Doctoral School of Multidisciplinary Medical Science
Faculty of Medicine
University of Szeged

Development of new increased fidelity SpCas9 variants

Péter István Kulcsár

Short Summary of PhD Thesis

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

Szeged
2019

List of Publications

Publications related to the thesis:

I. **Kulcsár P. I.**, Tálas A., Huszár K., Ligeti Z., Tóth E., Weinhardt N., Fodor E., Welker E., Crossing enhanced and high fidelity SpCas9 nucleases to optimize specificity and cleavage, Genome Biology (2017). **IF: 13.214**

II. Tálas A., **Kulcsár P. I.**, Weinhardt N., Borsy A., Tóth E., Szebényi K., Krausz S. L., Huszár K., Vida I., Sturm Á., Gordos B., Hoffmann O. I., Bencsura P., Nyeste A., Ligeti Z., Fodor E., Welker E., 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, DNA Research (2017). **IF: 5.415**

Cumulative impact factor of papers directly related to the thesis: 18.629

Publications indirectly related to the subject of the thesis:

III. Tóth E., Czene B. C., **Kulcsár P. I.**, Krausz S. L., Tálas A., Nyeste A., Varga É., Huszár K., Weinhardt N., Ligeti Z., Borsy A. É., Fodor E., Welker E., 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, Nucleic acids research (2018). **IF: 11.147**

IV. Tóth E., Weinhardt N., Bencsura P., Huszár K., **Kulcsár P. I.**, Tálas A., Fodor E., Welker E., Cpf1 nucleases demonstrate robust activity to induce DNA modification by exploiting homology directed repair pathways in mammalian cells, Biology Direct (2016). **IF: 2.856**

Publications not related to the thesis:

V. Billes V., Kovács T., Manzéger A., Lőrincz P., Szincsák S., Rezgős Á., **Kulcsár P. I.**, Korcsmáros T., Lukácsovich T., Hoffmann Gy., Erdélyi M., Mihály J., Takács-Vellai K., Sass M., Vellai T., Developmentally regulated autophagy is required for eye formation in *Drosophila*, Autophagy (2018). **IF: 11.059**

VI. Tóth E., Huszár K., Bencsura P., **Kulcsár P. I.**, Vodicska B., Nyeste A., Welker Zs., Tóth Sz., Welker E., Restriction Enzyme Body Doubles and PCR Cloning: On the General Use of Type IIS Restriction Enzymes for Cloning, Plos One (2014). **IF: 3.234**

VII. Kulcsár G., Gaál D., **Kulcsár P.I.**, Schulcz Á., Czömpöly T. A mixture of amino acids and other small molecules present in the serum suppresses the growth of murine and human tumors in vivo., Int J Cancer, (2013). **IF: 5.007**

VIII. Tóth E., **Kulcsár P. I.**, Fodor E., Ayaydin F., Kalmár L., Borsy A. E., László L., Welker E., The highly conserved, N-terminal (RXXX)(8) motif of mouse Shadoo mediates nuclear accumulation., Biochim Biophys Acta (2013). **IF: 5.297**

Cumulative impact factor of all papers: 57.229, h-index: 5, citations: 102

Introduction

The development of new genome modification techniques has greatly increased the feasibility of introducing targeted changes into the genome of any target organism. These techniques have great potential in medical treatments, in agriculture and in basic research applications as well. To get genome engineering up to the current level, one of the critical issue, needed to be addressed was to develop reliable methods for targeted DNA modification. Although meganucleases (MNs), zinc finger nucleases (ZFNs), and transcription activator-like effector nucleases (TALENs) were also an important step, the introduction of the CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 (CRISPR associated protein 9) system as a genome engineering tool clearly revolutionized the wider field of biotechnology.

CRISPR/Cas proteins are RNA-guided endonucleases that can be directed to cleave a chosen DNA or RNA sequence and they provide an adaptive immunity against mobile viruses, plasmids and transposons in archaea and bacteria. The CRISPR/Cas systems can be classified into two major classes to our knowledge today. Class 1 encompasses subgroups which have multiple subunit effector complexes, while Class 2 contains much simpler systems with single multifunctional and multidomain protein effector modules. Nucleases of type II group of Class 2 class contain cas proteins with similar domain architecture, including a RuvC-like and a HNH nuclease domains each cleaving one DNA strand. Since the first published results, many different CRISPR nucleases have proved their values for genome engineering applications, but among them presently *Streptococcus pyogenes* Cas9 (SpCas9) is perhaps the most commonly used genome engineering tool.

