Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Anna Maikova

PhD Program: Life Sciences

Title of Thesis: The CRISPR-Cas system of human pathogen *Clostridium difficile*: function and regulation

Supervisors: Prof. Konstantin Severinov;

Prof. Olga Soutourina, University of Paris-Saclay, France

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

Professor Harald Putzer, Paris Diderot University, France *Email:* putzer@ibpc.fr

Date of Thesis Defense: 30 September 2019

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

Dear jury members,
Thank you for the reviews and useful suggestions for my thesis improvement. Here are responses to your comments and thesis modifications.

Reviewer: Prof. Claire Janoir

- *The section on the pathogen she worked on during her PhD is very short and lacks detailed data on virulence factors and what is currently known on their regulation. To fully introduced her experimental work, Ms Maikova could have given information, as examples, on regulation of known or putative virulence factors by the second messenger c-di-GMP and on the suggested pleiotropic role of the RNA-chaperone Hfq protein in C. difficile.*

Answer: A new 1.1.3 section has been added to the Chapter 1. It describes *C. difficile* virulence factors and their regulation, including the second messenger c-di-GMP. Details on the suggested pleiotropic role of the RNA-chaperone protein Hfq in *C. difficile* have been added to the 1.1.4 section (p. 25).

- *In addition, importance of C. difficile adaptation during the infectious cycle should have been emphasized and illustrated by some examples.*

Answer: The example, describing the importance of *C. difficile* adaptation during the infectious cycle have been added to the 1.1.2 section (p. 20). Also, a legend was added to the figure 1.2 for the better understanding of *C. difficile* infection cycle.

- *Regrettably, the objectives, quickly mentioned in the abstract at the beginning of the manuscript, are not specified, but this can be easily corrected.*

Answer: The objectives have been duplicated in the Introduction parts of Chapters 2 and 3 (sections 2.1 and 3.1).

Reviewer: Dr. Ekaterina Semenova

- *Page 3, "in interference" should be replaced with "CRISPR interference".*

Answer: Corrected

- *Page 17, a reference on C. difficile database should be added. Is it correct that genomes of 3,000 C. difficile strains have been sequenced?*

Answer: References on databases, containing about 3,000 *C. difficile* genomes, have been included (p. 18).

- *"CrRNA" should be replaced with "crRNA" throughout the text.*

Answer: corrected

- *Page 33, Figure 1.9, it might be better to present more recent model for spacer integration indicating the first nucleophilic attack at the leader-repeat junction (Nunez et al., Molecular Cell, 2016). This model explains a polarity of spacer integration at the leader proximal end.*

Answer: Figure 1.9 has been replaced with a more recent model, proposed by Nunez et al., Molecular Cell, 2016.

- *Pages 10, 43, use "complementary" instead of "complimentary"; Page 40, replace "bought to the limelight" with "brought to the limelight; Page 40, replace "bought to the limelight" with "brought to the limelight".*

Answer: Corrected

- *Consider indicating PAM consensus as YCN rather than TCN/CCN.*

Answer: Corrected

- *Figures 2.13 and 2.14, please correct a description of a full cas operon mutant as "full cas operon deleted" instead of "full cas operon"; Page 66, replace "catalytic centers of the Cas2 protein are not required…" with a more precise "nuclease activity of the Cas2 protein is not required…"*

Answer: Corrected

- *Page 50 and Figures 2.5, 2.6, it was suggested that the efficiency of CRISPR interference provided by various arrays depends on the crRNA expression level. However, only RNA-seq profiles were presented. The candidate should be ready to discuss the correlation between interference efficiency and expression level in more quantitative terms*

Answer: Suggestions for the necessity of additional quantitative analysis of CRISPR array expression levels have been added to the Discussion part (section 2.4) of Chapter 2 and to the Chapter 5.

