scientific reports

OPEN

Developing small Cas9 hybrids using molecular modeling

Antoine Mangin[1,2], Vincent Dion[1,2]✉ & Georgina Menzies[3]✉

The contraction of CAG/CTG repeats is an attractive approach to correct the mutation that causes at least 15 neuromuscular and neurodegenerative diseases, including Huntington's disease and Myotonic Dystrophy type 1. Contractions can be achieved in vivo using the Cas9 D10A nickase from *Streptococcus pyogenes* (SpCas9) using a single guide RNA (sgRNA) against the repeat tract. One hurdle on the path to the clinic is that SpCas9 is too large to be packaged together with its sgRNA into a single adeno-associated virus. Here we aimed to circumvent this problem using the smaller Cas9 orthologue, SlugCas9, and the Cas9 ancestor OgeuIscB. We found them to be ineffective in inducing contractions, despite their advertised PAM sequences being compatible with CAG/CTG repeats. Thus, we further developed smaller Cas9 hybrids, made of the PAM interacting domain of *S. pyogenes* and the catalytic domains of the smaller Cas9 orthologues. We also designed the cognate sgRNA hybrids using molecular dynamic simulations and binding energy calculations. We found that the four Cas9/sgRNA hybrid pairs tested in human cells failed to edit their target sequences. We conclude that in silico approaches can identify functional changes caused by point mutations but are not sufficient for designing larger scale complexes of Cas9/sgRNA hybrids.

Keywords Cas9, Molecular dynamics, Binding energy, Gene editing

The clustered regularly interspaced short palindromic repeats (CRISPR) Cas system evolved as an immunity mechanism against phage invasion[1,2]. But once Cas9 from *Streptococcus pyogenes* was found to be a programmable RNA-guided nuclease creating DNA double-strand breaks, gene editing at high efficiency in mammalian cells became possible[3,4]. Type II CRISPR/Cas complexes are formed of a Cas9 nuclease guided to a target sequence using a CRISPR (cr)RNA. This contains a customizable sequence (~ 20 nucleotides), and a trans-activating crRNA (tracrRNA) formed of one or more stem loops and are unique to each Cas9 orthologue[5]. The two RNAs can be fused together to produce a convenient single guide (sg)RNA[3]. The sgRNA-Cas9 complex recognizes its complementary sequence in genomic DNA using base pairing when the target sequence is followed by a short (3 to 8 bp), species-specific, protospacer adjacent motif (PAM). Upon binding, the HNH and RuvCI domains of Cas9 are activated and induce a DNA double-strand break, with the HNH cutting the complementary DNA strand while the RuvCI cleaves the noncomplementary strand[3]. The double-strand break is repaired by the cell's endogenous machinery, leading to insertions and deletions[4,6]. Homologous recombination can also occur, provided that a template for repair is present[4,6].

Due to its high cutting efficiency in mammalian cells, the Cas9 enzyme from *Streptococcus pyogenes* (Sp) is the most widely studied of the Cas9 orthologues[4,7], and has been used in the clinic[8]. SpCas9 is most efficient when the PAM sequence is composed of 5'-NGG (with N being any nucleotide), but also tolerates 5'-NAG and, to a lower extent, 5'-NGA and 5'-NTG[9–11]. This ability to accommodate mismatches between the target sequence and the sgRNA means that some off-target sequences are more accessible and unwanted mutations can occur[12,13]. This has led to a flurry of studies aiming to modify Cas9 to make it more specific or to change the compatibility of its PAM sequence. These studies used a variety of methods, including structure-based rational design[11,14–18] and screening[19–21]. Luan et al.[22], for example, used molecular dynamics simulations and free energy perturbation to shorten the PAM recognition sequence of *Staphylococcus aureus* (Sa) Cas9. They found that free energy perturbation correlated well with gene editing efficiency in human cells. These results suggest that molecular dynamics simulations can assist the design of Cas9 variants quickly and cheaply.

SpCas9, with its 1368 amino acids, is too large to fit into an adeno-associated virus (AAV) together with its sgRNA for in vivo delivery. AAVs are one of best delivery systems because they display low immunoreactivity, have a wide tropism, and their viral DNA does not integrate easily within the host genome[23]. The size of SpCas9 means that two AAVs are required, one encoding the Cas9 enzyme, the other the sgRNA, resulting in increased

[1]UK Dementia Research Institute at Cardiff University, Cardiff CF24 4HQ, UK. [2]Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Maindy Road, Cardiff CF24 4HQ, UK. [3]School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK. ✉email: dionv@cardiff.ac.uk; menziesg@cardiff.ac.uk

cost, reduced efficiency, and an enhanced risk of an adverse immune reaction. To get around this problem, smaller Cas9 orthologues, like SaCas9 or *Campylobacter Jejuni* (Cj)Cas9, have been used in animal studies as they are small enough to be packaged into a single AAV with their sgRNAs[24–26]. However, they use a different PAM sequences. SaCas9 recognizes NNGRRT (where R is a purine) and CjCas9 recognizes NNNNACA[26,27]. This means that most sequences targeted by SpCas9 are inaccessible to these smaller orthologues. For many applications, this is not a problem as the sequence to be targeted can be adjusted to that of the smaller orthologues. However, this is not as trivial when there is a need for a specific target sequence. For example, we have previously showed that targeting CAG/CTG repeats with the SpCas9 nickase (D10A) leads to efficient contractions in patient-derived cells as well as in vivo[28]. This is dependent on targeting the repeat tract itself and so the sequence of the sgRNA cannot be changed[29]. Similarly, if single-nucleotide polymorphisms from patients need to be targeted, then there is no flexibility in the sequence to be targeted[30].

To overcome these issues, here we aimed to test whether the Cas9 orthologue SlugCas9[31] or the CRISPR ancestor, OgeuIscB[32], which are both predicted to target CAG/CTG repeats, could lead to repeat contractions. We found however that neither of them could induce contractions of CAG/CTG repeats in human cells. We therefore developed new compact Cas9 hybrids by replacing the PAM Interacting Domain (PID) of SaCas9 and CjCas9 with that of SpCas9. This was done to maintain the PAM properties of the SpCas9 while reducing the size of the protein. The main challenge was that sgRNAs are species-specific[33], and therefore we had to develop new sgRNA hybrids and ensure that they interact with the Cas9 hybrids. We used a two pronged approach for this: first a molecular modelling approach where we calculated the binding energy of the sgRNA to the Cas9 hybrids, and then tested four of the novel hybrid/sgRNA pairs for their ability to induce contractions of CAG/CTG repeats. We found that unlike with point mutations, the in silico approach was not sufficient to predict the function of large chimeric systems.

