nature structural & molecular biology

ARTICLES

https://doi.org/10.1038/s41594-019-0188-z

DNA stretching induces Cas9 off-target activity

Matthew D. Newton[1,2], Benjamin J. Taylor[3], Rosalie P. C. Driessen[4], Leonie Roos[5,6], Nevena Cvetesic[5,6], Shenaz Allyjaun[1,2], Boris Lenhard[5,6,7], Maria Emanuela Cuomo[3]* and David S. Rueda[1,2]*

CRISPR/Cas9 is a powerful genome-editing tool, but spurious off-target edits present a barrier to therapeutic applications. To understand how CRISPR/Cas9 discriminates between on-targets and off-targets, we have developed a single-molecule assay combining optical tweezers with fluorescence to monitor binding to λ-DNA. At low forces, the *Streptococcus pyogenes* Cas9 complex binds and cleaves DNA specifically. At higher forces, numerous off-target binding events appear repeatedly at the same off-target sites in a guide-RNA-sequence-dependent manner, driven by the mechanical distortion of the DNA. Using single-molecule Förster resonance energy transfer (smFRET) and cleavage assays, we show that DNA bubbles induce off-target binding and cleavage at these sites, even with ten mismatches, as well as at previously identified in vivo off-targets. We propose that duplex DNA destabilization during cellular processes (for example, transcription, replication, etc.) can expose these cryptic off-target sites to Cas9 activity, highlighting the need for improved off-target prediction algorithms.

Clustered regularly interspaced short palindromic repeat (CRISPR) RNAs and their associated proteins (Cas) are RNA-directed programmable endonuclease complexes that can cleave a specific DNA target in a programmable fashion. Originally discovered in bacteria and archaea, where they act as an acquired immunity mechanism against invasive viral DNA and RNA, the application of CRISPR/Cas to biological research has revolutionized genomic engineering, and offers the potential for novel therapeutic applications[1–5]. The *Streptococcus pyogenes*-derived Cas9 protein (SpCas9), has been instrumental in demonstrating the potential for genetic modification using CRISPR/Cas systems. Cas9 is an endonuclease that binds two RNAs to form an active holoenzyme (Fig. 1a). The trans-activating CRISPR RNA (tracrRNA) plays a structural role, whereas the CRISPR RNA (crRNA) contains a 20 nucleotide targeting sequence that directs the Cas9 complex to complementary double-stranded (ds) DNA targets. In addition, a protospacer adjacent motif (PAM, sequence 5′-NRG-3′ for SpCas9) is required immediately preceding the target[6]. Once bound at a target site, two endonuclease domains, HNH and RuvC, cleave the DNA to generate a double-strand break (DSB)[6].

Cas9 technology has been successfully utilized in a variety of in vivo applications to generate sequence deletions, insertions and to regulate gene expression by coupling to activator or repressor accessory proteins[1–3,7–9]. However, spurious off-target DSBs represent an important barrier that hinders its use for therapeutic applications[10–18].

The mechanism by which the Cas9 complex binds and cleaves targets, and off-targets with a small number of mismatches, has been extensively studied by a combination of in vitro bulk assays, crystallography, computational and single-molecule approaches[16,19–28]. The initial Cas9–DNA interaction is mediated by recognition and binding of the PAM site, which in turn facilitates PAM-proximal DNA melting, allowing directional probing of complementarity between the potential target site and the crRNA to form a stable R-loop[19,21,26]. Binding and cleavage exhibit differential requirements of complementarity across the 20 base pair (bp) target. High-stability binding requires at least 7–9 matched PAM-proximal bases; in contrast, as few as 4 mismatches in the PAM-distal end have been shown to hinder cleavage but not binding[22,29]. On the basis of these and other studies, algorithms have been developed to predict off-target activity and high specificity Cas9 proteins have been produced, with the aim of minimizing Cas9 promiscuity[20,24,29–38]. However, sequencing-based approaches have experimentally validated numerous Cas9 off-targets that remain unexplained by such algorithms[13–15], raising the interesting question of how the Cas9 complexes bind and cleave off-targets.

To address this question, we have used a combination of optical tweezers with confocal microscopy and microfluidics to monitor fluorescently labeled Cas9 complex binding in real time on force-stretched λ-DNA with single-molecule resolution (Fig. 1a). Our data show that mechanically distorting the DNA induces stable off-target binding at sequence-specific sites across the DNA. We validate three of these off-target binding sites, which contain as many as 10 mismatches, by smFRET binding and bulk cleavage assays. Together these experiments demonstrate that partially unwound or melted DNA helps recruit Cas9 to bind and cleave previously unseen cryptic off-target sites. We further test our model using four bona fide in vivo off-targets and demonstrate that these sites are only cleaved when the dsDNA structure is disrupted. In cells, such cryptic sites could be made accessible by processes such as transcription, replication, DNA repair or (local) DNA supercoiling.

Results

SpCas9 binds low force-stretched λ-DNA tightly and specifically. To monitor Cas9–DNA interactions in real time with single-molecule resolution, we combined optical tweezers with confocal fluorescence microscopy and microfluidics (Fig. 1a)[39–41]. We assembled Cas9 complexes with a 5′-Cy3-labeled crRNA, unlabeled tracrRNA

[1]Molecular Virology, Department of Medicine, Imperial College London, London, UK. [2]Single Molecule Imaging Group, MRC-London Institute of Medical Sciences, London, UK. [3]Discovery Sciences, IMED Biotech Unit, Astra Zeneca, Cambridge, UK. [4]LUMICKS BV, Amsterdam, The Netherlands. [5]Computational Regulatory Genomics, MRC-London Institute of Medical Sciences, London, UK. [6]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK. [7]Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway.
*e-mail: emanuela.cuomo@astrazeneca.com; david.rueda@imperial.ac.uk

