Structural basis for Cas9 off-target activity

Martin Pacesa[1], Chun-Han Lin[2,3], Antoine Cléry[4], Katja Bargsten[1,5], Matthew J. Irby[2], Frédéric H.T. Allain[4], Peter Cameron[2,6], Paul D. Donohoue[2], Martin Jinek[1]

Address:

[1]Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[2]Caribou Biosciences, 2929 Seventh Street #105, Berkeley, CA 94710, United States

[3]Present address: LinkedIn, Sunnyvale, CA 94085, United States

[4]Institute of Biochemistry, ETH Zurich, Hönggerbergring 64, CH-8093 Zurich, Switzerland

[5]Present address: leadXpro AG, Villigen, Switzerland

[6]Present address: Spotlight Therapeutics, Hayward, CA 94545, United States

Corresponding author: Martin Jinek (jinek@bioc.uzh.ch)

Abstract

The target DNA specificity of the CRISPR-associated genome editor nuclease Cas9 is determined by complementarity to a 20-nucleotide segment in its guide RNA. However, Cas9 can bind and cleave partially complementary off-target sequences, which raises safety concerns for its use in clinical applications. Here we report crystallographic structures of Cas9 bound to *bona fide* off-target substrates, revealing that off-target binding is enabled by a range of non-canonical base pairing interactions and preservation of base stacking within the guide–off-target heteroduplex. Off-target sites containing single-nucleotide deletions relative to the guide RNA are accommodated by base skipping rather than RNA bulge formation. Additionally, PAM-distal mismatches result in duplex unpairing and induce a conformational change of the Cas9 REC lobe that perturbs its conformational activation. Together, these insights provide a structural rationale for the off-target activity of Cas9 and contribute to the improved rational design of guide RNAs and off-target prediction algorithms.

42 ## Introduction

43 Cas9, the effector nuclease of prokaryotic Type II CRISPR adaptive immune systems
44 (Makarova et al., 2020), cleaves double-stranded DNA substrates complementary to a guide
45 CRISPR RNA (crRNA) (Jinek et al., 2012). By changing the sequence of the guide RNA, the
46 target DNA specificity of the CRISPR-Cas9 system is readily programmable (Jinek et al.,
47 2012), a feature that has been widely exploited for genome engineering applications (Anzalone
48 et al., 2020). Cas9 functions in conjunction with a trans-activating crRNA (tracrRNA), which
49 is required both for crRNA loading and subsequent DNA binding and cleavage (Deltcheva et
50 al., 2011; Jinek et al., 2012). Target DNA binding and cleavage is further dependent on the
51 presence of a protospacer-adjacent motif (PAM) flanking the target sequence (Anders et al.,
52 2014; Jinek et al., 2012). Due to its high activity and 5'-NGG-3' PAM specificity,
53 *Streptococcus pyogenes* Cas9 (SpCas9) remains the most widely used CRISPR-Cas nuclease
54 for gene editing applications. However, despite a high intrinsic accuracy in generating targeted
55 DNA breaks, SpCas9 can nevertheless cleave genomic DNA sequences with imperfect
56 complementarity to the guide RNA, resulting in off-target editing (Cameron et al., 2017; Hsu
57 et al., 2013; Pattanayak et al., 2013; Tsai et al., 2015). The off-target activity of SpCas9, as well
58 as other Cas9 enzymes, thus presents a safety concern for their therapeutic applications.

59 Off-target sites typically contain one or several nucleobase mismatches relative to the
60 guide RNA (Cameron et al., 2017; Tsai et al., 2017; Tsai et al., 2015). Recent studies have
61 established that the type of mismatch, its positioning within the heteroduplex, and the total
62 number of mismatches are important determinants of off-target DNA binding and cleavage
63 (Boyle et al., 2017; Boyle et al., 2021; Doench et al., 2016; Jones et al., 2020; Zhang et al.,
64 2020). PAM-proximal mismatches within the seed region of the guide RNA-target DNA strand
65 heteroduplex typically have a dramatic impact on substrate DNA binding and R-loop formation
66 (Boyle et al., 2021; Ivanov et al., 2020; Singh et al., 2016; Zhang et al., 2020). In contrast,

67 PAM-distal mismatches are compatible with stable binding; however, their presence often
68 results in the formation of a catalytically incompetent complex (Boyle et al., 2021; Dagdas et
69 al., 2017; Ivanov et al., 2020; Jones et al., 2020; Sternberg et al., 2015; Yang et al., 2018; Zhang
70 et al., 2020). In addition, Cas9 has been shown to cleave off-target substrates containing
71 insertions or deletions relative to the guide RNA sequence, which have been proposed to be
72 recognized through the formation of nucleotide "bulges" in the guide RNA-target DNA
73 heteroduplex (Boyle et al., 2021; Cameron et al., 2017; Doench et al., 2016; Jones et al., 2020;
74 Lin et al., 2014; Tsai et al., 2015).

75 Numerous computational tools have been developed to predict possible genomic off-
76 target sites based on sequence similarity (Bae et al., 2014; Stemmer et al., 2015). However, the
77 majority of actual off-target cleavage events remain unpredicted (Cameron et al., 2017; Tsai et
78 al., 2015). Furthermore, although Cas9 is able to bind genomic sites harbouring as few as five
79 complementary nucleotides, only a relatively small number of off-target sites are actually
80 cleaved and result in detectable off-target editing in cells (Kuscu et al., 2014; O'Geen et al.,
81 2015; Wu et al., 2014). Several structures of target-bound Cas9 complexes have been
82 determined to date (Anders et al., 2014; Jiang et al., 2016; Nishimasu et al., 2014; Zhu et al.,
83 2019) that have shed light on the mechanism of on-target binding and cleavage. However, the
84 same processes for off-target sites remain poorly understood.

85 To elucidate the mechanism of mismatch tolerance of Cas9, we determined crystal
86 structures of a comprehensive set of *bona fide* off-target–bound complexes. These structures
87 reveal that the formation of non-canonical base pairs and preservation of heteroduplex shape
88 underpin the off-target tolerance of Cas9. We also observe that consecutive mismatches in the
89 seed region can be accommodated by base skipping of a guide RNA nucleotide, as opposed to
90 nucleotide bulging. Finally, the structure of an off-target complex containing three PAM-distal
91 mismatches exhibits REC2/3 domain rearrangements, which likely perturbs conformational

activation of Cas9 and thus modulates cleavage efficiency. Taken together, our structural data reveal the diversity of mechanisms enabling off-target recognition and lay the foundation for engineering optimized CRISPR-Cas9 complex designs for gene editing.

95 ## Results

96 ### *In vitro* profiling reveals diversity of Cas9 off-targets

97 Multiple studies have investigated the off-target activity of Cas9, suggesting context-dependent
98 tolerance of nucleobase mismatches between the guide RNA and off-target DNA sequences
99 (Boyle et al., 2021; Cameron et al., 2017; Lazzarotto et al., 2020; Tsai et al., 2015; Zhang et al.,
100 2020). To investigate the effect of mismatches on Cas9 binding and cleavage, we performed
101 the SITE-Seq® assay (Cameron et al., 2017) to define the off-target landscapes of 12 well-
102 studied guide RNAs to select suitable off-targets for further evaluation (**Table S1, Table S2**)
103 (**Figure S1A-B**). The SITE-Seq assay analysis revealed a total of 3,848 detectable off-target
104 sites at the highest Cas9 ribonucleoprotein (RNP) concentration, with a total of 21,732 base
105 mismatches and a median of 5 mismatches per off-target site (**Figure S1C**). The detected
106 mismatches covered all possible base mismatch combinations and were distributed throughout
107 the length of the guide RNA-target DNA heteroduplex (**Figure S1D-E**).

108 To probe the thermodynamics of on- and off-target substrate DNA binding and the
109 kinetics of DNA cleavage, we focused on a subset of four guide RNAs (*AAVS1*, *FANCF*,
110 *PTPRC-tgt2,* and *TRAC*) and a total of 15 *bona fide* off-target sites detectable *in vivo* (Cameron
111 et al., 2017; Donohoue et al., 2021; Tsai et al., 2017; Tsai et al., 2015) (**Figure 1A**) that covered
112 all 12 possible base mismatch types. Nuclease activity assays using synthetic DNA substrates
113 with fluorophore-labeled target strand (TS) revealed that all selected off-target sequences were
114 cleaved slower than the corresponding on-target substrates, with 20–500-fold reductions in the
115 calculated rate constants (**Figure 1B, Figure S2A**). To distinguish whether the cleavage defects
116 were due to slower R-loop formation or perturbations in downstream steps, including
117 conformational activation of the nuclease domains, we also quantified cleavage kinetics using
118 PAMmer DNA substrates (Anders et al., 2014; O'Connell et al., 2014) in which the on-/off-
119 target sequence was single-stranded. These experiments revealed that the slower cleavage

120 kinetics of most off-target substrates was due to perturbed R-loop formation (**Figure 1B, Figure**
121 **S2B**). However, for some off-targets, notably *AAVS1* off-targets #2 and #5, *FANCF* off-targets
122 #3, #4, and #6, and *PTPRC-tgt2* off-target #1, the rate of PAMmer substrate cleavage was more
123 than 100-fold slower as compared to their respective on-target sequences (**Figure 1B**, **Figure**
124 **S2B**). This indicates that these off-target mismatches may additionally cause perturbations in
125 the conformational activation checkpoint downstream of guide-target hybridization or inhibit
126 cleavage by direct steric hindrance of the Cas9 HNH domain (Chen et al., 2017; Dagdas et al.,
127 2017).

128 Complementary quantification of substrate DNA binding using DNA nanolever
129 (switchSENSE) methodology revealed perturbations in binding affinities for most off-targets
130 as compared to the respective on-target sequences (**Figure 1B, Data S1**). Notably, the
131 reductions in binding affinities were almost entirely due to increased dissociation rates (k_{off}),
132 while on-binding rates (k_{on}) were largely unperturbed (**Figure 1B**), indicating that most of the
133 off-target mismatches in our data set promote DNA dissociation, likely due to R-loop collapse.
134 However, there was little correlation between the observed reductions in cleavage rates and
135 binding constants (**Figure S2C**), confirming that the molecular basis for off-target
136 discrimination by Cas9 is not based on substrate binding alone, in agreement with prior studies
137 (Boyle et al., 2021; Chen et al., 2017; Dagdas et al., 2017; Jones et al., 2020; Yang et al., 2018;
138 Zhang et al., 2020). The dissociation rate (k_{off}) correlated significantly only with the number of
139 mismatches located in the seed region (R^2=0.46, p=0.001) (**Figure S2C**), suggesting that off-
140 targets with mismatches in the seed are regulated mainly through R-loop collapse and non-
141 target strand (NTS) rehybridization (Boyle et al., 2017; Gong et al., 2018; Singh et al., 2016;
142 Sternberg et al., 2014).

143 Finally, we correlated the measured cleavage rate constants (k_{obs}) with predicted data
144 based on a leading biophysical model of Cas9 off-target cleavage that accounts for mismatch

145 number and position using position-dependent penalties and includes position-independent
146 weights for mismatch type (Jones et al., 2020). Although there was good overall correlation
147 between the model and our data (R^2=0.46, p=0.004) (**Figure S2C**), there were nevertheless
148 several prominent outliers (*AAVS1* off-target #2 and off-target #3, and *FANCF* off-target #4),
149 suggesting that accurate modelling of off-target interactions will require accounting for
150 position-specific effects of individual mismatch types (**Figure S2D**).

151 Taken together, these results indicate that *bona fide* off-target substrates exhibit
152 significantly perturbed kinetics of substrate DNA binding and cleavage. Moreover, the
153 magnitude of the perturbation is dependent not only on the number and position of mismatches
154 in the off-target sequence but also on mismatch type, in agreement with recent studies (Boyle
155 et al., 2021; Doench et al., 2016; Jones et al., 2020; Zhang et al., 2020). Moreover, the off-target
156 activity cannot be accurately predicted by biophysical off-target activity models that account
157 for mismatch type in a position-independent manner, implying that mismatches have position-
158 specific and context-dependent effects. This further highlights the need to understand the
159 molecular principles of Cas9 off-target recognition at the structural level.

160 **Crystallographic analysis of off-target interactions**

161 To obtain insights into the structural basis of off-target recognition and mismatch tolerance, we
162 employed a previously described approach (Anders et al., 2014) to co-crystallize Cas9 with
163 sgRNA guides and partially duplexed off-target DNA substrates (**Figure 1C**). Focusing on our
164 set of *AAVS1*, *FANCF*, *PTPRC-tgt2* and *TRAC* off-targets (**Figure 1A**), covering all 12 possible
165 mismatch types, we determined a total of 15 off-target complex structures at resolutions of
166 2.25–3.30 Å (**Figure 1C**, **Table S3)**. For the *AAVS1*, *FANCF* and *TRAC* guide RNAs, we also
167 determined the structures of the corresponding on-target complexes; the structure of the
168 *PTPRC-tgt2* on-target complex could not be determined due to insufficient diffraction of the
169 crystals. Overall, the off-target complex structures have very similar conformations, with the

170 Cas9 polypeptide backbone superimposing with a mean root mean square deviation of 0.41Å
171 over 1330 Cα atoms (as referenced to the *FANCF* on-target complex structure, excluding
172 *FANCF* off-target #4, as discussed below). Of note, the *AAVS1* on-target complex structure
173 reveals substantial repositioning of the REC2 domain, where it undergoes a 12° rotation (as
174 compared to the *FANCF* and *TRAC* on-target complexes) (**Figure S3A**), with concomitant
175 shortening of the α-helix comprising residues 301-305 and restructuring of the loop comprising
176 residues 175-179 (**Figure S3B**), enabled by the absence of crystal contacts involving the REC2
177 and REC3 domains.

178 However, the structures display considerable local variation of the RNA-TS DNA
179 heteroduplex conformation. Base base pairing and base stacking are mostly preserved
180 throughout the guide RNA-TS DNA heteroduplexes (**Table S4**), with the exception of positions
181 1–3 within the PAM-distal end of the guide-TS duplex, where the presence of mismatches
182 results in duplex unpairing. This is observed in *AAVS1* off-target #2 and #4, *FANCF* off-target
183 #4 and #5, and *TRAC* off-target #1 complexes (**Figure S4**). Despite the observed
184 conformational variation, the off-target structures preserve almost all intermolecular contacts
185 between the Cas9 protein and the bound nucleic acids, further underscoring the structural
186 plasticity of Cas9 in accommodating mismatch-induced distortions in the guide RNA-TS DNA
187 heteroduplex.

188 Together, these observations indicate that the crystal form used for determination of the
189 off-target complex structures is sufficiently plastic to accommodate conformational changes
190 resulting from the presence of base mismatches in the guide RNA–TS DNA heteroduplex and
191 imply that the observed structural effects of guide RNA–TS DNA base mismatches provide a
192 true depiction of off-target DNA binding. In addition, the HNH and REC2/3 domain
193 conformations observed in the crystal structures are similar to those observed in the 16-bp
194 heteroduplex, pre-checkpoint state determined by cryo-EM (Pacesa and Jinek, 2021).

195 **Non-canonical base-pairing interactions facilitate off-target recognition**

196 Close inspection of the 15 Cas9 off-target complex structures reveals that a substantial fraction
197 of off-target base mismatches (34 out of 49) is accommodated by non-canonical base pairing
198 interactions that preserve at least one hydrogen bond between the guide and off-target bases.
199 The most common off-target mismatches, both in our data set (**Table S1, Table S2**) and as
200 reported by other studies (Boyle et al., 2017; Doench et al., 2016; Hsu et al., 2013; Jones et al.,
201 2020; Pattanayak et al., 2013; Zhang et al., 2020), are rG-dT (**Figure S5**) and rU-dG (**Figure**
202 **2A, Figure S6**), which have the potential to form wobble base pairs (Kimsey et al., 2015).
203 Indeed, all rG-dT mismatches in the determined structures are accommodated by wobble base
204 pairing. At duplex positions 4 (*AAVS1* off-target #1 and #5), 13 (*FANCF* off-target #2 and #7)
205 and 15 (*TRAC* off-target #1), the dT base undergoes a ~1 Å shear displacement into the major
206 groove of the guide–TS DNA heteroduplex to form the wobble base pair (**Figure S5**), whereas
207 at duplex position 2 (in *TRAC* off-target #1 and #2), wobble base pairing is enabled by a minor
208 groove displacement of the rG base. In contrast, rU-dG mispairs in the determined structures
209 exhibit considerable structural variation. At duplex position 10 in the *FANCF* off-target #2 and
210 #4 complexes, the rU base is able to undergo the major groove displacement required for
211 wobble base-pairing (**Figure 2A, Figure S6A**). In contrast, at duplex position 5 in *FANCF* off-
212 target #1, #3 and #6 complexes, the backbone of the RNA strand makes extensive contacts with
213 Cas9 (**Figure S6B-D**). As a result, the rU-dG mispairs are instead accommodated by
214 compensatory shifts of the dG base to maintain hydrogen-bonding interactions (**Figure S6B-**
215 **D**). At duplex position 9 in the *TRAC* off-target #1 complex, the rU-dG mismatch is
216 accommodated by wobble base-pairing enabled by a minor groove displacement of the dG base
217 (**Figure S6E**). At the same duplex position in the *AAVS1* off-target #1 and #2 complexes,
218 however, this mispair occurs next to rC-dA and rC-dT mismatches, respectively, and adopts the
219 sterically prohibited Watson-Crick geometry (**Figure S6F-G**), implying a tautomeric shift or

220 base deprotonation to accommodate this otherwise unfavorable base pairing mode (**Figure**
221 **S6H**). Collectively, these observations suggest that the ability of rU-dG (and likely rG-dT)
222 mismatches to form wobble base pairing interactions is determined not only by backbone
223 constraints at the specific position within the guide RNA-TS DNA heteroduplex, but also by
224 local sequence context and/or the presence of neighboring mismatches.

225 rA-dC or rC-dA mismatches can also form wobble-like base pairs when the adenine
226 base is protonated at the N1 position (Garg and Heinemann, 2018; Wang et al., 2011). In the
227 rA-dC mispairs found at duplex position 4 in the *FANCF* off-target #2 and #3 complexes, the
228 dC nucleotide undergoes a wobble displacement compatible with the formation of two
229 hydrogen bonds with adenine base, indicative of adenine protonation (**Figure 2B**, **Figure S7A**).
230 At other duplex positions in our data set, the rA-dC or rC-dA mispairs are instead
231 accommodated by slight displacements of the adenine base within the base pair plane resulting
232 in the formation of a single hydrogen bond in each case (**Figure S7B-D**).

233 Accommodating purine-purine mismatches by Watson-Crick-like interactions would
234 normally require severe distortion of the guide–off-target duplex to increase its width by more
235 than 2 Å (Leontis et al., 2002). At positions where the duplex width is constrained by Cas9
236 interactions, rG-dA and rA-dG mispairs are accommodated by *anti*-to-*syn* isomerization of the
237 adenine base to form two hydrogen-bonding interactions via its Hoogsteen base edge. This is
238 observed at duplex position 11 in the *AAVS1* off-target #4 complex (rG-dA mispair) (**Figure**
239 **2C**) and at position 7 in the *AAVS1* off-target #2 complex (rA-dG mispair) (**Figure S7E**).
240 Similarly, the rG-dG mispair at duplex position 13 in the *FANCF* off-target #5 complex is
241 accommodated by Hoogsteen base-pairing as a result by *anti*-to-*syn* isomerization of the guide
242 RNA base (**Figure 2D**). Overall, the observed Hoogsteen base pairing interactions are near-
243 isosteric with Watson-Crick base pairs and maintain duplex width without excessive backbone
244 distortion (**Table S4**).

245 Taken together, the prevalence of non-canonical base-pairing interactions, such as
246 wobble and Hoogsteen base pairing, in off-target structures indicates that they serve a
247 fundamental role in off-target recognition. These interactions preserve hydrogen bonding
248 between guide RNA and off-target DNA bases while simultaneously maintaining the integrity
249 of the guide RNA–off-target DNA heteroduplex and minimizing its structural distortions.

250 **Duplex backbone rearrangements accommodate otherwise non-permissive base mispairs**

251 Whereas wobble (G-U/T or A-C) and Hoogsteen (A-G or G-G) base pairs are generally
252 compatible with the canonical A-form geometry of an RNA-DNA duplex, other nucleotide
253 mismatches only form non-isosteric base pairs that require considerable distortion of the
254 (deoxy)ribose-phosphate backbone. The formation of pyrimidine-pyrimidine base pairs is
255 expected to occur by a substantial reduction in duplex width. This is observed at the rU-dC
256 mismatch at duplex position 9 in the *FANCF* off-target #1 complex (**Figure 3A**). Here, the
257 guide RNA backbone is able to shift towards the target DNA strand, resulting in a reduction of
258 the C1'–C1' distance to 8.65 Å as compared to 10.0 Å in the *FANCF* on-target complex. This
259 facilitates the formation of two hydrogen bonding interactions within the rU-dC base pair,
260 which is further enabled by a substantial increase in base propeller twist (**Figure 3A**). In
261 contrast, rC-dT mismatches remain unpaired at duplex positions 6 and 7 in *FANCF* off-target
262 #6 and #7 complexes (**Figure S8A-B**), respectively, or form only a single hydrogen bond at
263 position 8 in the *AAVS1* off-target #2 complex (**Figure S8C**), likely due to backbone steric
264 constraints at these positions imposed by Cas9 interactions. Of note, the *FANCF* off-target #7
265 rC-dT mispair is bridged by hydrogen bonding interactions with the side chain of Arg895
266 inserted into the minor groove of the heteroduplex (**Figure S8B**); however, the interaction is
267 not essential for the tolerance of rC-dT mismatches at this position (**Figure S9**). Backbone steric
268 constraints also likely influence the formation of rU-dT base pairs. At duplex position 7 in the
269 *TRAC* off-target #2 complex, the mismatch remains unpaired (**Figure S8D**), whereas

270 productive pairing is seen at duplex positions 8 (*PTPRC-tgt2* off-target #1 complex) and 9
271 (*FANCF* off-target #5 and *TRAC* off-target #2 complexes), facilitated by distortions of the guide
272 RNA and TS backbone, respectively (**Figure 3B, Figure S8E-F**).

273 rC-dC mispairs only form productive hydrogen-bonding interactions if bridged by a
274 water molecule or when one of the cytosine bases is protonated (Leontis et al., 2002). Only the
275 former is observed in the determined structures, at duplex position 5 in the *AAVS1* off-target #2
276 complex (**Figure S8G**). In contrast, at duplex position 8 and 15 in the *AAVS1* off-target #3
277 complex, the bases remain unpaired while maintaining intra-strand base stacking interactions
278 (**Figure 3C, Figure S8H**). Similarly, rA-dA mispairs are unable to form productive hydrogen-
279 bonding interactions within the constraints of an A-form duplex (Leontis et al., 2002).
280 Accordingly, the rA-dA mispair at duplex position 5 in the *PTPRC-tgt2* off-target #1 complex
281 is accommodated by extrusion of the dA nucleobase out of the base stack into the major groove
282 of the duplex (**Figure 3D**). As the duplex width is constrained at this position by Cas9, the base
283 extrusion is enabled by local distortion of the TS backbone (**Figure S10A**). Analysis of our
284 SITE-Seq assay data set revealed that off-target rA-dA mispairs occur at all positions within
285 the guide RNA–TS DNA heteroduplex (**Figure S10B**), in agreement with previous studies
286 (Boyle et al., 2017; Boyle et al., 2021; Doench et al., 2016; Jones et al., 2020; Zhang et al.,
287 2020). This suggests that rA-dA mismatches do not encounter steric barriers within Cas9 that
288 would disfavour their presence, which is consistent with the absence of specific contacts with
289 Cas9 along the length of the major groove of the guide RNA–TS DNA duplex.

