350–361 Nucleic Acids Research, 2018, Vol. 46, No. 1 Published online 14 November 2017
doi: 10.1093/nar/gkx1117

The initiation, propagation and dynamics of CRISPR-SpyCas9 R-loop complex

Yan Zeng[1,†], Yang Cui[1,2,†], Yong Zhang[1], Yanruo Zhang[1,2], Meng Liang[1,2], Hui Chen[1,2], Jie Lan[1,2], Guangtao Song[1,*] and Jizhong Lou[1,2,*]

[1]Key Laboratory of RNA Biology, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China and [2]University of Chinese Academy of Sciences, Beijing 100049, China

Received September 01, 2017; Revised October 20, 2017; Editorial Decision October 20, 2017; Accepted October 25, 2017

ABSTRACT

CRISPR-Cas9 system has been widely used for efficient genome editing. Although the structures of Cas9 protein in complex with single-guided RNA (sgRNA) and target DNA have been resolved, the molecular details about the formation of Cas9 endonuclease R-loop structure remain elusive. Here we examine the DNA cleavage activities of *Streptococcus pyogenes* Cas9 (SpyCas9) and its mutants using various target sequences and study the conformational dynamics of R-loop structure during target binding using single-molecule fluorescence energy transfer (smFRET) technique. Our results show that Cas9–sgRNA complex divides the target DNA into several distinct domains: protospacer adjacent motif, linker, Seed, Middle and Tail. After seed pairing, the Cas9 transiently retains a semi-active conformation and induces the cleavage of either target or non-target strand. smFRET studies demonstrate that an intermediate state exists in prior to the formation of the fully stable R-loop complex. Kinetics analysis of this new intermediate state indicates that the lifetime of this state increases when the base-pairing length of guide-DNA hybrid duplex increases and reaches the maximum at the size of 18 bp. These data provide new insights into the process of R-loop formation and reveal the source of off-targeting in CRISPR/Cas9 system.

INTRODUCTION

Clustered, Regularly Interspaced, Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins are widely found in bacteria and archaea genomes as adaptive immune systems against invasive genetic elements (1–8). Recently, these systems have attracted much attentions due to the discovery that CRISPR-Cas9, a class II CRISPR-Cas system, can be re-purposed as a powerful RNA-guided DNA targeting platform to easily create deletions, insertions and replacements in mammalian genome (9–13). In this widely used CRISPR-Cas9 system, *Streptococcus pyogenes* Cas9 (SpyCas9) functions together with a chimeric single-guide RNA (sgRNA) comprising CRISPR RNA (crRNA) and transactivating crRNA (tracrRNA) modules to bind with 20 bp DNA target sequence and induce DNA double-strand break. SpyCas9 contains two nuclease domains, a RuvC-like (RNase H fold) domain and an HNH (McrAlike fold) domain, for target DNA cleavage (9). The tracrRNA module in sgRNA serves as a scaffold around which Cas9 can fold into an active conformation, and the crRNA module guides Cas9 protein to cleave target DNA with sequence specificity provided by base-pairing between the 20 nt guide sequence and the target DNA (9,14). In addition, the efficient target cleavage by Cas9 requires the presence of a short protospacer adjacent motif (PAM) that flanks the target region (15).

Despite the wide usage of the powerful CRISPR-Cas9 system in genome editing, numerous studies have assessed off-target DNA binding and cleavage by the Cas9–RNA complex, both *in vitro* and *in vivo* (16–18). In the past few years, the field of CRISPR-Cas9 specificity is rapidly evolving, with several important improvements achieved in guide selection and Cas9 engineering (18–22). However, their overall on-target efficiency relative to the wild-type (WT) enzyme system remains to be studied in different applications and organisms. Efforts to understand the molecular mechanism of DNA targeting and cleavage of Cas9, and improve its specificity and efficiency have accordingly been of great interests in this area. Recent biochemical and structural studies of SpyCas9 have provided much insights into the RNA-guided DNA cleavage mechanism of the Cas9 enzymes (23–28). Upon complexed with sgRNA, Cas9 adopts a pre-ordered conformation, its PAM recog-

[*]To whom correspondence should be addressed. Tel: +86 10 6486 8183; Fax: +86 10 6486 8183; Email: jlou@ibp.ac.cn
Correspondence may also be addressed to Guangtao Song. Tel: +86 10 6488 9893; Fax: +86 10 6488 9893; Email: gsong@moon.ibp.ac.cn
[†]These authors contributed equally to the paper as first authors.

nition region initiates DNA interrogation by sampling the genome for PAM matches by 3D diffusion (29). Conformational changes of Cas9 upon initial DNA binding then accommodate guide RNA strand and trigger the formation of an R-loop in which the guide region of the crRNA invades the dsDNA target to form an RNA:DNA hybrid with the complementary strand, displacing the opposing non-complementary strand. This R-loop structure, which resembles those found within the transcription bubble, is stabilized by Cas9 protein through interacting with the non-target strand and RNA:DNA hybrid duplex (26). Recent smFRET experiments indicated that as short as 9 base-pairing in length between guide RNA and target strand is sufficient for stable target binding (30). When the size of R-loop reaches 14–17 bp, Cas9 undergoes additional structural changes and reach a cleavage competent state, in which the HNH domain rotates by around 180° and translate around 2 nm toward the RNA:DNA hybridduplex (31,32). The HNH and RuvC domains then simultaneously cleave the target and non-target strand, respectively (31). In spite of these advances, the structural dynamics of Cas9 complexes and the molecular details about its target recognition, kinetics, dynamics and cleavage process are still not fully understood (Figure 1A). In this work, by combined use of biochemical, smFRET, and molecular dynamics (MD) simulation tools, we study the structural dynamics of Cas9 complex and the process of R-loop formation to probe and characterize features of CRISPR-Cas9 system in genome editing.

MATERIALS AND METHODS

Recombinant SpyCas9 expression and purification

SpyCas9 and all other mutants were cloned into a custom pET-based expression vector encoding an N-terminal GST-tag followed by His6-tag and a thrombin protease cleavage site. Point mutations were introduced into SpyCas9 using site-directed mutagenesis and verified by DNA sequencing. Cas9 was expressed in *Escherichia coli* Rosetta 2 (DE3) (Novagen) and purified by chromatography on Ni-NTA Superflow (QIAGEN). After removing the tags by thrombin at 4°C overnight, cation exchange (SP sepharose) chromatography was used for further purification steps.

In vitro transcription and purification of RNA

sgRNAs were in vitro transcribed using T7 *in vitro* transcription kit (MEGAscript T7, Invitrogen) and polymerase chain reaction (PCR) generated DNA templates carrying a T7 promoter sequence. RNA product was gel-purified before use.

Plasmid DNA cleavage assays

pUC19 based protospacer plasmids for *in vitro* cleavage assays were generated by annealed oligonulceotides between digested EcoR I and Hind III sites in PUC19. Ligated plasmids DNA were transformed into *E. coli* DH5a cells according to a standard heat shock protocol. Plasmid cleavage assays were carried out by using a custom-designed plasmid containing a 20-bp DNA target sequence and a 5′-TGG-3′ PAM motif. Purified SpyCas9 (final concentration 125 nM) and synthesized sgRNA (final concentration 375 nM) were pre-incubated in the cleavage buffer (20 mM HEPES, 100 mM KCl, 2 mM $MgCl_2$) at room temperature for 15 min, 500 ng supercoiled plasmid were then added to the reaction mixtures and incubated at room temperature for various time points. The reactions were stopped with 6× sodium dodecyl sulphate loading buffer prior to loading on the agarose gel. Cleavage products were run on 1.2% agarose gels (1 × TAE buffer) and Midori Green (NIPPON Genetics Europe) staining. All experiments were replicated at least three times, and presented results are representative replicates. The gel images were further analyzed with ImageJ software. The cleavage rate constants (k) in Figure 5A were obtained from single-exponential fits of the relative amount of cleavage product at different time points. DNA cleavage experiments presented in Figures 3 and 5C were performed at 37°C for 2 h and contained 300 ng DNA and 200 nM Cas9–sgRNA complex.