Reviewer: Prof. Mikhail Gelfand

- *Section 2.3.1.2 "The consensus sequences of selected PAMs were then visualized by the WebLogo tool (Crooks et al., 2004). This analysis demonstrated that the -4 position of the PAM does not play any role in C. difficile CRISPR-Cas system functioning (data not shown)." — Visual analysis of Web-logo is not sufficient to prove that a position is not relevant, statistical analysis of the distribution at that position as compared to non-PAM positions in the same context is required.*

Answer: The results on the -4 position of the PAM confirm data, obtained previously using other methods of PAMs investigation, including experimental approaches (Boudry et al., mBio 2015). I have modulated the sentence on the -4 position relevance and added this remark to the section 2.3.1.2 (p. 53). Therefore, additional experimental verification of every PAM position is required. I have added this perspective to the Chapter 5 (p. 135).

- *Figure 2.3. There are clear differences in the PAM frequencies in two considered strains (A and B);the given explanation "Different patterns of CCN and TCN PAMs distribution in 630Δerm and R202091 strains could be a consequence of the different amount of good-quality selected reads in the libraries "after"." Is somewhat superficial: why would filtering of reads affect positional nucleotide frequencies in PAMs?*

Answer: During the depletion analysis some of the "After" reads are filtered due to their bad quality. It could affect the final results and give the impression, that some PAM nucleotides are more depleted than

other, since they were more presented in the eliminated reads. Therefore, an additional experimental verification of every PAM position is required. I have added this perspective to the Chapter 5 (p. 135).

- *Section 3.3.1.4 — many of identified ORFs could be spurious, clustered due to residual sequence conservation in genomes of closely related strains. It might be a good idea to analyze positions of nucleotide polymorphisms in the candidate genes relative to the reading phase*

Answer: The goal of this section was to perform a general alignment analysis to find variants of small proteins adjacent to CRISPR loci in *C. difficile* strains. Several functional small proteins have been further characterized as toxins within type I toxin-antitoxin systems, while the function of other small proteins remains to be investigated.

Reviewer: Prof. Konstantin Lukyanov

- *Some histograms on conjugation efficiencies (Figs. 2.5A, 2.6A, 2.14) do not show dispersion.*

Answer: Dispersions have been added to figures 2.5A and 2.6A. Since the work is still ongoing and results will be complemented by futures studies (as it was indicated in the beginning of the Chapter 2), we need more repeats of the conjugation efficiency experiments with 630Δ*erm*Δ*CD2975-2982* strain. Therefore Figure 2.14 lacks dispersion.

- *Page 50: "Overall, interference levels of C. difficile 630Δerm CRISPR arrays correlate with their expression levels (Figure 2.5B, C), detected by RNA-seq in the previous work (Boudry et al., 2015)"(also in Discussion, page 66: "Experiments with C. difficile 630Δerm strain demonstrated that defense levels of different arrays correspond to their expression rates, identified previously (Boudry et al.,2015))". Actually, such correlations are not obvious from Fig. 2.5. RNAseq figures in the panels B and C are very small and contain no Y-axes marks (are they the same for all graphs?). The legend provides almost no explanation (e.g., what do black diamond symbols mean?). Among three low-effective arrays, only CRISPR17 shows low expression level, while CRISPR11 and 12 look quite similar to the efficient arrays CRISPR6, 7, and 10. Possibly, author can choose some other way of data representation (e.g., as a table or a graph for relative expression levels) to convince readers in any correlations between CRISPR array efficiency and expression level. I think that current description does not support it.*

Answer: RNA-seq pictures have been corrected (Y-axes were marked and the legends of figures 2.5 and 2.6 have been supplemented with additional information). Suggestions for the necessity of additional quantitative analysis of CRISPR array expression levels have been added to the Discussion part (section 2.4) of Chapter 2 and to the Chapter 5.

- *Page 64: "Adaptation experiments using native endogenous expression levels of Cas proteins were unsuccessful for C. difficile 630Δerm (data not shown)". It would be helpful for readers to describe these experiments in more details. In particular, what stimulus (exogenous DNA) was used to induce adaptation? Can a bacteriophage induce a stronger activation of naïve adaptation (compared to plasmid)?*

Answer: Adaptation experiments using native endogenous expression levels have been described (p. 60-61). For these initial experiments, plasmids carrying mutant protospacers were used. A bacteriophage can potentially induce a stronger activation of naïve adaptation and experiments with phages were discussed as a perspective of future studies in the Chapter 5.