290 Collectively, these structural findings indicate that conformational rearrangements of
291 the (deoxy)ribose-phosphate backbone of the guide RNA or TS DNA facilitate interactions of
292 base mispairs that would otherwise be incompatible with canonical A-form duplex geometry.
293 The specific mechanism of base mismatch accommodation at a given position is governed by

294 local steric constraints on duplex width and the ability of the guide RNA or TS DNA to undergo
295 backbone distortions, which are in turn dictated by local interactions with Cas9.

296 **PAM-proximal mismatches are accommodated by TS distortion due to seed sequence**
297 **rigidity**

298 The seed sequence of the guide RNA (nucleotides 11-20) makes extensive interactions with
299 Cas9, both in the absence and presence of bound DNA (Anders et al., 2014; Jiang et al., 2015;
300 Nishimasu et al., 2014; Zhu et al., 2019). Structural pre-ordering of the seed sequence by Cas9
301 facilitates target DNA binding and contributes to the specificity of on-target DNA recognition
302 (Jiang et al., 2015; O'Geen et al., 2015; Wu et al., 2014). Conversely, binding of off-target
303 DNAs containing PAM-proximal mismatches is inhibited and results in accelerated off-target
304 dissociation (Boyle et al., 2017; Boyle et al., 2021; Ivanov et al., 2020; Jones et al., 2020; Singh
305 et al., 2016; Zhang et al., 2020). Nevertheless, Cas9 does tolerate most base mismatch types
306 within the seed region of the guide RNA, leading to detectable off-target DNA cleavage (Boyle
307 et al., 2021; Doench et al., 2016; Jones et al., 2020; Zhang et al., 2020). In particular, the first
308 two PAM-proximal positions display a markedly higher tolerance for mismatches than the rest
309 of the seed region (Cofsky et al., 2021; Doench et al., 2016; Hsu et al., 2013; Mekler et al.,
310 2017; Zeng et al., 2018); this is supported by our SITE-Seq assay data as the frequency of
311 mismatches at the first three PAM-proximal positions is roughly twice as high as at the other
312 seed sequence positions (**Figure S1D-E**).

313 Unlike the seed region of the guide RNA, the complementary PAM-proximal TS
314 nucleotides are not directly contacted by Cas9 in the pre-cleavage state and are thus under fewer
315 steric constraints, with the exception of duplex position 20 in which the phosphodiester group
316 of the TS nucleotide makes extensive interactions with the phosphate lock loop of Cas9 (Anders
317 et al., 2014) (**Figure S11A**). In agreement with this, our off-target complex structures reveal
318 that PAM-proximal base mismatches are accommodated solely by structural distortions of the

319 TS backbone, while the conformation of the guide RNA backbone and base stacking within the
320 seed region remain unperturbed. The presence of an rA-dA mismatch in the PAM-proximal
321 position 18 of *FANCF* off-target #6 results in the extrusion of the TS nucleobase into the major
322 groove (**Figure 4A**), likely due to steric constraints on duplex width at this position. In contrast,
323 the rA-dA mismatch at duplex position 19 in the *AAVS1* off-target #2 is instead accommodated
324 by a marked distortion in the TS backbone that results in increased duplex width, which
325 preserves base stacking within the duplex in the absence of productive pairing between the
326 adenine bases (**Figure 4B**, **Figure S11B**). Similarly, the rA-dG mismatch at position 19 in the
327 *AAVS1* off-target #5 is accommodated by a ~2 Å displacement of the TS backbone, increasing
328 duplex width. This not only preserves base stacking but also facilitates rA-dG base paring by
329 two hydrogen bonding interactions via their Watson-Crick edges (**Figure 4D**, **Figure S11C**).
330 This off-target complex also contains a rU-dG mispair at duplex position 20 which does not
331 undergo wobble base pairing as the rU20 nucleotide is extensively contacted by Cas9 and
332 unable to shift towards the major groove and is instead accommodated by a slight shift in the
333 dG nucleotide (**Figure 4D**). Finally, the rU-dT base mismatch at duplex position 20 in the
334 *AAVS1* off-target #4 complex remains unpaired and the dT base lacks ordered electron density
335 (**Figure 4C**). This is likely a result of the dT nucleotide maintaining contact with the phosphate-
336 lock loop of Cas9, which prevents a reduction in the duplex width and precludes productive
337 base pairing.

338 Overall, these observations indicate that off-target DNAs containing mismatches to the
339 seed sequence of the guide RNA can be accommodated by Cas9 due to limited interactions with
340 the TS DNA in the seed-binding region, which permits structural distortions of the TS backbone
341 to accommodate base mispairs without steric hindrance and may facilitate non-canonical base
342 pairing interactions. Conversely, the extensive interactions of Cas9 with the ribose-phosphate

343 backbone of the seed region of the guide RNA provide strong steric constraints that would be
344 expected to disfavour specific base mispairs.

345 Cas9 recognizes off-targets with single-nucleotide deletions by base skipping or via 346 multiple mismatches

347 A substantial fraction of *bona fide* off-target sites recovered in our SITE-Seq assay analysis
348 (46.4%, when not considering the possibility of nucleotide insertions or deletions) contained
349 six or more mismatched bases to the guide RNA (**Figure S1C**, **Table S1, Table S2**). Such off-
350 target sequences have previously been proposed to be accommodated by bulging out or
351 skipping of nucleotides (Boyle et al., 2021; Cameron et al., 2017; Doench et al., 2016; Jones et
352 al., 2020; Lin et al., 2014; Tsai et al., 2015), which would result in a shift of the nucleotide
353 register to re-establish correct base pairing downstream of the initially encountered mismatch.
354 The *PTPRC-tgt2* off-target #1, *FANCF* off-target #3 and *AAVS1* off-target #2 sites are predicted
355 to contain single nucleotide deletions at duplex positions 15, 17 and 9, respectively (**Figure 1A**,
356 **Figure 5C**). Structures of the *PTPRC-tgt2* off-target #1 and *FANCF* off-target #3 complexes
357 reveal that the single nucleotide deletions in these off-target substrates are not accommodated
358 by bulging out the unpaired guide RNA nucleotide. Instead, the conformations of the guide
359 RNAs remain largely unperturbed and the off-target TS DNAs "skip over" the unpaired RNA
360 bases to resume productive base-pairing downstream (**Figure 5A-B**). Comparisons with the
361 *FANCF* on-target complex structure show that the seed sequences of the guide RNAs are held
362 in place by interactions with the bridge helix and the REC1 domain, whereas the DNA target
363 strand phosphate backbones are displaced by almost 3 Å (**Figure S12A-B**). The base pair skips
364 are accommodated by considerable buckling and tilting of the base pairs immediately
365 downstream of the skip site. An additional consequence of the base pairing register shift is the
366 formation of non-canonical base pairs between the off-target DNA and the extra 5'-terminal
367 guanine nucleotides present in the guide RNA as a consequence of *in vitro* transcription by T7

368 RNA polymerase (**Figure S12C-D**). This potentially explains the impact of the 5'-guanines on
369 both R-loop stability and *in vitro* cleavage activity (Kulcsar et al., 2020; Mullally et al., 2020;
370 Okafor et al., 2019).

371 Originally, our SITE-Seq assay analysis annotated the *AAVS1* off-target #2 as a single-
372 nucleotide deletion at duplex position 9 (**Figure 5C**). Unexpectedly, the structure of the *AAVS1*
373 off-target #2 complex instead reveals that the off-target substrate is bound in the unshifted
374 register, resulting in the formation of five base mismatches in the PAM-distal half of the guide
375 RNA–TS duplex (**Figure 5D**), including a partially paired rC-dC mismatch at position 5, an
376 rA-dG Hoogsteen pair at position 7, a partially paired rC-dT mismatch at position 8, and a
377 tautomeric rU-dG pair at position 9. The backbone conformations of the guide RNA and the
378 off-target TS exhibit minimal distortions and are nearly identical with the corresponding on-
379 target heteroduplex (**Figure 5D**), suggesting an explanation for the tolerance of the multiple
380 mismatches in this off-target site, and implying that certain mismatch combinations might
381 cumulatively result in guide RNA and TS backbone conformations that mimic the on-target
382 situation. To test this hypothesis, we reverted the rC-dT mismatch at position 8 to the on-target
383 rC-dG pair, thereby reducing the total amount of off-target mismatches from 6 to 5 (**Figure**
384 **5C**). The resulting off-target substrate (*AAVS1* off-target #6) exhibited substantially reduced
385 cleavage rates in both dsDNA and PAMmer formats, as well as a significantly increased
386 dissociation rate (**Figure 5E-F**). These results suggest that for some *bona fide* off-target
387 substrates containing mismatch combinations, the reversal of one mismatch may affect the
388 structural integrity of the guide RNA–TS DNA heteroduplex and interfere with DNA binding
389 and/or conformational activation of Cas9, despite a reduction in the total number of
390 mismatches.

391 Collectively, these results indicate that deletion-containing off-target complexes are
392 accommodated either by RNA base skipping, as opposed to RNA nucleotide bulging, or by the

393 formation of multiple base mismatches. The precise mechanism appears to be dependent on the
394 position of the deletion. Because the seed sequence nucleotides 12-20 of the guide RNA are
395 extensively contacted by Cas9 (**Figure S13**), while the complementary DNA nucleotides are
396 able to undergo distortions to accommodate a shift in the base pairing register, deletions at
397 positions within the PAM-proximal region of the guide RNA–TS heteroduplex (positions 11-
398 20) result in RNA base skipping. In contrast, deletions at PAM-distal positions (1-10), where
399 the guide RNA–TS DNA heteroduplex is constrained by interactions with the REC3 and HNH
400 domains, are instead likely to be bound without a register shift via multiple mismatches.

401 In light of our structural findings, we computationally analyzed off-target sites
402 identified by the SITE-Seq assay for the presence of insertions or deletions in the target DNA
403 relative to the guide RNA sequence. Our initial off-target classification algorithm assumed that
404 deletions and insertions can occur along the entire guide RNA–off-target DNA heteroduplex.
405 Based on our structural data we subsequently constrained the algorithm to only consider single-
406 nucleotide deletions and insertions at heteroduplex positions 10-20 and 6-20, respectively
407 (**Figure S14**). This resulted in a substantial reduction in the number of off-targets containing
408 deletions (from 323 to 277) but no change in off-targets predicted to contain insertions (116
409 sites for both). When extrapolated, these results collectively suggest that up to 14% of off-target
410 sites previously annotated as containing deletions or insertions in off-target studies might
411 instead be recognized via multiple mismatches (**Figure S15**), which has implications for off-
412 target prediction, as discussed below.

413 **PAM-distal mismatches perturb the Cas9 conformational checkpoint**

414 *FANCF* off-target #4, which contains three PAM-distal mismatches at positions 1-3 and a G-U
415 mismatch in position 10 (**Figure 1A**), is reproducibly the top ranking off-target site for the
416 *FANCF* guide RNA, as detected by SITE-Seq assay analysis at the lowest Cas9 RNP
417 concentrations (**Table S1, Table S2**). The off-target substrate exhibits slow cleavage kinetics

418 *in vitro* with both dsDNA and PAMmer substrates (**Figure 2B, Figure S2A-B**), indicating a
419 perturbation of the conformational activation checkpoint of Cas9. The structure of the *FANCF*
420 off-target #4 complex reveals that the RNA–DNA heteroduplex is unpaired at positions 1-3 as
421 a result of the PAM-distal mismatches, with nucleotides 1-2 of the guide RNA and 19-20 of the
422 TS disordered (**Figure 6A**). Furthermore, Cas9 undergoes structural rearrangements of its REC
423 lobe and the HNH domain (**Figure 6B**), resulting in a root mean square displacement of the
424 REC2 and REC3 domains of 3.7 Å (1,315 Cα atoms) relative to the *FANCF* on-target complex
425 structure. The REC3 domain undergoes a 19-degree rotation (**Figure 6B**), facilitated by
426 extending the helix comprising residues 703-712 through restructuring of loop residues 713-
427 716 (**Figure S16A)**, to accommodate the altered guide RNA conformation. The REC2 domain
428 rotates 32 degrees away from the REC3 domain (**Figure 6B**). This is accompanied by
429 restructuring of the hinge loop residues 174-180 and disordering of loops 258-264, 284-285,
430 and 307-309. Concomitantly, the HNH domain rotates 11° away from the heteroduplex, as
431 compared to the *FANCF* on-target structure, to accommodate distortion of the TS DNA (**Figure**
432 **6B**).

433 The unpaired 5' end of the sgRNA is located at the interface between the REC3 and the
434 RuvC domain and maintains interactions with heteroduplex-sensing residues Lys510, Tyr515,
435 and Arg661 of the REC3 domain (**Figure S16B).** In contrast to the corresponding on-target
436 complex structure, the unpaired 3' end of the off-target TS breaks away from the REC3 lobe
437 and instead points towards the REC2 domain, forming unique interactions with Arg895,
438 Asn899, Arg905, Arg919 and His930 in the HNH domain (**Figure S16C**). These interactions
439 (**Figure S16D)** could be responsible for the observed repositioning of the REC lobe and HNH
440 domain, and they may impede the formation of a cleavage-competent complex.

441 The conformation of the *FANCF* off-target #4 complex is distinct from the
442 conformations observed in cryo-EM reconstructions of the pre- and post-cleavage states of the

443 Cas9 complex (Zhu et al., 2019) (**Figure S17A-B**). Instead, the off-target complex structure
444 most closely resembles that of a high-fidelity variant xCas9 3.7 containing amino acid
445 substitutions that disrupt interactions with the TS DNA (Guo et al., 2019). Although the xCas9
446 3.7 complex adopts a slightly different REC lobe conformation (**Figure S17C**), the PAM-distal
447 duplex also undergoes unpairing at positions 1–3 and displays a comparable degree of structural
448 disorder (**Figure S17D**). These structural observations thus suggest that the presence of
449 multiple mismatches in the PAM-distal region of a guide RNA–off-target DNA duplex leads to
450 conformational perturbations in the DNA-bound complex that resemble the structural
451 consequences of specificity-enhancing mutations in high-fidelity Cas9 variants.
452
453

454 ## Discussion

455 The off-target activity of Cas9 has been extensively documented in prior genome editing,
456 biochemical and biophysical studies (Boyle et al., 2017; Boyle et al., 2021; Doench et al., 2016;
457 Jones et al., 2020; Lazzarotto et al., 2020; Zhang et al., 2020). Although numerous methods
458 have been devised for computational prediction of genomic off-target sites and their
459 experimental validation, these have reported highly variable mismatch tolerance profiles
460 depending on the screening method and the target sequence. Thus, a comprehensive
461 understanding of this phenomenon is still lacking, particularly as to whether off-target tolerance
462 has an underlying structural basis. In this study, we used the SITE-Seq assay to examine the
463 off-target landscape of 12 well-studied guide RNAs, observing a broad variation of cleavage
464 activities associated with individual off-target substrates. To shed light on the molecular
465 mechanisms underpinning off-target activity, we determined atomic structures of a
466 representative set of *bona fide* off-target complexes, thus providing fundamental insights into
467 the structural aspects of off-target recognition.

468 ### Role of non-canonical base pairing in off-target recognition

469 The principal, and largely unexpected, finding of our structural analysis is that the majority of
470 nucleotide mismatches in *bona fide* off-target substrates are accommodated by non-canonical
471 base pairing interactions. These range from simple rG-dT/rU-dG wobble or Hoogsteen base
472 pairing interactions, to pyrimidine-pyrimidine pairs that rely on (deoxy)ribose-phosphate
473 backbone distortions that reduce duplex width. With the notable exception of rA-dA mispairs,
474 which are accommodated at certain positions within the guide–TS heteroduplex by base
475 extrusion, the structural rearrangements associated with base mismatch accommodation
476 preserve base stacking, which is the primary determinant of nucleic acid duplex stability
477 (Yakovchuk et al., 2006). For some off-target sequences, our structures are suggestive of base
478 protonation or tautomerization, which facilitate hydrogen bonding interactions in otherwise

479 non-permissive base pair combinations, such as rA-dC. These rare base pair forms have been
480 previously observed in both RNA and DNA duplexes and are thought to be important
481 contributors to DNA replication and translation errors (Kimsey et al., 2015; Kimsey et al.,
482 2018). Future studies employing complementary structural methods, such as nuclear magnetic
483 resonance, will help confirm the occurrence of non-canonical base states in off-target
484 complexes.

485 The mismatch tolerance of Cas9 can be explained primarily by two factors. Firstly, Cas9
486 does not directly contact the major- or minor-groove edges of the guide RNA–TS DNA
487 heteroduplex base pairs at any of the duplex positions and thus lacks a steric mechanism to
488 enforce Watson-Crick base pairing. This is further underscored by Cas9's tolerance of base
489 modifications in target DNA, including cytosine 5-hydroxymethylation and, at least at some
490 duplex positions, glucosyl-5-hydroxymethylation (Vlot et al., 2018). In this respect, Cas9
491 differs from other molecular systems, notably the ribosome and replicative DNA polymerases,
492 which enhance the specificity of base-pairing by direct readout of base-pair shape and steric
493 rejection of mispairs (Kunkel and Bebenek, 2000; Rodnina and Wintermeyer, 2001; Timsit,
494 1999). Secondly, Cas9 is a multidomain protein that displays considerable conformational
495 dynamics and is therefore able to accommodate local distortions in the guide–TS duplex
496 geometry by compensatory rearrangements of the REC2, REC3 and HNH domains. Indeed, in
497 most off-target structures reported in this study, almost all atomic contacts between Cas9 and
498 the guide–TS heteroduplex are preserved. Thus, Cas9 only detects guide-target mismatches by
499 indirect readout of the guide RNA–TS DNA heteroduplex width, except at the PAM-distal end
500 of the heteroduplex where base mismatches result in duplex unpairing, as discussed below. Our
501 observations are consistent with recent molecular dynamics simulation studies showing that
502 internally positioned mismatches within the guide RNA–TS DNA heteroduplex are readily
503 incorporated within the heteroduplex and have only minor effects on Cas9 interactions

504 (Mitchell et al., 2020). The lack of a steric base-pair enforcement mechanism and the resulting
505 off-target promiscuity likely reflects the biological function of Cas9 in CRISPR immunity,
506 where mismatch tolerance contributes to interference by enabling the targeting of closely
507 related viruses and hindering immune evasion by mutations or covalent base modifications
508 (Deveau et al., 2008; Semenova et al., 2011; van Houte et al., 2016; Yaung et al., 2014).

509 **Structural rigidity of the guide RNA seed region and implications for off-target**
510 **recognition**

511 The seed sequence of the Cas9 guide RNA (nucleotides 11-20) is the primary determinant of
512 target DNA binding, a consequence of its structural pre-ordering in an A-like conformation by
513 extensive interactions with Cas9 (Anders et al., 2014; Jiang et al., 2015; Nishimasu et al., 2014;
514 Zhu et al., 2019). Our data indicate that structural rigidity of the guide RNA seed sequence also
515 affects off-target recognition, as base mispairs in the seed region of the guide–off-target
516 heteroduplex can only be accommodated by conformational distortions of the TS DNA, which
517 is subject to only a few steric constraints, notably at position 20 due to interactions with the
518 phosphate lock loop (Anders et al., 2014). The rigidity of the guide RNA seed sequence
519 increases the energetic penalty of base mispairing in the seed region of the heteroduplex, and
520 thus contributes to mismatch sensitivity of Cas9 within the seed region. Although structural
521 distortions of TS DNA facilitate biding of off-target substrates containing seed mismatches,
522 they may nevertheless inhibit off-target cleavage by steric hindrance of the HNH domain,
523 thereby further contributing to the general mismatch intolerance of the guide RNA seed
524 sequence. The contrasting structural plasticities of the guide RNA and TS DNA strands are
525 manifested in the differential activities of Cas9 against off-targets containing rU-dG and rG-dT
526 mismatches within the seed region (Boyle et al., 2021; Doench et al., 2016; Hsu et al., 2013;
527 Jones et al., 2020; Zhang et al., 2020). Whereas rG-dT mismatches can be readily
528 accommodated by wobble base pairing, seed sequence rigidity is expected to hinder rU-dG

529 wobble base pairing. Combined with a lower energetic penalty associated with rG-dT mismatch
530 binding (binding an off-target with an rG-dT mismatch requires unpairing a dT-dA base pair in
531 the off-target DNA, while rU-dG off-target recognition requires dC-dG unpairing), these effects
532 thus help Cas9 discriminate against rU-dG mismatches in the seed region.

533 **Recognition of off-targets containing insertions and deletions**

534 *Bona fide* off-target sites containing insertions or deletions have been detected in a number of
535 studies (Boyle et al., 2021; Cameron et al., 2017; Doench et al., 2016; Jones et al., 2020; Tsai
536 et al., 2015). Nucleotide "bulging" has been proposed as a mechanism to recognize such an off-
537 target, which would otherwise result in a large number of consecutive base mismatches.
538 However, as Cas9 encloses the guide RNA–TS DNA heteroduplex in a central channel and
539 makes extensive interactions along the entire length of the guide RNA strand, the formation of
540 RNA bulges is precluded due to steric clashes, pointing to a different mechanism.

541 Indeed, the structures of *PTPRC-tgt2* off-target #1 and *FANCF* off-target #3 complexes
542 reveal that off-target sequences predicted to contain single-nucleotide deletions in the seed
543 region of the heteroduplex are instead recognized by base skipping, resulting in an unpaired
544 guide RNA base within the duplex stack. Due to the lack of extensive contacts of Cas9 with the
545 TS and the rigid coordination of the guide RNA in the seed region, these findings suggest that
546 single nucleotide deletions can only be accommodated within the seed region of the
547 heteroduplex and not elsewhere. This is supported by the observation that the *AAVS1* off-target
548 #2 site, which was previously predicted to contain an RNA bulge or skip in the PAM-distal
549 region (Cameron et al., 2017; Lazzarotto et al., 2020), is recognised via multiple mismatches.

