Published in final edited form as:
Nat Microbiol. 2022 April ; 7(4): 530–541. doi:10.1038/s41564-022-01074-3.

Spacer prioritization in CRISPR-Cas9 immunity is enabled by the leader RNA

Chunyu Liao[1], **Sahil Sharma**[*,2], **Sarah L. Svensson**[*,2], **Anuja Kibe**[*,1], **Zasha Weinberg**[*,3], **Omer S. Alkhnbashi**[4], **Thorsten Bischler**[5], **Rolf Backofen**[4,6], **Neva Caliskan**[1,7], **Cynthia M. Sharma**[2], **Chase L. Beisel**[1,7,#]

[1]Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Centre for Infection Research (HZI), 97080 Würzburg, Germany

[2]Department of Molecular Infection Biology II, Institute of Molecular Infection Biology, University of Würzburg, 97080 Würzburg, Germany

[3]Bioinformatics Group, Department of Computer Science and Interdisciplinary Centre for Bioinformatics, Leipzig University, Härtelstraße 16-18, 04107 Leipzig, Germany

[4]Bioinformatics group, Department of Computer Science, University of Freiburg, 79085 Freiburg, Germany

[5]Core Unit Systems Medicine, University of Würzburg, 97080 Würzburg, Germany

[6]Signalling Research Centres BIOSS and CIBSS, University of Freiburg, 79085 Freiburg, Germany

[7]Medical Faculty, University of Würzburg, 97080 Würzburg, Germany

Abstract

CRISPR-Cas systems store fragments of foreign DNA called spacers as immunological recordings used to combat future infections. Of the many spacers stored in a CRISPR array, the newest spacers are known to be prioritized for immune defense. However, the underlying mechanism remains unclear. Here we show that the leader region upstream of CRISPR arrays in CRISPR-Cas9 systems enhances CRISPR RNA (crRNA) processing from the newest spacer, prioritizing

[#]Correspondence: Chase.Beisel@helmholtz-hiri.de (to C.L.B.).
[*]These authors contributed equally

AUTHOR CONTRIBUTIONS STATEMENT

C.L. and C.L.B. conceived this study. C.L. and C.L.B. designed the experiments. C.L. performed the plasmid cloning, *in vivo* assays in *E. coli* and *L. rhamnosus*, *in vitro* RNA transcription and purification, RNase III cleavage assays, and Northern blotting. C.L. and C.L.B. analyzed the associated data. S. Sharma conducted Western blotting and RNA immunoprecipitation for RIP-seq and helped analyze the data. S. Svensson conducted the RNA structural probing and RNase III cleavage site mapping and helped analyze the data. C.M.S. supervised the work performed by S. Sharma and S. Svensson. A.K. designed and performed the *in vitro* assay for RNA-RNA binding affinity and analyzed the data, with supervision by N.C. O.S.A. identified the repeat-leaders and computed mutations, with supervision by R.B. Z.W. assessed base-pairing probabilities. T.B. analyzed the data from RIP-seq. C.L.B. and C.L. wrote the manuscript, which was read and approved by all authors. C.L.B. supervised the project.

COMPETING INTERESTS STATEMENT

C.L.B. is a co-founder and member of the scientific advisory board for Locus Biosciences and is a member of the scientific advisory board for Benson Hill. C.L.B. and C.M.S. have submitted patent applications on CRISPR technologies unrelated to this work. The other authors declare no conflicts of interest.

defense against the matching invader. Using the CRISPR-Cas9 system from *Streptococcus pyogenes* as a model, we found that the transcribed leader interacts with the conserved repeats bordering the newest spacer. The resulting interaction promotes tracrRNA hybridization with the second repeat, accelerating crRNA processing. Accordingly, disrupting this structure reduces the abundance of the associated crRNA and immune defense against targeted plasmids and bacteriophages. Beyond the *S. pyogenes* system, bioinformatics analyses revealed that leader-repeat structures appear across CRISPR-Cas9 systems. CRISPR-Cas systems thus possess an RNA-based mechanism to prioritize defense against the most recently encountered invaders.

Adaptive immune systems possess the ability to remember prior invaders, allowing each system to specifically recognize and clear an invader if it appears in the future. As the only known adaptive immune systems in bacteria and archaea, CRISPR-Cas systems recognize and clear nucleic-acid sequences associated with invading plasmids and bacteriophages [1–3]. The immunological memory is stored as DNA spacers acquired from short segments of an invader's genomic material [4–6]. Stored spacers fall between conserved repeats in a CRISPR array, where new spacers are sequentially added at one end of an array [7–9]. To recall the stored memories for immune defense, the array is transcribed as a precursor and processed into individual CRISPR RNAs (crRNAs) comprising portions of a spacer and flanking repeat [10,11]. The mature crRNA then guides Cas effector nucleases to spacer-complementary nucleic-acid sequences, resulting in a nuclease cleaving the target or enacting widespread collateral cleavage of RNA or DNA that induce cell dormancy [12–14]. Because the spacer is derived from an invader, the immune system is programmed to clear this invader in case it attempts to reinfect the cell at another point in the future.

Within the large set of acquired spacers, CRISPR-Cas systems appear to prioritize defense through the most recently acquired spacers. RNA-seq analysis of native CRISPR-Cas systems has repeatedly revealed that the most abundant crRNAs derive from the newest end of the CRISPR array [15–18]. Separately, defense against a high phage titer was enhanced when moving an anti-phage spacer from the fifth position to the first position of the system's CRISPR array [19]. Spacer prioritization can be rationalized because increasingly large arrays can create competition within the available pool of crRNAs for the processing machinery and nuclease binding [20,21]. Spacer prioritization would also be particularly important by conferring protection against targeted invaders most likely to be encountered again by the cell, whether still present in the surrounding environment or as part of an active phage outbreak. What has remained elusive is the underlying mechanism. Here, we report a common mechanism for spacer prioritization within Type II CRISPR-Cas subtypes encoding the Cas9 nuclease that promotes preferential processing of the first crRNA.

A leader-repeat stem-loop interferes with ecrRNA generation

Our investigation of spacer prioritization began with the first repeat, the repeat immediately after the leader and copied as part of spacer acquisition, within CRISPR arrays of Type II CRISPR-Cas systems. Transcription of the CRISPR array as a precursor crRNA (pre-crRNA) leads to pairing between each repeat and the anti-repeat portion of a trans-activating crRNA (tracrRNA) [18,22]. The hybridized repeat:anti-repeat duplex is processed by the host endoribonuclease RNase III and bound by Cas9 (Fig. 1a) [18,23] . The upstream spacer then serves as the guide for DNA target recognition.

The first repeat presents a curiosity. On one hand, it normally matches any internal repeat in the array and thus should base-pair with the tracrRNA. On the other hand, the sequence upstream is the leader sequence rather than an acquired spacer, so the resulting “extraneous” crRNA (ecrRNA) would direct Cas9 with a sequence located outside of the array and thus not contribute to immune defense. For the CRISPR-Cas9 system from *Streptococcus pyogenes*, RNA-seq analysis did not indicate any stable products resembling an ecrRNA [18] . However, RNA-seq analysis of different lactobacilli revealed a stable “leader-derived” RNA [17] potentially representing an active ecrRNA. We therefore asked to what extent CRISPR-Cas9 systems form active ecrRNAs and whether any mechanisms exist to prevent their formation.

We focused on the CRISPR-Cas system from *S. pyogenes* because the tracrRNA was first identified in this bacterium, and the associated Cas9 is a mainstay of CRISPR technologies [24,25] (Extended Data Fig. 1a-b). To facilitate manipulation and testing, we transferred the system’s genetic locus into a low-copy plasmid propagated in *Escherichia coli,* paralleling its use in many bacterial applications [26–28] . DNA targeting through the ecrRNA or any of the crRNAs was measured by transforming a plasmid encoding the associated DNA target [24,29] . Transformed cells were either plated directly or subjected to non-selective outgrowth prior to plating (Fig. 1b). The outgrowth step grants more time before the antibiotics are administered, thereby allowing detection of plasmid clearance when none occurred under direct plating conditions [30–32] . The transformation assay revealed that the plasmid with the ecrRNA target was negligibly cleared with direct plating (1.6-fold) compared to a non-target control. In contrast, the same plasmid encoding the crRNA1 target (i.e. matching the first spacer S1) was efficiently cleared with direct plating (1,300-fold). The ecrRNA guide sequence was not the culprit, as replacing S1 with this sequence resulted in robust plasmid clearance with direct plating (30,000-fold) (Fig. 1b). The long 5′ end upstream of the ecrRNA was also not the culprit, as replacing R1 with the sgRNA handle to bypass crRNA processing exhibited enhanced plasmid clearance compared to the original ecrRNA (Fig. 1b). We instead posited that an active ecrRNA is poorly produced--albeit through an unknown mechanism.

While considering different mechanisms that might affect ecrRNA-mediated plasmid clearance, we noticed a stem-loop structure predicted to form between the first repeat (R1) and upstream leader (ldr) in the pre-crRNA (Fig. 1b and Extended Data Fig. 1c) supported by *in vitro* structure probing (Extended Data Fig. 1d-e). One potential consequence is that the stem-loop could block hybridization between the first repeat and the tracrRNA, thereby inhibiting ecrRNA biogenesis. *In vitro* binding measurements between the tracrRNA and an RNA spanning the leader through the first spacer (S1) confirmed that disrupting the stem-loop through leader mutations increased the binding affinity by at least 10-fold (Figs. 1c-d and Extended Data Figs. 1c and 2a). Another potential consequence is that the stem-loop could serve as a substrate for RNase III [33] , which normally processes repeat:tracrRNA duplexes. Accordingly, the same native leader-repeat RNA underwent cleavage by RNase III *in vitro*, while the leader mutations diminished RNA cleavage (Fig. 1E). The principal locations of RNase III cleavage overlapped with the site cleaved by RNase III within the standard repeat:tracrRNA duplex (Fig. 1e and Extended Data Fig. 1d-e) [18] . We concluded that the stem-loop formed between the pre-crRNA leader and first repeat can interfere with ecrRNA biogenesis by obstructing hybridization to the tracrRNA and driving tracrRNA-independent processing by RNase III. RNase III cleavage would also replicate standard processing of a repeat:tracrRNA duplex, allowing the first spacer to be separated from its upstream repeat and trimmed to a mature crRNA similar to all other spacers in the array [18] .

Stem-loop disruption impairs defense by the newest spacers

We next asked how disrupting the formed stem-loop affects plasmid interference directed by the ecrRNA as well as the six encoded crRNAs. Repeating the plasmid clearance assay in *E. coli* (Fig. 2a), we found that mutating the leader did not affect clearance by the ecrRNA with direct plating but did enhance clearance from 40-fold to 1,800-fold with outgrowth (comparing ecr and ecr(mut), Fig. 2a). This enhancement is in line with restored access by the tracrRNA rather than the specific mutations to the ecrRNA guide (Extended Data Fig. 3a). Mutating the leader also had a positional effect on crRNA-mediated plasmid clearance: clearance was heavily disabled through crRNA1 and crRNA2, partially disabled for crRNA3 and crRNA4, and unperturbed for crRNA5 and crRNA6. This result was unexpected, as the leader RNA had only been implicated in spacer acquisition or initiating transcription of the CRISPR array [19,34–36] . The impact of mutating the leader could not be obviously explained by perturbed Cas9 levels, altered transcription of the array, or a transcriptional start site internal to the array (Extended Data Figs. 1b and 3b-d). Instead, our results indicate that the stem-loop formed between the transcribed leader and the first repeat is critical for immune defense through the adjacent spacers.

To further evaluate the role of the leader-repeat stem-loop in immune defense, we replaced the first spacer with one of two spacers targeting the genome of the filamentous *E. coli* phage M13. We then evaluated defense against M13 infection based on plaque formation on a lawn of *E. coli* cells (Fig. 2b). In line with our plasmid clearance results, the M13-targeting arrays with a mutated leader as well as the native array lacking an M13-targeting spacer yielded viral plaques, while M13-targeting arrays with the native leader prevented plaque formation. Both M13-targeting spacers exhibited leader-dependent phage defense, indicating that the effect of the leader does not depend on the sequence of the first spacer. These results connect the leader region to anti-plasmid and antiviral defense by Cas9 and implicate the leader-repeat stem-loop in promoting defense through the newest CRISPR spacers. Given the lacking mechanisms to explain spacer prioritization for immune defense [9] , we turned our focus from the ecrRNA to the role the stem-loop plays in immune defense.

We asked if mutating the leader disrupts production of the crRNAs encoded near the beginning of the array. We therefore evaluated the abundance of Cas9-bound RNAs with the native or mutated leader by immunoprecipitating Cas9 and sequencing bound RNAs using RIP-seq (RNA immunoprecipitation and sequencing) (Fig. 3a and Extended Data Fig. 4) [30,37] . RIP-seq enriched the expected ecrRNA and the six crRNAs at least 33-fold compared to the untagged control for both the native and mutated leader, in line with binding by Cas9. Strikingly, crRNA1 was the most abundant Cas9-bound crRNA with the native leader, while its abundance dropped by 14-fold with the mutated leader. Mutating the leader also reduced the abundance of Cas9-bound crRNA2 but to a lesser degree (2.1-fold) and increased the abundance of Cas9-bound ecrRNA (2.2-fold). Similar trends in crRNA abundance were observed by northern blotting analysis using total RNA (Extended Data Fig. 5a-b). The loss of plasmid clearance through crRNA1 therefore can be attributed to the marked reduction in crRNA abundance due to disrupting the leader-repeat stem-loop.

The stem-loop and second repeat promote tracrRNA hybridization

The ensuing question is how the leader-repeat stem-loop accounts for enhanced crRNA production from the first spacer. One important insight came from our RIP-seq and northern blotting analyses (Extended Data Figs. 4 and 5a). They revealed a stable RNA product of ~190 nts spanning the leader to the RNase III processing site in the second repeat, which was also present when probing for crRNAs in the native *S. pyogenes* strain [18] . This RNA product disappeared when mutating the leader or removing Cas9, the tracrRNA, or RNase III in an *E. coli* strain harboring the native leader (Fig. 3b and Extended Data Fig. 5b). The leader-repeat stem-loop therefore was important for processing of the second repeat, the exact repeat associated with crRNA1. Processing appeared to occur through the second repeat before the first repeat, as the ~190-nt stable RNA product contained an intact first repeat and processed second repeat. Another insight came from our attempts to restore formation of the central leader-repeat stem. Reforming the central stem through additional mutations did not restore plasmid clearance through the newest spacers (Fig. 3c and Extended Data Fig. 5c-d). However, inverting the central stem by mutating the leader and then the first repeat disrupted and then restored position-dependent plasmid clearance (Fig. 3d). The key difference between these sets of mutations was that inverting the central stem maintained the rest of the stem-loop structure, suggesting that the upper portion of the stem-loop was also important for enhanced crRNA production.

The importance of the upper portion of the leader-repeat stem-loop for efficient processing of the second repeat could reflect a direct interaction between these physically separate parts of the pre-crRNA. If co-transcriptional folding forms the leader-repeat stem-loop before the second repeat is transcribed and before tracrRNA can hybridize with the first repeat, then the protruding loops of the stem-loop would be most readily available to interact with the second repeat. Following this logic, *in silico* folding predicted that the two main protruding loops of the leader-repeat stem-loop can extensively base pair with the second repeat (Fig. 4a). To test these predictions, we created compensatory mutations in the loops and the second repeat to disrupt and then reform this interaction while preserving the predicted secondary structure of the leader-repeat stem-loop (Fig. 4a). When mutating the second repeat, the tracrRNA anti-repeat was also mutated to maintain the repeat:anti-repeat duplex for processing and utilization by Cas9 [37] . Mutating the protruding loops disrupted plasmid clearance by crRNA1 130-fold under direct plating (Fig. 4b), although clearance was also high with outgrowth. Similarly, mutating the second repeat and the tracrRNA fully eliminated any measurable clearance, even with non-selective outgrowth (Fig. 4c). Importantly, reestablishing the predicted interactions by mutating the loops and the second repeat restored plasmid clearance partially with direct plating (10-fold) and fully with non-selective outgrowth (2,900-fold) (Fig. 4c). The leader-repeat stem-loop therefore appears to interact with the second repeat, which promotes immune defense through the newest spacer.

The interaction between the leader-repeat stem-loop and the second repeat raises the question: how could this interaction promote processing of the second repeat? If anything, the interaction would interfere with tracrRNA binding by sequestering at least a portion of the second repeat. However, we did notice that the repeat itself is predicted to form an imperfect stem-loop that could also interfere with tracrRNA hybridization (Fig. 4a). As the predicted interaction between the leader-repeat stem-loop and the second repeat and the predicted internal hairpin of the second repeat are mutually exclusive (Fig. 4a), the interaction could disrupt the internal hairpin and promote hybridization with the tracrRNA. To test the possible benefit of such an interaction, we performed *in vitro* binding measurements between the tracrRNA and a pre-crRNA spanning the leader through most of the second spacer (Fig. 4d and Extended Data Fig. 2b). The pre-crRNA was mutated within the leader-repeat stem to maintain its secondary structure and ensure that the tracrRNA hybridizes to the second repeat. Mutating the two protruding loops of the leader-repeat stem-loop reduced binding between the second repeat and tracrRNA by 4.2-fold. From these results, we conclude that the interaction between the leader-repeat and the second repeat promotes preferential hybridization of the tracrRNA to the second repeat, thereby prioritizing biogenesis of the crRNA derived from the newest spacer.

Leader-repeat stem-loops found across CRISPR-Cas9 systems

Given the role of the leader-repeat stem-loop in prioritizing immune defense for the *S. pyogenes* CRISPR-Cas9 system, we hypothesized that this mechanism would exist in many other CRISPR-Cas9 systems. The predicted interactions between the protruding loops of the leaderrepeat stem-loop and the second repeat for the *S. pyogenes* system are likely weaker, transient, and dependent on co-transcriptional folding [38] . However, the extensive stem formed between the leader RNA and first repeat offers a key feature that could be systematically predicted across CRISPR-Cas9 systems. We began with the II-A subtype of CRISPR-Cas9 systems that includes the system from *S. pyogenes*. Using publicly available genome sequences, we extracted 211 unique CRISPR array sequences from bacteria possessing only a II-A system and evaluated the predicted folding between the first repeat and the upstream 180 nts. We found numerous arrays with extensive predicted base pairing between the first repeat and its upstream sequence. Furthermore, by calculating the base-pairing potential between the inferred leader and repeat for each native or 1,000 scrambled sequences, we found that helix formation occurred significantly more than expected by chance across the II-A subtype ($p = 3\times10^{-6}$, Fisher's Method) (Fig. 5a). These findings support the broad prevalence of the leader-repeat stem-loop, at least for the II-A subtype.

