Skoltech
Skolkovo Institute of Science and Technology

# Thesis Changes Log

**Name of Candidate:** Anna Shiriaeva

**PhD Program:** Life Sciences

**Title of Thesis:** Interference and primed adaptation intermediates in type I CRISPR-Cas systems

**Supervisor:** Prof. Konstantin Severinov

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

*I would like to thank all reviewers for critical analysis of our results and valuable comments that helped me further improve the thesis. I was happy to think about general questions related to CRISPR biology raised by reviewers. And specific questions about the methodology showed me what aspects of library preparation procedure or data analysis should be improved in our future work. You can find the answers to your questions below.*

## Professor Konstantin Lukyanov

**Q1: Fig.12A. In the Methods, this experiment is described as follows: “At various time points postinduction, cells were plated with serial dilutions on 1.5% LB agar plates for counting colony-forming units (CFU)”. At the same time, the curves in the panel are continuous and smooth; no experimental data points are shown. Please comment. Does it show average results of several independent experiments or one representative example? It should be described in the figure legend in more detail.**

**R1:** The initial figure 12A showed a smoothened plot for one representative replicate. The figure has been corrected and now shows mean±SEM values of CFU/ml obtained in four biological replicates without smoothening. Time points at which cells were collected are now specified.

**Q2: “To achieve this goal, we developed FragSeq – a protocol for strand-specific high-throughput sequencing of short (<700 nt) single-stranded and double-stranded fragments generated in vivo”. Why did you choose this length if the expected prespacer sizes are much less? Please add some explanation to the text for clarity. Did you check the size distribution in the obtained samples (e.g., by gel electrophoresis)? If so, please show these data as a Figure.**

**R2:** The first reason to choose this relatively high cutoff is technical. Considering the limited amount and presumable short-lived nature of prespacer fragments generated inside cells, we needed a protocol that would allow for the selection of short fragments from a large amount of total genomic DNA. We were looking for a fast and simple protocol that would suit this purpose and found the Select-a-Size DNA Clean & Concentrator kit from ZymoResearch. According to the instruction manual, the kit allows for the purification of fragments in a range of 50-700 bp from 3 µg of DNA using two different spin columns. One column separates fragments <700 bp from longer molecules. Since the manufacturer provides no options to change the 700-bp cutoff, we specified this cutoff in our FragSeq method. According to the manufacturer’s protocol, the second column binds fragments longer than 50 bp. However, we run preliminary tests and found that 33-nt oligonucleotides can also be purified on this column, so we used the described Select-a-

Size protocol to purify fragments for FragSeq. The text has been modified and now states: "*The procedure starts with the purification of total DNA using phenol/chloroform extraction to retain small fragments that may be lost if silica column-based DNA purification methods are applied. The purified DNA is then filtered using a commercial kit allowing for selection of fragments shorter than ~700 bp in length (see Materials and Methods)" (see pages 73-74).*

The second reason is that we only knew the size of mature spacers (33 bp) when we started our experiments. Spacers precursors could have been longer, perhaps much longer. Further, we also expected to detect fragments produced by Cas3 in the *wt* and *cas1* self-targeting mutant. *In vitro*, the *E. coli* Cas3 produces fragments in a range of several hundred nucleotides (the exact fragment length distribution is affected by ATP concentration but in any case, fragments below 700 bp constitute the major part) (Mulepati and Bailey, 2013). However, we did not detect Cas3-generated fragments *in vivo*, suggesting they are beyond the cutoffs of our size-selection procedure. Yet, the sequencing of a wide range of fragments allowed us to detect fragments produced by the RecBCD complex (Figs. 34-35).

The size distributions for fragments sequenced using FragSeq and mapped to the genome are now shown in Supplementary Figs. 1A,B, and 2. Representative examples of sequencing library size distributions checked by gel electrophoresis are shown in Supplementary Fig. 1C.