Electrophoretic mobility shift assays

Alexa-488 labeled DNA duplexes were prepared by PCR amplification of above plasmid DNA target with 5′-end Alexa-488 labeled primers. Binding assays were performed in buffer containing 20 mM HEPES pH 7.5, 100 mM KCl, 2 mM $MgCl_2$, in a total volume of 20 μl. dCas9 mutant was programmed with equal molar amounts of sgRNA and titrated from 100 pM to 1 μM. Alexa488-labeled DNA was added to a final concentration of 0.1 μM. Samples were incubated for 1 h at 37°C and resolved at 4°C on an 15% native polyacrylamide gel containing 1× TBE and 2 mM $MgCl_2$. DNA was visualized by phosphor imaging, quantified with ImageQuant (GE Healthcare).

Single-molecule FRET experiments

All DNA oligonucleotides used for single-molecule experiments were purchased from Invitrogen (Thermo Fisher Scientific, Shanghai). The entire panel of DNA targets used in our measurements is available in Supplementary Table S1. DNA targets were prepared by mixing the two complement strands and heating to 95°C for 5 min followed by cooling to room temperature over 1 h. Cy3/Cy5-labelled DNA targets were immobilized on the polyethylene glycol-passivated cover-glass surface using streptavidin–biotin interaction. We acquired fluorescence data using an objective-based TIRF microscope as previous described (33). Imaging was performed at room temperature in a buffer (20 mM HEPES, pH 7.5, 100 mM KCl, 2 mM $MgCl_2$. An oxygen scavenging system (2 units μl^{-1} glucose oxidase, 20 units μl^{-1} catalase, 0.8% β-D-glucose and 2 mM Trolox (Sigma-Aldrich)) was used in all experiments to prevent the organic fluorophores from severe photo-fatigue (34). The time resolution for all the experiments was 50 ms. Detailed methods of smFRET data acquisition and analysis were described in previous studies (33). The FRET efficiency of a single molecule was approximated as $FRET = I_A/(I_D+I_A)$, where I_D and I_A are the background and leakage-corrected emission intensities of the donor and acceptor, respectively. The

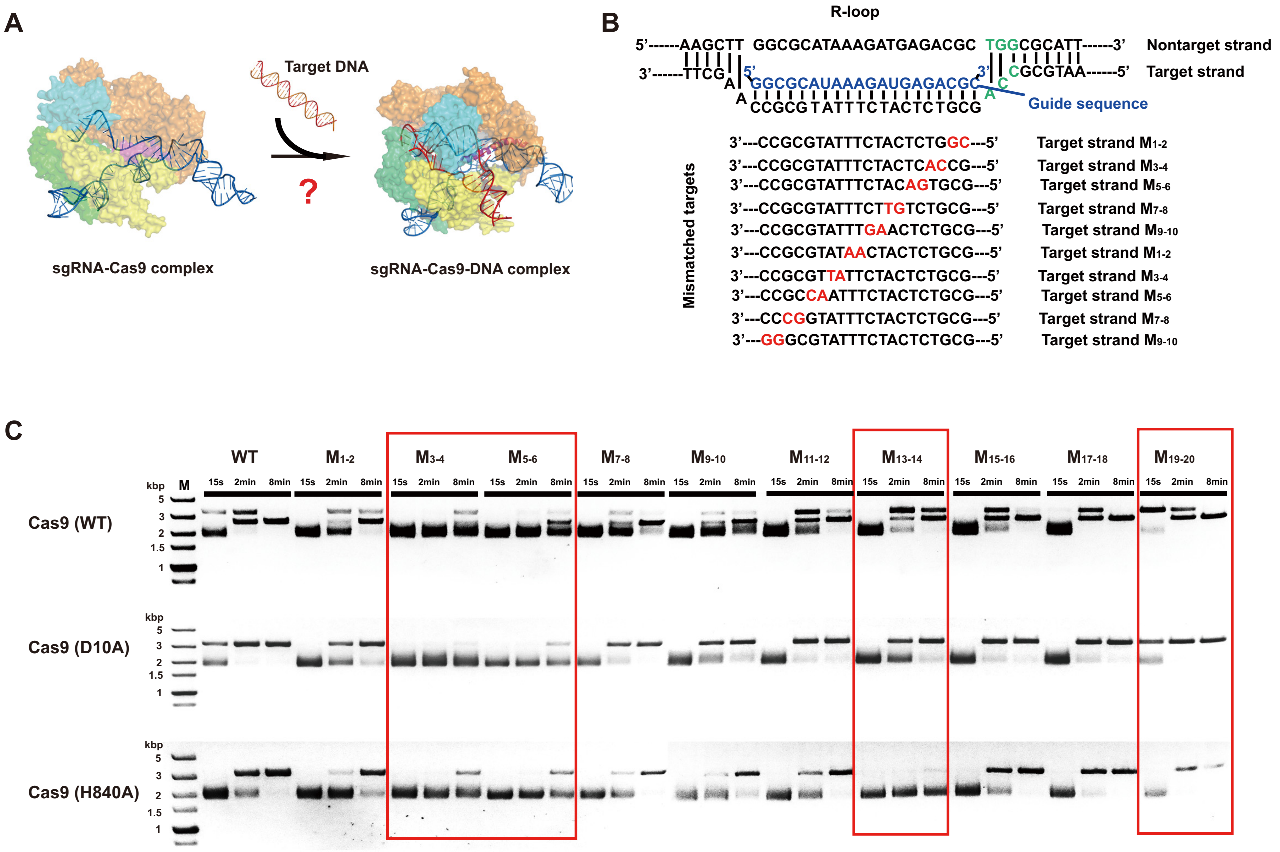

Figure 1. (**A**) Comparison between the Cas9 complex before and after target DNA binding. (**B**) Schematic representation of guide RNA and protospacer DNA sequences; The PAM sequence is shown in green; (**C**) Plasmids containing WT or mutant protospacer sequences shown in (B) were cleaved *in vitro* by Cas9–sgRNA complex.

histograms of the FRET upon Cas9–sgRNA binding were obtained by averaging the first 20 frames of each FRET trace for every individual molecule after manually filtering photo-bleaching effects. To determine the transition of FRET states and to calculate the transition rates between the states, we used hidden Markov modeling based on variational Bayesian expectation (35).

Molecular dynamics (MD) simulations

The initial structure of the Cas9–sgRNA–NA complex is from Protein Data Bank (5F9R). The complex model was solvated in a ∼1.2 × 1.4 × 1.6 nm^3 water box, which included K^+ and Cl^- ions (∼0.15 M) to neutralize the system. Amber ff14SB force field is applied for the protein and nucleic, including bsc0 and χOL3 modification for RNA and ϵ/ζOL1 and χOL4 modification for DNA (36,37). Under periodic boundary condition, a 12 Å cutoff (switching 10–12 Å) was used for van der Waals interactions, and Particle Mesh Ewald summation was used to calculate the electrostatic interactions. The NAMD package was used for energy minimizations and MD simulations. After multistep energy minimization to avoid possible clashes, the system was then equilibrated in three steps: 2 ns simulation with strong constraint of heavy atoms of protein/nucleic, 2 ns simulation with strong constraint of protein/nucleic backbone atoms and 4 ns simulation with weak constraint of protein/nucleic backbone atoms. Subsequently, ∼100 ns free MD simulations on NPT ensemble were performed to investigate the dynamics structure of Cas9–sgRNA–DNA complex. Three independent simulations (termed as MD1, MD2 and MD3) are performed to verify the simulation results. During the simulations, the temperature was controlled at 300K by Langevin method and pressure was controlled at 1atm by Langevin piston method. SHAKE method was used on all hydrogen-containing bonds to allow a 2 fs time step. The simulation trajectories were analyzed with VMD program (38).