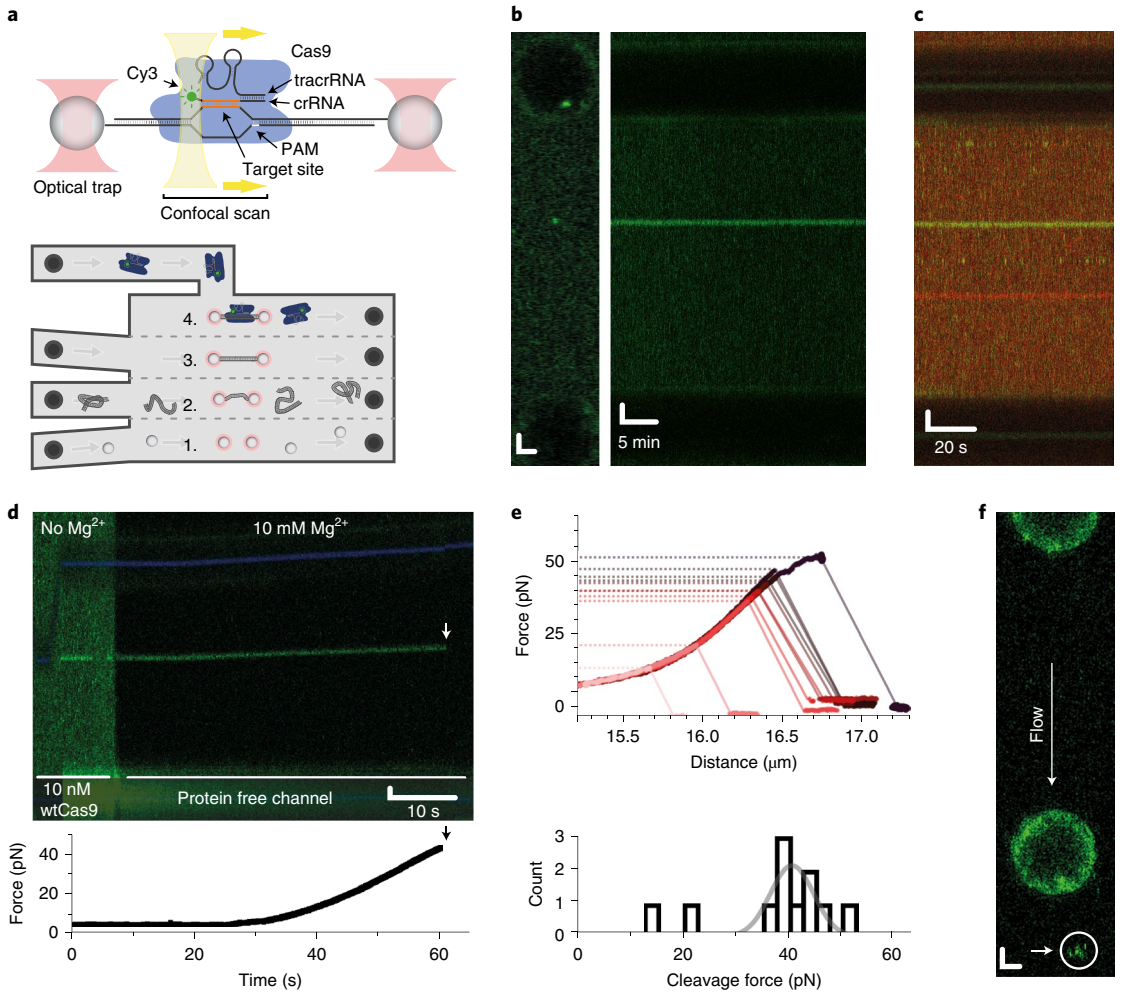

Fig. 1 | Cas9 binds and cleaves DNA tightly and specifically at low force. a, Diagram of force-stretched λ-DNA. DNA tethered between trapped beads, Cas9 complex is labeled using 5′-Cy3 on λ2-crRNA. Microfluidic set-up: 1. bead channel, 2. DNA channel, 3. buffer-only channel, 4. imaging channel. **b**, Confocal image of λ-DNA at 5 pN with 10 nM λ2-crRNA-dCas9 at expected binding site, and kymograph dCas9 bound >30 min. **c**, λ-DNA at 5 pN with 20 nM dCas9 with 5′-Cy3-labeled λ2-crRNA (green) or 5′-Cy5-labeled λ4-crRNA (red). **d**, Kymograph and force-ramp of λ-DNA with wtCas9 in the absence and presence of Mg^{2+}. Cleavage is observed at 60 s (43 pN) as loss of fluorescence (white arrow) and force (black arrow). **e**, Force–extension curves showing DNA cleavage and distribution of corresponding rupture forces ($n = 11$). **f**, Confocal image of post-cleavage flow-stretched λ-DNA with wtCas9 bound to DNA on lower bead. Scale bars 1 μm.

and catalytically inactive dCas9 (Fig. 1a). We chose a previously characterized crRNA sequence that specifically binds a unique target site (λ2, at 18 kilobases (kb)) on λ-DNA (Supplementary Table 1)[19]. Phage λ-DNA was linked to two 4.5 μm polystyrene beads through biotin–streptavidin interactions and stretched under 5 pN force (Fig. 1a), sufficient to stretch the DNA to its contour length (~16 μm) but without altering base pairing of the two strands[42]. In the presence of labeled λ2-crRNA:tracrRNA:dCas9 complex, we observed a single binding event at the expected location (Fig. 1b), which remained site-specifically bound for >30 min ($n = 5$), probably only limited by photobleaching of the fluorophore. This is consistent with previous observations that Cas9 complex binds target DNA with very high affinity[19]. No binding was observed with labeled crRNA alone, crRNA:tracrRNA or crRNA–dCas9 complex, confirming that these events represent stable binding events of the holoenzyme complex (Supplementary Fig. 1a–c). Binding was also not observed with a 5′-Cy3-crRNA lacking a target site (nt-crRNA) (Supplementary Fig. 1d and Supplementary Table 1). Then, we tested if we could monitor two dCas9 complexes targeting two sites using a second 5′-Cy5-labeled crRNA targeting a different site (λ4, at 30.5 kb) (Supplementary Table 1)[19]. In the presence of both dCas9 complexes, we also observed clear specific binding to their respective targets, as expected (Fig. 1c). At 5 pN, none of the crRNAs exhibited any off-target binding within the experimental time resolution (~250 ms). These data confirm that SpCas9 binds on-target sites on low force-stretched DNA with high affinity and specificity, and shows no off-target binding.

Specifically bound Cas9 cleaves DNA and remains bound across the DSB. Next, we tested whether the Cas9 complex was active on force-stretched λ-DNA. We assembled the complex with Cy3-labeled λ2-crRNA:tracrRNA and catalytically competent wild type Cas9 (wtCas9) in the presence of EDTA to prevent cleavage. Fluorescent labeling did not affect wtCas9 activity in control bulk assays (Supplementary Fig. 1e). The wtCas9 complex bound its specific site, as expected, and catalysis was activated by moving the DNA to a protein-free channel containing 10 mM Mg^{2+} (Fig. 1d). At low force (5 pN) cleavage was not observed within 10 s ($n = 11$). However, gradually increasing the force from 5 to 55 pN resulted in DNA rupture at a force of 40 ± 1 pN (mean ± s.e.m.), revealing