We used two methods to prepare sequencing libraries after short fragments purification. One method is shown in Fig. 15 and includes consecutive ligation of adapters and PAGE-purification of fragments from adapters and adapter dimers (fragment length distributions for libraries prepared by this method are shown in Supplementary Fig. 1). Another method relies on a commercially available Accel-NGS® 1S Plus DNA Library Kit (Swift Biosciences) and includes purification from adapter dimers on magnetic beads (fragment length distributions for libraries prepared by this method are shown in Supplementary Fig. 2). Currently, we do not know why the libraries prepared by the two methods look different, but we suggest that the difference may result from gel- vs. bead-purification.

**Q3: Fig. 16B,C: Y axes are marked "% fragments per 1 kb" and the values are up to 0.1. Is this really 0.1%? If so, it corresponds to 10 b per 1 kb only. Is it sufficient to draw any solid conclusions? Or did you mean 10% (i.e. 0.1 with no "%")? The same question is applicable to several other Figures showing similar experiments.**

**R3:** 100% on the Y-axis in these figures corresponds not to 1000 bp but to the total number of fragments mapped to the genome. The chromosome was divided into non-overlapping 1-kb (or 10-kb in some figures) bins and all fragments were grouped into these bins based on their coordinates (a read was assigned to a bin if the read's center lied within this bin). The resulting plots show the percentage of fragments assigned to each bin (compared to all fragments mapped to the genome). For clarity, the captions to Figs. 13, 14, 16, 21, 24, 26, 28, 32, 34 have been modified.

**Q4: Is it possible from your FragSeq data to roughly calculate the number of prespacer molecules per cell accumulated during primed adaptation?**

**R4:** Unfortunately, I do not think this is possible, because we only know the mass of total DNA before the selection of short fragments (9 μg) and the mass of fragments after the size selection (usually 5-25 ng) but we do not know how efficient the purification of short fragments is. Further, prespacers constitute only a small part of our sequencing libraries (there are plenty of non-specific fragments). Prespacers are short (~30-40 nt) and could be partially lost during purification of the sequencing library from adapter dimers. We are planning to address this issue in our future experiments by the addition of a fixed amount of a DNA ladder to purified total genomic DNA samples before the selection of small fragments. The ladder will be sequenced along with the cellular DNA fragments and will provide information on how fragments of different lengths are represented in final libraries and on their amounts, compared to total genomic DNA.

## Professor Dmitry Chudakov

**Q1: Not quite clear how self-ligation was avoided in FragSeq.**

**R1:** To prevent the self-ligation of fragments, we ligated 5' pre-adenylated adapters to fragments' 3' ends in a ligation reaction that did not include ATP. This is currently stated in chapter 3.8.1 of Materials and Methods *("Since ATP was omitted from the ligation mixture and the 5' end of the i116 adapter was pre-adenylated but the fragments were not, self-ligation of fragments was prevented").*
Self-ligation was possible during ligation of the second adapter. However, this scenario was unlikely because adapters were added in excess over fragments.

**Q2: In general, not quite clear what was going on in the bioinformatic analysis of FragSeq:**
**-Were the reads that looked like two or more ligated fragments filtered off? Or "uniquely aligned" barrier worked for this aim? Not clear.**
**-Was the reads-per-UMI coverage analyzed? Look like it was not high, then how errors in UMIs themselves were accounted for?**
**-How many UMIs obtained for different strains? (this information becomes only quantitative if sufficient coverage was achieved)**
**-Were the processed, UMI-clustered data normalized before further mapping, using UMI(that would be rational)?**