Building on these predictions, we investigated two well-characterized II-A CRISPR-Cas9 systems from *Lactobacillus rhamnosus* GG and *Streptococcus thermophilus* DGCC 7710 as representative examples [12,16,17,39] (Extended Data Fig. 6). Both are predicted to form distinct stem-loops between the leader and first repeat, which was supported by *in vitro* structural probing (Extended Data Fig. 7). Furthermore, the stem-loop structures block tracrRNA binding and undergo tracrRNA-independent processing by RNase III (Extended

Data Fig. 8), paralleling our observations from the *S. pyogenes* system. As the CRISPR-Cas9 system from *L. rhamnosus* was previously found to form a leader-derived RNA based on RNA-seq analyses [17] , we evaluated the formation and targeting activity of the ecrRNA in this strain. RIP-seq analysis using plasmid-expressed tagged and untagged *L. rhamnosus* (Lrh)Cas9 in the native strain revealed minimal bound ecrRNA (Extended Data Fig. 9a-c), suggesting that the previously reported leader-derived RNA was not bound by LrhCas9. In contrast, crRNA1 was one of the most abundant bound crRNAs. Finally, the ecrRNA and crRNA1 respectively yielded negligible and complete clearance of the target plasmid (Extended Data Fig. 9d). These examples support the common role of the leader RNA for promoting immune defense through the newest spacer, at least for II-A CRISPR-Cas systems.

Beyond II-A CRISPR-Cas9 systems, the more abundant II-C subtype offers a counter example. On top of inserting new spacers through the last repeat [40,41] , this subtype encodes a promoter within each repeat that initiates transcription within the downstream spacer [41,42] . This configuration obviates the need for a promoter upstream of the array, which would make prioritization of the first (and therefore oldest) spacer counterproductive. Accordingly, 636 assessed II-C CRISPR arrays collectively did not exhibit helix formation between the first repeat and upstream region more than that expected by chance ($p = 0.50$, Fisher's Method) (Fig. 5a). However, we did observe examples of II-C arrays with extensive base pairing predicted between the first repeat and the upstream sequence (Fig. 5b and Extended Data Fig. 10). A stem-loop between the first repeat and upstream sequence therefore can be found in II-C systems, potentially reflecting alternative modes of spacer acquisition and transcription initiation within the subtype.

As a final exploration, we performed a similar analysis with two subtypes (I-E, I-F) within the abundant Type I CRISPR-Cas systems (Fig. 5a). These systems also acquire spacers through the first repeat and initiate transcription near the beginning of the leader. However, the Cas6 endonuclease rather than a tracrRNA/RNase III is responsible for repeat processing, and the first repeat contributes the 5′ end of crRNA1 required for effector complex formation [10,43] . Therefore, a leader-repeat stem-loop would also be counterproductive to crRNA1 production and immune defense through the newest spacer. Accordingly, both subtypes were not predicted to exhibit helix formation between the leader region and first repeat more than expected by chance ($p = 1.0$, Fisher's Method) (Fig. 5a). Other mechanisms thus may exist to prioritize the newest spacers for immune defense across CRISPR-Cas immune systems.

Discussion

Through this work, we discovered an RNA-based mechanism that allows some CRISPR-Cas systems to prioritize immune defense against the most recently encountered invaders. As part of the proposed mechanism (Fig. 6), the leader RNA base pairs with the first repeat to form a stem-loop through co-transcriptional folding. The upper portion of the stem-loop then interacts with the second repeat, temporarily preventing formation of a predicted hairpin internal to the repeat that interferes with tracrRNA hybridization. Either by providing a less-structured repeat or adopting a structure that promotes seeding of base pairing with the tracrRNA, the interaction allows the tracrRNA to more readily hybridize with the second repeat, leading to accelerated processing by RNase III and binding by Cas9. Because the crRNAs bound to Cas9 are shielded from RNase attack, the crRNAs appear much more abundant than other crRNAs in the array. After the second repeat undergoes processing, the leader-repeat stem-loop can undergo tracrRNA-independent processing by RNase III, although this step does not appear to be necessary for DNA targeting by Cas9. This proposed mechanism would be a particularly exquisite example of symmetry breaking in biology [44] , as it allows the preferential biogenesis of the crRNA adjacent to the leader despite the associated repeat harboring virtually the same sequence as every other repeat in the CRISPR array. We did find that the elucidated mechanism did not extend to most II-C systems as well as Type I systems, suggesting that other mechanisms underlying spacer prioritization await discovery. Elucidating these mechanisms will also create the opportunity to harness crRNA prioritization as part of multiplexing applications with CRISPR technologies [45] .

While our mutational analyses support the predicted interactions between the leader-repeat stem-loop and the second repeat, a more complex structure likely exists and should be the subject of future work. The structure would be expected to depend on the dynamics of transcriptional co-folding in the cellular cytoplasm, where observing such dynamic structures would be less amenable to approaches such as crystallography or cryo-EM. Instead, approaches such as time-resolved microscopy using integrated fluorescent probes, single-molecule studies with optical tweezers, or in-cell SHAPE-seq could help resolve dynamic structures [46,47] . Regardless of the exact structure though, the interactions between the leader-repeat stem-loop and the second repeat have multiple implications for spacer prioritization. One implication is that the leader-repeat stem-loop could also interact with repeats downstream of the second repeat--particularly after the tracrRNA hybridizes to this repeat. These longer-range interactions possibly help explain why the downstream crRNAs are also negatively impacted by disrupting the leader-repeat stem-loop. Another implication is that the interaction could prevent the newest spacer from base pairing with the second repeat, thereby removing potential secondary-structure issues that could render a less effective spacer more effective while it exists at the beginning of the array. A third implication is that base pairing between any spacer and an adjacent repeat could prevent the repeat from forming an internal stem-loop, thereby promoting tracrRNA hybridization. We posit that this mechanism could help explain why some internal spacers give rise to highly abundant crRNAs.

The discovery of spacer prioritization began by exploring the fate of the first repeat in CRISPR-Cas9 arrays and its potential to yield an ecrRNA. We showed that the leader-repeat stem-loop actively reduced ecrRNA formation for three different CRISPR-Cas9 systems. The primary role of the leader-repeat stem-loop appears to be spacer prioritization, where the central stem ensures presentation of the loops for pseudoknot formation. However, it is intriguing that the central stem also blocks ecrRNA formation. Beyond CRISPR-Cas9 systems, many Type V-A CRISPR-Cas systems were shown to block ecrRNA formation [48] . For V-A systems, the last repeat would give rise to an ecrRNA, although many of these systems contain disruptive mutations in the last repeat that prevents ecrRNA processing. CRISPR-Cas9 systems of the II-A subtype are distinct because the putative ecrRNA derives from the first repeat. Because new spacers are acquired through this repeat, mutations that would disrupt ecrRNA formation would also disrupt defense by any acquired spacers. Therefore, the stem-loop offers a simple mechanism to prevent ecrRNA formation while still ensuring the function of any acquired spacers. Future work can elucidate the fate of ecrRNAs across CRISPR-Cas systems and whether they provide a hindrance to immune defense or confer potential benefits to cells through alternative functions [49].

Methods

Strains, plasmids and growth conditions

Supplementary Table 3 provides a list of the key resources used in this work. Supplementary Table 4 lists all strains, plasmids, oligonucleotides, and gBlocks.

E. coli cells were grown at 37°C in Luria Bertani (LB) broth (5 g/L NaCl, 5 g/L yeast extract, 10 g/L tryptone) with shaking at 250 rpm or on LB agar plates (LB broth, 18 g/L agar). The antibiotics ampicillin and/or kanamycin were added at 50 μg/mL to maintain any plasmids. *L. plantarum* and *L. rhamnosus* were grown at 37°C in De Man, Rogosa and Sharpe (MRS) broth (Becton Dickinson) without agitation or on MRS agar (Becton Dickinson). The antibiotics chloramphenicol and erythromycin were added at 10 μg/mL as necessary to maintain any plasmids.

The plasmid pCBS2225 expressing the tracrRNA, SpyCas9, and associated native CRISPR array was constructed by inserting the corresponding cassette amplified from the genomic DNA of *S. pyogenes* SF370 using Gibson assembly (New England Biolabs) into the backbone plasmid pCB902 following the manufacturer's instructions. The mutations in the leader and/or first repeat were introduced through Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions (pCBS2226). The first repeat was replaced by a sgRNA scaffold using Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions (pCBS2247). For Western blotting and RIP-seq analyses, the 3xFLAG tag was inserted downstream of the stop codon of the gene encoding SpyCas9 through Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions. The plasmid encoding the FLAG-tagged LrhCas9 was constructed by first PCR-amplifying the gene encoding LrhCas9 along with the upstream 437 bp containing the putative promoter from genomic DNA extracted from *L. rhamnosus* GG. The reverse primer included the FLAG tag. The resulting PCR product was inserted into the backbone plasmid pCB591 by Gibson Assembly (New England Biolabs) following the manufacturer's instructions. The targeted plasmids used in the plasmid clearance assay in *E. coli* and *L. rhamnosus* were constructed by performing Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions on plasmid pCB858 and pCB591 to insert the protospacer and PAM. *E. coli* TOP10 was used for the construction of plasmids used in *E. coli*. *L. plantarum* WCFS1 was used as the cloning strain for plasmids that can be propagated in *L. rhamnosus* but not in *E. coli*.

The plasmids used for interrogating if mutating leader affects transcription of array (pCBS2243 and pCBS2244) were constructed by first PCR-amplifying the fragments encoding the native promoter-native/mutated leader from plasmid pCBS2225 or pCBS2226.

The resulting PCR product was inserted into the backbone plasmid pCBS2242 by replacing the PJ23119 promoter using Gibson Assembly (New England Biolabs) following the manufacturer's instructions.

The plasmids used for M13 phage assay (pCBS2253, pCBS2254, pCBS2255, and pCBS2256) were constructed by replacing the first spacer on plasmid pCBS2225 or pCBS2226 with the corresponding spacers targeting gene VIII in genome of M13 phage through Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions.

The plasmids encoding single-spacer arrays for the ecrRNA and mutated ecrRNA (pCBS2245 and pCBS2246) were constructed by replacing the CRISPR array in plasmid pCBS2225 with a PCR amplicon encoding the corresponding repeat-spacer-repeat through Gibson assembly (New England Biolabs) following the manufacturer's instructions.

The plasmids with the stem-loop disrupted and restored by copying and flipping the sequences in the stem (pCBS2249 and pCBS2250) were constructed through Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions. The stem-loop is disrupted by replacing the portion of leader base-paring with the first repeat with the corresponding sequence of the first repeat using plasmid pCBS2225 as template for PCR. The resulting plasmid was used as a PCR template to restore the stem-loop by replacing the portion in the first repeat with the corresponding sequence of the leader.

Plasmids with mutated loops and/or the second repeat were constructed through Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions. The resulting plasmids were used as templates for PCR and Q5 mutagenesis for mutating the corresponding region on the tracrRNA encoded on the same plasmid.

Plasmid extraction and transformation of Lactobacilli

Plasmids constructed in *E. coli* TOP10 and used in *L. rhamnosus* were propagated in EC135 first before transfering into *L. plantarum* WCFS1. Plasmids used for transformation into *L. rhamnosus* were extracted from *L. plantarum* WCFS1. Cells were cultured in liquid MRS medium, pelleted by centrifugation, washed twice with water, and resuspended in 25 mg/mL lysozyme (Carl Roth) in lysozyme buffer (10mM Tris-HCl, pH 8.0 1mM EDTA, pH 8.0, 0.1M NaCl, 5% Triton X-100). After incubating at 37°C with shaking at 250 rpm for 40 min, the cells were pelleted by centrifugation and washed once with water. The washed cells were then used for plasmid extraction following the instructions for the ZymoPURE II™ Plasmid Midiprep Kit.

Electrocompetent cells were prepared and transformed for *L. plantarum* as described previously with modifications [50]. Briefly, *L. plantarum* cells grown to an ABS_{600} of ~0.8 in MRS broth with 2% glycine were collected by centrifugation and washed with 10mM $MgCl_2$ and 10% glycerol and resuspended in 10% glycerol for transformation. A total of 60 μL of competent cells and at least 2.5 μg of DNA were added to a 1-mm gap cuvette and electroporated at 1.8 kV, 200 Ω resistance, and 25 μF capacitance. Electrocompetent cells for *L. rhamnosus* were prepared using the same method as *L. plantarum*, only ampicillin was added into the culture to a final concentration of 10 μg/mL when the ABS_{600} reached ~0.2, then the cells were pelleted by centrifuging at 4°C at 5,000 x g for 15 min when the ABS_{600} reached ~0.4 and washed once using 10-mM ice-cold $MgCl_2$ solution and twice using ice cold 10% glycerol. Transformation for *L. rhamnosus* was performed by adding 100 μL of electrocompetent cells and at least 5 μg of plasmid (no more than 5 μL) to a 2-mm gap cuvette and electroporating at 2.5 kV, 200 Ω resistance, and 25 μF capacitance. Following electroporation, cells were recovered in 1 mL of MRS broth at 37°C without agitation for 3 h, plated on MRS agar plates with or without antibiotics, and incubated at 37°C for 48 h in an anaerobic chamber (80% N_2, 10% CO_2, and 10% H_2).

Transcription start site mapping

Total RNA was extracted from *L. rhamnosus* or from *E. coli* harboring the CRISPR cassette plasmid with the native leader (pCBS2225) as described above. Extracted RNA was then treated with Turbo DNase (Life Technologies) and cleaned using the RNA Clean & Concentrator (Zymo Research) following the manufacturer's instructions. The resulting RNA was treated with 5′ terminator exonuclease (TEX) (Epicentre) to degrade processed RNAs following the manufacturer's instructions, purified using the RNA Clean & Concentrator kit, and subjected to 5′ RACE using the Template Switching RT Enzyme Mix (New England Biolabs) following the manufacturer's instructions. PCR was performed using the Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs) following the manufacturer's instructions. The resulting PCR products were quality checked by electrophoresis on an agarose gel, purified using Zymo DNA Clean & Concentrator (Zymo Research) following the manufacturer's instructions, and inserted into the supplied linearized vector pMiniT 2.0 using the NEB PCR Cloning Kit (New England Biolabs) following the manufacturer's instructions. Transformed plasmids were then extracted from ten randomly-selected colonies using NucleoSpin Plasmid EasyPure (Macherey-Nagel) following the manufacturer's instructions and submitted for Sanger sequencing.

Plasmid clearance assays

Plasmid clearance assays in *E. coli* BW25113 were conducted as described previously [48] . Briefly, 50 ng of plasmid encoding the PAM-flanked target was electroporated into *E. coli* cells harboring the plasmid encoding the tracrRNA, SpyCas9, and the array with the native or mutated leader. After recovering for 1 h in SOC (20 g/L tryptone, 5 g/L yeast extract, 3.6 g/L glucose, 0.5 g/L NaCl, 0.186 g/L KCl, 0.952 g/L $MgCl_2$, PH 7.0) at 37°C with shaking at 250 rpm, cells were serially diluted and 5-μL droplets were plated on LB agar plates with ampicillin and kanamycin. Colony numbers were recorded for analysis after 16 h of growth. To increase the sensitivity of the plasmid clearance assay, 3 μL of the recovered culture was added to 3 mL of LB broth with kanamycin and cultured at 37°C with shaking at 250 rpm for 16 h. Cells were then serially diluted, and 5-μL droplets were plated on LB agar plates with ampicillin and kanamycin. Colony numbers were recorded for analysis after ~16 h of growth. All experiments represent three independent replicates starting from separate colonies.

For plasmid clearance assays in *L. rhamnosus*, 5 μg of plasmids encoding the PAM-flanked target was electroporated into *L. rhamnosus*. After recovering for 3 h in 1 mL of MRS at 37°C without agitation, cells were diluted and plated on MRS agar plates with chloramphenicol. Colony numbers were recorded for analysis after 60 h of growth in an anaerobic chamber.

RNA folding predictions

Equilibrium folding of the leader-repeat RNAs and repeat:tracrRNA duplexes were predicted using the online NUPACK algorithm [55,56] (http://www.nupack.org/partition/new). Default parameters were used as well as the following for folding the individual RNAs: Nucleic acid type: RNA, Temperature: 37°C. In the case of predicting pairing between the repeat and tracrRNA, a concentration of 1 μM was specified for each RNA. NUPACK considers both intermolecular and intramolecular base pairing. Interactions between the two protruding loops of the stem-loop and the second repeat were predicted using the online RNAfold algorithm [51,52] by fusing the two loops and flanking two nucleotides with the repeat. As part of the predictions, between one and four nucleotides were added between each of the loops and the repeat, and the algorithm was instructed to leave these nucleotides unpaired.

Western blotting analysis

As a quality control of the coIP, a volume of cell culture equivalent to an ABS_{600} of 1.0 were collected during different stages of the coIP (Lysate, Supernatant 1, Supernatant 2, Wash and coIP Eluate), boiled in protein loading buffer (62.5 mM Tris-HCl, pH 6.8, 100 mM DTT, 10% glycerol, 2% SDS, 0.01% bromophenol blue) at 95°C for 8 min, and stored at – 20°C for Western blot analysis. Overnight culture of CB414 *E. coli* cells harboring plasmid pCBS2225, pCBS2226, pCBS2240, or pCBS2241 were back-diluted to an ABS_{600} of ~0.05 in LB medium with kanamycin and shaken at 250 rpm at 37°C until the ABS_{600} reaches ~0.8. Pelleted cells equivalent to 1.44 ABS_{600} were resuspended in 144 μL of protein loading buffer, boiled at 95°C for 8 min, and stored at -20°C for Western blot analysis. Western blot analyses were conducted as described previously [30] . Briefly, the resulting samples corresponding to about 0.2 ABS_{600} of cells were resolved on an 10% SDS-polyacrylamide gel, transferred to Nitrocellulose 0.45 μM NC membrane (Amersham Protan), blotted using semi-dry blotter (VWR), washed using Tris-buffered saline (20 mM Tris, 150 mM NaCl) with 0.1 % Tween 20, and visualized on ImageQuant LAS 4000 (GE healthcare). Monoclonal ANTI-FLAG M2 (Sigma) antibody, anti-GroEL (Sigma) primary antibody, horseradish peroxidase-coupled anti-mouse IgG secondary antibody (Thermo Fisher), and anti-rabbit IgG secondary antibody (GE-Healthcare) were used for detection.