The ribonucleoprotein (RNP) complex of the Cas9 nucleases involve the Cas9 protein itself and two Cas9-associated RNAs [CRISPR RNA (crRNA) and the trans-activating crRNA (tracrRNA)] possessing sequences complementary to each other. Complementarity between the targeted DNA site and the spacer sequence of the crRNA and the presence of a short protospacer-adjacent motif (PAM) at the 3'-end of the target site are also required to the binding and cleavage to occur. It has been shown that SpCas9 nucleases can be guided to the desired target site by a fused cr- and tracrRNA, named single guide RNA.

Although the use of SpCas9 has a lot of advantages compared to the earlier systems (MNs, ZFNs, TALENs), some limitations still remain. The two major ones that have to be overcome are related to its varying efficacy (i.e. level of percentage modified at the desired on-target site) and specificity (i.e. on-target activity vs. off-target cleavage). SpCas9 nuclease

shows high level of on-target cleavage in case of most of the target sites, but also a considerable level of off-target activity; i.e. the nuclease also cleaves targets that show limited, imperfect complementarities with the associated sgRNA. The off-target sequences are difficult to predict *in silico* and have been shown to contain up to 5-6 base mismatches, a property that may interfere with many research applications as well as using this nuclease for therapeutic purposes.

Much effort has been devoted to circumventing the off-target effect of the SpCas9 nuclease with different approaches as well. There are two main directions for developing methods to attain higher specificity SpCas9 cleavage. (i) In case of the first direction one tries to limit the actions of SpCas9 in the cells such a way that it has still enough activity for on-target cleavage but no, or lower, activity for off-target cleavage. This is generally attempted by restricting the time or the extent of the exposure (e.g. applying inducible Cas9 or *in vitro* expressed ribonucleoprotein complex) or by limiting the activity of the protein through the use of sgRNAs containing truncated or extended spacer sequences. (ii) The other direction is to increase the recognition sequence by using paired SpCas9 nickases or using pairs of catalytically inactive SpCas9 fused to a non-specific *Fok*I nuclease domain.

The most promising approaches to decrease its off-target activity, however, are the generation of increased fidelity SpCas9 variants. The first two published increased fidelity variants are eSpCas9 and SpCas9-HF1 both of which generated by rational design. Both mutants exhibited considerably reduced off-target effects even when assessed by unbiased whole-genome off-target analysis, although some off-targets, mainly only with single-base mismatches, were still found. However, a subset of the targets, referred to as atypical in the SpCas9-HF1 publication, with repetitive or homopolymeric sequences were still cleaved with considerably high off-target effects. Two additional increased fidelity variants were also later published: HypaSpCas9 developed by rational design as well and evoSpCas9 by exploiting a selection scheme.

It has been shown that increased fidelity variants work poorly with some types of 5’ modified sgRNAs. This issue has practical aspects: to comply with the sequence requirement of the promoters (U6 and T7) commonly used to transcribe the sgRNAs, 5’ modified sgRNAs are frequently used with the wild type (WT) SpCas9 when appropriate 20G-N19-NGG targets cannot be identified bioinformatically. Another observation is that increased fidelity variants have a lower on-target activity on some target sites compared to WT SpCas9. Furthermore, it has also been reported that increased fidelity variants do not work effectively in the pre-assembled RNP form that would be relevant to their clinical application.

Aim of this study

While the results seen with increased fidelity variants are very encouraging, it has become clear that there are still several target sequences which can be cleaved by increased fidelity variants only with considerable off-target effect. Unfortunately, it is hardly predictable which sequences these are.

Furthermore, it is difficult to decide which SpCas9 variant is superior for applications where the avoidance of off-target activity is of paramount interest, because they were characterized in differing experimental setups: exploiting different sets of targets in different cells, and employing different methods to assess their genome-wide specificity.

