was applied.

2.6. Also, the color coding in Fig. 13 is not explained in the figure legend.

Explanation has been added to the figure legend: "On both panels, blue and red dots represent statistically significant over- and underrepresentation of pSTY-to-X mutations relative to non-phosphorylated STY, respectively. Darker shading represents higher statistical significance.".

3. Prof. Yurii Aulchenko's review

3.1. I would recommend formulating the aim and objectives of the thesis explicitly

The aim and objectives have been formulated at the end of the first passage in Abstract: "The aim of the work described in the present thesis is to assess the possible role of heritable modifications and their features such as clusterization on the expressed variability and rates of adaptation. In particular, we study the abundant A-to-I editing in soft-bodied cephalopods and protein phosphorylation in mammals. Our objectives include: 1) Constructing a theory of evolution of heritable modifications, 2) Assessment of the effects A-to-I RNA editing may have on the adaptation of soft-bodied

cephalopods, 3) Assessment of the effects of protein phosphorylation on evolutionary patterns at modified sites, 4) Study of clusterization of A-to-I editing sites in soft-bodied cephalopods and phosphorylated amino acid residues in mammals and effects this clusterization may have on evolution.".

3.2. *In the introduction, at the end of section 1.1, the question is posed about "the optimal rate of production of novel variants". This question is not really addressed by the research performed and the results obtained*

Indeed. The question has been removed.

3.3. *Sorry for perhaps nitpicking, but I find the statement "evolution may be regarded as the loop of information with two steps: i. Genotype that is decoded and phenotype is produced, ii. Mutation, selection and drift influence the frequencies of phenotypes and hence alleles underlying the phenotypic values and the system returns to step (i)." somewhat incorrect. In the context, in (ii) it is only the selection that affects phenotypes, while drift and mutation affect genotypes*

The sentence has been rephrased, as suggested.

3.4. *It seems that you sometimes equate the transcriptome and proteome variability (e.g. the statement 3, "Clustered editing contributes almost a half to the total transcript and proteome variability generated by editing"). In general, the changes observed on the transcriptome level are far from being guaranteed to be passed to the proteome level. I would recommend that you avoid, or experimentally substantiate, this claim*

The sentence has been rephrased and now it explicitly states that the mentioned proteome variability stems from clustered non-synonimous editing: "Non-synonimous clustered editing in coleoids contributes almost a half to the total proteome variability and clustered editing in general contributes almost half to the transcriptome variability.". In addition, a statement in Ch. 4.3.1. has been rephrased to clarify the matter: "…we find almost half of this variance to be explained by correlated editing at pairs of sites, namely, up to 46.3% of the transcriptome variance and up to 46.5% of the proteome variance due to non-synonimous editing.".
However, these statements still presume that non-synonimous editing is faithfully translated to proteins, i.e., that A-to-I editing does not influence the translation of the sequence. This presumption is correct, as there is experimental evidence of non-synonimous editing being indeed faithfully translated into protein sequences. e.g. see Liscovitch-Brauer, 2017, Fig.2.

3.5. *I think some word(s) are either excessive or missing in the statement 5, "Clustered phosphosites have more acidic contexts and are substituted to negatively charged amino acids than individual ones"*

Some words were indeed missing. Now the statement is: "Clustered phosphosites have more acidic contexts and are substituted to negatively charged amino acids more frequently than individual ones.".

3.6. *For the sake of reader, the bibliography could have been formatted better by, e.g. having different indentation of the first and the next lines of each reference.*

Between-line intervals separating individual references have been increased.

4. Prof. Ekaterina Khrameeva's review

4.1. *The literature review seems somewhat excessive and though I have enjoyed reading it, I am not sure whether all the presented details are necessary. In several cases, it was difficult for me to trace the relevance of the presented material to the results described in the thesis. In my opinion, it would be better to exclude introductions from these chapters and incorporate them into the literature review instead. Perhaps it would help to tie these three papers together better.*

The presented literature review is more of a general theoretical overview of a scientific problem than introduction to the research. Thus, a lot of questions discussed in the literature review are beyond the scope of the thesis. That being the case, we deemed it better for the general theoretical framework connecting the presented research to constitute the literature review, with topic-specific introductions being in the beginning of each chapter.

