University of Szeged
Faculty of Science and Informatics
Doctoral School of Biology

Investigation of CRISPR/Cas systems and promoters used for the expression of their small RNAs

Summary of Ph.D. thesis

Nóra Weinhardt

Supervisor: Ervin Welker PhD, DSc

Institute of Biochemistry
Biological Research Centre
Hungarian Academy of Sciences

Institute of Enzymology
Research Centre for Natural Sciences
Hungarian Academy of Sciences

2020
Szeged

Introduction

The Cas9 endonuclease (SpCas9) from the bacteria *Streptococcus pyogenes* is currently the most widely applied CRISPR/Cas system used for genetic engineering. Its specificity is provided by the 20 bp ‘spacer’ sequence at the 5 'end of the guideRNA (gRNA). The spacer is used to identify the target sequence based on complementary DNA-RNA hybridization. Nuclease cleaves both strands of DNA. This triggers DNA repair pathways that can be used to specifically modify DNA. The 2 main pathways are homologous recombination (HR) and non-homologous end joining (NHEJ) error correction mechanisms. Recognition and cleavage of the target sequence also requires the presence of a few base pairs in addition to the target, the PAM sequence.
With today's molecular biology techniques, the sequence of the spacer can be easily changed, thanks to which the system has explosively spread. Almost any section of the genome can be freely modified with the Cas9 protein, so it has significant biotechnological, medical and scientific uses.
Cas12a nucleases were first published a few years ago as a method for genome editing. They are very interesting because they recognize a different PAM sequence than SpCas9, so they are able to cleave different targets, and in a different way, because in contrast to Cas9, cleavage of two strands of DNA creates sticky ends rather than blunt ones. They have also been shown to have lower off-target activity than SpCas9.
CRISPR/Cas systems have limitations. One of these is the off-target effect, i.e. the cleavage of unwanted targets similar to the target sequences. There were several attempts to address this, one of which is to modify the nuclease with mutations at specific sites. To date, several SpCas9 variants have been developed that have increased specificity.

There are two common methods for producing gRNAs in practice than the more cost-effective RNA synthesis. One is using the T7 promoter *in vitro* for RNA transcription, and the other is the use of promoters suitable small RNA expression in different organisms. Numerous studies have shown that the activity of Cas nucleases is influenced by the amount and length of gRNAs, so it is extremely important to have an accurate knowledge of the promoters used.
Several studies suggest that nucleotides downstream of the promoter can also affect RNA expression and the position of transcription initiation. Although for some of the promoters we studied, the role of the first nucleotide downstream of the promoter has already been systematically investigated, but the role of additional nucleotides has not yet been studied. Based on these studies, both the human 7sk and U6 promoters can be used with A and G at the +1 position, and RNAs with the expected length were obtained from them.

The transcriptional start site of the T7 promoter is known to be affected by the first nucleotide downstream of the promoter and several results suggest that at least one Guanine should be used for proper expression. However, it is not consistent in the literature that the T7 promoter requires one, two or 3 guanines.
The U6T7 hybrid promoter was created by modifying the U6 promoter by replacing the end of the promoter with the T7 promoter. According to the article introducing the hybrid promoter, it is recommended to use at least one G nucleotide. To the best of our knowledge, no result for the +1 position has been published for bacterial J23119 promoters.
Although the U6 and 7sk promoters have been shown to be efficient in expressing RNAs with an A at position +1, in practice targets for SpCas9 gRNAs are generally selected in which the base corresponding to the 5' first base of the spacer is G, which narrows the number of the potential targets. As a solution, various gRNA modifications are used, which in turn can affect the activity of nucleases.
These +1 position studies were generally not performed with next-generation sequencing (NGS), which currently provides the most accurate picture of the length and amount of small RNAs, but with other less accurate methods and the results obtained with NGS so far used a very small sample size.

Objectives

1. With regard to Cas12a nucleases (like As and LbCas12a nucleases, which come from the *Acidaminococcus* sp. BV3L6 and *Lachnospiraceae* identified ND2006 bacterias) we were looking for answers to their ability to effectively induce HR repair in N2A (mouse neuroblastoma) cells, and to their efficiency compared to the most commonly used Cas9 nucleases.

