Determining the Specificity of Cascade Binding, Interference, and Priming *in vivo*

Lauren A. Cooper[1], Anne M. Stringer [2] and Joseph T. Wade[1,2,3]

[1]Department of Biomedical Sciences, School of Public Health, University at Albany, Albany, New York, USA.

[2]Wadsworth Center, New York State Department of Health, Albany, New York, USA

[3]Corresponding author:

joseph.wade@health.ny.gov

Tel: (518) 474 5727

Fax: (518) 474 3181

SIGNIFICANCE

Many bacterial and archaeal species encode CRISPR-Cas immunity systems that protect against invasion by foreign DNA. In the *Escherichia coli* CRISPR-Cas system, a protein complex, Cascade, binds 61 nt CRISPR RNAs (crRNAs). The Cascade-crRNA complex is directed to invading DNA molecules through base-pairing between the crRNA and target DNA. This leads to recruitment of the Cas3 nuclease that destroys the invading DNA molecule, and promotes acquisition of new immunity elements. We show that Cascade-crRNA binding to DNA is highly promiscuous *in vivo*. Consequently, endogenous *E. coli* crRNAs direct Cascade binding to >100 chromosomal locations. In contrast, target degradation and acquisition of new immunity elements requires highly specific association of Cascade-crRNA with DNA, limiting CRISPR-Cas function to the intended targets.

ABSTRACT

In CRISPR immunity systems, short CRISPR RNAs (crRNAs) are bound by CRISPR-associated (Cas) proteins, and these complexes target invading nucleic acid molecules for degradation in a process known as interference. In Type I CRISPR systems, the Cas protein complex that binds DNA is known as Cascade. Association of Cascade with target DNA can also lead to acquisition of new immunity elements, in a process known as priming. The sequence determinants for protospacer binding and interference have been well characterized for Type II CRISPR systems such as the Cas9 system of *Streptococcus pyogenes*. In contrast, relatively little is known about the requirements for Cascade-DNA binding, interference, and priming in Type I systems. Here, we use genome-scale approaches to assess the specificity determinants for Cascade-DNA interaction, interference, and priming *in vivo* for the Type I-E system of *Escherichia coli*. Remarkably, as few as 5 bp of crRNA-DNA are sufficient for association of Cascade with a DNA target. Consequently, a single crRNA promotes Cascade association with numerous off-target sites, and the endogenous *E. coli* crRNAs direct Cascade binding to >100 chromosomal sites. In contrast to the low specificity of Cascade-DNA interactions, >18 bp are required for both interference and priming. Hence, Cascade binding to sub-optimal, off-target sites is inert. Our data support a model in which initial Cascade association with DNA targets requires little sequence complementarity at the crRNA 5□ end, whereas recruitment and/or activation of the Cas3 nuclease, a prerequisite for interference and priming, requires extensive base-pairing.

INTRODUCTION

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas (CRISPR-associated) systems are adaptive immune systems found in approximately 40% of bacteria and 90% of archaea (1). CRISPR-Cas systems are characterized by the presence of CRISPR arrays and Cas proteins. CRISPR arrays are genomic loci that consist of short repetitive sequences ("repeats"), interspaced with short sequences of viral or plasmid origin ("spacers") (2–7). Spacers are acquired during a process known as "adaptation", in which a complex of Cas1 and Cas2 integrates invading DNA into a CRISPR array, effectively immunizing the organism from future assault by the invader (8). In the archetypal Type I-E CRISPR system of *Escherichia coli*, immunity occurs by a process known as "interference". During interference, a CRISPR array is transcribed, and Cas6e processes the transcript into individual, 61 nt CRISPR RNAs (crRNAs) that each include a single, 32 nt spacer sequence flanked by partial repeat sequences (9, 10). Individual crRNAs are then incorporated into Cascade, a protein complex composed of five different Cas proteins (Cse1-$Cse2_2$-$Cas7_6$-Cas5-Cas6e) (9, 11). Cascade-crRNA complexes bind to target DNA sequences known as "protospacers" that are complementary to the crRNA spacer, and are immediately adjacent to a short DNA sequence known as a "Protospacer-Associated Motif" (PAM). The crRNA bound by Cascade forms an R-loop with one strand of the target DNA, which in turn leads to recruitment of the Cas3 nuclease, DNA cleavage, and elimination of the invader (12–21).

For Type I CRISPR systems, adaptation can occur by two mechanisms: "naïve" and "primed". Naïve adaptation requires only two Cas proteins: Cas1 and Cas2 (8). Primed adaptation, by

78 contrast, requires all Cas proteins and an existing crRNA (22, 23). The molecular details of
79 primed adaptation are not well understood. Priming requires association of Cascade with a target
80 DNA molecule, and the newly acquired spacers correspond to locations on the same DNA
81 molecule as the protospacer (22–25). Some Type I CRISPR systems acquire spacers
82 preferentially in one direction relative to the targeted protospacer, with the majority of spacers in
83 either direction coming from the same DNA strand (22–27). In *E. coli*, priming is primarily
84 unidirectional, and has been proposed to involve translocation of Cas3 away from the Cascade-
85 bound protospacer (22, 25).
86
87 There are conflicting reports on the relationship between interference and priming. Initially, it
88 was proposed that priming occurs only when Cascade-protospacer interactions are sub-optimal
89 and cannot lead to interference, e.g. with a sub-optimal PAM, or with mismatches in the PAM-
90 proximal region of the spacer/protospacer known as the "seed" (22, 24, 28–30). However, more
91 recent studies have shown that at least some protospacers can lead to both interference and
92 priming, indicating that the requirements for interference and priming overlap (31–33).
93
94 Prior to interference or priming, the Cascade-crRNA complex must bind to the target
95 protospacer. This requires an interaction between Cse1 and the PAM, and base-pairing between
96 the crRNA and protospacer DNA (13, 16, 19). PAM recognition is required for both Cascade
97 binding and later activation of Cas3 (14, 18, 19, 34). Changes to the optimal PAM weaken
98 Cascade binding to a protospacer (35). Nonetheless, some sub-optimal PAMs are sufficient for
99 interference, albeit with lower efficiency than the optimal PAM (30). Sequences within the
100 crRNA spacer are also required for initial binding of Cascade to a protospacer; mutations in

positions 1-5 and 7-8 of the protospacer (the "seed sequence") reduce the affinity of Cascade for the protospacer (28, 36).

The precise sequence determinants for Cascade binding, interference and priming are unclear. Moreover, association of Cascade with protospacer DNA has not previously been studied in an *in vivo* context. Here, we use ChIP-seq to perform the first *in vivo* assessment of Cascade binding to its DNA targets. Our data show that base-pairing between the crRNA and protospacer with as few as 5 nt in the seed region, coupled with an optimal PAM, is often sufficient for Cascade binding. Hence, crRNAs, including those transcribed from the native *E. coli* CRISPR loci, drive off-target binding at hundreds of chromosomal sites. If Cascade binding to DNA was sufficient for interference or priming, these off-target binding events would be catastrophic for the bacterium. However, we show that near-complete base-pairing between the crRNA and protospacer is required for efficient interference and priming. Thus, under native conditions, the Cascade-crRNA complex samples potential DNA target sites, but limits nuclease activity to protospacers that meet a higher specificity threshold that would only be expected of on-target sites.

RESULTS

An AAG PAM and seed matches are sufficient for Cascade binding to DNA target sites *in vivo*

Previous studies of Cascade association with protospacer DNA have been *in vitro*, using purified Cascade and crRNA. To determine the *in vivo* target specificity of *E. coli* Cascade, we used ChIP-seq to map the association of Cse1-FLAG$_3$ and FLAG$_3$-*cas5* (FLAG-tagged strains retain CRISPR function; Figure S1) across the *E. coli* chromosome in Δ*cas3* (interference-deficient) cells constitutively expressing all other *cas* genes, and each of two crRNAs that target either the *lacZ* promoter or the *araB* promoter (both targets are chromosomal; Figure S2A-B). ChIP-seq data for Cse1 and Cas5 were highly correlated (R^2 values of 0.93-0.99 for *lacZ*-targeting cells, and 0.99 for *araB*-targeting cells), consistent with Cse1 and Cas5 always binding DNA together in the context of Cascade. We detected association of Cascade with many genomic loci for each of the two spacers tested (Figure 1A+B; Table S1). In all cases, the genomic region with strongest Cascade association was the on-target site at *lacZ* or *araB*. Off-target binding events occurred with <20% of the ChIP signal of on-target binding. To determine the sequence requirements for off-target Cascade binding with each of the two crRNAs used, we searched for enriched sequence motifs in the Cascade-bound regions, excluding the on-target site (Table S2). For both the *lacZ* and *araB* spacers, the most enriched sequence motif we identified was a close match to an AAG PAM, followed by 5 nt of sequence complementarity at the start of the seed region (Figure 1C-D; c.f. Figure S2A-B). In some cases, we observed Cascade binding events associated with non-AAG PAMs; however, these sites were more weakly bound, and/or had matches in seed region beyond position 5. We conclude that as few as 5 bp in the seed region,

141 together with an AAG PAM, are sufficient for Cascade binding, with additional base-pairing in
142 the seed region increasing binding and/or overcoming the need for an AAG PAM.
143
144 **Extensive off-target Cascade binding driven by endogenous spacers**
145 We identified several sites of Cascade binding that were shared between cells targeting *lacZ* and
146 cells targeting *araB*. These bound regions were not associated with sequences matching the seed
147 regions of either crRNA. We reasoned that these off-target binding events may be due to
148 Cascade association with the endogenous *E. coli* crRNAs. To test this hypothesis, we performed
149 ChIP-seq of Cse1-FLAG$_3$, as described above, for cells expressing only the endogenous CRISPR
150 RNAs from their native loci. Thus, we identified 188 binding sites for Cascade (Figure 2A;
151 Table S1). These sites were associated with four enriched sequence motifs, with each motif
152 corresponding to an AAG PAM and 5-10 nt matching the seed region of a crRNA from the
153 CRISPR-I array (spacers #1, #3, #4, and #8; Figure 2B; Figure S2C; Table S2). The strongest
154 binding events were associated with spacer #8 of CRISPR-I (Figure 2B; Figure S2C). To
155 confirm that Cascade binding events were due to association with endogenous crRNAs, we
156 repeated the ChIP-seq experiment in cells lacking the CRISPR-I array and cells lacking the
157 CRISPR-II array. Deletion of CRISPR-II had little effect on the profile of Cascade binding
158 (Figure 2C; Table S1). In contrast, deletion of CRISPR-I resulted in loss of Cascade binding to
159 almost all sites bound in wild-type cells (Figure 2D; Table S1). Instead, low-level binding of
160 Cascade was observed at a small number of sites that were associated with a weakly enriched
161 sequence motif corresponding to a perfect PAM and 8 nt matching the seed region of spacer #2
162 of CRISPR-II (Figure S2D + S3; Table S2).
163

CRISPR-I spacer #8 is the major determinant of off-target Cascade binding in cells expressing endogenous crRNAs.