In vitro transcription and purification of RNA

gBlocks encoding the T7 promoter and desired RNA were ordered from IDT Technologies for PCR amplification. For RNAs spanning the leader through most of the second spacer, DNA templates for T7 transcription were amplified from the corresponding plasmid using a forward primer with the T7 promoter appended to the 5′ end. Amplicons were purified and concentrated using DNA Clean & Concentrator (Zymo Research). RNA was transcribed using the HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs) and treated with Turbo DNase (Life Technologies) according to the manufacturer’s instructions. The RNA was resolved on an 8% polyacrylamide gel (20 × 20 cm) containing 7 M urea at 300 V for 240 min, stained with SYBR Green II (Biozym), excised, and extracted using ZR small-RNA PAGE Recovery kit (Zymo Research) according to the manufacturer’s instructions.

The extracted RNAs eluted in nuclease-free water were quality checked by electrophoresis on a PAA-urea gel and stored in -80°C.

In vitro assay for RNA-RNA binding affinity

Binding affinities of the RNA transcripts and the respective tracrRNAs were measured by Microscale thermophoresis (MST). TracrRNAs 3′-labeled with a Cy5 fluorophore were ordered from IDT Technologies. The leader-repeat-spacer transcripts were *in vitro*-transcribed and purified as described above. After boiling at 90°C for 2 min and cooling down to room temperature by sitting on a bench for 10 min, RNAs were serially diluted 2-fold for 16 rounds in MST buffer (50 mM Tris-HCl, 150 mM NaCl, 10 mM $MgCl_2$ and 0.05% (v/v) Tween-20, pH 7.8), each mixed with one volume of 10 nM Cy5-labeled tracrRNA, and incubated at 37°C for 10 min. The 16 samples were then loaded into Monolith NT.115 Premium capillaries (NanoTemper Technologies) and measured using a Monolith NT.115Pico instrument (NanoTemper Technologies) at an ambient temperature of 25°C with 5% LED power and medium MST power. Binding affinity data of three independently pipetted measurements was analyzed (MO.Affinity Analysis software version 2.3, NanoTemper Technologies) using the signal from an MST-on time of 20 s for Sth1Cas9-related RNA, 5 s for the LrhCas9-related and SpyCas9-related RNAs for testing the first repeat, and 1.5 s for SpyCas9 related RNA for testing the second repeat.

In vitro RNase III cleavage assay

In vitro-transcribed and purified RNAs were boiled in a thermocycler at 95°C for 10 min, cooled down to room temperature by sitting on a bench for 10 min, and kept on ice. Cleavage reactions were prepared by adding 40 ng of RNA; 1, 0.2, 0.04, 0.008, or 0 units of RNase III (Invitrogen); and water in the supplemented buffer to a total volume of 10 μL. After incubation for 5 min at 37°C, the reaction was stopped by adding an RNA loading buffer (0.025% bromophenol blue, 0.025% SDS, 0.025% xylene cyanol, 18 mM EDTA (pH 8.0), 93.64% formamide) on ice. The mixture then was boiled in a thermocycler at 95°C for 10 min, resolved on an 8% polyacrylamide gel (20 × 20 cm) containing 7 M urea at 300 V for 210 min, stained with SYBR Green II (Biozym), and visualized on a Phosphorimager (Typhoon FLA 7000, GE Healthcare). The Low Range ssRNA Ladder (New England Biolabs) was used as a marker. For the assays with the leader-repeat-spacer RNA for LrhCas9, the RNA was truncated within the leader and the spacer to avoid cleavage of irrelevant secondary structures formed internally within either domain.

Northern blotting analysis

Overnight culture of CB414 or CL536 (RNase III-deficient) *E. coli* cells harboring plasmid pCBS2225, pCBS2226, pCBS3416, or pCBS3417 were back-diluted to an ABS_{600} of ~0.05 in LB medium with kanamycin and shaken at 250 rpm at 37°C until the ABS_{600} reaches ~0.8. Total RNAs were extracted from 4 ABS_{600} of pelleted cells using the hot-acid phenol chloroform as described previously [53]. Northern blotting analysis was carried out as described previously [48]. Oligodeoxyribonucleotides used for end labeling by γ-32P-ATP and probing can be found in Supplementary Table 4.

RNA structural probing and RNase III cleavage site mapping

I*n vitro*-transcribed and purified RNAs were dephosphorylated with Antarctic Phosphatase (New England Biolabs), 5′-end-labelled with γ32P) using T4 polynucleotide kinase (Thermo Fisher Scientific), and purified by gel extraction as previously described [54] . Sequences of the resulting T7 transcripts are listed in Supplementary Table 5. Inline probing assays for RNA secondary structure were performed as described previously with minor modifications [55] . End-labeled RNAs (0.2 pmol) in 5 μL of water were mixed with an equal volume of 2× Inline buffer (100 mM Tris-HCl, pH 8.3, 40 mM $MgCl_2$, and 200 mM KCl) and incubated for 40 h at room temperature to allow spontaneous cleavage. Reactions were stopped with an equal volume of 2× Colorless loading buffer (10 M urea and 1.5 mM EDTA, pH 8.0). For RNase III cleavage assays, the same 5′ end-labeled *in vitro* transcripts were briefly denatured and snap-cooled on ice, followed by the addition of RNase III buffer to a final concentration of 1× and yeast tRNA (Ambion) to a final concentration of 0.1 mg/ml. RNA samples were then incubated at 37°C for 10 min followed by the addition of 0, 0.0016, 0.008, 0.04, 0.2, or 1 U RNase III (Invitrogen) and further incubated at 37°C for 5 min. Reactions were stopped by adding an equal volume of Gel-loading buffer II (95% (vol/vol) formamide, 18 mM EDTA, and 0.025% (wt/vol) SDS, 0.025% xylene cyanol, and 0.025% bromophenol blue). Inline probing and RNase III cleavage reactions were then separated on a 6-10% PAA-urea sequencing gel, which were dried and exposed to a PhosphorImager screen. RNA ladders were prepared using alkaline hydrolysis buffer (OH ladder) or Sequencing buffer (T1 ladder) (Ambion) according to the manufacturer's instructions.

Flow cytometry analysis

Overnight cultures of CB414 cells harboring plasmid encoding GFP gene driven by the promoter of endogenous CRISPR array of SpyCas9 followed by the native leader (pCBS2243), mutated leader (pCBS2244) or empty vector (pCB908) were back-diluted to an ABS_{600} of ~0.05 in LB medium supplemented with kanamycin and shaken at 250 rpm at 37°C until reaching an ABS_{600} of ~0.8. The GFP fluorescence of single cells was then measured as described previously [48] . Briefly, cultures were diluted 1:100 in 1x phosphate-buffered saline (PBS) and analyzed on an Accuri C6 Plus flow cytometer with BD CSampler Plus (Becton Dickinson), a 488-nm laser, and a 530/30-nm bandpass filter. Forward scatter (cut-off of 11,500) and side scatter (cut-off of 600) were used to eliminate non-cellular events. The mean FITC-A value of 30,000 events within a gate set for live *E. coli* cells were used for data analysis after subtracting autofluorescence of the cells.

Phage sensitivity assay

Overnight cultures of NEB Turbo cells harboring the CRISPR cassette plasmid with the native leader (pCBS2225) or mutated leader (pCBS2226) were back-diluted to an ABS_{600} of ~0.05 in LB medium supplemented with kanamycin and shaken at 250 rpm at 37°C until the ABS_{600} reached ~0.5. The cells were then collected by centrifugation and resuspended in 1/10 volume of LB with kanamycin. Petri dishes (Ø 90 × 16.2 mm) with 24 mL of LB agar supplemented with kanamycin were overlaid with 4 mL of soft LB agar (7.5 g/L) with kanamycin containing 0.75 mL of the cell suspension. After solidifying for 10 min, 3 μL of

10-fold serial dilutions of phage lysates were spotted onto the surface of the soft agar. Plates were dried at room temperature under a flame until no liquid was visible on the surface of the agar and incubated at 37°C for 15 h. Plagues were visualized using an ImageQuant LAS 4000 imaging system (GE Healthcare).

RNA immunoprecipitation for sequencing

Cas9-3xFLAG co-immunoprecipitation (coIP) combined with RNA-seq (RIP-seq) was performed in *E. coli* and *L. rhamnosus* as described previously with minor modifications [30] . Briefly, overnight cultures of CB414 harboring the plasmid pCBS2225, pCBS2226, pCBS2240, or pCBS2241 were back-diluted to an ABS600 of ~0.05 in LB medium with kanamycin and shaken at 250 rpm at 37°C until the ABS_{600} reaches ~0.8. Overnight cultures of *L. rhamnosus* with or without the plasmid encoding 3xFLAG-tagged LrhCas9 pCBS2227 were back-diluted to an ABS_{600} of ~0.05 in MRS medium with or without chloramphenicol, incubated at 37°C without agitation until the ABS_{600} reaches ~0.5. The equivalent of 37-40 ABS_{600} of cells were washed using Buffer A (20 mM Tris-HCl, pH 8.0, 150 mM KCl, 1 mM $MgCl_2$, 1 mM DTT) and subsequently pelleted at 4°C for 3 min at 11,000 x g.

The pellets were snap-frozen in liquid nitrogen and stored at −80°C until further use. Frozen pellets were thawed on ice and resuspended in 1.5 mL of lysis buffer (957 μL buffer A, 1 μL 1 mM DTT, 10 μL 0.1 M PMSF, 2 μL triton X-100, 20 μL DNase I, 10 μL Superase-In RNase Inhibitor) and distributed onto two pre-cooled fast-prep tubes for lysis (750 μL each). Quick lysis was performed with FastPrep homogenizer twice (6.5 M/s; 1 min), and the resulting lysate from both tubes was centrifuged at 4°C for 10 min at 13,000 rpm. Following centrifugation, the supernatant (i.e. the lysate fraction) from both tubes was combined and transferred to a new tube. The lysate was incubated with 35 μL of anti-FLAG antibody (Monoclonal ANTI-FLAG M2, Sigma) for 90 min at 4°C on a rocker (supernatant 1). Next, 75 μL of Protein A-Sepharose (Sigma) prewashed with Buffer A was added and the mixture was rocked for another 90 min at 4°C (supernatant 2). After centrifugation, the supernatant was removed and the pelleted beads were washed five times with 0.5 mL of Buffer A (wash). Finally, 500 μL of Buffer A was added to the beads. RNA and proteins were separated using phenol-chloroform-isoamyl alcohol (P:C:I). For each coIP, RNA was recovered from the aqueous phase, precipitated overnight using a 30:1 mix of ethanol and 3M sodium acetate at -20°C and eluted after centrifugation in 30 μL of RNase-free water. The resulting RNA was treated by DNase I. For protein samples in the organic phase, 1.4 mL ice-cold acetone was added and incubated overnight at -20°C. Samples were centrifuged at 15,000 rpm for 1 h to precipitate the protein and washed twice with 1 mL acetone without disturbing the pellet. A total of 100 μL of 1x protein loading buffer (62.5 mM Tris-HCl, pH 6.8, 100 mM DTT, 10% (v/v) glycerol, 2% (w/v) SDS, 0.01% (w/v) bromophenol blue) was then added to the pellet to obtain the final protein sample (eluate). To determine whether the coIP was successful, protein samples equivalent to 1.0 ABS_{600} of cells were collected during different stages of the coIP (lysate, supernatant 1, supernatant 2, wash and coIP eluate). A total of 100 μL of 1x protein loading buffer was added to each of the collected protein samples and boiled for 8 min. Protein samples corresponding to an ABS_{600} of 0.2 (lysate, supernatant 1, supernatant 2, and wash fraction) and 10 (for eluate fraction) were used for Western blotting analysis.

cDNA library preparation and deep sequencing

The extracted RNA was treated with DNase I (Thermo Scientific, EN0525) following the manufacturer's instructions. cDNA libraries for Illumina sequencing were constructed by Vertis Biotechnologie AG, Germany (http://www.vertis-biotech.com). Briefly, the resulting RNA was subjected to oligonucleotide adapter ligation on the 3′ end, first-strand cDNA synthesis using M-MLV reverse transcriptase (Agilent), and Illumina TruSeq sequencing adapter ligation on the 3′ end of the antisense cDNA. The resulting cDNA was PCR-amplified using Herculase II Fusion DNA Polymerase (Agilent) with 13 amplification cycles following the manufacturer's instructions, purified using Agencourt AMPure XP kit (Beckman Coulter Genomics) following the manufacturer's instructions, and analyzed by capillary Electrophoresis. The resulting samples were then run on an Illumina NextSeq 500 instrument with 76 cycles in single-read mode. Sequences of the oligonucleotide adapter, the 5′ Illumina TruSeq sequencing adapter, and the oligonucleotides used for PCR can be found in Supplementary Table 4.

Bioinformatics analysis of RIP-seq

Illumina reads were quality and adapter trimmed with Cutadapt [56] version 2.5 using a cutoff Phred score of 20 in NextSeq mode and reads without any remaining bases were discarded (command line parameters: --nextseq-trim=20 -m 1 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC). Afterwards, we applied the pipeline READemption [57] version 0.4.5 to align all reads longer than 11 nt (-l 12) to the respective reference sequences using segemehl [58] version 0.2.0 with an accuracy cut-off of 95% (-a 95). For *E. coli* K-12 BW25113, we applied RefSeq assembly GCF_000750555.1 with plasmid pCBS2225 (NL-Tagged-SpCas9-Plasmid) for libraries with native leader and plasmid pCBS2226 (ML-Tagged-SpCas9-Plasmid) for libraries with mutated leader. For *L. rhamnosus* GG libraries, we utilized RefSeq assembly GCF_000026505.1 together with the sequence of plasmid pCBS2227 (Tagged-LrCas9-Plasmid) for mapping. We used READemption gene_quanti to quantify aligned reads overlapping genomic features by at least 10 nts (-o 10) on the sense strand (-a). For this, we supplemented annotations for the respective RefSeq assembly (antisense_RNA, CDS, ncRNA, riboswitch, Rnase_P_RNA, rRNA, SRP_RNA, tmRNA, tRNA; GCF_000026505.1: annotation date 06/07/2020, GCF_000750555.1: annotation date 02/10/2020) in GFF format with annotations for crRNA, ecrRNA, tracrRNA, and other genes located on the plasmids (e.g., 3xFLAG-tagged *cas9*). Links to plasmid sequences and annotations can be found under Supplementary Table 4. In addition, READemption was applied to generate coverage plots representing the numbers of mapped reads per nucleotide. Here, we used sequencing depth-normalized files from output folder coverage-tnoar_mil_normalized for visualization.

To generate coverage plots and read counts for the ecrRNA and mature crRNAs, we applied a filtering step to the READemption BAM files after mapping. Specifically, all read alignments with a reference length >50 nts overlapping the respective CRISPR region (*E. coli* K-12 BW25113: NL/ML-Tagged-SpCas9-Plasmid: 7346 - 8170, *L. rhamnosus* GG: NC_013198.1: 2265656 - 2267803) were removed utilizing pysam (https://github.com/pysam-developers/pysam) [59] version 0.16.0.1. All subsequent steps were conducted as described above, and the total number of aligned reads before filtering was used to normalize filtered as well as unfiltered read counts and coverage. The normalized filtered read counts were compared directly when evaluating relative (e)crRNA abundance between or within samples.

To visualize read coverage in CRISPR regions, we applied pyGenomeTracks [60] version 3.5 after converting normalized coverage files to BigWig format [61] using wigToBigWig v4.

Bioinformatic identification of CRISPR-Cas systems

Complete and draft bacterial genomes were downloaded from NCBI. CRISPR-Cas systems were annotated using CRISPRcasIdentifier [62] and Casboundary [63] , and CRISPR arrays were extracted from genomes only containing I-E (4,991 arrays), I-F (2,632 arrays), II-A (211 arrays), and II-C (636 arrays) systems using CRISPRidentify [64] . Array orientations were then detected using CRISPRstrand[65] followed by manual curation. The most frequent repeat in each CRISPR array was assigned as the consensus repeat. See Supplementary Table 2 for all leader-repeat sequences.

Bioinformatic assessment of leader-repeat structure formation

To gain insight into potential mechanisms for the inactivation of the first repeat in Type II-A CRISPR arrays, we initially interrogated four CRISPR arrays in Streptococcus pyogenes M1 GAS, Lactobacillus rhamnosus GG, Streptococcus thermophilus CNRZ1066 and Streptococcus thermophilus ND07. Based on these observations, we studied a larger set of CRISPR arrays. In this analysis, we assumed that leader sequences would extend 180 nucleotides 5′ to the first repeat, since information on the correct transcription start site was generally unavailable. We split Type II-A examples into two groups whose inferred leader sequences were at least 50% different in pairwise alignments. We used cd-hit version 4.8.1 to cluster sequences by percent identity [66] . Within these two groups, we removed sequences so that they were less than 70% similar to one another. We performed the same procedure for Type II-C examples. We then conducted our initial analysis on the first subset of CRISPR arrays, which comprised 38 Type II-A and 112 Type II-C examples. As a statistic to represent pairing potential, we first considered the average probability that a nucleotide in the repeat would bind a nucleotide in the leader. We also considered the probability of forming helices in the first repeat with different numbers of base pairs and different numbers of mismatches or bulges. Base-pairing probabilities were calculated using version 2.4.14 of the ViennaRNA library for Python. Since an efficient algorithm to determine the probability of helix formation has not been published, we used ViennaRNA to sample random structures from the Boltzmann probability distribution, which corresponds to the probability of different structures forming at thermodynamic equilibrium. This strategy has been used previously to estimate probabilities of complex events [67] . We used 1,000 random samples. In all cases, we performed our calculations such that base pairs fully contained within the repeats and base pairs fully contained within the leader did not contribute to base-pairing probabilities or to helix-formation probabilities. To estimate the statistical significance of the base-pairing or helix-formation probabilities, we generated random samples by permuting the nucleotides within the leader sequence randomly. Because dinucleotide frequencies can bias RNA folding energies, we permuted the sequences in such a way as to exactly preserve the dinucleotide frequencies, using Peter Clote’s implementation (available through the link below) of a previously published method [68] . We used 1,000 random samples to estimate empirical p-values.

