Core Hairpin Structure of SpCas9 sgRNA Functions in a Sequence- and Spatial Conformation–Dependent Manner

Mingjun Jiang¹, Yanzhen Ye¹, and Juan Li¹

Abstract
The clustered regularly interspaced short palindromic repeat (CRISPR)/Cas9 system is a widely used genome-editing tool with great clinical potential. However, its application is limited because of low editing efficiency of some target sequences and off-target effects. As this system contains only the Cas9 protein and a single-guide RNA (sgRNA; engineered from crRNA and tracrRNA), the structure and function of these components should be studied in detail to address the current clinical needs. Consequently, we investigated the structural and sequence features of the core hairpin (the first stem loop of sgRNA) of SpCas9 sgRNA. We showed that the core hairpin structure of sgRNA is essential for SpCas9/sgRNA-mediated DNA cleavage and that the internal loop structure in the core hairpin plays a vital role in target DNA cleavage. We observed that the root stem structure within the core hairpin preferentially forms Watson-Crick base pairs and should be of a specific length to maintain an appropriate spatial conformation for Cas9 binding. However, the length of the leaf stem structure of the core hairpin is flexible, having a variable nucleotide composition. Furthermore, extension of the leaf stem structure enhances the DNA cleavage activity of the Cas9/sgRNA complex, and this could be used to enhance the efficiency of gene editing. These observations provide insight into the sgRNA/Cas9 interaction, indicating that sgRNA modification could be a strategy for improved DNA editing efficiency, and optimized sgRNA can be further used for genome-wide functional screening and clinical application.

Keywords
CRISPR/Cas9, screening, sgRNA, core hairpin

Introduction
The clustered regularly interspaced short palindromic repeats (CRISPR) system is an adaptive immunity system of bacteria and archaea that defends the host against invading foreign plasmids or viruses.¹,² Based on the number of effector proteins participating in the nucleic acid interference stage of CRISPR editing, two CRISPR classes are recognized: class 1 systems containing multiple effector proteins and class 2 systems, which include only one effector protein.³–⁵

The class 2 type II CRISPR/Cas9 system is the most extensively studied CRISPR system and the first-ever CRISPR system to be applied to genome editing in mammalian cells.⁶ Its functional complex includes the Cas9 protein, a CRISPR-related RNA (crRNA), and a transactivating crRNA (tracrRNA).⁷ Cas9 is a large multidomain protein that facilitates RNA binding and guides target DNA binding and cleavage. Further, crRNA hybridizes with tracrRNA to form double-stranded RNA. The sake of convenience of CRISPR/Cas9 use, crRNA and tracrRNA are linked by an artificial GAAA loop to form a single-guide RNA (sgRNA).⁷ The RNA duplex binds to the Cas9 protein, which results in a conformational change of the protein. Then, the ribonuclease protein (RNP) complex searches the genome for the cleavage site. This requires the presence of a protospacer-adjacent motif close to the target sequence, to differentiate the invading DNA from host DNA. After interaction with the protospacer adjacent motif, the RNP complex undergoes a structural change, the sgRNA target region pairs with the

¹State Key Laboratory of Natural and Biomimetic Drugs, Department of Biomedical Engineering, College of Engineering, Peking University, Beijing, China

Received Nov 28, 2019, and in revised form Apr 4, 2020. Accepted for publication Apr 8, 2020.

Supplemental material is available online with this article.

Corresponding Author:
Juan Li PhD, State Key Laboratory of Natural and Biomimetic Drugs, Department of Biomedical Engineering, College of Engineering, Peking University, Liaoakaiyuan Building 2-103B, NO.5 Yiheyuan Road, Haidian District, Beijing, China.
Email: lijuanlj@pku.edu.cn
complementary DNA sequence, and a double-strand break is induced at the DNA in the target locus. DNA double-strand breaks are then majorly repaired via nonhomologous end-joining or homology-directed repair, leading to DNA modification.9,9

High editing efficiency is essential for the application of CRISPR/Cas9, and it is extensively used in many research efforts.10,11 Previous studies mainly focused on the structure of the Cas9 protein, Cas9/sgRNA binary complex, and Cas9/sgRNA/DNA ternary complex, to elucidate the key functional sites and interactional mechanisms of the CRISPR/Cas9 system.10–12 Nevertheless, additional features of the sgRNA sequence and structure associated with the activity of Cas9/sgRNA complex are yet to be explored.

