Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Karyna Karneyeva

PhD Program: Life Sciences

Title of Thesis: Exploring type III CRISPR-Cas immunity in *Thermus thermophilus*

Supervisor: Prof. Konstantin Severinov

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

- **Prof. Yuri Kotelevtsev** did not have any questions.
- Comments from **Prof. Mikhail Gelfand:**

1. *Chapter 1 "Introduction", 1st paragraph: "The sophisticated relationships between prokaryotes and MGEs have intrigued researchers for over a century" _– _indeed? Who was that genius who had been intrigued by MGEs back in 1924?*

The phrase has been modified to be more precise. It now reads: "The sophisticated relationships between prokaryotes and MGEs have intrigued researchers since the middle of the 20th century".

2. *Page 80: 50% phages in supernatants of collapsed cultures contained intact protospacers. Does that mean that phages with intact protospacers still could overcome CRISPR-Cas and proliferate**?***

It is unlikely that phages containing intact protospacers could overcome CRISPR-Cas defense. The presence of reads corresponding to a full-length protospacer could be explained by the presence of initially added viral particles in the media that could not efficiently infect bacteria, which is why they retained intact DNA in the capsid and remained in the supernatant. The infected culture did not collapse completely, indicating that the efficiency of escaper phage proliferation is not as high as for the wild type, and the total amount of reads corresponding to escaper DNA is not as abundant as one might expect.

3. *Page 83, the last paragraph: is faster elongation rate of the phage RNA polymerase an established fact or a conjecture? In the formed case a reference should be provided; in the latter case it might be better to reformulate in softer terms.*

The faster elongation rate of the PhiFa phage RNAP was not established. Thus, the phrase was modified: "Due to co-transcriptional Type III targeting, the efficiency of CRISPR-Cas immunity could vary depending on the activities of the host's or viral RNAP. For example, differences in elongation rates or specific features of the transcription bubble can influence Type III interference."

4. *Page 115: "We assume that the free energy of binding of a pair of complementary nucleotides ΔG is additive and identical for all complementary base pairs" — how valid is this assumption? Is it crucial in the sense that the modeling results are robust with regards to the value of ΔG?*

While considering the difference between A-T and G-C pairs and the lack of additivity might offer more precision, in our case, rounding the final calculation to an integer number of nucleotides makes this distinction irrelevant. We simplified the model for this reason. However, this model may not be suitable when the number of nucleotides is significantly higher or the nucleotide composition is heavily skewed towards to G-C or A-T pairs.

5. *Page 122: "Our findings demonstrate the effectiveness of Type III-A and III-B CRISPR-Cas systems in protecting T. thermophilus cells from phiFa and phiKo phage infections solely" — I'm not sure I understand this sentence. In what sense "solely"?*

The phrase was modified and the revised sentence reads: "Our findings demonstrate that Type III-A and III-B CRISPR-Cas systems can each independently protect *T. thermophilus* cells from phage infections".

6. *A more detailed review of their variety might be beneficial (in particular, the author does not mention toxin-antitoxin systems that were one of the first discovered suicidal systems). Similarly, non-defense functions of CRISPR-Cas systems deserve more attention.*

The literature review was updated.

7. The author mentions the site of proteomic analysis (para. 3.19) and some sequencing analysis (para. 3.12, but without data on the fragment and read lengths), but not other sequencing rounds (para 3.13; 3.14, no detail; 3.16 and 3.17, Oxford Nanopore – _where?).

The information was added.

8. Chapters 4 and 5 are preceded by acknowledgements of the colleagues' help with the bioinformatics, proteomics, and modeling analysis; these notes *should be more detailed (what analyses have been performed by whom?).*

Notes were adjusted.

9. The Conclusions are clear, although a bulleted list might be useful to understand what particular results the author selects as the most important ones.

A bulleted list was added.

- **Prof. Edze Westra** did not have any questions.

- Comments from **Prof. Olga Soutourina:**

1. *Page 17. Please check for the new phage family nomenclature.*

Changed for *Caudoviricetes* and *Tectiliviricetes* classes. The literature review was also updated. However, the term "siphovirus" was kept as morphological (non-taxonomic) identifiers such as "podovirus", "myovirus", or "siphovirus" are still recognized, even though they do not have any formal taxonomic meaning or significance.

