Dynamics changes of CRISPR-Cas9 systems induced by high fidelity mutations

Zheng, Liangzhen; Shi, Jiahai; Mu, Yuguang

2018

Zheng, L., Shi, J., & Mu, Y. (2018). Dynamics changes of CRISPR-Cas9 systems induced by high fidelity mutations. Physical Chemistry Chemical Physics, 20(43), 27439-27448. doi:10.1039/C8CP04226H

https://hdl.handle.net/10356/139057

https://doi.org/10.1039/C8CP04226H

Journal Name

ARTICLE

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Dynamics changes of CRISPR-Cas9 system induced by high fidelity mutations

Liangzhen Zheng[a], Jiahai Shi[b,c], Yuguang Mu[a,†]

CRISPR-Cas9 as a powerful genome editing tool has widely been applied in biological fields. Ever since the discovery of CRISPR-Cas9 as an adaptive immune system, it has been gradually modified to perform precise genome editing in eukaryotic cells by creating double-strand breaks. Though being robust and efficient, current CRISPR-Cas9 system faces a major flaw: the off-target effect, which has not been well understood. Several Cas9 mutants show significant improvement, with very low off-target effect, however relatively lower cleavage efficiency for on-target sequences as well. In this study, the dynamics of the wild-type Cas9 from *Streptococcus pyogenes* and the high fidelity Cas9 mutant has been explored using molecular dynamics simulations. It turns out that the mutations cause reduction of electrostatic interactions between Cas9 and R-loop. Consequently, the flexibility of the tDNA/sgRNA heteroduplex is reduced, which may explain the reason of less tolerance of mismatches in the heteroduplex region. The mutations also affect the protein dynamics and the correlation networks among Cas9 domains. In mutant Cas9, weakened communications between two catalytic domains, as well as the mutations induced slight opening of the conformation, account for the lower on-target cleavage efficiency, and probably lower off-target as well. These findings would facilitate more precise Cas9 engineering in future.

Introduction

Many bacteria and archaea adopt a clustered regularly inter spaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) genes based adaptive immune system to discriminate against invading phages and other foreign DNAs.[1-4] The identified three types (type I, II and III) of CRISPR-Cas systems widely exist across various microbial species. Among the three types of CRISPR-Cas systems, type II system was first modified to become the genome editing tool. Unlike type I and type III, type II system is composed of a single, smaller-sized (around 1380 amino acids, Figure 1a), *Streptococcus pyogenes* DNA endonuclease Cas9 (called Cas9 hereafter), a small CRISPR RNA (crRNA), and a trans-activating crRNA (tracrRNA).[5, 6] The crRNA and tracrRNA could bind to Cas9 ahead of the subsequent DNA binding and form partial complementary tertiary structure.[7] The first step of foreign double strand DNA (dsDNA) targeting is formation of an R-loop), the complex entity forming by the Cas9 related guide RNAs as well as the dsDNA (or the cleaved DNA products) loading in position (Figure 1b and 1c). The target strand (tDNA) of the external dsDNA binds with a 20 nucleotide (nt) segment of the crRNA through Watson-Crick base-pairing. The dsDNA thus is unwounded and the non-target DNA strand (ntDNA) is displaced. The two endonuclease domains RuvC and HNH (Figure 1a) in Cas9 then catalyze the cleavage of the two strands in dsDNA, thus produce a double-strand break (DSB). The cleavage process is guided by a protospacer-adjacent motifs (PAM), generally a 5’-NGG-3’ sequence 1 or 2 nt upstream the DSB cleavage site in the tDNA strand.

CRISPR-Cas9 generated DSBs induce the common DSB repairing mechanisms, such as non-homologous end joining, which creates small deletions and insertions, as well as homologous recombination, which could be adopted to facilitate precise genome editing given homologous repair template.[7, 8] Further engineering makes the simplest form of type-II CRISPR-Cas9 consisting of only Cas9 protein and single-strand guide RNAs (sgRNA),[9, 10] which is a fusion chimeric form of crRNA and tracrRNA complex. Hence, CRISPR-Cas9 systems are widely utilized for site-specific genome editing.

However, these systems still suffer a systematic drawback, the off-target effect.[11-13] The commonly used sgRNA/Cas9 based genome editing tool tolerates up to 6 mismatches within the sgRNA complementary region of tDNA,[9, 14] and Cas9 also cleaves dsDNA at sites in absence of the extract 5’-NGG-3’ PAM sequence in dsDNA.[10, 15] Insertions and deletions in tDNA in the 20 bp tDNA/sgRNA complementary region (region 1 in Figure 1b) are also common events of the off-target effects.[16]

[a] *School of Biological Sciences, Nanyang Technological University, Singapore*
[b] *Department of Biomedical Sciences, College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong SAR*
[c] *City University of Hong Kong Shenzhen Research Institute, Shenzhen, P. R. China*
† Corresponding author, ygmu@ntu.edu.sg

Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

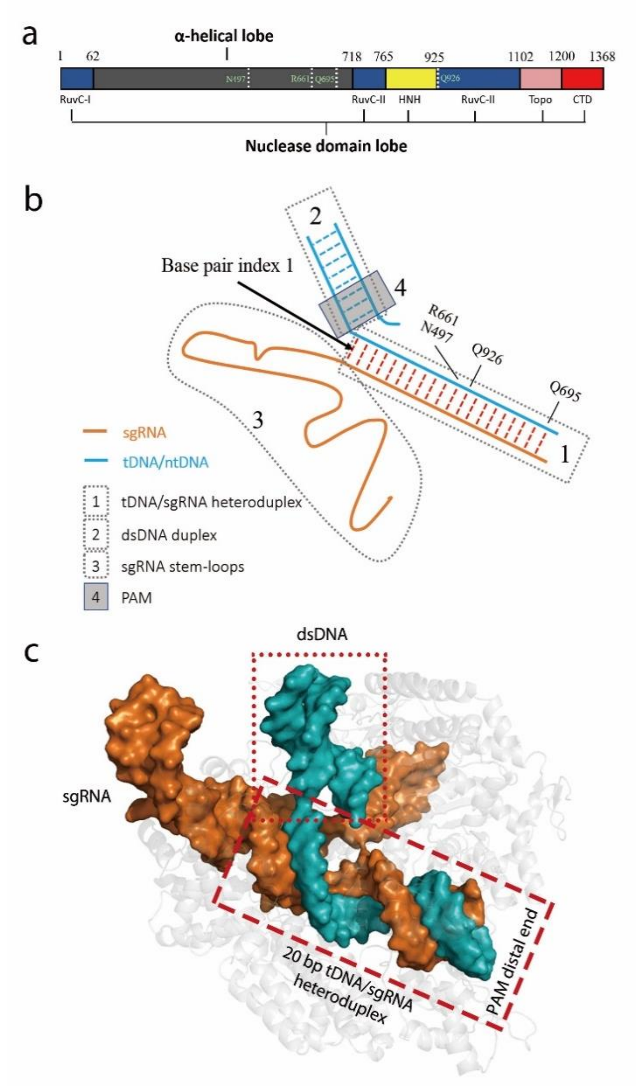

Figure 1. Domains of Cas9 and regions in R-loop. a), domains of Cas9 and the locations of the 4 HF mutations (white dash lines). RuvC, α-helical lobe, HNH domain, RuvC domain, Topoisomerase (Topo), and C-terminal domain (CTD) are in iceblue, gray, yellow, pink and red respectively; b), the relative positions of nucleic acids components in R-loop, where tDNA/ntDNA and sgRNA are in cyan color and orange color respectively. c) the R-loop surface view, with DNA in cyan and RNA in orange, and the tDNA/ntDNA duplex and tDNA/sgRNA heteroduplex are highlighted with dot-dashed box and line-dashed box, respectively.