**R2:** As discussed in the answer to the first question, self-ligation of fragments was prevented during ligation of the first adapter to fragments' 3' ends. It is possible that self-ligation artifacts appeared when the second adapter was added for ligation to fragments' 5' ends, and unligated fragments from the first step could compete with the adapter. However, this is unlikely since adapters were added in excess over fragments. In any case, self-ligated fragments were, indeed, filtered off during alignment. The alignment was performed using matchPDict function from the ShortRead R package with 2 mismatches and no indels allowed; self-ligated fragments will not be mapped using this function. If a read was mapped to several locations in the genome, it was also excluded from the analysis. The rest of the mapped reads we call "uniquely aligned".
Identical reads (with the same insert and UMI) were assigned into families and their number was reduced to 1 read per UMI prior to alignment. If an insert or UMI had nucleotides with Phred quality < 20, we considered these bases as potentially erroneous, substituted them with N and, when comparing these reads to others, used an argument fixed=FALSE to allow matching of an N to any letter. Other errors in fragments and UMIs were not considered. We believe this procedure reduced amplification biases but, due to errors incorporated by DNA polymerase in inserts and UMIs, some fragments could have been counted several times. To test if this had any effect on our results, I performed UMI-clusterization with up to 5 errors allowed in inserts and UMI and found that, indeed, the total amount of families decreased, on average, by 10±2% among all libraries. However, this did not affect the interpretation of our results in any way. For each UMI family a consensus sequence was extracted and mapped to the genome. The distribution of reads along the genome did not change and all values corresponding to the percentage of fragments near the PPS were highly similar to the originally reported values (the difference with originally reported values was within 3±1%). The correlation coefficients also remained the same.
We had low reads-per-UMI coverage (mean 2.9±0.4) and the amount of UMIs obtained for various samples varied between approximately 100 thousand and 2 million UMIs. We could not reach higher coverage because a high amount of adapter dimers was present in our libraries that were difficult to get rid of even after multiple PAGE-purification steps.
Overall, we admit that there are several aspects that should be improved in FragSeq. It will be optimized in our further work to make it more efficient (reduce the amount of adapter dimers in libraries) and quantitative (higher reads-per-UMI coverage will be reached due to decreased amount of adapter dimers in our libraries; errors will be accounted for; we are also planning to account for differences in amplification/purification efficiency of fragments with different lengths and possible biases emerging during ligation).

**Q3: "Only reads with a length 16 to 100 nt uniquely aligned to the genome were further analyzed". So this is artificial selection, since the library was like up to 500 bp.**
**Then: "The lengths of fragments mapped to the genome in all strains varied in the range of ≈16-100 nt with a median length of fragments outside the PPS-containing region of 45-47 nt."**

**Formally sounds strange: since the upper limit was artificially selected, one can say that estimation of median length is a nonsense. In reality, Figure 20 shows that this ok, but should be better discussed, starting from the point why 100 nt was selected as a length threshold.**

**R3:** The distribution of fragment lengths in the sequenced libraries is now shown in Supplementary Figs. 1 and 2. FragSeq libraries for the type I-E samples described in Chapter 4.1 were sequenced in a single-end 1x150 bp HTS run (Supplementary Fig. 1A). 24 bp of 150 bp sequenced were UMIs and barcodes. Therefore, we could not map 3’ ends of fragments longer than 126 in these HTS runs. The fragment length distributions had a maximum at ~40 nt and very few fragments were longer than 100 nt, so we took this cutoff for all libraries. However, to make sure that we did not lose valuable information by analyzing only fragments below 100 bp we re-sequenced the libraries for the most interesting samples (*wt* and the *cas1* mutant) using paired-end sequencing (2x75 bp or 2x150 bp) so that both 5’ and 3’ ends of fragments could be sequenced regardless of fragments’ lengths. Then we mapped each pair of a forward and reverse read, found the coordinates of 5’ ends, and extracted the sequence between these coordinates as the sequence of the insert. Fragments longer than 100 nt constituted only a small fraction of all fragments (Supplementary Fig. 1B) so we believe the sequencing of these libraries in a single-end 1x150 bp mode did not have any influence on the results. It is now discussed in the Materials and methods (see page 65).
The sentence “The lengths of fragments mapped to the genome in all strains varied in the range of ≈16-100 nt with a median length of fragments outside the PPS-containing region of 45-47 nt”, indeed, does not provide any valuable information. However, its purpose was to show that there is a difference between the PPS region and the rest of the genome. The next sentence states: “In comparison, fragments mapped to the PPS-containing region of the *wt* self-targeting strain were shorter with the median length of 35 nt.” A figure showing that fragment length distribution remains similar throughout the genome except for the PPS region in the *wt* strain is now included in the thesis (Fig. 20). The wording of the discussed sentence has been changed to emphasize that we report the median lengths of selected 16-100-nt fragments, not a median of all fragments present in cells. The text says now: “*In all strains, median lengths of selected (16-100-nt) fragments mapping outside the PPS-containing region were 45-47 nt (Figure 20). In comparison, fragments mapping to the PPS-containing region of the wt self-targeting strain were shorter with the median length of 35 nt (Figure 20)*” (see page 82).