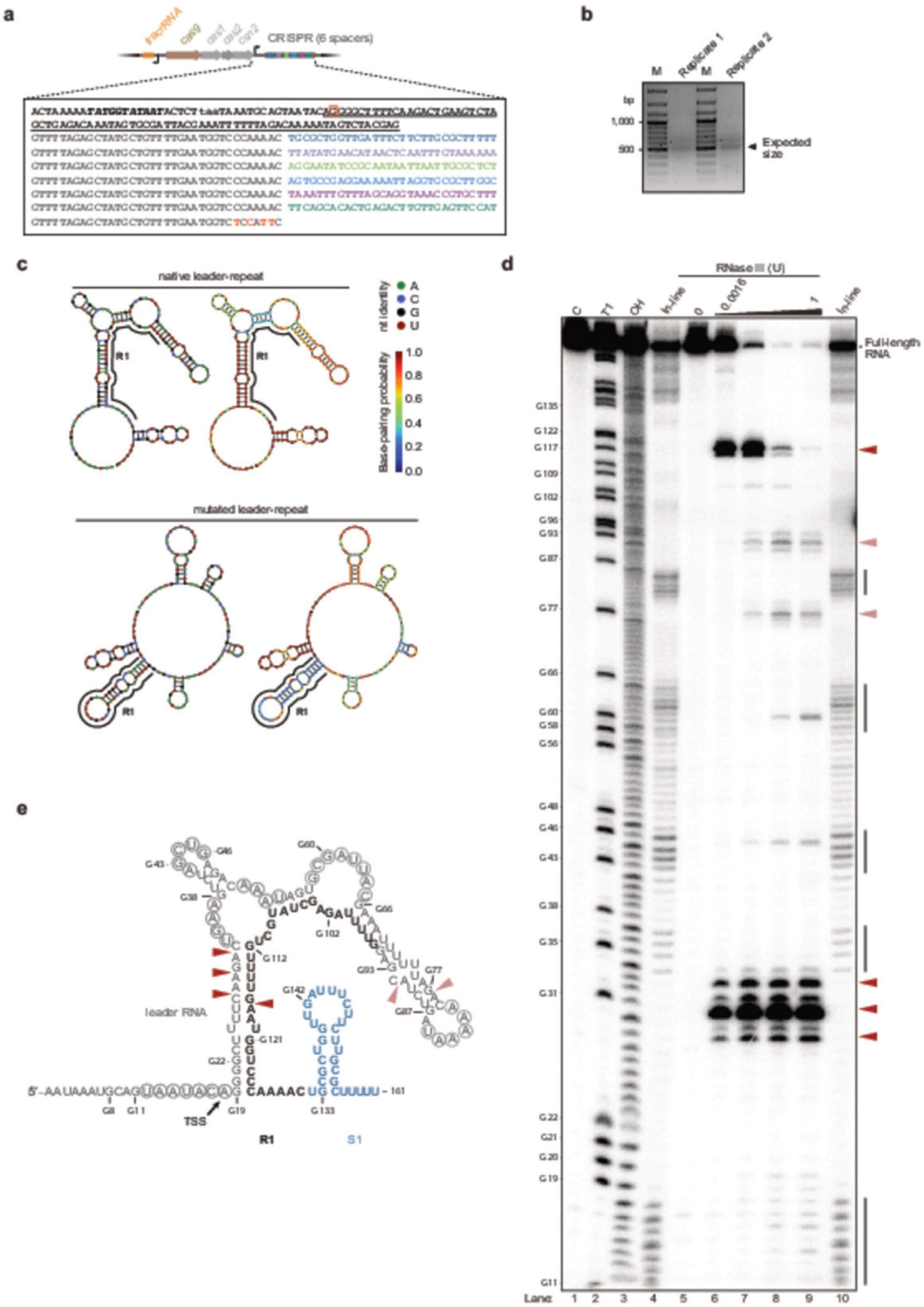
http://clavius.bc.edu/~clotelab/RNAdinucleotideShuffle/ShuffleCodeParts/altschulEriksonDinuclShuffle.txt. For each of the II-A or II-C leader/repeat pairs, and for each statistic (e.g. helix-formation probability), we calculated a corresponding p-value. We combined the p-values for II-A examples and for II-C examples using Fisher's Method, as implemented by the scipy.stats.combine_pvalues function in Python3 (using version 1.4.1 of the scipy library). We refer to these as aggregate p-values. We faced two technical issues in our use of Fisher's Method. First, the method uses the sum of the logarithms of the individual p-values. Because our empirical p-values are based on 1,000 samples, some estimated p-values will be zero (in cases of a very strong helix), leading to logarithms of negative infinity. To address this issue, we replaced empirical p-values of zero with the value 1/1,000. This value is slightly larger than 1/1,001, which would be the estimate according to Laplace's Rule of Succession. We did not adjust other empirical p-values. Second, Fisher's Method assumes that the p-values are independent, but our p-values are based on sequences that presumably are evolutionarily related. We hoped that elimination of sequences that are more than 70% identical will eliminate this problem. It was not practical to more aggressively eliminate similar sequences (e.g. at 50% identity), because of the relatively low number of II-A systems currently available. Based on our experiments with the first subset of CRISPR arrays, we found that one of the statistics that was most elevated in the II-A leader/repeat sequences was the probability of forming a helix containing at least eight uninterrupted base pairs, and we decided to use this statistic for further analysis. We considered the possibility of using only 80 or 100 nucleotides upstream of the first repeat as the leader. We also considered treating the last 15 nucleotides of the leader as part of the repeat, such that helices in this region would contribute towards the helix-formation probability. However, we ultimately decided that variant methods did not significantly change the overall statistics, and we continued to use the original formulation. We used the second subset of CRISPR arrays to test our method. This subset consisted of 30 Type II-A and 173 Type II-C leader/repeat pairs. We determined an aggregate p-value using Fisher's Method of 3.19×10-6 for the 30 type II-A examples and 0.495 for the 173 Type II-C examples. Although we decided not to treat the last 15 nucleotides of the leader as if it were part of the repeat, we noticed that we obtained significant aggregate p-values for Type II-C examples. Therefore, there might be pairing propensity in some of the Type II-C leaders. For the diagram in Fig. 5A, we used all 68 type II-A leader/repeat examples that are less than 70% identical to one another. To make the II-C examples a similar height, we clustered them at 51.1% identity, which also resulted in 68 examples. A similar evaluation was also performed with I-E and I-F CRISPR-Cas systems. Each subtype was split into two groups at 50% identity. We then removed systems that were more than 70% identical. We thus arrived at 379 I-E systems and 151 I-F systems in the initial set. We used this set to analyze our results, and quickly found that our previously applied procedure did not lead to statistically significant aggregate p-values. We then analyzed the second, independent dataset, which consisted of 123 I-E and 142 I-F systems. We also arrived at insignificant aggregate p-values in this case. The p-values for the I-E and I-F systems were both 1 because, in a high proportion of I-

E and I-F CRISPR-Cas systems, the probability that eight consecutive base pairs would form was very low. This fact led to some high individual empirical p-values, and thus a very high aggregate p-value. For the diagram in Fig. 5A, we used the 67 I-E systems that clustered at 51.7% identity as well as the 70 I-F systems that clustered at 53% identity. Both of these numbers (67 and 70) were similar to the 68 systems used for the II-A and II-C depictions.

Statistical analyses

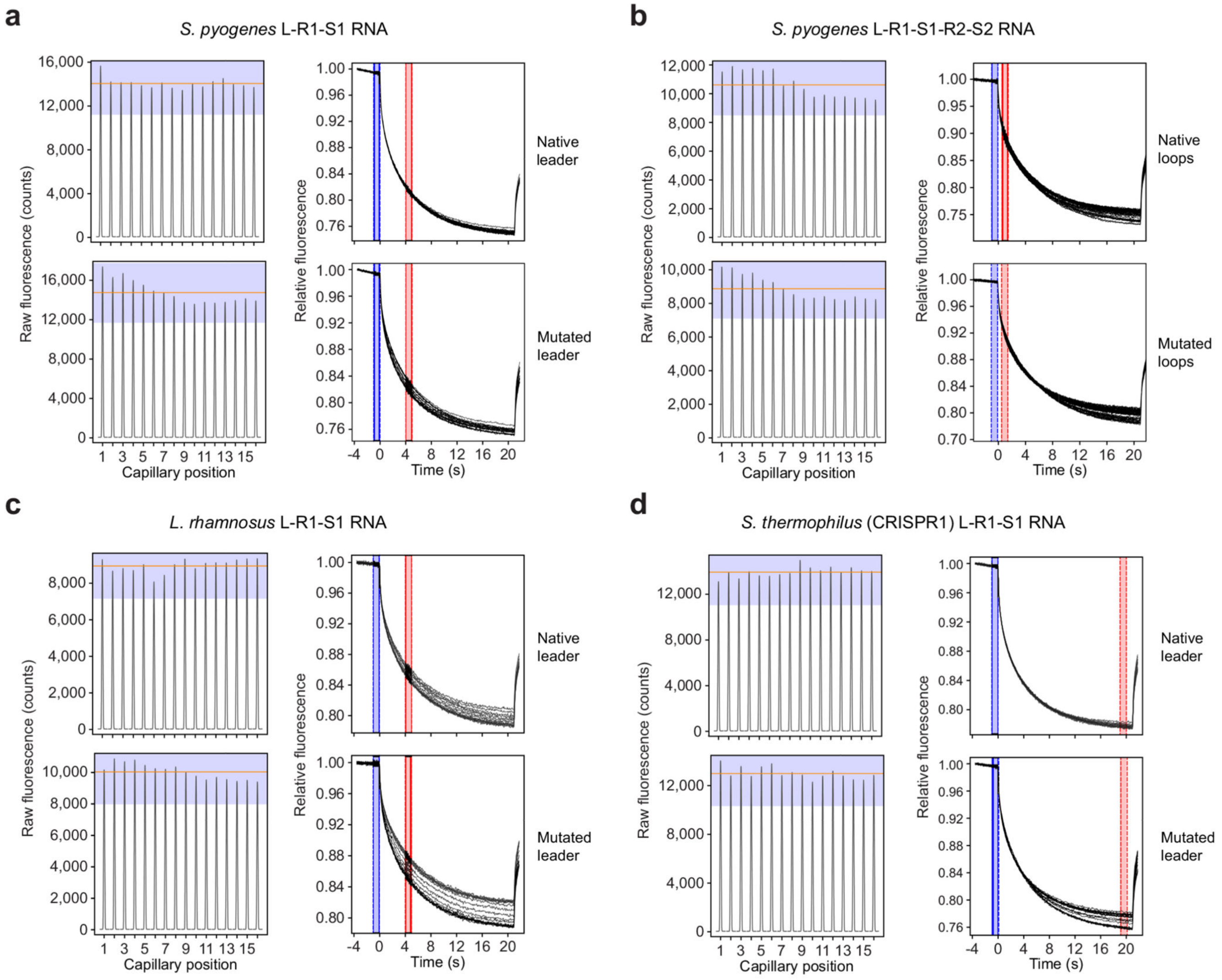
Statistical comparisons of experimental data were performed using a Student's two-tailed t-test assuming unequal variance. Values were assumed to be normally distributed with the exception of transformation efficiencies, which were assumed to be normally distributed only after applying the logarithm. To analyze the folding predictions for the sets of leader-repeat RNAs, empirical p-values were calculated using randomly shuffled leader sequences and then combined into a single p-value using Fisher's Method. The threshold of significance was set as 0.05 in all cases.

Extended Data


a
tracrRNA
cas9
cas1
cas2
csn2
CRISPR (6 spacers)
ACTAAAAATATGGTATAATACTCT taaTAAATGCAGTAATACAGGGGCTTTTCAAGACTGAAGTCTA
GCTGAGACAAATAGTGCGATTACGAAATTTTTTAGACAAAAATAGTCTACGAG
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC TGCGCTGGTTGATTTCTTCTTGCGCTTTTT
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC TTATATGAACATAACTCAATTTGTAAAAAA
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC AGGAATATCCGCAATAATTAATTGCGCTCT
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC AGTGCCGAGGAAAAATTAGGTGCGCTTGGC
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC TAAATTTGTTTAGCAGGTAAACCGTGCTTT
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC TTCAGCACACTGAGACTTGTTGAGTTCCAT
GTTTTAGAGCTATGCTGTTTTGAATGGTCTCCATTC
b
M
Replicate 1
M
Replicate 2
bp
1,000
500
Expected size
c
native leader-repeat
nt identity
A
C
G
U
Base-pairing probability
1.0
0.8
0.6
0.4
0.2
0.0
R1
R1
mutated leader-repeat
R1
R1
d
RNase III (U)
C
T1
OH
In-line
0
0.0016
1
In-line
Full-length RNA
G135
G122
G117
G109
G102
G96
G93
G87
G77
G66
G60
G58
G56
G48
G46
G43
G38
G35
G31
G22
G21
G20
G19
G11
Lane 1 2 3 4 5 6 7 8 9 10
e
G43
G46
G38
G60
G66
G102
G112
G142
G93
G77
G87
G121
G22
leader RNA
5'-AAUAAAUGCAGUAAUACAGGCAAAACUGCUUUUU - 161
G8
G11
G19
G133
TSS
R1
S1

Extended Data Fig. 1. The leader-repeat stemloop from the CRISPR-Cas9 system native to *Streptococcus pyogenes* SF370.

Accession #: NC_002737.2. **a**, Array sequence and context within the CRISPR-Cas system. Repeats are in gray, spacers match the corresponding color in the cartoon, and mutations to the consensus repeat are shown in red. The underlined sequence encodes the transcribed RNA leader as determined in *S. pyogenes* SF370 [18] . The bold and italicized sequence is the putative -10 promoter element, while the lowercase letters designate the stop codon of *csn2*. The red box indicates the mapped transcriptional start site in *E. coli* determined using 5′ RACE. **b**, PCR product generated by 5′ RACE. Biological duplicates are shown. M: DNA marker. **C**, Predicted minimal free-energy structure of the native and mutated leader-repeat RNA predicted by NUPACK. Left: nucleotide (nt) identities. Right: base-pairing probabilities. **d**, *In vitro* determination of the secondary structure and RNase III cleavage sites for the leader-repeat RNA associated with SpyCas9. The transcription start site was extended by 17 nts using the sequence from *S. pyogenes* to allow visualization of shorter RNAs. Vertical bars: unstructured regions. C: full-length (untreated) control. T1: Ladder of G's generated by incubating the RNA with RNase T1. OH: single-nucleotide ladder generated by incubating the RNA under basic conditions. Dark and light red arrows indicate the most and second most preferred sites of RNase III cleavage, respectively. Results are representative of triplicate independent experiments. **e**, Corresponding secondary structure of the leader-repeat RNA. Circles indicate unstructured bases identified by in-line probing. The preferred site of RNase III cleavage lies within one nt of the equivalent site within the crRNA:tracrRNA duplex (see Figure 1c). R1: first repeat. S1: first spacer.

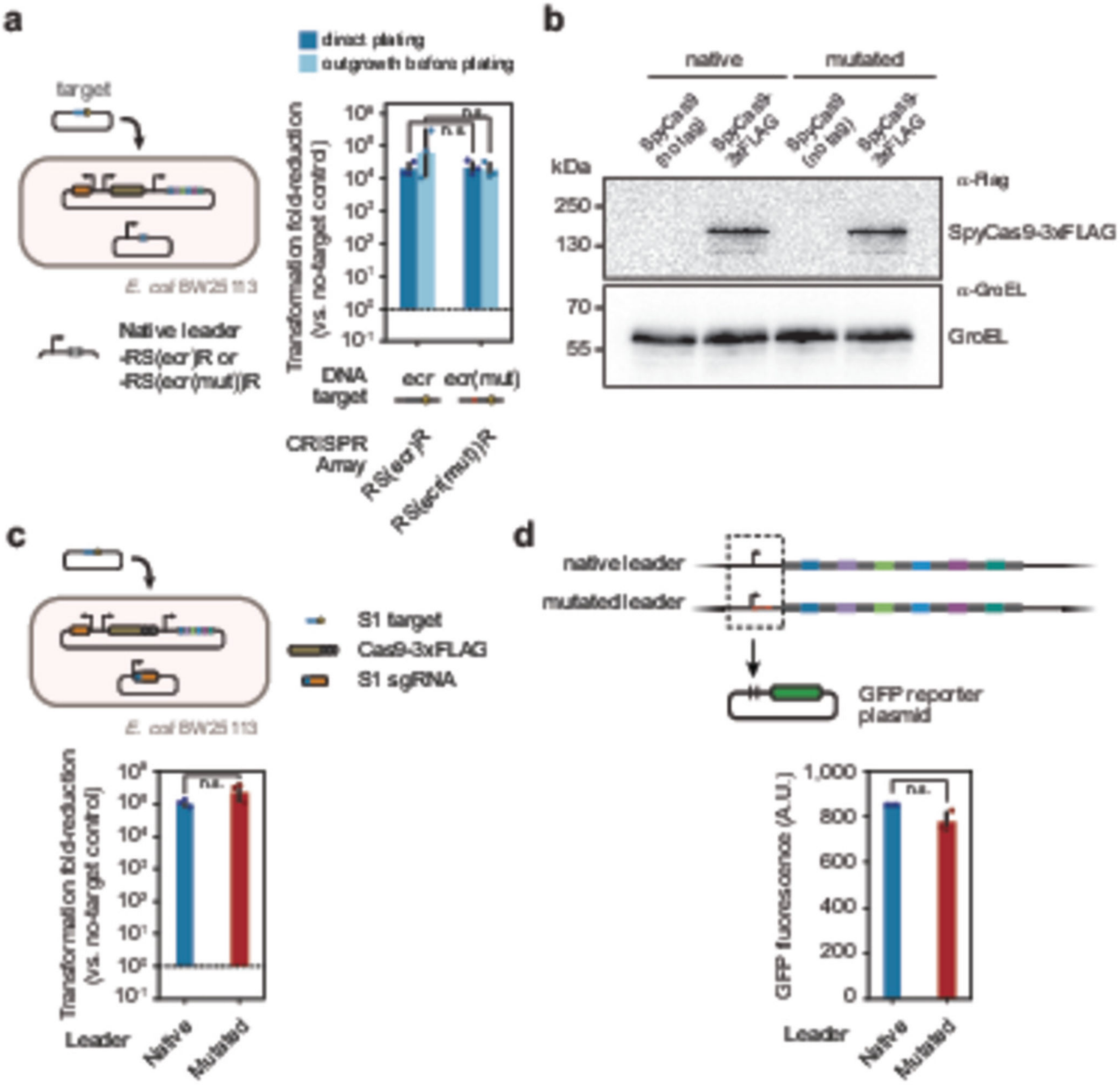


Extended Data Fig. 2. Capillary scans and thermophoretic time-traces of microscale thermophoresis (MST) measurements of binding between the leader-repeat RNA and tracrRNA associated with different CRISPR-Cas9 systems.

a, *Streptococcus pyogenes* SF370 with an RNA spanning the leader to the first spacer.