Our data suggested that the majority of Cascade binding associated with endogenous crRNAs is due to CRISPR-I, and that the dominant spacer from CRISPR-I is spacer #8 ("sp8"). To confirm this, we measured Cascade binding by ChIP-seq in cells lacking CRISPR-I but expressing a plasmid-encoded sp8 crRNA. Most of the Cascade binding sites we observed were identical to those seen in cells expressing both CRISPR arrays, or cells expressing only CRISPR-I (Figure 3A; Table S1), and corresponded to regions containing strong matches to sp8 (orange dots in Figure 3A correspond to regions containing a match to the sp8 motif shown in Figure 2B). As expected, and unlike for cells expressing CRISPR-I, we detected only a single strongly enriched sequence motif (Figure S4A; Table S2). This motif, as expected, corresponds to an AAG PAM and 9 nt matching the seed region of sp8 (Figure S2C). We also detected a weakly enriched sequence motif (Figure S4B; Table S2) that corresponds to an AAG PAM and the 11 nt immediately downstream of the second repeat on the plasmid encoding the sp8 crRNA. This is likely due to formation of a non-canonical crRNA that consists of the sequence between the second repeat and the transcription terminator (Figure S2E). A transcription terminator hairpin has previously been shown to function analogously to repeat sequence in the *E. coli* crRNAs (37).

The most enriched Cascade target region in cells with CRISPR-I, and cells expressing sp8 crRNA, was inside the *yggX* gene. We identified a sequence in this region with an AAG PAM and matches to positions 1-5 and 7-10 of sp8 (Figure 3B). We used targeted ChIP-qPCR to measure Cascade binding to this site in cells lacking CRISPR-I but expressing plasmid-encoded

187 sp8. We compared binding of Cascade to *yggX* in wild-type cells, and cells where the putative
188 protospacer was mutated in the region predicted to bind the sp8 crRNA seed. As expected, we
189 observed greatly reduced Cascade binding at the mutated site relative to the wild-type site.
190 Similarly, we observed greatly reduced Cascade binding at the wild-type site when we expressed
191 a mutant sp8 with changes in the seed region (Figure 3C). However, when we combined the
192 mutant spacer with the mutant protospacer, base-pairing potential was restored, and we observed
193 wild-type levels of Cascade binding (Figure 3C). We conclude that sp8 is the major determinant
194 for off-target Cascade binding in cells expressing endogenous crRNAs.
195
196 **Off-target Cascade binding events do not affect local gene expression**
197 Cascade binding events can lead to transcription repression by preventing initiating RNA
198 polymerase binding to a promoter, or acting as a roadblock to elongating RNA polymerase
199 within a transcription unit (38, 39). To determine if off-target events driven by endogenous
200 spacers affect local gene expression, we measured global RNA levels using RNA-seq in Δ*cas3*
201 cells with other *cas* genes constitutively expressed, with either intact CRISPR arrays or a
202 ΔCRISPR-I deletion. We detected few differences in RNA levels between the two strains, and
203 none of the differences correspond to genes within 1 kb of a Cascade binding site identified by
204 ChIP-seq. We conclude that off-target binding by a Cas3-deficient complex does not impact
205 local gene expression.
206
207 **Off-target Cascade binding is not associated with interference**
208 Previous studies have suggested that extensive mismatches at the 3□ end of the
209 spacer/protospacer prevent interference (12, 18). To determine whether off-target Cascade

binding events lead to interference, we constructed a Δ*yggX* Δ*cas3* strain expressing all other *cas* genes, with both CRISPR arrays intact. We introduced a plasmid with the off-target protospacer from *yggX* that is an imperfect match to sp8, or an equivalent plasmid with a protospacer that is a perfect match to sp8. We transformed each of these strains with a plasmid expressing *cas3*, or an equivalent empty vector, simultaneously selecting for retention of the protospacer-containing plasmid. We reasoned that the number of viable transformants with the *cas3*-containing plasmid would be low for cells where interference caused loss of the protospacer-containing plasmid, since these cells would be killed by the antibiotic selection. In contrast, the number of viable transformants with the empty vector should be high in all cases. Thus, we measured the relative level of interference for each of the two protospacers. As expected, the protospacer that perfectly matches sp8 resulted in highly efficient interference, whereas the protospacer with the native *yggX* sequence (i.e. imperfect match to sp8) resulted in no detectable interference (Figure 4A). We conclude that off-target Cascade binding events do not cause interference.

Off-target Cascade binding is not associated with priming

The molecular determinants for priming have not been well studied. However, protospacers with multiple mismatches to a crRNA can still result in priming (24), and a recent study suggested that binding of Cascade to a protospacer with extensive mismatches, including in the seed, is sufficient to cause priming (12). To test whether off-target Cascade binding is sufficient for priming, we used the strains described above that contained a plasmid with a protospacer that is either an imperfect or a perfect match to sp8. We then introduced a plasmid with an inducible copy of *cas3*, under non-inducing conditions, to avoid interference. Following induction of *cas3* expression, we harvested cells and PCR-amplified the 5□ end of the CRISPR-II array to

233 determine whether new spacers had been acquired because of priming. We observed robust
234 primed spacer acquisition for the protospacer with a perfect match to sp8, but no detectable
235 spacer acquisition for the off-target protospacer with an imperfect match to sp8 (Figure 4B). We
236 conclude that off-target Cascade binding events do not cause priming.
237
238 **Strong Cascade binding to protospacers with extensive mismatches at the crRNA 3□ end**
239 To further delineate the protospacer sequence requirements for Cascade binding, interference and
240 priming, we constructed 13 variants of a protospacer that matches sp8. We selected sp8 because
241 it elicits robust Cascade binding, interference, and priming (Figures 3 + 4). The protospacer
242 variants (Figure 5A) included those with (variant i) complete sequence complementarity and an
243 optimal, AAG PAM; (variants ii - iii) non-optimal PAMs: CCG, which is expected to completely
244 abolish Cascade binding (35), and ATT, a sub-optimal sequence previously shown to cause
245 priming but not detectable interference (40); (variants iv - viii) two or three mismatches in the
246 first three positions of the seed; and (variants ix – xiii) stretches of ≥6 nt mismatches at various
247 positions within the protospacer.
248
249 We pooled cells containing each of the protospacer variants. We used ChIP of Cse1-$FLAG_3$ in
250 Δ*cas3* cells to measure association of Cascade with all protospacers within the pool (see
251 Methods). As expected, the protospacer with a CCG PAM (variant ii) had far less Cascade
252 association than did the optimal protospacer (variant i) (Figure 5A). We presume that the level of
253 ChIP signal for the protospacer with the CCG PAM (variant ii) represents the background of this
254 experiment. The protospacer with a sub-optimal, ATT PAM (iii), showed reduced Cascade
255 binding relative to the optimal protospacer (variant i), but was well above the experimental

256 background (Figure 5A). Similarly, mismatches in the seed region (variants iv - viii) resulted in
257 partial or complete loss of Cascade association, depending on the specific sequence mismatch
258 (Figure 5A). Our data for PAM and seed mutants are consistent with earlier studies showing that
259 these sequences are important for Cascade binding (21, 28, 35, 36).
260
261 Mismatches in the protospacer from positions 1-6 (variants xi and xii) or 7-20 (variant xiii)
262 abolished Cascade binding (Figure 5A). This is consistent with the observation from our ChIP-
263 seq data that sequence matches in positions 1-8 appear to be required for Cascade binding to off-
264 target sites using sp8 (Figure 2B + S4A). Strikingly, mismatches across positions 25-32 (variant
265 ix) or positions 19-32 (variant x) did not reduce Cascade association relative to the optimal
266 protospacer (variant i) (Figure 5A). In fact, these protospacer variants showed a modest increase
267 in Cse1 association relative to the optimal protospacer (variant i; Figure 5A), suggesting
268 conformational differences in the Cascade-DNA complex when the 3□ end of the crRNA is
269 mismatched with the protospacer.
270
271 **Near-complete crRNA-protospacer base-pairing is required for priming and interference**
272 We next determined which of the protospacer variants lead to interference. Using a modification
273 of a previously described assay (see Methods) (24, 40), we measured the level of interference
274 with a plasmid target for each of the 13 protospacers, using Δ*cas1* cells that cannot acquire new
275 spacers; primed spacer acquisition cannot contribute to the level of interference in these cells. As
276 expected, the optimal protospacer (i) was associated with robust levels of interference, whereas
277 protospacer variants that do not bind Cascade (variants ii, iv, v, xi, xii, and xiii; Figure 5A) were
278 not associated with detectable interference (Figure 5B). Protospacers with PAM and seed

279 variants that showed reduced but not abolished Cascade binding (variants iii, vi, vii, and viii;
280 Figure 5A) were associated with a range of interference levels that correlate well with the level
281 of Cascade binding. However, the ability of protospacers to cause interference did not always
282 correlate with the level of Cascade association. Specifically, we detected no interference for
283 either of the protospacer variants with mismatches only at the 3□ end (variants ix and x; Figure
284 5B), even though these protospacers bind Cascade at least as well as the optimal protospacer
285 (Figure 5A).
286
287 Previous studies have proposed that some protospacers with sub-optimal PAMs or mismatches in
288 the seed region are not subject to detectable interference, but are subject to priming (12, 22, 24,
289 40). We determined whether the 13 protospacer variants caused priming in a plasmid context.
290 Specifically, we introduced an inducible copy of *cas3* into cells containing each of the
291 protospacers on a high-copy plasmid. We then induced expression of *cas3*, and PCR-amplified
292 the CRISPR-II array to determine whether new spacers had been added. We observed robust
293 primed spacer acquisition for all protospacers associated with interference (variants i, iii, vii, and
294 viii; Figure 5C). By contrast, we observed no spacer acquisition for protospacers that do not bind
295 Cascade (variants ii, iv, v, xi, xii, and xiii; Figure 5C). Strikingly, we observed primed spacer
296 acquisition for two protospacers that were not associated with detectable interference (Figure
297 5C). One of these protospacers (variant vi) has the seed mismatch with the lowest level of
298 Cascade binding that is above the experimental background (Figure 5A). The other protospacer
299 has mismatches across positions 25-32 (variant ix). Thus, for these protospacers, we detected
300 robust Cascade binding and priming but we were unable to detect interference. For the

301 protospacer with mismatches across positions 19-32 (variant x), we detected no priming. Thus,

302 for this protospacer, we detected robust Cascade binding, but no priming or interference.