**Q4: Figure 16.**
**The text says that:**
**“No enrichment was detected in *cas1* mutant, *cas3* mutant, and the nontargeting strain (Figure 16B, C).”**
**In reality, there is enrichment in *cas1* mutant strain, if we look at a wider region, as it should be according to observed in whole genome sequencing, shown in Figure 13,**
**At the same time, there is a nice gap in the middle of this wide hill, specific for the *cas1* mutant strain. This observation is not discussed, while it could be (if it is not an artifact) a critical piece of the whole puzzle.**
**Like, in my naïve interpretation: Cas1 stabilizes the fragments produced by Cas3, otherwise degraded, and this mechanism works in proximity to PPS region, while further DNA degradation beyond +/-200,000 bp region is mostly driven by other enzymes.**
**Or more complicated story, but anyway the nice gap in the hill not discussed.**

**R4:** Indeed, a small enrichment with fragments was present in the *cas1* mutant strain. The sentence “No enrichment was detected in *cas1* mutant, *cas3* mutant, and the nontargeting strain (Figure 16B, C).” has been changed and now states that no such peak (as in the *wt*) was detected in the *cas1* mutant, the *cas3* mutant, and the nontargeting strain. Fig. 16B has been substituted with another plot showing differences in genomic DNA content and fragment distribution along the chromosome. Small enrichment with fragments observed around the PPS in the *cas1* mutant is now discussed on pages 74, 85-87.

**Q5: Figure 22.**
**Motifs are still enriched upstream in the *cas3* nuclease mutant. But relative abundance of these molecules is low, according to Figure 16. No comparative analysis here in terms of numbers, hard to interpret what it means.**

**R5:** A new figure (Fig. 24) has been added showing that primed adaptation occurs in the *cas3* nuclease mutant but spacers are mainly selected from a short ~1.5-kbp region upstream of the PPS. It is, at present, difficult to interpret these data because the precise mechanism of primed adaptation initiation, of Cas3 changing its direction to move downstream of the PPS, and its function in prespacer generation remain unknown.

## Professor Scott Bailey

**Q1: How do you interpret the results of Xue et al (2016), conformational changes in the NTD of Cse1, with your model? Or more generally what role, if any, do you believe these conformation changes play in priming?**