2. *Page 24. The sentence "involved in various processes" should be improved, please specify the processes.*

The phrase was modified and the revised sentence reads:
"Further, it has been demonstrated that CRISPR-associated (*cas*) genes encode proteins that are essential for adaptive immunity. At the interference stage of CRISPR-Cas immunity, Cas proteins help recognize and destroy the genetic material of the invader. At the adaptation stage, Cas proteins capture and incorporate fragments of invader's genetic material into the host genome. These fragments are then used to provide immunity against future attacks by the same or similar invaders. Cas proteins are also involved in processing of CRISPR array transcripts, generating CRISPR RNAs (crRNAs). These crRNAs guide Cas proteins to their targets."

3. *Page 26. "Approximate positions range between 6-12 base pairs", please check the location in nucleotides for the distance?*

The phrase was modified and the revised sentence reads: "Approximate positions range usually between 5-10 base pairs".

4. *Page 26. Please check the reference citation format throughout the manuscript: for example (35-38) instead of (35), (36), (37), (38).*

Changed as suggested throughout.

5. *Page 27. Please check the reference (44), is it Science 1979 paper?*

There was an issue with automatic reference generation where all science articles had the additional "(1979)" added to the name of the journal. However, the rest of the data, including article names, authors, and years of publication, was correct. This has now been corrected.

6. *Page 32. Please replace "contradicting" by "contradictory"*

Changed.

7. *Page 40. "to the conclusion:" please remove ":"*

Changed.

8. *Page 41. "ATP-to-cOAs" please add "conversion"*

Changed.

9. *Page 42-43. More Acr to cite now, please provide a citation/link to the complete updated list of acr, and also briefly indicate the strategies for their identification.*

The paragraph was revised.

10. *Page 46. The gene/mutation names to put in italic in the E. coli strain genotype description, please check*

Changed.

11. *Page 46 Please avoid using "rpm" that depends on the apparatus, rather use more universal "g" for centrifugation parameters.*

Changed.

12. *Page 46 and throughout the M&M section OD600 ("600" in subscript format)*

Changed.

13. *Page 48-49. Please include the abbreviation for antibiotics used as well as concentrations to specify once in the growth conditions part.*

Changed.

14. *Page 49. Absorption test, please check, we are using a different protocol with a large excess of bacteria over phage particles for quantification of not absorbed phages remaining in the supernatant.*

In our case, we used an absorption test to determine the time it took for cells to become infected after phage addition, in order to select appropriate time points for the RNA sample collection. This is why we counted infected cells rather than the number of phages that were absorbed. However, for single-step growth curves, we did indeed test infected cultures with a low MOI to determine the burst size of the phages (data not shown).

15. *Page 50-51 and throughout the manuscript "0,2" please use a comma for decimal numbers.*

Changed.

16. *Page 51. Please specify the sequencing technology and the kit used for library preparation.*

Details of sequencing technologies, with data on the read lengths, are added throughout the chapter.

17. *Page 61. Even if it was not a focus of this study, a more detailed description of type I-B and type I-C CRISPR-Cas components could be provided on the figure 4.1*

A description of the Type I-B and Type I-C CRISPR-Cas components has been added to the introduction of the chapter.

18. *Page 62. In the sentence "present in type III CRISPR arrays and not in active type I CRISPR arrays" please specify if both type III and type I are active.*

The introduction to chapter 4 was adjusted as follows: "Prior research in our laboratory has revealed that both the I-B and I-C systems are active and provide defense against phages". To avoid misunderstanding, the word "active" was removed from the phrase "present in type III CRISPR arrays and not in active type I CRISPR arrays".

19. *Page 86 Please correct "we refer to these strains are refereed to as", "located downstream of a downstream of a"*

Changed.

20. *Page 86 the promoter name should be in italic*

Changed.

21. *Page 97 Please include the full-name for « intensity-based absolute quantification (iBAQ) signals »*

Changed.

22. *Page 122. The conclusion part seems rather short and could be completed with more detailed description of future directions and summary figure.*

The conclusion part was adjusted. Summary figures (Figure 6.1 and Figure 6.2) were added.