303

DISCUSSION

Base-pairing in the seed region together with an AAG PAM is sufficient for Cascade to bind DNA

Relatively little is known about the sequence determinants for Cascade-DNA binding, interference, and priming. Moreover, no previous studies have measured Cascade binding to protospacer DNA *in vivo*. Our ChIP data indicate that an AAG PAM and as little as 5 nucleotides of base-pairing at the start of the seed region are sufficient for *E. coli* Cascade to bind DNA targets. The sequence requirements for protospacer binding in Type II systems are similarly relaxed (41–43). The affinity of Cascade for a protospacer increases as the extent of base-pairing increases, but maximal affinity occurs with no more than an 18 bp match at the 5□ end (Figure 5A).

AAG is the optimal PAM in *E. coli*

Two previous studies proposed that AAG, GAG, TAG, AGG, and ATG are optimal PAMs in *E. coli* (24, 44), while another study suggested that AAG, ATG and GAG PAMs were associated with moderately higher affinity Cascade binding than an AGG PAM (35). Our data clearly indicate that AAG is the optimal PAM for off-target sites, with most off-target Cascade binding events being associated with an AAG PAM. Specifically, 65% of Cascade binding sites associated with a detectable motif have an AAG PAM for the crRNAs targeting *lacZ* and *araB*, and the plasmid-encoded sp8 crRNA. Moreover, off-target Cascade binding events with higher enrichment scores, suggestive of higher Cascade affinity, were more likely to be associated with an AAG PAM than Cascade binding events with lower enrichment scores (76% vs 61% for the

top 20% and bottom 80% of bound regions, respectively, after sorting by Cse1 enrichment level). We hypothesize that the dependence on the PAM for Cascade binding is increased in situations where base-pairing only occurs in the seed region. According to this model, complete or near-complete base-pairing between the crRNA and protospacer would weaken the requirement for an optimal PAM, obscuring differences in PAM affinity. This would explain why previous studies suggested that there are at least three optimal PAMs (24, 35, 44).

Defining the crRNA seed

The seed region of a crRNA has been previously defined as positions 1-5 and 7-8, with position 1 being immediately adjacent to the PAM (28). However, our data suggest that the length of the seed varies between crRNAs, since we observed off-target binding with some crRNAs that requires base-pairing in positions 1-5, whereas off-target binding for other crRNAs requires base-pairing up to position 9 (Figures 1-2, S3-S4). We propose that the crRNA sequence determines the length of the seed, and that this reflects the initial binding mode, prior to extended base-pair formation. Every 6^{th} position of the crRNA is flipped out in the Cascade-crRNA complex, and hence does not contribute to base-pairing (16, 45, 46). Consistent with this, the importance of position 6 for off-target binding is substantially less than that of positions 1-5 (Figures 1-2, S3-S4). Nonetheless, off-target protospacers had a sequence match to the crRNA at position 6 far more frequently than expected by chance (45% for the crRNAs targeting *lacZ* and *araB*, and the plasmid-encoded sp8 crRNA; Binomial Test *p*-value = $2.4e^{-10}$). We hypothesize that the initial binding of Cascade to a protospacer includes base-pairing interactions at position 6, but that the complex rapidly transitions to a conformation in which the 6^{th} position is flipped out of the helix. Our data are consistent with an *in vitro* study of another Type I-E system, where

position 6 was also shown to contribute to off-target Cascade binding (47). The apparent requirement for a sequence match at position 6 is not consistent across all crRNAs we tested, suggesting that the pathway towards stable seed base-pairing differs in a sequence-dependent manner.

Interference and priming require near-complete R-loop formation

Although binding of Cascade to a DNA target requires relatively little sequence identity, our data indicate that robust interference and priming require at least 18-25 bp, beginning in the seed region. This is consistent with *in vitro* data showing that near-complete R-loop formation is required to license Cas3 activity (12, 18). Thus, although Cascade binds DNA promiscuously, functional binding occurs with high specificity. Our data support a previously proposed model in which complete R-loop formation triggers a conformational change in Cascade at the 3′ end of the spacer, which is then transmitted, presumably through Cse2 to PAM-associated Cse1, 5′ to the spacer (18, 48). This change in Cse1 conformation then recruits Cas3, and/or activates the nuclease activity of Cas3, as suggested by a recent structural study (48). In support of this model, we detected higher ChIP signal for Cascade bound to protospacers without complete R-loop formation than those with complete R-loop formation (Figure 5A), suggesting that the conformation of Cascade with respect to the DNA changes upon R-loop completion, moderately decreasing ChIP efficiency.

Evidence that interference and priming are obligately coupled processes

Priming was initially proposed to be an alternative pathway to interference, with optimal PAM/seed sequences leading to interference, and sub-optimal sequences leading to priming (12,

17, 22, 24, 40, 49). However, primed spacer acquisition has been observed in situations where interference occurs, suggesting that priming and interference can be coupled processes (Figure 5, variants i, iii, vii, and viii) (23, 31–33). While these data show that priming and interference can occur at the same time at a population level, they do not necessarily indicate that individual priming and interference events are coupled. Moreover, while it has been proposed that interference and priming are obligately coupled (50), this has not been tested, and there are many examples where primed spacer acquisition has been observed in the absence of detectable interference (12, 22, 24, 31, 40, 49). Our data show that protospacers with seed sequence mismatches can cause detectable priming but not detectable interference when the protospacer is present on a multi-copy plasmid (Figure 5). Strikingly, for protospacers with seed mismatches, the levels of interference and priming correlate well with the level of Cascade binding (Figure 5). We detected primed spacer acquisition but not interference for the weakest-bound seed variant that has above-background levels of Cascade binding (Figure 5, variant vi). This is consistent with the expectation that primed spacer acquisition is a more sensitive readout of Cascade/Cas3 function since (i) it is an irreversible process, and (ii) it does not require destruction of all copies of the plasmid. Our data are consistent with a model in which low levels of interference are undetectable when plasmid replication outpaces plasmid degradation (50). We also observed primed spacer acquisition in the absence of detectable interference for a protospacer with mismatches across positions 25-32 (Figure 5, variant ix). We propose that this degree of mismatch at the 3□ end of the crRNA greatly reduces, but does not abolish, the isomerization of Cascade into the “active” state that recruits/activates Cas3.

Extensive, inert, off-target binding of Cascade

Cascade has many off-target binding sites due to its ability to bind DNA with low sequence-specificity. Consequently, the endogenous crRNAs transcribed from the bacterial genome result in extensive off-target binding, even in the absence of an on-target site. Since off-target binding does not involve complete R-loop formation, it has no deleterious effects on genome integrity. We also observed no impact on transcription associated with any of the off-target binding events, despite that fact that targeted Cascade binding is known to repress transcription by occluding promoters or acting as a roadblock for elongating RNA polymerase (38, 39). Transcription repression by Cascade is considerably weaker when targeting within a transcribed region (i.e. acting as a roadblock) (38). Given that the location of off-target Cascade binding sites is essentially random with respect to genome organization, and that genes make up ~90% of the *E. coli* genome, off-target Cascade binding is expected to be primarily intragenic. This may partly explain the lack of transcriptional impact. Moreover, a recent study showed that the level of repression by Cascade occlusion of a promoter is greatly reduced with as few as 6 bases mismatched at the 3□ end of the spacer/protospacer (51), suggesting that even intergenic off-target Cascade binding sites would be transcriptionally inert. We propose that incomplete R-loop formation results in an unstable Cascade-DNA complex with a relatively high rate of dissociation, such that it cannot compete effectively with initiating or elongating RNA polymerase. Consistent with this model, stable association of Cascade with DNA *in vitro* has been shown to require near-complete R-loop formation (20). We conclude that Type I CRISPR systems have evolved to tolerate off-target binding driven by the endogenous crRNAs, and are only functional at on-target sites. Given the length of crRNA spacers in Type I systems, there is no expectation of complete or near-complete spacer-protospacer base-pairing by chance. It is important to note that self-targeting by Type I CRISPR systems has been described previously,

419 but these would be considered "on-target" events, likely caused by acquisition of spacers from
420 the chromosome. As expected for spacers with perfect sequence complementarity, these self-
421 targeting crRNAs are typically functional in gene regulation and interference (52–54).
422
423 **Not all crRNAs are created equal**
424 The *E. coli* genome encodes at least 19 crRNAs, yet our data suggest that only four crRNAs
425 contribute to off-target binding of Cascade. All four of these crRNAs are encoded in the
426 CRISPR-I array, and the majority of off-target binding is driven by just one, sp8. The lack of off-
427 target binding driven by CRISPR-II crRNAs is likely due to weak transcription of this array,
428 which is repressed by H-NS (55). In contrast, the CRISPR-I array is likely co-transcribed with
429 the upstream *cas* genes, which are strongly transcribed in the strain used in this study. The
430 preference for specific spacers within CRISPR-I cannot be explained by differences in
431 expression levels, since the crRNAs are transcribed as a single RNA. Rather, biases in spacer
432 usage are more likely due to differential assembly of specific crRNAs into Cascade. Consistent
433 with this, a previous study surveyed crRNAs associated with Cascade. Spacers #2, #4 and #8
434 represented 68% of the Cascade-associated crRNAs (9). The cause of this bias is unclear, but
435 may in part be due to differences in RNA secondary structure between spacers, which could
436 impact the efficiency of RNA processing by Cas6e. Consistent with this, RNA secondary
437 structure of repeat sequences, and associated processing by Cas6, has been shown to be impacted
438 by spacer sequences in the Type I-D system of *Synechocystis* sp. PCC 6803 (56). Nonetheless, it
439 is likely that other factors influence the level of off-target binding, since the relative association
440 of crRNAs for spacers #2, #4 and #8 with Cascade is likely to be similar (9), but sp8 drives a
441 disproportionately high level of off-target binding.