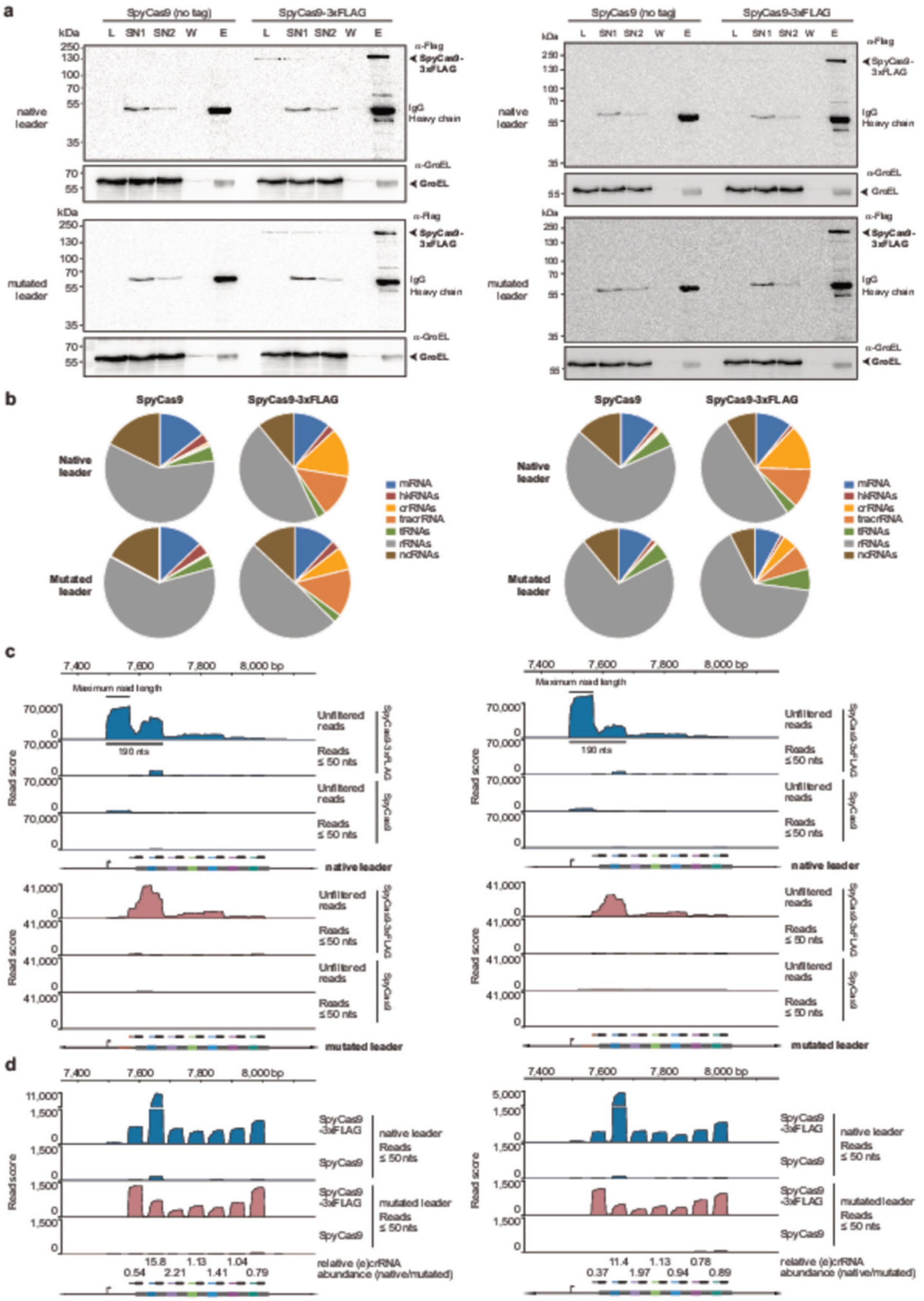
b, *Streptococcus pyogenes* SF370 with an RNA spanning the leader to the second spacer.

c, *Lactobacillus rhamnosus* GG with an RNA spanning the leader to the first spacer. **d**, *Streptococcus thermophilus* DGCC 7710 with an RNA spanning the leader to the first spacer. In all cases, the tracrRNA was fluorescently labeled while unlabeled leader-repeat RNA was added at different concentrations. Capillary scans and traces of one of three independent experiments are shown. The gray boxes in the capillary scans mark 20% above and below the average peak fluorescence indicated in orange, the acceptable limit of deviations across the fluorescence scans. Blue and red boxes in the time-course traces represent the temperature jump and MST-on time, respectively. In all cases, there is no adsorption of the labeled tracrRNAs to the capillaries, and the time traces indicate no aggregation. See Figures 1d and 4d and Extended Data Figures 8b and 8e for the resulting binding curves. Values in a-d represent the mean and standard deviation of triplicate independent measurements.



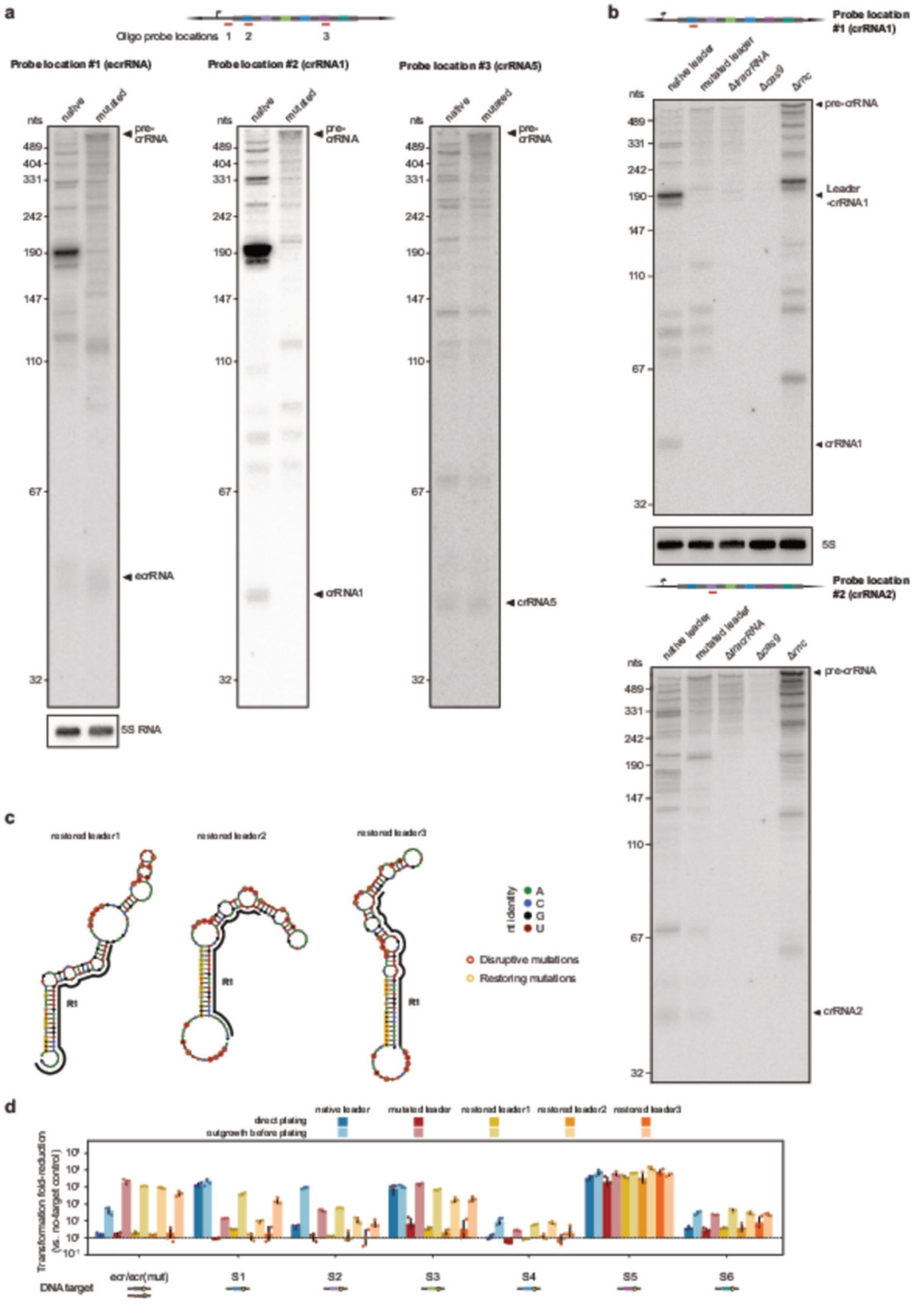
Extended Data Fig. 3. Data rejecting alternative explanations for the impact of mutating the leader region associated with SpyCas9.

a, Assessing targeting by the mutated ecrRNA guide by plasmid clearance in *E. coli*. The native and mutated ecrRNAs were encoded as single-spacer arrays with the native leader. There was no significant difference in plasmid clearance with (Student’s two-tailed t-test with unequal variance, p = 0.36, n = 3) or without (Student’s two-tailed t-test with unequal variance, p = 0.80, n = 3) outgrowth. **b**, Western blotting analysis of SpyCas9-3xFLAG levels with the native or mutated leader. Results are representative of two independent experiments. **c**, Plasmid clearance with SpyCas9-3xFLAG in *E. coli*. The SpyCas9-3xFLAG fusions were tested using an sgRNA with a guide derived from spacer 1 (S1) in the native array. The transformations were conducted without non-selective outgrowth. The results showed that the fusion did not compromise clearance activity by SpyCas9, and introducing the mutations into the CRISPR leader did not significantly affect SpyCas9 activity (Student's two-tailed t-test with unequal variance, p = 0.168, n = 3). **d**, Assessing transcription of the CRISPR array with the mutated leader. The native or mutated leader through the first spacer was cloned upstream of *gfp* in the pUA66 plasmid. *E. coli* cells harboring either plasmid were then subjected to flow cytometry analysis. There was no significant difference (Student's two-tailed t-test with unequal variance, p = 0.103, n = 3) in the background-subtracted GFP fluorescence between the constructs. Values represent the mean and standard deviation of triplicate independent measurements starting from separate colonies. Values in a, c and d represent the geometric mean and standard deviation from independent experiments starting from three separate colonies. n.s.: not significant. n.s.: p > 0.05. Statistical tests were performed using a two-tailed Student's t-test with unequal variance, n = 3.



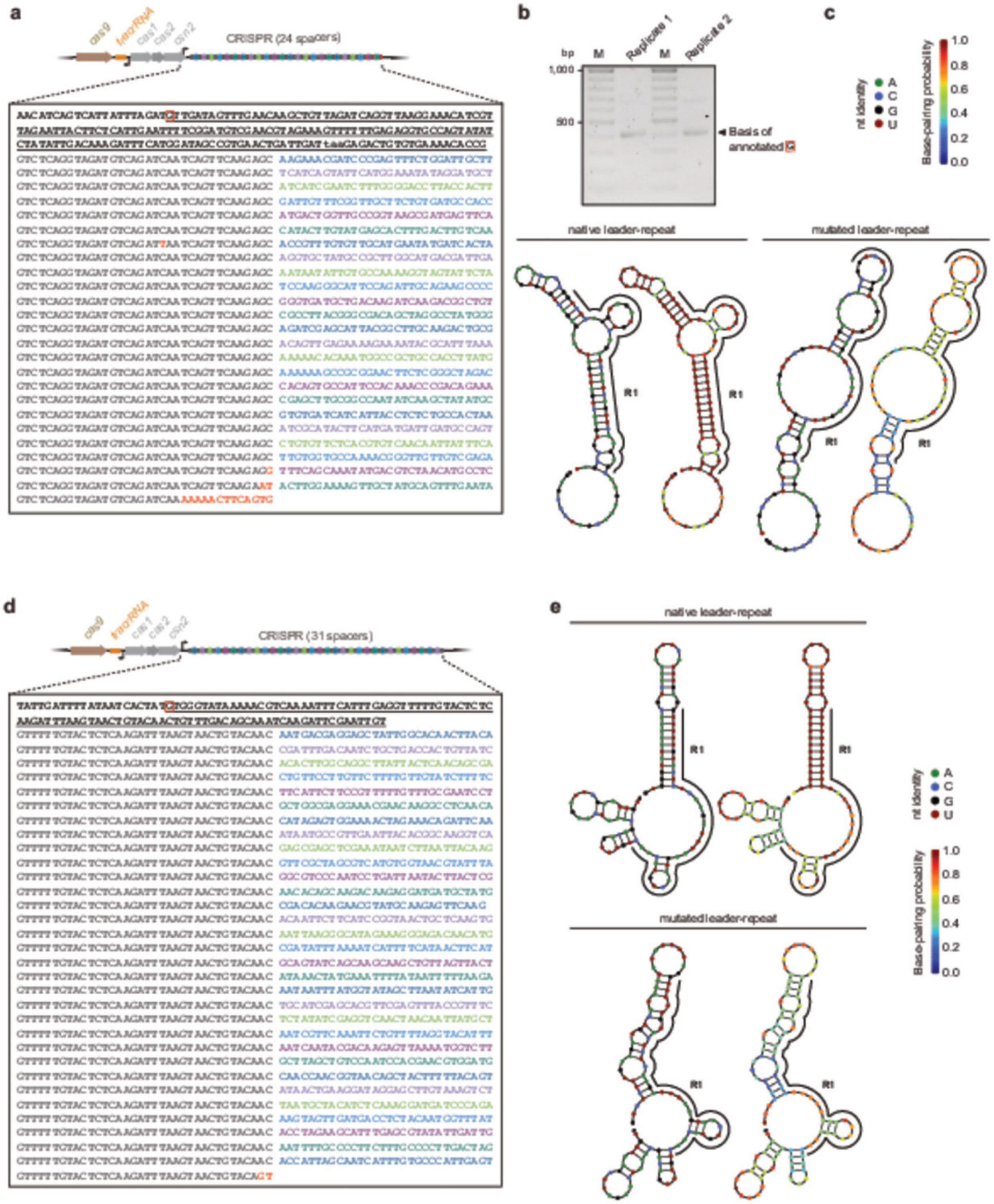
Extended Data Fig. 4. RIP-seq analysis using SpyCas9 combined with the native or mutated leader in *E. coli*.

The left and right sides of the figure represent the results from two independent experiments. RIP-seq was performed using *E. coli* BW25113 harboring the SpyCas9/tracrRNA/CRISPR or SpyCas9-3xFLAG/tracrRNA/CRISPR plasmid. **a**, Western blotting confirmed enrichment of SpyCas9-FLAG. Co-immunoprecipitated RNAs were isolated and subjected to next-generation sequencing. **b**, Distribution of RNA classes based on total mapped reads.**c**, Mapped reads for the CRISPR locus with the native or mutated leader. The scale above the plot indicates the location in the plasmid. Positional coverage for total aligned reads and reads aligning with a reference length ≤ 50 nts was normalized based on the total number of aligned reads in each sample. The reduction in reads upon applying the size filter indicates an excess of pre-crRNA and immature crRNAs, which parallels Northern blotting analysis for the ecrRNA and individual crRNAs (see Figure 3b and Extended Data Figure 5a-b). We also note that the reads begin ~12 nts upstream of the transcriptional start site mapped by 5' RACE (see Extended Data Figure 1), suggesting that a slightly upstream transcriptional start site or processing site from a longer transcript also exists. **d**, Direct comparison of mapped reads with the native or mutated leader. The plot corresponds to that shown in Figure 3a. The read score for the first crRNA downstream of the native leader extends above the vertical limit of 1,500. The relative read scores for the ecrRNA and each crRNA are indicated below the plots. Values below one indicate a reduction in (e)crRNA abundance with the introduced mutations. See Supplementary Table 1 for statistics about the RIP-seq analyses.