RESULTS

SpyCas9 divides target DNA into distinct functional domains

Upon binding with sgRNA, Cas9 protein forms a DNA surveillance complex to achieve site-specific recognition and cleavage of target DNA. In this complex, the tracrRNA plays a crucial role in structural rearrangement of Cas9 into an active conformation, and the 20-nt spacer sequence of crRNA confers DNA targeting specificity. Previous studies have shown that the spacer region of crRNA contains a PAM-proximal seed sequence with 7–10 bp in length located at its 3′-end (9,16,17,39), which generally defines the target specificity of sgRNA–Cas9 complex. In addition, it was found that the 5′-end truncated sgRNAs, with shorter regions of target complementarity, can decrease undesired mutagenesis at some off-target sites without sacrificing on-target genome editing efficiencies. These results indicated that the spacer region of crRNA or protospacer region of target DNA could be further divided into subdomains with distinct roles in target binding and cleavage. To probe the relative importance of these domains, we first analyzed a series of protospacer-containing plasmid DNAs harboring consecutive transversion double-nucleotide mismatches (Target-1, Figure 1B) for their ability to be cleaved by SpyCas9, its mutants SpyCas9 (D10A) and SpyCas9 (H840A). SpyCas9 (D10A) mutant merely digests target strand, whereas SpyCas9 (H840A) only digests non-target strand (9). The cleavage efficiency of Cas9 and its mutants show different mismatch sensitivity at different region (Figure 1C). The 3–6 region proximal to the PAM is the most sensitive positions, confirming its role as seed region. Furthermore, it could be found that the sensitivity of trimming activity to mismatches in this region for D10A nickase is more obvious than that of H840A mutant. All Cas9 variants are not very sensitive to mismatches at first 2 nt proximal to the PAM under our conditions, which is quite consistent with previous *in vitro* results (40). Target binding assay using dCas9 ('dead' Cas9, a catalytically inactivated form of Cas9 resulting from the mutations D10A and H840A) also indicated that mismatches in this region do not dramatically reduce target binding affinity (Supplementary Figure S1A). The results are also proved with constructs with different target sequence (Target-2, Supplementary Figure S2). These results indicated that the first 2-nt sequence proximal to PAM (position 1–2) may not serve as the part of the seed region.It was also found that mismatches at the last two nucleotides at the PAM distal region (position 19–20) show higher cleavage efficiency than that of WT target (Figure 1C and Supplementary Figure S2B). In addition, mismatches at 13–14 position may also lead to deficient non-target strand cleavage activity by Cas9 (Figure 1C). However, this result is not observed for Target-2 DNA (Supplementary Figure S2B), suggesting that mismatch sensitivity in this region is sequence dependent. Collectively, these results indicate that the Cas9 protein, the same as argonaute protein in RNA silencing system (41), divides the target DNA sequence into distinct functional domains.

A PAM/protospacer linker exists in SpyCas9 system

It had been demonstrated that a short (1–4 nt), non-conserved linker separate the PAM and the adjacent protospacer in several other Cas9 systems such as StCas9 and NmeCas9 (42,43). Lengthening or shortening of this linker sequence will affect cleavage efficiency and cleavage site selection. Considering the result that SpyCas9 is insensitive to the mismatches in the first 2-bp PAM-proximal region, we speculate that this 2-bp sequence in SpyCas9 system may also serve as a linker between the seed region of protospacer and PAM sequence. To test this hypothesis, we designed two other DNA substrates with different linker length (Figure 2). Our results show that 3-bp linker length displays higher cleavage efficiency than that of the WT sequence, while 1-bp linker length displays lower cleavage efficiency than that of the WT sequence (Figure 2B). Sequencing of the cleavage product indicated the cleavage site of target strand by HNH domain is fixed at a position determined by guide RNA. For the non-target strand, two cleavage sites by RuvC domain exist, one between 3 and 4 nt from the PAM, the other between 4 and 5 nt from the PAM (Figure 2A). These cleavage sites are not affected when increasing the linker length to 3-bp (Figure 2A). In contrast, when shorter linker is used, although the cleavage efficiency is much lower (Figure 2B), the cleavage site is fixed between position 3 and 4 nt from the PAM (Figure 2A), suggesting an increase of cleavage specificity with shorter linker. These results indicated that SpyCas9 is tolerable to the mismatches or extra bases on PAM proximal 1–2 bp region.

The arginine-rich bridge helix and the immobilized seed region of sgRNA plays key role in the initiation of R-loop

To examine the detailed requirement of the seed region in protospacers, consecutive dinucleotides transversion mismatch sequences with 1 nt step size of target DNA were designed to study the tolerance of SpyCas9 for mismatched target sites (Supplementary Figure S1B). It could be found that nucleotides 3–7 of PAM-proximal region in the target strand are positions most sensitive to mismatches (Supplementary Figure S1C). High resolution structures of sgRNA–Cas9 complex with and without target DNA indicated that the nucleotides of sgRNA in this region were immobilized by a highly conserved arginine-rich bridge helix (24,25,27). Direct docking of standard B-form duplex DNA onto the Cas9–sgRNA complex indicates the clear interference of DNA by the bridge helix (Figure 3A). We thus speculated that the initiation of R-loop structure requires the disruption of DNA duplex by this bridge helix and its associated sgRNA sequences. To test this hypothesis, we conducted several single-site mutation studies on Cas9 in this region. As seen in Figure 3B, the E60A, R66A, R70A and R74A mutations reduced the cleavage activities of SpyCas9. And R66A/R70A/R74A triple mutation totally eliminated the cleavage activities of Cas9. These results strongly suggest that the arginine-rich bridge helix and its combined seed-region of sgRNA are important for the R-loop initiation.

R-loop has a critical size for efficient double-stranded DNA cleavage

Previous biochemical and single-molecule studies revealed that the sgRNA–Cas9 complex could efficiently bind to its target with as short as 8 nt base pairing between guide RNA and PAM proximal region of target strand (30,31). However, at least 12 bp RNA:DNA hybrid duplex is required for efficient nicking of target DNA (26), and 14–17 bp is necessary for efficient double duplex cleavage (44). The molecular details about this R-loop propagation process are still not very clear. We next examined the cleavage activities of Cas9 and its mutants (Cas9 (WT), Cas9 (D10A), Cas9 (H840A)) on DNA target sequences with 2–10 bp mismatches at the PAM distal end (Supplementary Table S1). It was found that, as short as 10 nt base-pairing between RNA and target DNA could induce nicking of target DNA, with cleavage on either target or non-target strand (Supplementary