Results:

SlugCas9-KH and OgeuIscB nickases do not induce the contraction of CAG/CTG repeats

The variant of the *Staphylococcus lugdunensis* orthologue, SlugCas9-KH, is one of very few CRISPR enzymes that has been reported to use CAG or CTG as its PAM sequence[31]. Its small size (1054 aa) and its PAM preference for NNRG (R = A or G) made it an ideal orthologue to test for its ability to induce contraction of CAG repeats. We therefore made the D10A mutation to turn this nuclease into a nickase. We used it in combination with a well characterized assay for the contraction of CAG repeats in GFP$(CAG)_{101}$[34]. The assay works because CAG repeats interfere with the splicing of the GFP reporter in a size-dependent manner with longer repeats producing less GFP (Fig. 1A,B). The SpCas9 nickase (D10A mutant) together with a sgRNA that targets the repeat tract itself induces contractions of the CAG/CTG repeat, thereby increasing GFP levels in these cells[29,35]. When we expressed the SlugCas9, together with a sgRNA against the CAG/CTG repeats (Fig. 1C,D), we found no increase in GFP levels, suggesting no activity of this orthologue on the repeats, unlike the SpCas9 D10A nickase (Fig. 1E).

Another potential CRISPR enzyme that should be able to target CAG/CTG repeats is the Cas9 ancestor, OgeuIscB[32]. It is only 429 amino-acid long and recognizes a target-adjacent motif (TAM) compatible with CAG/CTG repeats when paired with an ω-RNA. However, expressing the OgeuIscB nickase (E193A) in GFP$(CAG)_{101}$ cells failed to increase GFP levels over background levels, despite its robust expression compared to SpCas9 (Fig. 1F,G). We conclude that in its current form, OgeuIscB cannot induce contractions of CAG/CTG repeats.

Molecular dynamics and binding energy identify functional changes in Cas9

To achieve our goal of finding small CRISPR enzymes that can target CAG/CTG repeats, we turned to a different paradigm. Specifically, we used molecular dynamics, testing the hypothesis that changes in binding energy will correlate with changes in gene editing efficiency, as previously suggested for point mutations in the SaCas9[22]. First, we aimed to reproduce these data and test whether the approach could identify meaningful functional changes. To do so, we ran molecular dynamics simulations of the SaCas9 complex with and without its sgRNA. We reasoned that the presence of the sgRNA would have a large impact on the dynamics of the Cas9 enzyme. Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) values were extracted from the simulations. RMSD is a measure of the average distance for a group of atoms between a reference structure and the resulting structure after a specified amount of time. The RMSF relates to the movement of each individual residue over the time of the simulation. The SaCas9 without its sgRNA reached a maximum RMSD of 1.22 nm but remained stable beyond 40 ns (Fig. 2A), with an overall RMSD average of 1.1 nm. In contrast, the SaCas9 in a complex with its sgRNA was more stable with an average RMSD of 0.33 nm, and a maximum of 0.37 nm (Fig. 2A). Consistent with these data, the RMSF analysis showed that every domain of the protein is stabilized by the presence of its sgRNA (Fig. 2B). The REC lobe, the bridge helix, RuvCI, and HNH domains, were the most impacted. We conclude that changes in RMSD and RMSF values reflect large scale changes in SaCas9.

To determine whether we could find functional differences using binding energy measurements, we compared SaCas9 to the D10A mutant, which turns the nuclease into a nickase[3]. To do so, we used the MMPBSA approach[36–38], which works by calculating the overall energy for Cas9 in solution compared to that of Cas9 in a vacuum. The same is done for the sgRNA. From these data, we can calculate the energy difference between Cas9 and its sgRNA in solution as well as in a vacuum, which gives us an approximation of the binding affinity between the two molecules for each amino-acid (Supplementary Fig. 1A). We reasoned that the binding energy for the mutated residue should change if this method is predictive of a change in function. Moreover, the rest of the protein should change very little, if at all, by this amino acid change. To test this directly, we calculated the binding energy associated with every residue extracted from the SaCas9 simulation and compared them to that of the SaCas9 D10A variant. We found that the G binding of D10 was 912 times greater than that of A10 (Fig. 2C). This is in contrast to the other residues, which on average changed by 0.39 kJ/mol (Supplementary

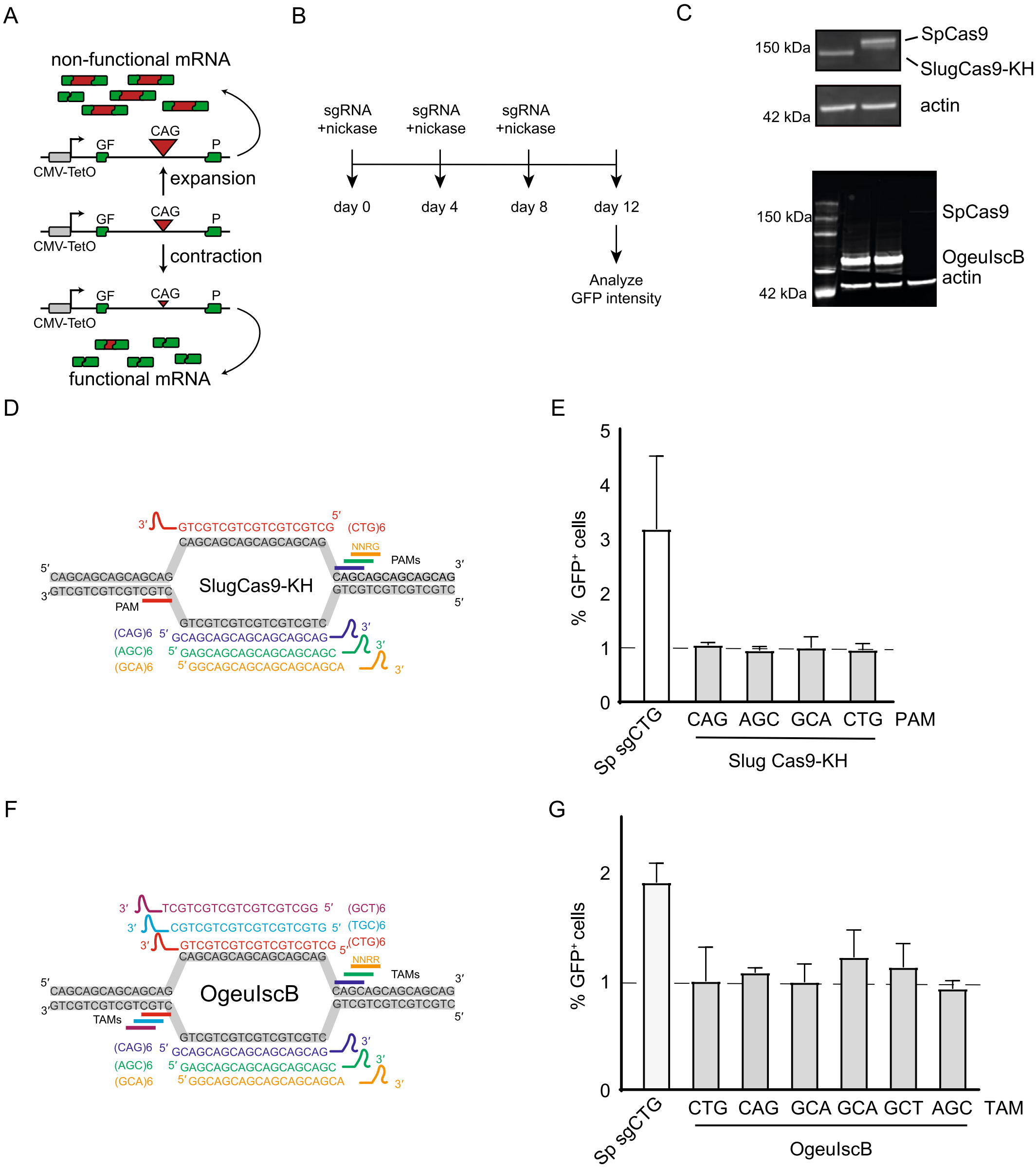
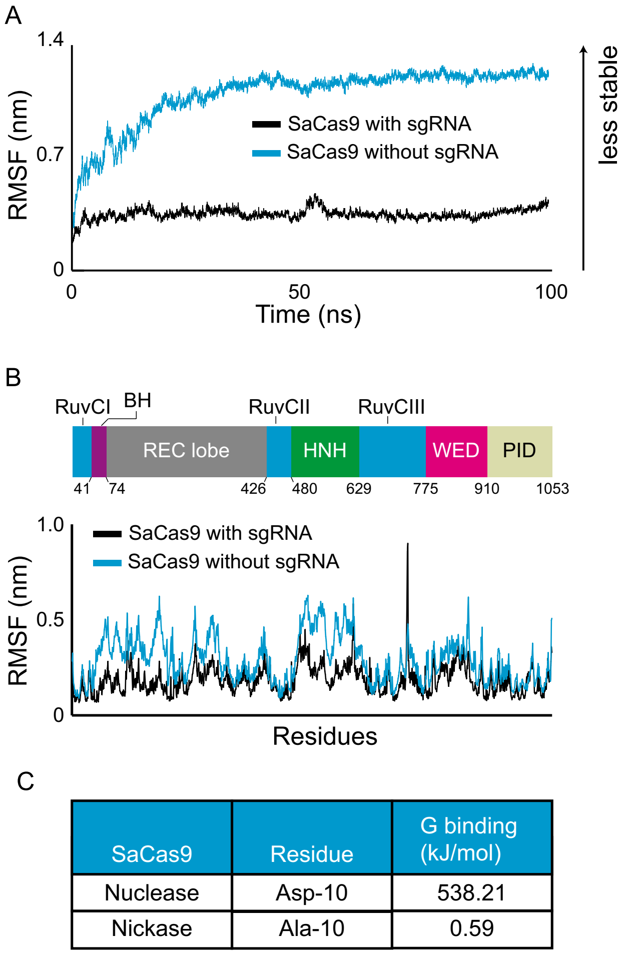