sgRNA consists of two regions: a 5′-target sequence and a 3′-scaffold sequence. The conserved modules of sgRNAs have been identified and characterized, and the orthogonality between sgRNAs and Cas9 proteins has been explored.13 Studies of sgRNA target sequence have identified target features for high editing efficiency and have established tools for target sequence selection.14–18 Furthermore, the sgRNA target sequence was truncated, sgRNA basic scaffold was artificially reformed, and sgRNA sequence was chemically modified to improve the genome-editing efficiency. In addition to the improvements of efficacy, studies of the sgRNA secondary structure helped to reduce the off-target effects of this editing system.19–26

Regardless of the above efforts, however, little is known about the features of the sgRNA scaffold sequence. Consequently, in the current study, we designed a screening system in Escherichia coli to investigate the sequence composition of the sgRNA core hairpin structure, which is the first stem loop in sgRNA and is the complementary region between crRNA and tracrRNA. We also explored the nucleotides of the core hairpin structure that influence the cleavage activity of SpCas9/sgRNA complex in mammalian cells. We subsequently demonstrated the feasibility of promoting the cleavage efficiency via sgRNA sequence optimization and presented modified sgRNAs that improve CRISPR/Cas9 genome-editing efficiency in mammalian cells. We anticipate that the presented findings will provide a strategy for improving the editing efficiency for therapeutic applications.

Materials and Methods

Plasmid Construction

SpCas9-D10A protein and mutant sgRNA library used for screening in E. coli were individually expressed under the control of the T7 promoter from the pACYCDuet-1 vector (Novagen, Darmstadt, Germany). SpCas9-D10A sequence was inserted into the plasmid using the restriction enzymes Pst I and Not I to generate pACYCDuet1-SpCas9-D10A. The original second T7 promoter on pACYCDuet-1 plasmid was deleted using Not I and Avr II to remove the unnecessary ribosome-binding site. Constructs containing T7 promoter and mutant sgRNA sequences were synthesized by using random nucleotides according to different designs of the experiments (Biomed, Beijing, China) and inserted into the pACYCDuet1-SpCas9-D10A plasmid using Not I and Avr II, generating pACYCDuet1-SpCas9-D10A-sgRNA libraries using Gibson Assembly Master Mix (New England BioLabs, Ipswich, MA). Single-strand annealing (SSA) red fluorescent protein (RFP) reporter plasmid was constructed based on vector pEDTduet-1 (Novagen). The incomplete RFP (396 bp) sequence was placed in front of two TAA stop codons and the target sequence. Next to the target sequence, an out-of-frame complete RFP sequence was inserted. The sequence encoding for wild-type SpCas9 protein was codon-optimized for mammalian cells by Lasergene. It was inserted in the expression vector pcDNA3.1 (Invitrogen, Carlsbad, CA) with restriction enzymes Kpn I and Xba I, and nuclear localization sequences were inserted at both the C- and N-termini of the protein. sgRNA was expressed under the U6 promoter from plasmid pGPU6/GFP/Neo (GenePharma, Shanghai, China). sgRNA was inserted in the vector using restriction enzymes Bbs I and BamH I. All sequences are provided in the Supplemental Note.

Core Hairpin Structure Screening in E. coli

E. coli BL21 strain was transformed with Cas9-D10A/sgRNA plasmid pool and the SSA RFP reporter plasmid by electroporation. Briefly, 150 μL of electrocompetent E. coli BL21 cells (Biomed) were mixed with 300 ng Cas9/sgRNA plasmid pool and 100 ng SSA RFP reporter plasmid and electroporated using Lonza Amaxa Nucleofector 2b (Program 7). subsequently, 500 μL of room-temperature SOC medium was added and the sample agitated vigorously (220 rpm) at 37 °C for 1 h. Next, 1 mM isopropyl β-D-1-thiogalactopyranoside was added, and the cells were cultured for another 1 h. The transformants were spread on LB agar plates supplemented with ampicillin and chloramphenicol and incubated at 37 °C; 18 h later, red clones were picked using a fluorescence microscope. The total plasmids were extracted (Biomed), and the RFP plasmids were degraded by Xho I digestion for a second round of screening. We then checked for complete elimination of RFP plasmids after transformation into BL21 cells. The process of the second screen was the same as the primary screen. The red clones were then picked and sequenced using Sanger sequencing (Biomed).