METHODS

Strains and plasmids

All strains, plasmids, oligonucleotides and purchased, chemically synthesized dsDNA fragments are listed in Table S3. All strains are derivatives of MG1655 (57). CB386 has been previously described (38). CB36 contains a chloramphenicol resistance cassette in place of *cas3*. We removed this cassette using Flp recombinase, expressed from plasmid pCP20 (58), to generate strain AMD536. Epitope tagged strains AMD543 and AMD554 (Cse1-$FLAG_3$ and $FLAG_3$-Cas5, respectively), were generated using the previously described FRUIT method of recombineering (59). Cse1 was C-terminally tagged in AMD543 by inserting a $FLAG_3$ tag immediately upstream of codon 495 using oligonucleotides JW6364 and JW6365. Tagging of Cse1 resulted in an 8 amino acid C-terminal truncation. We predicted based on phylogenetic comparisons and on structural data (46) that this truncation would not impact the function of Cse1. Cas5 was N-terminally tagged in AMD554 by inserting $FLAG_3$ using oligonucleotides JW6272 and JW6273. LC060 is a derivative of was generated using (i) FRUIT (59) with oligonucleotides JW7537-JW7540 to delete the CRISPR-II locus, (ii) P1 transduction of the CB386 (Δ*cas3* P*cse1*)::(cat::P_{J23199}) region, (iii) FRUIT (59) to C-terminally tag Cse1 with $FLAG_3$ (as described above for AMD543), and (iv) pCP20-expressed Flp recombinase (58) to remove the *cat* cassette. LC074 is a derivative of AMD536 in which the CRISPR-I array was deleted using FRUIT (59) with oligonucleotides JW7529 and JW7530 and a synthesized dsDNA fragment (gBlock 14148263; Integrated DNA technologies). LC077 is a derivative of LC074 in which Cse1 was C-terminally tagged with $FLAG_3$ (as described above for AMD543). AMD566 is a derivative of AMD536 in which Cse1 was C-terminally tagged with $FLAG_3$ (as described above for

AMD543). LC099 is a derivative of AMD566 in which the off-target binding site for Cascade in *yggX* was mutated using FRUIT (59) with oligonucleotides JW7635-8. LC103 is a derivative of AMD536 in which the the *yggX* gene was replaced with a kanamycin resistance cassette using P1 transduction from the Keio Collection Δ*yggX*::*kan*^{R} strain (60). LC106 is a derivative of LC103 with an unmarked, scar-free deletion of *cas1* made using FRUIT with oligonucleotides JW7898-JW7901.

Plasmids that express crRNAs targeting the *lacZ* promoter (pCB380) and *araB* promoters (pCB381) have been described previously (38). All other crRNA-expressing plasmids are derivatives of pAMD179. pAMD179 was constructed by amplifying a DNA fragment from plasmid pAMD172 (Integrated DNA Technologies) using with oligonucleotides JW6421 and 6513. This DNA fragment was cloned into pBAD24 (61) cut with *Nhe*I and *Hin*dIII (NEB) using the In-Fusion method (Clontech). The inserted fragment contains two repeats from the CRISPR-I array, separated by a stuffer fragment containing *Xho*I and *Sac*II restriction sites, and an intrinsic transcription terminator downstream of the second repeat. To clone individual spacers, pairs of oligonucleotides were annealed, extended, and inserted using In-Fusion (Clontech) into the *Xho*I and *Sac*II sites of pAMD179 to generate pLC008 (with oligonucleotides JW6518 and JW7911), pLC010 (with oligonucleotides JW6518 and JW7912), and pAMD189 (with oligonucleotides JW7598 and JW7693).

pLC021 and pLC022 are derivatives of pBAD24 (61) containing a protospacer matching the off-target Cascade binding site in *yggX* (pLC021) or a protospacer with a perfect match to sp8 (pLC022). These plasmids were constructed by annealing and extending pairs of

oligonucleotides (JW7913 and JW7914 for pLC021, and JW7924 and JW7925 for pLC022), and cloning the resultant DNA fragments into the *Eco*RV and *Sph*I sites of pBAD24. pAMD191 is a derivative of pBAD33 (61) that expresses *cas3* under arabinose control. To construct pAMD191, *cas3* was amplified by colony PCR using oligonucleotides JW7736 and JW7738. The PCR product was cloned into the *Sac*I and *Hin*dIII sites of pBAD33 using In-Fusion (Clontech). All protospacers described in Figure 5 are cloned into plasmid pLC020, the "pre-protospacer plasmid", which is a derivative of pBAD24 (61). pLC020 was generated by cloning the ~500 bp region upstream of *E. coli thyA* (amplified by colony PCR using oligonucleotides JW8040 and JW8128) and the ~500 bp region downstream of *E. coli thyA* (amplified by colony PCR using oligonucleotides JW8042 and JW8043) into the *Eco*RI site of pBAD24 using In-Fusion (Clontech), simultaneously generating a new *Eco*RI site between the upstream and downstream regions of *thyA*. The *thyA* gene was then amplified by colony PCR using a universal forward primer (oligonucleotide JW8129) and each of 13 reverse primers (oligonucleotides JW8130, JW8139, JW8145, JW8169, JW8499-JW8502, and JW8675-JW8679) containing the 13 protospacer variants described in Figure 5. The resulting PCR products were cloned into the *Eco*RI site of the pBAD24 derivative using In-Fusion (Clontech) to generate plasmids pLC023-pLC035 (see Table S3 for details).

ChIP-qPCR

For all ChIP-qPCR and ChIP-seq experiments, cells were grown overnight in LB, subcultured in LB supplemented with 0.2% arabinose and 100 μg/mL ampicillin (for experiments where a crRNA was expressed from a plasmid) at 37 °C with aeration to an OD_{600} of ~0.6. AMD566 and LC099 with either pLC008 or pLC010 were used for ChIP-qPCR. ChIP-qPCR was performed as

511 described previously (62), except that 2 µL anti-FLAG M2 monoclonal antibody (Sigma) and 1
512 µL anti-σ^{54} monoclonal antibody (NeoClone) were included simultaneously in the
513 immunoprecipitation step. qPCR was performed using oligonucleotides JW7490-1 (amplifies the
514 off-target site in *yggX*) and JW7922-3 (amplifies the region upstream of *hypA*). Since σ^{54} is
515 known not to bind within *yggX* (63), we were able to normalize binding of Cse1 within *yggX* to
516 the binding of σ^{54} upstream of *hypA*.
517
518 **ChIP-seq**
519 Strains AMD543, AMD554, LC060, LC077, AMD543/AMD554 with pCB380/pCB381, and
520 strain LC074 with pLC008, were used for ChIP-seq of Cse1-$FLAG_3$ and $FLAG_3$-Cas5. Cells
521 were grown and processed as described for ChIP-qPCR. ChIP-seq was performed in duplicate,
522 following a previously described protocol (64) using 2 µL anti-FLAG M2 monoclonal antibody
523 (Sigma). Sequencing was performed on an Illumina High-Seq 2000 Instrument (Next-Generation
524 Sequencing and Expression Analysis Core, State University of New York at Buffalo) or an
525 Illumina Next-Seq Instrument (Wadsworth Center Applied Genomic Technologies Core). ChIP-
526 seq data analysis was performed as previously described (65), with reads mapped to the updated
527 MG1655 *E. coli* genome (accession code U00096.3). Relative sequence coverage values were
528 calculated by dividing the sequence read coverage at a given genomic location by (total number
529 of sequence reads in the run/100,000). Values plotted in Figures 1A-B, 2A and 2D are the
530 maximum values in 1 kbp regions across the genome. R^2 values comparing ChIP-seq datasets
531 were calculated by comparing read coverage at peak centers for all peaks identified for the
532 analyzed datasets. Read coverage at peak centers was determined using a custom Python script.
533 Sequence motifs were identified using MEME (version 4.12.0) (66) with default parameters.

534

535 **RNA-seq**

536 RNA-seq was performed in duplicate with strains AMD536 and LC074. Cells were grown
537 overnight in LB, subcultured in LB supplemented with 0.2% arabinose at 37 °C with aeration to
538 an OD_{600} of ~0.6. RNA was purified using a modified hot phenol method, as previously
539 described (67). Purified RNA was treated with 2 µL DNase (TURBO DNA-free kit; Life
540 Technologies) for 45 minutes at 37 °C, followed by phenol extraction and ethanol precipitation.
541 The RiboZero kit (Epicure) was used to remove rRNA, and strand-specific cDNA libraries were
542 created using the ScriptSeq 2.0 kit (Epicure). Sequencing was performed using an Illumina Next-
543 Seq Instrument (Wadsworth Center Applied Genomic Technologies Core). Differential RNA
544 expression analysis was performed using Rockhopper (version 2.03) using default parameters
545 (68). Differences in RNA levels were considered statistically significant for genes with *q*-values
546 ≤ 0.01.

547

548 **Plasmid transformation efficiency assay**

549 LC103 was transformed with either pLC021 or pLC022. These strains were then transformed
550 with either empty pBAD33 or pAMD191 (expresses *cas3*), and cells were plates on M9 medium
551 supplemented with 0.2% glycerol, 0.2% arabinose, 100 µg/mL ampicillin and 30 µg/mL
552 chloramphenicol at 37 °C. After overnight growth, colonies were counted, and the ratio of
553 pAMD191-transformed cells to pBAD33-transformed cells was calculated for each of the two
554 strains.

555

556 **PCR to assess primed spacer acquisition**

Primed spacer acquisition was assessed for AMD536 with pAMD191 and either pLC021 or pLC022 (Figure 4B), LC103 with pAMD191 and each of pLC023-pLC035 (Figure 5C), and AMD543/AMD544 with pAMD191 and pAMD189 (expresses a self-targeting crRNA; Figure S1). Cells were grown overnight in LB supplemented with 100 µg/mL ampicillin and 30 µg/mL chloramphenicol at 37°C with aeration, and sub-cultured the next day in LB supplemented with chloramphenicol and 0.2% arabinose at 37°C with aeration for six hours. Cells were pelleted from 1 mL of culture by centrifugation, and cell pellets were frozen at -20°C. PCRs were then performed on the cell pellets, amplifying the CRISPR-II array using oligonucleotides JW7818 and JW7819. PCR products were visualized on acrylamide gels.

Sequence analysis of protospacers from a pooled ChIP library

LC099 with each of the 13 protospacer variant plasmids (pLC23-pLC035), was grown overnight in LB supplemented with 100 µg/mL ampicillin. 10 mL subcultures were grown in LB supplemented with 100 µg/mL ampicillin and 0.2% arabinose at 37°C with aeration to an OD_{600} of ~0.6. 3 mL from each culture was combined. ChIP was performed on mixed cultures 2 µL M2 anti-FLAG monoclonal antibody (Sigma), as previously described (62). A Zymo PCR Clean and Concentrate kit was used to purified ChIP and input DNA. A 50 µL FailSafe (Epicentre) PCR reaction using FailSafe PCR 2X PreMix “C” and 5.48 ng of ChIP DNA was performed following the manufacturer’s instructions, using oligonucleotide JW8567 and each of oligonucleotides JW8537, JW8556, JW8557, JW8558, JW8559, JW8561, JW8562, JW8563, JW8564, and JW8565 (these incorporate different Illumina indices). PCR products were purified and concentrated using 0.8X Ampure Beads (Beckman Coulter Life Sciences) and sequenced on an Illumina Mi-Seq Instrument (Wadsworth Center Applied Genomic Technologies Core).