23. *On my opinion the addition of figures describing different cycles of phages (it would be interesting to evoke alternative cycles), diversity of defense systems and general features of CRISPR-Cas systems including for example classification and PAM definition before focusing on type III CRISPR-Cas features would be helpful for the readers.*

The literature review was updated. New figures (Figure 2.1, Figure 2.2, Figure 2.3) and a new table (Table 2.1) were added.

- **Prof. Raymond Staals** did not have any questions.

- Comments from **Dr. Robert Fagerlund:**

1. *It is stated that Thermus thermophilus is a convenient model for Type III CRISPR-Cas research. What are the pros and cons of this system?*

Thermus thermophilus offers several great advantages as a model system, including its natural competence for DNA uptake, which allows for easy genetic manipulation. Additionally, its thermostability makes it convenient for procedures involving protein manipulation. Moreover, it is well-suited for studying Type III CRISPR-Cas systems since it allows us to compare subtypes III-A and III-B in the native genomic context. Furthermore, a number of distinct phages (other than phiKo and phiFa) are available to study defense mechanisms. However, there are also limitations to using *T. thermophilus*. For example, some genomic manipulations are relatively challenging, such as constructing strains with point mutations in domains of interest. Nevertheless, for future studies, methods for genomic engineering in *T. thermophilus* using endogenous CRISPR systems are currently being developed in our laboratory. Additionally, there is a lack of a wide choice of vectors, convenient inducible promoters, and other routine reagents that are stable at +70 ºC.

2. *Page 32 The thesis writes that type III CRISPR-Cas investigations have had contradicting results. Highlighting holes in the literature is important, but care is needed when describing past contradictory results with present-day hindsight. For example, early work on type III-A and -B systems did appear contradictory; however, these papers were published before the mechanism of accessory nuclease activation was revealed, which may explain many of these early results. Best to focus on what is known now.*

I agree that it is important to use past results with caution, especially when considering the context of early research conducted before key mechanisms were fully understood. The intention of this part was to discuss the history of Type III CRISPR-Cas research and to highlight the complexities and challenges associated with interpreting data from such multicomponent systems.

3. *Page 35 Note Csm and Cmr complexes are comprised of 5 and 6 different complexes. Total number of subunits per complex is often >10.*

The literature review was adjusted.

4. *Pg 35,37 There are two Palm domains but only one is active and has the GGDD motif.*

The literature review was adjusted.

5. *Chapter 2 would be considerably enhanced if it ended with a summary and direction towards the presented work. Why was this introduction important? What are the key questions this thesis addresses?*

The literature review was adjusted. A section "2.9 Summary" was added.

6. *It was not clear to me that both type III systems recognise the same repeat sequence. I don't believe this is common for co-occuring type III-A and -B systems, and therefore made understanding and interpreting some of these results difficult. It needs to be stated early that all arrays have the same repeat sequence (is this the case?) and both systems can use all arrays. If repeats are different, consider a table showing each sequence.*

The introduction to Chapter 4 was adjusted.

7. *Fig 4.1 Highlight the accessory proteins. An early mention of each one present should be stated. This is an important part of interpreting the results as it wasn't clear which immunity mechanisms were in play. Do both systems have adaptation genes?*

The introduction to Chapter 4 was adjusted. Six of eleven CRISPR arrays are shared by the III-A and III-B subtypes. *T. thermophilus* encodes two full-sized Cas1 proteins and one Cas2 protein. One of the *cas1* genes is located in the subtype I-B CRISPR–Cas locus, while the second one is located near the CRISPR-11 array, away from either Type III *cas* operons. Yet, it is this gene that is responsible for acquiring new spacers by Type III CRISPR arrays of *T. thermophilus*.