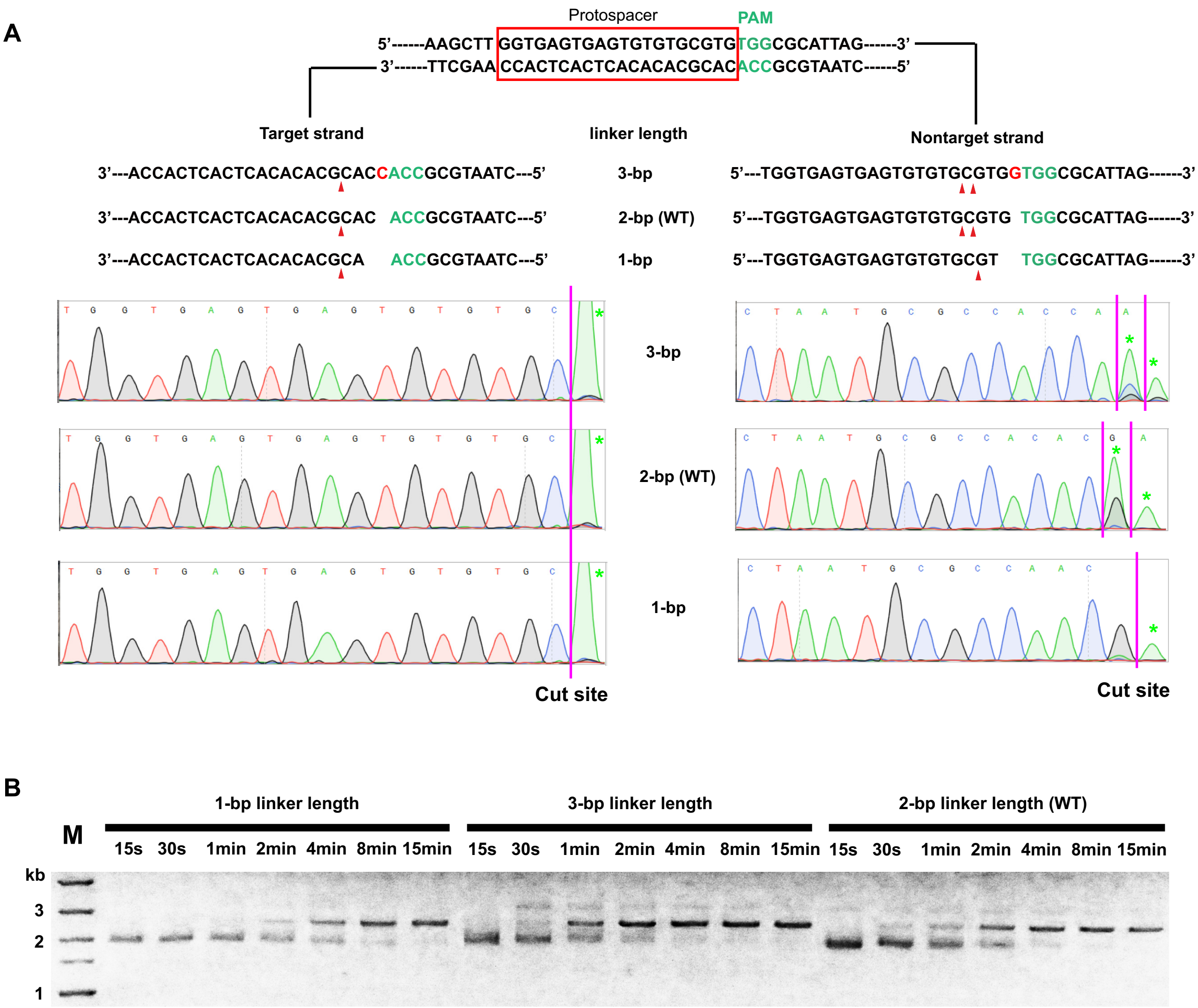

Figure 2. (**A**) Mapping of the cleavage sites in target and non-target strand of the target DNA with different linker length; (**B**) Digestion of the target DNA with different linker length by Cas9–sgRNA complex.

Figure S3A). These results suggest that the nicking reaction happens once Cas9–sgRNA stably binds to the target DNA after seed pairing. It can also be speculated that the cleavage of target and non-target strand DNA by Cas9–sgRNA complex is highly coordinated, even though the cleavage efficiency for target strand by HNH domain is a little higher than that for non-target strand by RuvC domain. Cas9 protein adopts a 'partial active conformation' when the base pairing length is less than 12 nt, rather than an 'inactive conformation' mentioned before (44). Double-strand cleavage product appeared when the base-pairing length reaches 12 bp, indicating that a minimal size for the R-loop structure exists for efficient target cleavage by Cas9 complex. Similar results were also observed when using another DNA target (Target-2, Supplementary Figure S3B), except that 15 bp guide-DNA pairing is needed for efficient double-strand breaks in this target.

Structural dynamics of Cas9 complex during R-loop propagation

The existence of 'partial active conformation' in Cas9 during R-loop propagation indicates that the Cas9 complex is highly dynamic. To further investigate the dynamics of R-loop complex and the precise mechanism of Cas9 activation, we used single-molecule fluorescence Energy Transfer (smFRET) technique to directly observe the the formation of RNA:DNA hybrid in real time upon dCas9–sgRNA complexes binding. Biotinylated Target-2 DNA was labeled with Cy3 (donor) at −5 position on the target strand and Cy5 (acceptor) fluorophores at +8 position on the non-target strand, respectively (Figure 4A). This labeling strategy does not affect the binding and cleavage activity of Cas9–sgRNA complex (data not shown). Stable Cas9 R-loop complex formation results in a decrease of FRET between Cy3 and Cy5. (Supplementary Figure S4A). The target DNA was first immobilized on the biotin functionalized glass surface of microfluidic chamber through biotin-

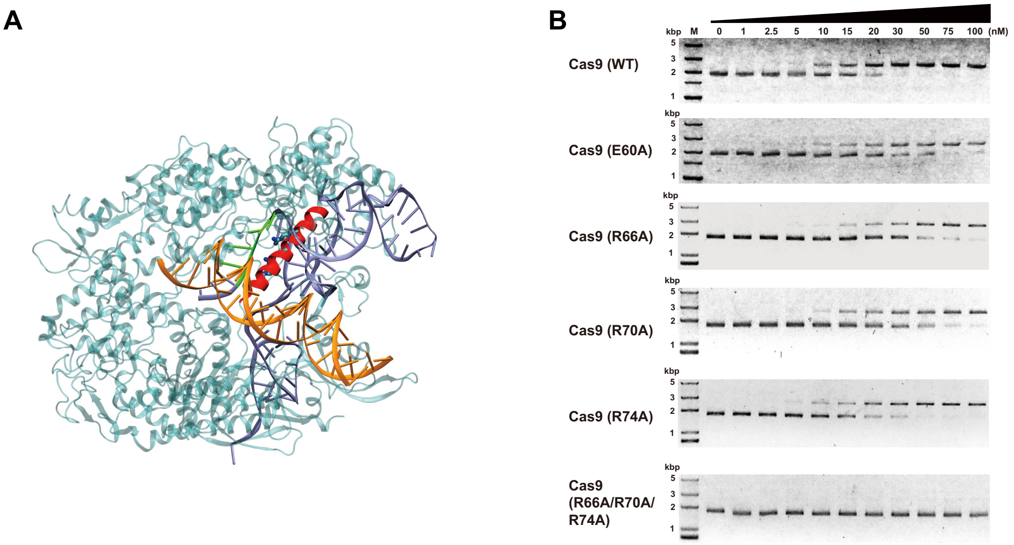

Figure 3. Effect of arginine-rich bridge helix in Cas9 on target digestion. (**A**) Docking of dsDNA on sgRNA–Cas9 complex (PDB ID: 4UN3); (**B**) Target digestion of target DNA by Cas9 or its mutants at different concentrations.

streptavidin interaction (Figure 4A). The sgRNA–Cas9 complex (50 nM) was then flowed through the chamber. Individual interactions of target DNA with Cas9–sgRNA complex were recorded immediately after flowing sgRNA–Cas9 complex (50 nM) was then flowed through the chamber using a total internal reflection microscope. Progression of DNA binding and R-loop formation was monitored by measuring the FRET signal changes (Figure 4B and D) between Cy3 and Cy5. DNA alone shows a single FRET state at around 0.5 (folded state). After the addition of sgRNA–Cas9 complex, the FRET signal disappeared quickly, which is much faster than that of intrinsic photo-bleaching (Figure 4C and E; Supplementary Figure S5A). This indicates that FRET between Cy3 and Cy5 reached around 0 (unfolded state) after R-loop formation, which is quite consistent with the ensemble measurements (Supplementary Figure S4A). Interestingly, we occasionally observed a new lower FRET state (0.35) in some single-molecule traces (Supplementary Figure S4B). This new FRET state may reveal the intermediate state during R-loop complex formation detected before (30,44).