Sequence reads were mapped to each of the 13 protospacer variants using a custom Pythom script. Relative protospacer abundance in input and ChIP samples for each protospacer were normalized to the total sequence reads. Values for normalized protospacer abundance were further normalized to values from the input sample. Protospacer abundance values are reported relative to those for the optimal protospacer (variant i in Figure 5).

Measuring interference for a pooled protospacer library

Overnight cultures of LC106 strains with each of the 13 protospacer plasmids (pLC23-pLC035) were grown in LB with 100 μg/mL ampicillin and 30 μg/mL kanamycin. All 13 cultures were combined to make a single subculture; 7.7 μL of each strain into a 10 mL culture. Electrocompetent cells were made and transformed with either empty pBAD33 or pAMD191 (pBAD33-*cas3*). Transformants were plated onto M9 agar supplemented with 0.2% glycerol, 0.2% arabinose, and 30 μg/mL chloramphenicol, and grown overnight at 37°C. Cells were scraped off plates, washed in LB, and protospacers were PCR amplified from cell pellets with oligonucleotide JW8567 and each of oligonucleotides JW8537, JW8558, JW8559, JW8562, JW8563, and JW8566 (these incorporate different Illumina indices). PCR products were purified and concentrated with 0.8X Ampure Beads (Beckman Coulter Life Sciences), and sequenced using a Illumina Mi-Seq Instrument (Wadsworth Center Applied Genomic Technologies Core). Sequence reads were mapped to each of the 13 protospacer variants using a custom Pythom script. Individual protospacer abundances were compared between Cas3-expressing cells and cells containing empty pBAD33. Protospacer abundances were normalized to those for the protospacer with a CCG PAM (variant ii in Figure 5).

603 **ACKNOWLEDGEMENTS**

604

605 We thank Chase Beisel for sharing strains and plasmids. We thank the Wadsworth Center
606 Applied Genomic Technologies Core Facility and the University at Buffalo Genomics and
607 Bioinformatics Core Facility for Sanger and MiSeq sequencing. We thank the Wadsworth Center
608 Media and Tissue Culture and Glassware Core Facilities. We thank Todd Gray, Keith
609 Derbyshire, and Shailab Shrestha for helpful discussions. This study was supported by NIH
610 Grant AI126416 (to J.T.W.) and a University at Albany, SUNY, RNA Fellowship (to L.A.C.).

611

612 **REFERENCES**

613

614 1. Grissa I, Vergnaud G, Pourcel C (2007) The CRISPRdb database and tools to display
615 CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8(1):1–
616 10.
617 2. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD (2005) Clustered regularly interspaced short
618 palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiol Read*
619 *Engl* 151(Pt 8):2551–2561.
620 3. Kunin V, Sorek R, Hugenholtz P (2007) Evolutionary conservation of sequence and
621 secondary structures in CRISPR repeats. *Genome Biol* 8(4):R61.
622 4. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV (2006) A putative RNA-
623 interference-based immune system in prokaryotes: computational analysis of the predicted
624 enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical
625 mechanisms of action. *Biol Direct* 1:7.
626 5. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E (2005) Intervening sequences
627 of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*
628 60(2):174–182.
629 6. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C (2009) Short motif
630 sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiol*
631 *Read Engl* 155(Pt 3):733–740.
632 7. Pourcel C, Salvignol G, Vergnaud G (2005) CRISPR elements in Yersinia pestis acquire
633 new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for
634 evolutionary studies. *Microbiology* 151(3):653–663.

635 8. Nuñez JK, et al. (2014) Cas1–Cas2 complex formation mediates spacer acquisition during
636 CRISPR–Cas adaptive immunity. *Nat Struct Mol Biol* 21(6):528–534.
637 9. Brouns SJJ, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes.
638 *Science* 321(5891):960–964.
639 10. Carte J, Wang R, Li H, Terns RM, Terns MP (2008) Cas6 is an endoribonuclease that
640 generates guide RNAs for invader defense in prokaryotes. *Genes Dev* 22(24):3489–3496.
641 11. Jore MM, et al. (2011) Structural basis for CRISPR RNA-guided DNA recognition by
642 Cascade. *Nat Struct Mol Biol* 18(5):529–536.
643 12. Blosser TR, et al. (2015) Two distinct DNA binding modes guide dual roles of a CRISPR-
644 Cas protein complex. *Mol Cell* 58(1):60–70.
645 13. Hayes RP, et al. (2016) Structural basis for promiscuous PAM recognition in type I–E
646 Cascade from E. coli. *Nature* advance online publication. doi:10.1038/nature16995.
647 14. Hochstrasser ML, et al. (2014) CasA mediates Cas3-catalyzed target degradation during
648 CRISPR RNA-guided interference. *Proc Natl Acad Sci U S A* 111(18):6618–6623.
649 15. Mulepati S, Bailey S (2013) In vitro reconstitution of an Escherichia coli RNA-guided
650 immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J Biol*
651 *Chem* 288(31):22184–22192.
652 16. Mulepati S, Héroux A, Bailey S (2014) Structural biology. Crystal structure of a CRISPR
653 RNA-guided surveillance complex bound to a ssDNA target. *Science* 345(6203):1479–
654 1484.
655 17. Redding S, et al. (2015) Surveillance and Processing of Foreign DNA by the Escherichia
656 coli CRISPR-Cas System. *Cell* 163(4):854–865.

18. Rutkauskas M, et al. (2015) Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. *Cell Rep*. doi:10.1016/j.celrep.2015.01.067.

19. Sashital DG, Wiedenheft B, Doudna JA (2012) Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol Cell* 46(5):606–615.

20. Szczelkun MD, et al. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc Natl Acad Sci U S A* 111(27):9798–9803.

21. Westra ER, et al. (2012) CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol Cell* 46(5):595–605.

22. Datsenko KA, et al. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat Commun* 3:945.

23. Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ (2012) CRISPR Interference Directs Strand Specific Spacer Acquisition. *PLoS ONE* 7(4):e35888.

24. Fineran PC, et al. (2014) Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci* 111(16):E1629–E1638.

25. Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. *RNA Biol* 10(5):716–725.

26. Li M, Wang R, Zhao D, Xiang H (2014) Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res* 42(4):2483–2492.

679 27. Richter C, et al. (2014) Priming in the Type I-F CRISPR-Cas system triggers strand-
680 independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids*
681 *Res* 42:8516–8526.
682 28. Semenova E, et al. (2011) Interference by clustered regularly interspaced short palindromic
683 repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A*
684 108(25):10098–10103.
685 29. Wiedenheft B, et al. (2011) Structures of the RNA-guided surveillance complex from a
686 bacterial immune system. *Nature* 477(7365):486–489.
687 30. Xue C, et al. (2015) CRISPR interference and priming varies with individual spacer
688 sequences. *Nucleic Acids Res* 43:10831–10847.
689 31. Künne T, et al. (2016) Cas3-Derived Target DNA Degradation Fragments Fuel Primed
690 CRISPR Adaptation. *Mol Cell* 63(5):852–864.
691 32. Semenova E, et al. (2016) Highly efficient primed spacer acquisition from targets destroyed
692 by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proc Natl Acad*
693 *Sci*:201602639.
694 33. Staals RHJ, et al. (2016) Interference-driven spacer acquisition is dominant over naive and
695 primed adaptation in a native CRISPR–Cas system. *Nat Commun* 7:12853.
696 34. Sinkunas T, et al. (2013) In vitro reconstitution of Cascade-mediated CRISPR immunity in
697 Streptococcus thermophilus. *EMBO J* 32(3):385–394.
698 35. Westra ER, et al. (2013) Type I-E CRISPR-Cas Systems Discriminate Target from Non-
699 Target DNA through Base Pairing-Independent PAM Recognition. *PLoS Genet* 9(9).
700 doi:10.1371/journal.pgen.1003742.

36. Wiedenheft B, et al. (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci U S A* 108(25):10092–10097.

37. Semenova E, et al. (2015) The Cas6e ribonuclease is not required for interference and adaptation by the E. coli type I-E CRISPR-Cas system. *Nucleic Acids Res* 43(12):6049–6061.

38. Luo ML, Mullis AS, Leenay RT, Beisel CL (2014) Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. *Nucleic Acids Res* 43:674–681.

39. Rath D, Amlinger L, Rath A, Lundgren M (2015) The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* 117:119–128.

40. Xue C, et al. (2015) CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Res*:gkv1259.

41. Duan J, et al. (2014) Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Res* 24(8):1009–1012.

42. Kuscu C, Arslan S, Singh R, Thorpe J, Adli M (2014) Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol* 32(7):677–683.

43. Wu X, Kriz AJ, Sharp PA (2014) Target specificity of the CRISPR-Cas9 system. *Quant Biol* 2(2):59–70.

44. Leenay RT, et al. (2016) Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Mol Cell* 62(1):137–147.

45. Jackson RN, et al. (2014) Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science* 345(6203):1473–1479.

724 46. Zhao H, et al. (2014) Crystal structure of the RNA-guided immune surveillance Cascade
725 complex in Escherichia coli. *Nature* 515(7525):147–150.
726 47. Jung C, et al. (2017) Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes
727 on Next Generation Sequencing Chips. *Cell* 170(1):35–47.e13.
728 48. Xiao Y, et al. (2017) Structure Basis for Directional R-loop Formation and Substrate
729 Handover Mechanisms in Type I CRISPR-Cas System. *Cell* 170(1):48–60.e11.
730 49. Xue C, Whitis NR, Sashital DG (2016) Conformational Control of Cascade Interference
731 and Priming Activities in CRISPR Immunity. *Mol Cell* 64(4):826–834.
732 50. Severinov K, Ispolatov I, Semenova E (2016) The Influence of Copy-Number of Targeted
733 Extrachromosomal Genetic Elements on the Outcome of CRISPR-Cas Defense. *Front Mol*
734 *Biosci* 3:45.
735 51. Luo ML, et al. (2016) The CRISPR RNA-guided surveillance complex in Escherichia coli
736 accommodates extended RNA spacers. *Nucleic Acids Res* 44:7385–7394.
737 52. Heussler GE, O'Toole GA (2016) Friendly Fire: Biological Functions and Consequences of
738 Chromosomal Targeting by CRISPR-Cas Systems. *J Bacteriol* 198(10):1481–1486.
739 53. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R (2010) Self-targeting by CRISPR: gene
740 regulation or autoimmunity? *Trends Genet TIG* 26(8):335–340.
741 54. Vercoe RB, et al. (2013) Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can
742 Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. *PLoS Genet*
743 9(4):e1003454.
744 55. Pul Ü, et al. (2010) Identification and characterization of E. coli CRISPR-cas promoters and
745 their silencing by H-NS. *Mol Microbiol* 75(6):1495–1512.