Figure 1. SlugCas9-KH and OgeuIscB nickases do not induce the contraction of CAG/CTG repeats. (**A**) GFP-based assay used to detect contraction of the CAG repeats length. The repeat interferes with the splicing of the reporter in a size-dependent manner with longer repeats inducing more missplicing and lower GFP expression. (**B**) Timeline experiment for the GFP-based assay to detect CAG contractions. (**C**) Western blots of GFP(CAG)$_{101}$ cells transfected with SpCas9D10A, SlugCas9-KH D10A, and OgeuIscB E193A. Full uncropped blots are found in Supplementary Fig. 7. (**D**) Cartoon of the sgRNAs used together with the SlugCas9-KH. (**E**) Percentage of GFP positive cells for each sgRNA. Dashed lines represent the gate containing the brightest 1% of the cells transfected with Cas9 only. Previous work[29] has shown that this population is enriched in shorter repeats. (**F**) Cartoon of the ω-RNAs targeting the CAG/CTG repeat. (**G**) Percentage of GFP positive cells for each sgRNA. Dashed lines represent the gate containing the brightest 1% of the cells transfected with Cas9 only. Previous work[29] has shown that this population is enriched in shorter repeats.

Fig. 1B). Together, these analyses show that the molecular dynamics simulation together with the binding energy calculations robustly predicts functional changes.

The Cas9 hybrid complexes are stable in silico

Next we tested whether we could use this computational approach to design multiple compact Cas9/sgRNA hybrid pairs to be tested in human cells. This was done in two steps. First, we performed molecular dynamics simulations and determined the RMSD and RMSF of the hybrid pairs and then determined which ones to take forward using binding energy calculations. We initially designed 9 such pairs in silico, 7 SaSp pairs and 2 CjSp



SaCas9	Residue	G binding (kJ/mol)
Nuclease	Asp-10	538.21
Nickase	Ala-10	0.59

Figure 2. Molecular dynamics simulations and binding energy calculation identify functionally relevant changes in Cas9 activity. (**A**) Root Mean Square Deviation (RMSD) of SaCas9 with or without its sgRNA over 100 ns simulations. (**B**) Root Mean Square Fluctuation (RMSF) of SaCas9 with or without its sgRNA. (*BH* bridge helix, *WED* wedge domain, *PID* PAM interacting domain). (**C**) G binding values at position 10 of SaCas9 comparing the nuclease (D10) to the nickase (A10).

hybrid pairs (Fig. 3A-B). For all these, as well as for the wild type versions of SpCas9, SaCas9, and CjCas9, we performed molecular dynamics simulations and analyzed the stability of these hybrids (Fig. 3A-J and Supplementary Fig. 2A-F and supplementary 3A-B). We found that in the presence of a sgRNA the RMSD for the SaSpCas9 or CjSpCas9, regardless of the sgRNA hybrid used, was systematically higher than the SaCas9, SpCas9 or CjCas9 coupled to their wild type sgRNAs (Fig. 3C,D – P value < 0.0001). However, for both hybrids, a plateau was reached, showing that the proteins arrived at a stable conformation.

When comparing the protein RMSD for the 7 SaSp models, we found that Large Sam (LSam) showed the highest maximum average at 0.93 nm whereas Pippin had the lowest maximum average at 0.60 nm (Fig. 3E). For the CjSp, Rosy had the lowest value with 0.73 nm. The same protein hybrids showed different RMSD values and were significantly different from one to another (P < 0.0001) presumably, because of how they interacted with their respective sgRNAs. Analysis of the RNA RMSD showed that Pippin had the highest value with 1.37 nm whereas the lowest values were seen with Merry (0.40 nm, Fig. 3F). For the CjSp hybrids, again, Rosy had the lowest value with 0.33 nm. The DNA RMSD did vary significantly between the SaSp systems (P < 0.0001) and ranged between 0.36 nm (Msam) and 0.41 nm (Merry; Fig. 3G). For the CjSp, Rosy had the lowest value of 0.52 nm. Altogether, the RMSD data show that the hybrids had consistently higher values than the three wild types, yet they were below what we observed for SaCas9 without the presence of a sgRNA (Fig. 2).