4.2. *Regarding the presentation of the results, it is not clear what result the Fig. 1B illustrates. It is not discussed in the text. It is hard to see the differences between species because medians are not clearly presented. What conclusion is the reader supposed to make from this figure?*

The reader is supposed to make two conclusions. Firstly, the considered editing site sets have consistent distributions of editing levels. Secondly, that the editing levels are quite low on average, however there is a large number of strongly edited sites. Two sentences have been added: 1) In the caption of Figure 1B: "Note that editing sites have quite low (<10%) editing levels on average, however the distributions of editing levels across sites have heavy right tails, which correspond to large sets of heavily edited sites" and 2) In Ch. 3.2.1.: "The sets of editing sites for the four coleoid species are consistent with respect to the distributions of editing levels"

4.3. *The p-values corresponding to Fig. 2C are a bit confusing: p=10-22 is specified in the panel, p=10-33 is specified in the legend, and p<10-3 is specified in the text. Probably, the differences between them should have been explained better.*

P-value specified in the legend is a typo, which has been corrected. P-value in the text refers to two similar figures obtained from comparisons of two: Fig. 2C and Suppl. Fig. S6, as specified in the text. To avoid possible confusion, the sentence has been changed: "For NCES in both species pairs we have observed significantly more cases when the edited site in a pair is more structured than the unedited one while the control CES set shows no bias (binomial test p<1.1×10–3 for two species pairs: *O. vulgaris/O. bimaculoides* and squid/cuttlefish; Fig. 2C, Suppl. Fig. S6).".

4.4. *In Fig. 2D, it would be good to see confidence intervals in some form.*

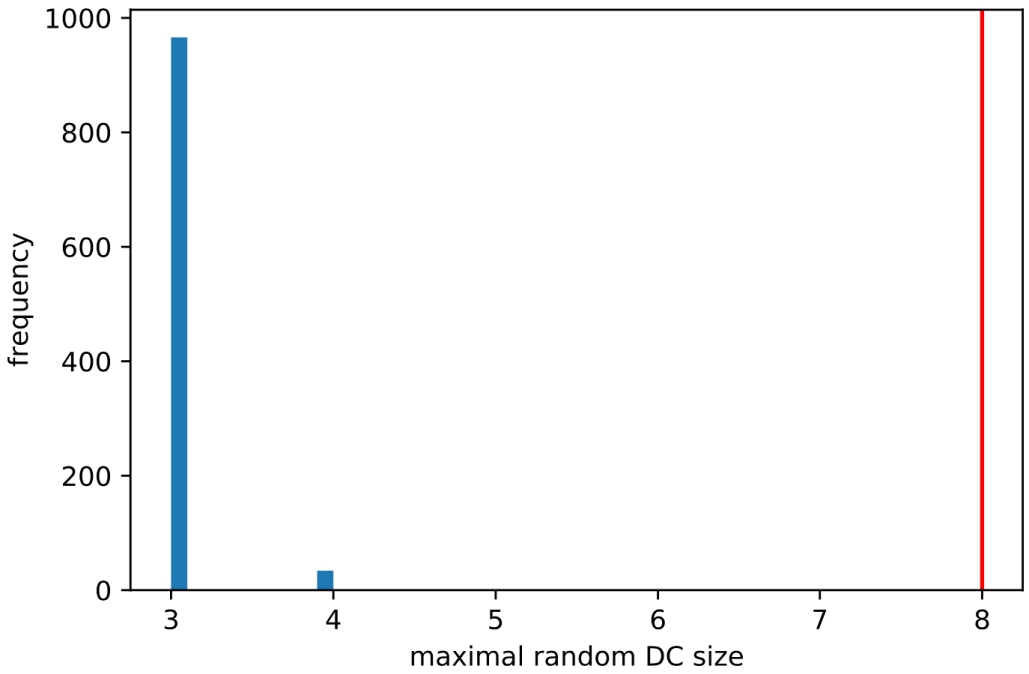
Confidence intervals for the rightmost trio of bars span from 0 to 1, and when the figure is scaled, minor significant differences between the leftmost bars almost cannot be seen. Thus, we have marked statistically significant differences with sets of asteriscs.

4.5. *Chapter 3.3.5 is probably more suitable for the Discussion.*

Chapter 3.3.5. has been moved to Discussion, now it is Chapter 3.4.2.

