Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Sofia Medvedeva

PhD Program: Life Sciences

Title of Thesis: Natural diversity of CRISPR spacers

Supervisor: Prof. Konstantin Severinov

Co-Supervisor: Dr. Mart Krupovic

Chair of PhD defense Jury: Prof. Mikhail Gelfand ***Email**: m.gelfand@skoltech.ru*

Co-Chair of PhD defense Jury: Prof. Guennadi Sezonov

***Email**: guennadi.sezonov@sorbonne-universite.fr*

Date of Thesis Defense: 3 June 2019

Name of the Reviewer:

I confirm the absence of any conflict of interest	**Signature:**  **Date: 01-05-2019**

Reviewer's Report
The thesis essentially is a collection of papers published (in one case, submitted) in good journals, and as such there is no doubt in its high quality. The overall structure of the research project is well though-over; several marginal aspects are listed in the Annex. The quality of publications is very high, and the author has played a key role in most of them, while providing some essential contribution to the rest. The aims, listed in page 29, are logical and well-connected. The dissertation is written on a very hot topic, the diversity of CRISPR spacers in natural environments. At that, the study is more than purely descriptive, as the author has identified a number of new, important phenomena, the most interesting of which clearly is the conflict between closely related viruses mediated by mini-arrays. Another interesting observation is the apparent avoidance of PAM-like motifs in prespacers involved in the primed adaptation — this observation, made using experimenta data, has allowed the author to demonstrate that the primed adaptation plays a role in the evolution of CRISPR cassettes of bacteria in the wild. The paper collection is preceded by a well-written introductory section and followed by a conclusion where several directions for further research are outlined. I have very few substantial comments, rather

suggestions than criticisms, and some purely editorial ones; I list the later just to prove that I've indeed carefully read the manuscript.

Pages 24-25, point 2. A paper about reconstruction of *Yersinia pestis* phylogeny based on CRISPR cassettes is mentioned. In fact, as we've recently demonstrated, deletions in the cassettes very rapidly obliterate the phylogenetic signal, making reconstructions rather questionably (Bochkareva et al., Genome rearrangements and phylogeny reconstruction in *Yersinia pestis*, PeerJ, 2018).

Page 25, beginning of Section 8: in the same vein, while indeed CRIPSR cassettes reflect past viral infections, the catalog is highly incomplete due to deletions.

Generally, It is a wise policy to cite relevant papers by members of one's defence committee. Here is one (missed) opportunity. While indeed ref. 179 (Sorokin et al., AEM, 2010) has demonstrated the propensity of CRIPSRomes to reflect local virus populations (and hence this reference might have been mentioned to support the discussion in the first paragraph of Sec. 8), no clear correlation between spacers and protospacers was seen in the human gut microbiomes (Gogleva et al., Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs, BMC Genomics, 2014).

Page 41 mentions "fragments that must correspond to CRISPR arrays/array fragments that are either extinct or that have not been isolated yet in contemporary *E. coli*" — an additional possibility is, of course, that they are chimeras.

Page 54 says that "clustering-based subsystems" with unknown function are "related to houskeeping functions" — not necessarily; it could easily be, e.g., resistance or virulence.

Same page, beginning of the second column. I do not understand how addition of Arctic samples could make Antarctic samples indistinguishable from soil and mats. One possible explanation for this parados could be that the separation seen without Arctic samples would be seen if the third principal axis were considered. At that, visual analysis of PCA is not the best way to assess the existence of clusters; direct clustering methods would serve better.

Page 56, discussion of repeats with mismatches. Would such repeats be seen by the used primers? If not (or if yes, but with lower efficiency), spacers adjacent to such repeats would be underrepresented. This probably is a minor matter not influencing the overall conclusions, but still, it could have been taken into account (and I think this should be seen in the data, without additional experiments).

The analysis of PAMs (the last paragraph of the Results section, page 58) would be more instructive if not only the prevailing motifs were identified, but their cross-occurrence (NNAAAG in published genomes and NNATAT in studied metagenomes) were analyzed. The key question here is whether there is an absolute preference for the respective PAMs in these two datasets, or only a tendency. Similarly, it could easily happen that the lack of a strong signal in the env_nt sample is caused by the fact that it contains a mixture of several motifs. Without such analysis the conclusion that Antarctic and Northern hemisphere strains of *F. psychrophilum* evolved different PAM specificities (page 59) is somewhat premature.

The observation about the possible role of self-targeting spacers in preventing exchange between strains is interesting; one could also recall that related strains often harbor different sets of restriction-modification systems.

One of the results of Chapters III and IV is that different phages tend to be targeted by different CRISPR-Cas systems (pages 70 and 90, respectively; see also the first paragraph of page 167). It could be a nice direction of research, to check systematically, what features of phages (taxonomy, infection mode, etc.) are predictive of what systems would target these phages (if there indeed exist general correlations). The same applies to the conclusion of Chapter IV (page 90) about the preference of CRISPR types towards different types of mobile elements.

The prediction that “it is conceivable that complete viral genomes could be assembled using this approach, provided sufficient depth of CRISPRome sequencing and abundant CRISPR targeting” (page 156) is too optimistic: the obstacle would be the presence of many slightly different phage strains.

Page 165: “The results for SPV1 and SPV2 viruses were biased by super-abundant spacers from mini-arrays” — but could this bias be estimated, and hence the problem be resolved by using a separate set of virus-specific primers?

Editorial comments.

A note to Chapter I says: “Contribution: As stated in the paper”. However, the statement in the paper is merely that “S.M. and S.S. analysed data”. In other cases, though, the contribution by the author is described in sufficiently explicit terms.

Page 56 and legend to Figure 6: “transposes genes” and “IS110 family transposes” — probably, “transposases”.

The first paragraph in page 58 contains an m-dash (“—“, correct), an n-dash (“–“, incorrect) and an l-dash (“-“, awful). The late V.A. Uspensky turns in his grave.

Page 67: “fewer than two mismatches” — that is, one or none?

Page 72. The logic behind the conclusion that “the observed location of type III protospacers suggests that phages do exert pressure on Thermus communities, for in the absence of such pressure non-functional type III spacers targeting the non-transcribed strand of phage DNA could have been expected” is not clear.

Page 90, line 123: “local adaptation of the Sulfolobales viruses” — probably, not “of viruses”, but “to viruses”.

Page 153: “population from Beppu a thermal field in Beppu, Japan” contains a spurious repeat.

Page 154: “facets pf CRISPR arrays” contains a misprint, should be “of”.

Page 159: “3 independent events” — better, “three independent events”. Similarly, on page 162, “two spacers” would be better than “2 spacers”.

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*~~