To complement these findings, we determined the RMSF values of each residue in the proteins and sgRNAs for each hybrid pair. We found that the N-terminal of the SaSpCas9 coupled to Merry (Fig. 2H) and to Gollum (Supplementary Fig. 2A) matched closely the wild type SaCas9 coupled to its sgRNA, with the HNH domain being the most fluctuating domain for both proteins. The PID of the SaSp hybrids (Fig. 3I and Supplementary Fig. 2A-F) were more mobile than the Sp wild type pair, especially for some atoms around residue 1000. This may be explained by the energy minimization step that moved the initial position of these residues away from other atoms, giving it more space to move. The RMSF for SaSp Merry sgRNA was closest to that of the Sp and Sa sgRNAs (Fig. 3J) with the exception of stem loop 2 from Sp wild type showing more fluctuations. Similar to

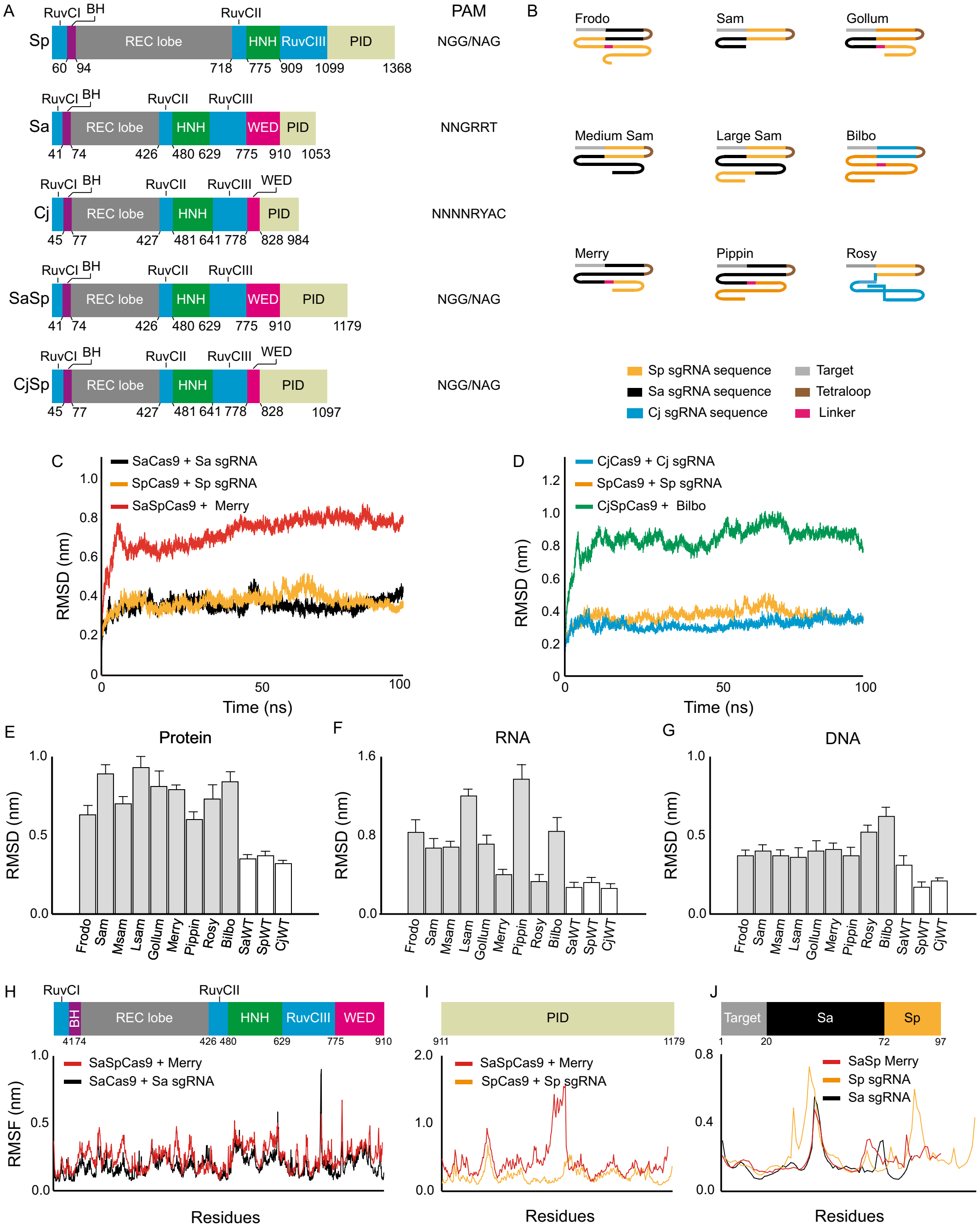


Figure 3. Molecular dynamics simulations of Cas9 hybrids. (**A**) Domain organization of the Sp, Sa, and Cj Cas9 as well as the hybrids tested here. The PAM sequence is either the reported ones or, for the hybrids, the expected PAM. The numbers indicate the position of amino acids at domain boundaries (*BH* bridge helix, *WED* wedge domain, *PID* PAM interacting domain). (**B**) Schematic representation of the sgRNA hybrids tested here. (**C**) RMSD of SpCas9, SaCas9, and SaSpCas9 with their specific sgRNAs during 100 ns. The SaSp sgRNA presented is Merry. (**D**) RMSD of SpCas9, CjCas9, and CjSpCas9 with their specific sgRNAs over a simulation of 100 ns. The CjSp sgRNA presented is Bilbo. (**E**–**G**) Average RMSD for the protein (**E**), sgRNA (**F**), and DNA (**G**) in all 12 systems modeled. (**H**) RMSF of SaCas9 with its sgRNA (black) and SaSpCas9 in combination with Merry (red) for the N-terminal common to both proteins. (**I**) RMSF of the PID of SpCas9 with its sgRNA (yellow) and SaSpCas9 in combination with Merry (red). (J) RMSF of SpCas9 sgRNA (yellow), SaCas9 sgRNA (black) and SaSpCas9 sgRNA (red).

the SaSpCas9 hybrids, we found that the CjSpCas9 hybrids were consistently more mobile than the wild types (Supplementary Fig. 3A-B). Altogether, the RMSD and RMSF data suggest that the Cas9 hybrids that we designed are stable during molecular dynamics simulations.

Binding energy of the Cas9/sgRNA hybrids closely match that of wild type complexes

To determine which hybrid to test in cellular systems, we analyzed the binding energy of every SaSp hybrid pair. The analysis of the SaCas9 complex could readily identify nearly all residues previously shown to interact with the sgRNA in the crystal structure[27], as well as additional ones (Fig. 4A and Supplementary Fig. 4). This added confidence that the residues with the smallest binding energy were indeed structurally important. Comparing the binding energy in the residues interacting with the sgRNA, we found that some SaSp sgRNAs, especially, Merry and Gollum, showed very similar binding energy compared to the SaCas9 and its sgRNA (Fig. 3A and Supplementary Fig. 4A). By contrast, Frodo showed more extreme values (Fig. 4B). Similarly, the CjSp complexes Rosy and Bilbo showed similar values as the CjCas9 complex, but Rosy had one amino acid that interacted with the sgRNA in CjCas9, Arg-63, which lost its binding affinity (Supplementary Fig. 5B). We conclude from these binding energy calculations that the best designs, as defined as those that most closely resemble the structural characteristics of SaCas9, SpCas9, and CjCas9, were Merry and Gollum for the SaSpCas9 hybrid and Bilbo for the CjSpCas9 hybrid. We also tested Frodo in cells as an example of a less optimal design.