Optimizations of the wildtype Cas9 (referred as *wt*Cas9) system to reduce the off-target effects are invaluable for high-specificity genome editing. Kleinstiver et al discovered that high-fidelity (HF) Cas9 mutants were capable to diminish the off-target effects efficiently.[15] Among them, the NRQQ mutant type (called HF-1, with 4 mutations, N497A, R661A, Q695A and Q926A), referred as *nrqq*Cas9 hereafter, reduces all or nearly all genome-wide off-target effects to undetectable levels,[15] however prosses a slightly weaken on-target cleavage efficiency meanwhile.[17] Although it has been reported that the "excess energy" model and unwound dsDNA instability may explain the high specificity of the *nrqq*Cas9 and another Cas9 variant eCas9,[15, 17-20] the atomic level explanations of the lower off-target effect as well as the lower on-target cleavage efficiency of *nrqq*Cas9 has not been well understood. Therefore, molecular dynamics (MD) simulations were employed to study the mutation induced dynamics changes of CRISPR-Cas9 system, with an emphasis on the comparison of *wt*Cas9 and *nrqq*Cas9 on their structural and dynamical properties with on-target dsDNA binding. We find that, in *nrqq*Cas9 and nucleic acids (R-loop) complex, electrostatic interaction network changes and global dynamics changes in both the protein and nucleic acids. Decreased polar contacts between *nrqq*Cas9 and R-loop (especially in the tDNA/sgRNA heteroduplex) leads to a more rigid heteroduplex. And the conformational ensembles of *wt*Cas9 and *nrqq*Cas9 are also quite different. With R-loop bound, *wt*Cas9 conformation population adopts the "closed" state, while *nrqq*Cas9 shifts towards the slightly "open" state. Meanwhile, in *nrqq*Cas9/R-loop complex, one of the three ribosome recognition motifs (REC1, REC2 and REC3 respectively), REC2, has a small increased connection and information flow with HNH domain, whereas RuvC and HNH domain has a greatly decreased information flow, while the strong correlation and information flow between HNH domain and RuvC domain are required for efficient dsDNA cleavage.[21, 22] These findings reveal the dynamics changes of CRISPR-Cas9 system induced by HF mutations and lower cleavage efficiency of the *nrqq*Cas9 with on-target dsDNA and are also valuable to engineer Cas9 with higher accuracy and efficiency in future.

Experiment

System setup

We set up several simulation systems to study the Cas9 and R-loop dynamics. We simulated wild-type and several mutants Cas9 in R-loop bound form. There are quite a few x-ray structures, such as 5F9R, 4UN3,[23] are deposited in the RCSB Protein Data Bank (PDB). The PDB 5F9R were used in several MD studies, however the 4 key residues (N497, R661, Q695, and Q926) are not all form direct contacts with the 20 base pairs tDNA/sgRNA heteroduplex region, and in this structure, there are missing Mg^{2+} ions which are crucial for the functional dynamics of CRISPR/Cas9 system. The incomplete crystal structure PDB 4UN3 captured the PAM recognition and ntDNA cleavage state of CRISPR-Cas9 system, with mutated Cas9 and R-loop (sgRNA and tDNA, partial ntDNA). The high-fidelity Cas9 mutant was previously hypothesized to have low off-target effect based on this 4UN3 model, and the 4 residues both form hydrogen bonds or salt-bridges with the tDNA chain in the heteroduplex region. Therefore, we determined to use both the PDB 4UN3 as the initial structure, and mutated back important residues and modelled missing loops.

In this initial structure PDB 4UN3, the inactivating mutation H840A was mutated back to H840. To construct the HF mutants,[15] the missing residues and loops in Cas9 were modelled with Swiss-Model online server (https://swissmodel.expasy.org/) by structure and sequence alignment with other solved *sp*Cas9 structures, while the mutants were created with Pymol mutagenesis tool.[24] The 8 Mg^{2+} ions were kept unchanged.

MD simulation protocol

The simulations were all performed with GPU accelerated Gromacs 5.1.2 package.[25] Amber99SB-ILDN force field was applied for Cas9 proteins and ions,[26] and Amber ParmBSC0 [27] for nucleic acids to describe the atomic potentials, while the commonly used TIP3P [28] explicit water model was also adopted. The systems firstly went through 1000-step energy minimization, following by a 5-ns equilibration at 300 K under NVT ensemble with all heavy atoms fixed by a 1000 kJ/(mol·nm^2) force constant except for the Swiss Model server modelled loops regions. Afterwards, the system was subjected to another 10 ns NPT equilibrium at 300 K with all the protein and nucleic acid atoms constrained with the same force constant. The last structure from the previous step was utilized as the initial structure for product NPT ensemble simulation at 300 K and 1 bar for each system. In NVT and NPT ensemble simulations, velocity rescaling algorithm[29] and berendesen pressure coupling algorithm[30] were used where applicable. For non-bonded interactions, a 1.2 nm distance cutoff was applied for both long-range van de Waals interactions and the long-range electrostatic interactions, which were realized by particle mesh Ewald (PME) summation scheme[31] Covalent bonds between heavy atoms (non-hydrogen atoms) and heavy atoms, hydrogen atoms and heavy atoms were maintained by LINC algorithm[32] and SHAKE algorithm respectively. The simulation time step was 2 fs and conformation frames were saved to trajectory files every 2 ps. The RMSDs of Cas9 αCarbon atoms and the R-loop atoms along simulation time have been plotted to assess the convergence of the systems (Support Figure 1).

Table 1. MD Simulation systems and setup of *wt*Cas9

S/N	System name	Mutations	Product Run	#Reps
S1	*wt*	None	500 ns	3
S2	*nrqq*	N497A, R661A, Q695A, Q926A	500 ns	3
S3	*N497A*	N497A	50 ns	1
S4	*R661A*	R661A	50 ns	1
S5	*Q695A*	Q695A	50 ns	1
S6	*Q926A*	Q926A	50 ns	1
S7	*R661A/Q695A*	R661A, Q695A	100 ns	1
S8	*Q695A/Q926A*	Q695A, Q926A	100 ns	1

Trajectory analysis

Number of polar contacts were defined as the coordination number between Cas9 with the nitrogen, oxygen, and phosphate atoms of the R-loop. In detail, if any one of the oxygen, nitrogen atoms in Cas9 is close to any one of the nitrogen, oxygen, and phosphate atoms of the R-loop within a 0.35 nm distance cutoff, it is counted as 1 polar contact. Number of all the polar contacts were calculated along the simulation trajectories with plumed 2.3 driver tool with a stride of 2 ps for the first repeats of simulation *wt* and *nrqq* systems (simulation systems S1 and S2 in Table 1).[33] For two polar residues, if in one residue any one of the oxygen or nitrogen atoms is within a 0.35 nm of that of the other for 80% of the time, there exists a salt-bridge between these two residues.