550 Our structural observations indicate that *bona fide* off-targets predicted to contain single
551 deletions within the seed region of the heteroduplex are recognized by base skipping, which
552 incurs a large energetic penalty. As the seed region of the TS DNA is devoid of Cas9 contacts
553 in the pre-cleavage state (Zhu et al., 2019), off-target sequences containing single-nucleotide

554 insertions in the seed region of the heteroduplex are likely to be recognized by DNA nucleotide
555 bulging, likewise incurring a large energetic penalty as unwinding an off-target DNA sequence
556 containing an insertion requires breaking an extra base pair. Additionally, TS DNA distortion
557 might inhibit cleavage by steric hindrance of the HNH domain. These observations thus explain
558 why Cas9 appears to tolerate mismatches better than insertions or deletions (Boyle et al., 2021;
559 Cameron et al., 2017; Doench et al., 2016; Jones et al., 2020) and why deletions and insertions
560 within the seed region are particularly deleterious. In contrast, off-target sequences containing
561 insertions or deletions in the PAM-distal region of the heteroduplex, where both the guide RNA
562 and TS DNA strands are contacted by Cas9, are instead likely to be bound in the unchanged
563 register, with multiple base mispairs accommodated by non-canonical base pairing interactions.
564 Our analysis suggests that a significant fraction of off-target sites previously predicted to
565 contain insertions or deletions may be recognized in this manner.

566 **PAM-distal base pairing and the conformational checkpoint of Cas9**

567 Upon substrate DNA hybridization and R-loop formation, Cas9 undergoes conformational
568 activation of its nuclease domains (Zhu et al., 2019). The Cas9 REC3 domain plays a key role
569 in the process, as it senses the integrity of the PAM-distal region of the guide RNA–TS DNA
570 heteroduplex and allosterically regulates the REC2 and HNH domains, providing a
571 conformational checkpoint that traps Cas9 in a conformationally inactive state in the absence
572 of PAM-distal hybridization (Chen et al., 2017; Dagdas et al., 2017; Palermo et al., 2018; Zhu
573 et al., 2019). Our structural data confirm that mismatches at the PAM-distal end of the
574 heteroduplex (positions 1-3) result in heteroduplex unpairing, incomplete R-loop formation and
575 structural repositioning of the REC3 domain, indicating a perturbation of the Cas9
576 conformational checkpoint. We envision that the observed conformational state mimics the
577 structural effect of 5'-truncated guide RNAs, which have been shown to improve targeting
578 specificity (Fu et al., 2014). Furthermore, similarities with the structure of a high-fidelity Cas9

579 variant (Guo et al., 2019) suggest a shared underlying mechanism for increased specificity. In
580 both cases, disruption of REC3 contacts with the PAM-distal heteroduplex modulates REC2/3
581 domain positioning, hindering allosteric activation of the HNH nuclease domain (Chen et al.,
582 2017; Dagdas et al., 2017; Palermo et al., 2018). This is also consistent with observations that
583 REC2/3 domain repositioning in Cas9 complexes with chimeric RNA-DNA guides modulates
584 cleavage efficiency and results in increased specificity by slowing down conformational
585 nuclease activation and promoting substrate DNA dissociation (Donohoue et al., 2021). In
586 addition, the establishment of new HNH protein contacts with the heteroduplex, as observed in
587 *FANCF* off-target #4, has been proposed to affect the active site positioning of the HNH domain
588 (Mitchell et al., 2020; Ricci et al., 2019; Zeng et al., 2018). Indeed, it has been demonstrated
589 that truncated guides result in reduced cleavage rates due to impaired HNH docking (Dagdas et
590 al., 2017).

591 **Implications for off-target prediction**

592 Our structural data reveal that Cas9 plays a limited steric role in off-target discrimination insofar
593 as only sensing the integrity and general shape of the guide–target heteroduplex. Off-target
594 activity is thus largely determined by the kinetics and energetics of R-loop formation, *i.e.,* off-
595 target DNA strand separation and guide RNA–TS DNA hybridization, and the Cas9
596 conformational activation checkpoint. We observe on multiple occasions in the determined off-
597 target complexes that a given base mismatch adopts different conformational arrangements
598 depending on its position along the guide RNA–TS DNA heteroduplex. This poses a challenge
599 for *ab initio* modelling of off-target activity, as biophysical models of off-target binding and
600 cleavage are bound to be of limited accuracy unless they incorporate position-dependent
601 energetic penalties for each base mismatch type and for deletions, as well as position- and base-
602 specific penalties for insertions (Boyle et al., 2021; Jones et al., 2020; Zhang et al., 2020). In
603 addition, as certain off-target sequences that are incompatible with dsDNA cleavage can

604 undergo NTS nicking (Fu et al., 2019; Jones et al., 2020; Murugan et al., 2020; Zeng et al.,
605 2018), future bioinformatic models need to be able to predict off-target nicking activity as well.
606 Furthermore, accurate modelling of off-target interactions remains difficult due to context-
607 dependent effects, as documented in previous studies showing that the binding and cleavage
608 defects of consecutive mismatches deviate from additivity (Boyle et al., 2021; Cameron et al.,
609 2017; Lazzarotto et al., 2020; Zhang et al., 2020). Indeed, our structural data rationalize this by
610 showing that the conformation of a given base mismatch is highly sensitive to the presence of
611 neighbouring mismatches. As seen in the case of *AAVS1* off-target #2 complex, multiple
612 mismatched bases can synergistically combine to preserve an on-target-like heteroduplex
613 conformation that passes the REC3 conformational checkpoint, supporting nearly on-target
614 efficiencies of cleavage (Zhang et al., 2020). This is in line with recent cryo-EM structural
615 studies suggesting that indirect readout of heteroduplex conformation is coupled to nuclease
616 activation, while the presence of mismatches disrupts this coupling (Bravo et al., 2021; Pacesa
617 and Jinek, 2021). Critically, reversion of one of the mismatches in this off-target substrate
618 impairs cleavage activity. Similar effects have been described for other DNA binding proteins
619 such as transcription factors, where mismatches modulate the binding activity of the protein by
620 affecting the conformation of the DNA duplex (Afek et al., 2020). In an analogy with Cas9,
621 these proteins check for correct binding sites through indirect sequence readout by sampling
622 for the correct duplex shape rather than base sequence (Abe et al., 2015; Kitayner et al., 2010;
623 Rohs et al., 2009a; Rohs et al., 2009b).

624 In conclusion, structural insights presented in this study establish an initial framework
625 for understanding the molecular basis for the off-target activity of Cas9. In conjunction with
626 ongoing computational studies, these findings will help achieve improved energetic
627 parametrization of off-target mismatches and deletions/insertions, thus contributing to the
628 development of more accurate off-target prediction algorithms and more specific guide RNA

629 designs. In doing so, these studies will contribute towards increasing the precision of CRISPR-

630 Cas9 genome editing and the safety of its therapeutic applications.

631

632 ## Author contributions

633 M.P., P.C., P.D.D., and M.J. conceived the study. M.P. purified wild-type Cas9, performed *in*
634 *vitro* cleavage assays, crystallized ternary Cas9 complexes, solved the structures, and
635 performed structural analysis along with M.J.; A.C. performed switchSENSE binding
636 measurements, under the supervision of F.H.T.A; M.J.I. performed the SITE-Seq assay; C-H.L.
637 wrote the computational off-target classification model and P.D.D. and P.C. analysed the
638 output; K.B. purified dCas9, transcribed sgRNAs, and prepared DNA substrates for in vitro
639 cleavage assays; M.P., F.H.T.A., P.C., P.D.D., and M.J. wrote the manuscript.

640

641 ## Conflict of interest statement

642 P.D.D. and M.J.I are current employees of Caribou Biosciences, Inc., and C-H.L. and P.C. are
643 former employees of Caribou Biosciences, Inc. M.J. is a co-founder of Caribou Biosciences,
644 Inc. M.J., C-H.L., M.J.I., P.C. and P.D.D. are named inventors on patents and patent
645 applications related to CRISPR-Cas technologies.

646

647 ## Acknowledgements

648 We thank members of the Jinek laboratory and Caribou Biosciences, Inc. for discussion and
649 critical reading of the manuscript. We thank Vincent Olieric, Meitian Wang, and Takashi
650 Tomizaki (Swiss Light Source, Paul Scherrer Institute) for assistance with crystallographic data
651 collection and Nena Matscheko from Dynamic Biosensors for support with switchSENSE
652 experiments. We thank the NCCR RNA and Disease for providing infrastructural support. This
653 work was supported by the Swiss National Science Foundation Grant 31003A_182567 (to
654 M.J.). M.J. is an International Research Scholar of the Howard Hughes Medical Institute and
655 Vallee Scholar of the Bert L & N Kuggie Vallee Foundation.

Figures and Legends

Figure 1

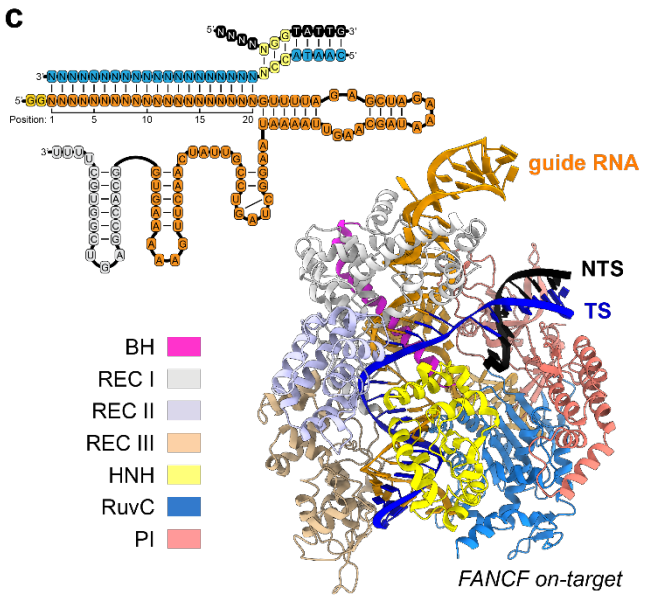
a

Gene	Sequence	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	P	A	M
AAVS1	guide RNA	G	G	G	G	G	C	C	A	C	U	A	G	G	G	A	C	A	G	G	A	U			
AAVS1	on-target		C	C	C	C	G	G	T	G	A	T	C	C	C	T	G	T	C	C	T	A	A	C	C
AAVS1	off-target1		·	·	·	T	·	·	·	A	G	·	·	·	·	·	·	·	·	·	·	·	C	C	C
AAVS1	off-target2		A	·	·	·	C	·	G	T	G	·	·	·	·	·	·	·	·	·	A	·	A	C	C
AAVS1	off-target3		·	·	·	·	·	·	·	C	·	·	·	·	·	·	C	·	·	·	·	·	T	C	C
AAVS1	off-target4		A	·	·	·	·	·	·	·	·	·	A	·	·	·	·	·	·	·	·	T	A	C	C
AAVS1	off-target5		·	·	·	T	·	·	·	·	·	C	·	·	·	·	·	·	·	·	G	G	A	C	C
FANCF	guide RNA	G	G	G	A	A	U	C	C	C	U	U	C	U	G	C	A	G	C	A	C	C			
FANCF	on-target		C	C	T	T	A	G	G	G	A	A	G	A	C	G	T	C	G	T	G	G	A	C	C
FANCF	off-target1		·	·	·	·	G	·	·	·	C	·	·	·	·	·	·	·	·	·	·	·	T	C	C
FANCF	off-target2		·	·	·	C	·	·	·	·	·	G	·	·	T	·	·	·	·	·	·	·	T	C	C
FANCF	off-target3	G	·	·	·	C	G	·	·	·	·	·	·	·	·	·	·	·	–	·	·	·	G	C	C
FANCF	off-target4		T	G	G	·	·	·	·	·	·	G	·	·	·	·	·	·	·	·	·	·	A	C	C
FANCF	off-target5		A	·	·	·	·	·	·	·	T	·	·	·	G	·	·	·	·	·	·	·	A	C	C
FANCF	off-target6		·	·	·	·	G	T	·	·	·	·	·	·	·	·	·	·	·	A	·	·	T	C	C
FANCF	off-target7		·	·	·	·	·	·	T	·	·	·	A	·	T	·	·	·	·	·	·	·	T	C	C
PTPRC	guide RNA	G	G	C	A	A	A	A	C	U	C	A	A	C	C	C	U	A	C	C	C	C			
PTPRC	on-target		C	G	T	T	T	T	G	A	G	T	T	G	G	G	A	T	G	G	G	G	A	C	C
PTPRC	off-target1	T	·	·	·	·	A	·	·	T	·	·	·	·	·	·	–	·	·	·	·	·	A	C	C
TRAC	guide RNA	G	A	G	A	G	U	C	U	C	U	C	A	G	C	U	G	G	U	A	C	A			
TRAC	on-target		T	C	T	C	A	G	A	G	A	G	T	C	G	A	C	C	A	T	G	T	G	C	C
TRAC	off-target1		C	T	·	·	·	·	·	·	G	·	·	·	·	·	T	·	·	·	·	·	A	C	C
TRAC	off-target2		·	T	·	·	·	·	T	·	T	·	·	·	·	·	·	·	·	·	·	·	A	C	C

b

Gene	Target	(%) 24h cleavage	k_{obs} (s^{-1})	k_{on} ($M^{-1} \cdot s^{-1}$)	k_{off} (s^{-1})	K_d (pM)	24h cleavage PAMmer (%)	k_{obs} (s^{-1}) PAMmer
AAVS1	on-target	92.2	1.6240	3.95E+06	5.74E-05	14.5	95.0	0.5622
AAVS1	off-target1	92.7	0.0713	4.73E+06	6.29E-05	13.3	87.5	0.2390
AAVS1	off-target2	95.1	0.0337	8.76E+06	3.38E-03	386.0	85.5	0.0040
AAVS1	off-target3	96.7	0.0511	3.30E+06	2.51E-03	761.0	82.7	0.0806
AAVS1	off-target4	94.4	0.0039	1.09E+07	3.28E-03	301.0	89.6	0.0645
AAVS1	off-target5	18.9	0.0001	ND	ND	ND	70.1	0.0068
FANCF	on-target	97.5	0.2383	3.45E+06	7.46E-05	21.6	98.3	0.5654
FANCF	off-target1	35.1	0.0006	3.97E+06	2.09E-03	528.0	97.4	0.0693
FANCF	off-target2	62.4	0.0006	1.42E+06	2.45E-03	1730.0	92.9	0.2333
FANCF	off-target3	0.0	0.0000	1.22E+07	2.37E-03	193.0	4.2	0.0000
FANCF	off-target4	53.0	0.0005	3.35E+06	1.91E-03	571.0	38.9	0.0011
FANCF	off-target5	80.4	0.0010	1.27E+06	2.55E-03	2010.0	92.9	0.0584
FANCF	off-target6	8.2	0.0001	1.50E+06	2.03E-03	1350.0	66.6	0.0007
FANCF	off-target7	5.2	0.0036	2.95E+06	3.21E-03	1090.0	94.5	0.0134
PTPRC	on-target	96.8	0.4588	6.08E+06	2.19E-04	36.0	95.5	0.0741
PTPRC	off-target1	0.0	0.0000	1.22E+07	2.39E-03	196.0	91.4	0.0012
TRAC	on-target	97.7	0.3808	1.02E+07	3.23E-04	31.8	93.5	0.1812
TRAC	off-target1	95.8	0.0195	1.37E+06	1.77E-04	130.0	90.7	0.0807
TRAC	off-target2	65.0	0.0007	9.43E+06	3.27E-04	34.6	88.4	0.0260

c



658 **Figure 1. Biochemical and structural analysis of Cas9 off-targets.**

659 (**A**) Guide RNA and (off-)target DNA sequences selected for biochemical and structural
660 analysis. Matching bases in off-targets are denoted by a dot; nucleotide mismatches and
661 deletions (–) are highlighted. (**B**) Kinetic and thermodynamic parameters of off-target
662 substrates. The cleavage rate constants (k_{obs}) were derived from single-exponential function
663 fitting of measured cleavage rates. The binding and dissociation rate constants (k_{on} and k_{off}) and
664 the equilibrium dissociation constant (K_d) were determined using a DNA nanolever binding
665 (switchSENSE) assay. (**C**) Top: Schematic representation of the guide RNA (orange), TS
666 (blue), and NTS (black) sequences used for crystallisation. The PAM sequence in the DNA is
667 highlighted in yellow. Bottom: Structure of the Cas9 *FANCF* on-target complex. Individual
668 Cas9 domains are coloured according to the legend; substrate DNA target strand (TS) is
669 coloured blue, non-target strand (NTS) black, and the guide RNA orange.

670

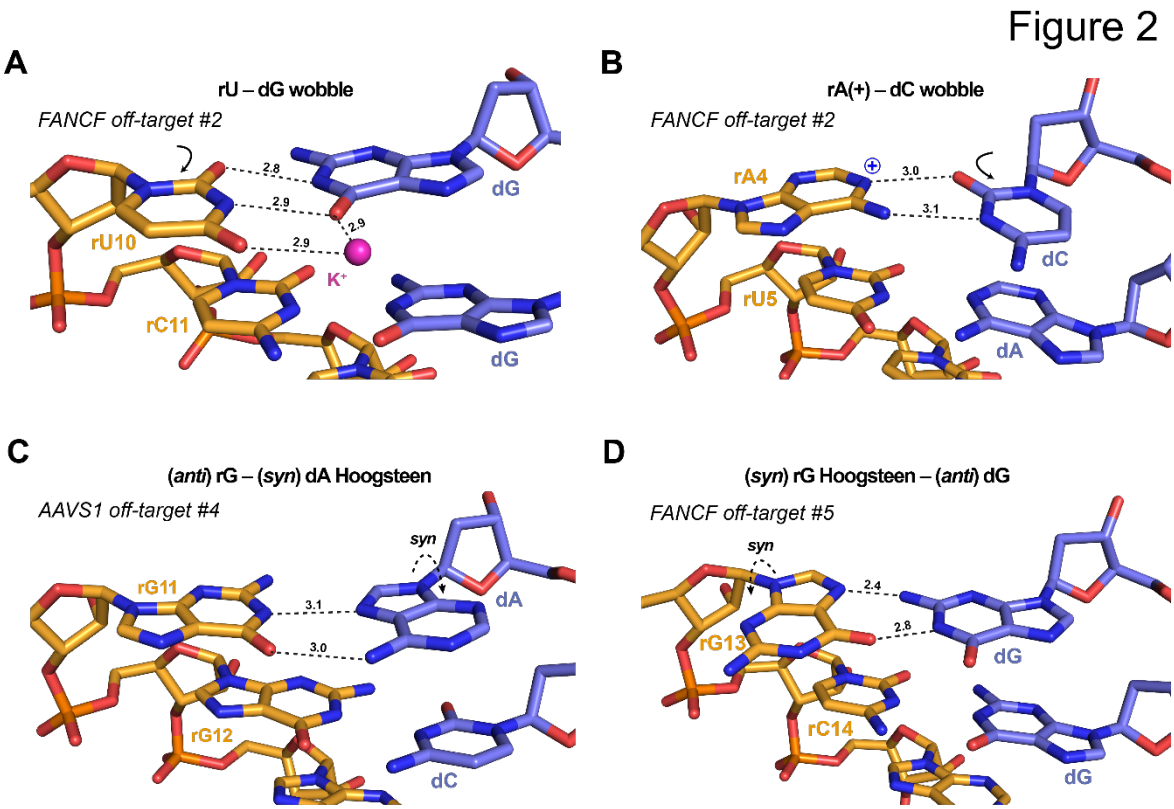
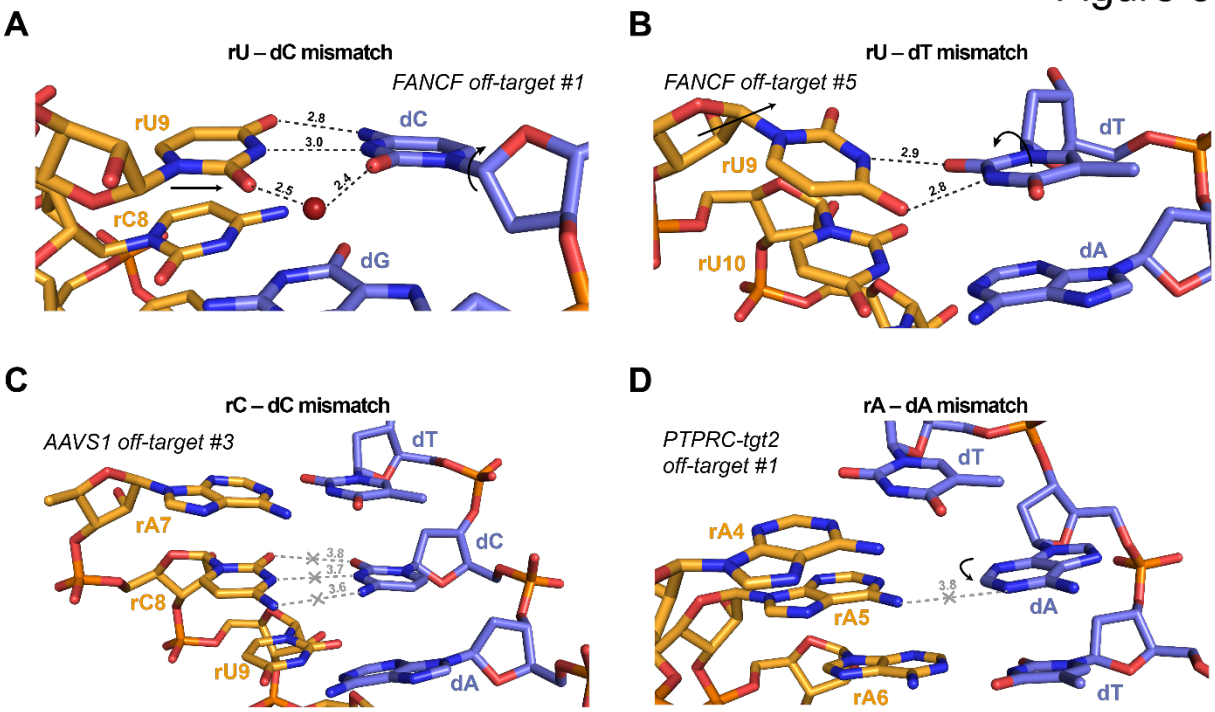


Figure 2. Cas9 off-target binding is enabled by non-canonical base pairing. Close-up views of (**A**) rU-dG wobble base pair at duplex position 10 in *FANCF* off-target #2 complex, (**B**) rA-dC wobble base pair at position 4 in *FANCF* off-target #2 complex, (**C**) rG-dA Hoogsteen base pair at duplex position 11 in *AAVS1* off-target #4 complex and (**D**) rG-dG Hoogsteen base pair at duplex position 13 in *FANCF* off-target #5 complex. Hydrogen bonding interactions are indicated with dashed lines. Numbers indicate interatomic distances in Å. Solid arrows indicate conformational changes relative to the corresponding on-target complex structures. Dashed arrows indicate *anti-syn* isomerization of the dA and rG bases to enable Hoogsteen-edge base pairing. A bound monovalent ion, modelled as K^+, is depicted as a purple sphere.