746 56. Reimann V, et al. (2017) Structural constraints and enzymatic promiscuity in the Cas6-
747 dependent generation of crRNAs. *Nucleic Acids Res* 45:915–925.
748 57. Blattner FR, et al. (1997) The complete genome sequence of Escherichia coli K-12. *Science*
749 277(5331):1453–1462.
750 58. Cherepanov PP, Wackernagel W (1995) Gene disruption in Escherichia coli: TcR and KmR
751 cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant.
752 *Gene* 158(1):9–14.
753 59. Stringer AM, et al. (2012) FRUIT, a scar-free system for targeted chromosomal
754 mutagenesis, epitope tagging, and promoter replacement in Escherichia coli and Salmonella
755 enterica. *PloS One* 7(9):e44841.
756 60. Baba T, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout
757 mutants: the Keio collection. *Mol Syst Biol* 2:2006.0008.
758 61. Guzman LM, Belin D, Carson MJ, Beckwith J (1995) Tight regulation, modulation, and
759 high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol*
760 177(14):4121–4130.
761 62. Bonocora RP, Fitzgerald DM, Stringer AM, Wade JT (2013) Non-canonical protein-DNA
762 interactions identified by ChIP are not artifacts. *BMC Genomics* 14:254.
763 63. Bonocora RP, Smith C, Lapierre P, Wade JT (2015) Genome-Scale Mapping of Escherichia
764 coli σ54 Reveals Widespread, Conserved Intragenic Binding. *PLoS Genet*
765 11(10):e1005552.
766 64. Singh SS, et al. (2014) Widespread suppression of intragenic transcription initiation by H-
767 NS. *Genes Dev*. doi:10.1101/gad.234336.113.

768 65. Fitzgerald DM, Bonocora RP, Wade JT (2014) Comprehensive Mapping of the Escherichia
769 coli Flagellar Regulatory Network. *PLOS Genet* 10(10):e1004649.
770 66. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to
771 discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28–36.
772 67. Stringer AM, et al. (2013) Genome-Scale Analyses of Escherichia coli and Salmonella
773 enterica AraC Reveal Non-Canonical Targets and an Expanded Core Regulon. *J*
774 *Bacteriol*:JB.01007-13.
775 68. McClure R, et al. (2013) Computational analysis of bacterial RNA-Seq data. *Nucleic Acids*
776 *Res* 41(14):e140.
777
778

FIGURE LEGENDS

Figure 1. Extensive off-target Cascade binding in *E. coli*. (A) Binding profile of Cse1 across the *E. coli* genome, as determined by ChIP-seq, for Cse1-$FLAG_3$ cells (AMD543) expressing a plasmid-encoded crRNA targeting the *lacZ* promoter region (pCB380). The graph indicates the relative sequence read coverage (see Methods for details) across the genome in a Cse1 ChIP-enriched sample. A scale-bar is shown for 1 Mbp. **(B)** Binding profile of Cas5 across the *E. coli* genome, as determined by ChIP-seq, for $FLAG_3$-Cas5 cells (AMD554) expressing a plasmid-encoded crRNA targeting the *araB* promoter region (pCB381). **(C)** Enriched sequence motif associated with off-target Cascade binding sites when targeting the *lacZ* promoter, as determined by MEME. The number of identified motifs and the MEME E-value are shown. **(D)** Enriched sequence motif associated with off-target Cascade binding sites when targeting the *araB* promoter, as determined by MEME.

Figure 2. Endogenous crRNAs drive Cascade association with hundreds of chromosomal sites. (A) Binding profile of Cse1 across the *E. coli* genome, as determined by ChIP-seq, for Cse1-$FLAG_3$ cells (AMD543) expressing only endogenous crRNAs. **(B)** Enriched sequence motifs associated with Cascade binding sites in cells expressing only endogenous crRNAs. The four motifs are associated with four of the CRISPR-I spacers, as indicated. The likely PAM sequence is also indicated. The number of identified motifs and the MEME E-value are shown. **(C)** Comparison of Cascade binding events in Cse1-$FLAG_3$ cells with both CRISPR arrays intact (AMD543) and or CRISPR-II deleted (LC060). Sequence read coverage is shown for CRISPR-1^+ CRISPR-2^+ (AMD543) and CRISPR-I^+ ΔCRISPR-II cells (LC060), for all ChIP-seq peaks

identified for either strain. **(D)** Binding profile of Cse1 across the *E. coli* genome, as determined by ChIP-seq, for Cse1-$FLAG_3$ cells expressing only endogenous crRNAs, but with CRISPR-I deleted (LC077).

Figure 3. CRISPR-I Spacer #8 is responsible for the majority of Cascade binding in cells expressing only endogenous crRNAs. (A) Comparison of Cse1-$FLAG_3$ binding events in cells with both CRISPR arrays intact (AMD543), and cells with CRISPR-I deleted (LC077) that express CRISPR-I spacer #8 from a plasmid (pLC008). Sequence read coverage is shown for all ChIP-seq peaks identified for either strain. ChIP-seq peaks associated with the CRISPR-I spacer #8 motif (first motif listed in Figure 2B) are shown in orange. **(B)** Predicted base-pairing interaction between CRISPR-I spacer #8 and a protospacer within *yggX*. **(C)** ChIP-qPCR measurement of Cse1 binding at wild-type (i and iii; AMD566) and mutant (ii and iv; LC099) protospacers in *yggX* for cells expressing wild-type (i and ii; pLC008) or mutant (iii and iv; pLC010) CRISPR-I spacer #8 from a plasmid. The mutations in spacer #8 restore base-pairing potential with the mutant protospacer, as indicated. Values represent the average of three independent replicate experiments. Error bars show one standard deviation from the mean.

Figure 4. Off-target Cascade binding events are not associated with interference or priming. (A) Relative efficiency of transformation of a *cas3*-expressing plasmid (pAMD191) into cells expressing CRISPR-I spacer #8 (LC103), and either (i) a protospacer that base-pairs perfectly with spacer #8 (pLC022), (ii) the protospacer from *yggX* that has only partial base-pairing with spacer #8 (pLC021), or (iii) no protospacer (pBAD24). Transformation efficiency was calculated relative to that of empty pBAD33, as described in the Methods. Values represent

825 the average of three independent replicate experiments. Error bars show one standard deviation
826 from the mean. **(B)** PCR-amplification of the start of the CRISPR-II array to detect primed
827 spacer acquisition in cells expressing CRISPR-I spacer #8 (AMD536), *cas3* (pAMD191), and
828 with (i) a protospacer that base-pairs perfectly with spacer #8 (pLC022), or (ii) the protospacer
829 from *yggX* that has only partial base-pairing with spacer #8 (pLC021). L = molecular weight
830 ladder, with marker sizes (bp) indicated. The expected PCR product sizes are indicated. Note that
831 there is a non-specific PCR product of ~380 bp for both samples i and ii.
832
833 **Figure 5. Assessment of Cascade binding, interference and priming for a panel of**
834 **protospacer variants. (A)** Relative Cse1 association (in strain LC099) for each of 13
835 protospacer variants: (i) optimal protospacer that has a perfect match to CRISPR-I spacer #8 and
836 an AAG PAM (pLC023); (ii) CCG PAM (pLC027); (iii) ATT PAM (pLC029); (iv) mismatches
837 at positions 1-3 (pLC031; wild-type is CTG); (v) mismatches at positions 1 and 3 (pLC033); (vi)
838 mismatches at positions 1 and 3 (pLC035); (vii) mismatches at positions 2 and 3 (pLC032);
839 (viii) mismatches at positions 2 and 3 (pLC034); (ix) mismatches across positions 25-32
840 (pLC024); (x) mismatches across positions 19-32 (pLC025); (xi) mismatches across positions 1-
841 6 (pLC026); (xii) mismatches across positions 1-6 and 25-32 (pLC028); (xiii) mismatches across
842 positions 7-24 (pLC030). Values represent the average of five independent replicate
843 experiments. Error bars show one standard deviation from the mean. **(B)** Relative efficiency of
844 transformation of a *cas3*-expressing plasmid (pAMD191) into a pool of cells (LC106) containing
845 each of the indicated protospacer variants (pLC023-pLC035). See Methods for details. Values
846 represent the average of three independent replicate experiments. Error bars show one standard
847 deviation from the mean. **(C)** PCR-amplification of the start of the CRISPR-II array to detect

848 primed spacer acquisition in cells expressing CRISPR-I spacer #8 (LC103 + pLC008), *cas3*
849 (pAMD191), and with each of the 13 indicated protospacer variants (pLC023-pLC035).
850
851

852 **LIST OF SUPPLEMENTARY FIGURES AND TABLES**

853

854 **Figure S1. $FLAG_3$-tagged Cse1 and Cas5 are fully functional for primed spacer acquisition.**
855 PCR-amplification of the start of the CRISPR-II array to detect primed spacer acquisition in cells
856 expressing *cas3* (pAMD191) and a plasmid-encoded crRNA that perfectly targets a sequence on
857 the same plasmid (pAMD189). Spacer acquisition was assessed for (i) MG1655 (no *cas* gene
858 expression, except for plasmid-encoded *cas3* (pAMD191)), (ii) MG1655 with constitutive
859 expression of *cas* genes (AMD536), (iii) MG1655 *cse1*-$FLAG_3$ with constitutive expression of
860 *cas* genes (AMD543), and (iv) MG1655 $FLAG_3$-*cas5* with constitutive expression of *cas* genes
861 (AMD554).

862

863 **Figure S2. crRNA spacers used in this study. (A)** Sequence of the crRNA spacer targeting the
864 *lacZ* promoter (pCB380). **(B)** Sequence of the crRNA spacer targeting the *araB* promoter
865 (pCD381). **(C)** Sequence of the CRISPR-I array. Spacers #1, #3, #4 and #8 are underlined. **(D)**
866 Sequence of the CRISPR-II array. Spacer #2 is underlined. **(E)** Sequence of a portion of the
867 spacer #8 crRNA-expressing plasmid (pLC008). Note that the sequence downstream of the
868 second repeat (underlined) can be used as a spacer.