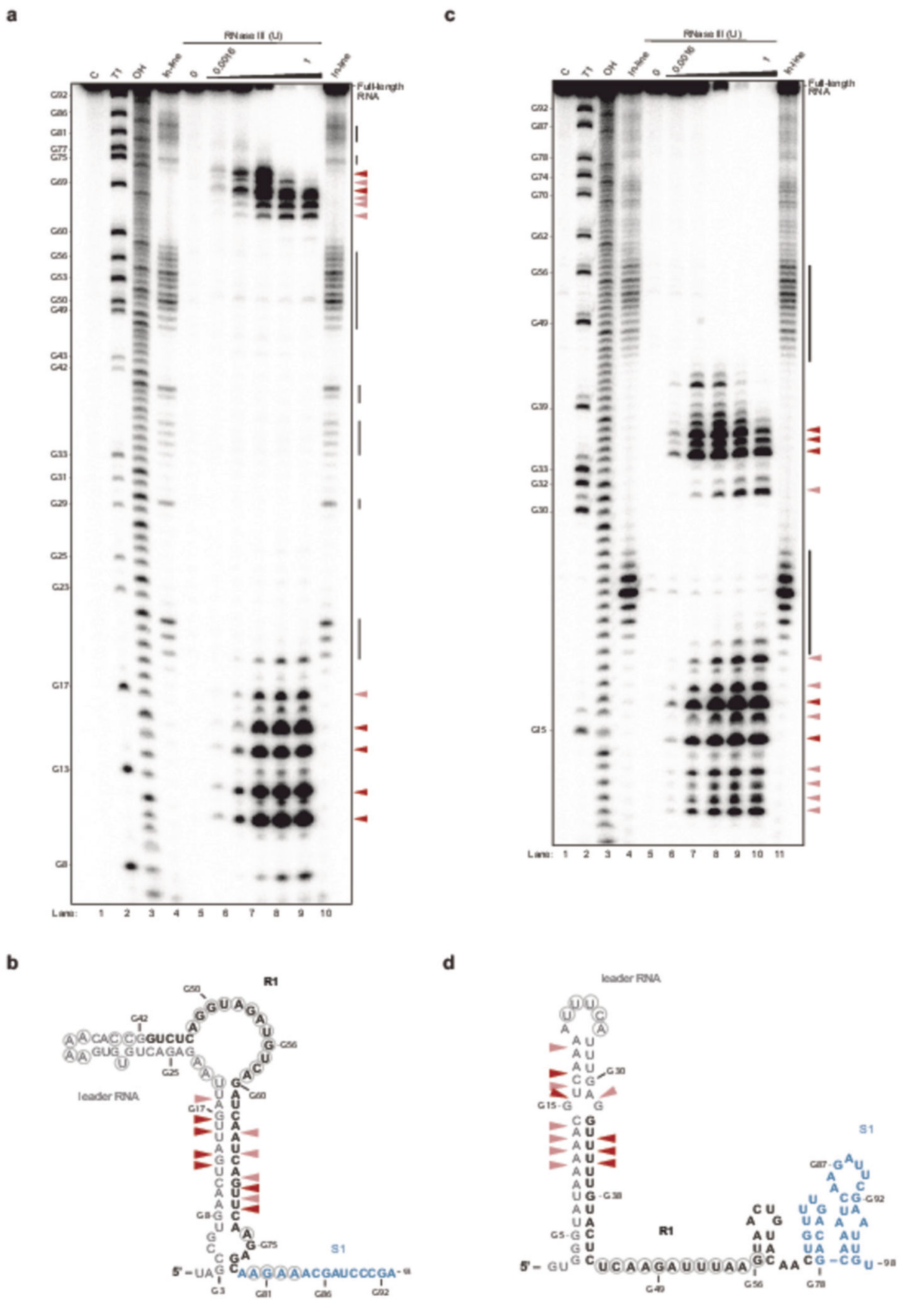
Extended Data Fig. 5. Impact of mutating the leader-repeat stem-loop from the CRISPR-Cas9 system from *Streptococcus pyogenes* SF370.

a, Northern blotting analysis of the produced crRNAs with the native or mutated RNA leader. The system's CRISPR array was expressed in *E. coli* with SpyCas9 and the tracrRNA, and the ecrRNA (probe #1), crRNA1 (probe #2), and crRNA5 (probe #3) were detected. The ecrRNA and mecrRNA were detected using an equimolar mixture of both probes. **b**, Northern blotting analysis of the produced crRNAs with different mutant backgrounds. See a for details. Experiments were conducted with the native or mutated leader or with the *tracrRNA*, *cas9*, or *rnc* deleted. The results for probe #1 are those shown in Figure 3b. All probing was performed with the same blot. The indicated RNA spanning the leader through the processed crRNA1 corresponds to that observed by RIP-seq (see Extended Data Figure 4c) and is supported by the band's absence when probing for crRNA2. Results in a and b are representative of duplicate independent experiments. **c**, Predicted secondary structures of three different restoring mutant sets. Disruptive mutations were made to the mutated leader depicted in Figure 1c. In each case, a stable stem was created by making restoring mutations, although the upper structure deviates from that found in the native leader-repeat. **d**, Impact of the mutations on plasmid clearance by SpyCas9 in *E. coli*. The clearance assays were conducted with or without a non-selective outgrowth, where the non-selective outgrowth improves the extent of plasmid clearance. Values represent the geometric mean and standard deviation from independent experiments starting from three separate colonies.



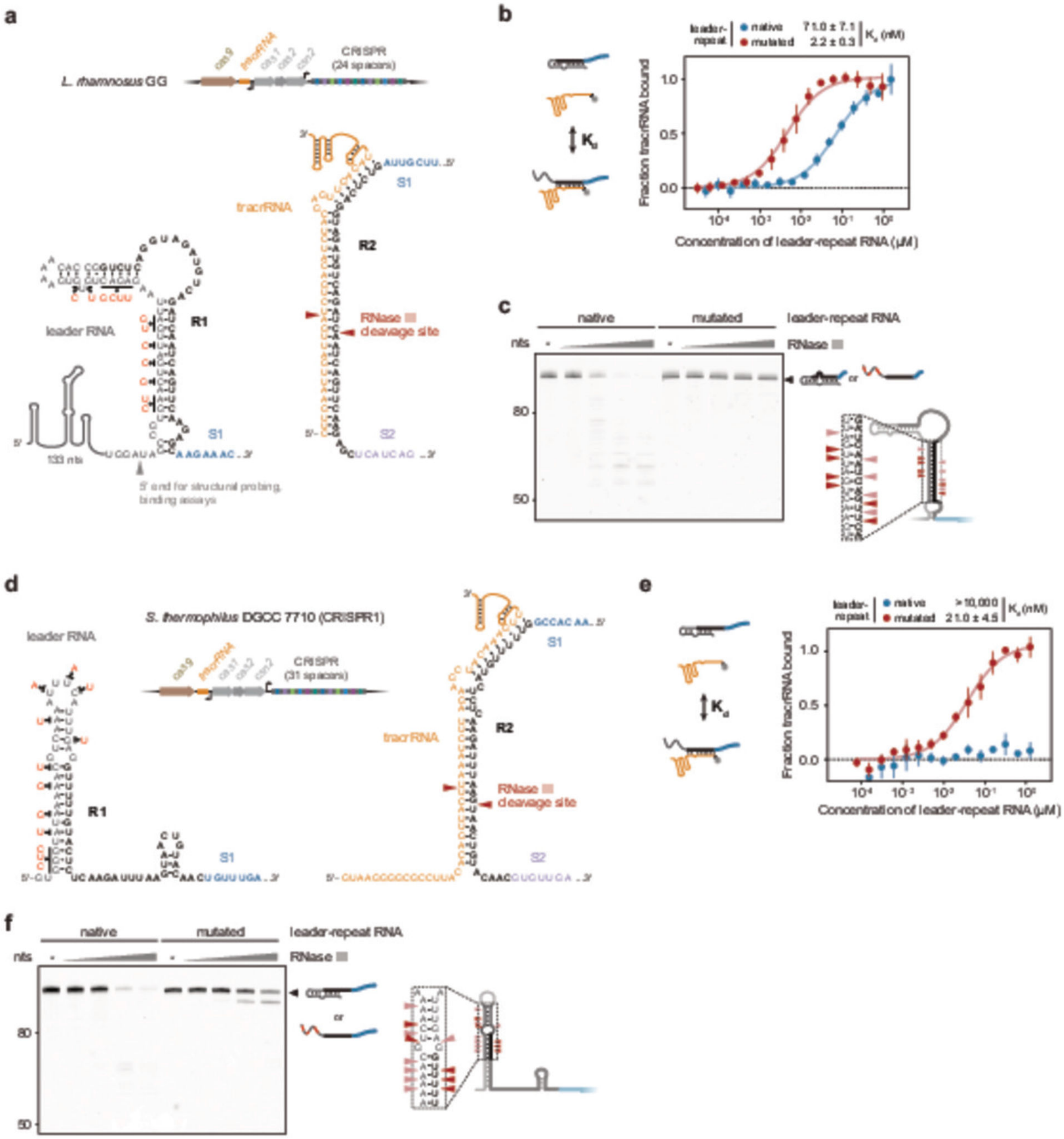
Extended Data Fig. 6. CRISPR arrays from other CRISPR-Cas9 systems within the II-A subtype that appear to possess a leader-repeat stem-loop.

a, Array sequence and context within the CRISPR-Cas system native to *Lactobacillus rhamnosus* GG. Accession #: GCF_000026505.1. The sequence begins within *csn2* (annotated as LGG_02201) and ends after the terminal repeat. See Extended Data Figure 1a for details. The underlined sequence encodes the transcribed RNA leader as determined by 5′ RACE in *L. rhamnosus* in this work. Lowercase letters designate the stop codon of *csn2*. The promoter(s) driving expression of the *cas* genes has not been mapped. **b**, PCR product as part of 5′ RACE using total RNA from *L. rhamnosus* GG. See Extended Data Figure 1b for details. Only one major product was visible in both replicates. Biological duplicates are shown. M: DNA marker. Results from duplicate independent experiments are shown. **c**, Secondary structure of the native and mutated leaderrepeat RNA predicted by NUPACK. See Extended Data Figure 1c for details. The 5′ of the leader was truncated to match the sequence used in the structural probing and RNase III cleavage assays (see Extended Data Figure 7b). Mutations were selected to disrupt the original secondary structure of the native leader-repeat RNA. **d**, Array sequence and context within the CRISPR-Cas system native to *Streptococcus thermophilus* DGCC 7710 (CRISPR1 locus). Accession #: CP025216.1. The sequence begins downstream of *csn2* and ends after the terminal repeat. See Extended Data Figure 1a for details. The underlined sequence encodes the transcribed RNA leader as determined previously by RNA sequencing analysis of transcripts [16]. The promoter(s) driving expression of the *cas* genes has not been mapped. **e**, Secondary structure of the native and mutated leaderrepeat RNA predicted by NUPACK. See Extended Data Figure 1c for details.



Extended Data Fig. 7. *In vitro* determination of the secondary structure and RNase III cleavage sites for the leader-repeat RNA associated with LrhCas9 and Sth1Cas9.

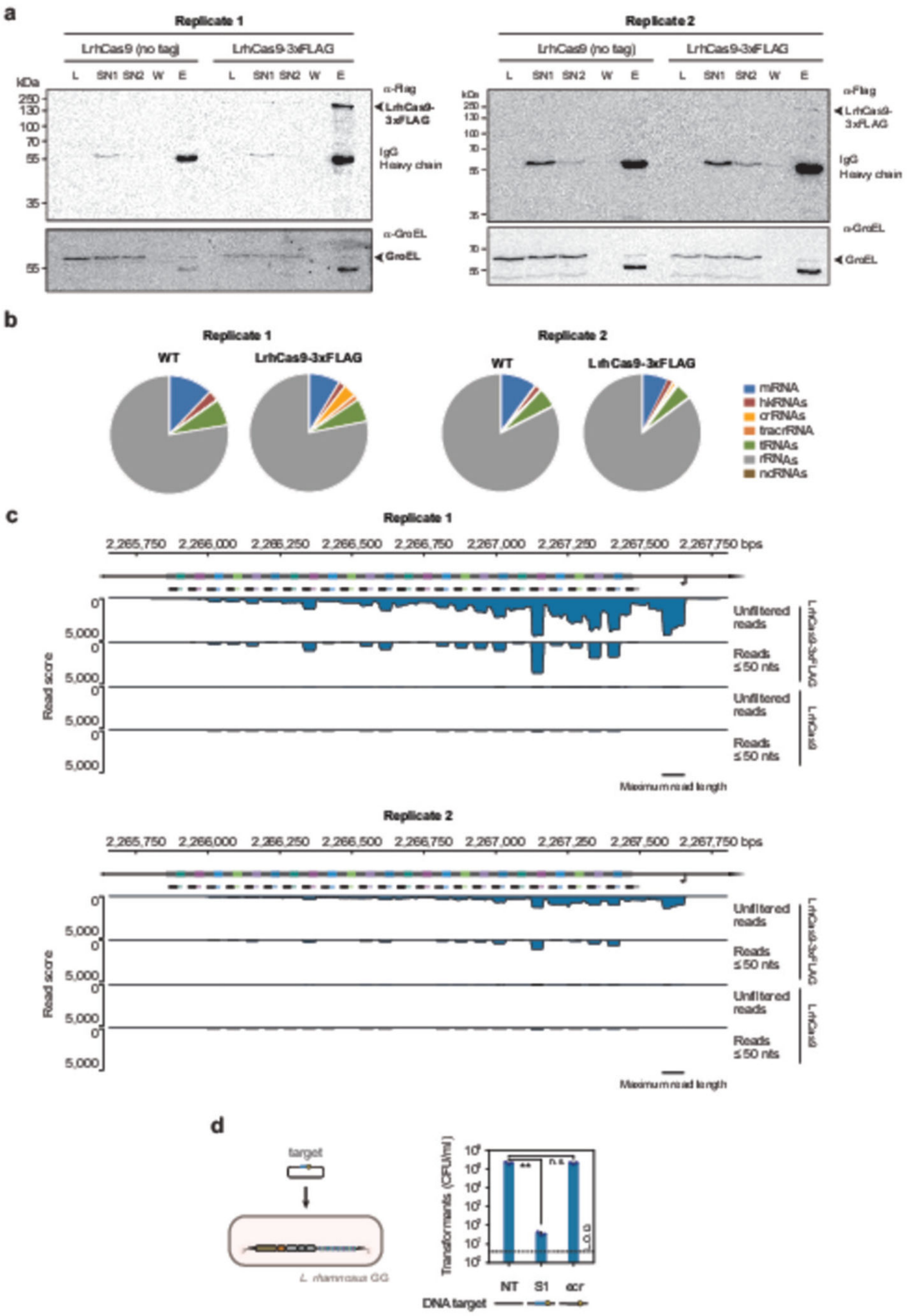
a, *In vitro* determination of the secondary structure and RNase III cleavage sites for the leader-repeat RNA associated with LrhCas9. The probed RNA was 5′ radiolabeled and resolved by denaturing PAGE. The 5′ end was truncated to focus on the predicted secondary structure involving the repeat. Vertical bars on the right indicate unstructured regions. C - full-length control. T1: Ladder of G’s generated by incubating the RNA with RNase T1. OH: single-nucleotide ladder generated by incubating the RNA under basic conditions. RNase III: the RNA was incubated with the indicated units of *E. coli* RNase III (0, 0.0016, 0.008, 0.04, 0.2, 1) for 5 min at 37°C. Dark and light red arrows indicate the most preferred and second most preferred sites of RNase III cleavage, respectively. Results are representative of triplicate independent experiments. **b**, Corresponding secondary structure of the leader-repeat RNA. Circles indicate unstructured bases identified by in-line probing. The preferred site of RNase III cleavage lies below the equivalent site within the crRNA:tracrRNA duplex (see Extended Data Figure 8a). R1: first repeat. S1: first spacer. **c**, *In vitro* determination of the secondary structure and RNase III cleavage sites for the leaderrepeat RNA associated with Sth1Cas9. See a for details. The 5′ end was truncated to focus on the predicted secondary structure involving the repeat. Results are representative of triplicate independent experiments. **d**, Corresponding secondary structure of the leader-repeat RNA. Circles indicate unstructured bases identified by inline probing. The preferred site of RNase III cleavage lies above the equivalent site within the crRNA:tracrRNA duplex (see Extended Data Figure 8d). R1: first repeat. S1: first spacer.



Extended Data Fig. 8. II-A CRISPR-Cas9 systems form distinct leader-repeat stemloops.

a, The CRISPR-Cas system from *L. rhamnosus* GG and the secondary structure of the leader-repeat RNA. The structure was predicted by NUPACK and confirmed *in vitro* (see Extended Data Figures 6c and 7a-b). Mutations indicated in red were made to disrupt stems formed between the leader RNA and the first repeat. **b**, Measured equilibrium binding between the tracrRNA and native or mutated RNA leader-repeat RNA. See Extended Data Figure 2c for supporting data. Values represent the mean and standard deviation of triplicate independent measurements. **c**, RNase III cleavage of the native and mutated leader-repeat

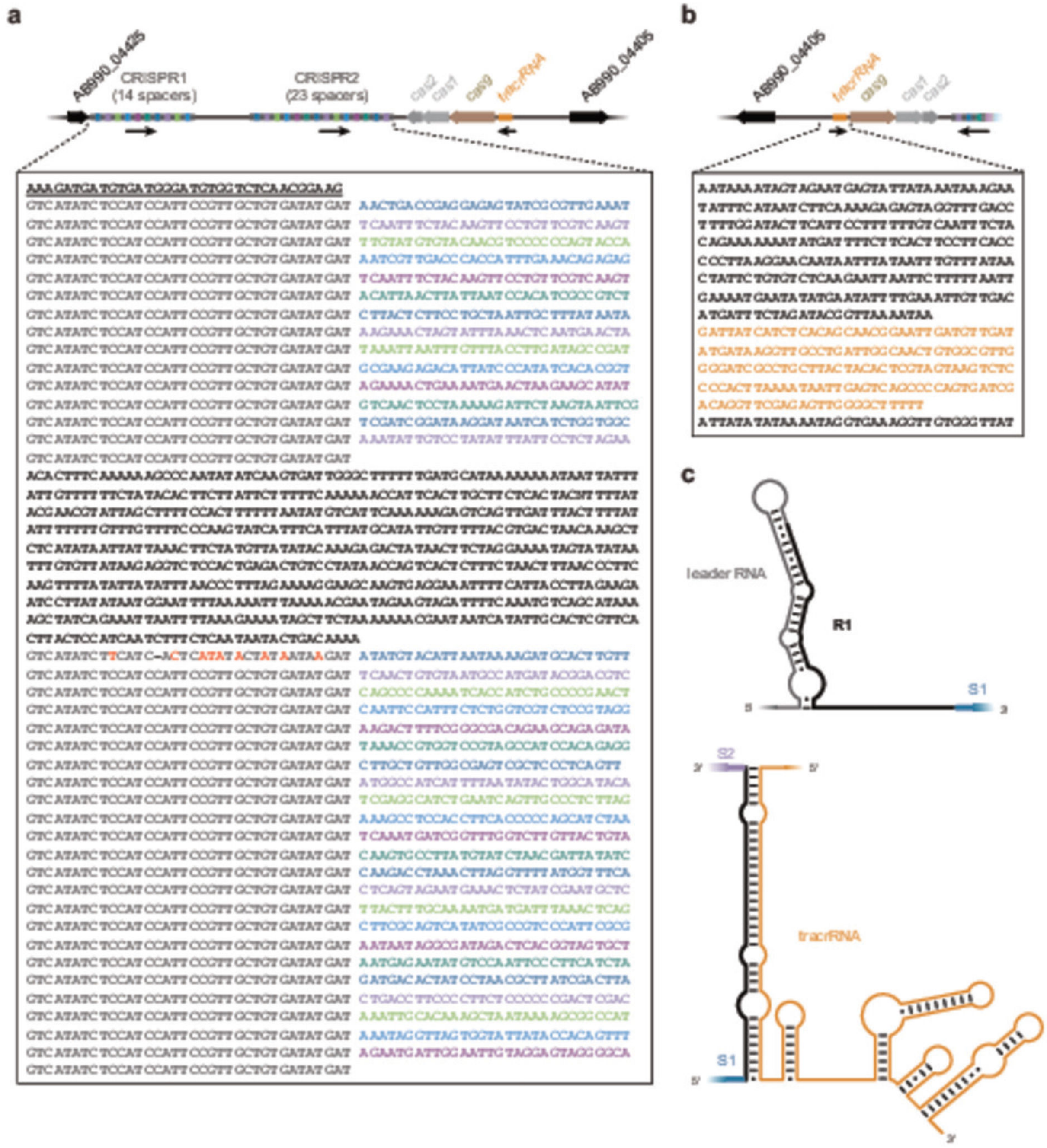
RNA *in vitro*. See Extended Data Figure 7a-b for the mapped secondary structure and RNase III cleavage sites. Results are representative of duplicate independent experiments. **d**, The CRISPR-Cas system associated with the CRISPR1 locus of *S. thermophilus* and the secondary structure of the leader-repeat RNA. The structure was predicted by NUPACK and confirmed *in vitro* (see Extended Data Figures 6e and 7c-d). Indicated mutations in red were made to disrupt the stem formed between the leader RNA and first repeat. The three mutations in the loop were introduced to disrupt alternative structures formed by the other mutations. Pairing between the repeat and the tracrRNA is provided as a basis of comparison. Red arrows indicate the previously mapped site cleaved by RNase III [16] . R1: first repeat. R2: second repeat. S1: first spacer. S2: second spacer. **e**, Measured equilibrium binding affinity between the leader-repeat and the tracrRNA under *in vitro* conditions. See Extended Data Figure 2d for supporting data. Values represent the mean and standard deviation of triplicate independent measurements. **f**, RNase III cleavage of the native and mutated leader-repeat RNA *in vitro*. Results are representative of duplicate independent experiments.