Figure 3



683

684 **Figure 3. Duplex backbone distortions facilitate formation of non-canonical base pairs.**

685 (**A**) Close-up view of the rU-dC base pair at duplex position 9 in *FANCF* off-target #1 complex,

686 facilitated by lateral displacement of the guide RNA backbone (solid arrow). Hydrogen bonding

687 interactions are indicated with dashed lines. Solid arrows indicate conformational changes

688 relative to the on-target complex. Numbers indicate interatomic distances in Å. Bound water

689 molecule is depicted as red sphere. (**B**) Zoomed-in view of the rU-dT base pair at position 9 in

690 *FANCF* off-target #5 complex. Solid arrows indicate lateral displacement of the rU nucleotide

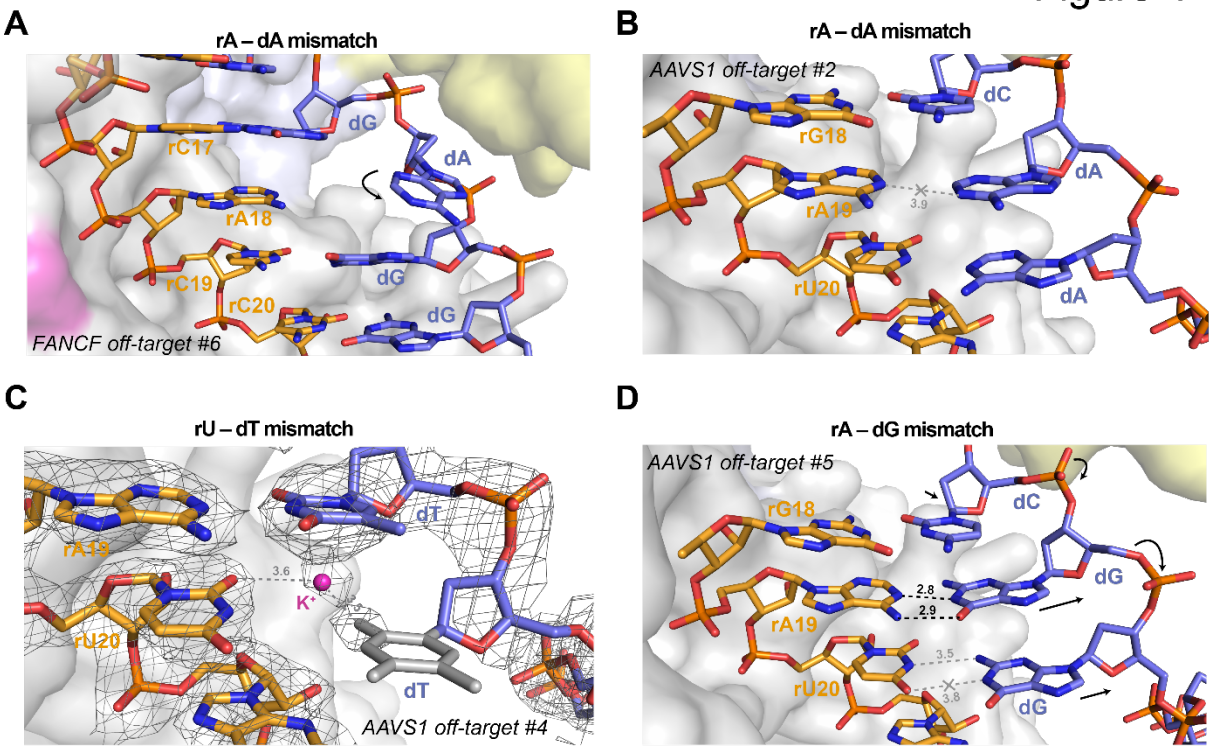
691 and propeller twist of the dT base. (**C**) Zoomed-in view of the rC-dC mispair at duplex position

692 8 in *AAVS1* off-target #3 complex. The distances between the cytosine bases indicate lack of

693 hydrogen bonding. (**D**) Zoomed-in view of the rA-dA mispair at duplex position 5 in *PTPRC-*

694 *tgt2* off-target #1 complex.

Figure 4



695

696 **Figure 4. TS distortion facilitates mismatch accommodation in the seed region of the**

697 **guide–off-target heteroduplex.**

698 (**A**) Close-up view of the rA-dA mismatch at position 18 in *FANCF* off-target #6 complex,

699 showing major groove extrusion of the dA base. (**B**) Close-up view of the rA-dA mismatch at

700 position 19 in *AAVS1* off-target #2 complex, showing retention of the dA base in the duplex

701 stack. (**C**) Close-up rU-dT mispair at the PAM-proximal position 20 in *AAVS1* off-target #4

702 complex. Residual electron density indicates the presence of an ion or solvent molecule.

703 Refined 2mF_o−DF_c electron density map of the heteroduplex, contoured at 1.5σ, is rendered as

704 a grey mesh. Structurally disordered thymine nucleobase for which no unambiguous density is

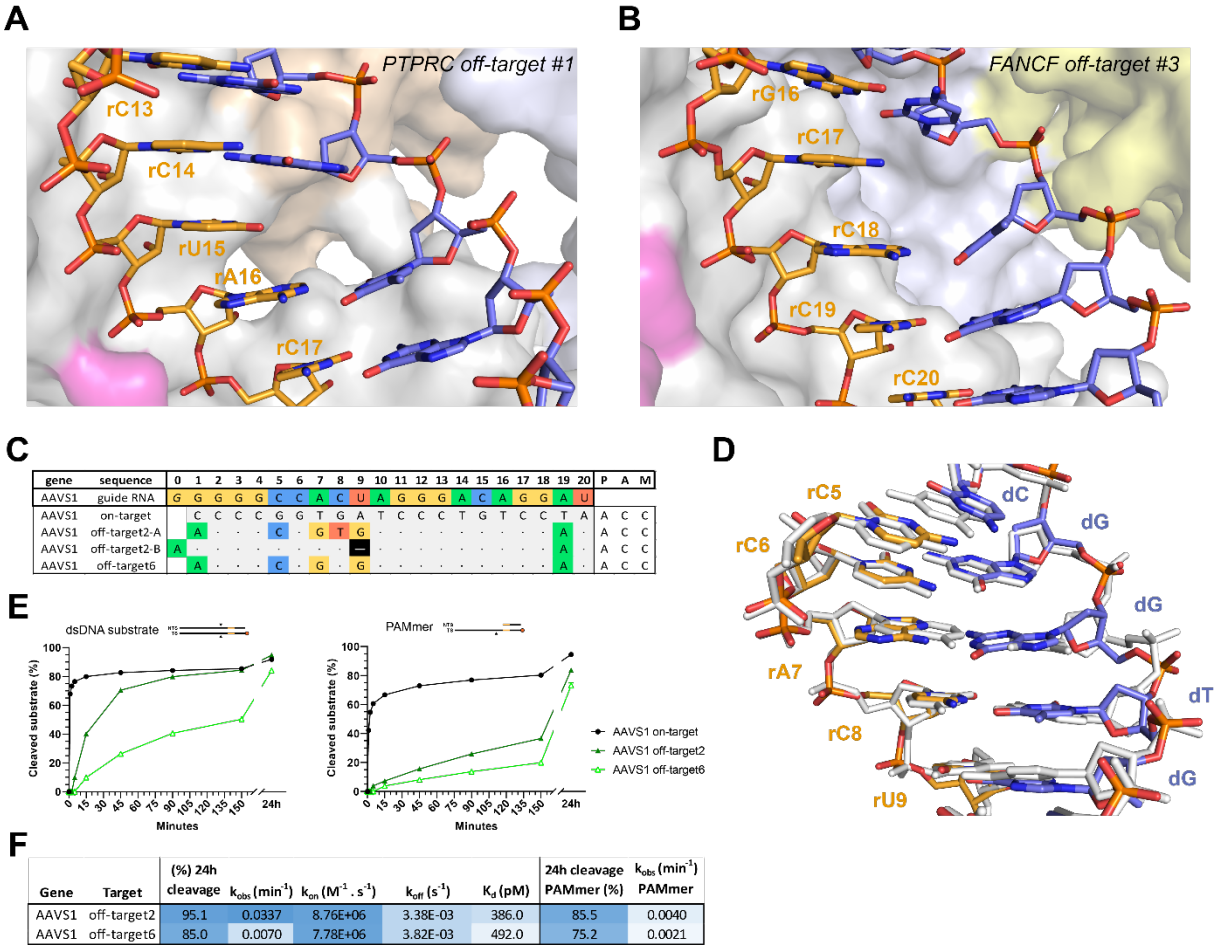
705 present is in grey. (**D**) Zoomed-in view of the rA-dG base pair at position 19 and the unpaired

706 rU-dG mismatch at position 20 in *AAVS1* off-target #5 complex. Arrows indicate

707 conformational changes in the TS backbone relative to the on-target complex.

708

Figure 5



gene	sequence	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	P	A	M
AAVS1	guide RNA	G	G	G	G	G	C	C	A	C	U	A	G	G	G	A	C	A	G	G	A	U			
AAVS1	on-target		C	C	C	C	G	G	T	G	A	T	C	C	C	T	G	T	C	C	T	A	A	C	C
AAVS1	off-target2-A		A	·	·	·	C	·	G	T	G	·	·	·	·	·	·	·	·	·	A	·	A	C	C
AAVS1	off-target2-B	A	·	·	·	·	·	·	·	·	–	·	·	·	·	·	·	·	·	·	A	·	A	C	C
AAVS1	off-target6		A	·	·	·	C	·	G	·	G	·	·	·	·	·	·	·	·	·	A	·	A	C	C

Gene	Target	(%) 24h cleavage	k_{obs} (min^{-1})	k_{on} ($M^{-1} . s^{-1}$)	k_{off} (s^{-1})	K_d (pM)	24h cleavage PAMmer (%)	k_{obs} (min^{-1}) PAMmer
AAVS1	off-target2	95.1	0.0337	8.76E+06	3.38E-03	386.0	85.5	0.0040
AAVS1	off-target6	85.0	0.0070	7.78E+06	3.82E-03	492.0	75.2	0.0021

Figure 5. Off-targets with single-nucleotide deletions are accommodated by base skipping or multiple consecutive mismatches.

(**A**) Zoomed-in view of the base skip at duplex position 15 in the *PTPRC-tgt2* off-target #1 complex. (**B**) Zoomed-in view of the base skip at duplex position 17 in the *FANCF* off-target #3 complex. (**C**) Schematic depiction of alternative base pairing interactions in the *AAVS1* off-target #2 complex. *AAVS1* off-target #6 substrate was designed based on the *AAVS1* off-target #2, with the reversal of a single mismatch in the consecutive region back to the corresponding canonical base pair. (**D**) Structural overlay of the *AAVS1* off-target #2 (coloured) and *AAVS1* on-target (white) heteroduplexes. (**E**) Cleavage DNA kinetics of AAVS1 on-target, off-target #2 and off-target #6 substrates. (**F**) Kinetic and thermodynamic parameters determined for *AAVS1* off-target #2 and #6 substrates. The apparent cleavage rate constants (k_{obs}) were derived

721 from a single-exponential function fitting of measured cleavage. Substrate binding (k_{on}) and

722 dissociation (k_{off}) constants were determined using SwitchSENSE assay.

723

Figure 6

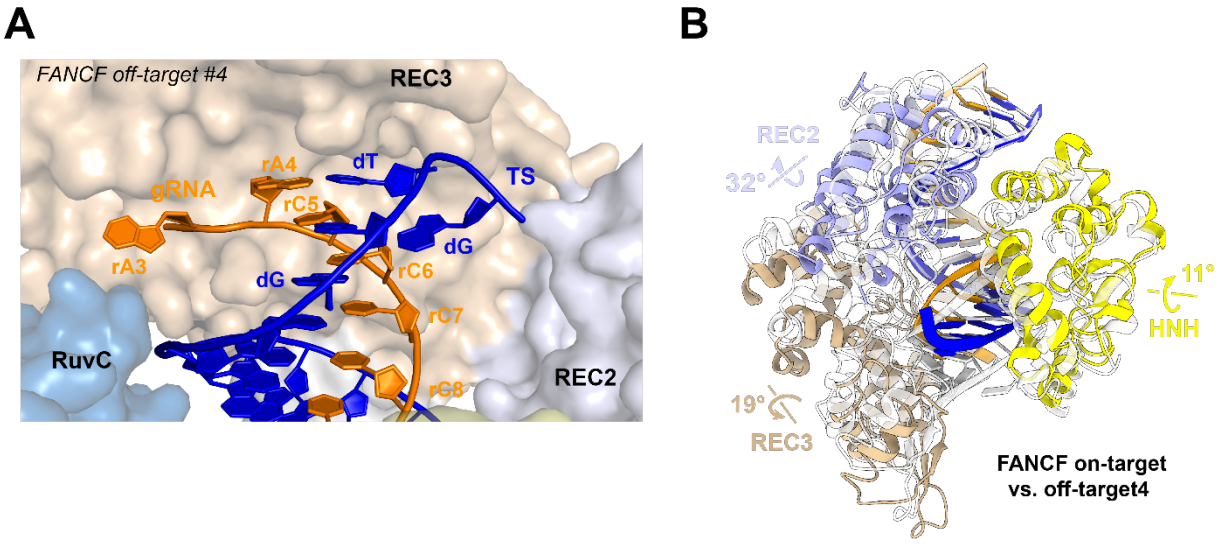


Figure 6. *FANCF* off-target #4 exhibits conformational changes in the REC2/3 and HNH domains due to PAM-distal duplex unpairing.

(**A**) Close-up view of the unpairing of mismatched bases at the PAM-distal end of the *FANCF* off-target #4 heteroduplex. The last two nucleotides on each strand could not be modelled due to structural disorder. (**B**) Overlay of the *FANCF* off-target #4 and *FANCF* on-target complex structures. The *FANCF* off-target #4 complex is coloured according to the domain legend in **Figure 1A**, *FANCF* on-target complex is shown in white. The REC1, RuvC, and PAM-interaction domains have been omitted for clarity, as no structural differences are observed.

Figure S1

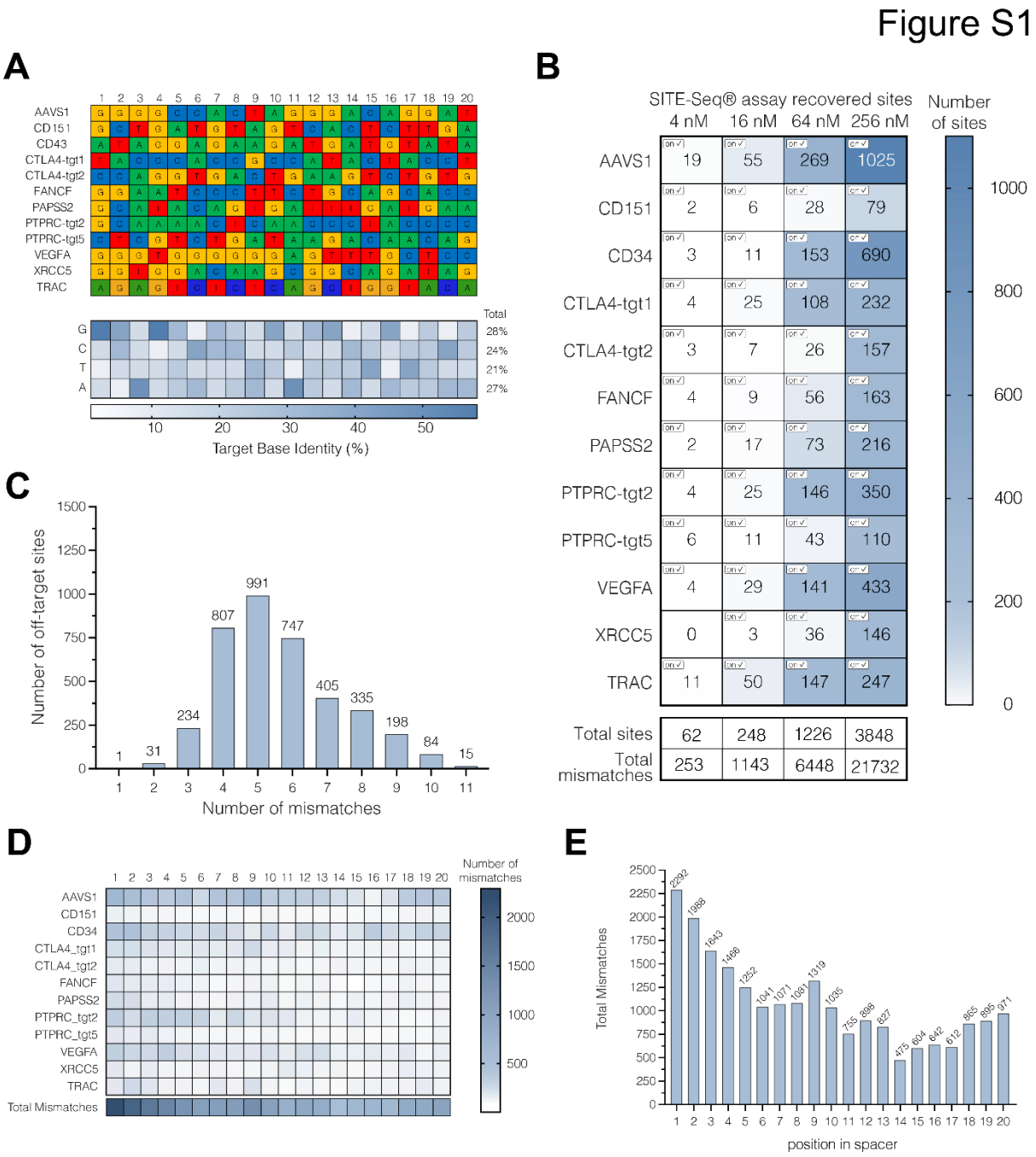


Figure S1. Off-target profiling of selected genomic sites using SITE-Seq.

(**A**) Selected genomic targets and the corresponding guide RNA sequences selected for the SITE-Seq assay off-target profiling. Heatmap indicates frequency of nucleotide identity across each position for the selected targets. (**B**) SITE-seq assay analysis for RNPs assembled with indicated crRNAs. The numbers of detected off-target sites are shown as a function of RNP concentration. Checked boxes indicate recovery of the on-target site. n=3 replicates per sample. (**C**) Number of off-target sites recovered by the SITE-Seq assay are shown as a function of the

741 number of mismatches between the guide RNA and the off-target sequence. (**D**) Frequency of
742 nucleotide mismatches at each guide RNA–off-target DNA heteroduplex position for all off-
743 target sites identified in (B). (**E**) Number of total identified mismatches per heteroduplex
744 position.
745

Figure S2

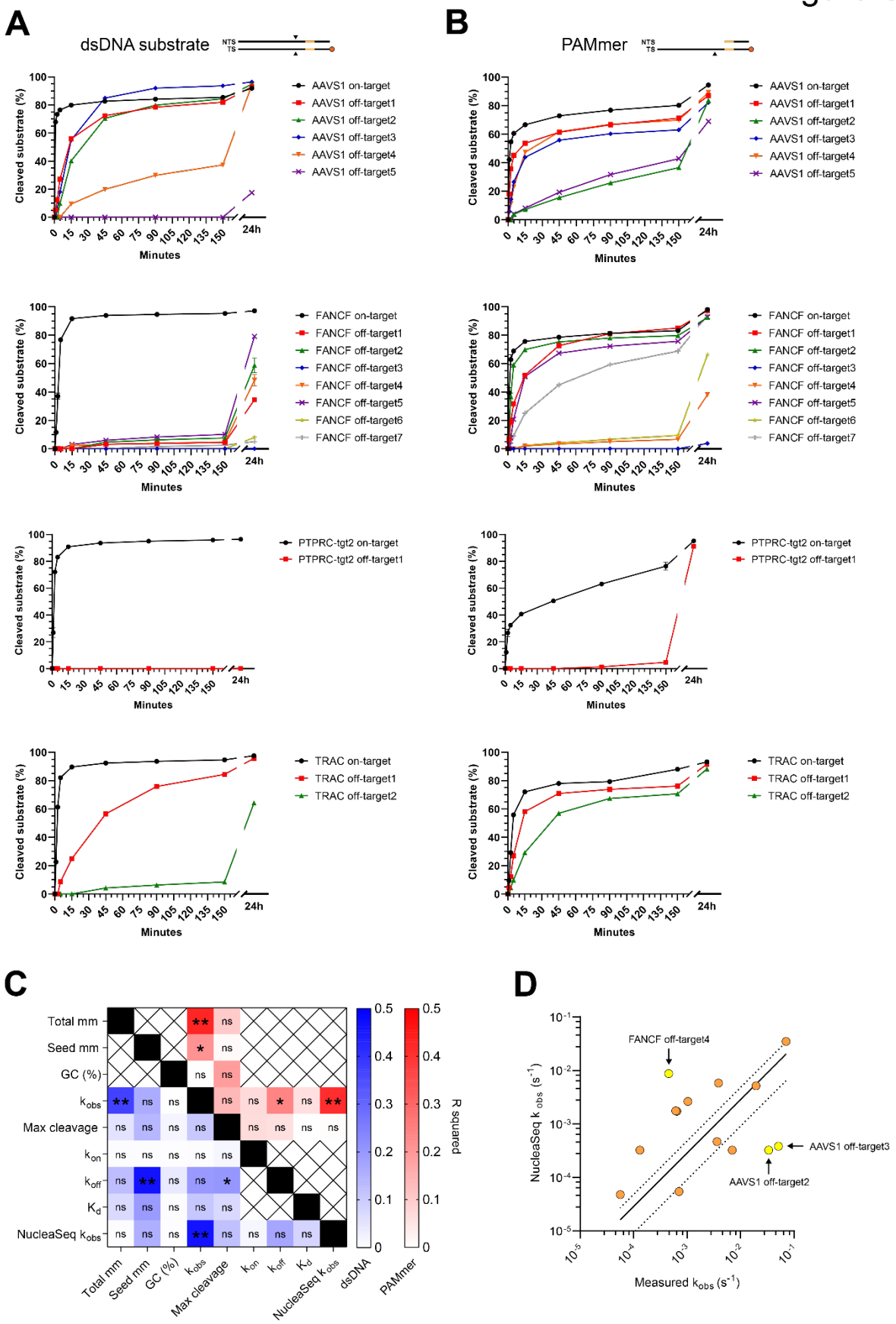
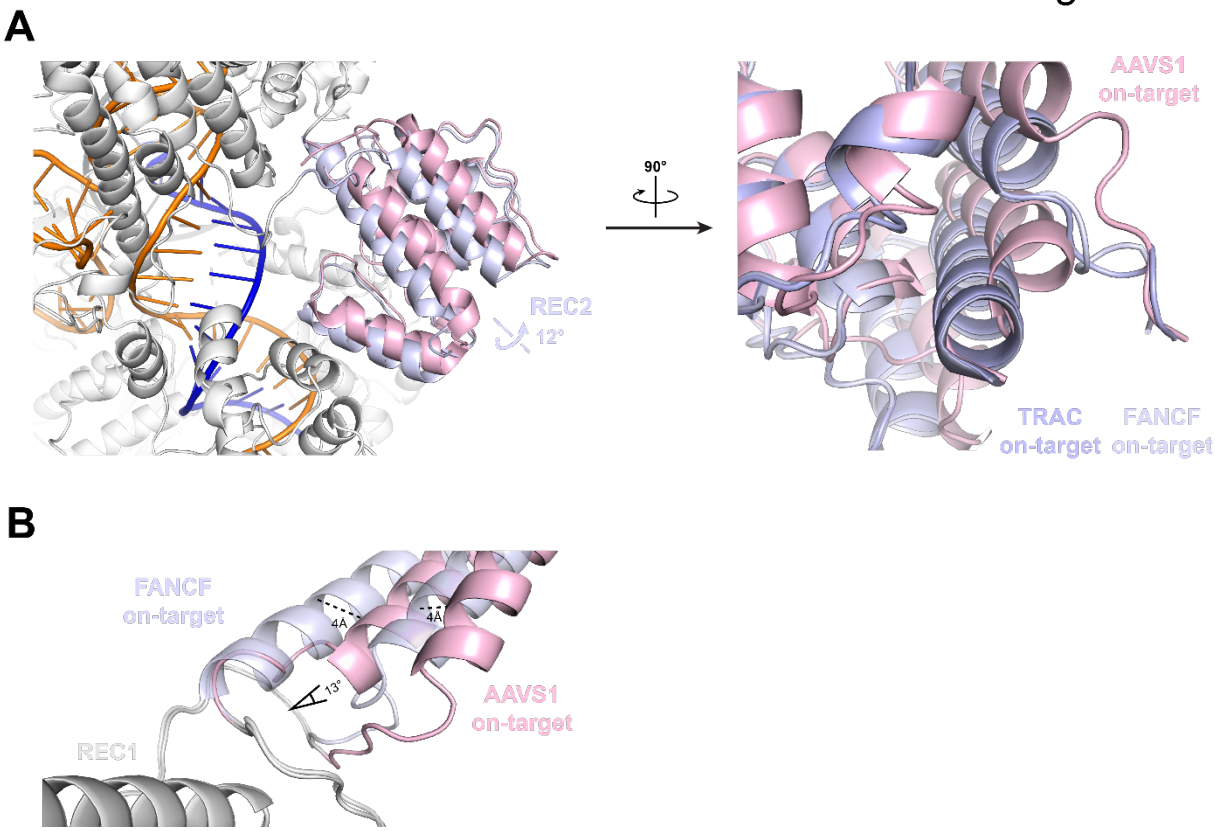