Mammalian Cell Culture and Transfection

HEK293T cells were cultured in high-glucose Dulbecco’s modified Eagle medium (Gibco, Carlsbad, CA) supplemented
products were denatured by incubating at 95 °C for 5 min with 10% fetal bovine serum in an incubator set at 37 °C and under 5% CO₂ atmosphere. The cells were transfected in 48-well plates using Lipofectamine 2000 (Thermo Fisher, Waltham, MA) with 300 ng Cas9 expression plasmid and 100 ng sgRNA plasmid. The cells were harvested 72 h after transfection for T7 endonuclease I (T7EI) assay.

**T7EI Assay**

For the experiment, 72 h posttransfection, genomic DNA was extracted from HEK293T cells using TIANamp Genomic DNA kit (TIANGEN, Beijing, China). The T7EI experiment was performed as previously described with some modification.27 Briefly, Q5 high-fidelity DNA polymerase (New England BioLabs) was used to amplify genomic sequence containing the target sequence using 100 ng genomic DNA as a template. *HBB* gene primers are shown in the Supplemental Note. The PCR products were analyzed by agarose gel electrophoresis and then purified using the EasyPure Quick Gel Extraction kit (Transgene, Beijing, China). Then, 200 ng PCR products were denatured by incubating at 95 °C for 5 min and reannealed to form heteroduplex DNA. This was followed by digestion with T7EI (New England BioLabs) at 37 °C for 10 min. The digestion products were purified and quantified using Quantity One to determine the editing efficiencies of the Cas9/sgRNA complex.

**Statistical Analysis**

Data are represented as mean ± SD of two or three independent experiments. An unpaired two-tailed *t* test was applied to analyze the *p* value between the two groups (*p* < 0.05, **p** < 0.01).

**Results and Discussions**

**Screening of sgRNA Core Hairpin Sequence**

**Features in E. coli**

According to previous studies, the unique structure of sgRNA allows it to bind Cas9 protein.10,13 However, little is known about the specific features of sgRNA. Based on its RNA secondary structure, we divided the sgRNA structure into three regions: the spacer region, core hairpin structure, and other hairpin structures. The core hairpin structure refers to the complementary region between crRNA and tracrRNA, and it contains three modules: root stem, internal loop, and leaf stem. The secondary structure of the sgRNA core hairpin was predicted by using RNAfold Server (Fig. 1A).28 We chose *E. coli* as the screen host because of the ease of culture. Further, each *E. coli* clone contained only one type of plasmid with the same replicon, which facilitated obtaining the positive clone.

Accordingly, we performed a large-scale screening of the features of the sgRNA core hairpin sequence by using an SSA-based RFP reporter (Fig. 1B). SSA is also a type of repair pathway of DNA break; it uses the resected ends to anneal exposed complementary sequences to repair.29 For the experiment, sgRNA and Cas9 were expressed separately from the pACYCDuet-1 plasmid under the control of the T7 promoter. The resistance gene on the pACYCDuet-1 was different from that on pETDuet-1 carrying the RFP reporter. Initially, we co-electrotransformed *E. coli* strain BL21 with wild-type SpCas9/sgRNA plasmid and the SSA RFP reporter plasmid, but the approach yielded no clones (data not shown). We speculated that the repair efficiency of the double-strand break in *E. coli* was low, which was consistent with previous research.30 We therefore repeated the experiment using SpCas9 with a D10A substitution. This protein only nicks the target DNA,7,31 and we used SpCas9-D10A for subsequent screening.