The Cas9 hybrids failed to edit DNA in human cells

Next we wondered whether the designed Cas9 hybrids could edit DNA in cells. We cloned our SaSpCas9 together with either Merry, Gollum, or Frodo. We transfected them into GFP(CAG)$_0$ cells and showed that they all expressed to comparable levels (Supplementary Fig. 6A,B). Then, we cloned each sgRNA hybrid with four different target sequences against the GFP gene (Fig. 5A,B and Supplementary table 4). This is convenient because gene editing produces insertions and deletions that inactivate GFP. We could then readily detect the drop in GFP

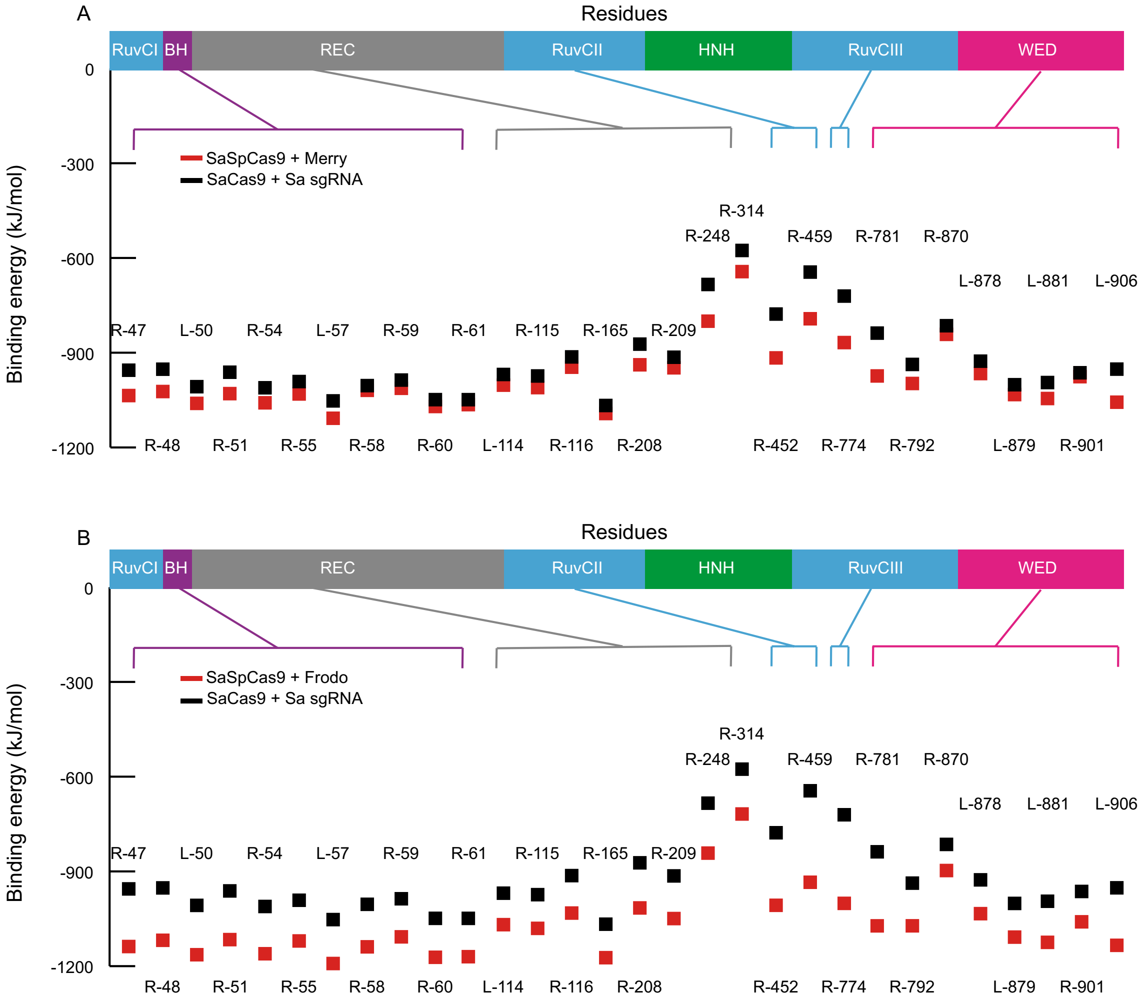


Figure 4. Binding energy calculations for SaSp hybrids. (**A**-**B**) Binding energy for each amino acid known to interact with the sgRNA for SaCas9 with its sgRNA (black) versus SaSpCas9 with Merry (**A**) or Frodo (**B**) (red). *BH* bridge helix, *WED* wedge domain, *PID* PAM interacting domain.