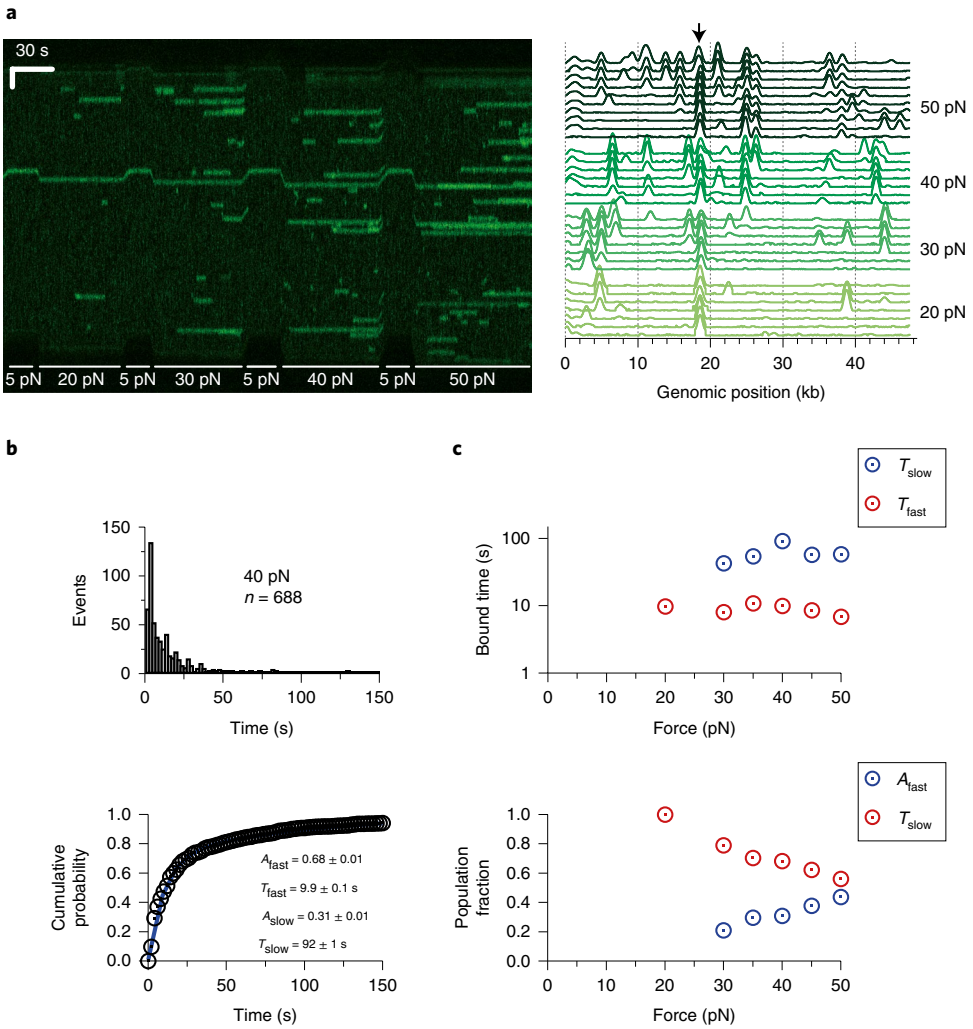

Fig. 2 | DNA stretching induces Cas9 off-target binding. a, Kymograph of force-stretched λ-DNA in the presence of 10 nM λ2-crRNA-dCas9 (green) between 5 and 50 pN, with off-target binding >20 pN. Time-binned intensity histogram (right) shows force dependence and genomic off-target binding locations (arrow marks on-target site). **b**, Dwell-time histogram of λ2-crRNA-dCas9 off-target bound times at 40 pN and integrated cumulative probabilities with double exponential fit (bottom) yield lifetimes (τ) and amplitudes (A). **c**, Force dependence of off-target lifetimes (top) and amplitudes. Data for 20 pN were fit to a single exponential. Source data for panels **b** and **c** are available online.

that Cas9 had cleaved the DNA but remained tightly bound to the product complex across the DSB ($n = 11$, Fig. 1e), consistent with previous observations[19,26,43]. While cleavage may occur rapidly after specific binding at low force, substantial mechanical force is required to disassemble the post-cleavage complex. It has been suggested that on dissociation Cas9 remains preferentially bound to the PAM-proximal side of the break[43]. In a few instances, where photobleaching had not occurred before cleavage, we were able to test this by flow-stretching the cleaved DNA, which revealed the tagged wtCas9 complex still bound to the cleaved strand (Fig. 1f). These results demonstrate that the Cas9 complex is active under our experimental conditions, confirm previous observations that the complex holds together cleaved DNA ends with high affinity, and also give the first estimates of the mechanical strength of the post-cleavage complex.

DNA stretching induces off-target binding. Next, we investigated how increasing the tension on the DNA affects Cas9 binding. Using the Cy3-labeled λ2-crRNA:tracrRNA:dCas9 complex, we obtained an on-target binding event at low force (5 pN) before increasing the force to 20 pN (Fig. 2a). At the higher force, we immediately started to see additional binding events at off-target locations (Fig. 2a, left, and Supplementary Video 1). Reducing the force to 5 pN results in dissociation of all off-target-bound complexes from the DNA, with only the on-target Cas9 complex remaining bound. To determine the off-target binding locations, we time-binned the kymographs, and mapped the inter-bead distance to the λ-DNA sequence (Fig. 2a, right, and Supplementary Fig. 1f). The resulting intensity profiles at 20 pN show the on-target peak at the expected site (Fig. 2a, right, arrow), and six off-target peaks distributed over the DNA. Increasing the force to 30, 40 and 50 pN results in more off-target binding events (Fig. 2a). To characterize the off-target kinetics, we obtained dwell times for hundreds of λ2-crRNA:tracrRNA:dCas9 off-target binding events across λ-DNA held at forces from 20–50 pN, and constructed dwell-time histograms (Fig. 2b and Supplementary Figs. 1f and 2). The resulting cumulative dwell-time histograms reveal a fast and a slow population of off-target dissociation events (Fig. 2b), except for 20 pN, which only exhibits the fast population (Supplementary Fig. 2). Interestingly, the lifetime of both populations remains approximately constant ($\tau_{fast} = 9.0 \pm 0.6$ s and $\tau_{slow} = 61 \pm 8$ s, mean ± s.e.m.) (Fig. 2c) but with a growing slow population as force increases (Fig. 2c). This suggests that Cas9 binds off-targets in two binding modes, a transient (short-lived) and a meta-stable (longer-lived) mode, the latter becoming dominant at

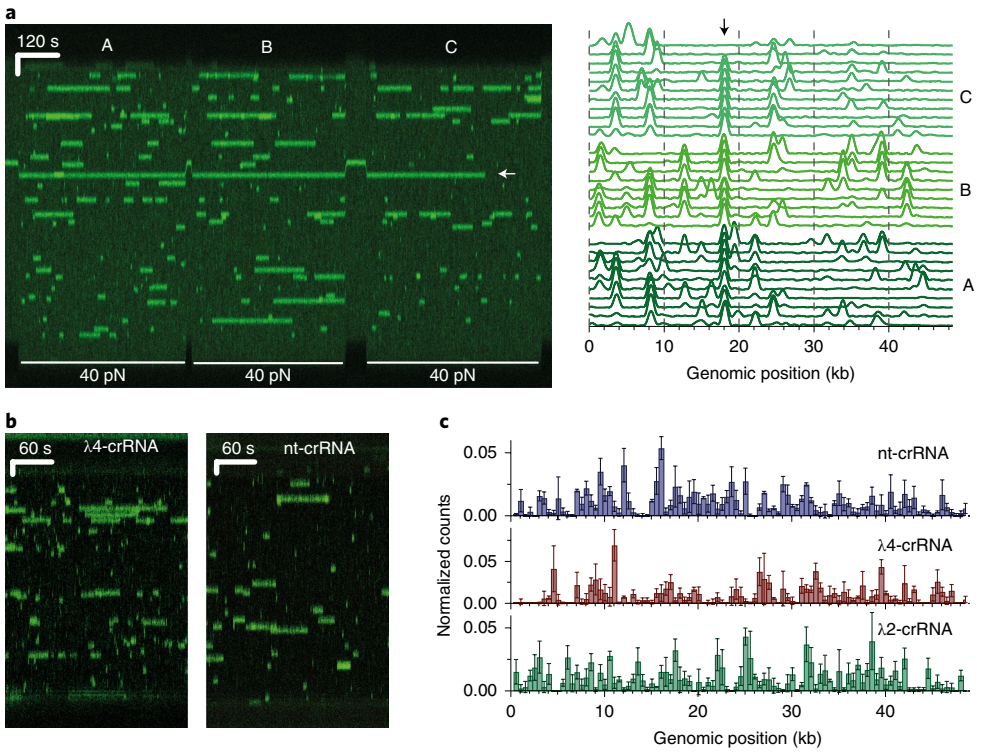

Fig. 3 | Off-target binding is crRNA sequence specific. a, Kymograph (left) of force-stretched λ-DNA in the presence of 10 nM λ2-crRNA-dCas9 (green) repeatedly stretched to 40 pN and relaxed to 5 pN. At 5 pN, only on-target remains bound. Time-binned intensity histogram (right) shows recurrent off-target binding locations (arrow marks on-target site). **b**, Kymograph of λ-DNA force-stretched to 40 pN in the presence of 1 nM λ4-crRNA- or 1 nM nt-crRNA-dCas9. **c**, Mean count, normalized histograms of mapped binding events for the three guides tested (n = 1,234, n = 797 and n = 1,950 for λ2, λ4 and nt (no target), respectively). Bin size = 500 nucleotides, error bars, ±s.d. Source data for **c** are available online.

higher forces. Together, these data show that stretching the DNA induces frequent off-target Cas9 binding in a force-dependent fashion at previously unobserved sites.