The principle component analysis (PCA) based essential dynamics were performed using python, and were visualized with VMD.[34] PCA analysis is a particularly useful method to identify slow motion dynamics of macromolecules to rule out the high frequency atomic level motions. By applying PCA analysis, we could observe the relative movement of the domains in Cas9 in different simulation systems. The calculations of PCA were based on αC atom coordinates of superimposed trajectories of the three repeats of the *wt* and *nrqq* systems (see Table 1), a stride of 50 ps was applied to save computation time and memory. For each domain, the eigenvectors of each residue αC atom were summed up to form a single vector representing the global motion of the domain in 3-dimentional space. The analysis could identify the important motion (domain level) modes of the systems and visualized in VMD to represent the motions and directions by arrows. Only the first (the largest variance dimension, Support Figure 2) PC was used to visualize the dynamics of Cas9.

In general, larger entropy change of a molecule during a period would indicate that the molecule is more flexible.[35-38] In this study, we used this configurational entropy change to compare the flexibility of molecules, or different domains of molecules. The configurational entropy [39] calculation was performed using Gromacs 4.6.7 g_covar and g_anaeig. The configurational entropies calculated with Gromacs are estimated based on the Quasi-harmonic oscillator approximation, assuming that the motions probabilities of the 3*N* atoms are multi-variate Gaussian distributed and produce an upper bound of the true configurational entropy S_{true},[39, 40] through configurational probability distribution density after removing the translational and rotational motions of the macromolecules by superimposing the conformations to a reference structure.

$$S_{true} < S_q^C = \frac{3N}{2}k_B + \frac{k_B}{2}ln(2\pi)^{3N}\det(\sigma)$$

where k_B is the Boltzmann constant, and σ is the mass weighted covariance matrix. In this study, we only used snapshots every 10 ps to save calculation time. For *wt* and *nrqq* systems, the final entropy value is the mean of several simulation periods (250-500ns, 250-230ns, and 250-210ns) for simulation repeat #1, and the standard error is represented by standard deviation of these entropy values.

The community network analysis procedures were adopted from other researches.[41, 42] The residue-residue contact maps were constructed based on normalized side chains atomic interaction percentage $C_{i,j}$ between two residues, therefore for each contact map, it is a symmetry *N* by *N* matrix $M_{initial} = \begin{pmatrix} I_{1,1} & \cdots & I_{1,M} \\ \vdots & \ddots & \vdots \\ I_{M,1} & \cdots & I_{M,M} \end{pmatrix}$, where *M* is total number of Cas9 residues. And

$$I_{i,j} = \frac{n_{i,j}}{\sqrt{N_i N_j}}$$

where $n_{i,j}$ is total number of close contacts (distance cutoff 0.5 nm) between atoms in side chains of two residues (residue i and j). The contact maps of each conformations used were summed up and averaged, if the averaged value $I_{average,\ ij} < I_{critic,\ ij}$, then a provisional link is formed between residues i and j. An actual link (value 1) is formed if the provisional link exists in 48% of the frames, otherwise a 0 will be assigned, resulting in a binary contact matrix, M_{tight}, based on which, the communities, containing a group of residues of Cas9, then were generated using VMD plugin gncommunities, and the output betweenness matrices (N by N). In this process, only far apart residues (separated by at least 4 residues) are considered only. The edge connectivity between a community pair was computed by summing up all residues pairs' (formed between these two communities) betweenness. The community network plots were generated with python networkx library[43] and matplotlib library. The frames were extracted from the three repeats of the *wt* system and *nrqq* trajectories respectively every 100 ps to save computation time.

Binding energy is a fundamental quantity to assess the binding strength between molecules. However, the determination of the absolute binding energy is rather troublesome, if not computational accessible. In order to compare the binding energies of the simulation systems, we used the approximated binding energy method Molecular Mechanics (MM) Poisson-Boltzmann (PB) Surface Area (SA) (MMPBSA) to calculate the binding energy between Cas9 and the R-loop with the g_mmpbsa tool for the various simulation systems.[44] The theory of MMPBSA has been thoroughly covered in many researches, and we are not going to explain in detail here.[44-46] In the process, we ignored the entropic contribution to the binding energy, and only included the MM, PB and SA parts. For calculating MM and PB, the dielectric constant was set as 4.0 for solvent for all the systems because of the highly-charged instinct of the nucleic acids. Other parameters were chosen as the default values. Only the last 10 ns trajectory of each system with a stride 10 ps was adopted for obtaining binding free energies. However, MMPBSA method may not be accurate to describe the highly charged systems such as nucleic acids, we still believe that it would be useful to compare the interaction energies for similar systems, such as Cas9 variant/R-loop complexes.

The double-strand nucleic acids conformational analysis were performed with 3DNA package together with a Python API.[47, 48] The helical parameters such as helical rise, helical bending angles, and the six parameters of base pairs were calculated along the *wt* and *nrqq* simulation trajectories (system S1 and S2 in table 1), whose last 80% simulation trajectories were collected with a 10 ps stride. The 12 base pair parameters (shift, slide, rise, tilt, roll, twist, shear, stretch, stagger, buckle, propeller and opening) of both the three repeats of *wt* and *nrqq* simulations were collected and merged for further PCA analysis using home-made python scripts.

The potential of the mean force (PMF) along the first two PCs was calculate based on the following equation:

$$PMF = -RT \ln P_i$$

While P_i is the probability distribution of the coordination to calculate, R and T are the gas constant and temperature. The representative structures of a low energy basin in PMF plot were selected through gromos clustering method in Gromacs g_cluster with a 0.2 nm cutoff.

Results

The polar interactions between Cas9 and nucleic acids changes caused by mutations

The numbers of polar contacts between Cas9 and the dsDNA (Figure 2a) and between Cas9 and the sgRNA stem-loop region (Figure 2b) are similar in both *wt* and *nrqq* simulations (see experiment part), whereas the *wt*Cas9 forms more polar contacts with the tDNA/sgRNA heteroduplex than *nrqq*Cas9 (Figure 2c). And interestingly, the base-pairs (bp index 9-20) in PAM distal end of the heteroduplex (Figure 2d, index of base-pairs starting from PAM distal end to proximal end) contributes the excess contacts with *wt*Cas9. Similarly, in