The screening process involved four steps: sgRNA library construction, primary screen, secondary screen, and sequencing (Fig. 1C). First, a group of mutant sgRNA libraries (design A–R) were constructed, as shown in Figure 1E and Supplemental Figure S1. All plasmids encoding mutant sgRNA and SSA RFP reporter plasmid were co-electrotransformed into *E. coli*. If a mutated sgRNA could interact with the Cas9 protein to nick the target site, the SSA RFP reporter plasmid would be repaired, with red fluorescence emitting upon green light excitation (Fig. 1B, D). From among 16,854 clones, more than 700 red clones were picked in the primary screen, as we wanted to ensure the authenticity of the results and reduce the ratio of false-positives. Prior to the secondary screen, all sgRNA plasmids from these red clones were extracted and mixed. The mixtures also contained RFP reporter plasmids. Therefore, the RFP reporter plasmids were first removed by endonuclease digestion, and then all sgRNA plasmids were used to transform *E. coli* for the secondary screen. Finally, 193 red clones were obtained, and 162 sgRNA sequences were confirmed by Sanger sequencing (Suppl. Table S1 and S3). We also verified the screening results using the T7EI assay in mammalian cells (Suppl. Fig. S2). This proved the accuracy of our screening in *E. coli*.

We then used statistics to analyze the base pairing status of all identified positive hits. A heat map was constructed to show the base-pairing pattern in Design A–E. Most of the bases preferentially formed pairs in the root stem (Design A–C), indicating that the root stem structure is necessary for the sgRNA function (Fig. 1E; Table 1). Two sites close to the internal loop, 7/24 and 9/20 (bases 7 and 24, or 9 and 20, accordingly), were preferentially engaged in weak or non–Watson-Crick base-pairing interactions, whereas the other sites in the root stem tended to show high Watson-Crick base-pairing tendency. In contrast, the leaf stem (Design D, E) was not inclined to be rigorously base paired as compared with the root stem.
The distribution of paired bases in the sgRNA core hairpin also depends on the structure of the root stem and leaf stem. We observed a relatively low proportion of purines in the 5′-end of the sgRNA core hairpin and a relatively high proportion of purines in the 3′-end of the sgRNA core hairpin (Fig. 1F). The complementarity of the composition of the 5′- and 3′-ends of the sgRNA core hairpin sequence coincided with the Watson-Crick base-pairing principle. We also observed two sites with disparate base composition: all sgRNAs harbored a purine at position 21, and all sgRNAs harbored a pyrimidine at position 24 (Fig. 1F). We speculated that these two sites are important for sgRNA function and should be further investigated.

We also investigated the loop size and composition of the internal loop and the leaf stem. We inserted additional nucleotides into the internal loop and the leaf stem loop, as
part of the screening results of Design H to Design R (Suppl. Table S3). We observed that the presence of additional nucleotides in the internal loop and leaf stem did not appreciably affect the Cas9/sgRNA complex. Sequencing results of Design P to Design R also exhibited that there might be mismatched base pairs in the leaf stem, which indicated that the leaf stem could be modified for specific applications.

Collectively, the core hairpin structure of sgRNA tended to form stable Watson-Crick base pairs in the root stem but had a loose conformation in the leaf stem. The bases adjacent to the internal loop had tendency to be unpaired. Although it is not accurate to evaluate the editing efficiency of different sgRNA scaffolds in E. coli as fluorescence intensity is slightly affected by clone size or bacterial growth, our screening method in E. coli can also be used to screen for mutant Cas9 protein with PAM sequences other than NGG, target sequence preferences for gene editing, and so on.

Internal Loop of the Core Hairpin Structure Is Critical for the DNA-Editing Activity of Cas9/sgRNA Complex

Because the editing efficiency of different sgRNAs could not be compared using screening results in E. coli, we investigated the editing efficiency of different sgRNAs by using the T7EI assay in HEK293T cells. We chose a target on the HBB gene for Cas9/sgRNA cleavage in mammalian cells (Supplemental Data). For the internal loop, we introduced changes by mutation, deletion, and insertion of nucleotides.

First, we replaced the internal loop 8A:21AAG with 8AC=21AG and 7G with 7A to destabilize the internal loop structure. We used the T7EI assay to evaluate DNA cleavage activity of the Cas9/sgRNA complex in HEK293T cells. The results revealed that the complex failed to cut the relevant genomic target in this scenario, indicating that the internal loop was important for the activity of Cas9/sgRNA complex (Fig. 2A). As internal loops expand the groove of the RNA stem and provide necessary space for amino acids of a protein to interact with nucleotides via base stacking or hydrogen bonding, we speculated that the groove of sgRNA was necessary for interaction with the Cas9 protein, even though perfectly matched base pairs in the stem might facilitate base stacking to maintain a steady secondary structure of sgRNA.

