Skoltech

Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Sofia Medvedeva

PhD Program: Life Sciences

Title of Thesis: Natural Diversity of CRISPR Spacers

Supervisor: Prof. Konstantin Severinov, Dr. Mart Krupovic

Chair of PhD defense Jury: Prof. Guennadi Sezonov, Prof. Mikhail Gelfand

Email*: sof.medv@gmail.com*

Date of Thesis Defense: 03/06/19

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

Dear jury members,
Thank you for the suggestions on how to improve my thesis and useful comments. Here are modifications I made in the final version of thesis text in response to your reviews.

Reviewer: Prof. Dmitri Pervouchine

- *"I have only minor comments regarding the use of abbreviations. For example, at the bottom of page 12, the term "tracRNA" is used without definition. Similarly, the term "PAM sequence" first occurs on figure 3 on page 13, but it is introduced in the text only on page 14. All the acronyms need to be introduced before the first use, because otherwise it creates difficulties for a reader with little experience in CRISPR-Cas systems (like myself). It always helps reading when the author presents the table of acronyms beforehand."*
  Answer: A list of acronyms has been added before the Introduction.
- *"Also, I believe that the information on how CRISPR spacers are recycled is missing from the introduction"*
  Answer: A paragraph about spacer turnover has now been added in the Introduction, in the section "CRISPR arrays".
- *"I noticed that the contribution section on page 42 says that the defendant and another coauthor analyzed the data. However, since this is a qualification work, it would be necessary to know which exact part of the data was analyzed by the defendant, and which analyses were carried out by the other coauthor."*
  Answer: I added more detailed contribution section before each chapter.
- *"However, the manuscript has no bibliographic reference and it would be good to know to which journal it is submitted. There is no reference in the list of publications on page 5 either. Perhaps the time that passed since the defendant prepared the manuscript allows to insert a valid bibliographic reference, or a letter of acceptance from a peer-reviewed journal."*
  Answer: The manuscript has been submitted in the beginning of April and is currently under review.
- *"A comment before this publication states that the defendant "obtained preliminary results of PAM avoidance in spacer sequences in primed adaptation experiments with different plasmids and lack of avoidance in CRISPRome mammoth data". From this remark it is not clear whether*

the analysis of the preliminary data is different from the analysis of the published data, and whether the defendant participated in writing the manuscript."
Answer: I added a more detailed contribution section before this chapter. Preliminary results were not different from the published data. No, I did not participate in the writing of the manuscript.

- *"I recommend the following changes to be implemented in the manuscript:*
  1. A careful revision of the introduction in regard of acronym use, possibly adding a table of acronyms.
  2. In order to highlight the contribution, I suggest a one-paragraph summary before each chapter in the Results section that would summarize the actual contribution of the defendant to that chapter, without references to display items therein, but rather explaining in plain language the analysis that was actually done."
  Answer: Thank you for the comments, list of acronyms and contribution summary have now been added.

Reviewer: Prof. Olga Soutourina

- *"The candidate has chosen extremely condensed and synthetic presentation of the scientific literature. On my opinion, the literature review appears too short and some additional information could be provided together with related illustrations to facilitate the reading of the manuscript before the description of the results of these studies."*
  Answer: I decided to use the condensed and synthetic presentation of literature, because each result chapter is preceded with a related and specific introduction within the text of the paper. I added a paragraph describing naïve and primed adaptation (section "Adaptation module") to facilitate the reading of Chapter VI.
- *"I have some concerns on the presentation of this results section. On my opinion, some additional sentences for transitions between different papers should be helpful to link these different pieces of work together."*
  Answer: Thank you for the suggestion. I have now added transition sentences before each chapter to link different chapters together.

Reviewer: Dr. Tamara Basta-Le Berre

- *"A minor issue that should be addressed before the defense concerns the general introduction chapter of the manuscript. Most of the figure legends in this chapter are incomplete and some are missing (exemple figure 5), this should be corrected before publication of the manuscript."*
  Answer: Thank you for your suggestion, I added more detailed figure legends.

Reviewer: Prof. Mikhail Gelfand

- *"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)."*
  Answer: Thank you for the reference, it now has been properly cited. I corrected the related text in the Introduction.
- *"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."*
  Answer: I added the comment in Introduction about deletions in CRISPR arrays.
- *"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)"*
  Answer: We found a strong correlation between spacer content and local viral populations for *Thermus* and *Sulfolobus* communities, described in Chapters III and IV. In agreement with your

comment, no correlation with local viruses was found in a CRISPRome study of human metagenome (Robles-Sikisaka et al., 2013), which can be a characteristic feature of this environment.

- "*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."*
  Answer: Although this possibility cannot be formally excluded, we tried to reduce the number of chimeras by filtering out low-abundance pairs of spacers.
- *"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."*
  Answer: Yes, clustering-based subsystems could be related to other functions. However, "virulence and resistance" genes form a separate group.
- *"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 paradox 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."*
  Answer: With addition of Arctic samples, two components described only ~35% of variance, which could not be significantly improved with a third axis. Probably, we should have tried other clustering methods.
- *"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)."*
  Answer: Repeats with a few mismatches are recognized with our primers, but with lower efficiency. This has been investigated experimentally in the course of the project described in Chapter I.
- *"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."*
  Answer: PAM sequence for Northern hemisphere strains is based on only 12 found protospacers, which might not be enough for cross-occurrence analysis. Additional analysis with recently sequenced genomes of *Flavobacterium psychrophilum* could provide further information on this issue.
- *"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."*
  Answer: Indeed, it would be an interesting direction of the research. In our datasets, we tested the correlation with a lifestyle of viruses and did not find it.
- *"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."*
  Answer: I agree that it is an optimistic, but not impossible prediction. All reconstruction methods designed for metagenomic data I tried were unable to assemble viral contigs with such high level of sequence diversity. I think that with more sensitive assemblers and higher spacer coverage the reconstruction of the genome is conceivable.

- *“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?”*
  Answer: In addition to four mini-CRISPR arrays found in genomes of SPV viruses, we identified several mini-array candidates in CRISPRome data. The majority of spacers from mini-arrays target closely related SPV genomes. As a result, CRISPRome spacers which target SPV are ‘contaminated’ by highly abundant spacers from viruses and cannot be used for studying *Sulfolobus* strain dynamics in enrichment cultures.
- *“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.”*
  Answer: More detailed contribution section was added.
- *“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”. “3 independent events” — better, “three independent events”. Similarly, on page 162, “two spacers” would be better than “2 spacers”.”*
  Answer: Corrected.
- *“Page 67: “fewer than two mismatches” — that is, one or none?”*
  Answer: Zero, one or two mismatches.
- *“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.”
  Answer: Unfortunately, these mistakes are in already published paper.
- *“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.”*
  Answer: The adaptation of new spacers by type III CRISPR-Cas system is not specific to the strand (transcribed or non-transcribed). We observed uneven distribution of spacers between strands, which can be explained by selective pressure of the phage.