Extended Data Fig. 9. RIP-seq analysis of RNAs bound to Cas9 from *Lactobacillus rhamnosus* GG. LrhCas9 with or without a 3xFLAG affinity was expressed from a plasmid, and the lysate was subjected to RIP-seq analysis.

LrhCas9 with or without a 3xFLAG affinity was expressed from a plasmid, and the lysate was subjected to RIP-seq analysis. **a**, Western blotting analysis of samples for RIP-seq using LrhCas9 in *Lactobacillus rhamnosus* GG. Western blotting confirmed enrichment of LrhCas9-FLAG. Co-immunoprecipitated RNAs were isolated and subjected to next-generation sequencing. Results from duplicate independent experiments are shown on the left and right. **b**, Distribution of RNA classes based on total mapped reads. hkRNAs:

house-keeping RNAs. ncRNAs: non-coding RNAs. **c**, Mapped reads for the CRISPR locus with the genome of *L. rhamnosus* GG (NC_013198.1). The scale above the plot indicates the location in the genome. The CRISPR locus is encoded on the negative strand. Positional coverage for total reads and reads aligning with a reference length ≤ 50 nts was normalized based on the total number of aligned reads in each sample. The maximum read length for the NGS run was 76 nts, explaining the drop in unfiltered read counts shortly downstream of the transcriptional start site. See Supplementary Table 1 for statistics from the RIP-seq analyses. Results in b and c are representative of duplicate independent experiments. **d**, Plasmid clearance by the CRISPR-Cas9 system in *L. rhamnosus* GG. The corresponding target of the ecrRNA or crRNA1 was encoded within the transformed plasmid. L.O.D.: limit of detection. There was no detectable ecrRNA-directed plasmid clearance. Values represent the geometric mean and standard deviation from three independent experiments starting from separate colonies. **: $p < 0.01$. n.s.: $p > 0.05$. Statistical tests were performed using a two-tailed Student’s t-test with unequal variance, $n = 3$.



Extended Data Fig. 10. The CRISPR array from the CRISPR-Cas9 system native to *Alkalihalobacillus pseudalcaliphilus* DSM 8725.

The system falls within the II-C subtype. Accession #: LFJO01000002.1. **a**, Array sequence and context within the CRISPR-Cas9 system. The sequence begins immediately downstream of the AB990_04425 gene unrelated to the CRISPR-Cas9 system and ends after the last repeat of the CRISPR2 array. Repeats are in gray, spacers match the corresponding color in the cartoon, and mutations to the consensus repeat are shown in red. The underlined sequence denotes the upstream region used for the folding predictions for the CRISPR1 array. The transcriptional start sites for both arrays are unknown, although there is a clear Rho-independent terminator downstream of each array. The promoters driving expression of the *cas* genes, the CRISPR arrays, or the tracrRNA have not been mapped. The predicted direction of transcription for the tracrRNA and CRISPR array are indicated with black arrows. **b**, tracrRNA sequence and context within the CRISPR-Cas9 system. The sequence begins ~2.7 kb upstream of the AB990_04405 gene unrelated to the CRISPR-Cas9 system and ends immediately upstream of *cas9*. The sequence in orange corresponds to the putative tracrRNA used in the folding predictions. **c**, Predicted stem-loop between the first repeat and upstream region for the CRISPR1 array. The predicted stem-loop is part of the minimal-free energy structure and reflects base-pairing probabilities principally between 90% and 100%. Pairing between the second repeat and the tracrRNA is provided as a basis of comparison. The tracrRNA ends with a canonical Rho-independent terminator.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Tatjana Achmedov for extensive assistance with RNA preparation and northern blotting, Franziska Tippel from NanoTemper Technologies (Munich) for technical support, and Jörg Vogel and Gisela Storz for critical feedback on the manuscript. This work was supported by funding through the European Research Council Consolidator Award (865973 to C.L.B.), the Deutsche Forschungsgemeinschaft SPP 2141 (BE 6703/1-1 to C.L.B., SH 580/9-1 to C.M.S., BA 2168/23-1 to R.B.), and the Interdisciplinary Center for Clinical Research Würzburg (IZKF) project Z-6.

Data Availability Statement

Next-generation sequencing data for RNA immunoprecipitation sequencing is accessible through NCBI Gene Expression Omnibus (GEO) accession number GSE158637 using the link https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158637. See Supplementary Table 4 for details. Source data for Figures 1b,d,e, 2a,b, 3b,c,d and 4b,c,d and Extended Data Figures 1b,d, 3a,c,d, 4a,b,c,d, 5a,b,d, 6a, 7a,c, 8b,c,e,f, and 9a,d are included in the Source Data files.

Code Availability Statement

Custom scripts to analyze folding of the leader-repeat region of different CRISPR-Cas systems are available on GitHub at https://github.com/zashaweinberglab/type-II-A-leader-repeat.

References

1. Barrangou R, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007; 315: 1709–1712. [PubMed: 17379808]
2. van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat Rev Microbiol. 2014; 12: 479–492. [PubMed: 24909109]
3. Jackson SA, et al. CRISPR-Cas: Adapting to change. Science. 2017; 356

4. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005; 151: 2551–2561. [PubMed: 16079334]

5. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol. 2005; 60: 174–182. [PubMed: 15791728]

6. Sorek R, Kunin V, Hugenholtz P. CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Microbiology. 2008; 6: 181–186. [PubMed: 18157154]

7. Arslan Z, Hermanns V, Wurm R, Wagner R, Pul Ü. Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system. Nucleic Acids Research. 2014; 42: 7884–7893. [PubMed: 24920831]

8. Xiao Y, Ng S, Nam KH, Ke A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. Nature. 2017; 550: 137–141. [PubMed: 28869593]

9. McGinn J, Marraffini LA. Molecular mechanisms of CRISPR-Cas spacer acquisition. Nat Rev Microbiol. 2019; 17: 7–12. [PubMed: 30171202]

10. Brouns SJJ, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008; 321: 960–964. [PubMed: 18703739]

11. Charpentier E, Richter H, van der Oost J, White MF. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev. 2015; 39: 428–441. [PubMed: 25994611]

12. Garneau JE, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010; 468: 67–71. [PubMed: 21048762]

13. Meeske AJ, Nakandakari-Higa S, Marraffini LA. Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. Nature. 2019; 570: 241–245. [PubMed: 31142834]

14. Rostøl JT, et al. The Card1 nuclease provides defence during type III CRISPR immunity. Nature. 2021; 590: 624–629. [PubMed: 33461211]

15. Elmore JR, et al. Programmable plasmid interference by the CRISPR-Cas system in *Thermococcus kodakarensis* . RNA Biol. 2013; 10: 828–840. [PubMed: 23535213]

16. Carte J, et al. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus* . Mol Microbiol. 2014; 93: 98–112. [PubMed: 24811454]

17. Crawley AB, Henriksen ED, Stout E, Brandt K, Barrangou R. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. Scientific Reports. 2018; 8 11544 [PubMed: 30068963]

18. Deltcheva E, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011; 471: 602–607. [PubMed: 21455174]

19. McGinn J, Marraffini LA. CRISPR-Cas systems optimize their immune response by specifying the site of spacer integration. Mol Cell. 2016; 64: 616–623. [PubMed: 27618488]

20. Martynov A, Severinov K, Ispolatov I. Optimal number of spacers in CRISPR arrays. PLoS Comput Biol. 2017; 13 e1005891 [PubMed: 29253874]

21. Rao C, Chin D, Ensminger AW. Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. RNA. 2017; 23

22. Liao C, Beisel CL. The tracrRNA in CRISPR biology and technologies. Annu Rev Genet. 2021; 55: 161–181. [PubMed: 34416117]

23. Karvelis T, et al. crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus* . RNA Biol. 2013; 10: 841–851. [PubMed: 23535272]

24. Jinek M, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337: 816–821. [PubMed: 22745249]

25. Pickar-Oliver A, Gersbach CA. The next generation of CRISPR–Cas technologies and applications. Nature Reviews Molecular Cell Biology. 2019; 20: 490–507. [PubMed: 31147612]

26. Bikard D, et al. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res. 2013; 41: 7429–7437. [PubMed: 23761437]

27. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 2013; 31: 233–239. [PubMed: 23360965]

28. Citorik RJ, Mimee M, Lu TK. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat Biotechnol. 2014; 32: 1141–1145. [PubMed: 25240928]

29. Leenay RT, Beisel CL. Deciphering, communicating, and engineering the CRISPR PAM. J Mol Biol. 2017; 429: 177–191. [PubMed: 27916599]

30. Dugar G, et al. CRISPR RNA-dependent binding and cleavage of endogenous RNAs by the *Campylobacter jejuni* Cas9. Mol Cell. 2018; 69: 893–905. e7 [PubMed: 29499139]

31. Xue C, et al. CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Res. 2015; 43: 10831–10847. [PubMed: 26586800]

32. Collias D, et al. A positive, growth-based PAM screen identifies noncanonical motifs recognized by the Cas9. Sci Adv. 2020; 6 eabb4054 [PubMed: 32832642]

33. Altuvia Y, et al. *In vivo* cleavage rules and target repertoire of RNase III in *Escherichia coli* . Nucleic Acids Res. 2018; 46: 10530–10531. [PubMed: 30184218]

34. Wei Y, Chesne MT, Terns RM, Terns MP. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus* . Nucleic Acids Res. 2015; 43: 1749–1758. [PubMed: 25589547]

35. Pougach K, et al. Transcription, processing and function of CRISPR cassettes in *Escherichia coli* . Mol Microbiol. 2010; 77: 1367–1379. [PubMed: 20624226]

36. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli* . Nucleic Acids Res. 2012; 40: 5569–5576. [PubMed: 22402487]

37. Jiao C, et al. Noncanonical crRNAs derived from host transcripts enable multiplexable RNA detection by Cas9. Science. 2021; doi: 10.1126/science.abe7106

38. Jabbari H, Wark I, Montemagno C. RNA secondary structure prediction with pseudoknots: Contribution of algorithm versus energy model. PLoS One. 2018; 13 e0194583 [PubMed: 29621250]

39. Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev. 2015; 29: 356–361. [PubMed: 25691466]

40. Laanto E, Hoikkala V, Ravantti J, Sundberg L-R. Long-term genomic coevolution of host-parasite interaction in the natural environment. Nat Commun. 2017; 8

41. Zhang Y, et al. Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis* . Mol Cell. 2013; 50: 488–503. [PubMed: 23706818]

42. Dugar G, et al. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. PLoS Genet. 2013; 9 e1003495 [PubMed: 23696746]

43. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science. 2010; 329: 1355–1358. [PubMed: 20829488]

44. Li R, Bowerman B. Symmetry breaking in biology. Cold Spring Harb Perspect Biol. 2010; 2 a003475 [PubMed: 20300216]

45. McCarty NS, Graham AE, Studená L, Ledesma-Amaro R. Multiplexed CRISPR technologies for gene editing and transcriptional regulation. Nat Commun. 2020; 11 1281 [PubMed: 32152313]

46. Al-Hashimi HM, Walter NG. RNA dynamics: it is about time. Curr Opin Struct Biol. 2008; 18: 321–329. [PubMed: 18547802]

47. Watters KE, Strobel EJ, Yu AM, Lis JT, Lucks JB. Cotranscriptional folding of a riboswitch at nucleotide resolution. Nat Struct Mol Biol. 2016; 23: 1124–1131. [PubMed: 27798597]

48. Liao C, et al. Modular one-pot assembly of CRISPR arrays enables library generation and reveals factors influencing crRNA biogenesis. Nat Commun. 2019; 10 2948 [PubMed: 31270316]

49. Wimmer F, Beisel CL. CRISPR-Cas systems and the paradox of self-targeting spacers. Front Microbiol. 2019; 10 3078 [PubMed: 32038537]

50. Leenay RT, et al. Genome editing with CRISPR-Cas9 in *Lactobacillus plantarum* revealed that editing outcomes can vary across strains and between methods. Biotechnol J. 2019; 14 e1700583 [PubMed: 30156038]

51. Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Research. 2008; 36: W70–W74. [PubMed: 18424795]

52. Lorenz R, et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology. 2011; 6

53. Sharma CM, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature. 2010; 464: 250–255. [PubMed: 20164839]

54. Papenfort K, et al. σ^E-dependent small RNAs of *Salmonella* respond to membrane stress by accelerating global *omp* mRNA decay. Mol Microbiol. 2006; 62: 1674–1688. [PubMed: 17427289]

55. Pernitzsch SR, Tirier SM, Beier D, Sharma CM. A variable homopolymeric G-repeat defines small RNA-mediated posttranscriptional regulation of a chemotaxis receptor in *Helicobacter pylori* . Proc Natl Acad Sci U S A. 2014; 111: E501–10. [PubMed: 24474799]

56. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011; 17: 10.

57. Förstner KU, Vogel J, Sharma CM. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. Bioinformatics. 2014; 30: 3421–3423. [PubMed: 25123900]

58. Hoffmann S, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009; 5 e1000502 [PubMed: 19750212]

59. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. [PubMed: 19505943]

60. Lopez-Delisle L, et al. pyGenomeTracks: reproducible plots for multivariate genomic data sets. Bioinformatics. 2020; 37: 422–423.

61. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010; 26: 2204–2207. [PubMed: 20639541]

62. Padilha VA, Alkhnbashi OS, Shah SA, de Carvalho ACPLF, Backofen R. CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. Gigascience. 2020; 9

63. Padilha VA, et al. Casboundary: Automated definition of integral Cas cassettes. Bioinformatics. 2020; 37: 1352–1359.

64. Mitrofanov A, et al. CRISPRidentify: identification of CRISPR arrays using machine learning approach. Nucleic Acids Res. 2021; 49: e20. [PubMed: 33290505]

65. Alkhnbashi OS, et al. CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. Bioinformatics. 2014; 30: i489–96. [PubMed: 25161238]

66. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the nextgeneration sequencing data. Bioinformatics. 2012; 28: 3150–3152. [PubMed: 23060610]

67. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. 2003; 31: 7280–7301. [PubMed: 14654704]

68. Altschul SF, Erickson BW. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol. 1985; 2: 526–538. [PubMed: 3870875]

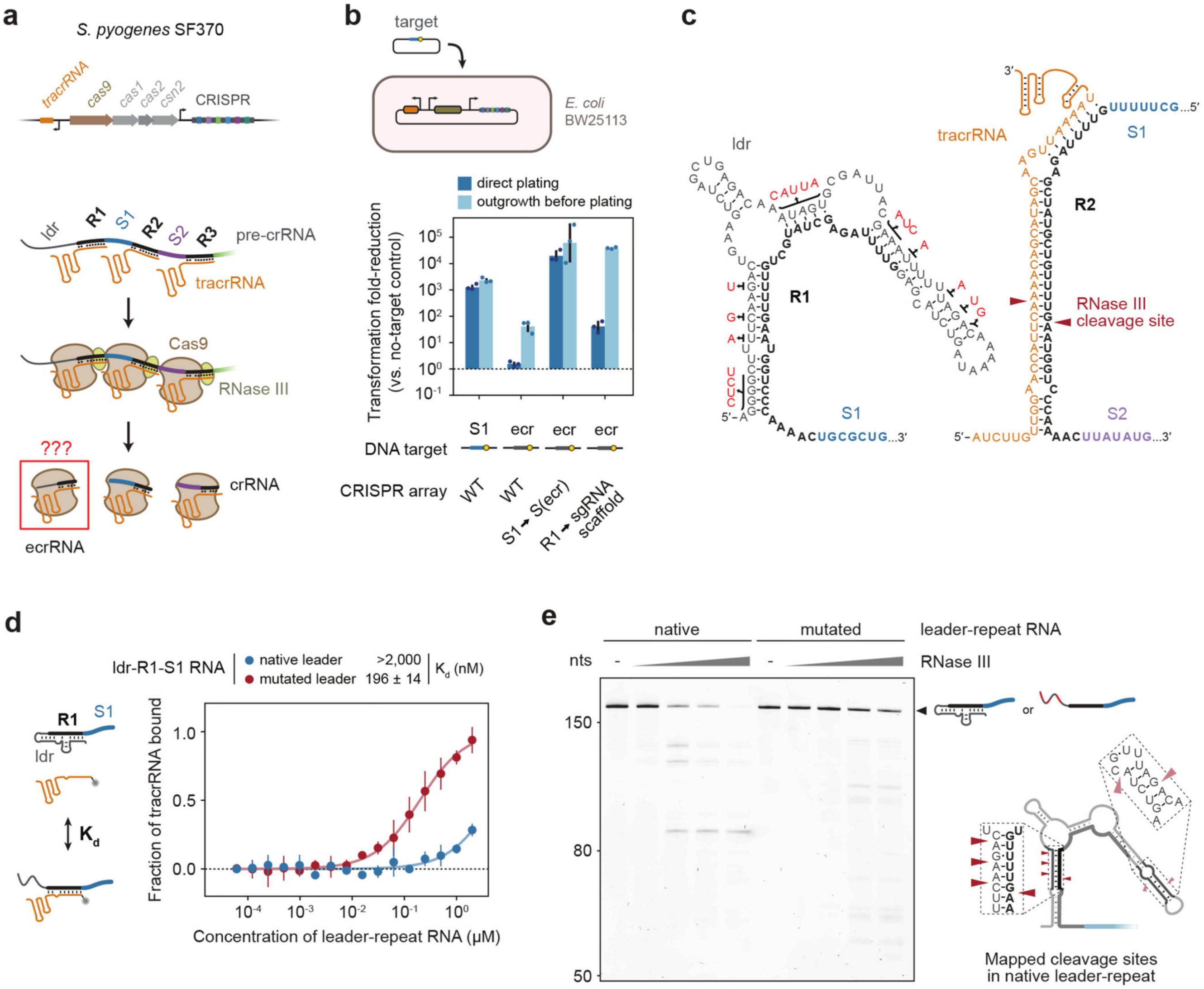


Figure 1. The pre-crRNA from the CRISPR-Cas system in *S. pyogenes* forms a stem-loop between the leader RNA and first repeat that interferes with extraneous crRNA function.

a, The CRISPR-Cas system from *S. pyogenes* and the process of generating crRNAs. The first repeat could give rise to an extraneous crRNA (ecrRNA) from the pre-crRNA. ldr: leader RNA. R: conserved repeat. S: invader-derived spacer. See Extended Data Figure 1a for the annotated sequence of the CRISPR array. **b**, Measured plasmid clearance by the leader-encoded ecrRNA in *E. coli*. Clearance can be improved by including an outgrowth lacking selection for the target plasmid prior to plating. WT: CRISPR array from *S. pyogenes* with the native leader. One tested construct encoded a single-spacer array with the native leader and the spacer derived from the ecrRNA (S(ecr)), effectively replacing the first spacer (S1) with this spacer. Another construct replaced the first repeat (R1) of the CRISPR array with a fused version of the processed repeat:tracrRNA (sgRNA scaffold), thereby creating an sgRNA with an elongated 5′ end comprising the leader RNA. The target (blue bar) is flanked by a recognized PAM (yellow circle). Values represent the geometric mean and standard deviation from independent experiments starting from three separate colonies. **c**, Predicted secondary structure of the leader-repeat for the CRISPR-Cas9 system from *S. pyogenes*. See Extended Data Figure Figure 1c for base-pairing probabilities. Mutations indicated in red were made to disrupt stems formed between the leader RNA and the first repeat. Pairing between the second repeat and the tracrRNA is provided as a basis of comparison. Red arrows indicate the established RNase III cleavage site. **d**, Measured equilibrium binding affinity between the leader-repeat RNA and the tracrRNA under *in vitro* conditions. See Extended Data Figure 2a for additional data. We consider the difference in binding affinities to be smaller than that *in vivo* due to co-transcriptional folding, RNase III processing, and standard RNA turnover. Values represent the mean and standard deviation of triplicate independent measurements. **e**, RNase III cleavage of the native and mutated leader-repeat RNA *in vitro*. RNAs were stained with SYBR Green II. Right: Preferred (red arrows) and less-preferred (light red arrows) sites of RNase III cleavage within the native leader-repeat RNA. See Extended Data Figure 1d-e for the mapped secondary structure and RNase III cleavage sites. Results are representative of duplicate independent experiments.