- *Page 82: "For both TA modules (CD2517.1 and CD2907.1/CD0956.2), the overexpression of the toxins in strain 630/pT led to a significant increase in cell length for about 9% and 5.4% of the cells, respectively. The length of these cells was above the value of 630/p mean length with 2 standard deviations (10.5 µm) (Figure 3.5D and Figure S3.4 in Supplementary materials)." These Figures show microphotographs; it would be helpful to add some histograms for quantification of cell lengths in these populations (obviously, it is hard to see by eyes just 5-10% differences, especially when intrasample heterogeneity is high).*

Answer: Cell length was estimated for more than 100 cells for each strain and the percentage was counted to show the different levels of cell heterogeneity in strains' populations. An additional histogram has been added to the Supplementary part of the Chapter 3 (Figure S3.6).

- *Page 90: "When the CDIP634 strain grows in medium supplemented with ATc to induce the Ptet, intracellular c-di-GMP levels significantly increase (data not shown)". Why it is not shown? How*

it was measured? How large the increase was?

Answer: The explanation has been added to the text (p. 95). Intracellular c-di-GMP levels significantly increase due to the abundance of active diguanylate cyclase inside the cells. The elevated expression of the *dccA* gene was detected in the CDIP634 strain by qRT-PCR.

- *Histogram of RT-qPCR in Fig. 3.10 (page 93) does not show dispersion.*

Answer: Work on the regulation of *C. difficile* CRISPR-Cas system by c-di-GMP is still ongoing. We need more repeats qRT-PCR experiments. Therefore Figure 3.10 lacks dispersion.

- *A general question to the section 3.3.2 "Regulation of C. difficile CRISPR-Cas system by c-di-GMP": On the base of overlapping of the CRISPR12 array with a c-di-GMP-I riboswitch in genome, the author suggested a plausible model of regulation of CRISPR12 activity by c-di-GMP (Fig. 3.7). At the same time, further experiments showed "that expression of both cas-operons and CRISPR6, 12 and 16/15 arrays increased in high c-di-GMP level conditions (Figure 3.10)". As only CRISPR12 array overlaps with a c-di-GMP-I riboswitch, these data cast doubt on the proposed model for CRISPR12, suggesting some other general mechanisms of c-di-GMP-dependent expression regulation.*

Answer: The c-di-GMP-dependent riboswitch was found only close to the CRISPR12 of 630 strain, and it is a unique feature of this array (as it was designated in the section 3.4.2). Therefore, we supposed that this array could be directly regulated by c-di-GMP while other *C. difficile* CRISPR-Cas system components could be indirectly regulated by this second messenger, which has global effects on bacterial cells' physiology.

- *Figure 2.13 (page 64). It would be clearer to place both curves on the same plot to demonstrate its similarity*

Answer: The curves are highly similar, and they fuse with each other on the same plot, causing difficulties for their distinguishing. Therefore, they are presented on different plots.

- *In Fig. 4.1B dashed lines are somewhat misleading as they connect not corresponding ends.*

Answer: These lines were used to designate homologous recombination and this figure has been accepted in the final version of our article in Applied and Environmental Microbiology Journal.

- *Fig.4.4B (page 122): The minor tics on Y-axes mark 1/9 of the intervals that is quite unusual for the decimal system.*

Answer: Minor ticks have been removed.

- *Page 19: "The first ncRNAs (RNAs (tRNAs) and the rRNA ribosomal RNAs) were identified in the 1960s transfer in the 1960s." --> The first ncRNAs (transfer RNAs (tRNAs) and the rRNA ribosomal RNAs)were identified in the 1960s; Page 91: "The difference between CDIP634 and 630Δerm strains was clear only in the plasmid, carrying protospacer, corresponding to CRISPR12 spacer1, case." --> The difference betweenCDIP634 and 630Δerm strains was clear only in the case of the plasmid, carrying protospacer, corresponding to CRISPR12 spacer1."; Page 125: "The previous work, showed that CRISPR repeats …" --> The previous work showed that CRISPR repeats*

Answer: Corrected.