4.6. *In Fig. 7A, it is not clear how the random set of clusters was constructed. I believe the details of the procedure are important to obtain a correct result. Moreover, only two control sets were analyzed. I would suggest to make 1000 control sets here and perform a classical permutation test, to demonstrate that the effect was consistent and significant, and calculate the permutation p-value.*

As suggested, we have constructed 1000 control sets and assessed the stability of results portrayed on Fig. 7A, i.e. much larger dense cluster sizes in the actual data compared to the uniform expectation. The following passage has been added to the methods chapter 4.2.3: "To check for the stability of constructed controls with respect to the sampling of adenines, we have constructed 1000 control sets of adenines for *O. vulgaris* and calculated the maximal DC size in these sets (Suppl. Fig. S32). Maximal DC size was 3 in 97% of cases and 4 – in the remaining 3%. This shows that the procedure is robust with respect to random sampling variance.". And in the results chapter 4.3.2.: "The results obtained for the two control sets did not differ (Supplementary Figure S14) and the results obtained for multiple permutation rounds did not differ (permutation $p < 10^{-3}$, Supplementary Figure S32).". Also, we have added Supplementary Figure S32 which shows the maximal dense cluster size in a thousand permutations of a uniform expectation and compares them to the maximal dense cluster size observed in the respective species (*O. vulgaris*):

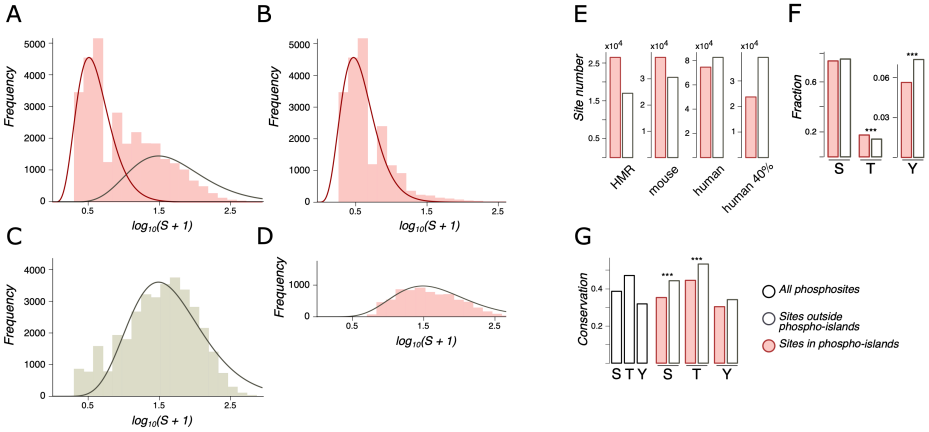


Supplementary Figure S32 | Maximal DC sizes in randomly computed uniform expectations (blue histogram) and the actual maximal editing DC size in O. vulgaris (red line).

The reason for such consistency of a general property of a random adenine sample may be the large numbers of editing sites, which are equal to the numbers of randomly sampled adenines by construction of the latter set.

4.7. It seems like Fig. 12B and D are missing Y-axes. Or are they the same as in panels A and C? Perhaps thin horizontal lines would help to clarify that

Figure 12 has been changed:



4.8. Why is the Chapter 5.3.3 title highlighted in red?

The coloring has been changed to black.

4.9. From Conclusions (Chapter 6), it is not clear how the mRNA editing story is tied together with the phosphorylation story. It should have been articulated better.

The following passage has been added to the beginning of Ch. 6: “Both A-to-I RNA editing and protein phosphorylation discussed in the present thesis heritably change the encoded information beyond the genomic blueprint. These modifications, while changing the expressed genomic

information in an organism, are also associated with divergent mutational patterns in populations in respective sites or, as is the case with coleoid A-to-I editing, with apparent positive selection at heavily modified sites. Both of these modifications also tend to cluster along the transcripts of proteins, thus adding to the expressed variability they generate.".

4.10. The thesis contains very few typos and grammar issues. I think I've spotted one at page 39, line 4 ("decoding yielding"). Also, there is a missing comma in Fig. 11A legend ("human, mouse and rat").

The mentioned issues have been corrected.

4.11. But what is the meaning of square brackets at page 39 ([neutral]) and page 38 ([function], [usage])?

In the first case, it is a clarification. In the second case, square brackets mark generic terms, which may be substituted with any specific terms in the provided sentence structure.

In the new version of manuscript, the table of contents, page numbers and the list of supplementary figures have been updated in accordance with implemented changes.