Figure S2. ***in vitro*** **cleavage of selected Cas9 off-target substrates.**

748 (**A**) *In vitro* clevage kinetics of fully double stranded on- and off-target DNA substrates for each
749 guide RNA used in the study. Black triangles in the substrate schematic (top) indicate position
750 of cleavage sites. Each data point represents a mean of four independent replicates. Error bars
751 represent standard deviation for each time point. (**B**) *In vitro* clevage kinetics of partially single
752 stranded (PAMmer) on- and off-target substrates. (**C**) Heatmap representation of mutual
753 correlations between measured kinetic and thermodymamic parameters including cleavage
754 (k_{obs}), substrate DNA binding (k_{on}), substrate dissociation (k_{off}) rate constants, equilibrium
755 dissociation constant (K_d) with numbers of nucleotide mismatches in the off- target sites (total
756 and within seed), the GC content of the spacer (%GC) and cleavage rate predicted using the
757 NucleaSeq algorithm (NucleaSeq k_{obs}). The values were calculated across all off-targets for
758 both dsDNA (lower left half, in blue), and partially single stranded (PAMmer) substrates (upper
759 right half, in red). ns, no significant correlation. (**D**) Correlation between measured and
760 NucleaSeq-predicted k_{obs} rate constants. Off-target sites with significant deviations are
761 highlighted in yellow.
762

Figure S3



763

764 **Figure S3. Alternative REC2 conformation in AAVS1 on-target.**

765 (**A**) Overlay of REC2 domain conformations in the *AAVS1* (pink), *FANCF* (purple) and *TRAC*

766 (light blue) on-target complexes (**B**) Close-up view of helix REC2 helix spanning Cas9 residues

767 174-180. Linear and angular displacements of the helix in the *AAVS1* on-target complex relative

768 to the *FANCF* and *TRAC* on-target complexes are indicated.

769

770

Figure S4

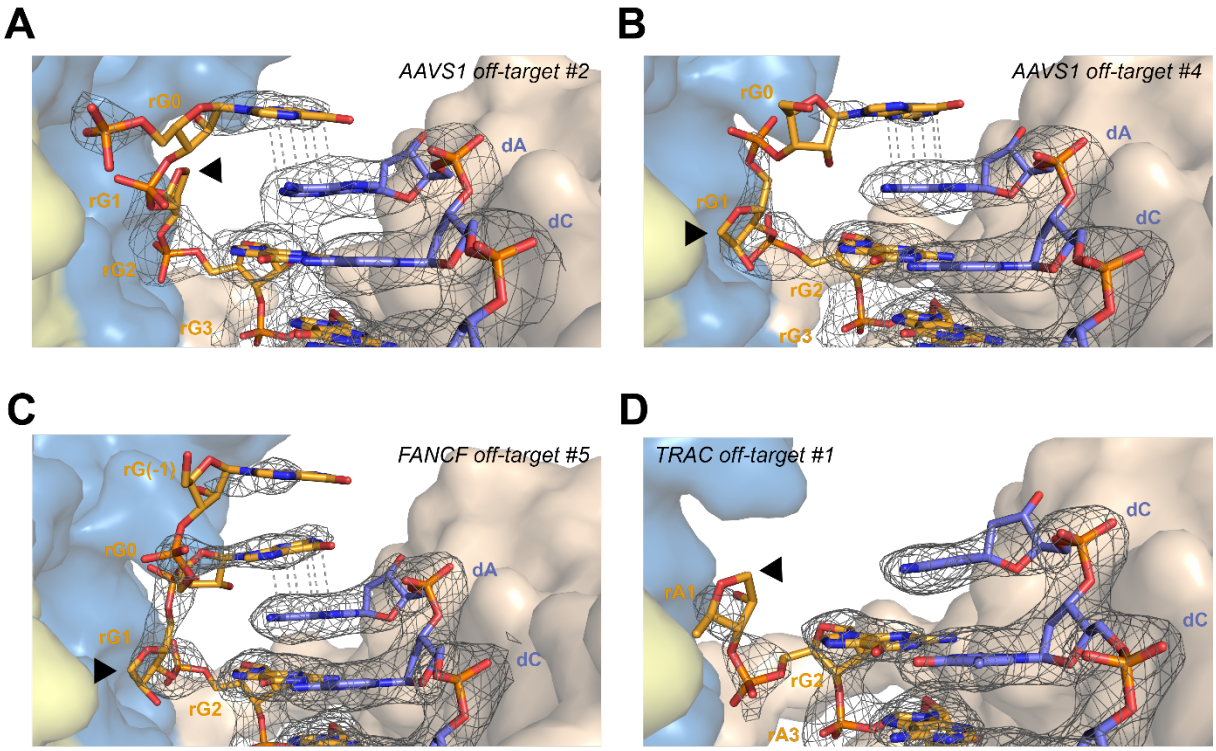


Figure S4. PAM-distal mismatches result in unpairing and disordering of guide RNA nucleobase in position 1.

Close-up views of the PAM-distal end of the guide RNA-TS heteroduplex in (**A**) *AAVS1* off-target #2, (**B**) *AAVS1* off-target #4, (**C**) *FANCF* off-target #5 and (**D**) *TRAC* off-target #1 complexes. Arrowheads indicate nucleotides with disordered bases. Refined $2mF_o$–DF_c electron density maps of the heteroduplexes are rendered as a grey mesh and contoured at 1.2σ for (A) and 1.0σ for (B)-(D).

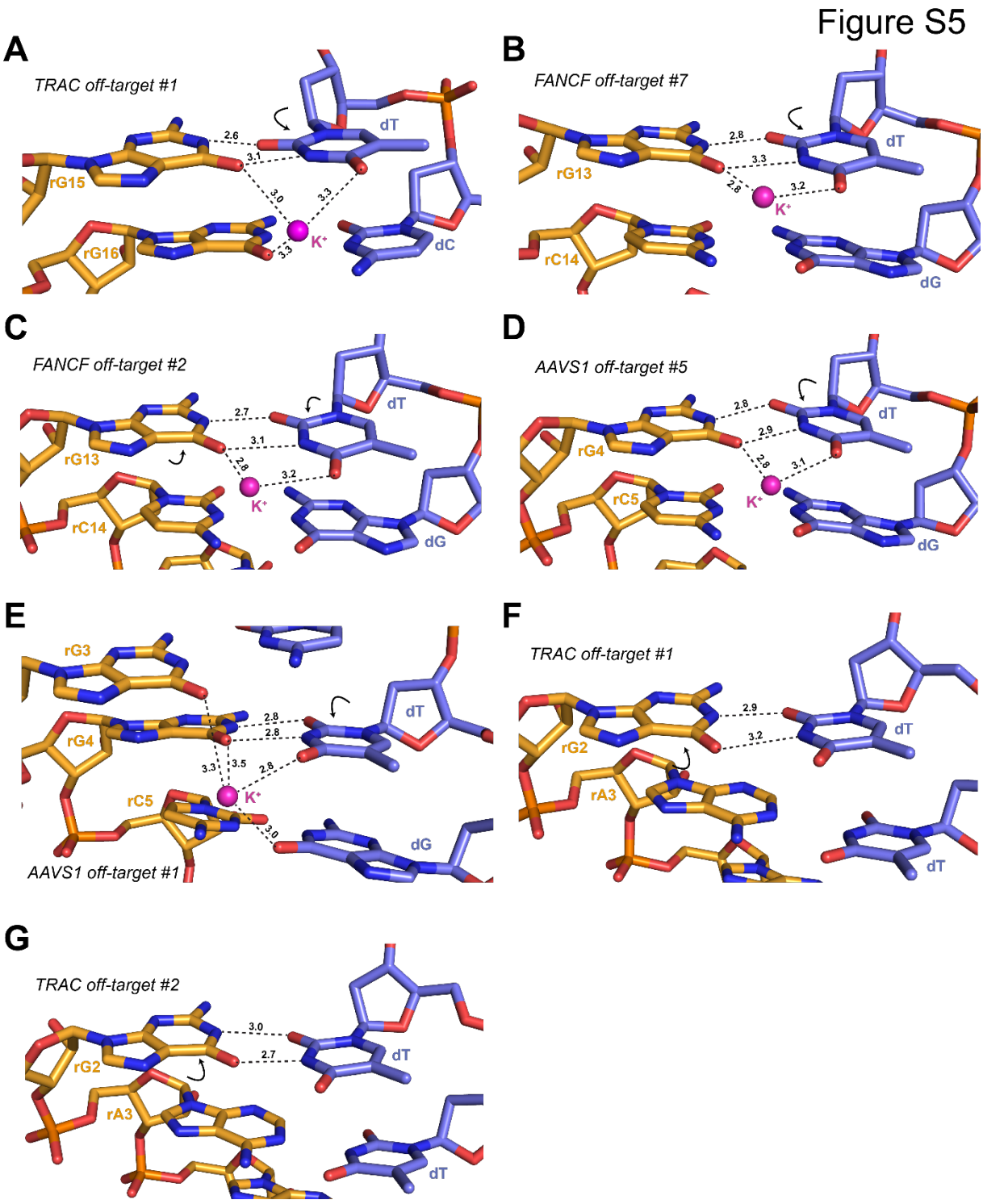
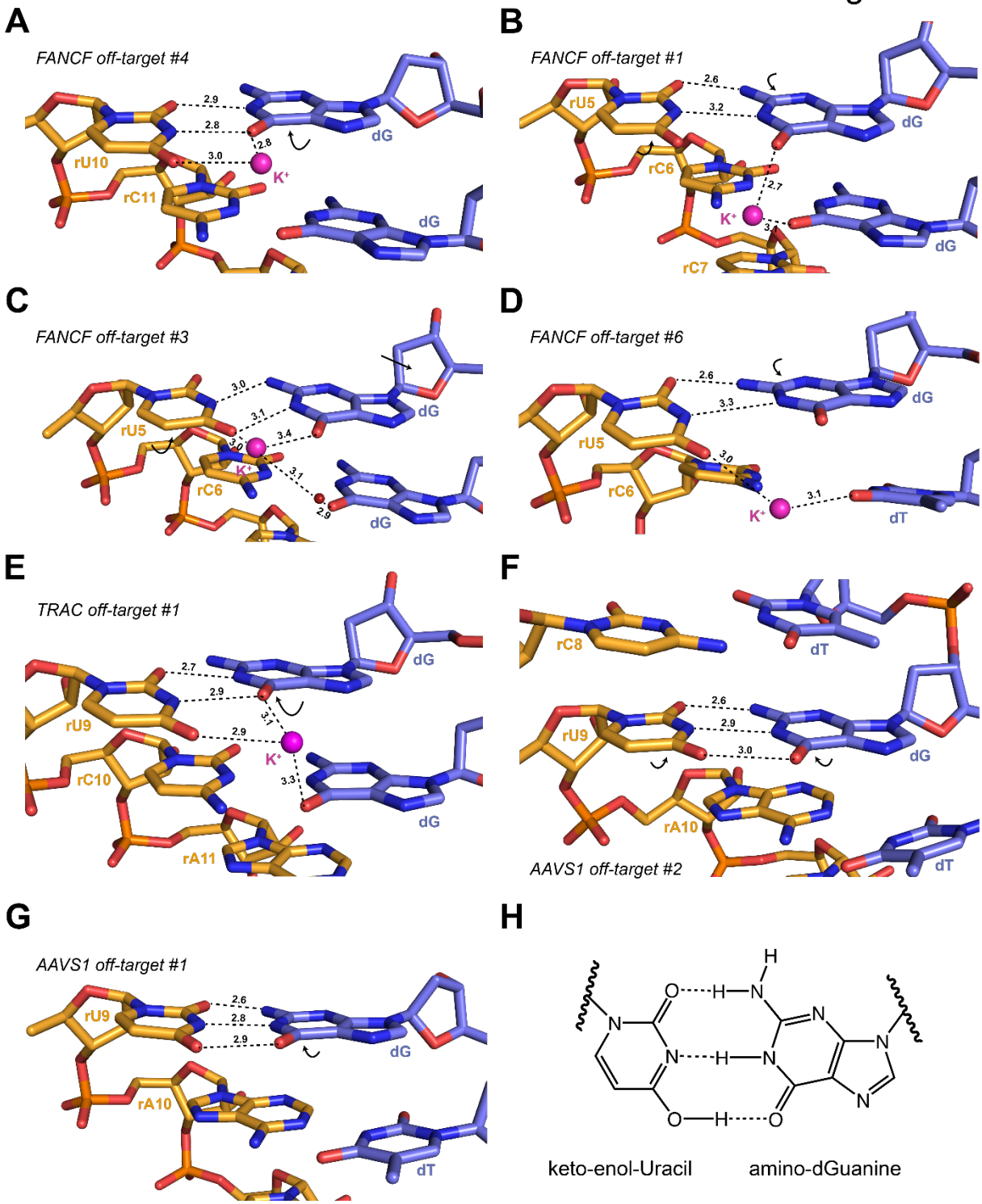


Figure S5. Wobble base pairing of rG-dT mismatches.

Close-up views of rG-dT mismatches at (**A**) heteroduplex position 15 in the *TRAC* off-target #1 complex, (**B**) position 13 in *FANCF* off-target #7 complex, (**C**) position 13 in *FANCF* off-target #2 complex, (**D**) position 4 in *AAVS1* off-target #5 complex, (**E**) position 4 in *AAVS1* off-target #1 complex, (**F**) position 2 in *TRAC* off-target #1 complex and (**G**) position 2 in *TRAC*

786 off-target #2 complex. Arrows indicate conformational changes relative to the corresponding
787 on-target complex structures. Monovalent ions, modeled as K^+, are depicted as purple spheres.
788 In (A)-(E), the dT base is displaced into the major groove and forms a canonical wobble base
789 pair with the rG base. In (F)-(G), the the rG base instead shifts towards the minor groove to
790 facilitate wobble pairing.
791
792
793
794

Figure S6



795

796 **Figure S6. rU-dG wobble base pairs adopt duplex position-dependent conformations.**

797 Close-up views of of rU-dG mispairs at (**A**) heteroduplex position 10 in *FANCF* off-target #4

798 complex, (**B**) position 5 in *FANCF* off-target #1 complex, (**C**) position 5 in *FANCF* off-target

799 #3 complex, (**D**) position 5 in *FANCF* off-target #6 complex, (**E**) *TRAC* off-target #1 complex,

800 and (**F**) *AAVS1* off-target #2, and (**G**) *AAVS1* off-target #1 complex. Arrows indicate
801 conformational changes relative to the corresponding on-target complex structures. Bound
802 potassium ions are depicted as purple spheres. In (A), rU-dG wobble base pairing is achieved
803 by minor groove displacement of the guanine base. In (B)-(D), the rU-dG mispairs adopt
804 atypical conformations. In (E), the guanine base is shifted into the minor groove to form a
805 wobble base pair, whereas at the identical heteroduplex position in (F) and (G), the rU-dG base
806 pairs do not engage in wobble pairing, instead adopting alternative tautomeric forms. **(H)**
807 Schematic depicting hydrogen bonding interactions between rU and dG bases in (F) and (G).
808

Figure S7

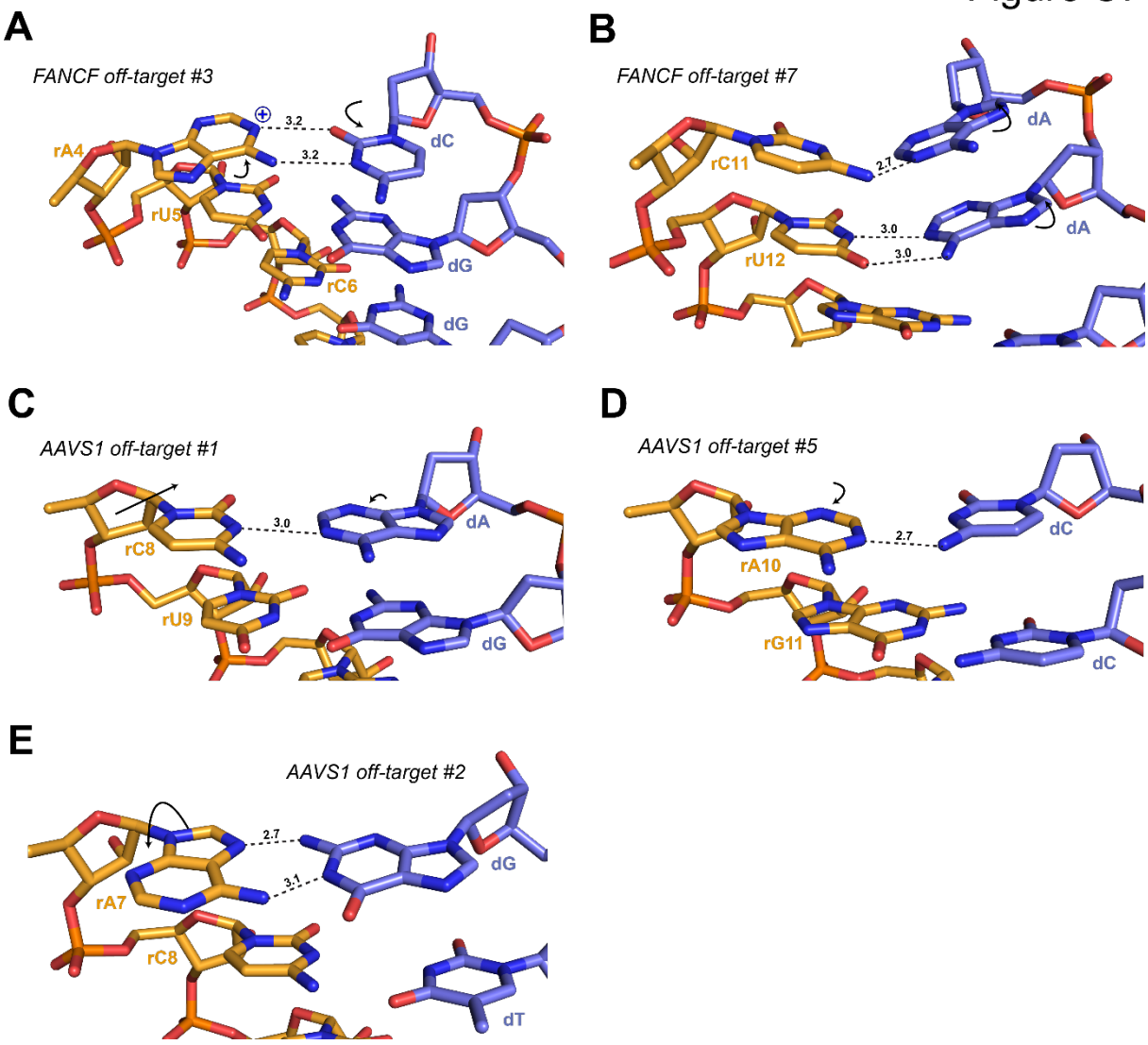


Figure S7. Additional non-canonical base pairs within Cas9 off-target complexes. (**A**) Close-up view of rA-dC wobble base pairing at position 4 in *FANCF* off-target #3 complex. The base pair geometry is consistent with base protonation or tautomerism to enable productive hydrogen bonding between the bases. (**B**) Close-up view of rC-dA mismatch at position 11 of *FANCF* off-target #7 complex, facilitated by base tilting at positions 11 and 12. (**C**) Close-up view of partially paired rC-dA mismatch at position 8 in *AAVS1* off-target #1 complex. (**D**) Close-up view of rA-dC mispair at position 10 in *AAVS1* off-target #5 complex. (**E**) Close-up view of Hoogsteen-edge rA-dG base pair at position 7 in *AAVS1* off-target #2 complex. Arrows indicate conformational changes relative to the corresponding on-target complexes.