Even though the propensity for off-target activity has been considerably decreased of SpCas9 in general by using its increased fidelity variants, there are several limitations detailed in the introduction (e.g. their lower activity with 5' modified sgRNAs or when used in RNP form, as well as their reduced target space) which prevent the exploitation of their full potential.

Therefore, I pursued the following aims:

1. Systematically characterizing the increased fidelity variants in the same experimental sets and conditions.
2. Making a new, even better working increased fidelity SpCas9 variant with even higher fidelity than the already published ones.
3. Examining the compatibility between increased fidelity variants and different type of 5' modified sgRNAs.
4. Examining how the specificity of genome modification using the increased fidelity variants could be further increased by combining it with other fidelity increasing approaches.

Methods

Plasmid construction

Vectors were constructed using standard molecular biology techniques including the one-pot cloning method, E. coli DH5α-mediated DNA assembly method, NEBuilder HiFi DNA Assembly and Body Double cloning method.

Cell culturing and transfection

The cells employed in these studies are mouse N2a, human HEK293, N2a.dd-EGFP (cell line developed by us containing a single integrated copy of an EGFP-folA dihydrofolate reductase destabilization domain fusion protein coding cassette driven by the Prnp promoter), N2a.EGFP and HEK-293.EGFP (cell line developed by us containing a single integrated copy of an EGFP cassette driven by the Prnp promoter) cells. Transfections were performed with TurboFect transfection reagent according to the manufacturer's recommended protocol.

EGFP disruption assay

SpCas9 variants were tested in an EGFP disruption assay, where we detected the loss of fluorescence signal in cells caused by frameshift mutations from the error-prone non-homologous end joining (NHEJ)-mediated repair after the targeted double-stranded breaks in a single copy integrated EGFP reporter gene. EGFP disruption experiments were conducted in N2a.EGFP or in N2a.dd-EGFP cells. SpCas9 variants and sgRNAs were transfected either in plasmid or in RNP form. Transfected cells were analyzed by flow cytometry. EGFP disruption values were calculated as follows: the average EGFP background loss from control transfections made in the same experiment was subtracted from each individual treatment in that experiment and the mean values and standard deviation (s.d.) were calculated from it.

Western blot

SpCas9 variants were tested by immunoblot analysis to test whether the mutations alter the expression level of SpCas9 or not, and if the amounts of the proteins expressed at steady state are comparable. N2a.dd-EGFP cells were transfected and four days post-transfection cells were analyzed for transfection efficiency by flow cytometry and resuspended in Harlow buffer. The blots were incubated overnight at 4 °C in primary anti-FLAG and anti-β-actin antibodies. The next day, after washing steps, the membranes were incubated with HRP-conjugated

secondary anti-mouse antibody. The signal from detected proteins was visualized by ECL using a CCD camera.

Transcriptional activation

One of the applications of SpCas9 nucleases, besides being a programmable nuclease, is using it for delivering effector domains precisely to a chosen locus within the genome for modulating endogenous gene expression. For transcriptional activation we exploited a method published by Konermann and colleagues, that uses a modified sgRNA containing two minimal hairpin aptamer - which selectively binds dimerized MS2 bacteriophage coat proteins- in the tetraloop and stem loop 2. The sgRNA complexed with catalytically inactive SpCas9 binds to the target site and the hairpins recruit the MS2-p65-HSF1 fused protein which can mediate transcriptional upregulation.

GUIDE-seq and TIDE

We performed GUIDE-seq experiments (Genome-wide, Unbiased Identification of DSBs Enabled by sequencing) - which allows the unbiased identification of the off-target sites occurred after a targeted DSB- with different SpCas9 variants on 13 different target sites either in plasmid or in RNP form in HEK293.EGFP cells. The method relies on the efficient integration of an end-protected, short, double-stranded oligodeoxynucleotide (dsODN) into sites of DSBs in living cells. The genomic context around each off-target site is amplified using this short sequence and the ligated adapters, followed by NGS and *in silico* identification of the occurring off-target cleavage sites. On-target cleavage activity was checked in each experiment either with an EGFP disruption assay by flow cytometry or with the Tracking of Indels by DEcomposition (TIDE) method.