Off-target binding occurs at non-random, crRNA sequence-dependent sites. Interestingly, the kymographs and intensity profiles show many recurrent off-target peak locations at high force, raising the possibility that off-target binding is non-random (Fig. 2a). To further test this idea, we pulled a single DNA molecule three consecutive times at 40 pN (Fig. 3a). The resulting kymographs and intensity profiles again show the same recurrent binding locations at each pull. Surprisingly, none of the commonly used Cas9 off-target prediction tools (for example, Cas-OFFinder, CCTop or CRISPOR)[30,37,38] were able to predict the large number of off-target sites observed here, suggesting that these tools may miss many off-targets. Previous studies have suggested that λ-DNA off-target binding locations correlate with PAM site densities (Supplementary Fig. 3a), which are unevenly distributed due to the polar GC distribution across the genome (Supplementary Fig. 3a), rather than crRNA:DNA sequence complementarity[19]. In our experiments, however, the observed off-target binding locations appear to be highly specific, suggesting that intrinsic features of the λ-DNA sequence are determining binding locations and indicating that off-target binding may depend on DNA–crRNA complementarity.

To explore the off-target binding guide-sequence dependence, we compared the λ2-crRNA with λ4-crRNA and nt-crRNA. At 5 pN, no off-target binding is observed with either guide (Fig. 1c); however, increasing the force to 40 pN results in multiple off-target binding events across the whole DNA (Fig. 3b). These off-target binding locations are also non-random, occurring repeatedly at the same genomic locations (Supplementary Fig. 3b). Interestingly, comparing the off-target binding locations for all three guides reveals distinct binding distributions, with many sites unique to each guide (Fig. 3c and Supplementary Fig. 3c–f). These data further support the idea that off-target binding locations are determined, at least in part, by DNA–crRNA sequence complementarity.

Off-target binding requires both DNA strands. To determine whether off-target Cas9 binding is driven by the presence of force-induced nicks or single-strand (ss) DNA, we performed a competition experiment between nt-crRNA-dCas9 and the single-stranded binding human replication protein A (hRPA) labeled with eGFP[44]. Under our experimental conditions (10 nM), hRPA does not bind small (15–20 nucleotide) bubbles on DNA[45,46]. The DNA stretching force was ramped from 5–65 pN, generating increasingly large regions of ssDNA[42,47,48] (Supplementary Fig. 3g). At low to high forces (5–60 pN), hRPA-eGFP only binds a small number of DNA nicks and never competes with dCas9. At the highest forces (60–65 pN), the DNA overstretches, exposing large ssDNA regions that propagate from the nicks, where hRPA-eGFP binds readily (Supplementary Fig. 3g, blue patches), but never dCas9 (Supplementary Fig. 3g,h), showing that Cas9 off-target binding is not driven by nicks or ssDNA. These results imply that, for stable off-target binding, Cas9 requires both DNA strands in close proximity but not in B-form, raising the interesting possibility that off-target binding is driven by small disruptions in the DNA helical structure, such as transient melting bubbles or other forms of distorted DNA structures such as DNA breathing, unstacking, supercoiling, etc.

Off-target sequence identification. In an effort to determine the rules of sequence-dependent off-target binding, we first asked if sequences similar to the crRNA are enriched within regions

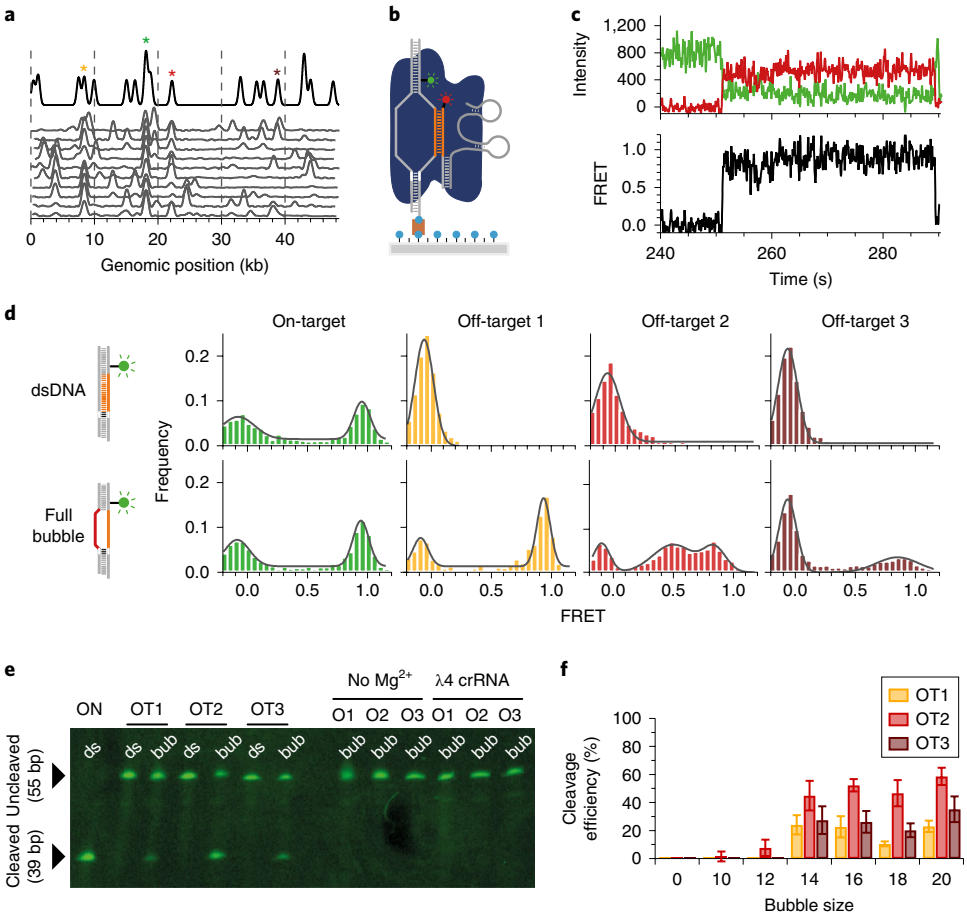

Fig. 4 | DNA bubbles induce off-target binding and cleavage. a, Experimental (gray) and calculated (black) time-binned intensity histograms for λ2-crRNA. Many predicted peaks align well with the data, including on-target and off-targets 1, 2 and 3 (green, 18 kb; orange, 8 kb; red, 22 kb; and brown, 39 kb, respectively). **b**, smFRET experimental design: surface-immobilized Cy3-DNA (green) containing target (orange) and PAM (white) sites is bound by Cy5-labeled (red) dCas9 (blue). **c**, Representative single-molecule trajectory with donor (green) and acceptor (red) intensities, and corresponding FRET (black). Binding results in a high FRET state for ~40 s before acceptor and donor photobleaching. **d**, FRET histograms of dsDNA and bubbled constructs with Gaussian fits (gray). **e**, Denaturing PAGE cleavage assay with double-stranded (ds) and bubbled (bub) λ2 off-target substrates. Bubbled substrates show cleavage, except in the absence of Mg^{2+} or with a non-related control λ4-crRNA guide. **f**, Relative cleavage efficiencies of off-target constructs with different bubble sizes. Quantification of the assay shown in **e**, $n = 3$ independent experiments, mean ±s.d. Uncropped gel images are shown in Supplementary Dataset 1. Source data for **f** are available online.

recurrently bound by the Cas9 complex. The λ-genome was scanned for each of the three 20-nucleotide crRNA sequences, initially without considering PAM requirement, at different matching thresholds. Apart from the on-target site for λ2 and λ4, additional sites were only identified when the match requirement was lowered to 60% sequence similarity to the crRNA. There was no substantial enrichment of 60% match sites within binding peaks compared with outside peaks. Furthermore, there were no location-specific nucleotide differences within the 20 nucleotide sequence, nor the expected PAM site location, which could distinguish recurrent Cas9-bound regions (Supplementary Fig. 4). De novo motif discovery on all peaks centered on the maximum intensity value did not reveal underlying motifs either. This analysis suggests that off-target binding to distorted DNA can tolerate too many sequence variations (relative to the on-target) to yield a specific sequence motif. Alternatively, peak location resolution (~500 bp) may impede the emergence of a sequence-specific motif.