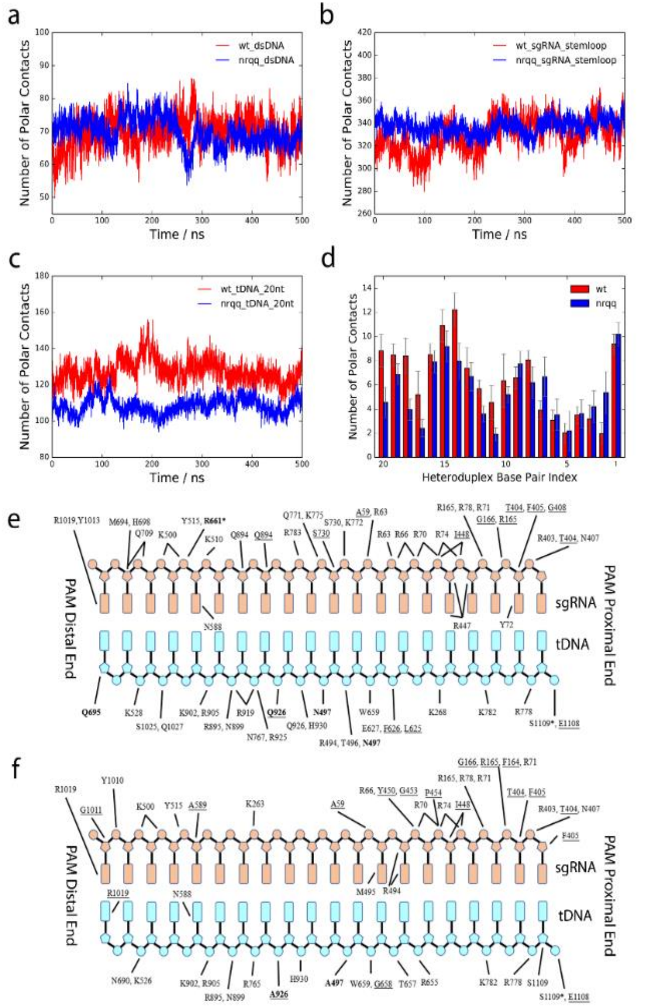

Figure 2. Electrostatic interactions between wtCas9 or nrqqCas9 and tDNA/sgRNA. a), Cas9 and the dsDNA; b), Cas9 and sgRNA stem-loop region; c), Cas9 and tDNA/sgRNA heteroduplex region; d), number of polar contacts per tDNA nucleotide. Blue and red colors are used for wt and nrqq simulation systems respectively. e) electrostatic interactions between wtCas9 and heteroduplex in last frame structure; f) electrostatic interactions between nrqqCas9 and heteroduplex in last frame structure. tDNA and sgRNA are indicated as cyan and orange color. In panel e and f, polar contacts between mainchain, or both sidechain and mainchain of Cas9 residues, are indicated using underline or asterisk.

single-site HF mutants and double-site HF mutants, the electrostatic interactions between tDNA and Cas9, as well as sgRNA and Cas9, are also decreased, but less than the decrease in *nrqq*Cas9/R-loop complex (Support Figure 3).

In *wt*Cas9 (Figure 2e, Supplementary Table), residues such as R63, K510, M694, H698, Q709, S730, N767, Q771, K772, K775, R783, Q894, R925 and Y1013, etc. all bind to the phosphate groups of sgRNA in heteroduplex region. S1025 and Q1027 form electrostatic interactions with the phosphate group of dT26 (bp index 18).

Among the Cas9 residues, Q771, K772 and K775 locate in a proposed signal transducer loop L1 linker (residues 765-780) between HNH domain and RuvC domain.[21, 22, 49] While R783, Q894 and R925 reside in HNH domain. The loss of electrostatic interactions in HNH domain of *nrqq*Cas9 with R-loop, in one hand, would distort the correlation and signalling transfer between the two catalytic domains. In another hand, it would weaken *nrqq*Cas9's ability of nucleic acid sensing as well as its cleavage efficiency when comparing to *wt*Cas9.[21, 22] R919 approaches dA23 (bp index 6) and dG22 (bp index 7) and forms electrostatic interactions in *wt*Cas9/R-loop complex with frequencies higher than those in *nrqq*Cas9/R-loop complex (Figure 2f and Supplementary Table). Interestingly, there are several residues (such as K526, R655 and R765) forming strong interactions with tDNA nucleotides only in the *nrqq*Cas9/R-loop simulations.

Among the "hot" residues, M694 and H698, as well as Q695, are classified as "cluster 1" residues (in REC3 domain) whose mutation combinations generated more on-target Cas9 variants.[17] The missing electrostatic interactions between "cluster 1" residues and tDNA/sgRNA heteroduplex, observed in the *nrqq*Cas9/R-loop simulation, indicate that the *nrqq*Cas9 REC3 domain loses the ability to sense the nucleic acid to a certain degree and thus may render *nrqq*Cas9 a lower cleavage efficiency with on-target dsDNA bound, as proved in previous *nrqq*Cas9 experiments.[15]

A detailed analysis of the interaction pattern in the heteroduplex region (Figure 2e) shows that *wt*Cas9 has more polar contacts in PAM distal region (bp index 1-12) and *nrqq*Cas9 has more contacts in PAM proximal region. Interestingly, from the genome-wide mismatch tests study,[12, 13, 15, 17, 50] more mismatches exist in PAM distal region than in PAM proximal region. The consistence between experimental data and our simulation results, indicates that, the loss of non-specific polar contacts with PAM distal region between *nrqq*Cas9 and tDNA is a key factor contributing to the lower off-target effect of *nrqq*Cas9.

Dynamical features of the tDNA/sgRNA heteroduplex

Several studies indicate that the charged amino acids which bind to the backbone phosphate groups may lead to the increased flexibility in the DNA double helix.[35-38]

The flexibility of biomolecules is well described by the configurational entropy.[51, 52] For the *wt*Cas9/R-loop complex, the configurational entropy of the tDNA/sgRNA heteroduplex (13178.0± 63.2 J/mol.K) is higher than that in the *nrqq* system (12372.2 ± 60.8 J/mol.K), while the entropies of Cas9 backbone and *ds*DNA duplex, however, are nearly the same for *wt*Cas9 and *nrqq*Cas9 (Table 2). The larger entropy of the tDNA/sgRNA heteroduplex in *wt* system indicates that the conformation of

this region is more flexible than that in *nrqq*Cas9/R-loop complex.

Table 2. The conformational entropies of the *wt* and *nrqq* systems.

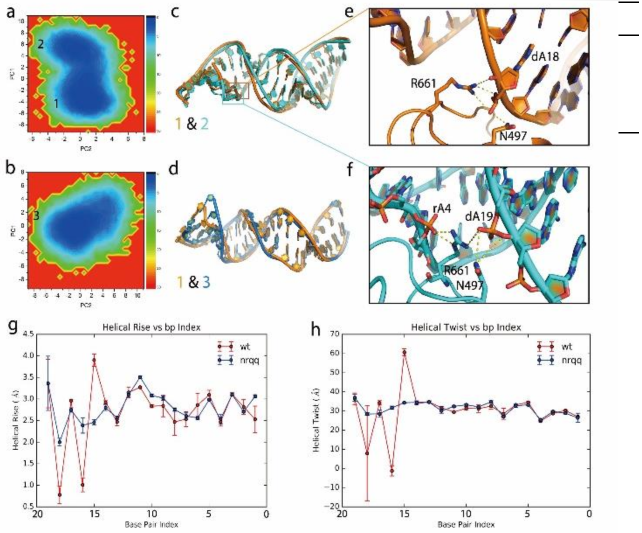

Figure 3. Conformation dynamics of tDNA/sgRNA heteroduplex. a) and b), PMF of 12 base-pair parameters projected on first 2 principle components of *wt* and *nrqq* simulation; the numbers (1, 2 and 3) in these two planes indicate the low energy basins c) and d), superimposed representative conformations in PCA PMF in panel a and b; the numbers in these two planes indicate where the representative structures locate in the PMF basins; the representative structures are selected through clustering algorithms as stated in experiment part; e) and f), zoomed two states (located in low energy basin 1 and 2 in panel a) of the heteroduplex in *wt* simulation; g) and h), helical rises and helical twists of the heteroduplex in *wt* (red color) and *nrqq* (blue color) simulations.