2. For the eSpCas9 and SpCas9HF1 variants with increased specificity, we wanted to know how their activity is affected by the different gRNA modifications used when the base corresponding to the first transcriptional base of the gRNA in the target is not guanine.

3. We intended to perform a much more comprehensive study of the promoters human U6 and 7sk, phage T7, U6T7 hybrid and bacterial J23119 compared to the previous ones with deep sequencing. We wanted to know how the 4 positions downstream of the promoter affect the position of transcription initiation and RNA expression.

Methods

To examine Cas12a nucleases and compare them with Cas9 nucleases, small RNAs complementary to different targets and plasmids expressing the nucleases were generated. These were cotransfected into N2a (mouse neuroblastoma) cells. Nuclease activity was assayed by GFxFP as well as genomic HR assays and the fluorescent signal was measured by flow cytometry.

GFxFP Assay

In this assay, the target sequence of nucleases is integrated between two repetitive GFP sequences that do not express intact GFP alone. Double-stranded DNA breakage created by nuclease cleavage results in the formation of intact GFP. The advantage of this assay is that cleavage efficiency is not affected by factors such as the genomic environment, the epigenetic state of the target sequence, or the relative distance of the target from homologous arms.

Genomic HR Assay

To induce genomic HR repair, a donor plasmid encoding a GFP sequence containing arms homologous to the target sequence was used. The donor plasmid was integrated behind the PRNP gene's promoter by nuclease cleavage-induced HR repair, and the GFP signal was measured cytometrically.

N2a.EGFP Assay

To examine the activity of nucleases with increased specificity, modified gRNAs and plasmids encoding the nucleases were generated and cotransfected into N2a.EGFP cells. N2a.EGFP cells contain EGFP integrated into the genome, which was targeted with nucleases, and the decrease in fluorescent signal was measured by flow cytometry to test the activity of the nucleases.

Construction of plasmid libraries

To examine the promoters, plasmids containing the appropriate promoters were first constructed to serve as vectors for the plasmid libraries. Then, 3 plasmid libraries were generated per promoter, encoding one SpCas9-modified gRNA with 3 different spacer sequences. The first 4 bases downstream of the promoters were randomized.
The sequence of the gRNA also contained a barcode to identify which RNA was derived from which template, and provided a variety of sequences for the 4 base sequences examined, which reduces the effect of external factors affecting the amount of RNA, such as sequence-specific degradation of RNAs. Fragments of the libraries were assembled by overlapping DNA segments using the Gibson cloning technique.

Generation of cDNA Libraries

Plasmid libraries for mammalian cell promoters were transfected into HEK-293 cells. From the T7 promoter libraries, the expression cassette containing the promoter and gRNA was amplified by PCR, and we used this as a template to generate RNAs *in vitro* with a T7 kit. Cells containing bacterial plasmid libraries were incubated in LB medium with shaking overnight. This was followed by isolation of RNAs followed by reverse transcription with a primer specific for gRNA.

NGS

Adapters were ligated to the 3 'end of the cDNAs using T4 DNA ligase, for which a unique method for sequencing gRNAs was set up. The adapter is a partially double-stranded DNA molecule. Both ends of the adapter and the 3' end of the oligo complementary to the adapter are phosphorylated. Phosphorylation of the 3' ends avoids the formation of ligation by-products because it inhibits ligase. At the 3' end of the complementary oligo, 4 randomized bases facilitate ligation with cDNA libraries containing also randomized bases.
The cDNA libraries, together with the plasmid libraries, were deep sequenced after PCR amplification, and the results were evaluated using bioinformatics methods.

Results and Discussion

We compared the efficacy of As and LbCas12a nucleases with the activity of Sp- and 3 other commonly used Cas9 nucleases. To examine the cleavage activity of nucleases, an assay called GFxFP was optimized from a previously published assay in which a background fluorescent GFP signal was obtained without cleavage of the nuclease. We successfully modified the reporter essay to reduce this non-cleaved GFP background. Using the assay, we wanted to exclude differences in activity caused by different genomic positions due to different PAM sequences. Based on our results, although SpCas9 proved to be more effective: it elicited a higher rate of HR improvement in cells than Cas12a, but compared to other Cas9s, Cas12a nucleases performed to a similar extent at some targets. We tested the suitability of Cas12a nucleases for genome editing not only with a reporter assay but also with genomic targets. Our results suggest that they are also suitable for inducing HR repair by targeting genomic targets. In our studies of Cas12a nucleases, we used another mammalian cell line (N2a) than Zetsche et al., and demonstrated that Cas12a nucleases are suitable for genome engineering, and while they only tested their activity using the NHEJ repair pathway, we induced HR repair in the genome and we targeted the GFxFP plasmid as well. Based on our results, Cas12a nucleases may provide a useful alternative in the range of genome editing tools, which is consistent with the fact that in recent years, several groups have also shown their suitability for gene editing in a variety of organisms.