In the absence of an emerging pattern in our de novo approach, we developed a simple thermodynamic off-target binding model that determines a binding score on the basis of sequence similarity, presence of PAM site and other factors that have previously been shown to influence Cas9 binding (Supplementary Note 1). A theoretical off-target binding histogram is calculated by thresholding and convoluting the resulting score matrix with an experimental Gaussian point spread function (Fig. 4a). The calculated intensity histogram predicts 18 potential binding locations, including the on-target site, and 17 off-targets (Supplementary Table 2). A visual comparison of the calculated and the experimental intensity histograms shows that at least 9 (50%) of the predicted peaks overlap with experimental ones, which enables us to identify potential candidates for off-target binding (Fig. 4a and Supplementary Table 2). We selected three of these off-target binding sites for further analysis (Fig. 4a and Supplementary Table 2).

To validate these candidates at zero force, we performed smFRET experiments with the correct target and the three off-targets from the thermodynamic analysis (Fig. 4a and Supplementary Tables 1–3). Using surface-immobilized, Cy3-labeled DNA and 5′-Cy5-λ2-crRNA:tracrRNA:dCas9 complex, we can detect binding as a FRET signal between the donor (Cy3) and the acceptor (Cy5), the magnitude of which depends on R-loop formation (Fig. 4b)[32]. With fully hybridized on-target DNA, about half of the molecules (46%, $n = 2{,}510$) reach a high FRET state (0.95 ± 0.02, mean ± s.d.) as

expected (Fig. 4c,d, green)[32]. To mimic a force-distorted DNA duplex, we mutate the non-target strand of the DNA to avoid complementarity with the target strand and create a bubble. The non-target strand sequence is replaced with polythymine, except where that would result in target-strand base pairing, in which case it is replaced with a cytosine (Supplementary Table 3). This full bubble, on-target construct exhibits similar binding levels (53%, $n = 2{,}986$) and FRET state (0.94 ± 0.02) as the fully dsDNA construct (Fig. 4b, green). This is consistent with the absence of base-specific interactions between the non-target strand and Cas9[25,49].

In agreement with the low force-stretched experiments (Fig. 1a), a dsDNA off-target construct (Off-Target 1, OT1; Fig. 4a and Supplementary Tables 2 and 3) completely abolishes Cas9 binding (Fig. 4d, yellow). Cas9 binding is rescued, however, with the corresponding full bubble construct (68%, $n = 893$, FRET $= 0.93 \pm 0.02$; Fig. 4d), supporting the idea that a melted bubble can recruit Cas9 binding to a cryptic, off-target site. Two additional dsDNA off-target sequences (OT2 and OT3; Fig. 4a and Supplementary Tables 2 and 3) confirm that Cas9 does not bind hybridized off-targets, but binding is rescued by the presence of a melted bubble (78%, $n = 1{,}069$, red, and 33%, $n = 2{,}062$, brown, respectively; Fig. 4d). No binding was observed between the full off-target bubble constructs and Cy5-λ2-crRNA:tracrRNA in the absence of protein (Supplementary Fig. 5a). Interestingly, the conformation of the bubble-bound complex varies depending on the off-target sequence (FRET $= 0.90 \pm 0.02$ for OT1, 0.50 ± 0.02 and 0.84 ± 0.02 for OT2, and 0.85 ± 0.02 for OT3; Fig. 4d and Supplementary Fig. 5b). The two conformations observed for OT2 are in dynamic equilibrium (Supplementary Fig. 5b). These differences probably arise from the location and number of mismatches (Supplementary Table 2) which can affect the structure of the R-loop[32]. Together these data validate some of the off-targets predicted by the thermodynamic model and show that pre-melting the target DNA can induce Cas9 binding to cryptic off-target sites.

DNA bubbles induce Cas9 off-target cleavage. We next tested whether these off-targets are active for cleavage. Labeled DNA constructs (Supplementary Table 3) were incubated with 5′-Cy5-λ2-crRNA:tracrRNA:wtCas9 and imaged by denaturing PAGE with fluorescence detection (Fig. 4e). As expected, double-stranded on-target DNA shows complete cleavage after 1 h at 37 °C (Fig. 4e, ON), whereas double-stranded off-target sequences did not cleave (Fig. 4e, OT1–OT3, ds). However, all three off-target bubble constructs showed substantial levels of cleavage (Fig. 4e; OT1–OT3, bub). No cleavage was observed in the absence of Mg^{2+} or with an unrelated control crRNA (Fig. 4e; no Mg^{2+}, λ4-crRNA). Cleavage efficiency of the full off-target bubbles ranges from 18–57% (Fig. 4e,f, "20"), but does not correlate with the number of mismatches, the PAM sequence (Supplementary Fig. 6a) or the fraction of molecules reaching the high FRET state (Fig. 4d), suggesting that even transient R-loop formation may be sufficient to result in Cas9 cleavage. Together, these results demonstrate that DNA bubbles induce Cas9 binding and cleavage at cryptic target sites containing as many as ten mismatches, indicating that Cas9 may be more promiscuous than previously thought.

To further characterize the effect of bubble size and location, we measured off-target cleavage with bubbles ranging from 0 to 20 nucleotides (Fig. 4f). The data show that only ≥14-nucleotide bubbles promote off-target cleavage (Fig. 4f and Supplementary Fig. 6b), and increasing the bubble size further does not increase cleavage efficiency. Interestingly, a small (6 nucleotide) PAM-proximal bubble is sufficient to promote off-target binding but not cleavage (Supplementary Figs. 5e and 6c). Small (6 nucleotide) bubbles in the middle or PAM-distal region of the DNA are not cleavage competent either (Supplementary Fig. 6c).

Finally, to determine whether the presence of the non-target strand or a double-stranded PAM are important for off-target binding, we measured binding and cleavage to a single-stranded DNA and a double-stranded PAM DNA (Supplementary Figs. 5c,d and 6d). The data show that the Cas9 complex does not bind or cleave single-stranded OT1, OT2 or OT3 (Supplementary Figs. 5c and 6d). A double-stranded PAM partially rescues binding and cleavage (Supplementary Figs. 5d and 6d). However, both the non-target strand and a double-stranded PAM sequence are necessary for robust off-target binding and cleavage (Fig. 4d–f).