To further decipher the flexible origin, a PCA based on 12 helical parameters of the base-pairs in tDNA/sgRNA heteroduplex region for *wt* system and *nrqq* system was performed. The potential of mean forces (PMF) (Figure 3a and 3b) as a function of the first two PCA components indicate that there are two distinct states (orange and cyan in Figure 3c) of the heteroduplex region in *wt* system. The major differences between the two states of the heteroduplex in *wt* system locate near the PAM distal end (bp index 13-20). In *wt*Cas9/R-loop complex, the positively charged residue R661 adopts two distinct orientations (Figure 3e and 3f). In one case, R661 approaches the phosphate group of dA19 in tDNA (Figure 3c); whereas, in another state, R661 rotates to form hydrogen bonds with rA4 and rC5 in sgRNA (bp index 4 and 5) (Figure 3f). N497 sidechain also forms salt-bridges with phosphate groups of dA18 and dA19 in the two states respectively (Figure 3e and 3f). The low energy gap (less than 3 kJ/mol) between the two states implies the frequent transitions between the two conformational states of the R-loop. Moreover, the interactions cause local twist and narrow heteroduplex major groove width (Support Figure 4a).

However, the tDNA/sgRNA heteroduplex in *nrqq*Cas9/R-loop complex has only one major type of conformations, and no twist form around the rA4 and rC5 in sgRNA (Figure 3b and 3d).

The heteroduplex region behaves differently in the base-pair rise and twists in *wt* and *nrqq* systems (Figure 3g and 3h). In *wt*Cas9/R-loop complex, the base-pair rises and twists fluctuate largely within the distal PAM end, which indicates a disordered PAM distal segment of bp index1-12. Consistently the standard deviations of the helical rises, are significantly larger in *wt* system than those in *nrqq* system.

Besides, it was proposed that bending of the tDNA/sgRNA heteroduplex region is required for R-loop formation.[49] Our simulation results observe a larger bending of the heteroduplex in *wt*Cas9/R-loop complex than *nrqq*Cas9/R-loop complex. The means of the global bending angle of the heteroduplex region are around 67˚ and 53˚ for the *wt* and *nrqq* systems. The averaged local helical bending angles are 82˚ and 59˚ for the distal PAM half (base-pairs 1-12), 35˚ and 28˚ for the PAM proximal half (base-pairs 13-20), in *wt* and *nrqq* systems respectively. The global and local helical bending angles (Support Figure 4b) also reveal that the heteroduplex region is less bent in *nrqq*Cas9/R-loop complex than those in *wt*Cas9/R-loop complex.

In summary, MD simulations clearly show that the tDNA/sgRNA heteroduplex region, especially the PAM distal end, is more flexible in *wt* system than *nrqq* system.

The "opening" dynamics of mutated Cas9

Domain dynamics in *wt*Cas9 and *nrqq*Cas9 systems were examined by PCA of αC atoms coordinates. The first two principal components (PC1 and PC2) of the combined trajectories in *wt* and *nrqq*Cas9 simulation systems shows that *wt*Cas9 and *nrqq*Cas9 sample quite different conformational spaces (Figure 4a).

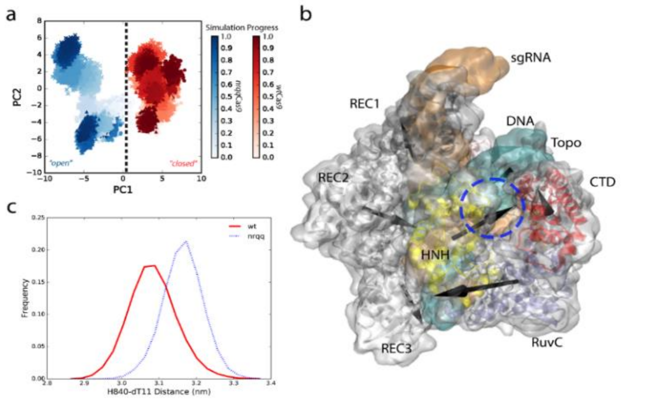

Figure 4. PCA analysis and domain motions of Cas9. a), the projection of the first 2 PC in PCA analysis for R-loop bound wtCas9 (red) and R-loop bound nrqqCas9 (blue). The projected points of three repeated simulations are shown as "x" symbols, dots, and triangles. b), The arrows indicate positive motion directions of domains in Cas9. c) the distribution frequencies of the distance between H840 and dT11 phosphate group in wt (red curve) and nrqq (blue dashed curve) simulations. RNA, DNA, α-helical lobe (including REC1, REC2 and REC3), RuvC domain, HNH domain, Topo domain and CTD domain are in orange, cyan, gray, ice blue, yellow, pink and red respectively. The active cleavage area is marked by blue dashed cycle in panel b.

To decipher the variable dynamic behaviours of Cas9 domains, the vector of PC1 is shown in Figure 4b. The vectors of all atoms in a domain were added and presented as a black arrow, whose positive direction thus indicates the dynamical features of a domain in *wt*Cas9. Clearly in comparison with *nrqq*Cas9, HNH domain of *wt*Cas9 moves towards the cleavage site, and RuvC domain shifts towards REC3 domain, (Support Movie, Support Figure 5) resembling the "open-to-closed" conformational transition which is required for efficient cleavage of dsDNA.[17, 21, 22, 49] However, in *nrqq*Cas9 (the opposite directions of the arrows in Figure 4b), HNH domain moves far apart from RuvC domain (Support Movie, Support Figure 5), as well as the active cleavage site. The conformational space sampled by *nrqq*Cas9, therefore, is more opened than that of *wt*Cas9.

Correspondingly the opening of HNH domain in *nrqq*Cas9 indeed causes a shift of the residue (H840) far from the tDNA cleavage site (dT11 in tDNA) (Figure 4c) in comparison to *wt*Cas9. The larger distance between H840 and dT11 phosphate group would lead to lower cleavage efficiency of this tDNA strand in *nrqq*Cas9

In summary, with on-target R-loop bound, *nrqq*Cas9 adopts slightly opening conformations with HNH domain moving away from the active cleavage area and large distance between H840 and cleavage site (dT11 phosphate group).[53] The opening dynamics of *nrqq*Cas9 and slightly larger distance between HNH

active cleavage residue and tDNA cleavage site echo the experimental finding that *nrqq*Cas9 has a lower cleavage efficiency for on-target dsDNAs.[15, 17]

Weaker allosteric effects caused by mutations

The community network analysis based on the contact map could cluster strong related residues into groups (or called communities), as well as their connections (or called betweennesses) with each other.[41, 42, 54] The community network analysis indicates that the allosteric effects, which are essential for highly efficient cleavage of on-target dsDNA, are weakened in the nrqqCas9.