We then altered the spatial conformation of the internal loop. We reversed the orientation of the internal loop by swapping nucleotides 8A and 21AAG. T7EI assay results suggested that the Cas9/sgRNA complex lost its DNA cleavage activity when sgRNA conformation was altered, emphasizing that the conformation of the groove formed by the internal loop was crucial for the interaction between Cas9 and sgRNA (Fig. 2B).

We also investigated the sequence of the internal loop. We replaced nucleotides 8A and 21AAG with 8T and 21TTC, respectively. This eliminated the DNA cleavage activity of the Cas9/sgRNA complex (Fig. 2B). To identify the key nucleotides in the internal loop, we deleted or altered nucleotide 8A by individually replacing it with the other three nucleotides. Based on the T7EI analysis, neither 8A deletion nor 8A replacement by other nucleotides negatively affected the activity of the Cas9/sgRNA complex. In fact, replacing 8A with 8C slightly enhanced the cleavage of target DNA (Fig. 2C). These observations suggested that nucleotide 8A was not important for the interaction between sgRNA and Cas9 protein. Further, deletion of site 21AAG resulted in the loss of DNA cleavage activity. Surprisingly, deletion of only 21A or 21AA did not affect the cleavage function (Fig. 2C). This indicated that nucleotide 23G was vital for sgRNA function. We next investigated the role of nucleotide 23G. First, its deletion resulted in a negative T7EI reading. We then replaced 23G with the other three nucleotides. Replacing 23G with 23C significantly reduced the DNA cleavage activity, whereas no significant changes in

<table>
<thead>
<tr>
<th>Design</th>
<th>Clone Number</th>
<th>Pair Sitea</th>
<th>Clones of Pair</th>
<th>Base-Pair Probabilityb</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>6</td>
<td>12=17</td>
<td>3</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>11=18</td>
<td>4</td>
<td>0.667</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>10=19</td>
<td>2</td>
<td>0.400</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>9=20</td>
<td>1</td>
<td>0.200</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>7=24</td>
<td>1</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6=25</td>
<td>5</td>
<td>1.000</td>
</tr>
<tr>
<td>B</td>
<td>17</td>
<td>5=26</td>
<td>12</td>
<td>0.706</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>4=27</td>
<td>13</td>
<td>0.765</td>
</tr>
<tr>
<td>A</td>
<td>7</td>
<td>3=28</td>
<td>5</td>
<td>0.714</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2=29</td>
<td>7</td>
<td>1.000</td>
</tr>
</tbody>
</table>

sgRNA, single-guide RNA.  
*aThe symbol “=” represents the base pair between interactive nucleotides.  
*bProbability is clones of pair divide clone number.
DNA cleavage activity were apparent after 23G to 23A and 23G to 23U replacements (Fig. 2C). We also replaced 21A with the other three nucleotides, but none of these changes significantly altered the Cas9/sgRNA complex activity (Fig. 2C). Collectively, these observations indicated that nucleotide 23G might directly interact with the Cas9 protein. Combined with the previous results, these observations suggested that the Cas9/sgRNA complex relies on both the sequence and appropriate conformation of the internal loop of the sgRNA core hairpin to exert its function.

Finally, we enlarged the size of the internal loop structure by adding one or more nucleotides in the loop region. We inserted one G in 21AAG, one A before 8A, and two As before 8A. Consequently, the DNA-editing efficiency was reduced to 18.9%, 11.2%, and 10.1%, respectively. By comparison, the editing efficiency of wild-type sgRNA was 27.8% (Fig. 2D). When a single A was inserted both before 8A and 21AAG in the internal loop, the editing efficiency was reduced to 11.1% (Fig. 2D). These observations demonstrated that enlarging the internal loop weakened the

---

**Figure 2.** Features of the internal loop structure of single-guide RNA (sgRNA) core hairpin. (A) T7EI results of sgRNA core hairpin without internal loop. Left is the predicted RNA secondary structure by RNAfold. The dashed box represents the mutant region of the core hairpin structure. (B) T7EI results of the internal loop with spatial configuration or sequence changes. Left are the predicted RNA secondary structures by RNAfold. (C) T7EI results of the mutation of different sites in the internal loop. (D) T7EI results of the internal loop with loop enlarged. Left are predicted RNA secondary structures by RNAfold.
binding between the sgRNA and Cas9 protein, subsequently reducing the editing activity of the Cas9/sgRNA complex.