Discussion

Owing to its ability to function as a straightforward, fully programmable gene-editing tool, CRISPR/Cas9 has transformed how cellular biologists edit and regulate genes in many cell types and in live organisms. It even holds great promise to function as a therapeutic tool against both genetic disorders and common diseases such as cancer[50,51]. However, its potential could be hindered by spurious off-target mutations that cannot be tolerated in clinical applications. It has, therefore, become essential to understand how Cas9 binds and cleaves off-targets. To this aim, we have developed an optical tweezers assay with fluorescence detection that enables us to monitor and manipulate Cas9 binding on single DNA molecules. The data show that Cas9 binds and cleaves low force-stretched DNA tightly and specifically at a unique target site (Fig. 1). However, distorting the DNA duplex structure with high stretching forces results in numerous off-target binding events at sites that depend on the guide sequence (Figs. 2 and 3). Indeed, three different crRNAs yield three distinct binding fingerprints (Fig. 3b,c), essentially equivalent to a real-time single-molecule chromatin-immunoprecipitation (ChIP) experiment in vitro. Competition experiments with hRPA show that off-target binding does not arise on DNA nicks or large ssDNA regions. Using a simple thermodynamic model, we identified at least nine of these off-targets, three of which we validated at zero force with smFRET and bulk cleavage assays. Interestingly, Cas9 only binds and cleaves these off-targets at zero force on bubbled substrates, confirming that it maintains target accuracy with fully paired substrates.

There are several factors that will probably influence off-target binding and cleavage. First, our data with three different crRNAs clearly show that the guide sequence plays an important role in binding location. Surprisingly, we find that bubbled substrates increase Cas9 cleavage tolerance to at least ten mismatches throughout the target sequence, including multiple mismatches in the seed region (Fig. 4e). Interestingly, cleavage efficiency is not proportional to the number of mismatches. For example, OT2 has the largest number of mismatches (ten) and the highest cleavage efficiency (Fig. 4e), even though the bound complex conformation is affected by the location and number of mismatches (as shown by the smFRET experiments; Fig. 4d). This implies that, in the context of a bubbled substrate, cleavage may be possible with an incomplete R-loop. Efficient Cas9 off-target binding and cleavage, however, requires both DNA strands and a double-stranded PAM. The number of tolerated mismatches, particularly in the seed region, is much greater than previously shown, and the lack of a consensus binding sequence emerging from the de novo motif search (Supplementary Fig. 4) further supports the idea of increased promiscuity with bubbled substrates. Second, the GC content of the target sequence is also probably a factor. GC-poor sequences are more prone to form bubbles, and therefore to facilitate Cas9 off-target binding and cleavage. In agreement with this idea, two of the guides (λ4- and nt-crRNA) exhibit more binding peaks in the AT-rich regions of λ-DNA (Supplementary Fig. 3b). Last, the density of PAM sequences will also probably play a role, but this could not be tested with λ-DNA because of its high density of PAM sites (Supplementary Fig. 3a). Future experiments with other DNA sequences and targets will help determine the

exact role of each factor and the rules governing off-target binding and cleavage.

It has been previously shown that the Cas9 target selection rate-limiting steps are duplex melting and R-loop formation[28,52]. Therefore, a probable mechanism for Cas9 decreased specificity is that force stretching lowers the DNA melting barrier, thereby facilitating R-loop formation, even in the presence of multiple mismatches. In agreement with this idea, recent experiments have shown that engineered Cas9 proteins achieve high specificity by hindering DNA unwinding and making it more sensitive to mismatches[31].

Cellular DNA is continually unwound during many fundamental processes, such as DNA transcription, replication, repair or chromatin remodeling. During these processes, transiently exposed 10–100-nucleotide-long bubbles may provide many opportunities to recruit Cas9 complexes at off-target sites, particularly under the high concentrations present in overexpression conditions. This idea is in agreement with a genome-wide Chip-seq analysis and GUIDE-seq data that show that off-targets are enriched in promoters, intron and exons (potentially transcribed regions)[14,53]. To further test this hypothesis, we selected four bona fide off-targets validated in vivo. Our cleavage data with these off-targets show that cleavage is only observed in the presence of bubbled substrates (Supplementary Fig. 6e,f), showing that Cas9 is unlikely to cut these off-targets unless the DNA structure is distorted, for example during transcription, DNA replication or DNA damage repair. This effect will be particularly relevant in actively replicating cell types (for example, liver cells or thymocytes), which are the desired target for many therapeutic applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41594-019-0188-z.

Received: 2 November 2018; Accepted: 18 January 2019;
Published online: 25 February 2019

References

1. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**, 347–355 (2014).
2. Liang, P. et al. CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein Cell* **6**, 363–372 (2015).
3. Maeder, M. L. & Gersbach, C. A. Genome-editing technologies for gene and cell therapy. *Mol. Ther.* **24**, 430–446 (2016).
4. Makarova, K. S. et al. Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).
5. Kim, K. et al. Highly efficient RNA-guided base editing in mouse embryos. *Nat. Biotechnol.* **35**, 435–437 (2017).
6. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
7. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
8. Gilbert, L. A. et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).
9. Schwank, G. et al. Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. *Cell Stem Cell* **13**, 653–658 (2013).
10. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).
11. Cameron, P. et al. Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nat. Methods* **14**, 600–606 (2017).
12. Duan, J. et al. Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Res.* **24**, 1009–1012 (2014).
13. Kim, D. et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243 (2015).
14. 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).
15. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).
16. Wu, X. et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).
17. Anderson, K. R. et al. CRISPR off-target analysis in genetically engineered rats and mice. *Nat. Methods* **15**, 512–514 (2018).
18. Akcakaya, P. et al. In vivo CRISPR editing with no detectable genome-wide off-target mutations. *Nature* **561**, 416–419 (2018).
19. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
20. Chen, J. S. et al. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
21. Szczelkun, M. D. et al. Direct observation of R-loop formation by single RNA-guided Cas9 and cascade effector complexes. *Proc. Natl Acad. Sci. USA* **111**, 9798–9803 (2014).
22. Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat. Commun.* **7**, 12778 (2016).
23. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl Acad. Sci. USA* **114**, 5461–5466 (2017).
24. Hirano, S., Nishimasu, H., Ishitani, R. & Nureki, O. Structural basis for the altered PAM specificities of engineered CRISPR-Cas9. *Mol. Cell* **61**, 886–894 (2016).
25. Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
26. Lim, Y. et al. Structural roles of guide RNAs in the nuclease activity of Cas9 endonuclease. *Nat. Commun.* **7**, 13350 (2016).
27. Knight, S. C. et al. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350**, 823–826 (2015).
28. Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depken, M. Hybridization kinetics explains CRISPR-Cas off-targeting rules. *Cell Rep.* **22**, 1413–1423 (2018).
29. Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Sci. Adv.* **3**, eaao0027 (2017).
30. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).
31. Singh, D. et al. Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis. *Nat. Struct. Mol. Biol.* **25**, 347–354 (2018).
32. Rueda, F. O. et al. Mapping the sugar dependency for rational generation of a DNA-RNA hybrid-guided Cas9 endonuclease. *Nat. Commun.* **8**, 1610 (2017).
33. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
34. Ran, F. A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
35. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
36. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
37. Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
38. Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS ONE* **10**, e0124633 (2015).
39. Heller, I. et al. STED nanoscopy combined with optical tweezers reveals protein dynamics on densely covered DNA. *Nat. Methods* **10**, 910–916 (2013).
40. Gross, P., Farge, G., Peterman, E. J. G. & Wuite, G. J. L. Combining optical tweezers, single-molecule fluorescence microscopy, and microfluidics for studies of DNA-protein interactions. *Methods Enzymol.* **475**, 427–453 (2010).
41. Brouwer, I. et al. Sliding sleeves of XRCC4–XLF bridge DNA and connect fragments of broken DNA. *Nature* **535**, 566–569 (2016).
42. Zhang, X. et al. Revealing the competition between peeled ssDNA, melting bubbles, and S-DNA during DNA overstretching by single-molecule calorimetry. *Proc. Natl Acad. Sci. USA* **110**, 3865–3870 (2013).
43. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34**, 339–344 (2016).
44. Modesti, M. Fluorescent labeling of proteins. *Methods Mol. Biol.* **783**, 101–120 (2011).
45. Krasikova, Y. S., Rechkunova, N. I., Maltseva, E. A., Petruseva, I. O. & Lavrik, O. I. Localization of xeroderma pigmentosum group A protein and replication protein A on damaged DNA in nucleotide excision repair. *Nucleic Acids Res.* **38**, 8083–8094 (2010).