869

870 **Figure S3. Spacer #2 of CRISPR-II directs Cascade binding in cells lacking CRISPR-I.**
871 Enriched sequence motif associated with Cascade binding sites in cells expressing only
872 endogenous crRNAs, where CRISPR-I is deleted (LC077). The motif is associated with
873 CRISPR-II spacer #2, as indicated. The likely PAM sequence is also indicated. The number of
874 identified motifs and the MEME E-value are shown.

875

876 **Figure S4. Sequence motifs associated with Cascade binding in cells expressing CRISPR-I**
877 **spacer #8 from a plasmid. (A)** Sequence of the most strongly enriched motif, as identified by
878 MEME, in ΔCRISPR-I cells (LC077) expressing spacer #8 from a plasmid (pLC008). The motif
879 is associated with CRISPR-I spacer #8, as indicated. The likely PAM sequence is also indicated.
880 The number of identified motifs and the MEME E-value are shown. **(B)** The second enriched
881 sequence motif, as identified by MEME, in ΔCRISPR-I cells (LC077) expressing spacer #8 from
882 a plasmid (pLC008). The motif is associated with the sequence immediately downstream of the
883 second repeat on the crRNA plasmid, as indicated.

884

885 **Table S1. Lists of ChIP-seq peak coordinates.**

886

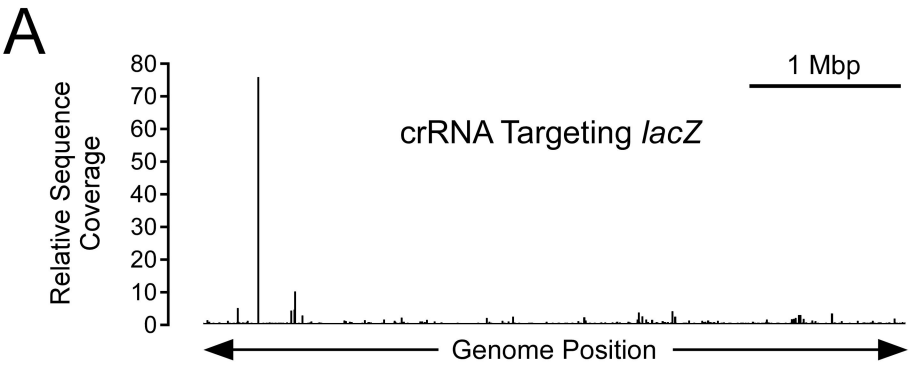
887 **Table S2. Lists of regions used to search for enriched sequence motifs.**

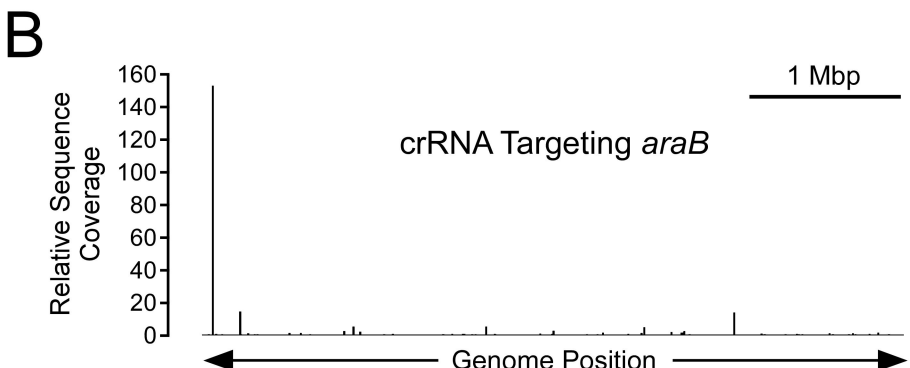
888

889 **Table S3. Strains, Plasmids, Oligonucleotides, and Chemically Synthesized dsDNA**
890 **fragments used in this study.**

891

Figure 1


A
Relative Sequence Coverage
80
70
60
50
40
30
20
10
0
1 Mbp
crRNA Targeting *lacZ*
Genome Position


B
Relative Sequence Coverage
160
140
120
100
80
60
40
20
0
1 Mbp
crRNA Targeting *araB*
Genome Position


C
PAM
n = 86
E = 5.3e-112
CTTTACAC . . .


D
PAM
n = 60
E = 8.6e-69
ATTAGCGGA . . .

Figure 2

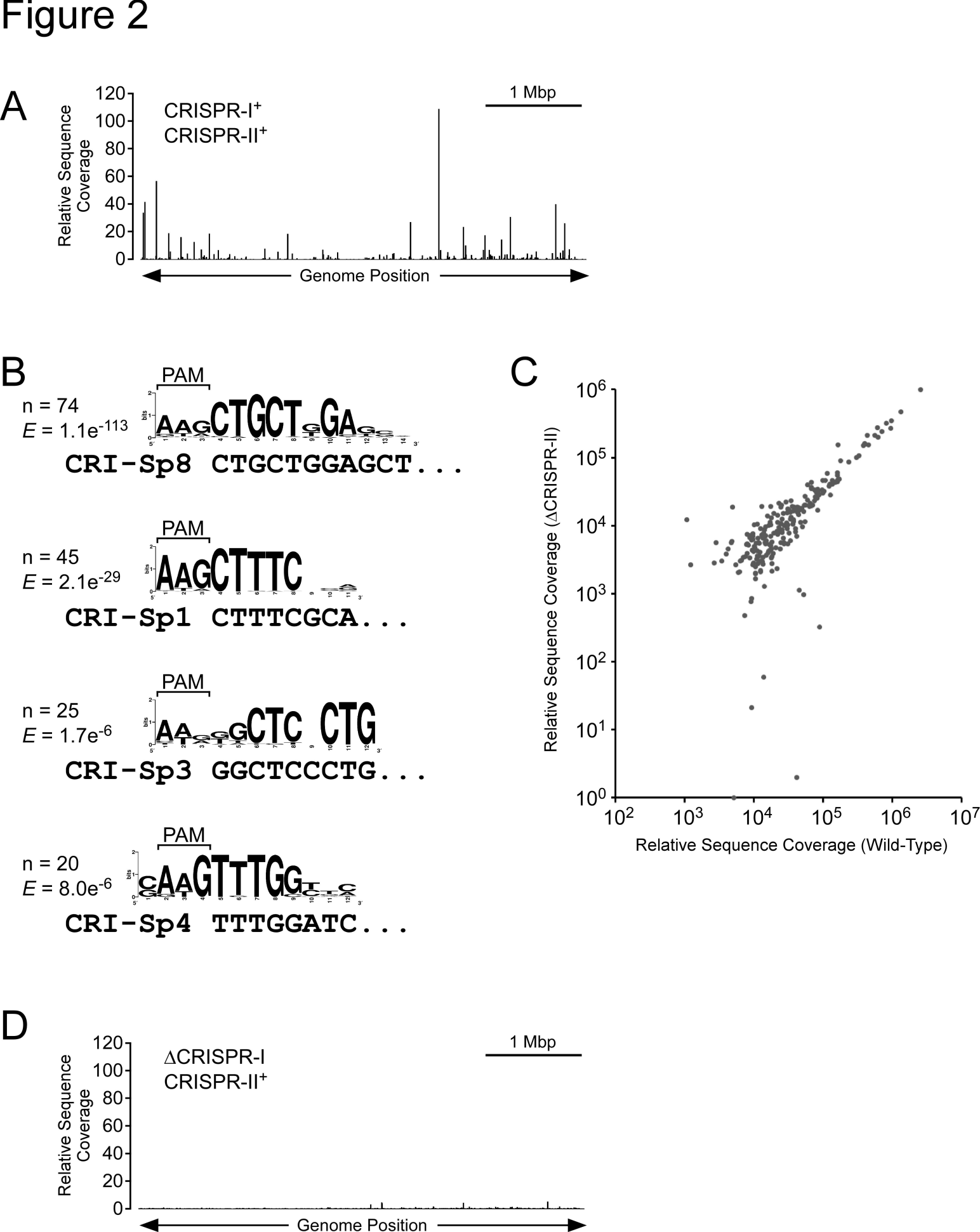
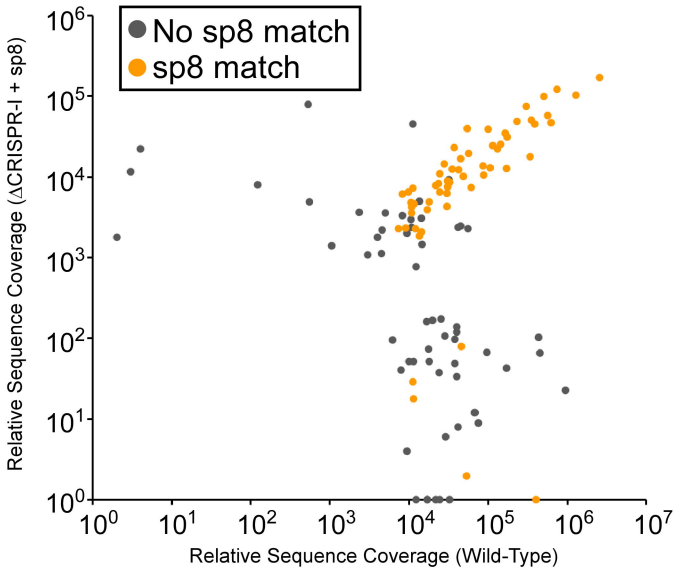
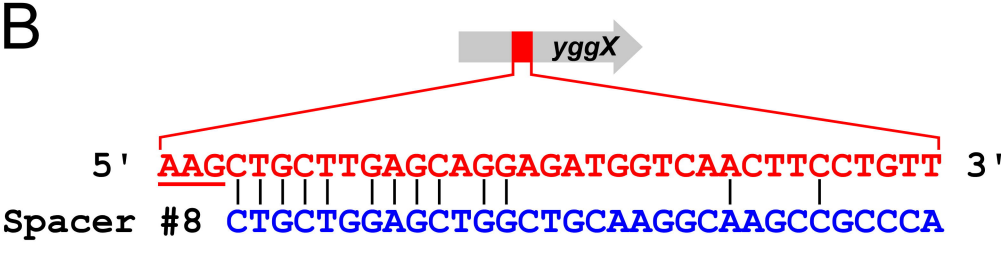