NHEJ-mediated integration using a ‘self-cleaving’ EGFP-expression plasmid

The key feature of this approach is the use of a ‘self-cleaving’ plasmid that enhances targeted integration. In case of this method we are transfecting two plasmids: (i) a plasmid expressing the SpCas9 and an sgRNA targeting the genomic site, and (ii) a ‘self-cleaving’ donor plasmid harboring the EGFP cassette and a - genomic-targetless – sgRNA targeting the donor plasmid itself. After the transfection the SpCas9 does not only cleave the genomic target site but cleaves the donor plasmid as well. During the repair of the broken genomic DNA, the linear donor plasmid will more efficiently integrate into the genome than a circular donor plasmid.

Results

There is a ranking by fidelity and target selectivity among the increased fidelity SpCas9 variants, and a ranking among the targets.

As the first step of this study, we aimed to characterize and compare the wild type (WT), the already existing increased fidelity SpCas9 variants (e-, -HF1, Hypa- and evoSpCas9), and those which we made by combining the mutations of different variants (Hypa2SpCas9 and HeFSpCas9) in the same experimental set ups. We performed our first direct comparisons of the SpCas9 variants using EGFP disruption assay. Firstly, we tested the on-target activity of the variants by employing 47 sgRNAs with perfectly matching 20 nucleotide-long spacers targeting the EGFP coding sequence. Secondly, we characterized the mismatch tolerance of the increased fidelity variants by comparing the effect of single base mismatches in the PAM distal region on their fidelity, because it is a sensitive approach capable of discerning small fidelity differences among the variants.

The on-target and mismatch screen experiments revealed a very intriguing pattern with important consequences. The fidelity of the variants gradually increases in parallel with increasing target-selectivity (i.e. decreasing average activity, shrinking target space) in the following order: WT SpCas9 < eSpCas9 < SpCas9-HF1 < HypaSpCas9 < Hypa2SpCas9 < evoSpCas9 < HeFSpCa9. Interestingly, these differences between the variants were not clear from the original publications.

Remarkably, the cleavability of different targets by the various increased fidelity variants is also strongly related to the fidelity of the variants, allowing the target sequences to be ranked as well. This target ranking is also apparent in the mismatch screen indicating that the efficiency and the specificity of target-cleavage by the increased fidelity SpCas9 variants are interrelated and highly determined by the targets. We suggest that this pattern seen in our results is caused due to the cumulative effect of (i) the contribution of the specific target, (ii) the increased-fidelity mutations, and (iii) the mismatches, if there are any. The resultant of these effects regulates the cleavage activity in each specific cleavage event. Taking a closer look at the target ranking, it can be seen that at one end of the ranking there are sequences (we named them "*high-contributing"* target sites) that all variants can cleave, but only the highest fidelity variant cleaves them with minimal off-target cleavage. At the other end of the ranking there are sequences ("*low-contributing*") that are cleaved efficiently only by the lower increased fidelity variants and by the WT SpCas9 but also with significantly lower level of off-target activity.

The higher fidelity variants, such as HeF-, evo-, Hypa2SpCas9 do not cleave these *low-contributing* target sites effectively.

This propensity of the targets makes efforts to develop a superior nuclease variant that cleaves all targets with high on-target activity and high specificity (without any off-target effect) futile. Instead, our results show that it is necessary to have a series of variants with increasing fidelity in order to have an optimal increased fidelity nuclease for targets in all cleavability ranks.

5'-extended sgRNAs diminish the activities of increased fidelity nucleases more with a matching than with a mismatching G nucleotide.

5’ modified sgRNAs are frequently used with WT SpCas9 when appropriate 20G-N19-NGG targets cannot be identified bioinformatically in the desired genomic region, because the most commonly used promoters (such as the human U6 promoter in mammalian cells or the T7 promoter *in vitro)* require a starting 5’-G nucleotide for efficient transcription. We examined the perhaps most commonly used approach to fulfill the G requirement: extending the sgRNA with an extra 21st 5’ G nucleotide (21G-sgRNA). The results of these experiments demonstrate that 21G-sgRNAs interfere with the activities of the increased fidelity nucleases (in line with papers published after our study were published). Interestingly, we found that extending the guide with a matching 5' G nucleotide is much more detrimental to the activities of these nuclease variants than extending it with a mismatching one. These results indicate that these increased fidelity nucleases are generally not compatible with these approaches and can be routinely used only with perfectly matching 20G-sgRNAs.