46. Chen, R., Subramanyam, S., Elcock, A. H., Spies, M. & Wold, M. S. Dynamic binding of replication protein a is required for DNA repair. *Nucleic Acids Res.* **44**, 5758–5772 (2016).
47. King, G. A., Peterman, E. J. G. & Wuite, G. J. L. Unravelling the structural plasticity of stretched DNA under torsional constraint. *Nat. Commun.* **7**, 11810 (2016).
48. van Mameren, J. et al. Unraveling the structure of DNA during overstretching by using multicolor, single-molecule fluorescence imaging. *Proc. Natl Acad. Sci. USA* **106**, 18231–18236 (2009).
49. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
50. Eyquem, J. et al. Targeting a CAR to the TRAC locus with CRISPR/Cas9 enhances tumour rejection. *Nature* **543**, 113–117 (2017).
51. Ren, J. et al. Multiplex genome editing to generate universal car T cells resistant to PD1 inhibition. *Clin. Cancer Res.* **23**, 2255–2266 (2017).
52. Gong, S., Yu, H. H., Johnson, K. A. & Taylor, D. W. DNA unwinding is the primary determinant of CRISPR-Cas9 activity. *Cell Rep.* **22**, 359–371 (2018).
53. O'Geen, H., Henry, I. M., Bhakta, M. S., Meckler, J. F. & Segal, D. J. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Res.* **43**, 3389–3404 (2015).

Acknowledgements

We thank E. Gordon (AstraZeneca) for preparing the Cas9 purified proteins, M. Modesti (CRCM, Marseille) for providing GFP-hRPA, and A. Candelli and J. Cabanas (LUMICKS B.V.) for assistance with the initial tweezers experiments. The Single Molecule Imaging Group is funded by a core grant of the MRC-London Institute of Medical Sciences (UKRI MC-A658-5TY10), a Wellcome Trust Collaborative Grant (206292/Z/17/Z), a Leverhulme Grant (RPG-2016-214), and a BBSRC CASE-studentship (to M.D.N.). The Computational Regulatory Genomics Group is supported by the Medical Research Council UK (MC UP 1102/1), L.R. is supported by the Wellcome Trust (106954) and N.C. is supported by EMBO LTF (EMBO ALTF 1279-2016).

Author contributions

M.D.N., B.J.T., M.E.C. and D.S.R. designed the studies. M.D.N. and R.P.C.D. performed the tweezers experiments. L.R., N.C. and B.L. carried out the de novo motif search. M.D.N. and S.A. did the smFRET and cleavage experiments. M.D.N. and D.S.R. analyzed the data and wrote the manuscript with input from all the authors.

Competing interests

R.P.C.D. is an employee and shareholder of LUMICKS.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-019-0188-z.

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

Correspondence and requests for materials should be addressed to M.E.C. or D.S.R.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

RNA and DNA purification, labeling and annealing. tracrRNA was synthesized by T7 transcription (Supplementary Table 1). Labeled crRNA guides were synthesized, labeled and HPLC purified by IDT (Supplementary Table 1). Synthetic, unmodified DNA non-target strand and modified oligonucleotides were synthesized by IDT (Supplementary Table 1). 5′-Biotin and internal amino linker-modified target-strand DNA oligonucleotides were labeled with Cy5-NHS ester and HPLC purified as previously described[54]. All oligonucleotides were further purified by 18% denaturing polyacrylamide gel electrophoresis. Equimolar concentrations of crRNA and tracrRNA were pre-annealed by heating to 90 °C for 1 min in annealing buffer (100 mN NaCl, 50 mM Tris–HCl pH 8 and 1 mM $MgCl_2$) followed by ~10 min cooling at room temperature. Target DNA synthetic oligonucleotides were annealed by mixing labeled target strand with unlabeled non-target strand at a ratio of 1:2 and heating to 90 °C for 2 min in annealing buffer, followed by slow cooling.

Protein purification, storage and complexing. Cas9 and catalytically dead Cas9 (dCas9; D10A and H840A) were cloned and purified as described previously[32]. Briefly, a synthetic gene coding for Cas9 or dCas9 with an amino (N)-terminal 6xHN tag and a carboxy (C)-terminal nucleoplasmin NLS sequence was synthesized and subcloned into pET24a to generate pET24a-Cas9. *Escherischia coli* BL21DE3* transformants of pET24a-Cas9 were selected on LB plates containing 100 μg ml^{-1} kanamycin. Large-scale cultures were performed by inoculating 750 ml of TB media at 37 °C, and the cultures grown at 37 °C until A_{600}~0.6, after which the culture temperature was lowered to 20 °C and protein production was induced by the addition of 100 μM IPTG. Incubation was continued overnight. Cas9 was purified from a cell lysate. After dilution, the Cas9-containing fractions were loaded onto a 5 ml Heparin HiTrap column (GE Healthcare) and eluted with a linear gradient of NaCl from 0.1 to 1 M, in a buffer containing 20 mM Tris pH 7.5, and 10% glycerol. The protein was polished by size-exclusion chromatography, on a Superdex 26/60 column (GE Healthcare) in a buffer containing 20 mM HEPES pH 7.5, 150 mM KCl, 10% glycerol and 1 mM tris(2-carboxyethyl)phosphine. Fractions containing Cas9 were pooled and concentrated to 10–20 mg ml^{-1} and aliquots flash-frozen in liquid nitrogen. For working stocks dCas9 was diluted to 1.6 μg μl^{-1} in 300 mM NaCl, 10 mM Tris–HCl, 0.1 mM EDTA, 1 mM dithiothreitol, 50% glycerol and stored at −20 °C.

Single-molecule experiments. For single-molecule experiments Cas9 was first complexed at 1 μM concentration with a 1:1 ratio of protein to pre-annealed crRNA:tracrRNA at room temperature for 10 min and subsequently diluted to the required working concentration with imaging buffer (100 mM NaCl, 50 mM Tris–HCl pH 8, 1 mM $MgCl_2$, 0.2 mg ml^{-1} BSA and the oxygen scavenger system with 5 mM protocatechuic acid, 100 nM protocatechuic dioxygenase). hRPA-eGFP was kindly provided by M. Modesti[44] and stored in 200 mM KCl, 20 mM Tris pH 7.5, 1 mM dithiothreitol, 0.5 mM EDTA, 10% glycerol at −80 °C.

Bulk cleavage assays. For bulk DNA cleavage assays Cas9 was first complexed at 1 μM concentration with a 1:1 ratio of protein to 5′-Cy3-crRNA:tracrRNA at room temperature for 10 min. Complexed Cas9 (100 nM) was incubated with annealed DNA (1 nM) for 1 h at 37 °C. The reaction was stopped by adding loading dye containing 96% formamide, 40 mM EDTA (20 mM at 1×) and 0.05% SDS (0.025% at 1×). DNA was loaded on a denaturing 18% polyacrylamide gel and run for ~3 h. Cy3 and Cy5 signals were visualized separately by fluorescence detection with a Fujifilm FLA-5100 gel imager at 200 μm resolution and quantified using Fiji gel analysis tools.