A
CRISPR-I+
CRISPR-II+
1 Mbp
Relative Sequence Coverage
120
100
80
60
40
20
0
Genome Position
B
PAM
n = 74
E = 1.1e-113
CRI-Sp8 CTGCTGGAGCT...
PAM
n = 45
E = 2.1e-29
CRI-Sp1 CTTTCGCA...
PAM
n = 25
E = 1.7e-6
CRI-Sp3 GGCTCCCTG...
PAM
n = 20
E = 8.0e-6
CRI-Sp4 TTTGGATC...
C
Relative Sequence Coverage (ΔCRISPR-II)
Relative Sequence Coverage (Wild-Type)
D
ΔCRISPR-I
CRISPR-II+
1 Mbp
Relative Sequence Coverage
Genome Position

Figure 3

A


No sp8 match
sp8 match
Relative Sequence Coverage (ΔCRISPR-I + sp8)
Relative Sequence Coverage (Wild-Type)
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7

B


yggX
5' AAGCTGCTTGAGCAGGAGATGGTCAACTTCCTGTT 3'
Spacer #8 CTGCTGGAGCTGGCTGCAAGGCAAGCCGCCCA

C

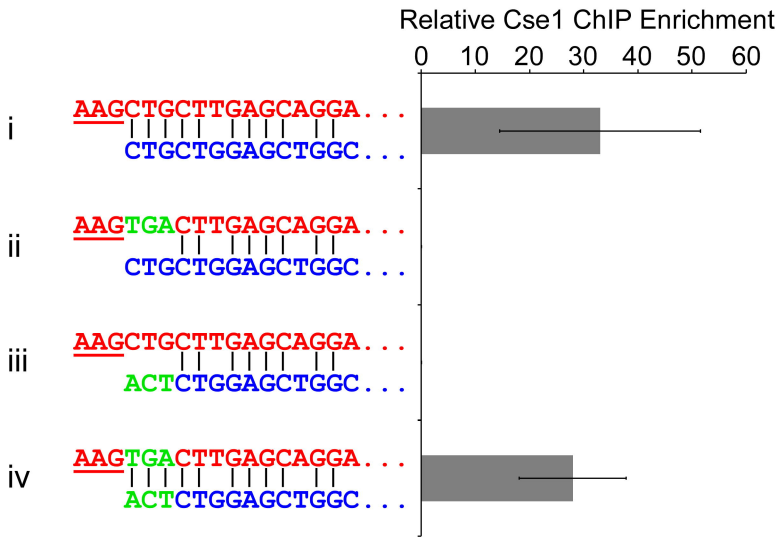
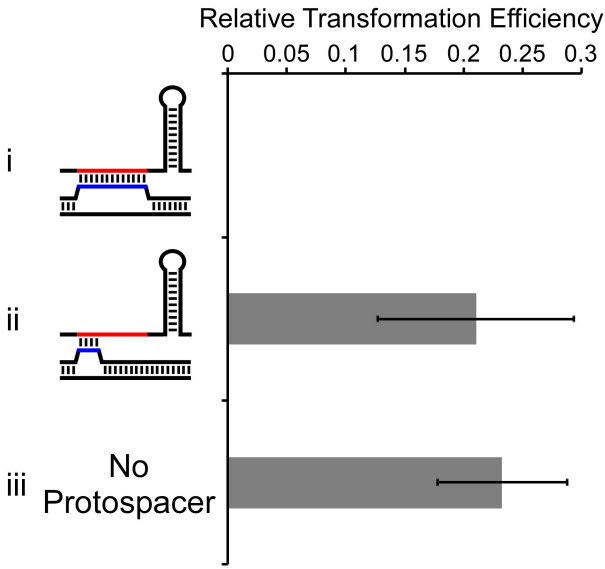

Relative Cse1 ChIP Enrichment
0 10 20 30 40 50 60
i
AAGCTGCTTGAGCAGGA...
CTGCTGGAGCTGGC...
ii
AAGTGACTTGAGCAGGA...
CTGCTGGAGCTGGC...
iii
AAGCTGCTTGAGCAGGA...
ACTCTGGAGCTGGC...
iv
AAGTGACTTGAGCAGGA...
ACTCTGGAGCTGGC...

Figure 4

A


Relative Transformation Efficiency
0
0.05
0.1
0.15
0.2
0.25
0.3
i
ii
iii
No
Protospacer

B

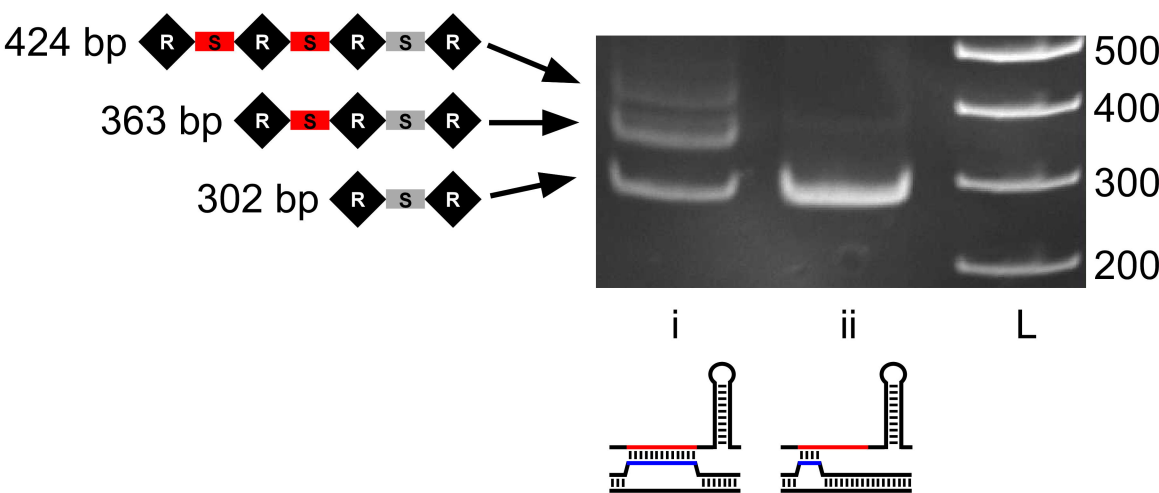
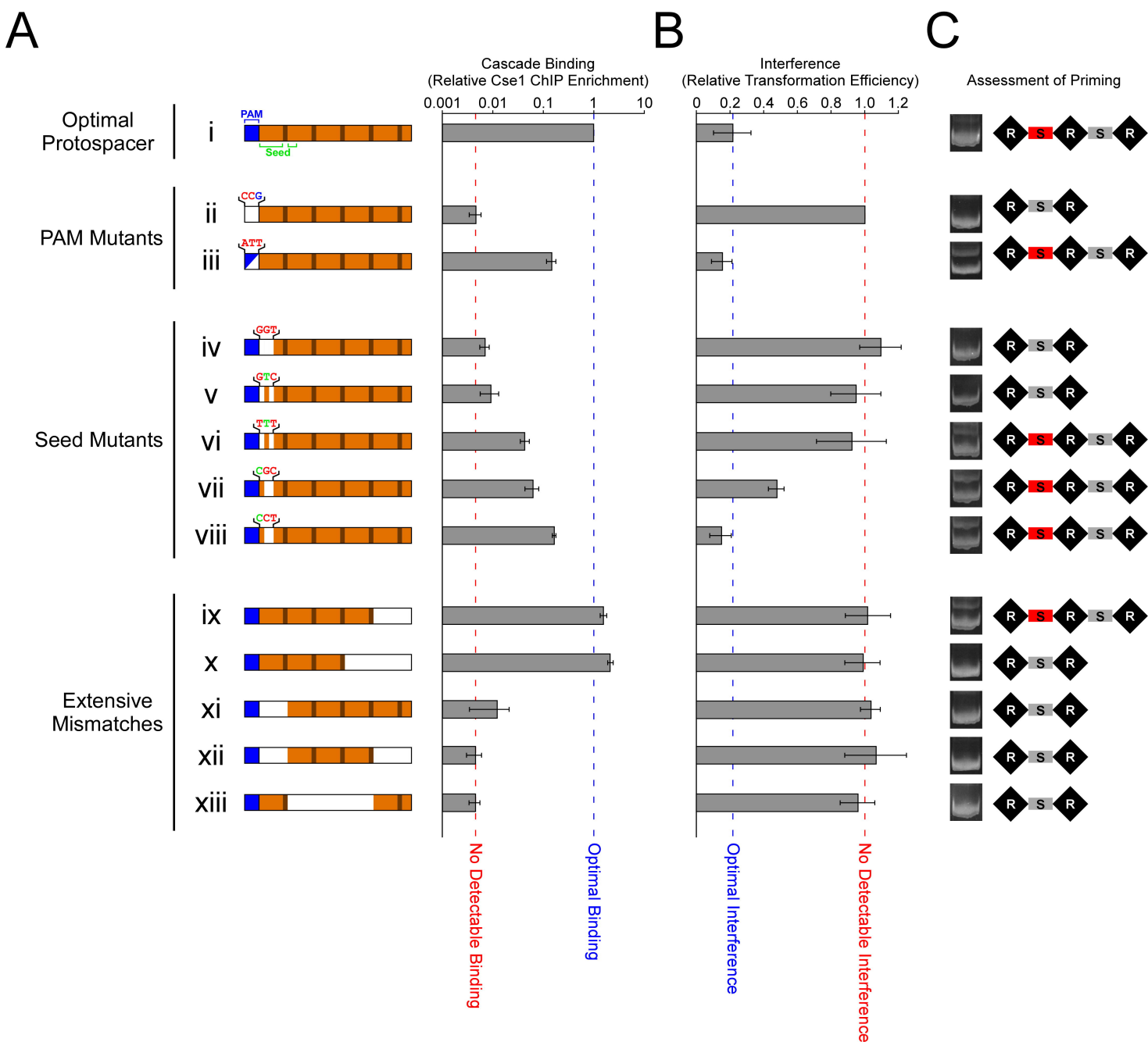

424 bp
363 bp
302 bp
R
S
500
400
300
200
i
ii
L

Figure 5


A
B
C
Cascade Binding
(Relative Cse1 ChIP Enrichment)
0.001
0.01
0.1
1
10
Interference
(Relative Transformation Efficiency)
0
0.2
0.4
0.6
0.8
1.0
1.2
Assessment of Priming
Optimal Protospacer
PAM Mutants
Seed Mutants
Extensive Mismatches
i
ii
iii
iv
v
vi
vii
viii
ix
x
xi
xii
xiii
PAM
Seed
CCG
ATT
GGT
GTC
TTT
CGC
CCT
R
S
No Detectable Binding
Optimal Binding
Optimal Interference
No Detectable Interference