Collectively, the presence and appropriate conformation of the internal loop in the sgRNA core hairpin was necessary for the DNA-editing activity of the Cas9/sgRNA complex. The functioning of the Cas9/sgRNA complex also depended on the internal loop sequence, and nucleotide in position 23 appeared to bind certain amino acids of the Cas9 protein directly. Indeed, recent studies of the structure of the Cas9/DNA/sgRNA ternary complex revealed that the guanine nucleotide at position 23 forms hydrogen bonds with Phe351 and Asp364 of Cas9 in a base-specific manner, confirming the above findings.

Cas9 Recognition of sgRNA Depends on the Appropriate Spatial Conformation and Specific Sequence of the Root Stem

The root stem of sgRNA core hairpin is a U-A base-pair repeat region close to the sgRNA target sequence. The root stem region is not required for Cas9 DNA cleavage activity in vitro. Considering the conserved nature of crRNA and tracrRNA sequences, we speculated that the U-A repeat region in the core hairpin binds to the Cas9 protein directly. To investigate the role of this region in DNA target cleavage, we shortened it. Initially, we removed one base pair at a time from the U-A repeat region (Fig. 3A). We observed that Cas9/sgRNA complex lost the DNA cleavage activity when any single base pair in the root stem region was removed. We also shortened the root stem by replacing successive double base pairs with a single base pair. The Cas9/sgRNA complex was inactive in all of these scenarios (Fig. 3B), which suggested that the interaction between Cas9 protein and sgRNA required the root stem to be sufficiently long. We then inserted several base pairs into the root stem to increase its length. As above, the insertion of base pairs resulted in the loss of DNA cleavage activity (Fig. 3C), which supported the notion of the specific length requirement of the stem sequence for the Cas9 and sgRNA interaction. Another possible explanation for this phenomenon was the requirement for specific nucleotide sequence. According to the structural analysis of the Cas9/sgRNA complex, nucleotides at sites U3 and A29 of the sgRNA root stem bind Arg1122 of Cas9 protein via a hydrogen bond, and Cas9 Rec and bridge helix domains interact with nucleotides from U25 to U50 in the sgRNA backbone. Collectively, we speculated that the root stem of the sgRNA core hairpin embedded itself in the groove of the Cas9 protein in a conformation- and sequence-dependent manner.

Because the screening analysis in E. coli suggested a preference for perfect Watson-Crick base pairing in the sgRNA root stem, we next explored the sequence features of the root stem. We replaced nucleotides at position 4U or 5U with A, G, or C; each replacement resulted in an unpaired nucleotide loop at that site. Changing 4U to A or G reduced the DNA-editing activity slightly, whereas changing 4U to C and 5U to A, G, or C increased the DNA-editing efficiency, based on the results of the T7EI assay (Fig. 3D). We obtained similar results after replacing 2U with C, 3U with A, and 6A with C (Fig. 3D). These observations suggested that sgRNA could tolerate one mismatched base pair in the root stem.

In the RNA secondary structure, G and U may be paired to maintain stability. Initially, we speculated that nucleotides G7 and U24 (at the edge of the root stem near the internal loop) could adopt wobble pairing or Watson-Crick pair and that replacing nucleotides at these locations would enhance the stability of sgRNA, thus increasing the editing activity of the Cas9/sgRNA complex. However, we observed that the replacement of both G7 and U24 with Watson-Crick base pairs (G7=G24) and opposite wobble pair (U7=G24) impaired the activity of the complex (Fig. 3E). We also performed a single-nucleotide exchange of G7 to U7, which was mismatched with U24; this change increased the editing efficiency of the Cas9/sgRNA complex (Fig. 3E). These observations led us to speculate that G7 and U24 did not pair in a wobbly manner but were separated to contact the amino acids of the Cas9 protein. However, the T7EI assay involving the U7=A24 replacement revealed that the construct had a similar editing efficiency as that of the wild-type sgRNA. We speculated that U7 and A24 did not form a pair while sgRNA was folding. According to the crystal structure of the Cas9 and sgRNA complex, nucleotide U44 faces outwards to the RNA helix and is embedded between the side-chain radical of Tyr325 and His328 for a stable-stack effect, and therefore it requires an unpaired conformation at site 7 and site 24 of the core hairpin.

To sum up the presented findings thus far: (1) the root stem of the sgRNA core hairpin could tolerate one base-pair mismatch and remain functional, (2) the root stem had to be exactly six base pairs in length to embed itself in the Cas9 protein groove, (3) nucleotides G7 and U24 did not form a wobble pair or a Watson-Crick pair, and (4) the base of G7 faced outward of the RNA helix axis to contact amino acids of the Cas9 protein.

Length of the Leaf Stem Affects the Activity of Cas9/sgRNA Complex

Jinek et al. demonstrated that the leaf stem of the sgRNA core hairpin is vital for Cas9-mediated DNA cleavage in vitro. Other studies similarly indicated the importance of the sequences near the loop region for the RNP complex. To investigate the function of the leaf stem in the sgRNA core hairpin in vivo, we designed a set of sgRNAs with leaf stem sequences that were either longer or shorter than the wild-type sgRNA. Notably, we observed that when the leaf stem length was shortened by one base pair, the DNA editing efficiency was higher than that with wild-type sgRNA. Further, instead
of losing DNA cleavage activity, sgRNA lacking the entire leaf stem structure also maintained DNA cleavage activity in mammalian cells (Fig. 4A). All of the tested mutant leaf stem sequences are listed in Supplemental Table S2. These observations were different from those obtained by Jinek et al., as the findings of the current study suggested that the leaf stem in the sgRNA core hairpin is not requisite for the function of Cas9/sgRNA complex in mammalian cells. Based on the Cas9/sgRNA structure, the Cas9 protein does not directly contact the leaf stem of sgRNA. We proposed that the leaf stem might function to promote the correct sgRNA folding before binding to Cas9 in vitro. Of note, mammalian cells contain chaperon proteins that assist in the correct RNA folding, even those with sequence mutation. Hence, sgRNA lacking the leaf stem sequence may fail to fold correctly in the absence of chaperone proteins in vitro. This could explain the discrepancy in the results from the in vivo and in vitro experiments. However, the reason behind the increased DNase activity of the Cas9/sgRNA complex following the loss or shortening of the leaf stem requires further study, as differences between cell-based cleavage activity are due to RNA folding, stability, complex formation, and/or stoichiometric differences in mammalian cells. Considering the assistance of the leaf stem in sgRNA folding, we hypothesized that a relatively long leaf stem may promote thermodynamic stability of sgRNA after folding, even those with sequence mutation. Hence, sgRNA lacking the leaf stem sequence may fail to fold correctly in the absence of chaperone proteins in vitro. This could explain the discrepancy in the results from the in vivo and in vitro experiments. However, the reason behind the increased DNase activity of the Cas9/sgRNA complex following the loss or shortening of the leaf stem requires further study, as differences between cell-based cleavage activity are due to RNA folding, stability, complex formation, and/or stoichiometric differences in mammalian cells. Considering the assistance of the leaf stem in sgRNA folding, we hypothesized that a relatively long leaf stem may promote thermodynamic stability of sgRNA after folding, even those with sequence mutation. Hence, sgRNA lacking the leaf stem sequence may fail to fold correctly in the absence of chaperone proteins in vitro. This could explain the discrepancy in the results from the in vivo and in vitro experiments. However, the reason behind the increased DNase activity of the Cas9/sgRNA complex following the loss or shortening of the leaf stem requires further study, as differences between cell-based cleavage activity are due to RNA folding, stability, complex formation, and/or stoichiometric differences in mammalian cells. Considering the assistance of the leaf stem in sgRNA folding, we hypothesized that a relatively long leaf stem may promote thermodynamic stability of sgRNA after folding, even those with sequence mutation. Hence, sgRNA lacking the leaf stem sequence may fail to fold correctly in the absence of chaperone proteins in vitro. This could explain the
observations indicated the possibility of enhancing the DNA-editing activity of the Cas9/sgRNA complex by shortening or extending the length of leaf stem.

In our sgRNA sequence screening in *E. coli*, we observed that the leaf stem did not tend to form Watson-Crick base pairs as compared with the root stem. We next tested the impact of inserting a loop structure into the leaf stem in mammalian cells. We added an additional internal loop structure into the leaf stem region based on the more effective sgRNA structure with a 4 bp leaf stem extension (Fig. 4B). We found that the additional internal loop did not affect the editing efficiency. Instead, the editing efficiency was improved, which meant that not only the length but also the structure was important for the RNP complex. Future work may focus on the structure modification of the leaf stem region for a high editing efficiency sgRNA scaffold.

We also attempted to alter the sequence of the leaf stem. We exchanged the wild-type sgRNA leaf stem 9GCUA=17UAGC with 9CUGA=17UCAG and found that the mutant sgRNA was still functional in mammalian cells (Fig. 4C). The editing efficiency was comparable with that of wild-type sgRNA scaffold. Overall, the presented findings demonstrated the conformation and sequence flexibility of the leaf stem and that sgRNAs, which are highly efficient in Cas9-mediated DNA cleavage, could be obtained by modifying the leaf stem sequence.

**Conclusions**

The CRISPR/Cas9 system has been successfully implemented to execute sequence-specific cleavage of target DNA in a variety of mammalian cells and model organisms, including *Drosophila*, zebrafish, frog, mouse, rat, and monkey. This DNA-editing approach has a great potential for disease treatment, but its clinical applications are limited by safety concerns and low efficiency. In the current study, we aimed to explore the sequence features of the sgRNA core hairpin structure, trying to further study the interaction between Cas9 and sgRNA. We showed that sgRNA could tolerate modification in some regions. These observations may be helpful for improving the editing efficiency of CRISPR/Cas9 system.

The structure of the Cas9/sgRNA complex has been recently analyzed, and the possible mechanism of target
DNA cleavage by the RNP complex has been deciphered. In the current study, we systematically analyzed the structural and sequence features of SpCas9 sgRNA core hairpin, in an attempt to identify new approaches for its optimization. The secondary structure of sgRNA contains many stem-loop structures. We focused on the core hairpin, which is the complementary region between the crRNA and tracrRNA. We divided the sgRNA core hairpin structure into three parts, according to its secondary structure: the root stem, the internal loop, and the leaf stem. We first established a screening system in E. coli to explore these sequence features. We selected different parts of the core hairpin and randomly synthesized mutant sequences to construct a mutant sgRNA library. To improve the screening accuracy, we performed two rounds of screening. We analyzed the effective clones and observed that Watson-Crick base pairs were typically present in the root stem region of the core hairpin structure and that there was a potential for sgRNA structure optimization in the upper leaf stem region of the core hairpin structure.

As it may be difficult to evaluate the editing efficiency of mutant sgRNAs in E. coli, we used the T7EI assay to test the cleavage efficiency of different sgRNAs in mammalian cells. We found that the internal loop contributes to sgRNA conformation and identified an important interaction position: the site 23G. In mammalian cells, the number of base pairs in the root stem region is fixed, only six base pairs in length, to maintain the interaction with Cas9 protein. Increasing or decreasing the number of base pairs will affect the cleavage activity of the complex. By contrast with the E. coli screening data, which indicated that the root stem tends to contain completely complementary base pairs, mismatches were tolerated in the root stem region in mammalian cells. According to the presented investigation of the leaf stem in the core hairpin structure, the matched base pairs are dispensable for the interaction with Cas9 protein, and although altering the number of base pairs affected the Cas9 cleavage activity, insertion of an unmatched bulge structure did not inactivate the functional RNP complex. We also found that the leaf stem could be potentially modified, which suggests that the region tolerates artificial reforming. According to other studies, MS2 RNA sequence can be inserted into this region, to allow the recruitment of MS2 coating protein without affecting the Cas9/gRNA complex assembly, which is consistent with the data from the current study.

We were unable to analyze the complete sgRNA structure because of the time- and cost-consuming nature of the experiments, but it remains a viable direction for further investigations. Further, we believe that aside from the target sequence, the sgRNA structure also affects the DNA-editing efficiency, and it is of great importance to develop screening systems to analyze other important sites of sgRNA. In recent years, an increasing number of CRISPR systems have been identified and exploited, such as the CRISPR/Cpf1 and CRISPR/C2C1 systems. We anticipate that future studies will focus on engineering the CRISPR system by altering the wild-type CRISPR crRNA to further improve the efficiency of editing and facilitate clinical applications.

Acknowledgments

We thank Hanshuo Zhang, Yuyang Dong, and Huinan Lu for their suggestions on the experiments.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References