8. *Fig 4.3 The analysis of this result on pg 65 would be strengthened with quantification of the EOP. Without this, some subtle effects may be missed. For example, are spacers #33 and #26 lower than the control? It appears #36 has approx. >1,000x decrease in EOP – is partial the best term to describe that reduction?*

Our focus was primarily on qualitative rather than quantitative analysis. However, I understand the importance of quantification, and I agree that it could provide a more detailed understanding of the results. Regarding your specific questions about spacers #33 and #26, it is indeed challenging to discern subtle effects and accurately describe the magnitude of reduction without quantification of the EOP. The *T. thermophilus* strain bearing plasmid-borne mini-CRISPR arrays with spacer #36 certainly demonstrates a more than 1,000-fold decrease in EOP. I agree the term "partial" does not properly describe such an effect; it was removed from the text. Ongoing projects in our laboratory are aimed at delving into more detailed analysis, including quantitative measurements of the EOP.

9. *Fig 4.4 Are these essential genes? What would happen if non-essential genes were targeted? Knowledge of the accessory genes is important to understand this experiment. In the legend, what do the colours on genes represent?*

Gene 44 is indeed essential as it encodes the phage RNA polymerase (RNAP). On the other hand, gene 42 (described further in the thesis) encodes a hypothetical protein of unknown function and is non-essential based on our results of phage escapers analysis. Still, the *T. thermophilus* strain bearing a spacer matching gene 42 transcript is resistant to phage infection. Regarding the colors on the genes in Figure 4.4, they represent the types of the predicted temporal class. The figure legend has been updated.

10. *Pg 67 Why were the three genes selected as "strong Acr candidates"? How were the Acr expression plasmids constructed? I don't see this in the methods. Is expression induced? Are you confident proteins were expressed?*

Regarding the selection of the genes as "strong Acr candidates", we used a machine-learning approach for Acr prediction, as referenced in the thesis (ref. 143). Fa_ORF42, Fa_ORF45, and Fa_ORF47 obtained scores that allowed us to recognize them as promising Acr candidates. As for the construction of Acr expression plasmids, the corresponding genes were cloned into the pMK18 vector. However, I apologize for the lack of detail in the Materials and Methods section regarding this process. I have now adjusted the methods to include more information on the construction of these plasmids. Unfortunately, we do not have vectors with inducible promoters, so expression was constitutive in our experiments. I also acknowledge that I did not check the expression of Fa_ORF45 and Fa_ORF47. However, this part of the work is ongoing, and we aim to investigate the expression of these additional candidates in future experiments. We are also improving our experimental procedures to discover the mechanisms employed by the phiFa phage to evade Type III CRISPR-Cas defense.

11. *Pg 69 Typo in the figure for pBAD33.*

Changed.

12. *Fig 4.9 and 4.10 Nice work on the RNA sequencing to determine early, middle and late expressed genes. In my experience, BPROM can miss promoter sites and other tools, like that from Silas' group, can be useful. Are all promoter sites identified? I expect some analysis and correlation of promoter sites to the heat map. Is there a promoter in front of genes 4 and 18? Genes coloured lighter may better differentiate some genes.*

Indeed, the accuracy of phage promoter prediction can vary depending on the tool used. Further, phage genomes can have unique features, such as promoter sequences or regulatory elements, that are distinct from host elements sequences. As with any bioinformatics tool, it is important to validate the predictions experimentally. I have marked only promoters predicted by BPROM that were consistent with results of RNA sequencing. On the other hand, according to results shown in the heatmap, it is expected that there are polycistronic phage transcripts, for example, one starting from gene 4. Yet, no promoters were predicted in front of gene 4 (or mentioned gene 18). In summary, I agree w that BPROM may miss some promoters and will explore other tools to compare and identify additional promoters.

13. *Pg 74 and in discussion The thesis claims there is no correlation between temporal class and protection efficiency. Can you be certain when only one spacer from late and middle were tested?*

While we acknowledge that testing multiple spacers could provide a more comprehensive view, our RNA-seq data support the assumption that genes in question are predominantly transcribed in the middle and late stages of infection. According to the RNA-seq data, middle and late genes likely produce polycistronic mRNAs. This implies that targeting other middle or late genes will target the same mRNAs as the ones that have been targeted with already tested spacers. Thus, we do not anticipate significant differences in protection efficiency when targeting other genes from the same temporal class. Therefore, we believe that the results obtained for a single target corresponding to late or middle temporal classes are representative.

14. *Pg 75 Quantification of EOP would be useful for the comparison of WT to the mutant strains. The thesis writes that they have comparable efficiency, but a closer look could reveal other interesting observations. Is there a difference between WT and the single mutants? Are these systems having an additive effect? Does this add some weight to the reasoning for why there are two systems*?