Figure S8

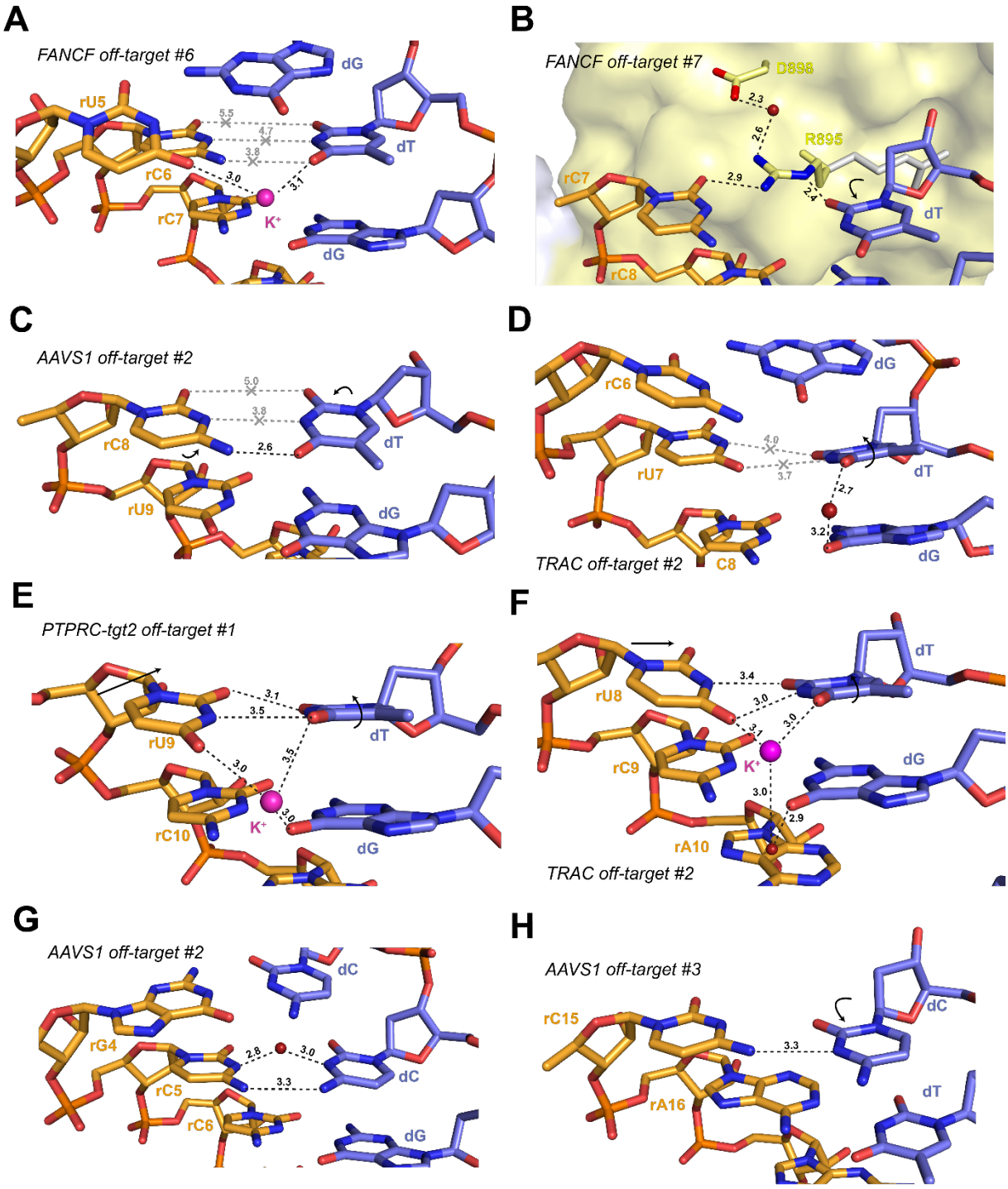
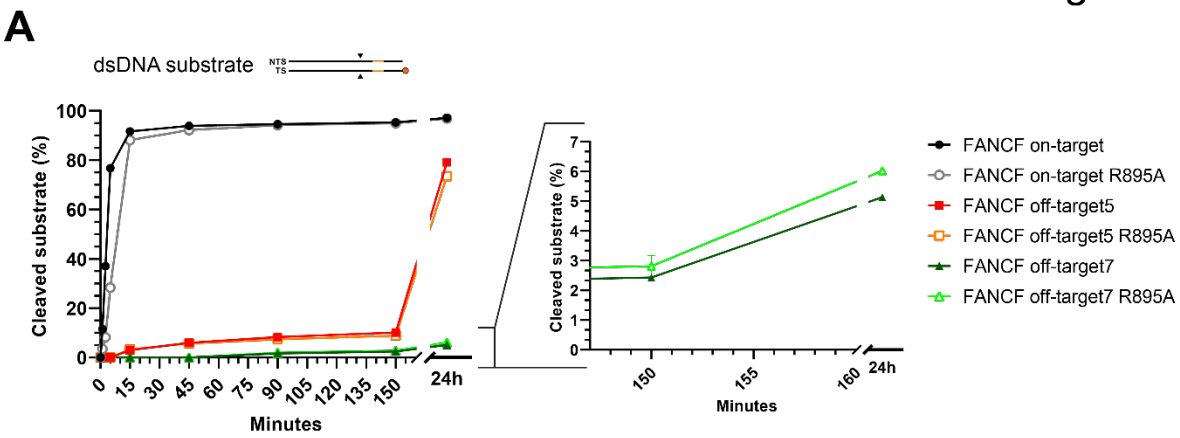


Figure S8. Preservation of base stacking in pyrimidine-pyrimidine off-target mismatches. (**A**) Close-up view of rC-dT mispair at position 6 in *FANCF* off-target #6 complex. (**B**) Close-up view of rC-dT mispair at position 7 in *FANCF* off-target #7 complex, bridged by Arg895. The arginine sidechain in the corresponding on-target complex is shown in white. (**C**) Close-up view of rC-dT base pairing at position 8 of *AAVS1* off-target #2. (**D**) Close-up view of rU-

825 dT mispair at position 7 in *TRAC* off-target #2 complex. (**E**) Close-up view of rU-dT pairing at
826 position 9 in *PTPRC*-tgt2 off-target #1 complex, facilitated by base propeller twisting. (**F**)
827 Close-up view of rU-dT pairing at position 8 in *TRAC* off-target #2 complex, enabled by
828 backbone shift of the RNA strand. (**G**) Close-up view of partially paired rC-dC mismatch at
829 position 5 in *AAVS1* off-target #2 complex, bridged by a water molecule. (**H**) Close-up view of
830 rC-dC mispair at position 15 in *AAVS1* off-target #3 complex.
831
832

Figure S9



B

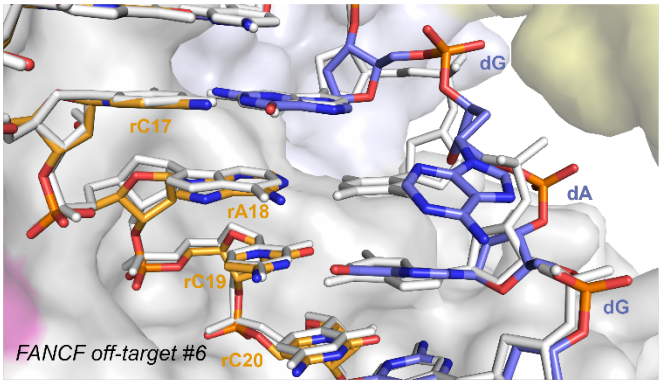
Gene	Target	Cas9	(%) 24h cleavage	k_{obs} (min^{-1})	k_{on} ($M^{-1} . s^{-1}$)	k_{off} (s^{-1})	K_d (pM)	24h cleavage PAMmer (%)	k_{obs} (min^{-1}) PAMmer
FANCF	on-target	WT	97.5	0.2383	4.02E+06	6.77E-05	16.9	98.3	0.5654
FANCF	on-target	R895A	97.1	0.0945	4.87E+06	7.03E-05	14.5	97.9	0.4214
FANCF	off-target7	WT	5.2	0.0036	5.23E+06	2.90E-03	555.0	94.5	0.0134
FANCF	off-target7	R895A	6.1	0.0036	5.06E+06	2.36E-03	466.0	96.5	0.0142

Figure S9. Cas9 **R895A mutation of Cas9 has no significant impact on FANCF off-target #7 cleavage or binding.**

(**A**) Kinetic analysis of *FANCF* on- and off-target substrate DNA cleavage by wild-type and R895A Cas9 proteins. (**B**) Kinetic and thermodynamic parameters of *FANCF* on- and off-target substrate DNA cleavage by wild-type and R895A Cas9. Cleavage rate constants (k_{obs}) were derived from single-exponential function fitting of plots shown in (A). Substrate binding and dissociation rate constants (k_{on} and k_{off}) and the equilibrium dissociation constant (K_d) were determined using a DNA nanolever (switchSENSE) binding assay.

Figure S10

a



b

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Total Mismatches	2029	1771	1453	1295	1100	882	888	924	1171	885	632	769	738	409	506	524	484	708	775	801
dA-rA mismatches	139	36	322	74	66	97	72	39	51	85	49	15	7	57	2	24	16	89	155	148
%dA-rA mismatches	10.8%	9.5%	7.8%	6.9%	5.9%	4.7%	4.7%	4.9%	6.3%	4.7%	3.4%	4.1%	3.9%	2.2%	2.7%	2.8%	2.6%	3.8%	4.1%	4.3%

Figure S10. Tolerance of adenine-adenine mismatches within the heteroduplex. (**A**) Close-up view of rA-dA mismatch at position 18 in *FANCF* off-target #6 complex, overlaid with the *FANCF* on-target structure (white). (**B**) Number of rA-dA off-target mismatches per heteroduplex position recovered in the SITE-Seq assay for all analysed genomic targets. Percentages indicate frequency of rA-dA mismatches recovered in the particular position as a fraction of total number of rA-dA mismatches.

Figure S11

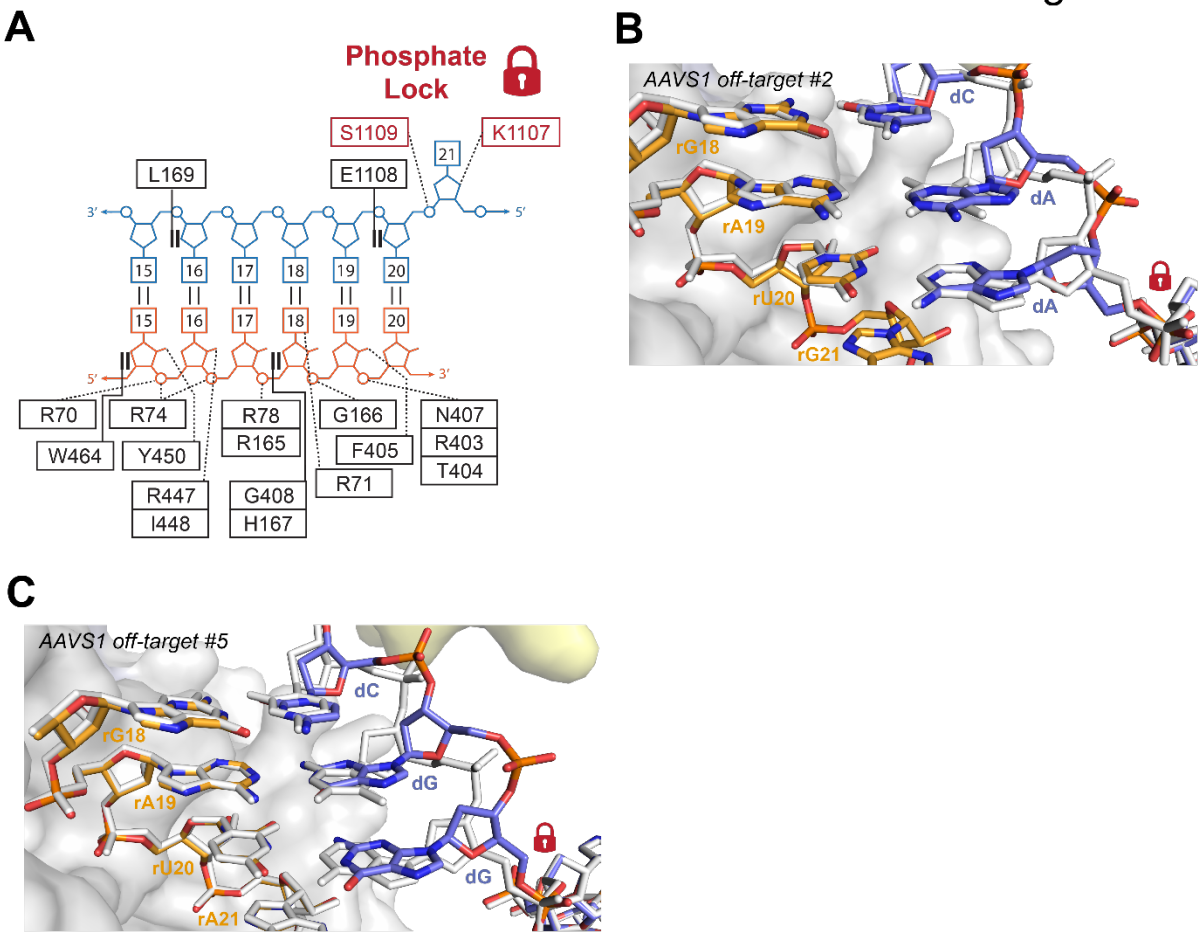


Figure S11. Lack of protein contacts with the target DNA strand in the seed region allows for large phosphate backbone distortions.

(**A**) Schematic overview of Cas9 interactions within the PAM-proximal seed region of the guide RNA-TS DNA heteroduplex. (**B**) Close-up view of the seed region in *AAVS1* off-target #2 complex, overlaid with the AAVS1 on-target heteroduplex (white), showing structural distortion of the TS due to rA-dA mispair at seed position 19. (**C**) Close-up view of the seed region in AAVS1 off-target #5 complex, overlaid with the AAVS1 on-target heteroduplex (white), showing structural distortion due to rA-dG and rU-dG mismatches at positions 19 and 20, respectively. Red lock icon indicates position of the phosphate lock residue in (B) and (C).

Figure S12

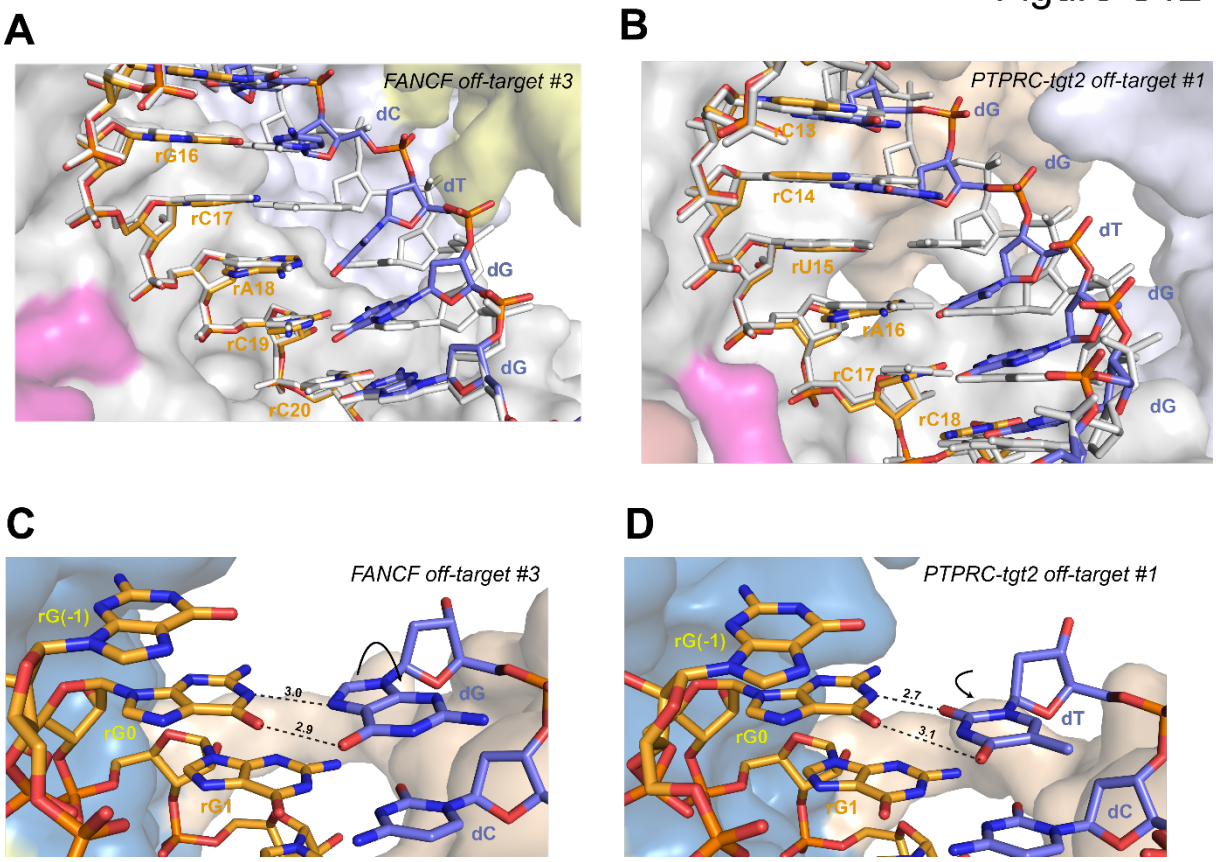
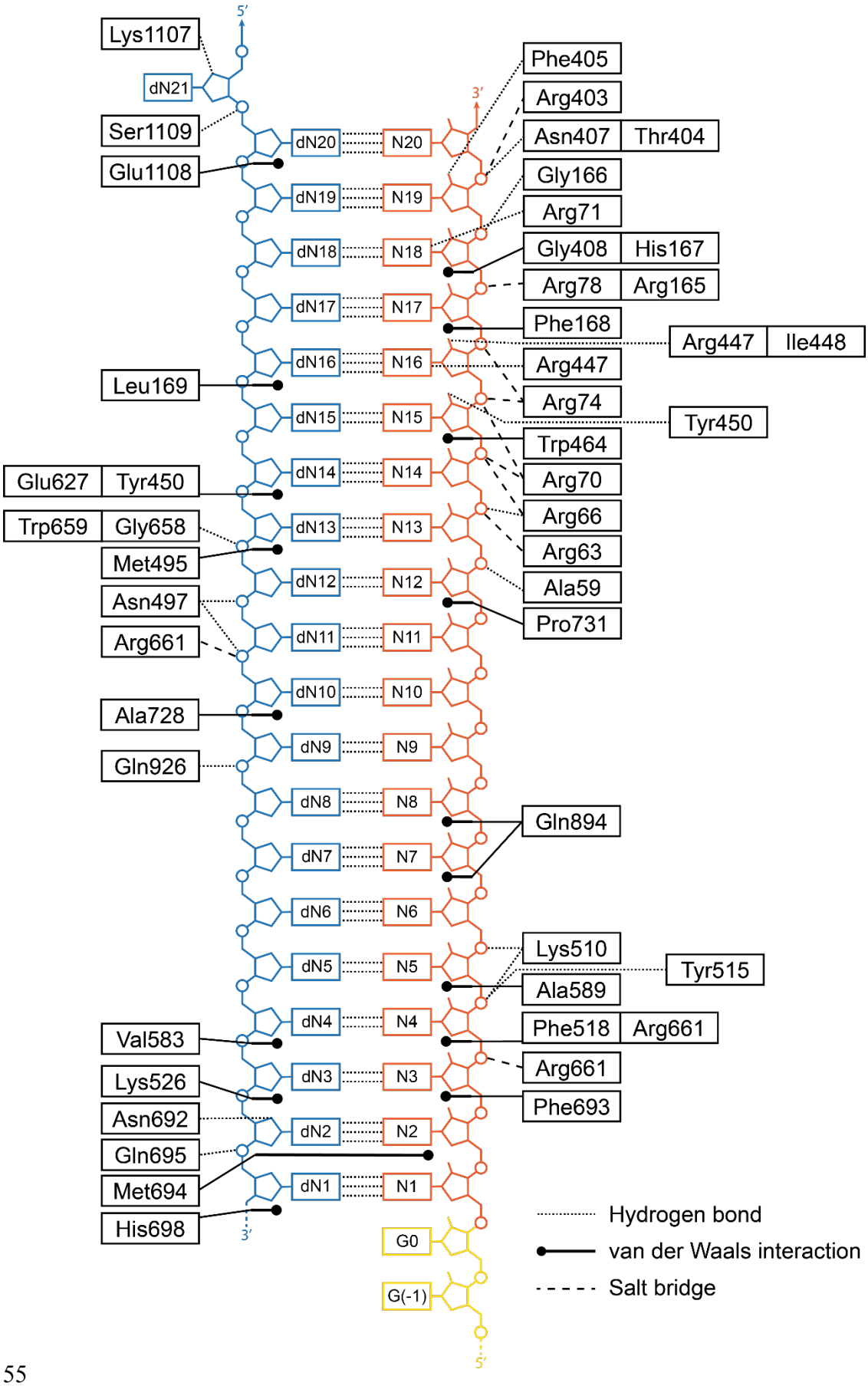


Figure S12. Recognition of off-target sites containing deletions in the seed region. (**A**) Close-up view of base skipping within the seed region of the guide RNA-off-target DNA heteroduplex in *FANCF* off-target #3 complex, overlaid with the on-target heteroduplex (white). (**B**) Close-up view of base skipping within the seed region of the guide RNA-off-target DNA heteroduplex in *PTPRC-tgt2* off-target #1 complex, overlaid with *FANCF* on-target heteroduplex (white). (**C**) Close-up view of non-canonical base pairs at the 5'-terminus of the guide RNA in *FANCF* off-target #3 complex involving guanosine nucleotides introduced during in vitro transcription of the guide RNA. (**D**) Close-up view of non-canonical base pairs at the 5'-terminus of the guide RNA in *PTPRC-tgt2* off-target #1 complex involving guanosine nucleotides introduced during in vitro transcription of the guide RNA.

Figure S13


5′
3′
Lys1107
dN21
Ser1109
Glu1108
Leu169
Glu627
Tyr450
Trp659
Gly658
Met495
Asn497
Arg661
Ala728
Gln926
Val583
Lys526
Asn692
Gln695
Met694
His698
dN20
dN19
dN18
dN17
dN16
dN15
dN14
dN13
dN12
dN11
dN10
dN9
dN8
dN7
dN6
dN5
dN4
dN3
dN2
dN1
N20
N19
N18
N17
N16
N15
N14
N13
N12
N11
N10
N9
N8
N7
N6
N5
N4
N3
N2
N1
G0
G(-1)
Phe405
Arg403
Asn407
Thr404
Gly166
Arg71
Gly408
His167
Arg78
Arg165
Phe168
Arg447
Ile448
Arg447
Arg74
Tyr450
Trp464
Arg70
Arg66
Arg63
Ala59
Pro731
Gln894
Lys510
Tyr515
Ala589
Phe518
Arg661
Arg661
Phe693
Hydrogen bond
van der Waals interaction
Salt bridge

875 **Figure S13. Cas9-nucleic acid interactions in on-target complexes.**

876 Schematic diagram depicting Cas9 residues interacting with the guide RNA-target DNA
877 heteroduplex. Dotted lines represent hydrogen bonding interactions; dashed lines represent salt
878 bridges; solid lines represent stacking/hydrophobic interactions. Target strand is coloured in
879 blue, guide RNA in orange. Phosphates are represented by circles, ribose moieties by
880 pentagons, and nucleobases by rectangles.

881

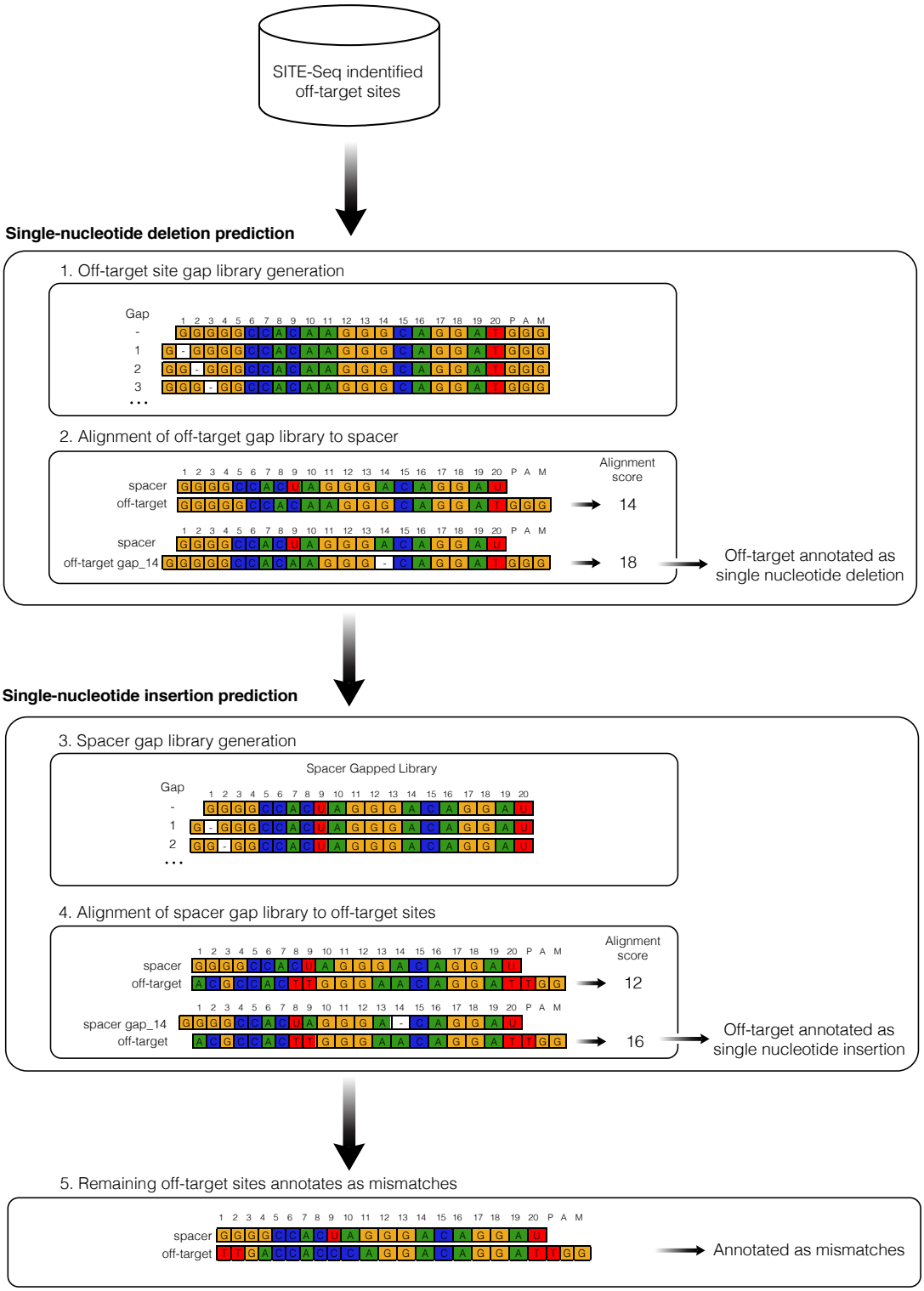


Figure S14. Schematic representation of mismatch, insertion, and deletion classification algorithm for the SITE-Seq assay analysis of off-target sites.

885 Schematic represents unrestricted classification algorithm of off-target sites with putative
886 insertions and deletions. In the final restricted pipeline, the positioning is limited to
887 heteroduplex positions 6-20 for insertions and positions 10-20 for deletions.
888

Figure S15

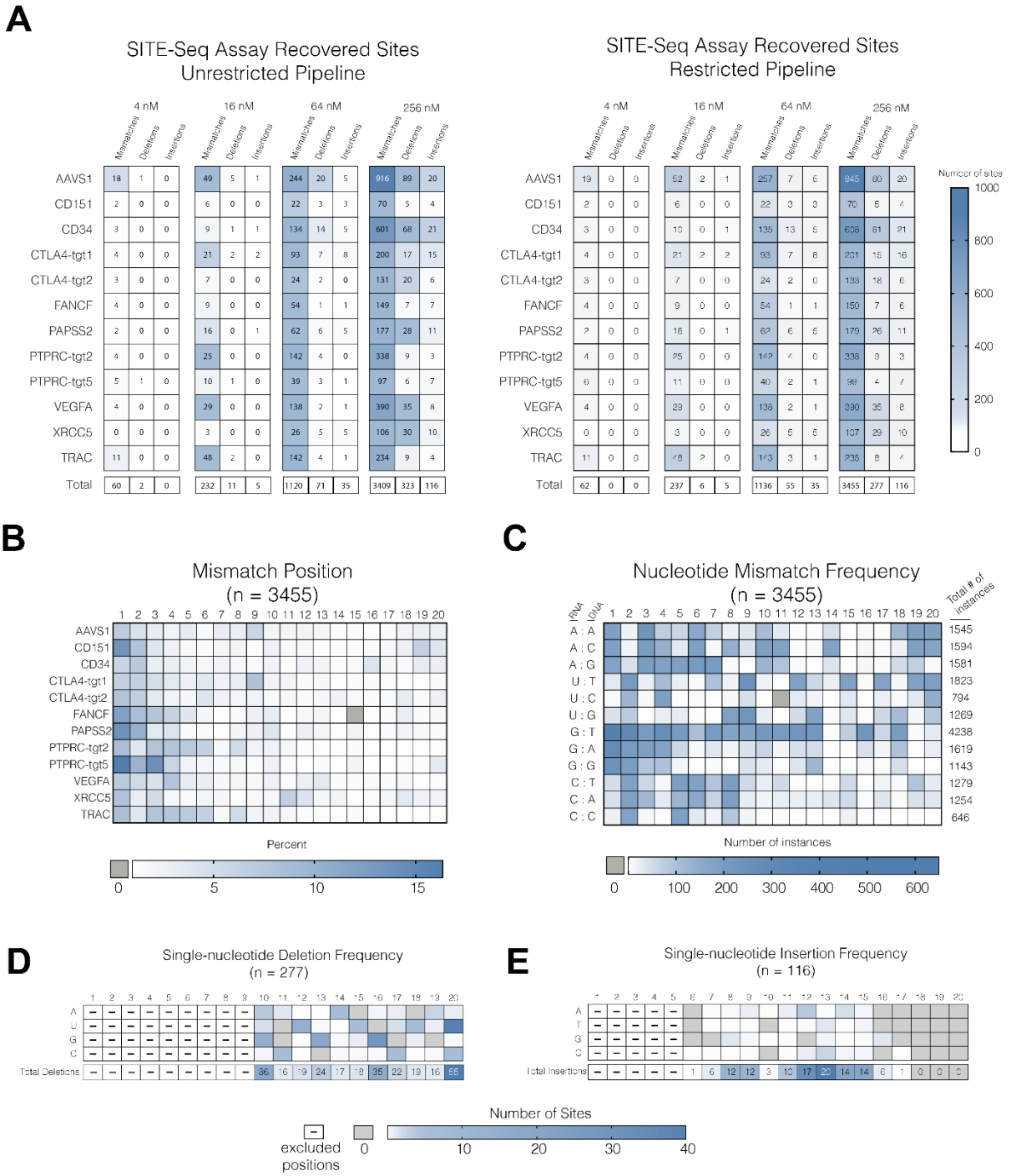


Figure S15. Positional restriction of nucleotide insertions/deletions during of SITE-Seq assay off-target profiling.

(**A**) Number of recovered off-target sites per genomic target as a function of RNP concentration classified as containing either only mismatches, single-nucleotide deletions, or single-nucleotide insertions. Left panel corresponds to classification using algorithm with no

895 positional restriction. Right panel corresponds to classification using algorithm restricting
896 deletions to positions 10-20 and insertions to 6-20 only. (**B**) Frequency of positional mismatch
897 occurrence per genomic target for mismatched off-targets with the positionally restricted
898 algorithm. (**C**) Frequency of nucleotide mismatches within the heteroduplex for all off-target
899 sites when classified with a positionally restricted pipeline (n=3445 sites for both (B) and (C)).
900 (**D**) Frequency of single-nucleotide deletions occurring within positions 10-20 of the
901 heteroduplex for all off-target sites when analysed with a positionally restricted pipeline.
902 (n=277 sites). (**E**) Frequency of single-nucleotide insertions occurring within positions 6-20 of
903 the heteroduplex for all off-target sites when analysed with a positionally restricted pipeline.
904 (n=116 sites).
905

Figure S16

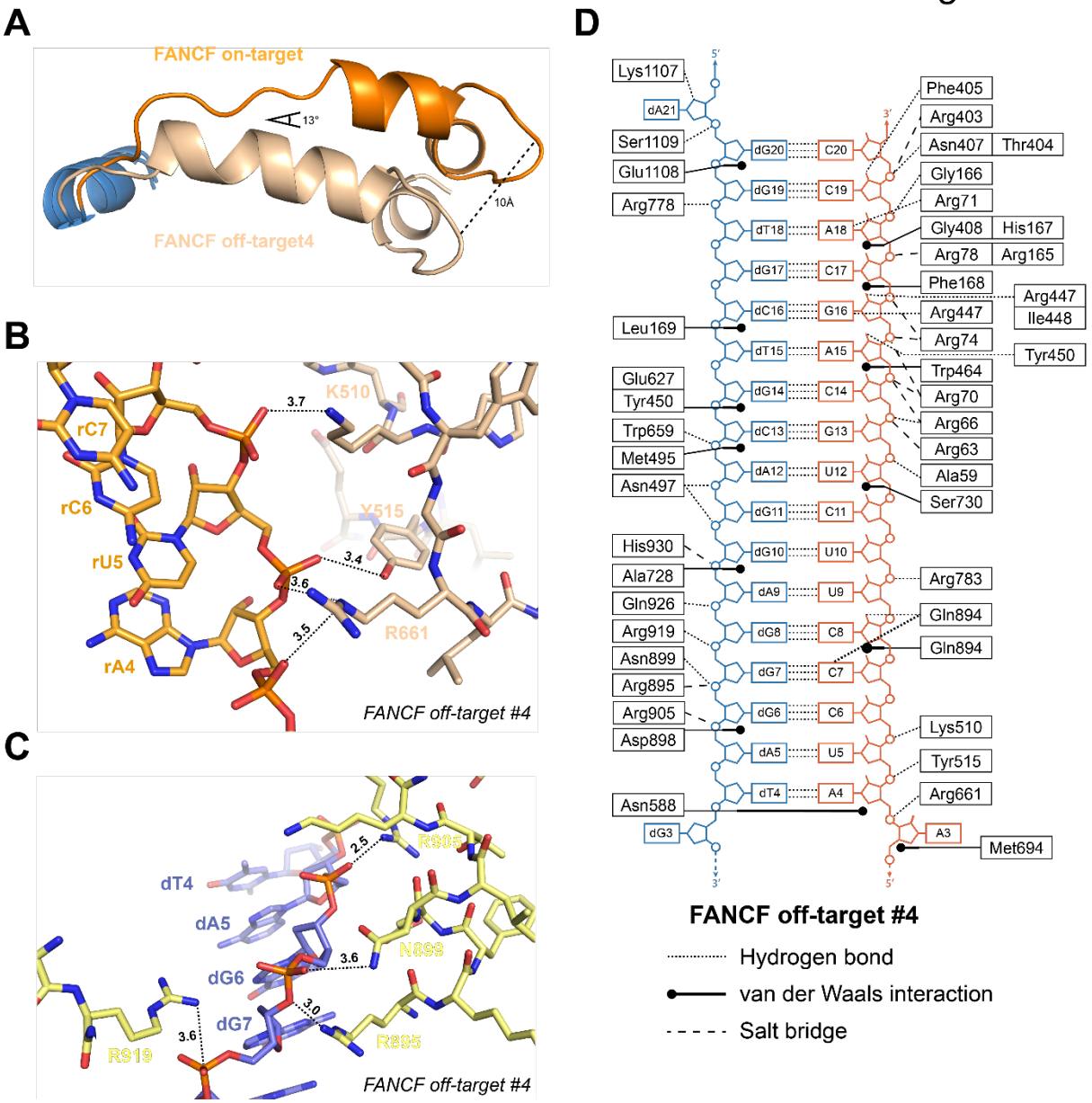


Figure S16. Altered heteroduplex interactions in *FANCF* off-target #4 complex. (**A**) Overlay of REC3 domain helix 703-712 in *FANCF* off-target #4 complex (wheat) with *FANCF* on-target complex (orange). (**B**) Close-up view of REC3 domain interactions with the guide RNA strand in *FANCF* off-target #4 complex. (**C**) Close-up view of TS DNA interactions established by HNH domain in *FANCF* off-target #4 complex. (**D**) Schematic diagram depicting Cas9 residues interacting with the guide RNA-off-target DNA heteroduplex in *FANCF* off-target #4 complex. Dotted lines represent hydrogen bonding interactions, dashed lines represent salt bridges, solid lines represent stacking/hydrophobic interactions. Target

915 strand is coloured blue, guide RNA orange. Phosphates are represented by circles, ribose
916 moieties by pentagons, and nucleobases by rectangles.
917

Figure S17

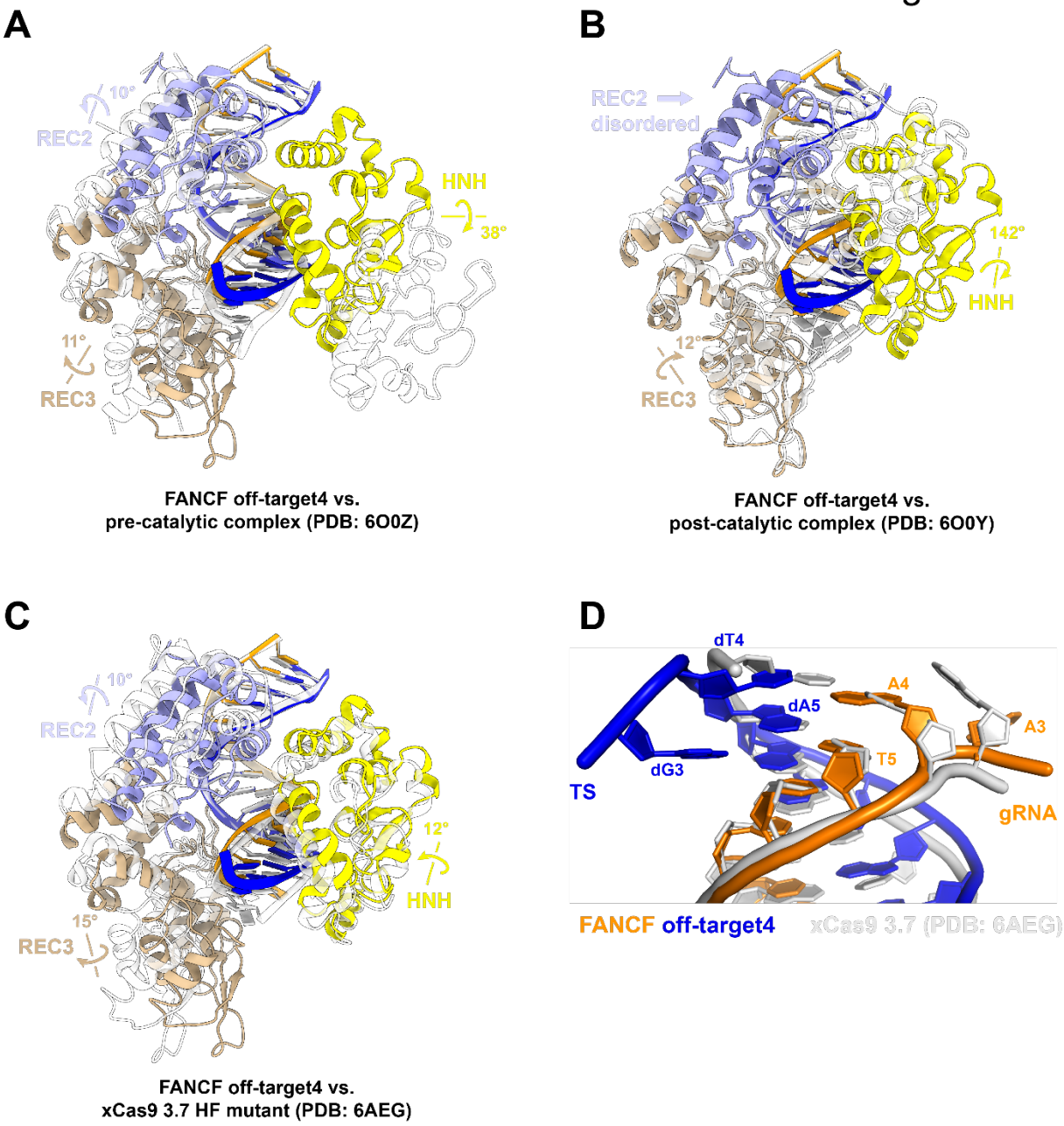


Figure S17. Conformational rearrangements of REC2/3 AND HNH domains in *FANCF* off-target #4 complex.

(**A**) Structural overlay of the *FANCF* off-target #4 complex with cryo-EM structure of a pre-catalytic (State I) Cas9 complex (PDB: 6O0Z). (**B**) Structural overlay of the *FANCF* off-target #4 complex with the cryo-EM structure of a post-catalytic (State II) Cas9 complex (PDB: 6O0Y). (**C**) Structural overlay of the *FANCF* off-target #4 complex with the crystallographic structure of the high-fidelity xCas9 3.7 variant (PDB: 6AEG). The REC1, RuvC, and PAM

interaction domains have been omitted for clarity in all panels, as no significant structural changes were observed in these domains. The *FANCF* off-target #4 complex domains are colored according to **Figure 1A**. The overlaid structures are coloured white. (**D**) Overlay of the PAM-distal heteroduplex region in *FANCF* off-target #4 and xCas9 3.7 on-target complexes. Target strand is coloured in blue, guide RNA is coloured orange.

932 **Table S1. SITE-Seq assay results for Cas9 off-target profiling of 12 selected genomic sites.**
933 Columns indicate the recovered off-target sequence; motif location; number of substitutions in
934 recovered target sequence compared to the on-target (substitutions); strand designation of
935 PAM; the lowest recovery concentration of each target; and whether the off-target is predicted
936 to contain inserts or deletions based on restricted pipeline paraments. Off-target sites recovered
937 at lower concentrations were also recovered at higher concentrations (e.g., all 4nM sites were
938 also recovered at 16nM, 64 nM, and 256 nM).
939
940 **Table S2. List of recovered off-target sequences aligned to the corresponding on-target**
941 **sequence.**
942 Off-target alignments classified by genomic target and by the presence of insertions, deletions
943 or purely mismatched targets, as based on restricted pipeline paraments. Indexes correspond to
944 off-target sequence numbering in Table S1.
945
946 **Table S3. Crystallographic data collection and refinement statistics of Cas9 on-target and**
947 **off-target complexes**
948
949 **Table S4. 3DNA 2.0 analysis of the helical parameters and sugar puckering of**
950 **characterised on-target and off-target duplexes**
951
952 **Table S5. List of oligonucleotides used in this study**
953
954
955

956 ## Methods

957 ### DNA oligonucleotides and substrates

958 Sequences of DNA oligonucleotides used in this study are summarised in **Table S5**.
959 Crystallisation substrates were synthesised by Sigma Aldrich without further purification,
960 sgRNA transcription templates and ATTO-532 labelled cleavage substrates were synthesised
961 by Integrated DNA Technologies, Inc., with PAGE and HPLC purification, respectively.
962 Partially double stranded crystallisation substrates were prepared by mixing complementary
963 oligonucleotides in a 1:1 molar ratio (as determined by 260 nm absorption), heating to 95 °C
964 for 5 minutes and slow cooling to room-temperature. Cleavage substrates were prepared
965 similarly, except that a 2-fold molar excess of the non-target strand was used.

966 ### Cas9 protein expression and purification

967 *Streptococcus pyogenes* Cas9 wild type protein and the nuclease dead mutant (D10A, H840A)
968 were both recombinantly expressed for 16 hours at 18 °C in *Escherichia coli* Rosetta 2 (DE3)
969 (Novagen) N-terminally fused to a hexahistidine affinity tag, the maltose binding protein
970 (MBP) polypeptide, and the tobacco etch virus (TEV) protease cleavage site. Cells were
971 resuspended and lysed in 20 mM HEPES-KOH pH 7.5, 500 mM KCl, 5 mM imidazole, and
972 supplemented with added protease inhibitors. Clarified lysate was loaded on a 10 ml Ni-NTA
973 Superflow column (QIAGEN), washed with 7 column volumes of 20 mM HEPES-KOH pH
974 7.5, 500 mM KCl, 5 mM imidazole, and eluted with 10 column volumes of 20 mM HEPES-
975 KOH pH 7.5, 250 mM KCl, 200 mM imidazole. Salt concentration is adjusted and protein is
976 loaded on a 10 ml HiTrap Heparin HP column (GE Healthcare) equilibrated in 20 mM HEPES-
977 KOH pH 7.5, 250 mM KCl, 1 mM DTT. The column is washed with 5 column volumes of 20
978 mM HEPES-KOH pH 7.5, 250 mM KCl, 1 mM DTT, and Cas9 is eluted with 17 column
979 volumes of 20 mM HEPES-KOH pH 7.5, 1.5 M KCl, 1 mM DTT, in a 0-32% gradient (peak
980 elution around 500 mM KCl). His_6-MBP tag was removed by TEV protease cleavage overnight

981 with gentle shaking. The untagged Cas9 was concentrated and applied to a Superdex 200 16/600
982 (GE Healthcare) and eluted with 20 mM HEPES-KOH pH 7.5, 500 mM KCl, 1 mM DTT.
983 Purified protein was concentrated to 10 mg/ml, flash frozen in liquid nitrogen and store
984 at -80 °C. DTT was omitted in the size-exclusion step of the purification when protein was used
985 for switchSENSE measurements.

986 sgRNA transcription and purification

987 sgRNAs are transcribed from a double stranded PCR product template amplified from a plasmid
988 in a 5 ml transcription reaction (30 mM Tris-HCl pH 8.1, 25 mM $MgCl_2$, 2 mM spermidine,
989 0.01% Triton X-100, 5 mM CTP, 5 mM ATP, 5 mM GTP, 5 mM UTP, 10 mM DTT, 1 µM
990 DNA transcription template, 0.5 units inorganic pyrophosphatase (Thermo Fischer), 250 µg
991 homemade T7 RNA polymerase. The reaction is incubated at 37 °C for 5 hours, and then treated
992 for 30 minutes with 15 units of RQ1 DNAse (Promega). The transcribed sgRNAs are
993 subsequently PAGE purified on an 8% denaturing (7 M urea) polyacrylamide gel, and lastly
994 ethanol precipitated and resuspended in DEPC treated water.