In connection with the eSpCas9 and SpCas9HF1 variants with increased specificity, we wanted to know how their activity is affected by the different gRNA modifications used when the base corresponding to the first transcribed base of gRNA in the target is not guanine. Of these modifications, the replacement of the last nucleotide at the 5 'end of the gRNA with G, or leaving the spacer unchanged, and the shortening of the 5' end of the gRNAs were tested. Their effects were tested in an EGFP-containing mouse neuroblastoma cell line with the efficacy of EGFP cleavage. Based on our results, while the activity of wild-type SpCas9 is only slightly affected by these modifications, the activity of the high-specificity Cas9 variants is greatly reduced to varying degrees. Thus, based on our results, these nuclease variants cannot be routinely used with such modifications. Several solutions are possible to resolve this, one of which is to modify the nuclease variants with additional mutations that better tolerate the modifications of the gRNA. In order to expand the range of targets for nucleases with increased specificity, there is a need for a more detailed study of promoters suitable for small RNA expression. This could be used to find other sequences with which these nucleases could be used without gRNA modifications.

In most of the work presented in my dissertation, we sought to answer the question of how bases located downstream of the promoters commonly used to express gRNAs are to affect RNA expression and the transcription start positions. In our experiments we performed a much more comprehensive examination of the selected promoters than others before. The role of not only the +1 position but also the 3 additional positions downstream of it was examined for RNA amounts and transcription initiation positions. In our study, plasmid libraries were generated for each promoter and a unique method for sequencing gRNAs was used to sequence RNAs expressed from plasmid libraries. Our experiments were performed with a large number of different RNA sequences, thanks to randomized barcodes built into the RNA coding region, in order to avoid external effects such as sequence-specific RNA degradation in addition to the 4 base sequence differences studied. Our studies of promoters were performed with high coverage per 4 basepair long sequences using next generation sequencing which is a state-of-the-art technique.

Based on our NGS results, it is true for all investigated promoters that the nucleotides of positions downstream of the promoter can affect the amount of RNA generated and the exact location of transcription initiation, and thus the length of the RNA. So far, only the +1 position has been systematically examined, also only for the T7, U6, and 7sk promoters. Based on our results, not only the +1 position but also the positions downstream of it are important for RNA expression. Thus, there may be orders of magnitude differences in RNA amounts depending on the additional positions, even though the nucleotide at position +1 is the same. It has been found for all promoters tested, that the starting position of transcription may also be affected by a sequence downstream of the promoter. For each promoter, we found that if pyrimidine is in the first position downstream of the promoter, the +1 position will be the starting position in a much smaller proportion, and in this case the transcription will not start from a fixed position. In the case of pyrimidine in +1, the starting position rate of each promoter was greatly influenced not only by the first nucleotide following the promoter but also by the other subsequent nucleotides.

We found that there was no significant difference in the amount of RNA expressed from the human U6 promoter between +1A and +1G, which is consistent with publications that the U6 promoter can also be used with an A. Examination of transcripts transcribed from the U6 promoter showed that if there is an A or G nucleotide at position +1, the starting position of transcription is in almost all cases the +1, which is not much affected by + 2-4 positions. Examining the first two positions, we found that AG and GG provide significantly higher expression, reaching twice the expression of the lowest expressed GC sequences.

The +3-4. positions influence the amount of RNA depending on the sequence upstream of it, but even the sequence of these positions can cause significant differences in the amount of RNA. By analyzing the sequences containing A or G at the +1 position, we showed that the amount of RNA expressed from the U6 promoter was affected by the sequence of + 2- +4 positions causing an order of magnitude difference.