Increased fidelity variants show closely WT-level binding activity on target sites cleaved inefficiently.

Another application of SpCas9, besides being a programmable nuclease, is using its inactive WT variant for delivering effector domains precisely to a chosen locus within the genome. eSpCas9 and SpCas9-HF1 were designed that way to have a weaker interaction either with the non-targeted DNA strand or with the targeted DNA strand, respectively. Therefore, it was not clear if they still have the same level of DNA binding as the WT SpCas9. We wondered how e- and SpCas9-HF1 and the higher fidelity HeFSpCas9 will perform in a transcriptional activation experiment which assumes DNA binding. Contrary to the expectations, all increased fidelity nucleases demonstrated comparable activities to WT, resulting a 15-20-fold activation,

both with 21G-sgRNAs or truncated sgRNAs. These results, although indirect, suggest that the binding of the increased fidelity nucleases to their targets are not impaired even when the on-target activity is diminished with altered sgRNAs.

Employing a more direct *in vitro* approach we performed a polyacrylamide-gel electrophoretic mobility shift assay (EMSA) of the target DNAs by the nucleases. We confirmed that although HeFSpCas9 does not show nuclease activity on selected target sites, it retains most of its DNA-binding abilities to these targets. Furthermore, to better understand the effect of appending an extra 5' G nucleotide to the end of the sgRNAs we examined the *in vitro* binding with matching 21G-sgRNAs to selected targets. We found that although the 5' G extension fully diminishes the cleavage activities of eSpCas9 and SpCas9-HF1 on these target sites in the EGFP disruption assay, their binding seems to remain unaffected.

RNP form of increased fidelity variants further increase the specificity

RNPs are the method of choice for prospective clinical applications. We were curious to see whether we could further increase the specificity of genome modifications with increased fidelity variants in RNP form compared to that when carried out on the same targets by the same optimized increased fidelity nuclease variants but expressed from plasmids. We showed on EGFP target site 20, EGFP target site 43 and *FANCF* site 2 that using the increased fidelity variants in RNP form can truly further decrease off-target events occurring in case of genome modifications. These results demonstrate the power and potential of using increased fidelity variants in combination with the RNP form.

Summary

We directly compared the already published increased fidelity nucleases (eSpCas9, SpCas9-HF1, HypaSpCas9, evoSpCas9) and the ones developed by us (Hypa2SpCas9, HeFSpCas9) in the same systems to understand the factors that affect their cleavage.

1. The comparison of the different increased fidelity variants revealed that these variants can be ranked by their fidelity that increases in parallel with their target-selectivity (i.e. the variant does not cleave or cleaves with reduced activity the target sites that are cleaved by the WT SpCas9 and that results in decreasing average activity and shrinking target space) as well, in the following order: WT SpCas9 < eSpCas9 < SpCas9-HF1 < HypaSpCas9 < Hypa2SpCas9 < evoSpCas9 < HeFSpCa9.

Our results suggest that not only the nucleases, but the target sequences can be ranked. At one end of the ranking there are *high-contributing* sequences that all variants can cleave, but only the highest fidelity variant cleaves them with minimal off-target cleavage. At the other end of the ranking there are *low-contributing* sequences that are cleaved efficiently only by the lower increased fidelity variants and by the WT SpCas9 but also with a significantly lower level of off-target activity.

2. These rankings identified by us makes efforts to develop a superior nuclease variant that cleaves all target sites with high on-target activity and high specificity futile. Instead, our results show that it is necessary to have a series of variants with increasing fidelity in order to have an optimal nuclease for targets in all cleavability ranks.

3. We showed that increased fidelity variants can routinely only be used with sgRNAs containing perfectly matching 20 nucleotide-long spacers (20G-sgRNAs). 5’ modified sgRNAs diminished the activities of increased fidelity variants.

4. We showed that by using the RNP form of the increased fidelity variants instead of the plasmid form, their specificity can be further increased.