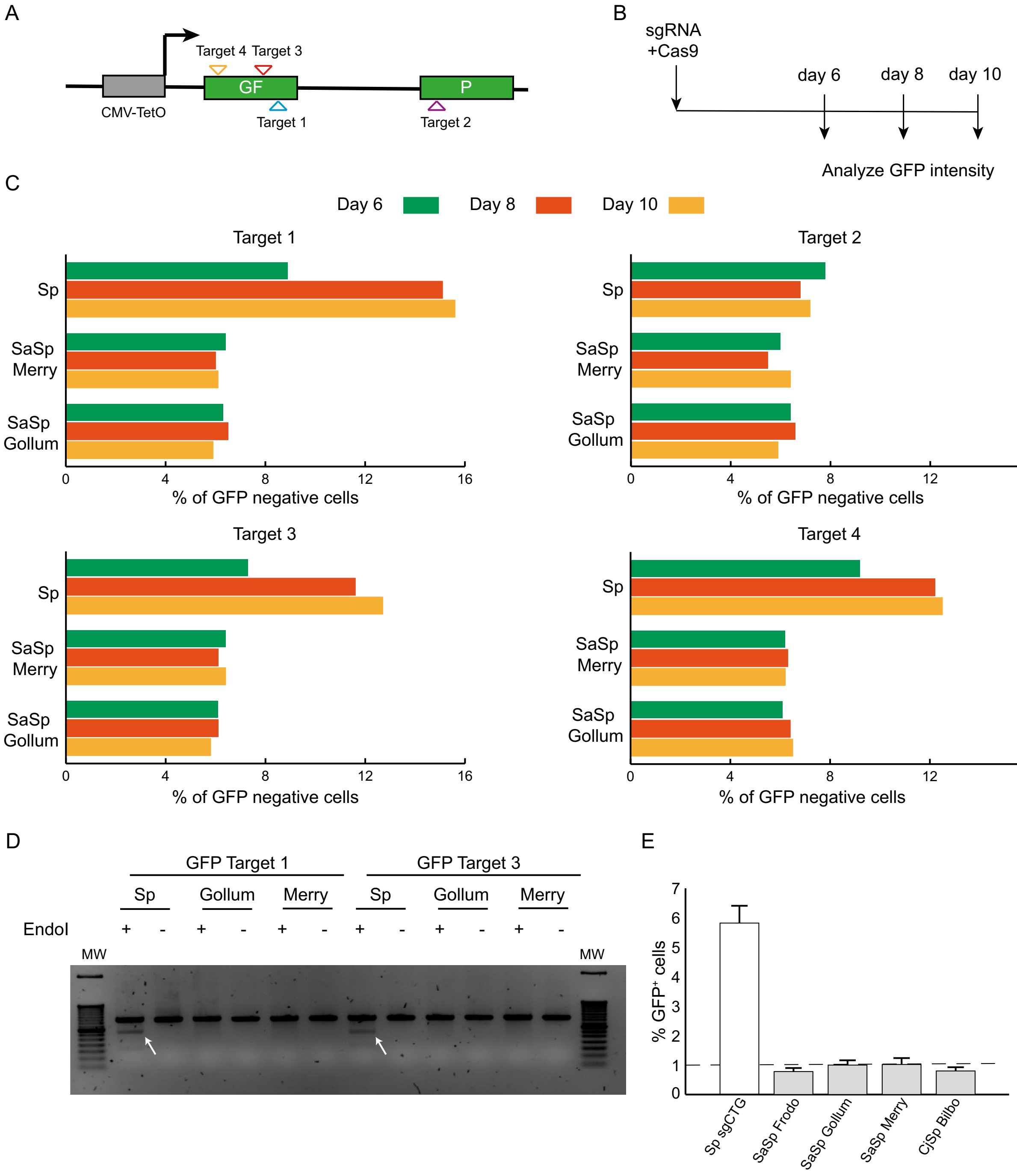


Figure 5. The Cas9 hybrids do not edit DNA in human cells. (**A**) Schematic representation of the GFP gene construction used for the editing assay using Cas9 nucleases. Each position targeted by the sgRNA used are shown with different triangles. This representation is not to scale. CMV-TetON is a doxycycline-inducible promoter. (**B**) Timeline of experiment used for the GFP assay. (**C**) Percentage of GFP negative cells for the 4 targets used at day 6 (green), 8 (orange) and 10 (yellow) for Sp, SaSp Merry and SaSp Gollum. (**D**) T7 endo I assay results for targets 1 and 3. The white arrows indicate the cleavage products indicative of efficient editing. Uncropped gels are found in Supplementary Fig. 7. (**E**) Percentage of GFP positive cells for each condition. Dashed lines represent the gate containing the brightest 1% of the cells transfected with Cas9 only.

intensity using flow cytometry. Using this approach, we confirmed that the SpCas9 was active, leading to a loss of GFP intensity of up to 15% for three of the four target sequences (Fig. 5C). By contrast, we found that the SaSp-Cas9/sgRNA hybrid pairs changed GFP expression very little, if at all, suggesting that they were not functional.

The GFP assay is an indirect measure of gene editing. To complement it, we used a second assay, the T7 endonuclease I assay, which detects edits at the DNA level. The insertions and deletions, once amplified and hybridized to unedited strands, form mismatches. These mismatches are then cut by the T7 endonuclease I, producing smaller DNA fragments that can be resolved on an agarose gel. We extracted DNA from the GFP assay at day 10 and performed the T7 endonuclease I assay for targets 1 and 3 as they showed the best results in editing the GFP gene. The results confirmed that whereas mismatches were generated upon SpCas9 transfection, the SaSpCas9 with Merry or Gollum induced no detectable edits (Fig. 5D), suggesting they are not functional.

To test their ability to contract CAG/CTG repeats, we transfected SaSpCas9 D10A together with Frodo, Merry, or Gollum, as well as CjSpCas9 D8A with Bilbo into GFP(CAG)$_{101}$ and measured GFP levels as before[29]. The SpCas9 enzyme increased GFP intensities, indicating contraction of the CAG repeat, whereas the Cas9 hybrids did not (Fig. 5E). Together, these results show that the SaSp and CjSp hybrid pairs are not functional in human cells.

Discussion

Very few RNA-guided nucleases can recognize repeats of CAG or CTG as a PAM sequence. This limits our ability to use the Cas9 nickase approach for contracting expanded CAG/CTG repeats in the clinic. Here we tested two CRISPR enzymes that were shown to recognize CAG or CTG PAM/TAM but failed to find evidence of contractions. This may be due to their lower editing efficiencies compared to SpCas9 or to the repeat sequences themselves, which may behave differently from non-repetitive sequences, by folding into secondary structures, for example.

We also tested the hypothesis that we could use molecular dynamics simulation and binding energy data to design in silico Cas9 hybrids that retained the PAM properties of SpCas9, but were small enough to fit within a single AAV. In addition, we found that the binding energy of the hybrids closely mirrored that of wild type Cas9 enzymes, thereby predicting edits in cells. This was not the case. Most other variants of SpCas9 and SaCas9 in the literature differ from the wild type by just a few amino acids and were designed to decrease the frequency of binding off-targets or to change their PAM recognition[11,14–21]. Similarly, four chimeric Cas9 enzymes have been created that change PAM sequence requirements[31,39,40]. At first glance these studies appear similar to what we have done here in that the PID of one orthologue was fused to the N-terminal of another. However, the fusions from those previous studies were between SpCas9 N-terminal and the PID of either *Streptococcus macacae* or *Streptococcus thermophilus* Cas9 or between the SlugCas9 PID and the N-terminal of SaCas9. All these fusions were done between closely related orthologues with interchangeable sgRNAs. Importantly, these hybrids do not reduce the overall size of the resulting proteins, which was our primary goal. Our results suggest that although powerful enough to predict the functional impact of one or a few residues, molecular dynamics simulation and binding energy values are not sufficient to predict function when the changes to the systems are on a larger scale.

It might be possible to further optimize the hybrid pairs by changing specific amino acids while minimizing the RMSD and RMSF. This may help improve the Cas9/sgRNA interactions, stabilize the hybrids further, and lead to a functional outcome. But this approach is predicated on the hypothesis that the parameters are predictive of function, which we have no evidence for here. It is also possible that the sgRNA hybrids were not expressed well enough to lead to robust gene editing.

Alternatively, the approach was perhaps successful, but the PAM requirements from the SpCas9 were not retained. Indeed, some SpCas9 with PAM variants are due to changes lying outside the PID[11,19]. If this was the case, then the target sequences we used here would not be recognized by our hybrids, explaining the lack of edits.

Further work will be required to make small Cas9 variants suitable for in vivo work that bind and edit CAG/CTG repeats. Improving the structure–function prediction of Cas9 variants would benefit a wide range of applications.