We found that the human 7sk promoter can also be used not only with +1 guanine but also with adenine for small RNA expression. We have shown that the presence of purine at the +1 position causes accurate transcription initiation at the +1 position, and the additional +2- +4 positions does not change this significantly. +1G produces significantly more RNA than + 1A. Examining the first two positions, we found that the highest expression was yielded from AG, GG, and GT, which produced twice the amount of RNA on average than the purine-initiated dinucleotides with the lowest RNA level. The +3-4 positions also have a significant effect on the expression level, depending on the sequence upstream of them. Based on our results, if nucleotide A is in position +1, the sequence of the additional 3 bases can cause differences in RNA level up to 27x and +1G can cause 7x. We have found that there is no significant difference in the amount of RNA produced from the U6 and 7sk promoters in HEK-293 cells, so the 7sk may be suitable for gRNA expression in parallel or even in place of the U6 promoter.

According to our measurements, the T7 promoter prefers guanine at the +1 position, similar to the previous data. However, there are also sequences starting with adenine from which the same amount of RNAs can be produced like with +1G, and transcription also largely starts from +1. Examining the RNAs generated from the T7 promoter, we found that the sequences starting with AG dinucleotides had almost as accuracy in the transciption initiating from the +1 position than the sequences starting with GN and the transcription start site was not greatly influenced by the other positions. It is inconsistent in the literature that the T7 promoter requires one, two or 3 guanines. According to our results, GG and GGG sequences generate half and a quarter amount of transcripts, respectively, than those initiating with GA. Interestingly, for AG dinucleotides in +1-2, we found expression levels reaching GA. The +3-4. positions also have a significant effect on the amount of RNA, depending on the sequence upstream of them. Based on our results, with a starting AG dinucleotide, the additional bases can cause differences in RNA level up to 5x and more than 13x with +1G.

Based on our results, the modified U6T7 hybrid promoter, although useful with T7 polymerase, should be considered for mammalian expression. Although we found that there was no significant difference in the amount of RNA generated from the U6T7, 7sk, and U6 promoters, but RNAs form U6T7 promoter are very heterogeneous in length, even from +1purine.
We examined the effect of the sequence downstream of the J23119 promoter on transcription and found that for +1purine, the site of transcription initiation is almost always the +1 position, which is not significantly affected by +2-4. positions. Furthermore, we have shown that most RNA is generated when purine is present in +1. Based on our results, the amount of RNA generated from the J23119 promoter is also significantly influenced by the sequence of +2- +4. positions. Examining the starting dinucleotides, most RNA is generated with an AG, which provides 3x the amount of RNA compared to the weakest starting dinucleotides. The +3-4 positions also have a significant effect on the expression level, depending on the sequence upstream of them. Based on our results, if nucleotide A is in position +1, the sequence of the additional 3 bases can cause differences in RNA level up to 50x and 5x with +1G.

Our results related to promoters can be significant for several reasons.
For each of the promoters from 3 different organisms we found that not only the first but also the 3 additional positions downstream of the promoter can affect the transcription start site, and thus the length of the RNAs as well as the expression. This may also be interesting in basic research and raises the need for investigating other promoters and additional positions.
Based on our results, higher RNA levels and more accurate lengths of RNAs can be obtained using the sequences provided with the promoters studied. This can be extremely useful for various applications such as RNA interference or CRISPR. Based on our results, the range of targets that can be selected for CRISPR may be expanded due to the finding of downstream-located sequences from promoters that were not previously known to be as suitable for gRNA expression as previously known variations. In addition, our results- together with the results from the examination of additional positions- could be suitable for the further development of the current CRISPR target prediction programs.
The results presented in my dissertation can contribute to a more efficient application of the CRISPR/Cas system.