Introducing NRQQ mutations in Cas9/R-loop complex causes a larger number of communities (Figure 5). More communities, which mean more fragile connections between residues, are observed in *nrqq*Cas9. This is also a signal of weakened allosteric effects and weaker correlations between domains, as found in another modeling study of PAM void Cas9/R-loop complex.[21]

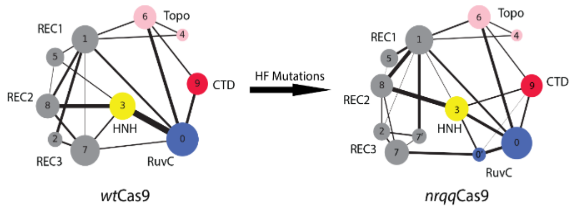

Figure 5. Community network analysis of *wt*Cas9 and *nrqq*Cas9. The information flow strength between communities are represented by the width of the edges connecting between them. The motifs in alpha-helical lobe, RuvC domain, HNH domain, CTD and Topo domain are in grey, ice blue, yellow, red and pink

The information flow (or connections) between the two catalytic domains, HNH domain (community 3) and RuvC domain (community 0) are much weaker in *nrqq*Cas9, while the connections between REC2 domain (community 8) and HNH domain is enhanced slightly (Figure 5). In *wt*Cas9, the betweenness of the largest edge (formed between residues T769 and N776) community 0 and community 3, is about twice of that in *nrqq*Cas9 (Supplementary Table). The weakened connections between these two residues T769 and N776, both located in the “allosteric transducer” L1 loop,[21, 22, 49] thus restrict the information flow between these two catalytic domains. Interestingly, the community network analys is consistent with PCA analysis and the essential dynamics of *nrqq*Cas9 which also suggest that the catalytic domain HNH shifts towards the alpha helical lobe motif REC2 but moves away from the cleavage site (Figure 4b).

Discussion

Early studies indicate that CRISPR-Cas9 system could cleave dsDNA with up to 6 mismatches in the tDNA/sgRNA heteroduplex region.[9, 14, 16, 55] Even mismatches in the PAM region could be tolerated.[15, 17] Genome wide searching for off-target sites further demonstrates that numerous DSBs were mistakenly created by CRISPR-Cas9 system.[12, 50, 56] Though Hus et al suggests that even in PAM proximal region, mismatches could be tolerated to some extent in a sequence dependent way,[10] it needs to mention that, mismatches are more tolerated in the PAM distal end than in the PAM proximal end of the tDNA/sgRNA heteroduplex region.[9, 10, 14, 15, 17, 56-58] And it is widely accepted that the PAM proximal region of the tDNA/sgRNA heteroduplex region is the primary seed sequence for binding affinity, specificity and cleavage.[9, 58, 59] And a computational modelling study of the off-target effects further validates the conclusion.[60] We have witnessed the larger magnitude of electrostatic interaction loss in PAM distal end than PAM proximal end in *nrqq*Cas9/R-loop complex, which is consistent with the experimental findings.

Many DNA binding proteins approach DNA helices resulting in distortions, such as unwinding, twisting, bending and kinking.[61-65] The local twisting, helical bending and base pair kinking, are due to the escaping of backbone torsion angles from the local torsional energy minima, thus increase the base pair flexibility.[66] Different from dsDNA, double strand RNA (dsRNA) helices maintain the A-form configuration unchanged upon protein binding,[67-69] and the main reason lies in the fact that the dsRNA helices are relatively more rigid than dsDNA helices.[70, 71] However, dsRNA helices also bend slightly upon protein binding. For example, the nonstructural protein 1 (NS1) of Influenza A binds to dsRNA and cause a ~40° bending in the dsRNA helix, however no much other forms of deformation in the dsRNA helix were found.[72] Similar case was also recorded for an endonuclease with dsRNA complex.[67] Though similar protein induced structure deformation has rarely been recorded for DNA/RNA hybrid helices, a computational study suggests that the rigidity of DNA/RNA hybrid helices lies between dsRNA and dsDNA helices.[71] Most of the biological relevant dsDNA helices adopt the B-form and dsRNA helices adopt the A-form, the conformations of DNA/RNA hybrid helices are more or less in the mediate form,[71] or sometimes the distorted B-form.[73]

The conformational changes of DNA helices are required for specific and non-specific binding, and these changes would also increase the overall flexibility of the DNA base pairs in binding interface, as seen in our simulations. It is known that in most cases mismatches in DNA duplexes and RNA/DNA heteroduplexes will increase the local flexibility and compromise the stability of the helix,[74, 75] whereas in a Cas9/R-loop complex, heteroduplex local bending and twisting may be required.[49, 76] As observed in references,[15, 56] Cas9 could tolerate mismatches in the sgRNA/tDNA heteroduplex region, therefore the flexibility of the heteroduplex would be induced by abundant electrostatic interactions and would be tolerated, it may be also rational to believe that the void of the electrostatic interactions in Cas9/R-loop interface then results in the rigidity of the heteroduplex, and thus would promote the less tolerance of the mismatches of the heteroduplex. Overall, for *nrqq*Cas9, less electrostatic interactions would, at least partially, hint the origin of the off-target effect.

It has been accepted that the selective mismatch tolerance for *nrqq*Cas9, as well as another Cas9 variant, eCas9, through an activation threshold mechanism enables higher cleavage specificity lower cleavage efficiency.[15, 17, 18] HNH domain of *wt*Cas9 could adopt active conformation with either on-target dsDNA or mismatched dsDNA, while *nrqq*Cas9 HNH domain adopts large ratio of inactive conformations with respect to active conformations even with on-target dsDNA sequences.[17] Our finding of the “opening” preference of *nrqq*Cas9 with on-target dsDNA binding, therefore is consistent with the low active conformation ratio discovered using FRET experiments. Meanwhile, REC3 domain, which senses the dynamics changes of tDNA/sgRNA PAM distal end, can relocate REC2 domain to further promote HNH domain adopting the activation state upon on-target dsDNA binding.[22] Whereas, upon off-target dsDNA binding, the conformational changes in PAM distal end of tDNA/sgRNA heteroduplex, inhibit the correlations between REC3 and REC2 domains, thus HNH domain is less able to adopt the activation conformation. It thus offered an allosteric hypothesis, where the correlations and cooperative conformational changes between Cas9 domains could sense mismatches in the heteroduplex PAM distal end through "cluster 1" residues in Cas9.[17] More recent researches concentrate on the importance of the unwinding of the dsDNA towards the cleavage.[19, 20] CRISPR-Cas9 system prosses a slow R-loop formation step and a fast cleavage step, where the stability of the unwound dsDNA determined the overall cleavage efficiency by HNH and RuvC domains.[19] We may thus hypothesize that the transient unwound dsDNA could form sequence complementary through its tDNA strand with sgRNA and promote the formation of the R-loop, and further is stabilized by interactions with Cas9. And the loss of electrostatic interactions between *nrqq*Cas9 and on-target tDNA strand and sgRNA strand, would not facilitate the stabilization of the transient unwound dsDNA, thus leads to low on-target cleavage efficiency. However, a more detailed atomic level mechanism of the selective mismatch tolerance is still void.