We next prepared a series of guide RNAs containing PAM-distal mismatches relative to the Target-2 DNA to characterize this new intermediate state during R-loop complex formation. As seen in Figure 5, the intermediate FRET state was clearly identified when the number of mismatches between guide RNA and target DNA increased. And frequent reversible transitions between the folded state (FRET = 0.5) and the intermediate state (FRET = 0.35) in single-molecule time trajectories were also observed (Figure 5A–D).These time trajectories were further analyzed based on hidden Markov model to obtain the transition kinetics between these two states. Survival probability distributions of dwell times in the two states were best described by a single-exponential decay (Figure 5E and Supplementary Figure S5B). The dwell time of folded state decreased constantly (Figure 5F), indicating that longer DNA–RNA base pairing induces easier opening of DNA duplex structure. Interestingly, the dwell time of intermediate state increases first and reaches highest value at 18 nt base-pairing, rather than 20-nt base-pairing. This is consistent with our biochemical results which clearly revealed that 2-nt mismatch between guide RNA and target DNA at the PAM-distal region results higher cleavage activity (Figure 1C and Supplementary Figure S2). Previous *in vivo* genome editing experiments also showed that truncated sgRNAs, with shorter length of target complementary nucleotides, could increase the target selectivity of Cas9 without sacrificing on-target genome editing efficiencies (20). We thus hypothesize that the intermediate FRET state of the target DNA we observed is closely related to the cleavage activities of Cas9–sgRNA complex. To test this hypothesis, we next quantitatively access the effect of mismatches at the PAM distal end of the guide region on the Cas9 cleavage activity. As seen in Figure 6A, the cleavage activity, both on target strand and non-target strand, increases first upon the increase of guide-DNA pairing and reaches a maximum value at 18-bp pairing, and then decreases subsequently. The results are consistent for both Target-1 and Target-2. In addition, it could also be observed that non-target strand cleavage is more sensitive to the guide-target pairing length than that of target strand. We speculate that the dynamics of HNH domain in Cas9 is controlled by the displaced non-target DNA, and a critical size of R-loop exists for efficient target DNA digestion.

Linker region between HNH and RuvC domain plays important role in sampling SpyCas9 into its active conformation

Structural and biochemical studies indicated that the HNH nuclease domain of SpyCas9 is highly dynamic and samples an inactive conformation if the PAM distal region of target DNA is not fully complimentary to RNA guide strand. It rotates by 180°and shifts about 20 Å toward the RNA:DNA hybrid duplex after the non-target stand was totally displace by guide RNA (27). Mutational analysis re-

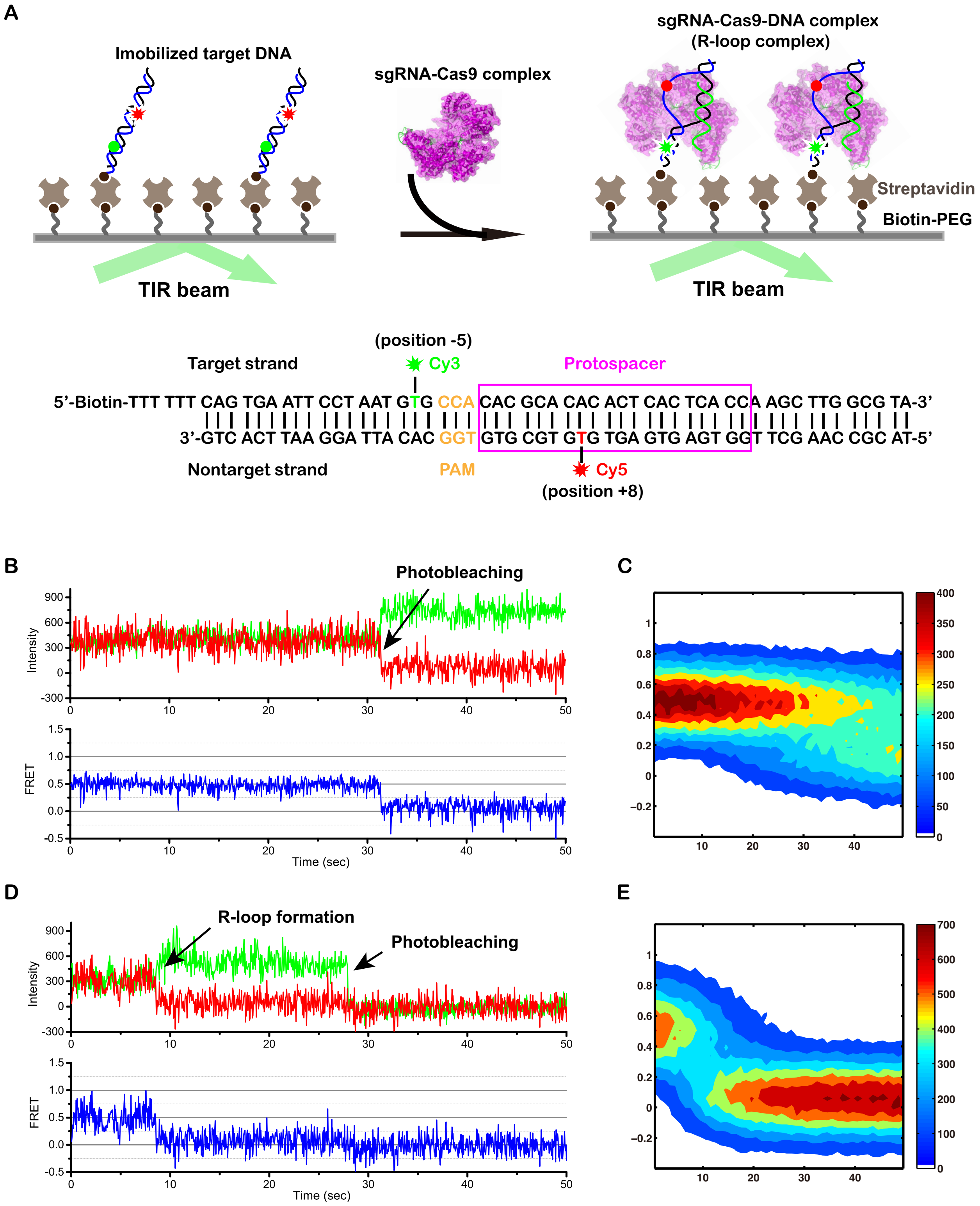

Figure 4. smFRET study on the R-loop formation. (**A**) Schematic representative of smFRET experimental setup; (**B** and **D**) Typical traces for target DNA with and without Cas9–sgRNA complex. (**C** and **E**) Contour plots of the time evolution of population FRET. Each plot was generated by superimposing the individual smFRET traces.