Experimental procedures

In silico models

The crystal structures required to run molecular dynamics were obtained from https://www.rcsb.org. PDB ID:4OO8 was used for SpCas9 in complex with its sgRNA and target DNA, this structure is at a 2.5Å resolution[5]. PDB ID:5CZZ was used for SaCas9 in complex with its sgRNA and target DNA this structure is at a 2.6Å resolution[27]. We note that the SaCas9 structure lacked stem loop 2 of the sgRNA. PDB:5X2G was used for the CjCas9 in complex with its sgRNA and a target DNA which was available at a 2.4Å resolution[41]. In this structure, the CjCas9 protein lacked the HNH domain, which was replaced with a GGGSGG linker. The SaSpCas9 hybrid is the hybrid between the N-terminal of SaCas9 (1-910AA) and the PID of SpCas9 (1100-1368AA). Similarly, the CjSp Cas9 hybrid is a fusion between the N-terminal of the CjCas9 (1-828AA) and the same Sp PID. To make the nickase mutations, we used the D10A mutation for the SaSpCas9 and the D8A for the CjSpCas9 (Supplementary Table 1), and created the change using the protein editing function in Molecular Operating Environment (MOE) 2019 (https://www.chemcomp.com/en/Products.htm). Cas9 hybrids and sgRNA hybrids (Supplementary Table 1 and 2) were made using MOE 2019 and PyMOL v2.5.4 (https://www.pymol.org/). An energy minimization was run on MOE to avoid any clashes between the added PID and rest of the protein prior to molecular dynamics simulations. We also designed the SaSp protein using ColabFold v1.5.3 (https://github.com/sokrypton/ColabFold). Alphafold2_multimer_v3 was used for complex prediction. A total of 3 recycles count (default) were used to predict the model. A total of 200 max iterations were used with a greedy pairing strategy. This approach, however, did not yield a useable model. Alphafold2 is currently unable to co-model with DNA or RNA structures and as such the model produced contained segments of protein which were positioned in such a way that the DNA/RNA would not be able to fit in the protein anymore. Further to this, whereas Alphafold2 is good at predicting proteins based on an evolutionary model, our hybrids are synthetic structures and not a natural protein. We chose instead to use the MOE software to design our structure being careful to preserve important structural features of both protein and DNA/RNA.

Molecular dynamics simulation

All the preparation steps and analysis for molecular dynamics simulation were carried out using GROMACS version 2018.2[42] (https://www.gromacs.org/). The forcefield AMBER03 was employed to create a stable environment for the system[43]. The protein-RNA–DNA complex was set in a 0.9 nm box filled with TIP3 water molecules. The system was neutralized either by adding Na + or Cl- ions depending on the overall charge. An integration time step of 0.001 ps was used. The complex was energy minimized to find the best position for the protein to avoid any clashes between the different structures. The energy minimization was run for a maximum of 3000 steps and a maximum energy barrier of 1000 kJ/mol/nm. During the molecular dynamics simulation, the v-rescale coupling method regulated the temperature to 310 K. A Berendsen barostat was coupled to the system to regulate the pressure with a reference of 1 bar. The whole simulation was composed of 5 different groups: protein, ions, solution, DNA, and RNA. The molecular dynamics simulation was run for a total of 100 ns.

Stability data analysis

The stability of every complex was studied by analyzing the Root Mean Square Deviation (RMSD) and the Root Mean Square Fluctuation (RMSF). Each of these are modules built-in GROMACS. The RMSD was extracted for every structure of the complex (protein-RNA-DNA) using the rms command and given every 10 ps. An average of the RMSD was made only including data from 10 to 100 ns as the first nanoseconds serve to warm up the system. The RMSF was extracted for every residue of every structure present in the complex using the rmsf command.

Binding energy

The Molecular Mechanics Poisson-Boltzmann surface area (MMPBSA) method was used to precisely determine the interaction between each base of the sgRNA hybrid and the protein hybrid. Recently, a GROMACS integrated tool was developed to perform these calculations[41]. The files and instructions needed to run it were found on: https://rashmikumari.github.io/g_mmpbsa/. This method combines calculations of potential energy in a vacuum, polar solvation energy, and non-polar solvation energy. The binding energy with g_MMPBSA, even if less computationally hungry than other methods, remains time-consuming. To reduce the computational burden and make the analysis possible, the last 100 frames of the molecular dynamics simulations were extracted and used to calculate the binding energy between the sgRNA and the protein.

Cell lines, plasmids, and culture conditions

HEK293-derived GFP$(CAG)_0$ and GFP$(CAG)_{101}$ cells were a gift from John H. Wilson[36]. They were maintained at 37 °C with 5% CO_2 in DMEM supplemented with 10% FBS, 100 U ml^{-1} penicillin, 100 μg ml^{-1} streptomycin, 15 μg ml^{-1} blasticidin, 150 μg ml^{-1} hygromycin, and, during the experiments, 1 μg ml^{-1} of doxycycline diluted in water. Both lines were regularly tested for mycoplasma using a service from Eurofins. The cells remained mycoplasma-free throughout the experiments. They were also genotyped using the Mycrosynth AG service and were determined to be HEK293.2sus as previously determined[44].

Transfections were done as before[45] by seeding 400,000 cells in a 12-well plate well on day 0 and transfecting 1 μg of plasmid using lipofectamine 2000. GFP intensities were measured using an Attune NxT flow cytometer and analyzed using the FlowJo software version 10.8.1 (https://www.flowjo.com/). The plasmids used in this study (Supplementary table 3) are available via Addgene (https://www.addgene.org/browse/article/28243489/).

For the GFP editing assays, we used GFP$(CAG)_0$ cells along with 4 well-characterized sgRNAs against GFP (Supplementary Table 4). The GFP levels in the cells were measured at day 6, 8, and 10 post-transfection. The flow cytometry data were analyzed using FlowJo v.10.8.1

The results were confirmed using a T7 endonuclease I assay using GFP$(CAG)_0$ cells transfected with GFP sgRNAs target 1 or 3 (Supplementary Table 4). The targeted regions were amplified using the primers oVIN-3474 and oVIN-3475 for sgRNA 1, 3 and 4 (Supplementary Table 5). The target region for sgRNA 2 was amplified using oVIN-3476 and oVIN-3477 (Supplementary Table 5). The PCR protocol used was 95 °C for 5min, followed by 35 cycles at 95 °C for 30", 55 °C for 30" and 72 °C for 1 min followed by a final step at 72 °C for 10 min. The annealing of PCR products was performed by mixing 200 ng of DNA in 1 × NEBuffer2. After annealing, 10 U of T7 endonuclease I (NEB) was added and incubated for 15 min at 37 °C. For repeat instability assays using the GFP reporter, GFP$(CAG)_{101}$ cells were cultured as above, but with dialyzed FBS (Merck). The cells were transfected on days 0, 4, and 8 before being analyzed for GFP intensity by flow cytometry on day 12. The resulting datasets were analyzed using FlowJo v.10.8.1 The plasmids used for transfection can be found in Supplementary Table 2.