As mentioned earlier, our focus in this study was primarily on qualitative rather than quantitative analysis. However, we agree that quantitative analysis can provide a more detailed understanding of the results. These experiments would be outside the scope of the current work but represent a valuable direction for future studies. Indeed, such analysis could provide additional insights into specific functions of Type III subtypes and potentially explain their dual presence in *Thermus*.

15. *Fig 4.17 Do the regions without mutations in panel C imply they are less important for targeting? How does this correlate with results from chapter 5?*

The regions without mutations in panel C of Figure 4.17 are indeed of interest, as it suggests that these regions are not sufficient for the recognition by the Type III CRISPR-Cas systems. These results align with findings from Chapter 5 of the thesis, where we determined that the minimal duplex length required for efficient immunity was 23 base pairs, while the maximal duplex length for escaper recognition from Figure 4.17 was 20 base pairs, allowing phage to escape immunity.

16. *Pg 83 Why is co-transcriptional targeting likely to be primarily directed towards nascent transcripts produced by the RNAP? What evidence is there that the single-subunit polymerase transcribes early genes? The cited paper refers to polymerase from an E. coli bacteriophage is expressed early to mediate late mRNA synthesis.*

Type III CRISPR-Cas systems are known to primarily target transcripts rather than ssDNA (ref. 76). However, we demonstrated that the activity of the HD domain and its ssDNA cleavage activity play a crucial role in Type III-A immunity. This implies that there must be ssDNA for the HD domain to act on. This could be the unwound ssDNA in the transcription elongation complex. Therefore, we hypothesized that there may be changes in Type III activity due to distinct features of viral RNA polymerase transcription. Regarding the evidence for single-subunit viral polymerase transcribing early genes, the cited paper referring to polymerase from an *E. coli* bacteriophage is not applicable to our study and this is a mistake. Instead, we assume that the phage single-subunit polymerase transcribes early genes as was shown for the closely related *Thermus* phage P23-45 (ref. 148). The reference was changed.

17. *Pg 86 typo repeated "downstream of".*

Changed.

18. *Fig 5.2. Are all repeats the same in the arrays? Could slight differences explain why some regions have no recombination? Are there spacer similarities between the "hot spots?*

While we did not specifically analyze the sequence similarity of repeats in the arrays, it is possible that slight differences in repeats (which do indeed exist) could contribute to the observed differences in recombination efficiency between different regions. Examining the similarities between repeats and spacers in "hot spots" and regions with no recombination could provide valuable insights.

19. *Pg 92 The percentage change in interference efficiency is recorded. Is this from EOP data? I recommend not referring to bases as residues as this could be confused with aminoacids.*

This is the relative transformation efficiency. The corresponding bars reflect the transformation efficiencies (based on CFU) for PS_dir and PS_rev plasmid-derivatives in comparison to the pMK18 control, as shown in Supplementary Figure S2. We believe that the suggested bars will not fully capture the effect of decreased colony sizes observed in "borderline" cases. Therefore, we submit that showing representative plates is more informative for our experiments. The term "residues" has been changed.

20. *Pg 94 The thesis concludes "The results support the fact that target abundance determines the degree of tolerance to mismatches", this is also discussed later. The experimental setup is complicated by the promoter on the bottom strand, which would result in duplex RNA forming and triggering a bacterial response to eliminate this. In hindsight, what features in your plasmid design would you include to manipulate target expression levels and prevent antisense expression?*

To address the complication caused by the promoter on the bottom strand, one approach could be to introduce in pMK18 terminators, that would terminate transcription before it reaches the region

complementary to the protospacer. Additionally, modifying promoter sequences to reduce their strength could also help manipulate target expression levels.

21. *Fig 5.5 CRISPR-Cas expression is typically tightly regulated to avoid fitness costs and autoimmunity. Would Cas levels be different in infected cells? I'm not sure how differences in subunit abundances correlate with stricter target complementarity requirements.*

Based on our RNA-seq data, we did not observe significant differences in *cas* gene transcription in infected cells compared to non-infected cells. Regarding your second point, according to our model, effective Type III immunity depends on two components: the quantity of bound RNA target and the energy of target binding to the effector complex. The ability to bind targets increases with the availability of more effector complexes. Even if these complexes have low binding energy and lead to less efficient cOAs synthesis or HD domain activation, their cumulative activity is sufficient to prevent infection or transformation.