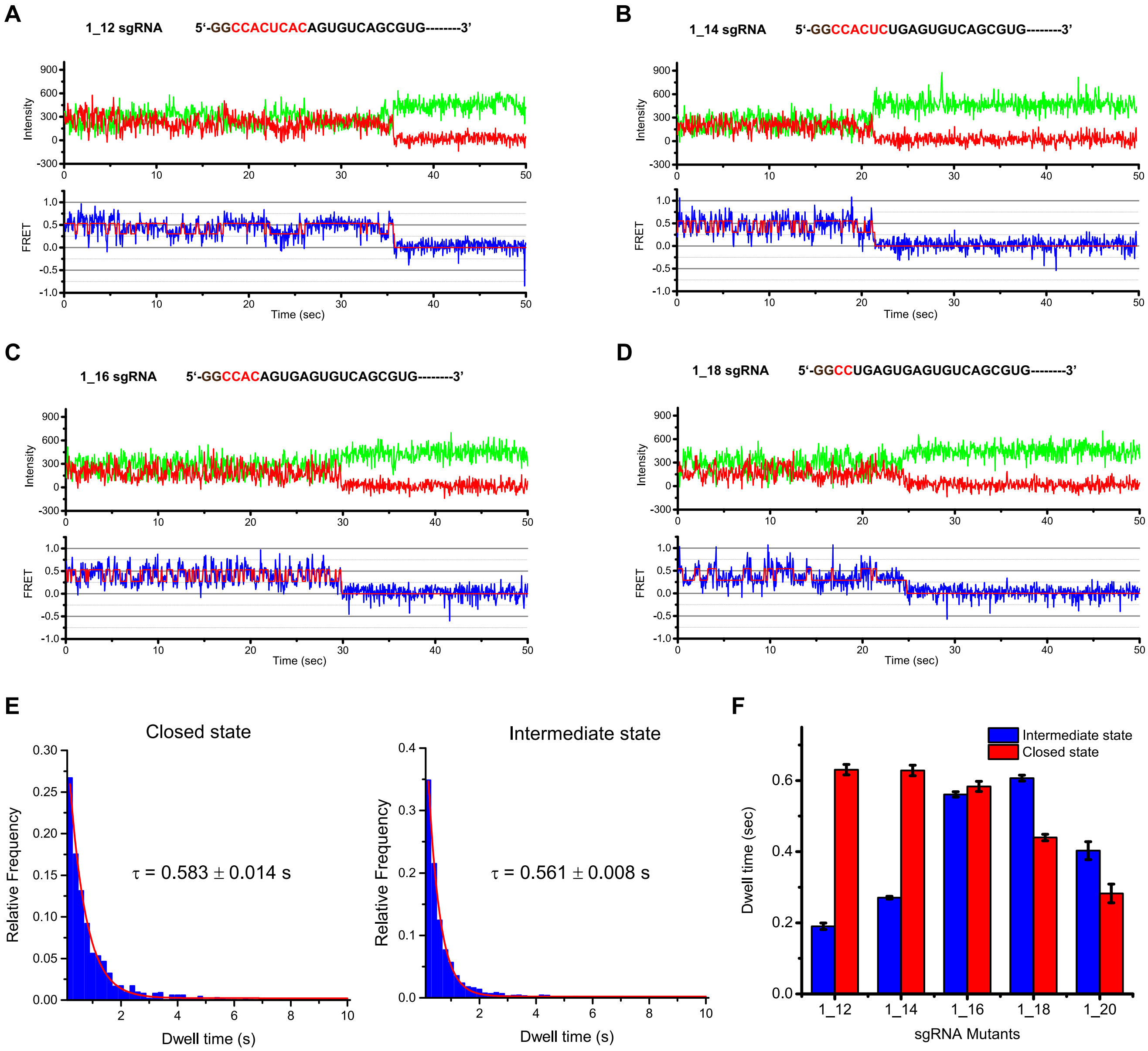

Figure 5. (**A**–**D**) Typical single-molecule fluoresce intensity and FRET time traces with different guide-target base-pairing length. (A) 12 bp, (**B**) 14 bp, (C) 16 bp, (D) 18 bp. (**E**) Dwell time analysis of folded and intermediate states of target DNA with 16 bp R-loop length. (**F**) Comparison of dwell time of folded and intermediate states between different R-loop length.

vealed that the HNH domain depleted Cas9 losses its target cleavage activity even retaining nearly WT DNA binding activity (31). Therefore, the dynamical property of HNH domain is an important feature of Cas9 complex in the process of target cleavage. HNH domain connects to RuvC domain through two linkers, linker-1 (residues 765–780) and linker-2 (residues 906–918). We next analyzed the flexibility of these two linkers using (MD) simulation. The result shows that several hydrogen bonds can be formed between the sidechain atoms of R780 and sgRNA backbone (A12) as well as the HNH domain (D809, L806) (Figure 6B). R780 thus may act as a bridge to restraint the body of HNH domain after R-loop complex formed. K913 and K918 were also found to act in similar way, although with less stable hydrogen bonding (Supplementary Figure S6). Target cleavage experiments further confirmed that longer guide RNA–target DNA base-pairing is needed for Cas9 mutants R780A and K918A to reach their active conformation than that of WT Cas9 (Figure 6C). smFRET analysis on dCas9(R780A) further demonstrated that the dwell time of the intermediate state decreases significantly comparing with that of Cas9 (WT) (Figure 6D).Homologous alignment of these two linkers between spyCas9 and other bacteria species showed that several highly conserved basic residues exist on both linkers, especially residue R780 on linker-1, residues K913 and K919 on linker-2 (Supplementary Figure S7). these results demonstrate that the interaction between SpyCas9 linkers and target DNA or guide

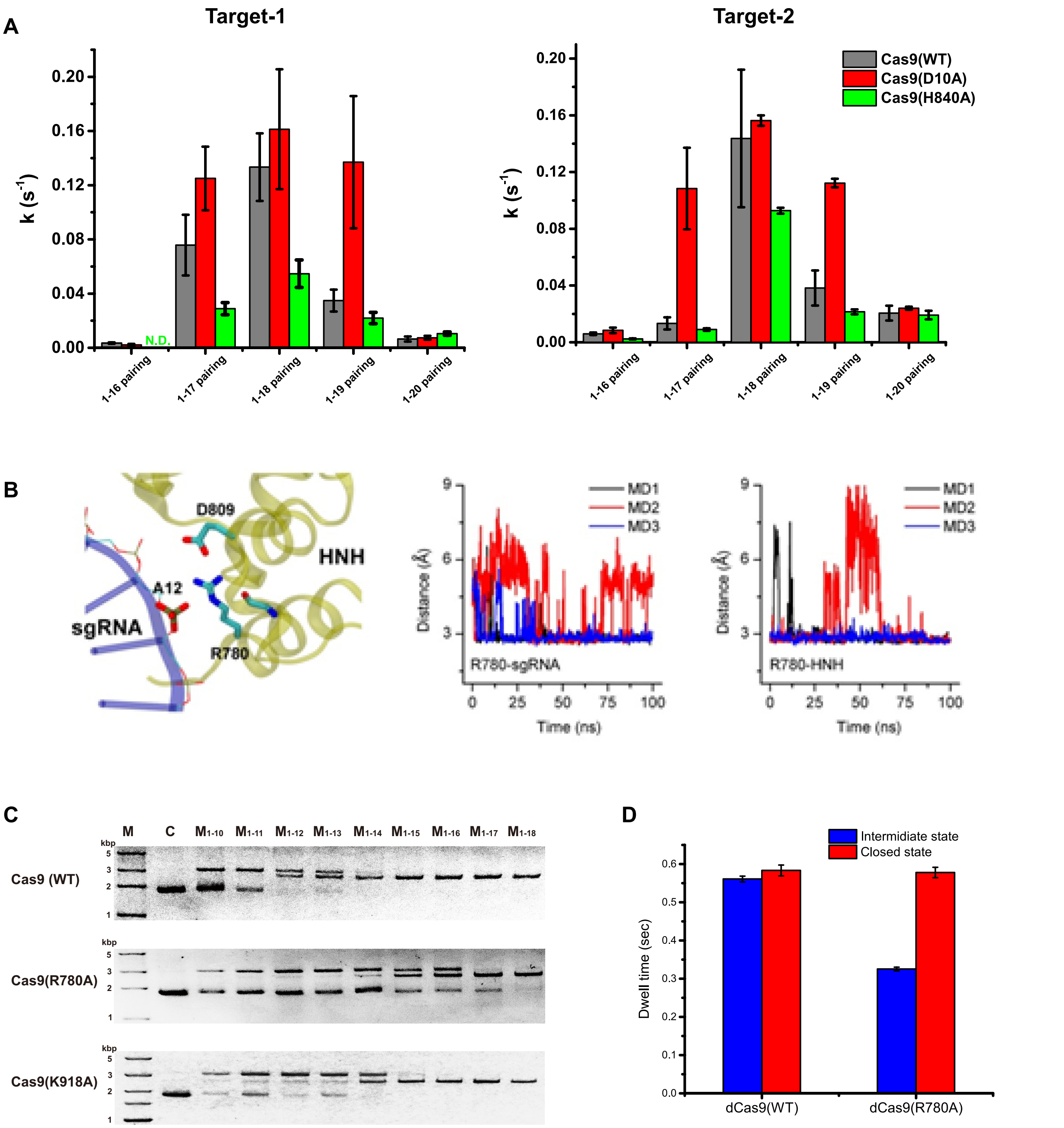

Figure 6. (**A**) The target cleavage activities in SpyCas9 system with different guide-target base-pairing length. (**B**) MD simulation of the dynamics of HNH domain in Cas9 complex. (**C** and **D**) Effect of linkers between HNH and RuvC domains on the the dynamics of Cas9 complex. (C) Cleavage of plasmids containing mutant protospacer sequence with different length of R-loop using WT Cas9 or mutants at linker region. (D) Dwell time analysis of folded and intermediate state of target DNA at the R-loop length of 16 bp using Cas9(WT) or Cas9 (R780A) mutant.

RNA affects the kinetics of HNH domain transiting into its active conformation.