Optical tweezers with confocal microscopy. Optical tweezer confocal microscopy experiments were performed on the commercially available Lumicks C-trap with integrated confocal microscopy and microfluidics. Protein channels of the microfluidics chip were first passivated with BSA (0.1% w/v in PBS) and Pluronics F128 (0.5% w/v in PBS), both flowed through over a period of 30 min. Bacteriophage λ-DNA was labeled at either end with biotin as previously described[40] and attached to 4.5 -μm SPHERO Streptavidin Coated polystyrene particles at 0.005% w/v using the laminar flow cell. DNA integrity was verified before each experiment by generation of force–extension curves from 0–55 pN. For confocal imaging, three excitation wavelengths were used, 488 nm for eGFP, 532 nm for Cy3 and 638 nm for Cy5, with emission detected in three channels with blue filter 512/25 nm, green filter 585/75 nm and red filter 640 LP.

Mapping kymograph peaks onto the λ-DNA genome. Analysis of the location of binding events was achieved by time-binning the fluorescent intensity signal between the two beads followed by mapping across the length of the known λ-DNA sequence. DNA orientation was determined on the basis of the location of the specific on-target binding event and, where possible slight misalignments may have occurred, were corrected by alignment with the location of the known on-target site. For binding localization of nt-crRNA-complexed Cas9, DNA orientation was first determined in a channel containing λ2-crRNA-complexed Cas9 by observing the location of on-target binding before moving to a separate channel containing the nt-crRNA–Cas9 complex. Mapped binned intensities were smoothed using fast Fourier transform and normalized for direct comparison. On the basis of the location of the raw mapping locations of the on-target λ2 site predicted using this method, we determined our resolution of determining binding location to be ~500 bp (Supplementary Fig. 1).

Kinetic and binding distribution analysis. For kinetic analysis, binding events were extracted from kymographs in ImageJ using the available macro Ridge Detection (https://imagej.net/Ridge_Detection) on the basis of a previously described line detection algorithm[55] (Supplementary Fig. 1). To minimize ridge detection errors that occur when multiple binding events occur in close proximity, low 1 nM concentrations of dCas9 were used for measurements at 35–50 pN and detected events shorter than 1 s were discarded. Detected binding events were processed using custom R and Igor Pro 7 scripts (available on request). Dwell-time histograms were constructed with 2 s bin widths, integrated and normalized to generate cumulative probability density plots (Supplementary Fig. 2). These were fit with either single or double exponential models to obtain time constants (τ) and amplitudes (A). To compare the off-target distributions between guides, binding events were identified on kymograms as described above (kinetic analysis), under low (1 nM) [dCas9] and 40 pN force. The position of each detected event was then mapped to the λ-DNA sequence and binned (500 nucleotide width) to generate normalized histograms (Supplementary Figs. 1 and 3). Each guide was repeated in triplicate and the corresponding histograms averaged to generate mean distribution histograms with error bars (Fig. 3c).

De novo motif discovery. To identify stable off-target locations, all technical replicates performed per single crRNAs were used in downstream analyses. First, the time-binned fluorescent intensity signals mapped to the λ-genome at a pulling force of 40 pN were averaged and normalized to the maximum intensity per unique experiment (Supplementary Fig. 3). This resulted in three (λ2-crRNA), four (λ4-crRNA), and five (nt-crRNA) unique λ-DNA intensity signals. Per crRNA, the peaks were aligned using the clupaSpectra function from the ChemoSpec R package (https://CRAN.R-project.org/package=ChemoSpec) to account for slight shifts per experiment, and the intensity signals were then smoothed using the filter fast Fourier transform function from the the nucleR Bioconductor package to reduce the noise[56]. Finally, the intensity profiles were averaged per crRNA. Peaks were identified using the peakDetection function from nucleR at a threshold of 20.

To identify if sequences similar to the crRNA are enriched within recurrent bound regions by the Cas9 complex (peaks), the λ-DNA sequence was scanned (1 bp overlap) for each of the three 20-nucleotide crRNA sequences without PAM at 100%, 80%, and 60% matching threshold. All >60% match regions were then selected and examined for location-specific stretches of nucleotide differences between regions inside and outside peaks. Per match region, each nucleotide was first coded 1 if identical to crRNA sequence and 0 if not. Subsequently, these were weighted; a higher weight was given to a nucleotide match if its two neighbors were also a match, and assigned 2. One neighbor would give 1.5, no neighboring matches remained at 1. Finally, to identify motifs within the peaks, we extracted sequence windows around the peak maxima to a total of 600 bp (300 upstream and 300 downstream from each peak maxima). De novo motif discovery on all peaks was performed using MEME-ChIP software[57].

Single-molecule FRET experiments and analysis. Flow chambers were prepared as previously described[58,59]. Briefly, quartz slides and coverslips were passivated with polyethylene glycol (5% biotinylated) and flow chambers constructed using double-sided sticky tape and sealed with epoxy. Pre-annealed dsDNA (final concentration 12.5 pM) was immobilized via biotin–streptavidin interactions. The flow chambers were imaged on a home-built, prism-based total internal reflection microscope with a 532-nm excitation laser (~2 mW), and images acquired on an EM-CCD camera (Andor) with a 30 ms exposure time. FRET efficiencies were calculated from integrated donor (I_D) and acceptor (I_A) intensities as $FRET = I_A/(I_D + I_A)$[54,59]. Images and data were analyzed by custom IDL, MATLAB and R scripts (available upon request). FRET efficiency histograms were constructed by averaging the first ten frames of each trajectory, filtering traces which gave an average FRET value >1.2 or <−0.2 and binning with bins of 0.05. FRET states were determined by Gaussian fitting and reported as mean ± s.d. When two distributions of bound and unbound populations were observed, percentage values were calculated from the area under the fitted Gaussian of one population as a proportion of the total fitted area of both populations.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability

All custom code is available on request.

Data availability

All data are available in the main text, the supplementary materials, or upon request. Source data for Figs. 2b, 3c and 4f are available with the paper online.

References

54. Senavirathne, G. et al. Activation-induced deoxycytidine deaminase (AID) co-transcriptional scanning at single-molecule resolution. *Nat. Commun.* **6**, 10209 (2015).
55. Steger, C. An unbiased detector of curvilinear structures. *IEEE. Trans. Pattern Anal. Mach. Intell.* **20**, 113–125 (1998).
56. Flores, O. & Orozco, M. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* **27**, 2149–2150 (2011).
57. Bailey, T. L. et al. Meme Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
58. Lamichhane, R., Solem, A., Black, W. & Rueda, D. Single-molecule FRET of protein–nucleic acid and protein–protein complexes: surface passivation and immobilization. *Methods* **52**, 192–200 (2010).
59. Zhao, R. & Rueda, D. RNA folding dynamics by single-molecule fluorescence resonance energy transfer. *Methods* **49**, 112–117 (2009).

nature research

Corresponding author(s): David S. Rueda

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	For optical tweezer data collection commercial LUMICKS C-Trap software was used. For smFRET data collection custom IDL scripts were used.
Data analysis	For analysis custom Matlab and R scripts were used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data is available in the main text, the supplementary materials, or upon request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined to obtain convergent results. For example, single molecule FRET trajectories were accumulated until distributions converged (or no longer changed with additional trajectories, typically over 100). For kymographs, each molecule was observed for long time periods, such that fewer trajectories were required to obtain convergent analyses.
Data exclusions	No data were excluded in most of our analyses. Single molecule fluorescence measurements were validated by intensity levels and single step photo-bleaching. smFRET trajectories not following these widely accepted rules were not included in the analysis.
Replication	Experiments were repeated on multiple DNA molecules, on different days and even on different instruments.
Randomization	No randomization was required in this study.
Blinding	No blinding was required in this study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging