Enhanced proofreading governs CRISPR-Cas9 targeting accuracy

Janice S. Chen[1]*, Yavuz S. Dagdas[2]*, Benjamin P. Kleinstiver[3,4]*, Moira M. Welch[3], Lucas B. Harrington[1], Samuel H. Sternberg[5†], J. Keith Joung[3,4], Ahmet Yildiz[1,6], Jennifer A. Doudna[1,5,7-8]

*These authors contributed equally to this work.

[1]Department of Molecular and Cell Biology, University of California, Berkeley, California, 94720, USA. [2]Biophysics Graduate Group, University of California, Berkeley, California 94720, USA. [3]Molecular Pathology Unit, Center for Cancer Research, and Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [4]Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Department of Chemistry, University of California, Berkeley, California 94720, USA. [6]Department of Physics, University of California, Berkeley, California 94720, USA. [7]Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. [8]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

[†]Present address: Caribou Biosciences, Inc. Berkeley, California 94710, USA.

The RNA-guided CRISPR-Cas9 nuclease from *Streptococcus pyogenes* (SpCas9) has been widely repurposed for genome editing[1-4]. High-fidelity (SpCas9-HF1) and enhanced specificity (eSpCas9(1.1)) variants exhibit substantially reduced off-target cleavage in human cells, but the mechanism of target discrimination and the potential to further improve fidelity were unknown[5-9]. Using single-molecule Förster resonance energy transfer (smFRET) experiments, we show that both SpCas9-HF1 and eSpCas9(1.1) are trapped in an inactive state[10] when bound to mismatched targets. We find that a non-catalytic domain within Cas9, REC3, recognizes target mismatches and governs the HNH nuclease to regulate overall catalytic competence. Exploiting this observation, we identified residues within REC3 involved in mismatch sensing and designed a new hyper-accurate Cas9 variant (HypaCas9) that retains robust on-target activity in human cells. These results offer a more comprehensive model to rationalize and modify the balance between target recognition and nuclease activation for precision genome editing.

Efforts to minimize off-target cleavage by CRISPR-Cas9 have motivated the development of SpCas9-HF1 and eSpCas9(1.1) variants that contain amino acid substitutions predicted to weaken the energetics of target site recognition and cleavage[8,9] **(Figure 1a)**. Biochemically, we found that these Cas9 variants cleaved the on-target DNA with rates similar to that of wild-type (WT) SpCas9, whereas their cleavage activity was significantly reduced on substrates bearing mismatches **(Extended Data Figure 1a, 2a)**. To test the hypothesis that SpCas9 with its single-guide RNA (sgRNA) might exhibit a greater affinity for its target than is required for effective recognition[9,11], we measured DNA binding affinity and cleavage of SpCas9-HF1 and eSpCas9(1.1) variants. Contrary to a potential hypothesis that mutating these charged residues to

46 alanine weakens target binding[11], the affinities of these variants for on-target and PAM-distal
47 mismatched substrates were similar to WT SpCas9 **(Figure 1b, Extended Data Figure 1a, 2b)**,
48 indicating that cleavage specificity is improved through a mechanism distinct from a simple
49 reduction of target binding affinity[11].

50 The HNH nuclease domain of SpCas9 undergoes a substantial conformational
51 rearrangement upon target binding[12-15], which activates the RuvC nuclease for concerted
52 cleavage of both strands of the DNA[12,16]. We have previously shown that the HNH domain
53 stably docks in its active state with an on-target substrate but becomes loosely trapped in an
54 catalytically-inactive conformational checkpoint when bound to mismatched targets[10],
55 suggesting that SpCas9-HF1 and eSpCas9(1.1) variants may employ a more stringent checkpoint
56 to promote off-target discrimination. To test this possibility, we labeled catalytically active WT
57 SpCas9 ($SpCas9_{HNH}$), SpCas9-HF1 ($SpCas9\text{-}HF1_{HNH}$) and eSpCas9(1.1) ($eSpCas9(1.1)_{HNH}$) with
58 Cy3/Cy5 FRET pairs at positions S355C and S867C to measure HNH conformational states
59 **(Figure 1c–f, Extended Data Figure 1c–e)**[12]. Whereas $SpCas9_{HNH}$ stably populated the active
60 state with both on-target and mismatched substrates in steady-state smFRET histograms **(Figure**
61 **1d)**, only ~32% of $SpCas9\text{-}HF1_{HNH}$ molecules occupied the HNH active state (E_{FRET} = 0.97)
62 with an on-target substrate, with the remaining ~68% trapped in the inactive intermediate state
63 (E_{FRET} = 0.45) **(Figure 1e)**. However, when $SpCas9\text{-}HF1_{HNH}$ was bound to a substrate with just a
64 single nucleotide mismatch at the PAM-distal end (20-20 bp mm), stable docking of the HNH
65 nuclease was entirely ablated **(Figure 1e)**. In addition, $eSpCas9(1.1)_{HNH}$ and other high fidelity
66 variants[8,9] reduced the HNH active state in the presence of mismatches **(Figure 1f, Extended**
67 **Data Figure 2c–d)**. We therefore propose that high fidelity variants of Cas9 reduce off-target

68 cleavage by raising the threshold for HNH conformational activation when bound to DNA
69 substrates.

70 Since the HNH domain does not directly contact nucleic acids at the PAM-distal end[13,17-
71 19], it is likely that a separate domain of Cas9 senses mismatches to govern HNH domain
72 mobility. Structural studies suggest that a domain within the Cas9 recognition (REC) lobe
73 (REC3) interacts with the RNA/DNA heteroduplex and undergoes conformational changes upon
74 target binding **(Extended Data Figure 3)**[13,14,17-19]. Because the function of this non-catalytic
75 domain was previously unknown, we labeled SpCas9 with Cy3/Cy5 dyes at positions S701C and
76 S960C ($SpCas9_{REC3}$) and observed that the conformational states of REC3 become more
77 heterogeneous as PAM-distal mismatches increase **(Extended Data Figure 4a-c)**. To determine
78 whether PAM-distal sensing precedes HNH activation, we deleted REC3 from WT Cas9
79 (SpCas9ΔREC3) **(Figure 2a)**. Deletion of REC3 decreased the cleavage rate by ~1000-fold
80 compared to WT Cas9, despite retaining near-WT binding affinity with a perfect target
81 **(Extended Data Figure 4d–f)**. Unexpectedly, *in vitro* complementation of REC3 domain *in*
82 *trans* rescued the on-target cleavage rate by ~100-fold in a concentration-dependent manner, but
83 had no effect on cleavage with a PAM-distal mismatched target **(Figure 2b, Extended Data**
84 **Figure 4e)**. Furthermore, smFRET experiments revealed that the HNH domain in
85 SpCas9ΔREC3 (SpCas9Δ $REC3_{HNH}$) occupied the active state only when REC3 was
86 supplemented *in trans* **(Figure 2c–d, Extended Data Figure 4f)**. We therefore propose that
87 REC3 acts as an allosteric effector that recognizes RNA/DNA heteroduplex to allow for HNH
88 nuclease activation.

89 We next considered allosteric interactions that could couple the discontinuous REC3 and
90 HNH domains. Structural studies suggested that REC2 occludes the HNH domain from the

91 scissile phosphate in the sgRNA-bound state[19], and undergoes a large outward rotation upon
92 binding to double-stranded DNA (dsDNA)[13,14] **(Figure 2e)**. To test whether the REC2 domain
93 regulates access of HNH to the target strand scissile phosphate, we labeled SpCas9 with
94 Cy3/Cy5 dyes at positions E60C and D273C ($SpCas9_{REC2}$) in order to detect REC2
95 conformational changes **(Extended Data Figure 1b–c)**. We observed reciprocal changes in bulk
96 FRET values ($(ratio)_A$)[20] between $SpCas9_{HNH}$ and $SpCas9_{REC2}$ across multiple DNA substrates
97 **(Extended Data Figure 4g)**, which suggest that the REC2 and HNH domains are tightly coupled
98 to ensure catalytic competence. smFRET experiments further confirmed a large opening of
99 REC2 during the transition from the sgRNA-bound state (E_{FRET} = 0.96) to the target-bound state
100 (E_{FRET} = 0.43) **(Figure 2e–f)**. In contrast to WT $SpCas9_{REC2}$, SpCas9-$HF1_{REC2}$ occupies an
101 intermediate state (E_{FRET} = 0.63) when bound to a target with just a single PAM-distal mismatch
102 **(Figure 2f–g)**. Together with the observation that the HNH domain of SpCas9-HF1 does not
103 occupy the active state in the presence of PAM-distal mismatches, these experiments suggest that
104 REC2 sterically occludes and traps the HNH nuclease domain in the conformational checkpoint
105 when SpCas9 is bound to off-target substrates.

106 Next, we investigated if this conformational proofreading mechanism could be rationally
107 exploited to design a suite of novel hyper-accurate Cas9 variants. We identified five clusters of
108 residues containing conserved amino acids within 5 Å of the RNA/DNA interface, four of which
109 are located within REC3 and one in the HNH-RuvC Linker 2 (L2) **(Figure 3a, Extended Data**
110 **Figure 5)**. Alone or in combination with Q926A, a substitution within L2 that confers
111 specificity[9], we generated alanine substitutions for each residue within five different clusters of
112 amino acids (Clusters 1-5 ± Q926A) **(Figure 3a)**. We tested whether these cluster mutations
113 affected off-target discrimination and equilibrium binding *in vitro*, and found that Cluster 1 alone

114 and Cluster 2 + Q926A exhibited the greatest suppression of off-target cleavage while retaining
115 target binding affinities comparable to WT **(Extended Data Figure 6)**. We next screened all
116 cluster variants in human cells using an enhanced GFP (*EGFP*) disruption assay[5]. On-target
117 activity for Cluster 1 was comparable to that of SpCas9-HF1 or eSpCas9(1.1), whereas Cluster 2
118 variants displayed generally lower activity **(Figure 3b, Extended Data Figure 7a)**. Furthermore,
119 Cluster 1 retained high on-target activity (> 70% of WT) at 19/24 endogenous gene sites tested,
120 compared to 18/24 for SpCas9-HF1 and 23/24 for eSpCas9(1.1) **(Figure 3c, Extended Data**
121 **Figure 8a)**.
122 We then focused on the specific contributions of mutations within Cluster 1 by restoring
123 each individual mutated residue to its wild-type identity, along with the Q926A mutation, and
124 testing the resulting variants for on-target editing efficiency in human cells. On-target activity
125 was significantly compromised when N692A/Q695A/Q926A mutations occurred together, but
126 restoring either N692 (Cluster 1 N692 + Q926A) or Q695 (Cluster 1 Q695 + Q926A) alone led
127 to robust on-target efficiency comparable to Cluster 1, signifying differential contributions from
128 these mutations to activity and specificity (**Extended Data Figure 7b-c**, **8a-b**). Using sgRNAs
129 with single mismatches against the endogenous human gene target *FANCF* site 1, we found that
130 Cluster 1 exhibited even greater specificity than both SpCas9-HF1 and eSpCas9(1.1) in the
131 middle and PAM proximal regions of the spacer, suggesting that mutating N692A and Q695A
132 together may induce specificity in parts of the spacer sequence that were previously susceptible
133 to off-target cleavage by high-fidelity Cas9 variants[9] **(Figure 3d, Extended Data Figure 8c)**.
134 Additional single mismatch tolerance assays on *FANCF* sites 4 and 6 further corroborated the
135 superior accuracy of Cluster 1 (N692A/M694A/Q695A/H698A, referred to as HypaCas9)

136 against mismatches at positions 1 through 18; however, single mismatches along *FANCF* site 2
137 were still tolerated across all SpCas9 variants tested **(Figure 3e, Extended Data Figure 8d, e)**.
138 To biochemically validate cleavage specificity in the middle region of the spacer with
139 HypaCas9, we measured cleavage rates against the *FANCF* site 1 sequence with or without
140 internal mismatches at positions 10-12 of the spacer. Although HypaCas9 retained on-target
141 activity comparable to WT and SpCas9-HF1 in human cells, its *in vitro* cleavage rate was
142 slightly reduced for the one target site examined **(Figure 4a)**. However, the cleavage rate with
143 internally mismatched substrates was considerably slower compared to WT and SpCas9-HF1
144 **(Figure 4a)**. This activity may be explained by the altered threshold of HNH domain activation;
145 whereas stable HNH docking was observed by $SpCas9_{HNH}$ and $SpCas9\text{-}HF1_{HNH}$ with both the
146 *FANCF* site 1 on-target and mismatched substrate at the 12th position, this HNH active state by
147 HypaCas9 ($HypaCas9_{HNH}$) was diminished with the on-target. Nevertheless, HNH docking was
148 nearly abolished when $HypaCas9_{HNH}$ was bound to a substrate with a single mismatch **(Figure**
149 **4b)**.
150 Our findings provide direct evidence to support previous speculation that Cas9 relies on
151 PAM-distal protospacer sensing to enable accurate targeting[21,22]. In particular, we define REC3
152 as an allosteric regulator of global Cas9 conformational changes to activate the nuclease
153 domains, whose conformational threshold can be tuned for high-fidelity cleavage. Mutation of
154 residues within REC3 that are involved in nucleic acid recognition, such as those mutated in
155 HypaCas9 or SpCas9-HF1, prevents transitions by the REC lobe, which more stringently traps
156 the HNH domain in the conformational checkpoint in the presence of mismatches **(Figure 4c,**
157 **Extended Data Figure 9)**. Curiously, nearly all of the amino acids within the cluster variants
158 were strongly conserved **(Extended Data Figure 5)**, suggesting that these residues may also be

involved in protospacer sensing and HNH nuclease activation across Cas9 orthologues. Furthermore, this observation may address how nature apparently has not selected for a highly precise Cas9 protein, whose native balance between mismatch tolerance and specificity may be optimized for host immunity. Our study therefore delineates a general strategy for improving Cas9 specificity by tuning conformational activation and offers innovative opportunities for rational design of hyper-accurate Cas9 variants that do not compromise efficiency.

182 **METHODS**

183

184 **Protein purification and dye labeling.** *S. pyogenes* Cas9 and truncation derivatives were cloned
185 into a custom pET-based expression vector containing an N-terminal His_6-tag, maltose-binding
186 protein (MBP) and TEV protease cleavage site. Point mutations were introduced by Gibson
187 assembly or around-the-horn PCR and verified by DNA sequencing. Proteins were purified as
188 described[23], with the following modifications: after Ni-NTA affinity purification and overnight
189 TEV cleavage at 4°C, proteins were purified over an MBPTrap HP column connected to a
190 HiTrap Heparin HP column for cation exchange chromatography. The final gel filtration step
191 (Superdex 200) was carried out in elution buffer containing 20 mM Tris-HCl pH 7.5, 200 mM
192 NaCl, 5% glycerol (v/v) and 1 mM TCEP. For FRET experiments, dye-labeled Cas9 samples
193 were prepared as described[12]. A list of all protein variants and truncations are listed in
194 **Supplementary Table 1**.

195

196 **Nucleic acid preparation**. sgRNA templates were PCR amplified from a pUC19 vector
197 containing a T7 promoter, 20 nt target sequence and optimized sgRNA scaffold. The amplified
198 PCR product was extracted with phenol:chloroform:isoamylalcohol and served as the DNA
199 template for sgRNA transcription reactions, which were performed as described[24]. DNA
200 oligonucleotides and 5′end biotinylated DNAs (Supplementary Table 2) were synthesized
201 commercially (Integrated DNA Technologies), and DNA duplexes were prepared and purified by
202 native PAGE as described[23].

203

204 **DNA cleavage and binding assays.** DNA duplex substrates were 5′-[^{32}P]-radiolabeled on both
205 strands. For cleavage experiments, Cas9 and sgRNA were pre-incubated at room temperature for
206 at least 10 min in 1X binding buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1
207 mM DTT, 5% glycerol, 50 μg ml^{-1} heparin) before initiating the cleavage reaction by addition of
208 DNA duplexes. For REC3 *in vitro* complementation experiments, SpCas9ΔREC3 and sgRNA
209 were pre-incubated with 10-fold molar excess of REC3 for at least 10 minutes at room
210 temperature before addition of radiolabeled substrate. DNA cleavage experiments were
211 performed and analyzed as previously described[12]. DNA binding assays were conducted in 1X
212 binding buffer without $MgCl_2$ + 1 mM EDTA at room temperature for 2 hours. DNA-bound
213 complexes were resolved on 8% native PAGE (0.5X TBE + 1 mM EDTA, without $MgCl_2$) at
214 4°C, as previously described[10]. Experiments were replicated at least three times, and presented
215 gels are representative results.
216
217 **Bulk FRET experiments.** All bulk FRET assays were performed at room temperature in 1×
218 binding buffer, containing 50 nM $SpCas9_{HNH}$ (C80S/S355C/C574S/S867C labeled with
219 Cy3/Cy5), $SpCas9ΔREC3_{HNH}$(M1–N497,GGS,V713–D1368 + C80S/S355C/C574S/S867C) or
220 $SpCas9_{REC2}$ (E60C/C80S/D273C/C574S labeled with Cy3/Cy5) with 200 nM sgRNA and DNA
221 substrate where indicated. Fluorescence measurements were collected and analyzed as
222 described[12]. For REC3 *in vitro* complementation FRET experiments, $SpCas9ΔREC3_{HNH}$ and
223 sgRNA were pre-incubated with 10-fold molar excess of REC3 for at least 10 minutes at room
224 temperature before measuring bulk fluorescence.
225

226 **Sample preparation for smFRET assay.** 99% PEG and 1% biotinylated-PEG coated quartz
227 slides were received from MicroSurfaces, Inc. Sample preparation was performed as previously
228 described[10]. To immobilize SpCas9 on its DNA substrate, 2.5nM biotinylated-DNA substrate
229 introduced and incubated in sample chamber for 5 min. Excess DNA was washed with 1×
230 binding buffer. SpCas9-sgRNA complexes were prepared by mixing 50 nM Cas9 and 50nM
231 sgRNA in 1× binding buffer and incubated for 10 min at room temperature. SpCas9-sgRNA was
232 diluted to 100 pM, introduced to sample chamber and incubated for 10 min. Before data
233 acquisition, 20 μL imaging buffer (1 mg ml^{-1} glucose oxidase, 0.04 mg ml^{-1} catalase, 0.8%
234 dextrose (w/v) and 2 mM Trolox in 1× binding buffer) was flown into chamber. The REC3 *in*
235 *vitro* complementation assay was performed similar to steady-state FRET experiments: 2.5nM
236 biotinylated-DNA substrate (on-target) was immobilized on surface, and excess DNA was
237 washed with 1× binding buffer. SpCas9-sgRNA complexes were prepared by mixing 50 nM
238 SpCas9ΔREC3 and 50nM sgRNA in 1× binding buffer and incubated for 10 min at room
239 temperature. SpCas9-sgRNA was diluted to 100 pM, introduced to the sample chamber and
240 incubated for 10 min. Before data acquisition, 20 μL imaging buffer was flown into chamber.
241 After data acquisition, the sample chamber was washed with 1× binding buffer. 20 μL imaging
242 buffer supplemented with 1μM REC3 was flown into sample chamber and incubated for 10min.
243 After incubation, data for REC3 complementation was collected.
244
245 **Microscopy and data analysis.** A prism-type TIRF microscope was setup using a Nikon Ti-E
246 Eclipse inverted fluorescent microscope equipped with a 60× 1.20 N.A. Plan Apo water objective
247 and the perfect focusing system (Nikon). A 532-nm solid state laser (Coherent Compass) and a
248 633-nm HeNe laser (JDSU) were used for Cy3 and Cy5 excitation, respectively. Cy3 and Cy5

249 fluorescence were split into two channels using an Optosplit II image splitter (Cairn Instruments)
250 and imaged separately on the same electron-multiplied charged-coupled device (EM-CCD)
251 camera (512×512 pixels, Andor Ixon EM^+). Effective pixel size of the camera was set to 267 nm
252 after magnification. Movies for steady-state FRET measurements were acquired at 10 Hz under
253 0.3 kW cm^{-2} 532-nm excitation. Data analysis was performed as described previously[10]. Briefly,
254 two fluorescent channels were registered with each other using fiducial markers (20 nm diameter
255 Nile Red Beads, Life Technologies) to determine the Cy3/Cy5 FRET pairs. Cy3/Cy5 pairs that
256 photobleached in one step and showed anti-correlated signal changes were used to build
257 histograms. FRET values were corrected for donor leakage and the histograms were normalized
258 to determine the percentage of distinct FRET populations.
259
260 **Human cell culture and transfection.** Descriptions of nuclease and guide RNA plasmids used
261 for human cell culture are available in **Supplementary Table 1 and 2**. Nuclease variants were
262 generated by isothermal assembly into JDS246 (Addgene #43861)[5], and guide RNAs were
263 cloned into BsmBI digested BPK1520 (Addgene #65777)[25]. Both U2OS cells (a gift from Toni
264 Cathomen, Freiburg) and U2OS-EGFP cells (encoding a single integrated copy of a pCMV-
265 EGFP–PEST cassette)[26] were cultured at 37 °C with 5% CO_2 in advanced DMEM containing
266 10% heat-inactivated fetal bovine serum, 2 mM GlutaMax, penicillin/streptomycin, and 400 µg
267 ml^{-1} Geneticin (for U2OS-EGFP cells only). Cell culture reagents were purchased from Thermo
268 Fisher Scientific, cell line identities were validated by STR profiling (ATCC) and deep-
269 sequencing, and cell culture supernatant was tested bi-weekly for mycoplasma. Transfections
270 were performed using a Lonza 4-D Nucleofector with the SE Kit and the DN-100 program on
271 ~200k cells with 750 ng of nuclease and 250 ng of guide RNA plasmids.

272

273 **Human cell EGFP disruption assay.** *EGFP* disruption experiments were performed as
274 previously described[5,26]. Briefly, transfected cells were analyzed ~52 hours post-transfection for
275 loss of EGFP fluorescence using a Fortessa flow cytometer (BD Biosciences). Background loss
276 was determined by gating a negative control transfection (containing nuclease and empty guide
277 RNA plasmid) at ~2.5% for all experiments.

278

279 **T7 endonuclease I assay.** Roughly 72 hours post-transfection, genomic DNA was extracted
280 from U2OS cells using the Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman
281 Coulter Genomics), and T7 endonuclease I assays were performed as previously described[26].
282 Briefly, 600–800 nt amplicons surrounding on-target sites were amplified from ~100 ng of
283 genomic DNA using Phusion Hot-Start Flex DNA Polymerase (New England Biolabs) using the
284 primers listed in Supplementary Table 2. PCR products were visualized (using a QIAxcel
285 capillary electrophoresis instrument, Qiagen), and purified (Agencourt Ampure XP cleanup,
286 Beckman Coulter Genomics), Denaturation and annealing of ~200 ng of the PCR product was
287 followed by digestion with T7 endonuclease I (New England Biolabs). Digestion products were
288 purified (Ampure) and quantified (QIAxcel) to approximate the mutagenesis frequencies induced
289 by Cas9-sgRNA complexes.

290

291

292

293

294

295 **REFERENCES**

296 1 Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome
297 engineering with CRISPR-Cas9. *Science* **346**, 1258096, doi:10.1126/science.1258096
298 (2014).

299 2 Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for
300 genome engineering. *Cell* **157**, 1262-1278, doi:10.1016/j.cell.2014.05.010 (2014).

301 3 Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology.
302 *Nat Methods* **10**, 957-963, doi:10.1038/nmeth.2649 (2013).

303 4 Barrangou, R. & Horvath, P. A decade of discovery: CRISPR functions and applications.
304 *Nat Microbiol* **2**, 17092, doi:10.1038/nmicrobiol.2017.92 (2017).

305 5 Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in
306 human cells. *Nat Biotechnol* **31**, 822-826, doi:10.1038/nbt.2623 (2013).

307 6 Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by
308 CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197, doi:10.1038/nbt.3117 (2015).

309 7 Tsai, S. Q. & Joung, J. K. Defining and improving the genome-wide specificities of
310 CRISPR-Cas9 nucleases. *Nat Rev Genet* **17**, 300-312, doi:10.1038/nrg.2016.28 (2016).

311 8 Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity.
312 *Science* **351**, 84-88, doi:10.1126/science.aad5227 (2016).

313 9 Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable
314 genome-wide off-target effects. *Nature* **529**, 490-495, doi:10.1038/nature16526 (2016).

315 10 Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A
316 Conformational Checkpoint Between DNA Binding And Cleavage By CRISPR-Cas9.
317 *bioRxiv* (2017).

318 11 Bisaria, N., Jarmoskaite, I. & Herschlag, D. Lessons from Enzyme Kinetics Reveal
319 Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-
320 Based Genome Editing. *Cell Syst* **4**, 21-29, doi:10.1016/j.cels.2016.12.010 (2017).
321 12 Sternberg, S. H., LaFrance, B., Kaplan, M. & Doudna, J. A. Conformational control of
322 DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110-113, doi:10.1038/nature15544
323 (2015).
324 13 Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage.
325 *Science* **351**, 867-871, doi:10.1126/science.aad8282 (2016).
326 14 Palermo, G., Miao, Y., Walker, R. C., Jinek, M. & McCammon, J. A. Striking Plasticity
327 of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular
328 Simulations. *ACS Cent Sci* **2**, 756-763, doi:10.1021/acscentsci.6b00218 (2016).
329 15 Palermo, G., Miao, Y., Walker, R. C., Jinek, M. & McCammon, J. A. CRISPR-Cas9
330 conformational activation as elucidated from enhanced molecular simulations. *Proc Natl*
331 *Acad Sci U S A*, doi:10.1073/pnas.1707645114 (2017).
332 16 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive
333 bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).
334 17 Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target
335 DNA. *Cell* **156**, 935-949, doi:10.1016/j.cell.2014.02.001 (2014).
336 18 Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent
337 target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569-573,
338 doi:10.1038/nature13579 (2014).

339 19 Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. STRUCTURAL BIOLOGY. A
340 Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348**, 1477-
341 1481, doi:10.1126/science.aab1452 (2015).
342 20 Majumdar, Z. K., Hickerson, R., Noller, H. F. & Clegg, R. M. Measurements of internal
343 distance changes of the 30S ribosome using FRET with multiple donor-acceptor pairs:
344 quantitative spectroscopic methods. *J Mol Biol* **351**, 1123-1145,
345 doi:10.1016/j.jmb.2005.06.027 (2005).
346 21 Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided
347 Cas9 and Cascade effector complexes. *Proc Natl Acad Sci U S A* **111**, 9798-9803,
348 doi:10.1073/pnas.1402597111 (2014).
349 22 Cencic, R. *et al.* Protospacer adjacent motif (PAM)-distal sequences engage CRISPR
350 Cas9 DNA target cleavage. *PLoS One* **9**, e109213, doi:10.1371/journal.pone.0109213
351 (2014).
352 23 Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational
353 activation. *Science* **343**, 1247997, doi:10.1126/science.1247997 (2014).
354 24 Wright, A. V. *et al.* Rational design of a split-Cas9 enzyme complex. *Proc Natl Acad Sci*
355 *U S A* **112**, 2984-2989, doi:10.1073/pnas.1501698112 (2015).
356 25 Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM
357 specificities. *Nature* **523**, 481-485, doi:10.1038/nature14592 (2015).
358 26 Reyon, D. *et al.* FLASH assembly of TALENs for high-throughput genome editing. *Nat*
359 *Biotechnol* **30**, 460-465, doi:10.1038/nbt.2170 (2012).
360
361

362 **ACKNOWLEDGEMENTS**

363 We thank Addison V. Wright, Stephen N. Floor, Joshua C. Cofsky, David Burstein, Christof
364 Fellman, Benjamin L. Oakes and Orestes Mavrothalassitis for discussions and critical reading of
365 the manuscript, and Michelle S. Prew for technical assistance. J.S.C. acknowledges support from
366 the National Science Foundation Graduate Research Fellowship program, and B.P.K. from
367 Banting (Natural Sciences and Engineering Research Council of Canada) and Charles A. King
368 Trust Postdoctoral Fellowships. J.A.D. is an Investigator of the Howard Hughes Medical
369 Institute. This work has been supported by NIH (GM094522 and GM118773 (A.Y.), R35
370 GM118158 (J.K.J.)), NSF (MCB-1617028 (A.Y.) and MCB-1244557 (J.A.D.)), and the
371 Desmond and Ann Heathwood MGH Research Scholar Award (J.K.J.).

372

373 **AUTHOR CONTRIBUTIONS**

374 J.S.C., Y.S.D. and B.P.K. conceived of and designed experiments with input from L.B.H., S.H.S,
375 J.K.J., A.Y. and J.A.D. J.S.C. performed protein expression, labeling and biochemical
376 experiments. Y.S.D. performed single-molecule fluorescence assays and data analysis. B.P.K.
377 and M.M.W. performed human cell-based experiments. J.S.C., Y.S.D., B.P.K., J.K.J., A.Y. and
378 J.A.D. wrote the manuscript.

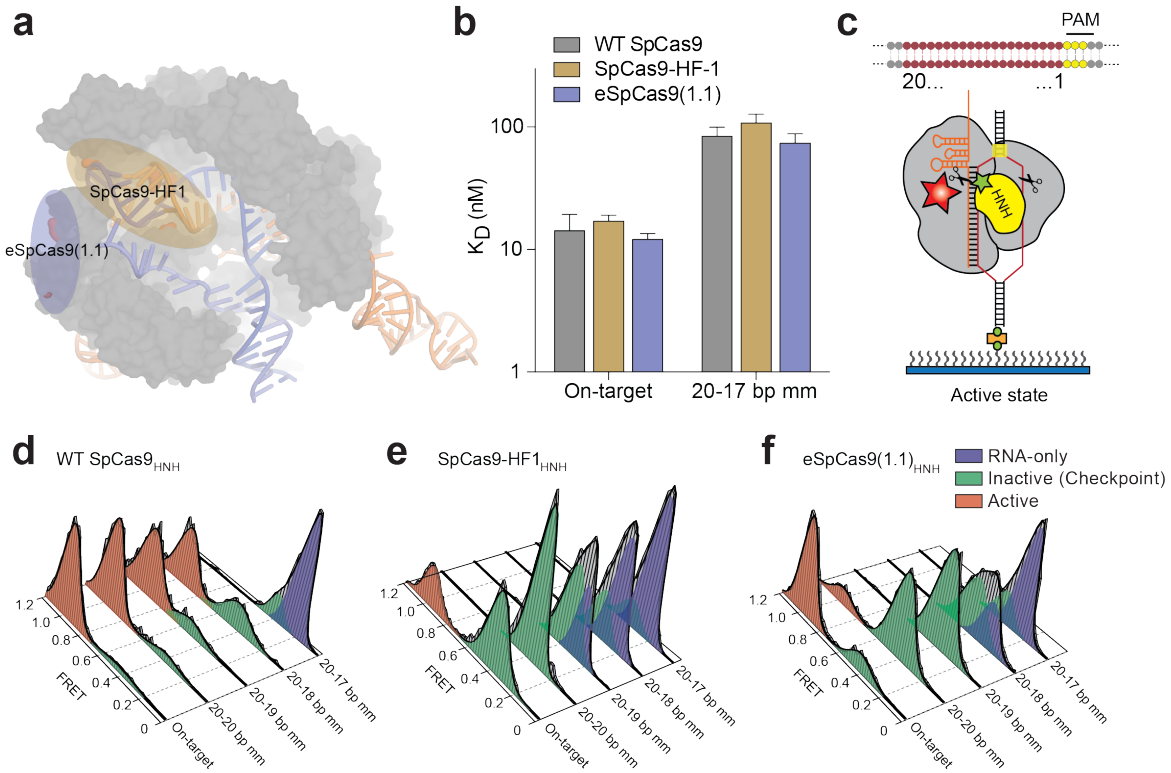
379

380 **COMPETING FINANCIAL INTERESTS**

381 J.K.J. has financial interests in Beacon Genomics, Beam Therapeutics, Editas Medicine, Pairwise
382 Plants, Poseida Therapeutics, and Transposagen Biopharmaceuticals. J.K.J.'s interests were
383 reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in
384 accordance with their conflict of interest policies. J.A.D. is a co-founder of Caribou Biosciences,

385 Editas Medicine, and Intellia Therapeutics; a scientific advisor to Caribou, Intellia, eFFECTOR
386 Therapeutics and Driver; and executive director of the Innovative Genomics Institute at UC
387 Berkeley and UCSF. S.H.S. is an employee of Caribou Biosciences, Inc. and an inventor on
388 patent applications related related to CRISPR-Cas systems and uses thereof. J.S.C, Y.S.D.,
389 B.P.K, L.B.H., S.H.S., A.Y., J.K.J. and J.A.D. are inventors on patents for CRISPR technologies.
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407

408 **FIGURE 1**



409

410 **Figure 1 | High-fidelity Cas9 variants enhance cleavage specificity through HNH**

411 **conformational control. a,** Locations of amino acid alterations present in existing high-fidelity

412 SpCas9 variants mapped onto the dsDNA-bound SpCas9 crystal structure (5F9R), with the HNH

413 domain omitted for clarity. **b,** Dissociation constants comparing WT SpCas9, SpCas9-HF1 and

414 eSpCas9(1.1) with perfect and a 20-17 bp mismatched target. Error bars, s.d.; $n = 3$. **c,** Cartoon

415 of DNA-immobilized SpCas9 complexes for smFRET experiments with DNA target numbering

416 scheme. **d–f,** smFRET histograms measuring HNH conformational activation with **d,** WT

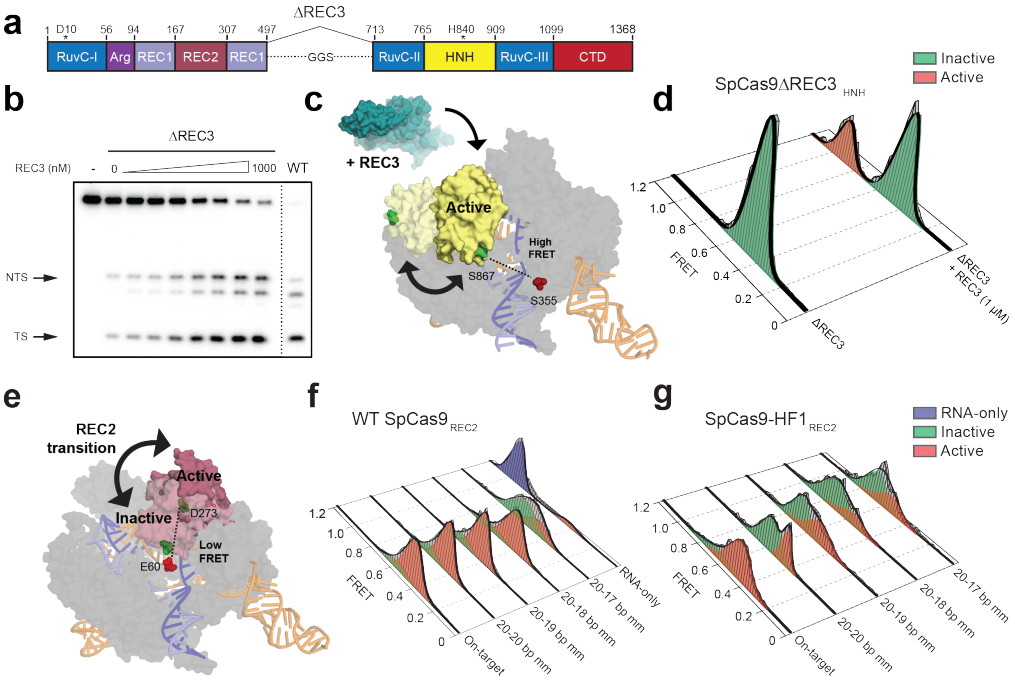
417 SpCas9$_{HNH}$, **e,** SpCas9-HF1$_{HNH}$ and **f,** eSpCas9(1.1)$_{HNH}$ bound to perfect and PAM-distal

418 mismatched targets. Black curves represent a fit to multiple Gaussian peaks.

419

420

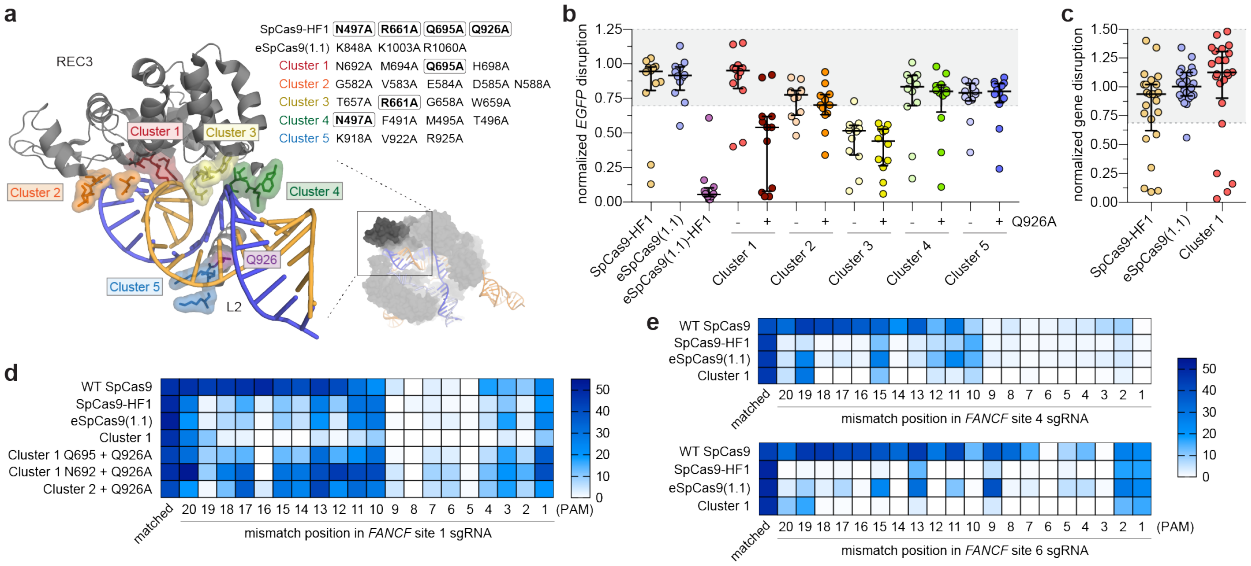
421 **FIGURE 2**



422

423 **Figure 2 | The alpha-helical lobe regulates HNH domain activation. a,** Domain organization
424 of SpCas9ΔREC3. **b,** Perfect target DNA cleavage assay using SpCas9ΔREC3 with increasing
425 concentrations of REC3 domain supplied in *trans*, resolved by denaturing PAGE. **c,** Schematic
426 of SpCas9ΔREC3$_{HNH}$ with FRET dyes at positions S355C and S867C, with the REC3 domain
427 added *in trans*. Inactive to active structures represent HNH in the sgRNA-bound (PDB ID:
428 4ZT0) to dsDNA-bound (PDB ID: 5F9R) forms, respectively. **d,** smFRET histograms measuring
429 HNH conformational states with SpCas9ΔREC3$_{HNH}$ in the absence and presence of the REC3
430 domain. **e,** Schematic of SpCas9$_{REC2}$ with FRET dyes at positions E60C and D273C, with HNH
431 domain omitted for clarity. Inactive to active structures represent REC2 in the sgRNA-bound
432 (PDB ID: 4ZT0) to dsDNA-bound (PDB ID: 5F9R) forms, respectively. **f–g,** smFRET
433 histograms measuring REC2 conformational states with **f,** WT SpCas9$_{REC2}$ and **g,** SpCas9-
434 HF1$_{REC2}$ bound to perfect and PAM-distal mismatched targets. For panels **d, f** and **g,** black
435 curves represent a fit to multiple Gaussian peaks.

436 **FIGURE 3**



437

438 **Figure 3 | Targeted mutagenesis within the REC3 domain reveals a SpCas9 variant with**

439 **hyper-accurate behavior in human cells. a,** Zoomed image of the REC3 domain and Linker 2

440 (L2) with Cluster variants indicated. Boxed residues indicate amino acids also present in

441 SpCas9-HF1. **b,** WT-normalized activity of SpCas9-HF1, eSpCas9(1.1) and Cluster variants,

442 using sgRNAs targeting 12 different sites within *EGFP*. **c,** WT-normalized endogenous gene

443 disruption activity measured by T7 endonuclease 1 (T7E1) assay across 24 sites. **d,** Activities of

444 WT and high-fidelity Cas9 variants when programmed with singly mismatched sgRNAs against

445 *FANCF* site 1 as measured by T7E1 assay. **e,** Activities of WT SpCas9, SpCas9-HF1,

446 eSpCas9(1.1) and Cluster 1 when programmed with singly mismatched sgRNAs against *FANCF*

447 site 4 and *FANCF* site 6. For panels **b** and **c**, error bars represent median and interquartile ranges,

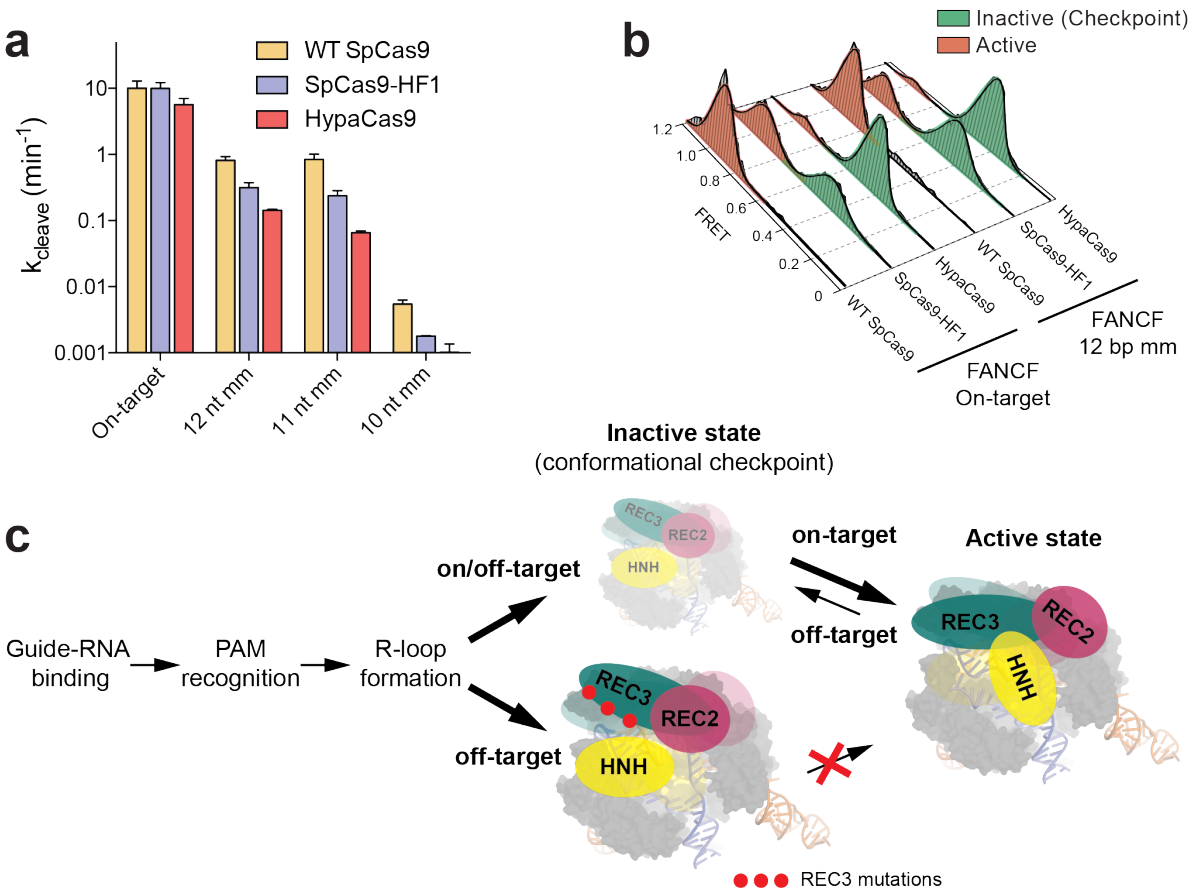
448 and n ≥ 3; the interval with > 70% of wild-type activity is highlighted in light grey.

449

450

451

452 **FIGURE 4**



453

454 **Figure 4 | Mutating residues involved in proofreading increases the threshold for**

455 **conformational activation to ensure targeting accuracy. a,** DNA cleavage kinetics of SpCas9

456 variants with the *FANCF* site 1 on-target and internally mismatched substrates. Error bars, s.d.; *n*

457 = 3. **b,** smFRET histograms measuring HNH conformational states for indicated SpCas9 variants

458 with a *FANCF* site 1 on-target and mismatched substrate at the 12th position; black curves

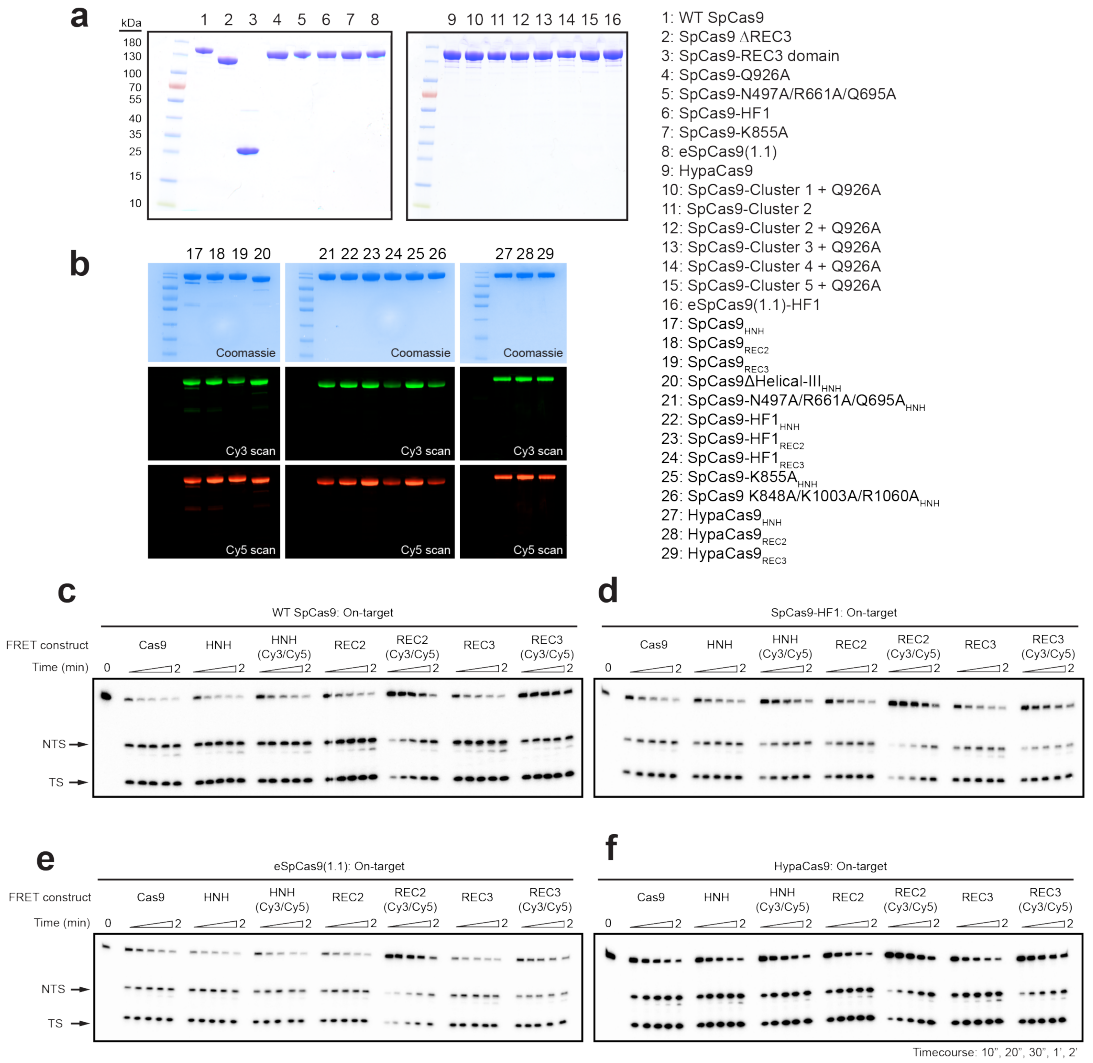
459 represent a fit to multiple Gaussian peaks. **c,** Model for alpha-helical lobe sensing and regulation

460 of the RNA/DNA heteroduplex for HNH activation and cleavage.

461

462

463 **EXTENDED DATA FIGURE 1**



464

465 **Extended Data Figure 1 | Dually-labeled SpCas9 variants are fully functional for DNA**

466 **cleavage. a,** Sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS–PAGE)

467 analysis of unlabeled Cas9 variants. **b,** SDS-PAGE analysis of Cy3/Cy5-labeled Cas9 variants.

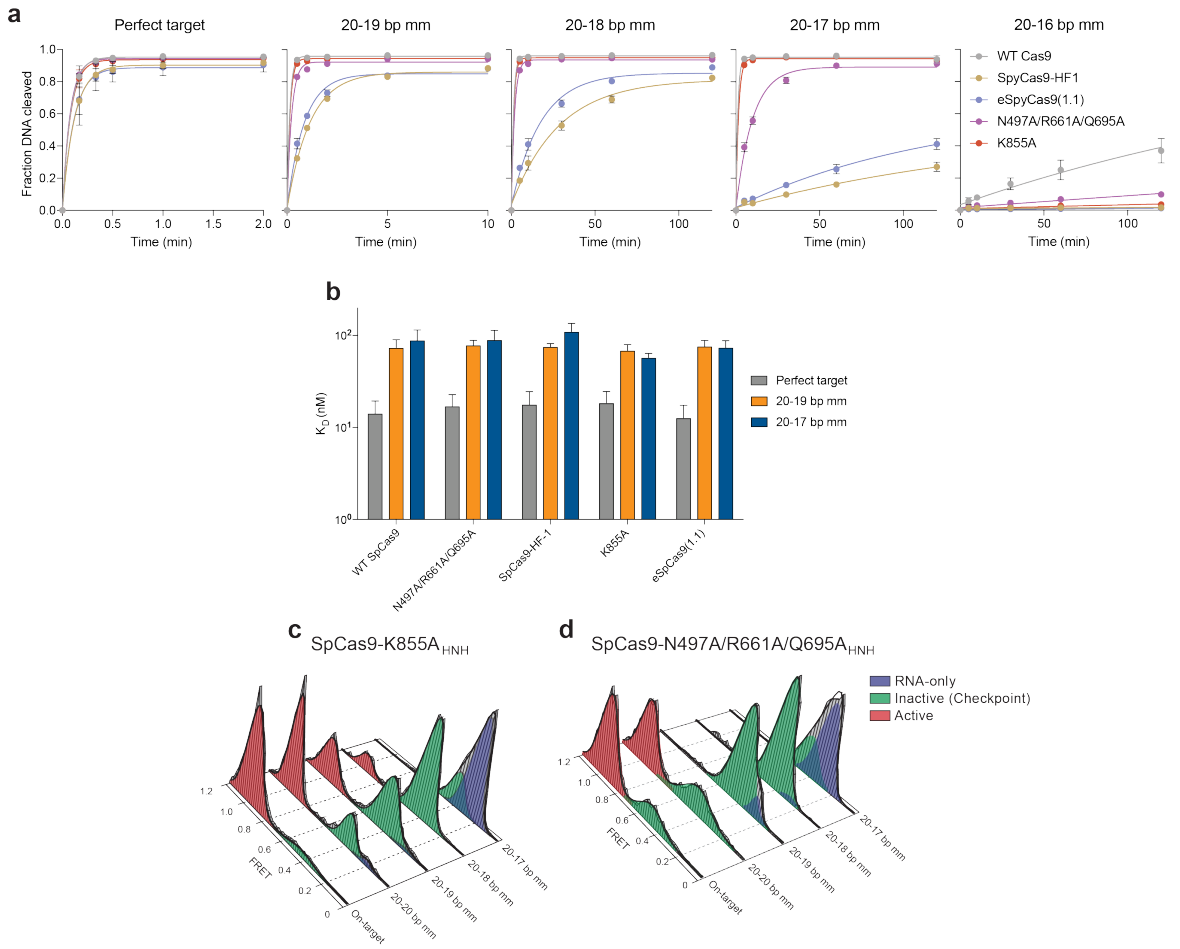
468 The gel was scanned for Cy3/Cy5 fluorescence (middle, bottom) before

469 staining with Coomassie blue (top). **c–f,** DNA cleavage time courses of Cas9 FRET constructs

470 and their dually-labeled counterparts for **c,** WT SpCas9, **d,** SpCas9-HF1, **e,** eSpCas9(1.1) and **f**,

471 HypaCas9.

472 **EXTENDED DATA FIGURE 2**



473

474 **Extended Data Figure 2 | HNH domain in eSpCas9 variants still populate the docked state**

475 **in the presence of PAM-distal mismatches. a,** Quantification of DNA cleavage time courses

476 comparing WT SpCas9, SpCas9-HF and eSpCas9(1.1) variants with perfect and PAM-distal

477 mismatched targets. **b,** Dissociation constants comparing WT SpCas9, SpCas9-HF and

478 eSpCas9(1.1) variants with perfect and PAM-distal mismatched targets, as measured by

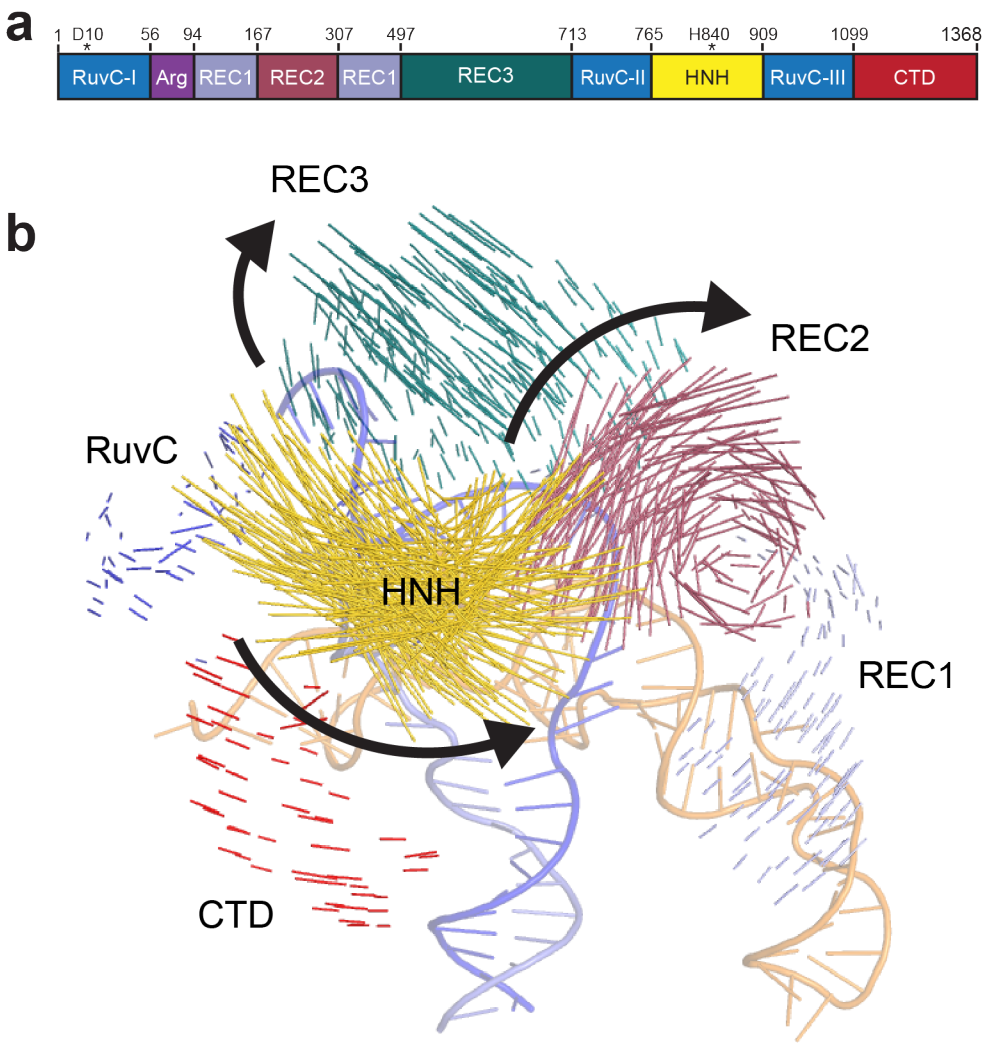
479 electrophoretic mobility shift assays. Error bars in **a** and **b**, s.d.; $n = 3$. **c–d,** smFRET histograms

480 for **c,** SpCas9-K855A and **d,** SpCas9-N497A/R661A/Q695A. For panels **c** and **d**, black curves

481 represent a fit to multiple Gaussian peaks.

482

483 **EXTENDED DATA FIGURE 3**



484

485 **Extended Data Figure 3 | The HNH nuclease, REC2 and REC3 domains undergo**

486 **substantial conformational changes upon binding to the dsDNA target. a,** Schematic of

487 SpCas9 domain structure with color coding for separate domains. **b,** Vector map of global

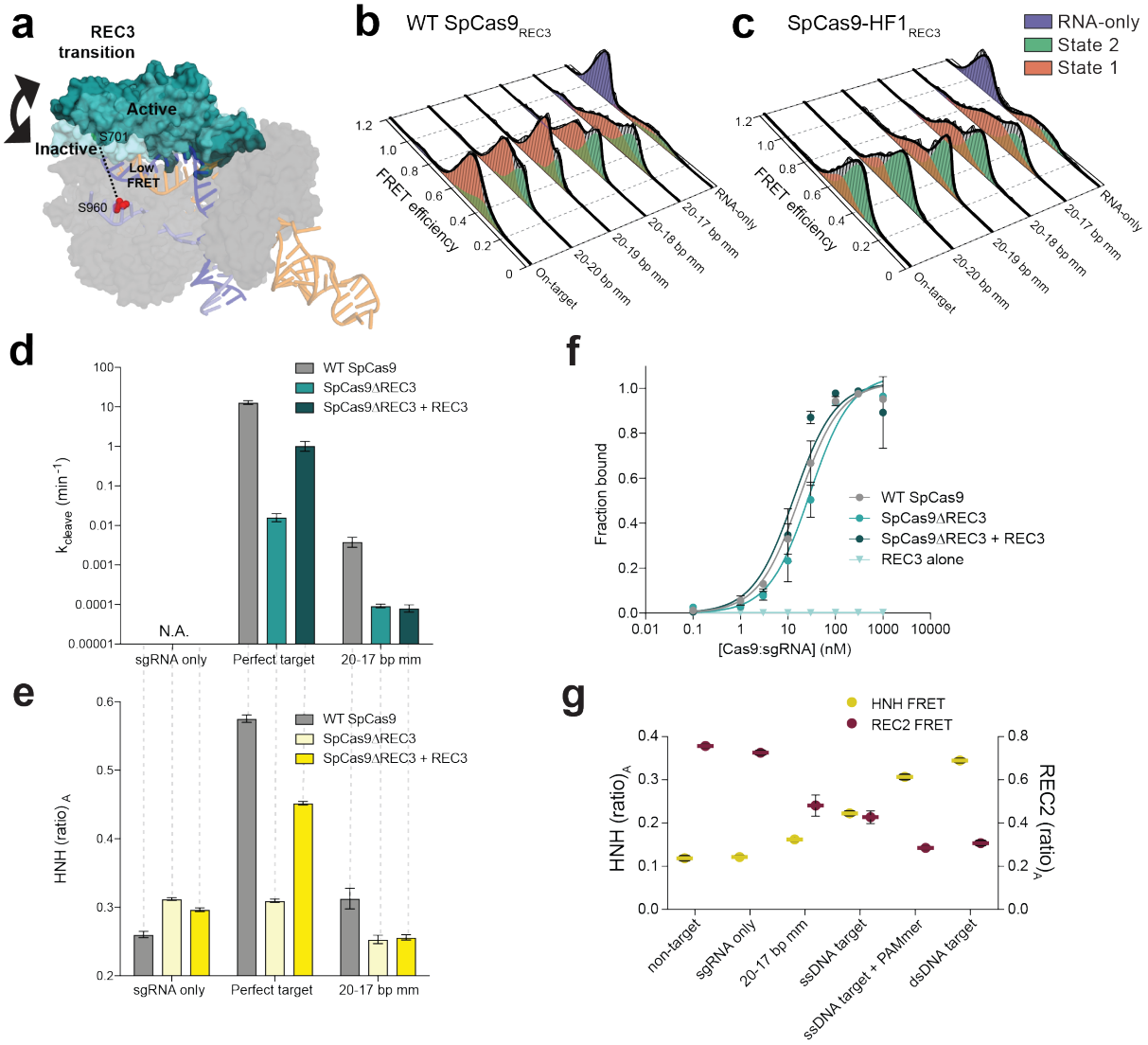
488 SpCas9 conformational changes from the sgRNA- (PDB ID: 4ZT0) to dsDNA-bound structures

489 (PDB ID: 5F9R), domains colored as in panel **a**.

490

491

492 **EXTENDED DATA FIGURE 4**



493

494 **Extended Data Figure 4 | Nucleic acid sensing requires engagement with the REC3 domain**

495 **and outward rotation of the REC2 domain. a,** Schematic of $SpCas9_{REC3}$ with FRET dyes at

496 positions S701C and S960C, with HNH domain omitted for clarity. Inactive to active structures

497 represent REC3 in the sgRNA-bound (PDB ID: 4ZT0) to dsDNA-bound (PDB ID: 5F9R) forms,

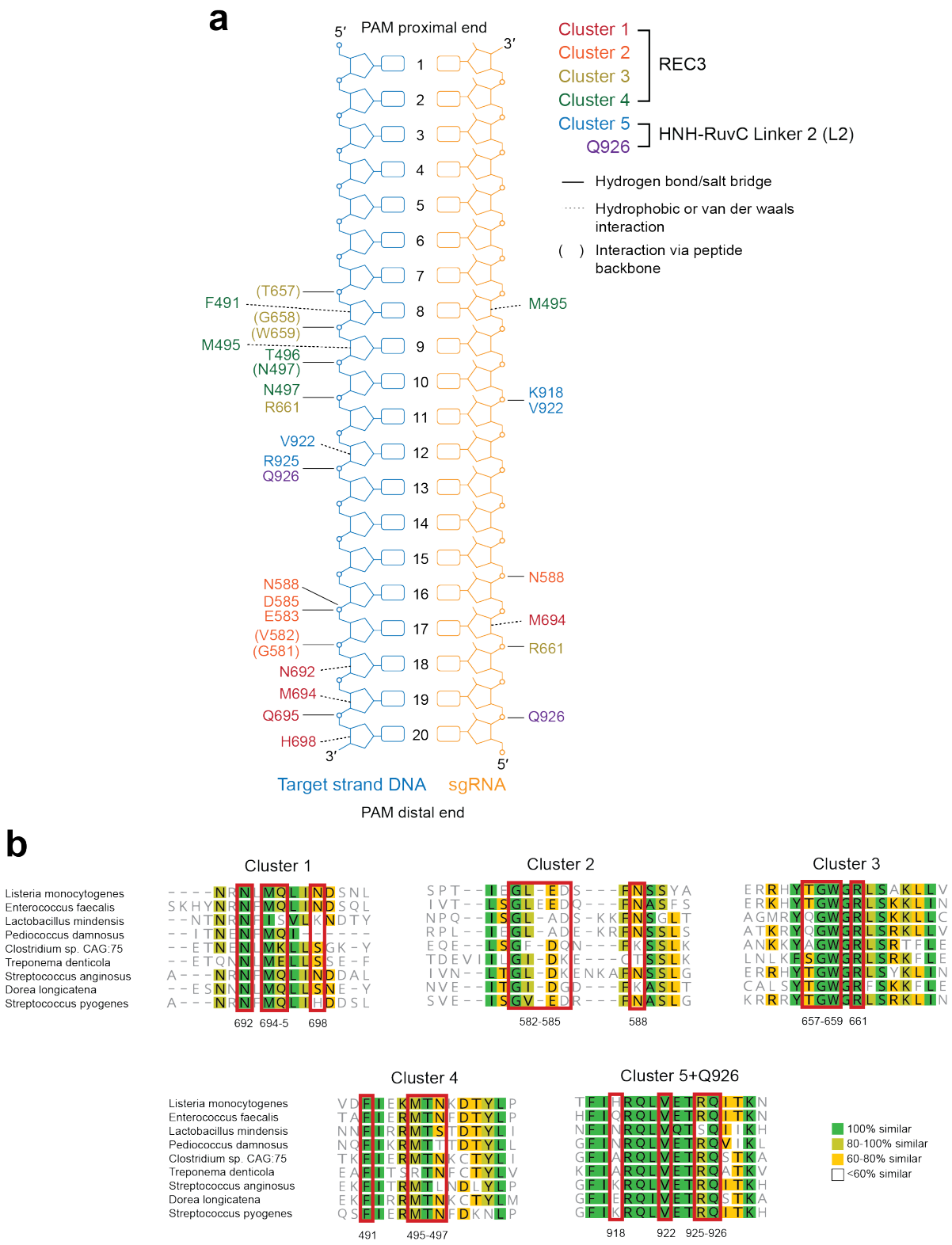
498 respectively. **b–c,** smFRET histograms measuring HNH conformational activation with black

499 curves representing a fit to multiple Gaussian peaks for **b,** WT $SpCas9_{REC3}$ and **c,** SpCas9-

500 $HF1_{REC3}$ bound to perfect and PAM-distal mismatched targets. The purple peak denotes the

501 sgRNA-only bound state, while the red and green peaks represent two states of REC3 with
502 conformational flexibility upon binding to DNA substrates. **d–e,** REC3 *in vitro* complementation
503 assay with SpCas9ΔREC3 by measuring **d**, cleavage rate constants and **e,** HNH activation with
504 $(ratio)_A$ values. **f,** Perfect target DNA binding assay in the presence or absence of the REC3
505 domain. **g,** $(Ratio)_A$ data with $SpCas9_{REC2}$ and $SpCas9_{HNH}$ showing reciprocal FRET states with
506 the indicated substrates. Error bars in **d–g**, s.d.; $n = 3$.
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523

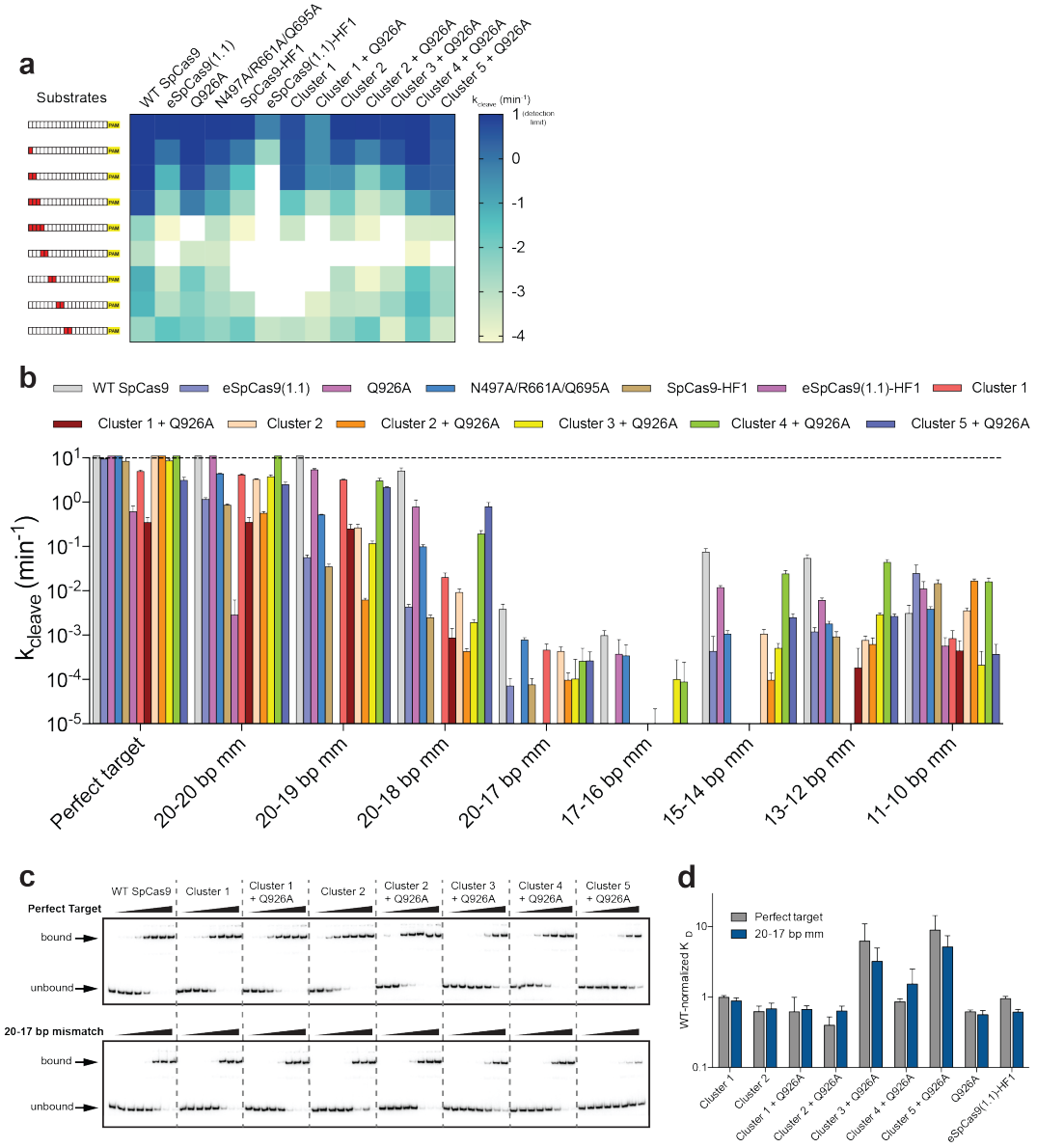
524 **EXTENDED DATA FIGURE 5**



525

526 **Extended Data Figure 5 | Identification of Cluster variants based on nucleic acid proximity**
527 **and multiple sequence alignment of residues within Clusters 1-5. a,** Schematic depicting
528 interactions of WT SpCas9 residues within Clusters 1-5 with the RNA/DNA heteroduplex, based
529 on PDB accession 5F9R (adapted from ref 8). **b,** Alignment of selected Cas9 orthologues using
530 MAFFT and visualized in Geneious 10.0, with red boxes outlining residues mutated to alanine
531 within each cluster variant.
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548

549 **EXTENDED DATA FIGURE 6**



550

551 **Extended Data Figure 6 | Mutation clusters in the REC3 domain along the RNA/DNA**

552 **heteroduplex demonstrate localized sensitivity to mismatches along the target sequence. a-**

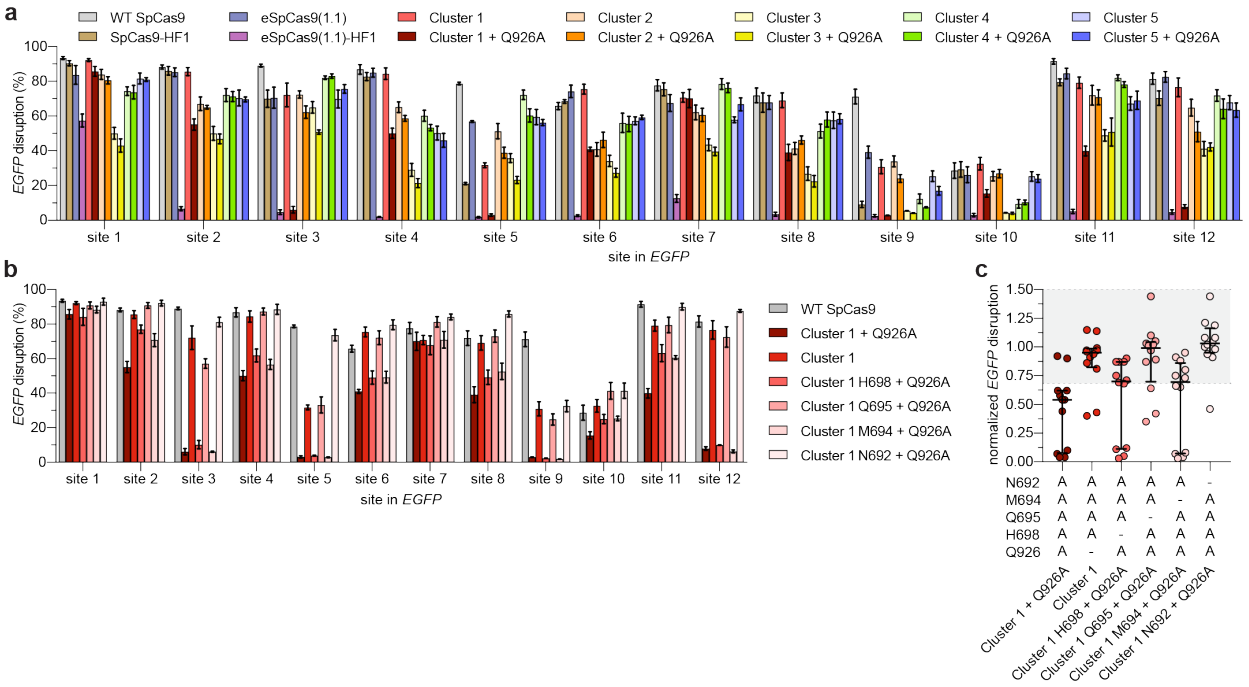
553 **b,** Quantified DNA cleavage rates (detection limit for k_{cleave} set at 10 min^{-1}) displayed as a **a,**

554 heatmap and **b,** bar graph. **c-d,** Target DNA binding assay **c,** resolved by native polyacrylamide

555 gel electrophoresis (PAGE) mobility shift assays and **d,** quantification with WT-normalized

556 dissociation constants. Error bars in **b** and **d**, s.d.; $n = 3$.

557 **EXTENDED DATA FIGURE 7**



558

559 **Extended Data Figure 7 | On-target activities of altered specificity variants using a human**

560 **cell *EGFP* disruption assay. a,** Summary of *EGFP* disruption activities for SpCas9-HF1,

561 eSpCas9(1.1), eSpCas9(1.1)-HF1 and Cluster variants ± Q926A with mean and s.e.m., where n ≥

562 3. **b,** Summary of *EGFP* disruption activities for the series of Cluster 1 variants with each

563 substituted residue restored to the canonical amino acid; mean and s.e.m. where n ≥ 3; WT,

564 Cluster 1, and Cluster 1 A926Q data from **panel a** is re-plotted for comparison. **c,** WT-

565 normalized plot of data in **panel b**; error bars represent median and interquartile range; the

566 interval with > 70% of wild-type activity is highlighted in light grey.

567

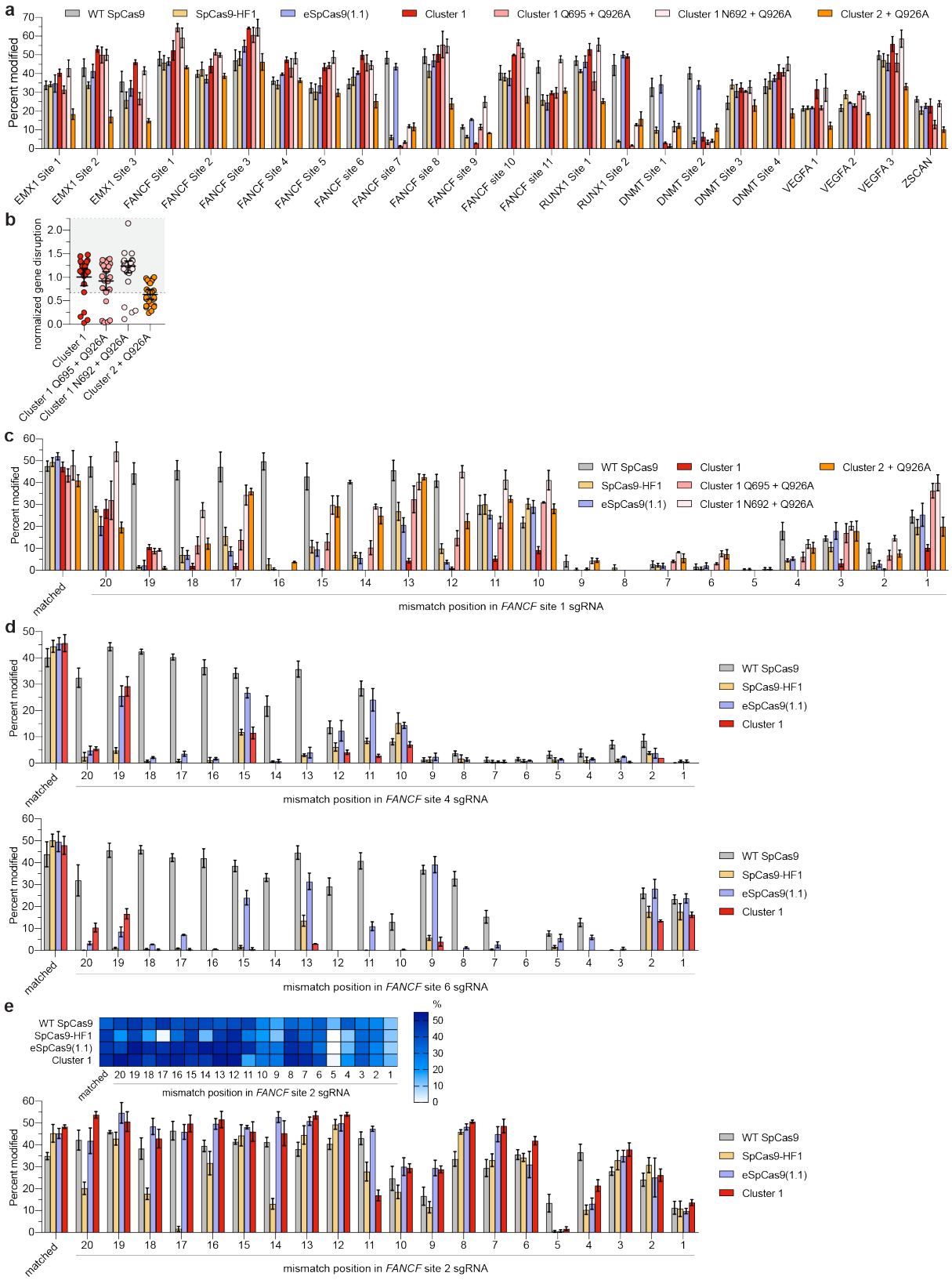
568

569

570

571

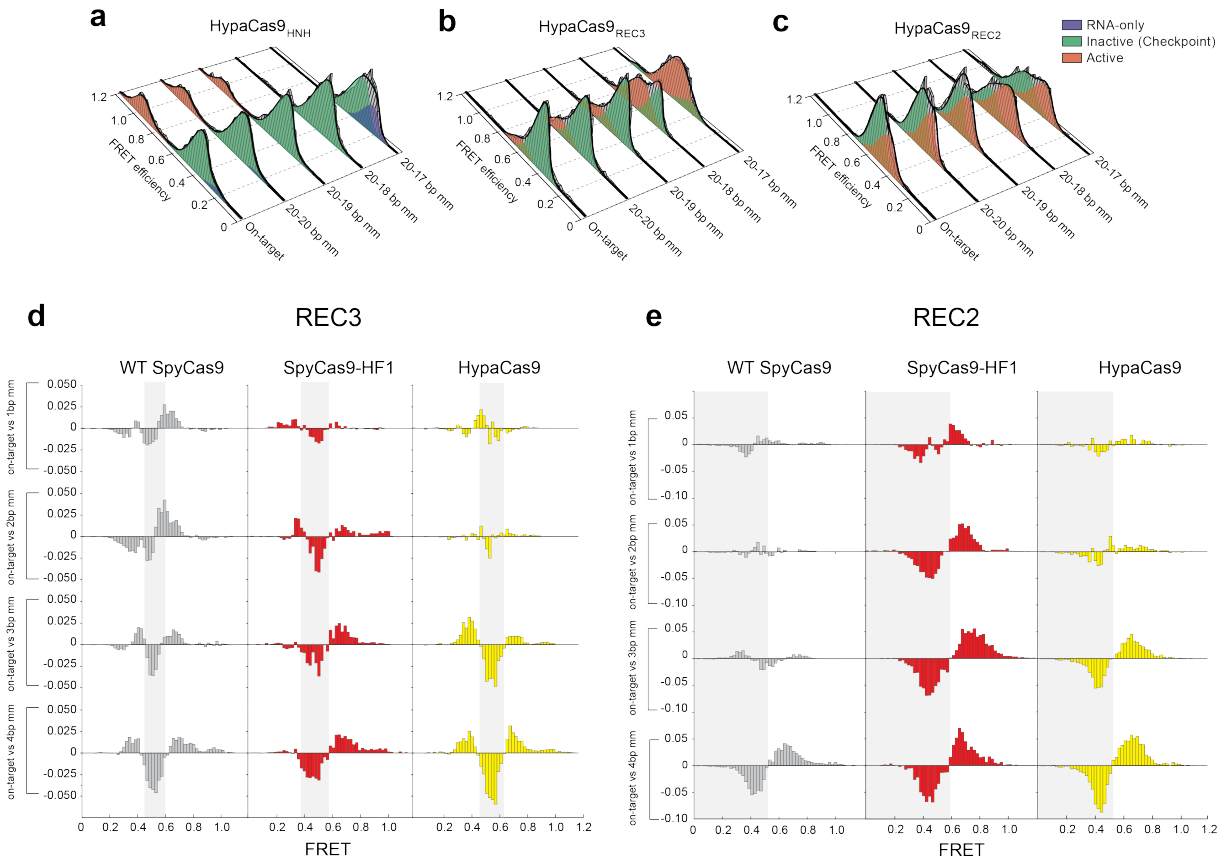
572 **EXTENDED DATA FIGURE 8**



573

574 **Extended Data Figure 8 | Activities and specificities of high-fidelity SpCas9 variants**
575 **targeted to endogenous human cell sites. a,** On-target activities of WT SpCas9, SpCas9-HF1,
576 Cluster 1 and Cluster 2 variants across 24 endogenous human genes, assessed by T7E1 assay.
577 Mean and s.e.m. shown; n ≥ 3. **b,** WT-normalized endogenous gene disruption data from panel **a**,
578 for Cluster 1 and 2 variants. Error bars represent median and interquartile ranges with the > 70%
579 interval of wild-type activity highlighted in light grey; Cluster 1 A926Q data from **Fig. 3b** is
580 replotted for comparison. **c-e,** Summary of single mismatch tolerance of WT SpCas9, SpCas9-
581 HF1, eSpCas9(1.1), and Cluster 1 and Cluster 2 variants on **c,** *FANCF* site 1 **d,** *FANCF* sites 4
582 and 6, and **e,** *FANCF* site 2. Percent modification assessed by T7E1 assay; mean and s.e.m.
583 shown; n ≥ 3.
584
585
586
587
588
589
590
591
592
593
594
595
596

597 **EXTENDED DATA FIGURE 9**



598

599 **Extended Data Figure 9 | Conformational gating drives targeting accuracy for SpCas9**

600 **variants. a-c, Steady state** smFRET histograms measuring **a,** HNH, **b,** REC2 and **c,** REC3

601 conformational states for HypaCas9 bound to on-target and PAM-distal mismatched substrates.

602 Black curves represent a fit to multiple Gaussian peaks. **d-e,** Steady state smFRET histograms of

603 Cas9 variants bound to PAM distal mismatched substrates were normalized to and subtracted

604 from that of on-target smFRET histograms. This analysis reveals transitions from one FRET

605 population (negative peak) to another population (positive peak) for **d,** REC3 and **e,** REC2.

606

607

608

609 **EXTENDED DATA TABLE 1**

		Inter-residue distance (Å)		
		HNH	REC2	REC3
Structure	**PDB**	**(S355-S867)**	**(E60-D273)**	**(S701-S960)**
sgRNA-only	4ZT0	81	12.8	28.4
sgRNA/TS DNA-bound	4OO8	61	35.9	45.1
sgRNA/TS DNA + PAMmer-bound	4UN3	59	37.2	46.5
sgRNA/dsDNA-bound	5F9R	30	32	40.3

610

611 **Extended Data Table 1 | Measured distances between residues labelled with Cy3/Cy5**

612 **FRET dyes for different substrate-bound Cas9 structures.** Residue pairs were designed to

613 report conformational changes of the specified domain (HNH, REC2 or REC3). The distances

614 were measured between Cα atoms of the indicated residues for the associated PDB structures.

615

616

617

618

619

620

621

622

623

624

625

626

627

628 **SUPPLEMENTARY TABLE 1**

	Substitution/Truncation						
Protein description	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Wild-type SpCas9	-	-	-	-	-	-	-
$SpCas9_{HNH}$	C80S	C574S	S355C	S867C	-	-	-
$SpCas9_{REC2}$	C80S	C574S	E60C	D273C	-	-	-
$SpCas9_{REC3}$	C80S	C574S	S701C	S960C	-	-	-
SpCas9ΔREC3	M1–N497,GGS,V713–D1368				-	-	-
$SpCas9ΔREC3_{HNH}$	M1–N497,GGS,V713–D1368				+ HNH FRET		-
$SpCas9ΔREC3_{REC2}$	M1–N497,GGS,V713–D1368				+ REC2 FRET		-
REC3 domain	K506–Q712		-	-	-	-	-
SpCas9-Q926A	Q926A	-	-	-	-	-	-
SpCas9-N497A/R661A/Q695A	N497A	R661A	Q695A	-	-	-	-
$SpCas9\text{-}N497A/R661A/Q695A_{HNH}$	N497A	R661A	Q695A	+ HNH FRET		-	-
SpCas9-HF1	N497A	R661A	Q695A	Q926A	-	-	-
$SpCas9\text{-}HF1_{HNH}$	N497A	R661A	Q695A	Q926A	+ HNH FRET		-
$SpCas9\text{-}HF1_{REC2}$	N497A	R661A	Q695A	Q926A	+ REC2 FRET		-
$SpCas9\text{-}HF1_{REC3}$	N497A	R661A	Q695A	Q926A	+ REC3 FRET		-
SpCas9-K855A	K855A	-	-	-	-	-	-
$SpCas9\text{-}K855A_{HNH}$	K855A	+ HNH FRET		-	-	-	-
eSpCas9(1.1)	K848A	K1003A	R1060A	-	-	-	-
$eSpCas9(1.1)_{HNH}$	K848A	K1003A	R1060A	+ HNH FRET		-	-
eSpCas9(1.1)-HF1	N497A	R661A	Q695A	Q926A	K848A	K1003A	R1060A
Cluster 1 (HypaCas9)	N692A	M694A	Q695A	H698A	-	-	-
$Cluster\ 1_{HNH}$	N692A	M694A	Q695A	H698A	Q926A	+ HNH FRET	
$Cluster\ 1_{REC2}$	N692A	M694A	Q695A	H698A	Q926A	+ REC2 FRET	
$Cluster\ 1_{REC3}$	N692A	M694A	Q695A	H698A	Q926A	+ REC3 FRET	
Cluster 1 + Q926A	N692A	M694A	Q695A	H698A	Q926A	-	-
Cluster 1 H698 + Q926A	N692A	M694A	Q695A	-	Q926A	-	-
Cluster 1 Q695 + Q926A	N692A	M694A	-	H698A	Q926A	-	-
Cluster 1 M694 + Q926A	N692A	-	Q695A	H698A	Q926A	-	-
Cluster 1 N692 + Q926A	-	M694A	Q695A	H698A	Q926A	-	-
Cluster 2	G582A	V583A	E584A	D585A	N588A	-	-
Cluster 2 + Q926A	G582A	V583A	E584A	D585A	N588A	Q926A	-
Cluster 3	T657A	G658A	W659A	R661A	-	-	-
Cluster 3 + Q926A	T657A	G658A	W659A	R661A	Q926A	-	-
Cluster 4	F491A	M495A	T496A	N497A	-	-	-
Cluster 4 + Q926A	F491A	M495A	T496A	N497A	Q926A	-	-
Cluster 5	K918A	V922A	R925A	-	-	-	-
Cluster 5 + Q926A	K918A	V922A	R925A	Q926A	-	-	-

629

630 **Supplementary Table 1 |** All enhanced specificity, high-fidelity, cluster and hyper-accurate

631 SpCas9 variants tested in this study. The HNH, REC3 or REC3 subscript designation with an

632 enhanced specificity, high-fidelity or cluster SpCas9 variant denotes combination of residue

633 substitutions with indicated FRET construct.

634

SUPPLEMENTARY TABLE 2

***in vitro* DNA substrates**	
substrate description	**target sequence with PAM (5'-->3')**
l1-targeting sgRNA	GACGCATAAAGATGAGACGCGTTTTAGAGCTATGCTGTTTTGGAAACAAAACAGCATAGCAAGTTAAAATAAG GCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGC
l1 TS on-target	AGCTGACGTTTGTACTCCAGCGTCTCATCTTTATGCGTCAGCAGAGATTTCTGCT
l1 NTS on-target	AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 NTS (5'biotin)	/5Biosg/AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 TS 20-20 bp mm	AGCTGACGTTTGTACTCCAGCGTCTCATCTTTATGCGTgAGCAGAGATTTCTGCT
l1 NTS 20-20 bp mm	AGCAGAAATCTCTGCTcACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 NTS (5'biotin) 20-20 bp mm	/5Biosg/AGCAGAAATCTCTGCTcACGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 TS 20-19 bp mm	AGCTGACGTTTGTACTCCAGCGTCTCATCTTTATGCGagAGCAGAGATTTCTGCT
l1 NTS 20-19 bp mm	AGCAGAAATCTCTGCTctCGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 NTS (5'biotin) 20-19 bp mm	/5Biosg/AGCAGAAATCTCTGCTctCGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 TS 20-18 bp mm	AGCTGACGTTTGTACTCCAGCGTCTCATCTTTATGCcagAGCAGAGATTTCTGCT
l1 NTS 20-18 bp mm	AGCAGAAATCTCTGCTctgGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 NTS (5'biotin) 20-18 bp mm	/5Biosg/AGCAGAAATCTCTGCTctgGCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 TS 20-17 bp mm	AGCTGACGTTTGTACTCCAGCGTCTCATCTTTATGgcagAGCAGAGATTTCTGCT
l1 NTS 20-17 bp mm	AGCAGAAATCTCTGCTctgcCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
l1 NTS (5'biotin) 20-17 -bp mm	/5Biosg/AGCAGAAATCTCTGCTctgcCATAAAGATGAGACGCTGGAGTACAAACGTCAGCT
FANCF site 1-targeting sgRNA	GGAATCCCTTCTGCAGCACCGTTTTAGAGCTATGCTGTTTTGGAAACAAAACAGCATAGCAAGTTAAAATAAG GCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGC
FANCF site 1 TS on-target	AGCTCGGAAAAGCGATCCAGGTGCTGCAGAAGGGATTCCATGAGGTGCGCGAAGG
FANCF site 1 NTS on-target	CCTTCGCGCACCTCATGGAATCCCTTCTGCAGCACCTGGATCGCTTTTCCGAGCT
FANCF site 1 TS 10 bp mm	AGCTCGGAAAAGCGATCCAGGTGCTGCAcAAGGGATTCCATGAGGTGCGCGAAGG
FANCF site 1 NTS 10 bp mm	CCTTCGCGCACCTCATGGAATCCCTTgTGCAGCACCTGGATCGCTTTTCCGAGCT
FANCF site 1 TS 11 bp mm	AGCTCGGAAAAGCGATCCAGGTGCTGCAGtAGGGATTCCATGAGGTGCGCGAAGG
FANCF site 1 NTS 11 bp mm	CCTTCGCGCACCTCATGGAATCCCTaCTGCAGCACCTGGATCGCTTTTCCGAGCT
FANCF site 1 TS 12 bp mm	AGCTCGGAAAAGCGATCCAGGTGCTGCAGAtGGGATTCCATGAGGTGCGCGAAGG
FANCF site 1 NTS 12 bp mm	CCTTCGCGCACCTCATGGAATCCCaTCTGCAGCACCTGGATCGCTTTTCCGAGCT
FANCF site 1 NTS (5'biotin) 12 bp mm	/5Biosg/CCTTCGCGCACCTCATGGAATCCCaTCTGCAGCACCTGGATCGCTTTTCCGAGCT

sgRNAs against *EGFP*	
plasmid description	**target sequence with PAM**
EGFP NGG site 1	GGGCACGGGCAGCTTGCCGGTGG
EGFP NGG site 2	GTCGCCCTCGAACTTCACCTCGG
EGFP NGG site 3	GTAGGTCAGGGTGGTCACGAGGG
EGFP NGG site 4	GGCGAGGGCGATGCCACCTACGG
EGFP NGG site 5	GGTCGCCACCATGGTGAGCAAGG
EGFP NGG site 6	GGTCAGGGTGGTCACGAGGGTGG
EGFP NGG site 7	GGTGGTGCAGATGAACTTCAGGG
EGFP NGG site 8	GTTGGGGTCTTTGCTCAGGGCGG
EGFP NGG site 9	GGTGGTCACGAGGGTGGGCCAGG
EGFP NGG site 10	GATGCCGTTCTTCTGCTTGTCGG
EGFP NGG site 11	GTCGCCACCATGGTGAGCAAGGG
EGFP NGG site 12	GCACTGCACGCCGTAGGTCAGGG
EGFP NGG site 1, mismatches at 1 & 2	GGGCACGGGCAGCTTGCCccTGG
EGFP NGG site 1, mismatches at 3 & 4	GGGCACGGGCAGCTTGggGGTGG
EGFP NGG site 1, mismatches at 5 & 6	GGGCACGGGCAGCTacCCGGTGG
EGFP NGG site 1, mismatches at 7 & 8	GGGCACGGGCAGgaTGCCGGTGG
EGFP NGG site 1, mismatches at 9 & 10	GGGCACGGGCtcCTTGCCGGTGG
EGFP NGG site 1, mismatches at 11 & 12	GGGCACGGcgAGCTTGCCGGTGG
EGFP NGG site 1, mismatches at 13 & 14	GGGCACccGCAGCTTGCCGGTGG
EGFP NGG site 1, mismatches at 15 & 16	GGGCtgGGGCAGCTTGCCGGTGG
EGFP NGG site 1, mismatches at 17 & 18	GGcgACGGGCAGCTTGCCGGTGG
EGFP NGG site 1, mismatches at 18 & 19	GccCACGGGCAGCTTGCCGGTGG
EGFP NGG site 7, mismatches at 1 & 2	GGTGGTGCAGATGAACTTgtGGG
EGFP NGG site 7, mismatches at 3 & 4	GGTGGTGCAGATGAACaaCAGGG
EGFP NGG site 7, mismatches at 5 & 6	GGTGGTGCAGATGAtgTTCAGGG
EGFP NGG site 7, mismatches at 7 & 8	GGTGGTGCAGATctACTTCAGGG
EGFP NGG site 7, mismatches at 9 & 10	GGTGGTGCAGtaGAACTTCAGGG
EGFP NGG site 7, mismatches at 11 & 12	GGTGGTGCtcATGAACTTCAGGG
EGFP NGG site 7, mismatches at 13 & 14	GGTGGTcgAGATGAACTTCAGGG
EGFP NGG site 7, mismatches at 15 & 16	GGTGcaGCAGATGAACTTCAGGG
EGFP NGG site 7, mismatches at 17 & 18	GGacGTGCAGATGAACTTCAGGG
EGFP NGG site 7, mismatches at 18 & 19	GcaGGTGCAGATGAACTTCAGGG

sgRNAs against endogenous sites	
plasmid description	target sequence with PAM
DNMT1 NGG site 1	GTCACTCTGGGGAACACGCCCGG
DNMT1 NGG site 2	GAGTGCTAAGGGAACGTTCACGG
DNMT1 NGG site 3	GAGACTGAACACTCCTCAAACGG
DNMT1 NGG site 4	GGAGTGAGGGAAACGGCCCCAGG
EMX1 NGG site 1	GAGTCCGAGCAGAAGAAGAAGGG
EMX1 NGG site 2	GTCACCTCCAATGACTAGGGTGG
EMX1 NGG site 3	GGGAAGACTGAGGCTACATAGGG
FANCF NGG site 1	GGAATCCCTTCTGCAGCACCTGG
FANCF NGG site 2	GCTGCAGAAGGGATTCCATGAGG
FANCF NGG site 3	GGCGGCTGCACAACCAGTGGAGG
FANCF NGG site 4	GCTCCAGAGCCGTGCGAATGGGG
FANCF NGG site 5	GAAGCTCGGAAAAGCGATCCAGG
FANCF NGG site 6	GCTTGAGACCGCCAGAAGCTCGG
FANCF NGG site 7	GACCAAAGCGCCGATGGATGTGG
FANCF NGG site 8	GGGGTCCCAGGTGCTGACGTAGG
FANCF NGG site 9	GGATCGCTTTTCCGAGCTTCTGG
FANCF NGG site 10	GGATTCCATGAGGTGCGCGAAGG
FANCF NGG site 11	GCGACTCTCTGCGTACTGATTGG
RUNX1 NGG site 1	GCATTTTCAGGAGGAAGCGATGG
RUNX1 NGG site 2	GGGAGAAGAAAGAGAGATGTAGG
VEGFA NGG site 1	GGGTGGGGGGAGTTTGCTCCTGG
VEGFA NGG site 2	GACCCCCTCCACCCCGCCTCCGG
VEGFA NGG site 3	GGTGAGTGAGTGTGTGCGTGTGG
ZSCAN2 NGG site	GTGCGGCAAGAGCTTCAGCCGGG
FANCF NGG site 1, mismatch at 1	GGAATCCCTTCTGCAGCACgTGG
FANCF NGG site 1, mismatch at 2	GGAATCCCTTCTGCAGCAgCTGG
FANCF NGG site 1, mismatch at 3	GGAATCCCTTCTGCAGCtCCTGG
FANCF NGG site 1, mismatch at 4	GGAATCCCTTCTGCAGgACCTGG
FANCF NGG site 1, mismatch at 5	GGAATCCCTTCTGCAcCACCTGG
FANCF NGG site 1, mismatch at 6	GGAATCCCTTCTGCtGCACCTGG
FANCF NGG site 1, mismatch at 7	GGAATCCCTTCTGgAGCACCTGG
FANCF NGG site 1, mismatch at 8	GGAATCCCTTCTcCAGCACCTGG
FANCF NGG site 1, mismatch at 9	GGAATCCCTTCaGCAGCACCTGG

638

FANCF NGG site 1, mismatch at 10	GGAATCCCTTgTGCAGCACCTGG
FANCF NGG site 1, mismatch at 11	GGAATCCCTaCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 12	GGAATCCCaTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 13	GGAATCCgTTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 14	GGAATCgCTTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 15	GGAATgCCTTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 16	GGAAaCCCTTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 17	GGAtTCCCTTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 18	GGtATCCCTTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 19	GcAATCCCTTCTGCAGCACCTGG
FANCF NGG site 1, mismatch at 20	cGAATCCCTTCTGCAGCACCTGG
FANCF NGG site 2, mismatch at 1	GCTGCAGAAGGGATTCCATcAGG
FANCF NGG site 2, mismatch at 2	GCTGCAGAAGGGATTCCAaGAGG
FANCF NGG site 2, mismatch at 3	GCTGCAGAAGGGATTCCtTGAGG
FANCF NGG site 2, mismatch at 4	GCTGCAGAAGGGATTCgATGAGG
FANCF NGG site 2, mismatch at 5	GCTGCAGAAGGGATTgCATGAGG
FANCF NGG site 2, mismatch at 6	GCTGCAGAAGGGATaCCATGAGG
FANCF NGG site 2, mismatch at 7	GCTGCAGAAGGGAaTCCATGAGG
FANCF NGG site 2, mismatch at 8	GCTGCAGAAGGGtTTCCATGAGG
FANCF NGG site 2, mismatch at 9	GCTGCAGAAGGcATTCCATGAGG
FANCF NGG site 2, mismatch at 10	GCTGCAGAAGcGATTCCATGAGG
FANCF NGG site 2, mismatch at 11	GCTGCAGAAcGGATTCCATGAGG
FANCF NGG site 2, mismatch at 12	GCTGCAGAtGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 13	GCTGCAGtAGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 14	GCTGCAcAAGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 15	GCTGCtGAAGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 16	GCTGgAGAAGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 17	GCTcCAGAAGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 18	GCaGCAGAAGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 19	GgTGCAGAAGGGATTCCATGAGG
FANCF NGG site 2, mismatch at 20	cCTGCAGAAGGGATTCCATGAGG
FANCF NGG site 4, mismatch at 1	GCTCCAGAGCCGTGCGAATcGGG
FANCF NGG site 4, mismatch at 2	GCTCCAGAGCCGTGCGAAaGGGG
FANCF NGG site 4, mismatch at 3	GCTCCAGAGCCGTGCGAtTGGGG
FANCF NGG site 4, mismatch at 4	GCTCCAGAGCCGTGCGtATGGGG

639

FANCF NGG site 4, mismatch at 5	GCTCCAGAGCCGTGCcAATGGGG
FANCF NGG site 4, mismatch at 6	GCTCCAGAGCCGTGgGAATGGGG
FANCF NGG site 4, mismatch at 7	GCTCCAGAGCCGTcCGAATGGGG
FANCF NGG site 4, mismatch at 8	GCTCCAGAGCCGaGCGAATGGGG
FANCF NGG site 4, mismatch at 9	GCTCCAGAGCCcTGCGAATGGGG
FANCF NGG site 4, mismatch at 10	GCTCCAGAGCgGTGCGAATGGGG
FANCF NGG site 4, mismatch at 11	GCTCCAGAGgCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 12	GCTCCAGAcCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 13	GCTCCAGtGCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 14	GCTCCAcAGCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 15	GCTCCtGAGCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 16	GCTCgAGAGCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 17	GCTgCAGAGCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 18	GCaCCAGAGCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 19	GgTCCAGAGCCGTGCGAATGGGG
FANCF NGG site 4, mismatch at 20	cCTCCAGAGCCGTGCGAATGGGG
FANCF NGG site 6, mismatch at 1	GCTTGAGACCGCCAGAAGCaCGG
FANCF NGG site 6, mismatch at 2	GCTTGAGACCGCCAGAAGgTCGG
FANCF NGG site 6, mismatch at 3	GCTTGAGACCGCCAGAAcCTCGG
FANCF NGG site 6, mismatch at 4	GCTTGAGACCGCCAGAtGCTCGG
FANCF NGG site 6, mismatch at 5	GCTTGAGACCGCCAGtAGCTCGG
FANCF NGG site 6, mismatch at 6	GCTTGAGACCGCCAcAAGCTCGG
FANCF NGG site 6, mismatch at 7	GCTTGAGACCGCCtGAAGCTCGG
FANCF NGG site 6, mismatch at 8	GCTTGAGACCGCgAGAAGCTCGG
FANCF NGG site 6, mismatch at 9	GCTTGAGACCGgCAGAAGCTCGG
FANCF NGG site 6, mismatch at 10	GCTTGAGACCcCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 11	GCTTGAGACgGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 12	GCTTGAGAgCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 13	GCTTGAGtCCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 14	GCTTGAcACCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 15	GCTTGtGACCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 16	GCTTcAGACCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 17	GCTaGAGACCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 18	GCaTGAGACCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 19	GgTTGAGACCGCCAGAAGCTCGG
FANCF NGG site 6, mismatch at 20	cCTTGAGACCGCCAGAAGCTCGG

sequence	description
CCAGAATGCACAAAGTACTGCAC	forward primer to amplify DNMT1 target sites in U2OS human cells
GCCAAAGCCCGAGAGAGTGCC	reverse primer to amplify DNMT1 target sites in U2OS human cells
GGAGCAGCTGGTCAGAGGGG	forward primer to amplify EMX1 target sites in U2OS human cells
CCATAGGGAAGGGGGACACTGG	reverse primer to amplify EMX1 target sites in U2OS human cells
GGGCCGGGAAAGAGTTGCTG	forward primer to amplify FANCF target sites in U2OS human cells (set #1)
GCCCTACATCTGCTCTCCCTCC	reverse primer to amplify FANCF target sites in U2OS human cells (set #1)
CAGCATGTGCACCGCAGACC	forward primer to amplify FANCF target sites in U2OS human cells (set #2)
TCATCTCGCACGTGGTTCCGG	reverse primer to amplify FANCF target sites in U2OS human cells (set #2)
CCAGCACAACTTACTCGCACTTGAC	forward primer to amplify RUNX1 target sites in U2OS human cells
CATCACCAACCCACAGCCAAGG	reverse primer to amplify RUNX1 target sites in U2OS human cells
ATCCCTGGACACTTCCCAAAGGAC	forward primer to amplify VEGFA target sites 1 and 3 in U2OS human cells
CTCGACCCCCACCAAGGTTCAC	reverse primer to amplify VEGFA target sites 1 and 3 in U2OS human cells
CGAGGAAGAGAGAGACGGGGTC	forward primer to amplify VEGFA target site 2 in U2OS human cells
CTCCAATGCACCCAAGACAGCAG	reverse primer to amplify VEGFA target site 2 in U2OS human cells
AGTGTGGGGTGTGTGGGAAG	forward primer to amplify ZSCAN2 target sites in U2OS human cells
ACGGGACTTGACTCAGACCACT	reverse primer to amplify ZSCAN2 target sites in U2OS human cells

Supplementary Table 2 | Sequences of all nucleic acids used in the study.