DISCUSSION

In this study, we characterized the biochemical properties of SpyCas9 during R-loop formation. We find that, similar to Argonate protein in RNA silencing processes, Cas9 protein, in complexed with sgRNA, divides target DNA into PAM, linker (1–2), seed (3–7), middle and tail functional domains. PAM sequence facilitates rapid identification of potential target DNA by sgRNA–Cas9 complex. Previous biochemical and structural results indicated that the PAM binding alone could initiate the target DNA strand separation (45). Our results showed that the mismatches between

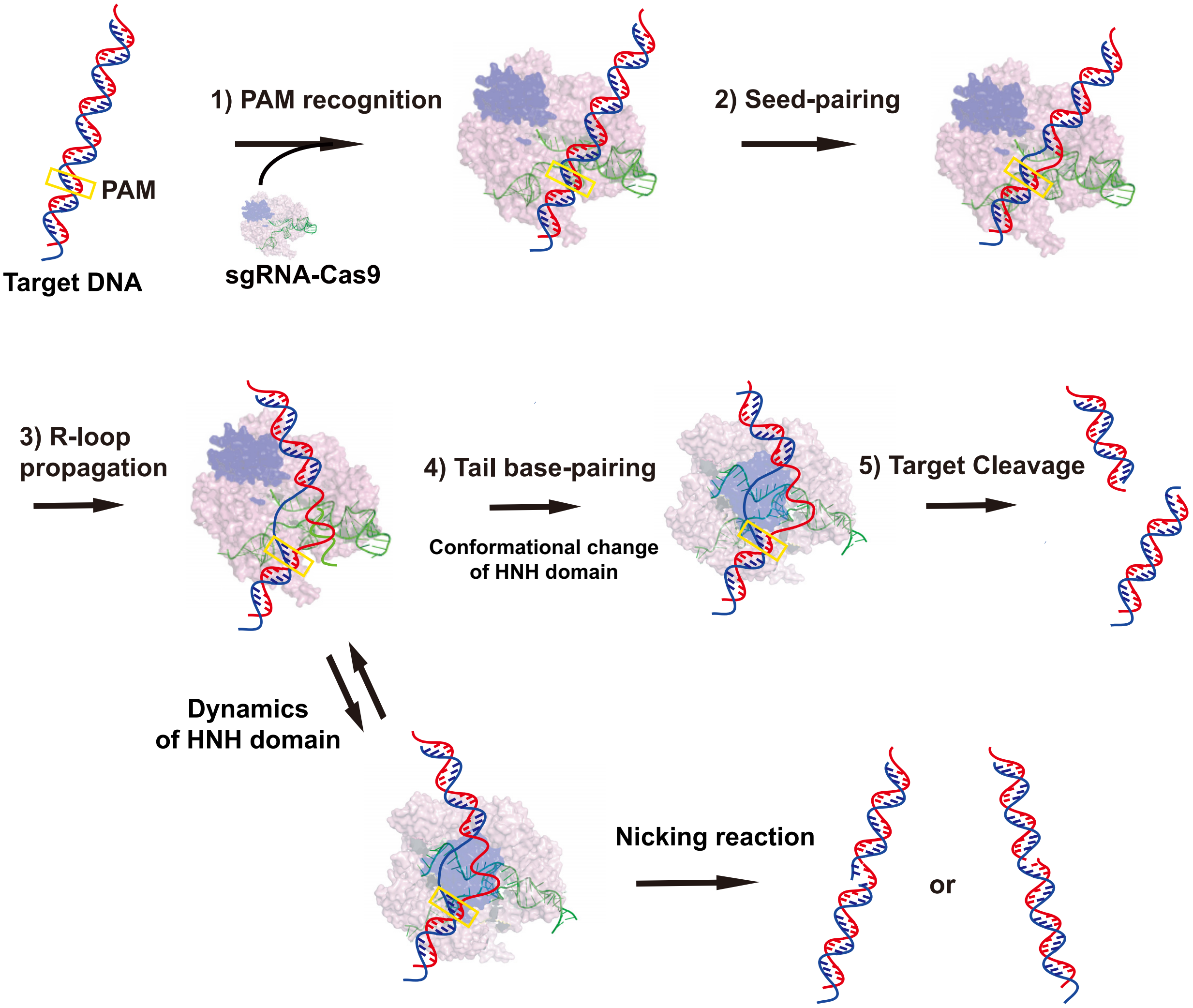

Figure 7. An updated model for the formation of R-loop complex in SpyCas9 system. In the first step, SgRNA–Cas9 complex binds to the DNA target through transient PAM recognition. R-loop structure is then initiated by both PAM recognition motif and Arginine-rich bridge helix in Cas9 protein. After seed pairing between guide RNA and target strand DNA, the HNH domain adopts a 'partial active conformation', inducing nicking reaction for either target strand or non-target strand. And once the R-loop structure reached a critical size, the HNH domain was able to adopt an active conformation and induced the double-strand cleavage.

guide RNA and the first two PAM proximal region of target DNA are tolerated by SpyCas9. Further studies indicated that the first 2-bp proximal sequence serves as a linker between PAM and seed region. Shortening the linker to 1-bp decreases the cleavage activity but increases the cutting site specificity for the RuvC domain. Lengthening the linker to 3-bp increases the cleavage activity. These results are similar to that of StCas9, indicating that the existence of linker sequence might be a universal property of target DNA for Cas9 system (42). Our data demonstrate that SpyCas9 is tolerable to the mismatches or extra bases at very proximal region of PAM. This leads us to speculate that the design of sgRNA is not strictly yield to the PAM position.

Seed base-pairing (3–7) between the guide and target strand is required for efficient target cleavage (both target and non-target strand). As observed in the crystal structures, the corresponding sequence of guide RNA for this region is immobilized by the arginine rich bridge helix in Cas9. Cas9 mutations in this region eliminate the DNA binding activity of sgRNA–Cas9 complex. We also test the sequence dependence of target cleavage activity, our results indicated that the target recognition in seed region is sequence insensitive (data not shown), which is in good agreement with previous studies. These facts suggested that Arginine-rich helix and the immobilized guide RNA play dominant roles in R-loop initiation.

Middle region in target DNA serves as enhancer for stable binding of Cas9–sgRNA to the target. The length of this region is sequence dependent. Previous single-molecule experiments revealed that sgRNA–Cas9 establishes stable complex with target DNA once DNA–RNA hybrid duplex of 8 bp or more is formed (30). The plasmid cleavage assay using Cas9 and its mutants revealed that the nicking of target DNA starts as short as 10-bp DNA–RNA hybrid duplex, both on target and non-target strand, which indicates that cleavage of target and non-target strand happened simultaneously. A total of 12–15 bp RNA:DNA paring with sequence dependent behavior are needed for efficient double strand cleavage. Further quantitative analysis indicated that the target cleavage reaction reaches highest efficiency at the R-loop size of 18-bp.

To further study the dynamics of R-loop complex, we used smFRET assay to probe the formation of RNA:DNA hybrid duplex during R-loop propagation. Our experiments with variable guide sequences for same target demonstrated that an intermediate state exists before the fully stable R-loop complex. Kinetics analysis of this new intermediate state indicated that the lifetime of this state increases when the base-pairing length of guide-DNA duplex increases and reaches the maximum at the size of 18 bp, which is quite consistent with our *in vitro* target cleavage results. Mutation R780A at the linker region between HNH domain and RuvC domain decreases the lifetime of this intermediate state and further decreases the target cleavage activity of Cas9. These results further confirmed the importance of the dynamics of HNH domain in target cleavage in SpyCas9 system. Recently, another smFRET study on the dynamic motions of the Cas9 HNH domain during target binding was reported (46). In that report, it was found that Cas9 adopted an intermediate state between target DNA binding and cleavage. And sequence mismatches between the DNA target and guide RNA at PAM distal region prevent transitions from the checkpoint intermediate to the active conformation. These results are quite consistent with our current data. We thus speculate that the structural intermediate of R-loop we identified may correspond to the Cas9 checkpoint intermediate before accessing its active conformation.