It was originally proposed that the HF mutations would reduce the interaction energies between Cas9 and R-loop,[15] however later study opposes the hypothesis.[17] Our simulation data show that the different levels of electrostatic interactions (Figure 2 and Supplementary Table) between tDNA and Cas9 in the distal and proximal PAM regions hint the diverse tolerance against mismatches. tDNA at distal PAM region (more mismatch detected) forms stronger polar interactions with Cas9, whereas tDNA at proximal PAM region (less mismatches detected) forms less polar contacts with Cas9. The *nrqq*Cas9 has less electrostatic interactions with tDNA in the tDNA/sgRNA heteroduplex region, which has lower conformational entropy and more rigid conformations comparing to the *wt*Cas9/R-loop simulations. The dismission of the interactions between R-loop and Cas9 is a general strategy for high specificity Cas9 engineering.[15, 18, 20] Meanwhile, we assessed the MMPBSA based binding energies between different Cas9 mutants (as well as the *wt*Cas9) and the R-loop (see Support Figure 6 and Supplementary table). It is quite clear that the *wt*Cas9 has the lowest binding free energies towards the R-loop, while the *nrqq*Cas9 has the highest binding free energies, thus indicating that the binding strength between *wt*Cas9 and R-loop would be excess for R-loop capture during cleavage. And the electrostatic interaction energies after HF mutations, as observed in the calculations, may partially explained by the loss of electrostatic contacts between *nrqq*Cas9 with the 20 bp heteroduplex comparing to the *wt*Cas9/R-loop system.

Therefore, the tightly interacting residues around the Cas9/R-loop interface as recorded in Supplementary Table are potential candidates. Except the 4 HF mutations, we propose that the following residues (mainly located in α-helical lobe and HNH domain), K526, K528, N588, G658, R895, N899, K902, R905, R919, R925, H930. There residues stably interact (over 80% of the simulations in either *wt* or *nrqq* systems) with tDNA in 20 bp heteroduplex PAM distal end, would be the potent target for mutations towards more precise and highly efficient Cas9 engineering.

The existence of mismatches usually induces more flexible conformations of nucleic acids. Our simulations provide valuable insights for understanding the low on-target cleavage efficiency, as well as the low off-target effects of *nrqq*Cas9: the rigidity of the tDNA/sgRNA heteroduplex region of *nrqq*Cas9 excludes the high probability of mismatching, and the *nrqq*Cas9 may favors the inactive conformations with on-target sequences, thus account for the lower on-target cleavage ability. Future work should further explore the mismatch tolerance of *nrqq*Cas9 with mismatched sequence binding.

Conclusions

We performed several MD simulations of *wt*Cas9 and *nrqq*Cas9 together with R-loop, to uncover mutation induced dynamics changes. The results indicate that the HF mutations cause loss of electrostatic interactions between Cas9 and tDNA/sgRNA heteroduplex, especially in PAM distal end. In *nrqq*Cas9/R-loop complex, the tDNA/sgRNA heteroduplex become more rigid and ordered than the same region in *wt*Cas9. It is also observed that the PAM distal end of the tDNA/sgRNA heteroduplex is more disordered and flexible in *wt*Cas9/R-loop, which is consistent with the previous experimental and computational studies. Meanwhile, after mutations, *nrqq*Cas9 samples slight open conformations, and the HNH domain has lower connections with RuvC domain. Thus, these changes render *nrqq*Cas9 less efficient for on-target dsDNA cleavage, and probably off-target dsDNA cleavage as well. These findings would be crucial for future high-efficient Cas9 engineering towards more robust and accurate genome editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The computational work for this article was fully performed on GPU resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

YM acknowledges the grant of MOE Tier 1 RG146/17. JS is supported by the Chinese Natural Science Foundation Grant (81770099) and the Shenzhen Science and Technology Innovation Commission (JCYJ20170413115637100 and

JCYJ20170412152916724).The authors would like to thank Dr. Ning Lulu, Shaanxi University of Science and Technology, China, for her assistance in generating Cas9 motion graphics using VMD.

Notes and references

1. R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath, *Science*, 2007, **315**, 1709-1712.
2. J. E. Garneau, M. E. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadan and S. Moineau, *Nature*, 2010, **468**, 67-71.
3. L. A. Marraffini and E. J. Sontheimer, *Nature reviews. Genetics*, 2010, **11**, 181-190.
4. R. Sorek, C. M. Lawrence and B. Wiedenheft, *Annual review of biochemistry*, 2013, **82**, 237-266.
5. K. S. Makarova, D. H. Haft, R. Barrangou, S. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost and E. V. Koonin, *Nature reviews. Microbiology*, 2011, **9**, 467-477.
6. A. E. Briner and R. Barrangou, *Cold Spring Harbor protocols*, 2016, **2016**, pdb top090902.
7. F. Jiang and J. A. Doudna, *Annual review of biophysics*, 2017, **46**, 505-529.
8. M. Takata, M. S. Sasaki, E. Sonoda, C. Morrison, M. Hashimoto, H. Utsumi, Y. Yamaguchi ‐ Iwai, A. Shinohara and S. Takeda, *The EMBO journal*, 1998, **17**, 5497-5508.
9. M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna and E. Charpentier, *Science*, 2012, **337**, 816-821.
10. P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu and O. Shalem, *Nature biotechnology*, 2013, **31**, 827-832.
11. A. Hruscha, P. Krawitz, A. Rechenberg, V. Heinrich, J. Hecht, C. Haass and B. Schmid, *Development*, 2013, **140**, 4982-4987.
12. D. Kim, S. Bae, J. Park, E. Kim, S. Kim, H. R. Yu, J. Hwang, J.-I. Kim and J.-S. Kim, *Nature methods*, 2015, **12**, 237-243.
13. Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung and J. D. Sander, *Nature biotechnology*, 2013, **31**, 822-826.
14. P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville and G. M. Church, *Science*, 2013, **339**, 823-826.
15. B. P. Kleinstiver, V. Pattanayak, M. S. Prew, S. Q. Tsai, N. T. Nguyen, Z. Zheng and J. K. Joung, *Nature*, 2016, **529**, 490-495.
16. Y. Lin, T. J. Cradick, M. T. Brown, H. Deshmukh, P. Ranjan, N. Sarode, B. M. Wile, P. M. Vertino, F. J. Stewart and G. Bao, *Nucleic acids research*, 2014, gku402.
17. J. S. Chen, Y. S. Dagdas, B. P. Kleinstiver, M. M. Welch, A. A. Sousa, L. B. Harrington, S. H. Sternberg, J. K. Joung, A. Yildiz and J. A. Doudna, *Nature*, 2017, DOI: 10.1038/nature24268.
18. I. M. Slaymaker, L. Gao, B. Zetsche, D. A. Scott, W. X. Yan and F. Zhang, *Science*, 2016, **351**, 84-88.
19. S. Gong, H. H. Yu, K. A. Johnson and D. W. Taylor, *Cell reports*, 2018, **22**, 359-371.
20. D. Singh, Y. Wang, J. Mallon, O. Yang, J. Fei, A. Poddar, D. Ceylan, S. Bailey and T. Ha, *Nat Struct Mol Biol*, 2018, **25**, 347-354.
21. G. Palermo, C. G. Ricci, A. Fernando, R. Basak, M. Jinek, I. Rivalta, V. S. Batista and J. A. McCammon, *Journal of the American Chemical Society*, 2017, DOI: 10.1021/jacs.7b05313.
22. S. H. Sternberg, B. LaFrance, M. Kaplan and J. A. Doudna, *Nature*, 2015, **527**, 110-113.
23. C. Anders, O. Niewoehner, A. Duerst and M. Jinek, *Nature*, 2014, **513**, 569-573.
24. T. Schwede, J. Kopp, N. Guex and M. C. Peitsch, *Nucleic acids research*, 2003, **31**, 3381-3385.
25. M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**, 19-25.
26. K. Lindorff ‐ Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins: Structure, Function, and Bioinformatics*, 2010, **78**, 1950-1958.
27. A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton and M. Orozco, *Biophysical journal*, 2007, **92**, 3817-3829.
28. P. Mark and L. Nilsson, *The Journal of Physical Chemistry A*, 2001, **105**, 9954-9960.
29. G. Bussi, D. Donadio and M. Parrinello, *The Journal of chemical physics*, 2007, **126**, 014101.
30. H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola and J. Haak, *The Journal of chemical physics*, 1984, **81**, 3684-3690.
31. J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, *Journal of Computational Physics*, 1977, **23**, 327-341.
32. B. Hess, H. Bekker, H. J. Berendsen and J. G. Fraaije, *Journal of computational chemistry*, 1997, **18**, 1463-1472.
33. M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci and R. A. Broglia, *Computer Physics Communications*, 2009, **180**, 1961-1972.
34. W. Humphrey, A. Dalke and K. Schulten, *Journal of molecular graphics*, 1996, **14**, 33-38.
35. T. M. Okonogi, S. C. Alley, E. A. Harwood, P. B. Hopkins and B. H. Robinson, *Proceedings of the National Academy of Sciences*, 2002, **99**, 4156-4160.
36. A. Podestà, M. Indrieri, D. Brogioli, G. S. Manning, P. Milani, R. Guerra, L. Finzi and D. Dunlap, *Biophysical journal*, 2005, **89**, 2558-2563.
37. A. Lebrun and R. Lavery, *Biopolymers*, 1999, **49**, 341-353.
38. L. D. Williams and L. J. Maher III, *Annual review of biophysics and biomolecular structure*, 2000, **29**, 497-521.
39. M. Karplus and J. N. Kushick, *Macromolecules*, 1981, **14**, 325-332.
40. U. Hensen, O. F. Lange and H. Grubmuller, *PLoS One*, 2010, **5**, e9179.
41. J. Guo and H.-X. Zhou, *Chemical reviews*, 2016, **116**, 6503-6515.
42. A. Sethi, J. Eargle, A. A. Black and Z. Luthey-Schulten, *Proceedings of the National Academy of Sciences*, 2009, **106**, 6620-6625.
43. A. Hagberg, P. Swart and D. S Chult, *Exploring network structure, dynamics, and function using NetworkX*, Los Alamos National Laboratory (LANL), 2008.
44. R. Kumari, R. Kumar and A. Lynn, *Journal of chemical information and modeling*, 2014, **54**, 1951-1962.