22. *Pg 109. It appears the HD mutant still has some interference activity (about 10-fold reduction). How does this result compare to type III-A systems from other bacteria? Were the accessory genes still present in this setup? How does that impact data interpretation?*

The recent preprint "The Cas10 nuclease activity relieves host dormancy to facilitate spacer acquisition and retention during type III-A CRISPR immunity" showed that the ssDNase activity of type III-A systems from *S. epidermidis* plays a crucial role in re-growth of arrested cells, likely by degrading phage DNA and inactivating the immune response through cOA-dependent accessory proteins (https://www.biorxiv.org/content/10.1101/2024.02.11.579731v1.full.pdf). In our setup, the presence of accessory genes and their activation by cOAs likely contribute to the residual interference activity observed in the HD mutant. This highlights the complexity of the CRISPR-Cas system and the need for further investigation into the role of accessory genes in modulating CRISPR immunity, including through mutation of the Palm domain, among others.

23. *Pg 111 and discussion It is proposed that a decrease in plasmid copy number explains the decreased growth rate. But why is there a drop in copy number? Rather than an alternative mechanism of type III function, could it just be that interference is still occurring but at a slower rate that doesn't clear the plasmid?*

We agree that the decrease in plasmid copy number observed in cells from small colonies indicates that interference was ongoing but not effective enough to eliminate the plasmid. We addressed this point in the discussion to Chapter 5 of the thesis, emphasizing the role of continuous interference in the observed growth inhibition and the potential mechanisms underlying the decrease in plasmid copy number. Our results align with previous observations for Type III-A systems from *S. epidermidis*. There, it has been shown that Csm6 (an accessory RNase) is required for immunity in cases of inefficient DNA degradation by the Cas10-Csm complex. Under these conditions, Csm6 induces a growth arrest in the host and prevents further plasmid replication by degrading host and plasmid transcripts. In contrast, when DNA degradation by Cas10 is sufficient for anti-plasmid immunity, Csm6 is dispensable (ref. 84).

24. *The discussion on why Thermo has two systems was very interesting. Investigation into the role of the accessory genes would be important towards this end. Could differences in gene regulation also be important, where the systems are turned on by different stimuli? What is known about the transcription profile of T. thermophilus under different stresses?*

Investigation into the role of accessory genes could indeed provide valuable insights into the functional significance of having two systems, since we don't know to what extent accessory gene influence the immunity and how exactly they are activated and disactivated. Regarding your question about differences in gene regulation and whether the systems are turned on by different stimuli, we currently do not have direct evidence for this in *T. thermophilus*. However, it is an intriguing possibility that would be explored in future research. As for the transcription profile of *T. thermophilus* under different stresses, we have RNA-seq data for non-resistant phiKo-infected cells and have not detected drastic changes in host transcription.

25. *Pg 120 What is meant by "distinct pathways" in the type III-B system? Is this speculation of noninterference roles? Or it must rely more on the activation of accessory proteins?*

The term "distinct pathways" includes speculation that the Type III-B system relies on more efficient activation of distinct accessory proteins (or different sets of accessory proteins) compared to Type III-A systems.

26. *The convention for labeling CRISPR-Cas systems is to include "type" before the Roman numerals.*

Changed.

27. *Generally referencing literature is done well. It is important to remember to include the reference when stating a fact.*

Revised.

28. *Pg 29, Thesis has "in the presence of antisense transcription". Is this in reference to CRISPR array antisense?*

Yes, "antisense transcription" refers to transcription from DNA strand opposite the one encoding the CRISPR array. This phenomenon is relevant in the context of autoimmunity avoidance, as it could potentially lead to the production of CRISPR RNA (crRNA) that targets the host transcripts, resulting in autoimmunity.

29. *Check the spelling of cOA, in places it is cAO*

Changed.

30. *Ref 22 has first names.*

Changed.