Summary

1. We have shown that Cas12a nucleases can effectively induce HR repair in N2a cells using both genome-integrated and ectopic GFP reporters.
2. Based on our results, the activity of the eSpCas9 and SpCas9HF1 variants is reduced by the examined gRNA modifications.
3. We have shown that the first base downstream of the promoter does not determine the expression level independently of the additional sequence. We have found that there may even be orders of magnitude differences in RNA amounts depending on the sequence of the + 2- +4 positions.
4. We have found that bases downstream of the promoter can also affect transcription start positions. At position +1, pyrimidine has a lower ratio of +1 position as the transcription start position, which is also affected by other positions.
5. Examining the U6 promoter we found that in the case of an A or G nucleotide at position +1, the transcription start position in almost all cases is the +1, which is not significantly influenced by the sequence of positions + 2- +4. We have shown that there is no significant difference in the amount of RNA expressed from the human U6 promoter between + 1A and + 1G.
6. We found that the presence of purine at position +1 in the human 7sk promoter causes accurate transcription initiation at position +1, which is not significantly altered by the sequence of the additional +2 + +4 positions. Overall, +1G generates significantly more RNA from the 7sk promoter than + 1A. Based on our results, there is no significant difference in the amount of RNA produced from U6 and 7sk promoters in HEK-293 cells.
7. According to our measurements, the T7 promoter prefers guanine at the +1 position, similar to the previous data. However, there are also sequences starting with adenine that can produce almost the same amount of RNAs than sequences with +1G, and with these transcription also largely starts from +1. According to our results of the sequences containing G in +1 fewer transcripts are generated from those starting with GG and GGG than from those starting with GA, which provide higher expression.
8. We have shown that modification of the U6 promoter to the U6T7 hybrid promoter resulted in a very heterogeneous position from which transcription begins even when there is a purine at position +1.
9. Regarding the J23119 promoter we found that for +1purin the transcription initiation site is almost always the +1 position, which is not as significantly influenced by the sequence of +2 + +4 positions as in the case of +1pyrimidines. Furthermore, we have shown that overall the most RNA is generated when there is purine in +1.

List of publications

Publications directly related to the topic of the dissertation:

1. Cpf1 nucleases demonstrate robust activity to induce DNA modification by exploiting homology directed repair pathways in mammalian cells
Toth, E; **Weinhardt, N***; Bencsura, P; Huszar, K; Kulcsar, P I; Talas, A; Fodor, E; Welker, E
BIOLOGY DIRECT 11 Paper: 46, 14 p. (2016) **IF:3,472 * shared first author**

2. Crossing enhanced and high fidelity SpCas9 nucleases to optimize specificity and cleavage
Kulcsar, PI; Talas, A; Huszar, K; Ligeti, Z; Toth, E; **Weinhardt, N**; Fodor, E; Welker, E
GENOME BIOLOGY 18 Paper: 190, 17 p. (2017) **IF:13,214**

3. Nucleotides downstream of the promoter affect start site and RNA level in promoters used for gRNA expression
(Results are currently in the form of a manuscript of which the candidate is the first author)

Cumulative impact factor of publications directly related to the topic of the dissertation: 16,686

Publications indirectly related to the topic of the dissertation:

4. Mb- and FnCpf1 nucleases are active in mammalian cells: activities and PAM preferences of four wild-type Cpf1 nucleases and of their altered PAM specificity variants.
Toth, E; Czene, B C; Kulcsar, P I; Krausz, S L; Talas, A; Nyeste, A; Varga, E; Huszar, K; **Weinhardt, N**; Ligeti, Z; Borsy A; Fodor, E; Welker, E
NUCLEIC ACIDS RESEARCH 46: 19 pp. 10272-10285., 14 p. (2018) **IF:11,147**

5. A convenient method to pre-screen candidate guide RNAs for CRISPR/Cas9 gene editing by NHEJ-mediated integration of a 'self-cleaving' GFP-expression plasmid
Talas, A; Kulcsar, P I; **Weinhardt, N**; Borsy, A; Toth, E; Szebenyi, K; Krausz, S L; Huszar, K; Vida, I; Sturm, A; Gordos, B; Hoffmann, O I; Bencsura, P; Nyeste, A; Ligeti, Z; Fodor, E ; Welker, E
DNA RESEARCH 24: 6 pp. 609-621., 13 p. (2017) **IF:5,415**

Publications not related to the topic of the dissertation:

6. Highly efficient RNAi and Cas9-based auto-cloning systems for C. elegans research.
Sturm, A; Saskoi, E; Kovacs, T; **Weinhardt, N**; Vellai, T
NUCLEIC ACIDS RESEARCH 46: 17 Paper: e105, 13 p. (2018) **IF:11,147**

MTMT ID: 10054747
Cumulative impact factor (IF) for all publications: 44,395
Total number of independent citations: 97
Hirsch index: 5