995 Crystallisation of Cas9 ternary complexes and structure determination

996 To assemble the Cas9 on-/off-target ternary complexes, the Cas9 protein is first mixed with the
997 sgRNA in a 1:1.5 molar ratio and incubated at room temperature for 10 minutes. Next, the
998 binary complex is diluted to 2 mg/ml with 20 mM HEPES-KOH 7.5, 250 mM KCl, 1 mM DTT,
999 2 mM $MgCl_2$ buffer, pre-annealed 100 µM DNA substrate is added in a 1:1.8 molar ratio and
1000 the complex is incubated another 10 minutes at room temperature. For crystallisation, 1 µl of
1001 the ternary complex (1-2 mg/ml) is mixed with 1 µl of the reservoir solution (0.1 M Tris-acetate
1002 pH 8.5, 0.3-0.5 M KSCN, 17-19% PEG3350) and crystals are grown at 20 °C using the hanging
1003 drop vapour diffusion method. In some cases, microseeding was be used to improve crystal
1004 morphology. Crystals are typically harvested after 2-3 weeks, cryoprotected in 0.1 M Tris-
1005 acetate pH 8.5, 0.4 M KSCN, 30% PEG3350, 15% ethylene glycol, 1 mM $MgCl_2$, and flash-

1006 cooled in liquid nitrogen. Diffraction data was obtained at beamlines PXI and PXIII of the
1007 Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland) and were processed using
1008 the XDS package (Kabsch, 2010). Structures were solved by molecular replacement through
1009 the Phaser module of the Phenix package (Adams et al., 2010) using the PDB ID: 5FQ5 model
1010 omitting the RNA-DNA target duplex from the search. Model adjustment and duplex building
1011 was completed using COOT software (Emsley et al., 2010). Atomic model refinement was
1012 performed using Phenix.refine (Adams et al., 2010). Protein-nucleic acid interactions were
1013 analysed using the PISA web server (Krissinel and Henrick, 2007). Characterisation of the
1014 guide-protospacer duplex was performed using the 3DNA 2.0 web server (Li et al., 2019).
1015 Structural figures were generated using PyMOL and ChimeraX (Pettersen et al., 2021).

1016 *In vitro* nuclease activity assays

1017 Cleavage reactions were performed at 37 °C in reaction buffer, containing 20 mM HEPES pH
1018 7.5, 250 mM KCl, 5 mM $MgCl_2$ and 1 mM DTT. First, Cas9 protein was pre-incubated with
1019 sgRNA in 1:1.25 ratio for 10 minutes at room temperature. The protein-RNA complex was
1020 rapidly mixed with the ATTO-532 labelled dsDNA, to yield final concentrations of 1.67 µM
1021 protein and 66.67 nM substrate in a 7.5 µl reaction. Time points were harvested at 1, 2.5, 5, 15,
1022 45, 90, 150 minutes, and 24 hours. Cleavage was stopped by addition of 2 µl of 250 mM EDTA,
1023 0.5% SDS and 20 µg of Proteinase K. Formamide was added to the reactions with final
1024 concentration of 50%, samples were incubated at 95 °C for 10 minutes, and resolved on a 15%
1025 denaturing PAGE gel containing 7M urea and imaged using a Typhoon FLA 9500 gel imager.
1026 Depicted error bars correspond to the standard deviation from four independent cleavage
1027 reactions. Rate constants (k_{obs}) were extracted from single exponential fits: [Product] = A*(1-
1028 exp(-k_{obs}*t))

1029 switchSENSE analysis

1030 The target strands (TS) containing a 3' flanking sequence complementary to the ssDNA
1031 covalently bound to the chip electrode, and the non-target strands (NTS) (**Table S5**) were
1032 resuspended in a buffer containing 10 mM Tris-HCl pH 7.4, 40 mM NaCl, and 0.05% Tween
1033 20. The matching TS:NTS duplex is pre-annealed and hybridised to the chip anode. The Cas9
1034 protein was mixed with the sgRNAs at a 1:2 protein:RNA molar ratio, and the complex was
1035 incubated for 30 min at 37 °C in association buffer containing 20 mM HEPES-KOH pH 7.5,
1036 150 mM KCl, 2 mM $MgCl_2$, 0.01% Tween 20. All switchSENSE experiments were performed
1037 on a DRX analyser using CAS-48-1-R1-S chips (Dynamic Biosensors GmbH, Martinsried,
1038 Germany). Kinetics experiments were performed at 25 °C in association buffer, with an
1039 association time of 5 min, dissociation time of 20 min, and a flow rate of 50 µl/min.

1040 SITE-Seq assay

1041 SITE-Seq assay reaction conditions were performed as described previously (Cameron et al.,
1042 2017). Briefly, high molecular weight genomic DNA (gDNA) was purified from human
1043 primary T cells using the Blood & Cell Culture DNA Maxi Kit (Qiagen) according to the
1044 manufacturer's instructions. RNPs comprising the guides were biochemically assembled for
1045 gDNA digestion. Specifically, equal molar amounts of crRNA and tracrRNA were mixed and
1046 heated to 95 °C for 2 min then allowed to cool at room temperature for ~5 min. Three-fold
1047 molar excess of the guides were incubated with *Streptococcus pyogenes* Cas9 (SpCas9) in
1048 cleavage reaction buffer (20 mM HEPES pH 7.4, 150 mM KCl, 10 mM MgCl2, 5% glycerol)
1049 at 37 °C for 10 min. In a 96-well plate format, 10 µg of gDNA was treated with 0.2 pmol (4
1050 nM), 0.8 pmol (16 nM), 3.2 pmol (64 nM), and 12.8 pmol (256 nM) of each RNP in 50 µL total
1051 volume in cleavage reaction buffer. Each cleavage reaction was performed in triplicate.
1052 Negative control reactions were assembled in parallel and did not include RNP. gDNA was
1053 treated with RNPs for 4 hours at 37 °C. SITE-Seq assay library preparation and sequencing

1054 was performed as described previously and the final library was loaded onto the Illumina
1055 NextSeq platform (Illumina, San Diego, CA), and ~1-3 M reads were obtained for each sample.

1056 SITE-Seq assay analysis and selection for cellular validation

1057 SITE-Seq assay recovered off-targets were filtered for sites that had read-pileups proximal to
1058 the expected cut site, a PAM comprising at least one guanine base, fewer than 12 mismatches
1059 (reasoning that sites with 12 or more mismatches are likely spurious peaks not resulting from
1060 Cas9-induced double-strand breaks), and all sites with 11 mismatches were visually inspected
1061 and included in analysis if a putative deletion or insertion would result in a reduction of >4
1062 mismatches relative to the spacer sequence.

1063 *In silico* mismatch, deletion, and insertion prediction algorithm

1064 Predictive classification of SITE-Seq assay recovered off-target sites as pure mismatches,
1065 deletions, or insertions was executed using a scoring algorithm which consisted of the following
1066 sequential steps (**Figure S14**):

1067 (i) For each off-target, a gap library was generated where a single nucleotide gap was
1068 introduced between each nucleotide in the off-target sequence.

1069 (ii) The off-target gap library was then aligned to the spacer sequence and each
1070 alignment was scored based on the number of matched bases between the spacer
1071 and gapped off-target pair. If the gapped off-target with the highest alignment score
1072 improved alignment by at least 4 nucleotides relative to the non-gapped spacer–off-
1073 target alignment, the off-target sequence was marked as a single-nucleotide deletion
1074 and removed from subsequent analysis.

1075 (iii) The remaining pool of off-targets were then aligned to a spacer gapped library where
1076 a single nucleotide gap was introduced at each positing in the spacer.

1077 (iv) The spacer gap library was then aligned to each off-target sequence and each
1078 alignment was scored based on the number of matched bases between the off-target

1079 and the gapped spacer pair. If the gapped spacer with the highest alignment score
1080 improved alignment by at least 4 nucleotides relative to the non-gapped spacer–off-
1081 target alignment, the off-target sequence was marked as a single-nucleotide insertion
1082 and removed from subsequent analysis.

1083 (v) The remaining off-target for which the spacer–off-target alignment was not
1084 improved by single-nucleotide deletions or insertions were annotated as a
1085 mismatched off-target.

1086 The prediction pipeline process was the same for the 'unrestricted' and structurally-
1087 informed 'restricted' pipelines, however in the 'restricted' pipeline the deletion gap library
1088 was restricted to positions 10-20 and the insertion gap library was restricted to positions 6-
1089 20.

1090

1091

1092 ## References

1093 Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and Mann, R.S. (2015).
1094 Deconvolving the recognition of DNA shape from sequence. Cell *161*, 307-318.

1095 Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung,
1096 L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based
1097 system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr *66*, 213-221.

1098 Afek, A., Shi, H., Rangadurai, A., Sahay, H., Senitzki, A., Xhani, S., Fang, M., Salinas, R., Mielko, Z.,
1099 Pufall, M.A.*, et al.* (2020). DNA mismatches reveal conformational penalties in protein-DNA
1100 recognition. Nature.

1101 Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target
1102 DNA recognition by the Cas9 endonuclease. Nature *513*, 569-573.

1103 Anzalone, A.V., Koblan, L.W., and Liu, D.R. (2020). Genome editing with CRISPR-Cas nucleases,
1104 base editors, transposases and prime editors. Nat Biotechnol *38*, 824-844.

1105 Bae, S., Park, J., and Kim, J.S. (2014). Cas-OFFinder: a fast and versatile algorithm that searches for
1106 potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics *30*, 1473-1475.

1107 Boyle, E.A., Andreasson, J.O.L., Chircus, L.M., Sternberg, S.H., Wu, M.J., Guegler, C.K., Doudna,
1108 J.A., and Greenleaf, W.J. (2017). High-throughput biochemical profiling reveals sequence determinants
1109 of dCas9 off-target binding and unbinding. Proc Natl Acad Sci U S A *114*, 5461-5466.

1110 Boyle, E.A., Becker, W.R., Bai, H.B., Chen, J.S., Doudna, J.A., and Greenleaf, W.J. (2021).
1111 Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target
1112 engagement. Science Advances *7*, eabe5496.

1113 Bravo, J.P.K., Liu, M.-S., McCool, R.S., Jung, K., Johnson, K.A., and Taylor, D.W. (2021). Structural
1114 basis for mismatch surveillance by CRISPR/Cas9. bioRxiv, 2021.2009.2014.460224.

1115 Cameron, P., Fuller, C.K., Donohoue, P.D., Jones, B.N., Thompson, M.S., Carter, M.M., Gradia, S.,
1116 Vidal, B., Garner, E., Slorach, E.M.*, et al.* (2017). Mapping the genomic landscape of CRISPR-Cas9
1117 cleavage. Nat Methods *14*, 600-606.

1118 Chen, J.S., Dagdas, Y.S., Kleinstiver, B.P., Welch, M.M., Sousa, A.A., Harrington, L.B., Sternberg,
1119 S.H., Joung, J.K., Yildiz, A., and Doudna, J.A. (2017). Enhanced proofreading governs CRISPR-Cas9
1120 targeting accuracy. Nature *550*, 407-410.

1121 Cofsky, J.C., Soczek, K.M., Knott, G.J., Nogales, E., and Doudna, J.A. (2021). CRISPR-Cas9 bends
1122 and twists DNA to read its sequence. bioRxiv, 2021.2009.2006.459219.

1123 Dagdas, Y.S., Chen, J.S., Sternberg, S.H., Doudna, J.A., and Yildiz, A. (2017). A conformational
1124 checkpoint between DNA binding and cleavage by CRISPR-Cas9. Sci Adv *3*, eaao0027.

1125 Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel,
1126 J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor
1127 RNase III. Nature *471*, 602-607.

1128 Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A.,
1129 Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus
1130 thermophilus. J Bacteriol *190*, 1390-1400.

1131 Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova,
1132 Z., Wilen, C., Orchard, R.*, et al.* (2016). Optimized sgRNA design to maximize activity and minimize
1133 off-target effects of CRISPR-Cas9. Nat Biotechnol *34*, 184-191.

1134 Donohoue, P.D., Pacesa, M., Lau, E., Vidal, B., Irby, M.J., Nyer, D.B., Rotstein, T., Banh, L., Toh,
1135 M.S., Gibson, J.*, et al.* (2021). Conformational control of Cas9 by CRISPR hybrid RNA-DNA guides
1136 mitigates off-target activity in T cells. Mol Cell *81*, 3637-3649 e3635.

1137 Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta
1138 Crystallogr D Biol Crystallogr *66*, 486-501.

1139 Fu, B.X.H., Smith, J.D., Fuchs, R.T., Mabuchi, M., Curcuru, J., Robb, G.B., and Fire, A.Z. (2019).
1140 Target-dependent nickase activities of the CRISPR-Cas nucleases Cpf1 and Cas9. Nat Microbiol *4*, 888-
1141 897.

1142 Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease
1143 specificity using truncated guide RNAs. Nat Biotechnol *32*, 279-284.

1144 Garg, A., and Heinemann, U. (2018). A novel form of RNA double helix based on G.U and C.A(+)
1145 wobble base pairing. RNA *24*, 209-218.

1146 Gong, S., Yu, H.H., Johnson, K.A., and Taylor, D.W. (2018). DNA Unwinding Is the Primary
1147 Determinant of CRISPR-Cas9 Activity. Cell Rep *22*, 359-371.

1148 Guo, M., Ren, K., Zhu, Y., Tang, Z., Wang, Y., Zhang, B., and Huang, Z. (2019). Structural insights
1149 into a high fidelity variant of SpCas9. Cell Res *29*, 183-192.

1150 Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu,
1151 X., Shalem, O.*, et al.* (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol
1152 *31*, 827-832.

1153 Ivanov, I.E., Wright, A.V., Cofsky, J.C., Aris, K.D.P., Doudna, J.A., and Bryant, Z. (2020). Cas9
1154 interrogates DNA in discrete steps modulated by mismatches and supercoiling. Proc Natl Acad Sci U S
1155 A *117*, 5853-5860.

1156 Jiang, F., Taylor, D.W., Chen, J.S., Kornfeld, J.E., Zhou, K., Thompson, A.J., Nogales, E., and Doudna,
1157 J.A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science *351*,
1158 867-871.

1159 Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). STRUCTURAL BIOLOGY. A Cas9-
1160 guide RNA complex preorganized for target DNA recognition. Science *348*, 1477-1481.

1161 Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A
1162 programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816-
1163 821.

1164 Jones, S.K., Jr., Hawkins, J.A., Johnson, N.V., Jung, C., Hu, K., Rybarski, J.R., Chen, J.S., Doudna,
1165 J.A., Press, W.H., and Finkelstein, I.J. (2020). Massively parallel kinetic profiling of natural and
1166 engineered CRISPR nucleases. Nat Biotechnol.

1167 Kabsch, W. (2010). Xds. Acta Crystallogr D Biol Crystallogr *66*, 125-132.

1168 Kimsey, I.J., Petzold, K., Sathyamoorthy, B., Stein, Z.W., and Al-Hashimi, H.M. (2015). Visualizing
1169 transient Watson-Crick-like mispairs in DNA and RNA duplexes. Nature *519*, 315-320.

1170 Kimsey, I.J., Szymanski, E.S., Zahurancik, W.J., Shakya, A., Xue, Y., Chu, C.C., Sathyamoorthy, B.,
1171 Suo, Z., and Al-Hashimi, H.M. (2018). Dynamic basis for dG*dT misincorporation via tautomerization
1172 and ionization. Nature *554*, 195-201.

1173 Kitayner, M., Rozenberg, H., Rohs, R., Suad, O., Rabinovich, D., Honig, B., and Shakked, Z. (2010).
1174 Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. Nat
1175 Struct Mol Biol *17*, 423-429.

1176 Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J
1177 Mol Biol *372*, 774-797.

Kulcsar, P.I., Talas, A., Toth, E., Nyeste, A., Ligeti, Z., Welker, Z., and Welker, E. (2020). Blackjack mutations improve the on-target activities of increased fidelity variants of SpCas9 with 5'G-extended sgRNAs. Nat Commun *11*, 1223.

Kunkel, T.A., and Bebenek, K. (2000). DNA replication fidelity. Annu Rev Biochem *69*, 497-529.

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., and Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat Biotechnol *32*, 677-683.

Lazzarotto, C.R., Malinin, N.L., Li, Y., Zhang, R., Yang, Y., Lee, G., Cowley, E., He, Y., Lan, X., Jividen, K.*, et al.* (2020). CHANGE-seq reveals genetic and epigenetic effects on CRISPR-Cas9 genome-wide activity. Nat Biotechnol.

Leontis, N.B., Stombaugh, J., and Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Res *30*, 3497-3531.

Li, S., Olson, W.K., and Lu, X.J. (2019). Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. Nucleic Acids Res *47*, W26-W34.

Lin, Y., Cradick, T.J., Brown, M.T., Deshmukh, H., Ranjan, P., Sarode, N., Wile, B.M., Vertino, P.M., Stewart, F.J., and Bao, G. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res *42*, 7473-7485.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P.*, et al.* (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol *18*, 67-83.

Mekler, V., Minakhin, L., and Severinov, K. (2017). Mechanism of duplex DNA destabilization by RNA-guided Cas9 nuclease during target interrogation. Proc Natl Acad Sci U S A *114*, 5443-5448.

Mitchell, B.P., Hsu, R.V., Medrano, M.A., Zewde, N.T., Narkhede, Y.B., and Palermo, G. (2020). Spontaneous Embedding of DNA Mismatches Within the RNA:DNA Hybrid of CRISPR-Cas9. Frontiers in Molecular Biosciences *7*.

Mullally, G., van Aelst, K., Naqvi, M.M., Diffin, F.M., Karvelis, T., Gasiunas, G., Siksnys, V., and Szczelkun, M.D. (2020). 5' modifications to CRISPR-Cas9 gRNA can change the dynamics and size of R-loops and inhibit DNA cleavage. Nucleic Acids Res.

Murugan, K., Seetharam, A.S., Severin, A.J., and Sashital, D.G. (2020). High-throughput in vitro specificity profiling of natural and high-fidelity CRISPR-Cas9 variants. bioRxiv, 2020.2005.2012.091991.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell *156*, 935-949.

O'Connell, M.R., Oakes, B.L., Sternberg, S.H., East-Seletsky, A., Kaplan, M., and Doudna, J.A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature *516*, 263-266.

O'Geen, H., Henry, I.M., Bhakta, M.S., Meckler, J.F., and Segal, D.J. (2015). A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. Nucleic Acids Res *43*, 3389-3404.

Okafor, I.C., Singh, D., Wang, Y., Jung, M., Wang, H., Mallon, J., Bailey, S., Lee, J.K., and Ha, T. (2019). Single molecule analysis of effects of non-canonical guide RNAs and specificity-enhancing mutations on Cas9-induced DNA unwinding. Nucleic Acids Res *47*, 11880-11888.

Pacesa, M., and Jinek, M. (2021). Mechanism of R-loop formation and conformational activation of Cas9. bioRxiv, 2021.2009.2016.460614.

1221 Palermo, G., Chen, J.S., Ricci, C.G., Rivalta, I., Jinek, M., Batista, V.S., Doudna, J.A., and McCammon,
1222 J.A. (2018). Key role of the REC lobe during CRISPR-Cas9 activation by 'sensing', 'regulating', and
1223 'locking' the catalytic HNH domain. Q Rev Biophys *51*.

1224 Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., and Liu, D.R. (2013). High-throughput
1225 profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat
1226 Biotechnol *31*, 839-843.

1227 Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and
1228 Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers.
1229 Protein Sci *30*, 70-82.

1230 Ricci, C.G., Chen, J.S., Miao, Y., Jinek, M., Doudna, J.A., McCammon, J.A., and Palermo, G. (2019).
1231 Deciphering Off-Target Effects in CRISPR-Cas9 through Accelerated Molecular Dynamics. ACS Cent
1232 Sci *5*, 651-662.

1233 Rodnina, M.V., and Wintermeyer, W. (2001). Fidelity of aminoacyl-tRNA selection on the ribosome:
1234 kinetic and structural mechanisms. Annu Rev Biochem *70*, 415-435.

1235 Rohs, R., West, S.M., Liu, P., and Honig, B. (2009a). Nuance in the double-helix and its role in protein-
1236 DNA recognition. Curr Opin Struct Biol *19*, 171-177.

1237 Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009b). The role of DNA shape
1238 in protein-DNA recognition. Nature *461*, 1248-1253.

1239 Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J.,
1240 Brouns, S.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic
1241 repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A *108*, 10098-10103.

1242 Singh, D., Sternberg, S.H., Fei, J., Doudna, J.A., and Ha, T. (2016). Real-time observation of DNA
1243 recognition and rejection by the RNA-guided endonuclease Cas9. Nat Commun *7*, 12778.

1244 Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J., and Mateo, J.L. (2015). CCTop: An
1245 Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. PLoS One *10*, e0124633.

1246 Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA
1247 target cleavage by CRISPR-Cas9. Nature *527*, 110-113.

1248 Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by
1249 the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62-67.

1250 Timsit, Y. (1999). DNA structure and polymerase fidelity. J Mol Biol *293*, 835-853.

1251 Tsai, S.Q., Nguyen, N.T., Malagon-Lopez, J., Topkar, V.V., Aryee, M.J., and Joung, J.K. (2017).
1252 CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets.
1253 Nat Methods *14*, 607-614.

1254 Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter,
1255 C., Iafrate, A.J., Le, L.P., *et al.* (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage
1256 by CRISPR-Cas nucleases. Nat Biotechnol *33*, 187-197.
1257 van Houte, S., Ekroth, A.K., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S.,
1258 Boots, M., Paterson, S., Buckling, A., *et al.* (2016). The diversity-generating benefits of a prokaryotic
1259 adaptive immune system. Nature *532*, 385-388.
1260 Vlot, M., Houkes, J., Lochs, S.J.A., Swarts, D.C., Zheng, P., Kunne, T., Mohanraju, P., Anders, C.,
1261 Jinek, M., van der Oost, J., *et al.* (2018). Bacteriophage DNA glucosylation impairs target DNA binding
1262 by type I and II but not by type V CRISPR-Cas effector complexes. Nucleic Acids Res *46*, 873-885.
1263 Wang, W., Hellinga, H.W., and Beese, L.S. (2011). Structural evidence for the rare tautomer hypothesis
1264 of spontaneous mutagenesis. Proc Natl Acad Sci U S A *108*, 17644-17648.
1265 Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E.,
1266 Konermann, S., Chen, S., *et al.* (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in
1267 mammalian cells. Nat Biotechnol *32*, 670-676.

1268 Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M.D. (2006). Base-stacking and base-pairing
1269 contributions into thermal stability of the DNA double helix. Nucleic Acids Res *34*, 564-574.
1270 Yang, M., Peng, S., Sun, R., Lin, J., Wang, N., and Chen, C. (2018). The Conformational Dynamics of
1271 Cas9 Governing DNA Cleavage Are Revealed by Single-Molecule FRET. Cell Rep *22*, 372-382.
1272 Yaung, S.J., Esvelt, K.M., and Church, G.M. (2014). CRISPR/Cas9-mediated phage resistance is not
1273 impeded by the DNA modifications of phage T4. PLoS One *9*, e98811.
1274 Zeng, Y., Cui, Y., Zhang, Y., Zhang, Y., Liang, M., Chen, H., Lan, J., Song, G., and Lou, J. (2018). The
1275 initiation, propagation and dynamics of CRISPR-SpyCas9 R-loop complex. Nucleic Acids Res *46*, 350-
1276 361.
1277 Zhang, L., Rube, H.T., Vakulskas, C.A., Behlke, M.A., Bussemaker, H.J., and Pufall, M.A. (2020).
1278 Systematic in vitro profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes. Nucleic
1279 Acids Research.
1280 Zhu, X., Clarke, R., Puppala, A.K., Chittori, S., Merk, A., Merrill, B.J., Simonovic, M., and
1281 Subramaniam, S. (2019). Cryo-EM structures reveal coordinated domain motions that govern DNA
1282 cleavage by Cas9. Nat Struct Mol Biol *26*, 679-685.

1283