Based on our results and previous other work (24,26,27,29–32,44–46), we proposed an improved model for the formation of R-loop complex in SpyCas9 system (Figure 7). SgRNA–Cas9 complex first rapidly identifies the target through PAM recognition. R-loop structure is then initiated by both PAM and Arginine-rich bridge helix. After stable sgRNA–Cas9–DNA complex formed (8–10 bp guide RNA:DNA hybrid duplex), the HNH domain adopts a ‘partial active conformation’, inducing nicking reaction for either target strand or non-target strand. As the size of RNA:DNA hybrid increases, the target DNA becomes highly dynamic, and adopts an intermediate state during this R-loop propagation. Once the R-loop reaches a critical size, the HNH domain will have enough space to reach its active conformation and induce the double-strand cleavage. The R-loop have an optimum size of 18 bp at which the HNH domain stays the longest time in its active conformation. Shorter RNA:DNA hybrid duplex leads to smaller R-loop bubble, blocking HNH domain to its active state; while longer RNA:DNA hybrid duplex leads to bigger R-loop bubble, constraining the linker region of the HNH and then the dynamics of HNH targeting to its active conformation. This model provides new insights into the process of R-loop formation and reveals the source of off-targeting in CRISPR/Cas9 system, and may shed new light on the sgRNA design and Cas9 engineering for optimizing CRISPR-Cas9 genome editing specificity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Lei Zhang for technical assistance with the single-molecule FRET experiments. We thank Dr Yanli Wang for critical reading and helpful comments on the manuscript.

FUNDING

National Basic Research Program of China [2014CB910202 to J.L.]; National Natural Science Foundation of China [91219103 to J.L., 31300772 to G.S., 31222022 to J.L.]; National Supercomputing Center Tianjin Center. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Haft,D.H., Selengut,J., Mongodin,E.F. and Nelson,K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.*, **1**, e60.
2. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
3. Brouns,S.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J., Snijders,A.P., Dickman,M.J., Makarova,K.S., Koonin,E.V. and van der Oost,J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.
4. Marraffini,L.A. and Sontheimer,E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, **322**, 1843–1845.
5. Sorek,R., Kunin,V. and Hugenholtz,P. (2008) CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.
6. Karginov,F.V. and Hannon,G.J. (2010) The CRISPR system: small RNA-guided defense in bacteria and archaea. *Mol. Cell*, **37**, 7–19.
7. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
8. Makarova,K.S., Wolf,Y.I., Alkhnbashi,O.S., Costa,F., Shah,S.A., Saunders,S.J., Barrangou,R., Brouns,S.J., Charpentier,E., Haft,D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.
9. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
10. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
11. Gilbert,L.A., Larson,M.H., Morsut,L., Liu,Z., Brar,G.A., Torres,S.E., Stern-Ginossar,N., Brandman,O., Whitehead,E.H., Doudna,J.A. *et al.* (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, **154**, 442–451.
12. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
13. Doudna,J.A. and Charpentier,E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**, 1258096.
14. Deltcheva,E., Chylinski,K., Sharma,C.M., Gonzales,K., Chao,Y., Pirzada,Z.A., Eckert,M.R., Vogel,J. and Charpentier,E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, **471**, 602–607.

15. Mojica,F.J., Diez-Villasenor,C., Garcia-Martinez,J. and Almendros,C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733–740.
16. Fu,Y., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,J.K. and Sander,J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.
17. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
18. Tycko,J., Myer,V.E. and Hsu,P.D. (2016) Methods for optimizing CRISPR-Cas9 genome editing specificity. *Mol. Cell*, **63**, 355–370.
19. Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,I., Sullender,M., Ebert,B.L., Xavier,R.J. and Root,D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.
20. Fu,Y., Sander,J.D., Reyon,D., Cascio,V.M. and Joung,J.K. (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.*, **32**, 279–284.
21. Kleinstiver,B.P., Pattanayak,V., Prew,M.S., Tsai,S.Q., Nguyen,N.T., Zheng,Z. and Joung,J.K. (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**, 490–495.
22. Slaymaker,I.M., Gao,L., Zetsche,B., Scott,D.A., Yan,W.X. and Zhang,F. (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science*, **351**, 84–88.
23. Jinek,M., Jiang,F., Taylor,D.W., Sternberg,S.H., Kaya,E., Ma,E., Anders,C., Hauer,M., Zhou,K., Lin,S. *et al.* (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, **343**, 1247997.
24. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**, 935–949.
25. Jiang,F. and Doudna,J.A. (2015) The structural biology of CRISPR-Cas systems. *Curr. Opin. Struct. Biol.*, **30**, 100–111.
26. Jiang,F., Zhou,K., Ma,L., Gressel,S. and Doudna,J.A. (2015) A Cas9-guide RNA complex preorganized for target DNA recognition. *Science*, **348**, 1477–1481.
27. Jiang,F., Taylor,D.W., Chen,J.S., Kornfeld,J.E., Zhou,K., Thompson,A.J., Nogales,E. and Doudna,J.A. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, **351**, 867–871.
28. Jiang,F. and Doudna,J.A. (2017) CRISPR-Cas9 structures and mechanisms. *Annu. Rev. Biophys.*, **46**, 505–529.
29. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.
30. Singh,D., Sternberg,S.H., Fei,J., Doudna,J.A. and Ha,T. (2016) Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat. Commun.*, **7**, 12778.
31. Sternberg,S.H., LaFrance,B., Kaplan,M. and Doudna,J.A. (2015) Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature*, **527**, 110–113.
32. Josephs,E.A., Kocak,D.D., Fitzgibbon,C.J., McMenemy,J., Gersbach,C.A. and Marszalek,P.E. (2016) Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic Acids Res.*, **44**, 8924–8941.
33. Heng,J., Zhao,Y., Liu,M., Liu,Y., Fan,J., Wang,X., Zhao,Y. and Zhang,X.C. (2015) Substrate-bound structure of the E. coli multidrug resistance transporter MdfA. *Cell Res.*, **25**, 1060–1073.
34. Roy,R., Hohng,S. and Ha,T. (2008) A practical guide to single-molecule FRET. *Nat. Methods*, **5**, 507–516.
35. Bronson,J.E., Fei,J., Hofman,J.M., Gonzalez,R.L. Jr. and Wiggins,C.H. (2009) Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.*, **97**, 3196–3205.
36. Perez,A., Marchan,I., Svozil,D., Sponer,J., Cheatham,T.E. 3rd, Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
37. Zgarbova,M., Luque,F.J., Sponer,J., Cheatham,T.E. 3rd, Otyepka,M. and Jurecka,P. (2013) Toward improved description of DNA backbone: revisiting epsilon and zeta torsion force field parameters. *J. Chem. Theory Comput.*, **9**, 2339–2354.
38. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
39. Kuscu,C., Arslan,S., Singh,R., Thorpe,J. and Adli,M. (2014) Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.*, **32**, 677–683.
40. Fu,B.X., Hansen,L.L., Artiles,K.L., Nonet,M.L. and Fire,A.Z. (2014) Landscape of target:guide homology effects on Cas9-mediated cleavage. *Nucleic Acids Res.*, **42**, 13778–13787.
41. Wee,L.M., Flores-Jasso,C.F., Salomon,W.E. and Zamore,P.D. (2012) Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*, **151**, 1055–1067.
42. Chen,H., Choi,J. and Bailey,S. (2014) Cut site selection by the two nuclease domains of the Cas9 RNA-guided endonuclease. *J. Biol. Chem.*, **289**, 13284–13294.
43. Nishimasu,H., Cong,L., Yan,W.X., Ran,F.A., Zetsche,B., Li,Y., Kurabayashi,A., Ishitani,R., Zhang,F. and Nureki,O. (2015) Crystal structure of Staphylococcus aureus Cas9. *Cell*, **162**, 1113–1126.
44. Lim,Y., Bak,S.Y., Sung,K., Jeong,E., Lee,S.H., Kim,J.S., Bae,S. and Kim,S.K. (2016) Structural roles of guide RNAs in the nuclease activity of Cas9 endonuclease. *Nat. Commun.*, **7**, 13350.
45. Anders,C., Niewoehner,O., Duerst,A. and Jinek,M. (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, **513**, 569–573.
46. Dagdas,Y.S., Chen,J.S., Sternberg,S.H., Doudna,J.A. and Yildiz,A. (2017) A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Sci. Adv.*, **3**, eaao0027.