Protein quantification and western blot

Proteins were extracted using commercial RIPA buffer (Fisher Scientific) and quantified using Pierce BCA protein assay (ThermoFisher Scientific). The proteins were separated on a 4–12% Bis tris gel (ThermoFisher Scientific) and transferred onto a nitrocellulose membrane (Bio-Rad). The membrane was blocked for 1 h in 5% milk in PBS-T then incubated overnight with an anti-flag antibody (Sigma) or actin antibody (Sigma) at 4°C. The secondary blotting was performed after 1 h incubation in an Alexa Fluor 680 anti-mouse antibody (Invitrogen). The blot was imaged using Licor Odyssey CLX.

Statistics

Two-tailed Mann–Whitney U tests were performed to determine statistical significance between the RMSD values of the hybrids and their relevant wild type Cas9 orthologues and between the two CjSpCas9. We used a Kruskal–Wallis test to determine statistical significance between the RMSD values of the SaSp hybrids. We used Graphpad Prism (version 10.0.0) to calculate the p-values.

Data availability

The datasets generated and analysed during the current study are available in the Zenodo repository, DOI: 10.5281/zenodo.10029983 (https://zenodo.org/records/10029983).

Received: 11 March 2024; Accepted: 19 July 2024
Published online: 26 July 2024

References

1. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).
2. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
3. Jinek, M. *et al.* A programmable Dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
4. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
5. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
6. Wang, J. Y. & Doudna, J. A. CRISPR technology: A decade of genome editing is only the beginning. *Science* **379**, eadd8643 (2023).
7. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).
8. Sharma, G., Sharma, A. R., Bhattacharya, M., Lee, S.-S. & Chakraborty, C. CRISPR-Cas9: A preclinical and clinical perspective for the treatment of human diseases. *Mol. Ther.* **29**, 571–586 (2021).
9. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
10. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
11. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290–296 (2020).
12. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
13. Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target effects in CRISPR/Cas9-mediated genome engineering. *Mol. Ther. Nucl. Acids* **4**, e264 (2015).
14. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
15. Kulcsár, P. I. *et al.* Blackjack mutations improve the on-target activities of increased fidelity variants of SpCas9 with 5′G-extended sgRNAs. *Nat. Commun.* **11**, 1223 (2020).
16. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
17. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
18. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).
19. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).
20. Lee, J. K. *et al.* Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat. Commun.* **9**, 3048 (2018).
21. Casini, A. *et al.* A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).
22. Luan, B., Xu, G., Feng, M., Cong, L. & Zhou, R. Combined computational-experimental approach to explore the molecular mechanism of SaCas9 with a broadened DNA targeting range. *J. Am. Chem. Soc.* **141**, 6545–6552 (2019).
23. Zhu, D., Schieferecke, A. J., Lopez, P. A. & Schaffer, D. V. Adeno-associated virus vector for central nervous system gene therapy. *Trends Mol. Med.* **27**, 524–537 (2021).
24. Ran, F. A. *et al.* In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
25. Tan, Y. *et al.* Rationally engineered *Staphylococcus aureus* Cas9 nucleases with high genome-wide specificity. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 20969–20976 (2019).
26. Kim, E. *et al.* In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat. Commun.* **8**, 14500 (2017).
27. Nishimasu, H. *et al.* Crystal structure of *Staphylococcus aureus* Cas9. *Cell* **162**, 1113–1126 (2015).
28. Murillo, A. *et al.* Cas9 nickase-mediated contraction of CAG/CTG repeats at multiple disease loci. *bioRxiv* https://doi.org/10.1101/2024.02.19.580669 (2024).
29. Cinesi, C., Aeschbach, L., Yang, B. & Dion, V. Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. *Nat. Commun.* **7**, 13272 (2016).
30. Merienne, N. *et al.* The self-inactivating KamiCas9 system for the editing of CNS disease genes. *Cell Rep.* **20**, 2980–2991 (2017).
31. Hu, Z. *et al.* Discovery and engineering of small SlugCas9 with broad targeting range and high specificity and activity. *Nucl. Acids Res.* **49**, 4008–4019 (2021).
32. Altae-Tran, H. *et al.* The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).
33. Fonfara, I. *et al.* Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucl. Acids Res.* **42**, 2577–2590 (2014).
34. Santillan, B. A., Moye, C., Mittelman, D. & Wilson, J. H. GFP-based fluorescence assay for CAG repeat instability in cultured human cells. *PLoS One* **9**, e113952 (2014).
35. Aeschbach, L. & Dion, V. Minimizing carry-over PCR contamination in expanded CAG/CTG repeat instability applications. *Sci. Rep.* **7**, 18026 (2017).
36. Kumari, R., Kumar, R., Lynn, A., Open Source Drug Discovery Consortium. *g_mmpbsa* —A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.* **54**, 1951–1962 (2014).
37. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Exp. Opin. Drug Discov.* **10**, 449–461 (2015).
38. Miller, B. R. *et al.* MMPBSA.py An efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* **8**, 3314–3321 (2012).
39. Chatterjee, P. *et al.* A Cas9 with PAM recognition for adenine dinucleotides. *Nat. Commun.* **11**, 2474 (2020).
40. Zhao, L. *et al.* PAM-flexible genome editing with an engineered chimeric Cas9. *Nat. Commun.* **14**, 6175 (2023).
41. Yamada, M. *et al.* Crystal structure of the minimal cas9 from campylobacter jejuni reveals the molecular diversity in the crispr-cas9 systems. *Mol. Cell* **65**, 1109-1121.e3 (2017).
42. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
43. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).

44. Yang, B. *et al.* Expanded CAG/CTG repeats resist gene silencing mediated by targeted epigenome editing. *Hum. Mol. Genet.* **31**, 386–398 (2022).

45. Cinesi, C., Yang, B. & Dion, V. GFP Reporters to Monitor Instability and Expression of Expanded CAG/CTG Repeats. In *Trinucleotide Repeats. Methods in Molecular Biology*, (eds Richard, G. F.) vol 2056. (Humana, New York, NY, 2020). https://doi.org/10.1007/978-1-4939-9784-8_16

Acknowledgements

We thank members of the Dion group for critical reading of the manuscript. This research was undertaken using the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff (ARCCA) on behalf of the Cardiff Supercomputing Facility and the HPC Wales and Supercomputing Wales (SCW) projects. We acknowledge the support of the latter, which is part-funded by the European Regional Development Fund (ERDF) via the Welsh Government.

Author contributions

AM performed all the experiments and designed them with the help of VD (for cell work) and GM (for in silico work). AM generated all figures and the manuscript was written in collaboration between all three authors.

Funding

This work is supported by the UK Dementia Research Institute [award numbers DRI-TAP2021-01 and UK DRI-3006] through UK DRI Ltd, principally funded by the UK Medical Research Council. VD is further supported by a professorship from the Academy of Medical Sciences (AMSPR1\1014) and the Moondance Foundation lab.

Competing interests

V.D. and G.M. have had a research contract with Pfizer Inc. within the last year on an unrelated project. A.M. declares no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-68107-1.

Correspondence and requests for materials should be addressed to V.D. or G.M.

Reprints and permissions information is available at www.nature.com/reprints.