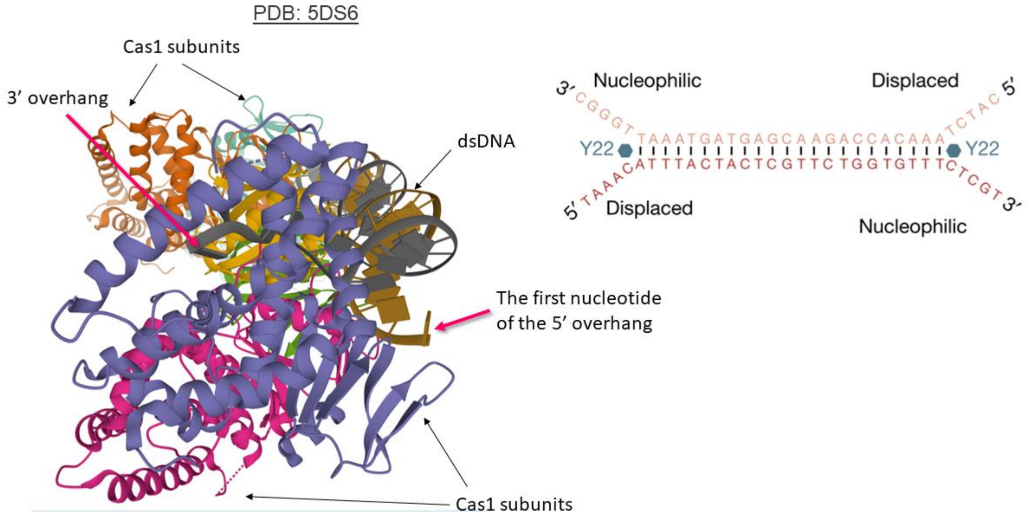
**R1:** In principle, our model should work even if there are no functionally distinct conformations of PPS-bound Cascade, with differences observed being solely due to the presence or the absence of Cas1-Cas2. But if the model presented by Xue et al. (2016) is true, our results do not contradict it. Xue et al. attempted to explain the apparent differences in interference and adaptation efficiencies on fully matched and partially matched protospacers through different ratios of two conformations of the Cascade (closed and open). Closed conformation is favored on fully matched protospacers leading to highly efficient interference while open conformation is favored on mismatched targets leading to the higher efficiency of priming. Following this reasoning, what I am showing in my model in Figure 37 applies to only those Cascades whose Cse1 subunit happened to be in the open conformation, irrespective of the kind of the bound target. One can suggest that when Cse1 is in the open conformation, it promotes the recruitment of Cas1-Cas2/Cas3 and the assembly of the primed acquisition complex (Redding et al., 2015; Dillard et al., 2018). Cas3 may function mostly as helicase in this case delivering Cas1-Cas2 to protospacer sequences located far away from the PPS. When Cse1 is in a closed conformation (not shown in my model), Cas3 alone is recruited and works in the interference mode degrading DNA around the PPS to fragments that are either very short-lived or so short (<16 nt) that we can not detect them with our method.
Overall, I think, the model proposed by Xue et al. could explain why we do not see any specific prespacer-related Cas3 products and why Cas3 nuclease activity may not even be required for prespacer generation. However, several contradictory observations have been published, for example, highly efficient primed adaptation on fully matching targets (Semenova et al. 2016), or highly efficient primed adaptation on partially matched protospacers with stable Cascade "locking" (presumably equal to conversion from open to close conformation) (Krivoy et al., 2018). Therefore, I think the mechanism of interference and priming initiation is not fully resolved and the model developed by Xue et al. requires additional testing.

**Q2: What is your hunch about the identity of nuclease X?**

**R2:** It should be some nuclease that progresses in the 3'→5' direction and releases oligonucleotides of several hundred nucleotides in length as products. It might be a nuclease involved in double-strand break repair that we did not test in this work (for example, ExoVII). Though the reported ExoVII products *in vitro* are shorter than 25 nt (Chase and Richardson, 1974a, 1974b), to my knowledge, its *in vivo* products have not been characterized.
It also might be Cas3 itself if we assume that it works differently in the immediate vicinity of the PPS and further away where we do detect several-hundred-nucleotide long fragments (Fig. 35). Cas3 can be involved through two possible scenarios. First, Cas3 may work differently when it initially reels the DNA while bound to Cascade and after the Cas3-Cascade interaction ruptures. Second, the observed long oligonucleotides could be the products of free Cas3 that was never recruited to the Cascade. It is known that purified Cas3 can work as a single-strand specific endonuclease without Cascade (Sinkunas et al., 2011), so after the initial degradation of the PPS-containing region via CRISPR interference mechanism the remaining DNA, if it has 3' overhangs, could be bound, unwound, and degraded by Cas3.