45. S. Genheden and U. Ryde, *Expert opinion on drug discovery*, 2015, **10**, 449-461.
46. B. R. Miller III, T. D. McGee Jr, J. M. Swails, N. Homeyer, H. Gohlke and A. E. Roitberg, *Journal of chemical theory and computation*, 2012, **8**, 3314-3321.
47. R. Kumar and H. Grubmüller, *Bioinformatics*, 2015, btv190.
48. X. J. Lu and W. K. Olson, *Nucleic acids research*, 2003, **31**, 5108-5121.
49. F. Jiang, D. W. Taylor, J. S. Chen, J. E. Kornfeld, K. Zhou, A. J. Thompson, E. Nogales and J. A. Doudna, *Science*, 2016, **351**, 867-871.
50. S. Q. Tsai, Z. Zheng, N. T. Nguyen, M. Liebers, V. V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A. J. Iafrate and L. P. Le, *Nature biotechnology*, 2015, **33**, 187-197.
51. B. Matthews, H. Nicholson and W. Becktel, *Proceedings of the National Academy of Sciences*, 1987, **84**, 6663-6667.
52. A. A. Rashin, *Biopolymers*, 1984, **23**, 1605-1620.
53. Z. Zuo and J. Liu, *Scientific reports*, 2016, **6**, 37584.
54. C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros and P. Csermely, *Febs Letters*, 2007, **581**, 2776-2782.
55. Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, J. K. Joung and J. D. Sander, *Nature biotechnology*, 2013, **31**, 822-826.
56. C. Kuscu, S. Arslan, R. Singh, J. Thorpe and M. Adli, *Nature Biotechnology*, 2014, **32**, 677-+.
57. R. Sapranauskas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath and V. Siksnys, *Nucleic acids research*, 2011, gkr606.
58. L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang and L. A. Marraffini, *Science*, 2013, **339**, 819-823.
59. W. Jiang, D. Bikard, D. Cox, F. Zhang and L. A. Marraffini, *Nature biotechnology*, 2013, **31**, 233-239.
60. I. Farasat and H. M. Salis, *Plos Comput Biol*, 2016, **12**.
61. J. A. Mcclarin, C. A. Frederick, B. C. Wang, P. Greene, H. W. Boyer, J. Grable and J. M. Rosenberg, *Science*, 1986, **234**, 1526-1541.
62. T. J. Richmond, *Nature*, 1987, **326**, 18-19.
63. F. H. Allain, Y. M. Yen, J. E. Masse, P. Schultze, T. Dieckmann, R. C. Johnson and J. Feigon, *The EMBO journal*, 1999, **18**, 2563-2579.
64. M. Mondal, S. Mukherjee and D. Bhattacharyya, *Journal of molecular modeling*, 2014, **20**, 2499.
65. S. Ferrari, V. R. Harley, A. Pontiggia, P. N. Goodfellow, R. Lovell-Badge and M. E. Bianchi, *The EMBO journal*, 1992, **11**, 4497.
66. J. W. Keepers, P. A. Kollman, P. K. Weiner and T. L. James, *Proceedings of the National Academy of Sciences*, 1982, **79**, 5537-5541.
67. H. Wu, A. Henras, G. Chanfreau and J. Feigon, *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**, 8307-8312.
68. B. Tian, P. C. Bevilacqua, A. Diegelman-Parente and M. B. Mathews, *Nature reviews Molecular cell biology*, 2004, **5**, 1013-1023.
69. X. Li, N. Zhou, W. Chen, B. Zhu, X. Wang, B. Xu, J. Wang, H. Liu and L. Cheng, *Journal of Molecular Biology*, 2017, **429**, 79-87.
70. P. Kebbekus, D. E. Draper and P. Hagerman, *Biochemistry*, 1995, **34**, 4354-4357.
71. T. E. Cheatham and P. A. Kollman, *Journal of the American Chemical Society*, 1997, **119**, 4805-4825.
72. A. Cheng, S. M. Wong and Y. A. Yuan, *Cell research*, 2009, **19**, 187-195.
73. P. Yin, D. Deng, C. Yan, X. Pan, J. J. Xi, N. Yan and Y. Shi, *Cell reports*, 2012, **2**, 707-713.
74. N. Sugimoto, M. Nakano and S. Nakano, *Biochemistry*, 2000, **39**, 11270-11281.
75. H. T. Allawi and J. SantaLucia, Jr., *Biochemistry*, 1997, **36**, 10581-10594.
76. M. Jinek, F. Jiang, D. W. Taylor, S. H. Sternberg, E. Kaya, E. Ma, C. Anders, M. Hauer, K. Zhou and S. Lin, *Science*, 2014, **343**, 1247997.