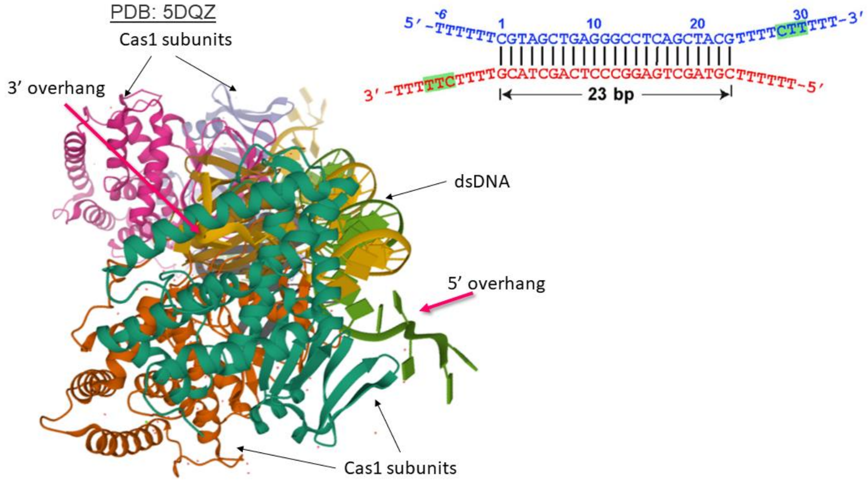
**Q3: How does the length of the trimmed 5'-end compare with the structure of Cas1-Cas2? In other words, does Cas1 protection set the length of 5' overhangs like it is believed to do with the 3' overhangs.**

**R3:** Cas1-Cas2 bound to a 33-bp dual-forked substrate with five-nucleotide frayed ends was crystallized by Nuñez et al. (2015). The authors observed electron density for only the first nucleotide of the 5-nt 5' overhang (the nucleotide closest to the duplex), which protruded out of the Cas1:



Nuñez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N., and Doudna, J.A. (2015). Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature *527*, 535–538.

In the structure obtained by Wang et al. (2015), the 5' overhangs are also displaced of the Cas1-Cas2 complex:


PDB: 5DQZ
Cas1 subunits
3' overhang
dsDNA
5' overhang
Cas1 subunits
-6
1
10
20
30
5'-TTTTTTCGTAGCTGAGGGCCTCAGCTACGTTTTCTTTTT-3'
3'-TTTTTCTTTTGCATCGACTCCCGGAGTCGATGCTTTTTT-5'
23 bp

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

Based on these structures, I would say that the 5' ends are less protected than the 3' ends and can, in theory, be trimmed up to the 23-bp duplex. In reality, it will also depend on the structure of the exonuclease that trims the 5' ends.

**Q4: What are your thoughts on 5' overhang cleavage by Cas1-Cas2 reported by Radovcic et al (2018).**

**R1:** I think that the design of DNA substrates used by Radovcic et al. (10- or 14-bp duplex with 40 nucleotides of single-stranded 5' or 3' ends) does not allow us to determine unequivocally where Cas1-Cas2 was bound to, especially after it was reported that Cas1-Cas2 can bind single-stranded DNA (Kim et al., 2020). Therefore, it is difficult to interpret the results by Radovcic et al.

## Professor Michael Terns

**Q1: I have just one minor suggestion to consider for improving the thesis. The orientation of the crRNA and DNA strands during R-loop formation is presented inconsistently throughout the thesis (i.e. two ways of representing the same structure are used in figures 1 and 2 vs. figures 3, 10, and 14). It seems unnecessary to force a reader to endure the mental gymnastics required to interconvert the two ways of representing the same structure (it slowed me down especially since there are many features to keep track of). Please consider having a uniform representation of the crRNA-bound DNA structure throughout this dissertation to avoid confusion and help the reader (it was particularly clumsy to be steered to follow the diagram in figure one only to have things turn upside down in figure 3…).**

**R1:** Figures 1 and 2 have been replaced and now show R-loops in the same orientation as in other figures throughout the dissertation.

## Professor Yuri Kotelevtsev

**Q1: I have no major or even minor issues with the dissertation which is merely a broader narrative of two excellent experimental papers and an original review containing discussion of the own experimental data of Anna Shiriaeva.**

**R1:** Thank you!