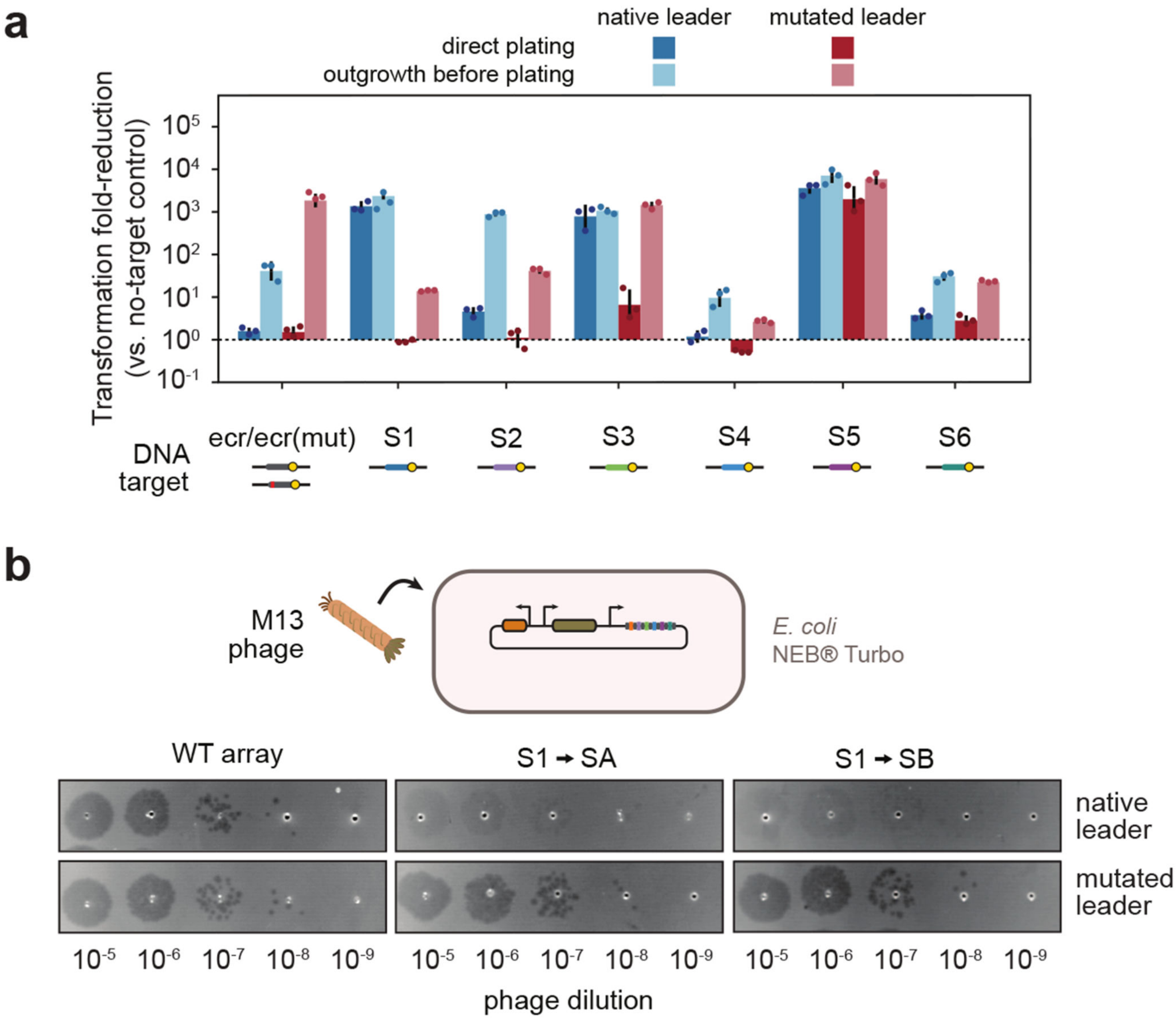


Figure 2. Disrupting the leader-repeat stem-loop impairs immune defense through the newest CRISPR spacers.

a, Impact of disrupting the stem-loop in the *S. pyogenes* CRISPR-Cas9 system on plasmid clearance in *E. coli*. The clearance assays were conducted with or without non-selective outgrowth, where the outgrowth enhances clearance. ecr(mut): the DNA target of the ecrRNA mutated to match the sequence in the mutated leader RNA. The mutated leader is the same as shown in Figure 1c. Data for the native leader with the ecr and S1 targets are the same as those in Figure 1c. The guides for ecr and ecr(mut) yield the same plasmid clearance activity with their cognate target in the context of a single-spacer array (see Extended Data Figure 3a). Values represent the geometric mean and standard deviation from independent experiments starting from three separate colonies. **b**, Impact of mutating the leader on defense against M13 phage. The first spacer in the array was replaced with the SA or SB spacer, each targeting the M13 genome. Visible plaques indicate successful infection by the phage. Results are representative of triplicate independent experiments.

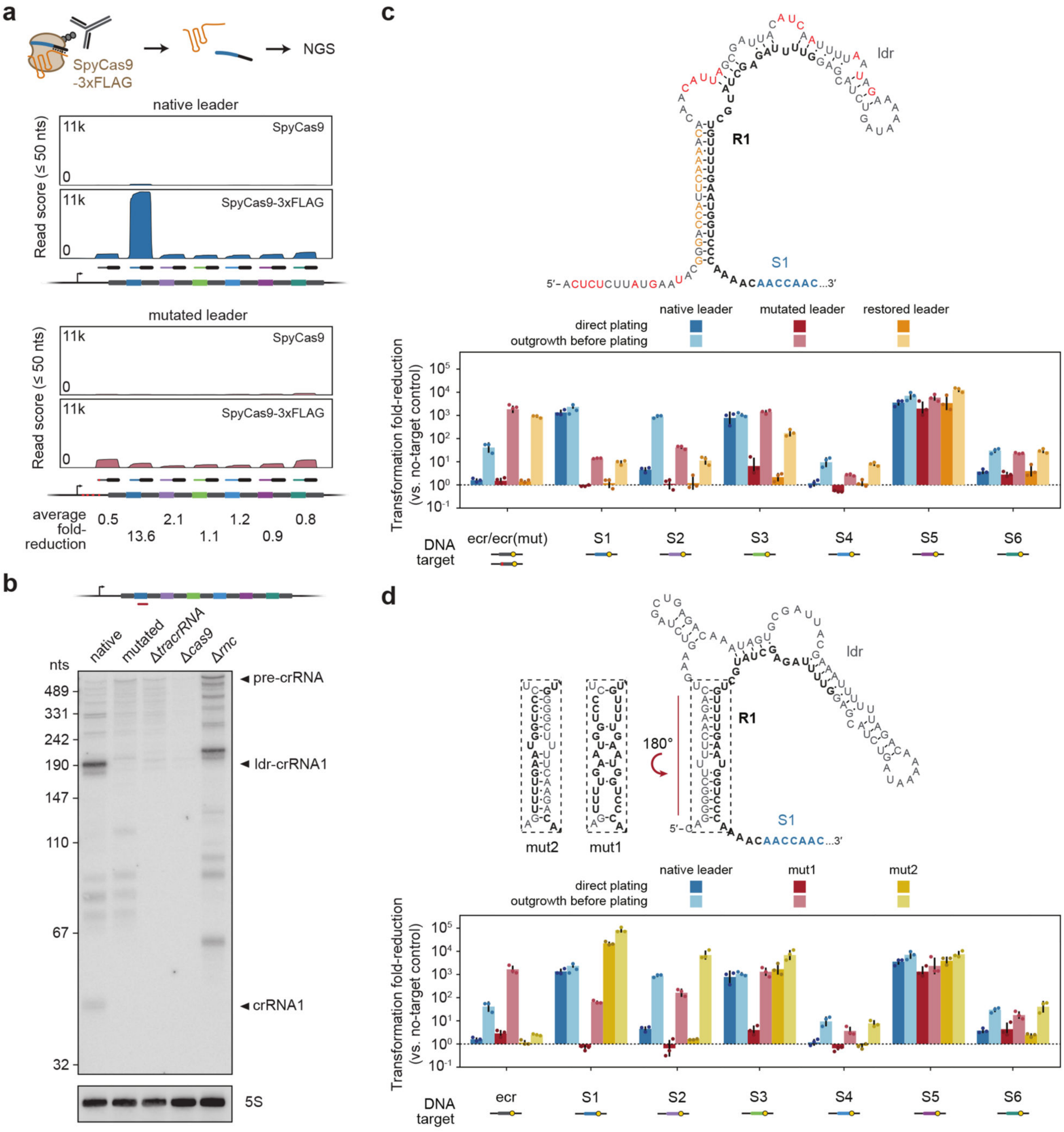


Figure 3. The leader-repeat stem-loop is important for the increased abundance and enhanced processing of the crRNA derived from the newest spacer.

a, RNA co-immunoprecipitation and sequencing (RIP-seq) analysis from the *S. pyogenes* CRISPR-Cas9 system expressed in *E. coli*. Adapter-trimmed reads representing processed products for the ecrRNA and crRNAs were mapped. The ratio of normalized read counts between the mutated leader and native leader for the ecrRNA and each crRNA are indicated below the plots. See Extended Data Figure 4 for additional analyses. Extended Data Figure 4d is a rescaled version of the plot, while Extended Data Figure 4c plots the same data without size filtering. Results are representative of duplicate independent experiments. **b**, Northern blotting analysis of the pre-crRNA from the *S. pyogenes* CRISPR-Cas9 system expressed in *E. coli*. All gene deletion mutants utilize the native leader. The RNA was probed through the first spacer. ldr-crRNA1: RNA corresponding to the leader through the processed second repeat. A similar RNA species was observed as part of RIP-seq analyses (see Extended Data Figure 4). Results are representative of duplicate independent experiments. **c**, Impact of mutations intended to restore the central stem of the leader-repeat stem-loop on plasmid clearance by SpyCas9 in *E. coli*. Top: predicted secondary structure of the mutated leader-repeat with additional mutations to restore the central stem of the leader-repeat stem-loop. Red letters correspond to the mutated nts in the mutated leader in Figure 1c. Yellow letters correspond to the nts that were mutated to restore the central stem. Bottom: impact of mutations on plasmid clearance. Mutations in red correspond to the mutated leader, while mutations in red and yellow correspond to the restored leader. ecr(mut): the DNA target of the ecrRNA mutated to match the sequence in the mutated leader RNA. Results for two additional sets of restoring mutations are located in Extended Data Figure 5c-d. **d**, Impact of inverting the central stem of the leader-repeat stem-loop. Top: Inversion of the central stem. mut1: inverting only the leader. mut2: inverting both the leader and repeat. Bottom: impact of inverting the central stem on plasmid clearance. Values in c and d represent the geometric mean and standard deviation from independent experiments starting from three separate colonies.

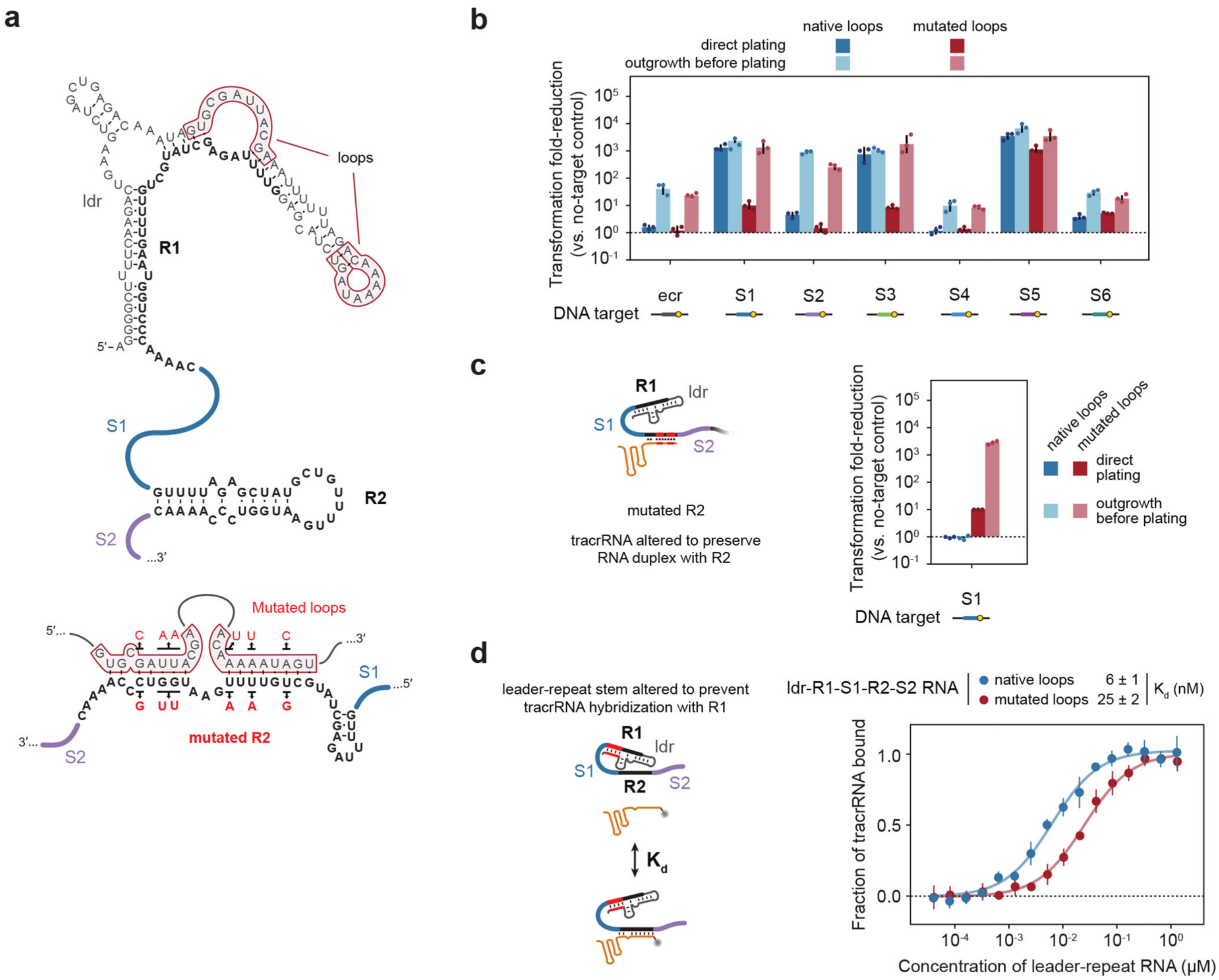


Figure 4. An interaction between the leader-repeat stem-loop and the second repeat promotes tracrRNA hybridization to the second repeat.

a, Predicted interaction between the leader-repeat stem-loop and the second repeat in the pre-crRNA. The second repeat is predicted to fold into a hairpin, while the predicted interactions with the two loops disrupt this hairpin. Red base pairs indicate mutations made in b-d to mutate the loops, the second repeat, or both. **b**, Impact of mutating both loops of the stem-loop on plasmid clearance through the ecrRNA and crRNAs. Results for the native leader are the same as those in Figure 2a. See Figure 2a for more information. **c**, Impact of mutating the second repeat to restore the predicted interactions with the mutated loops on plasmid clearance through the first spacer. The tracrRNA was mutated to maintain the crRNA:tracrRNA duplex. The other crRNAs were not tested because the mutations in the anti-repeat portion of the tracrRNA would prevent efficient hybridization to the corresponding repeats. Values in c and d represent the geometric mean and standard deviation from independent experiments starting from three separate colonies. **d**, Measured equilibrium binding affinity between the tracrRNA and the RNA spanning the leader through the beginning of the second spacer under *in vitro* conditions. The leader-repeat stem was mutated to prevent hybridization between the first repeat and the tracrRNA. See Extended Data Figure 2b for additional data. Results are representative of triplicate independent measurements.

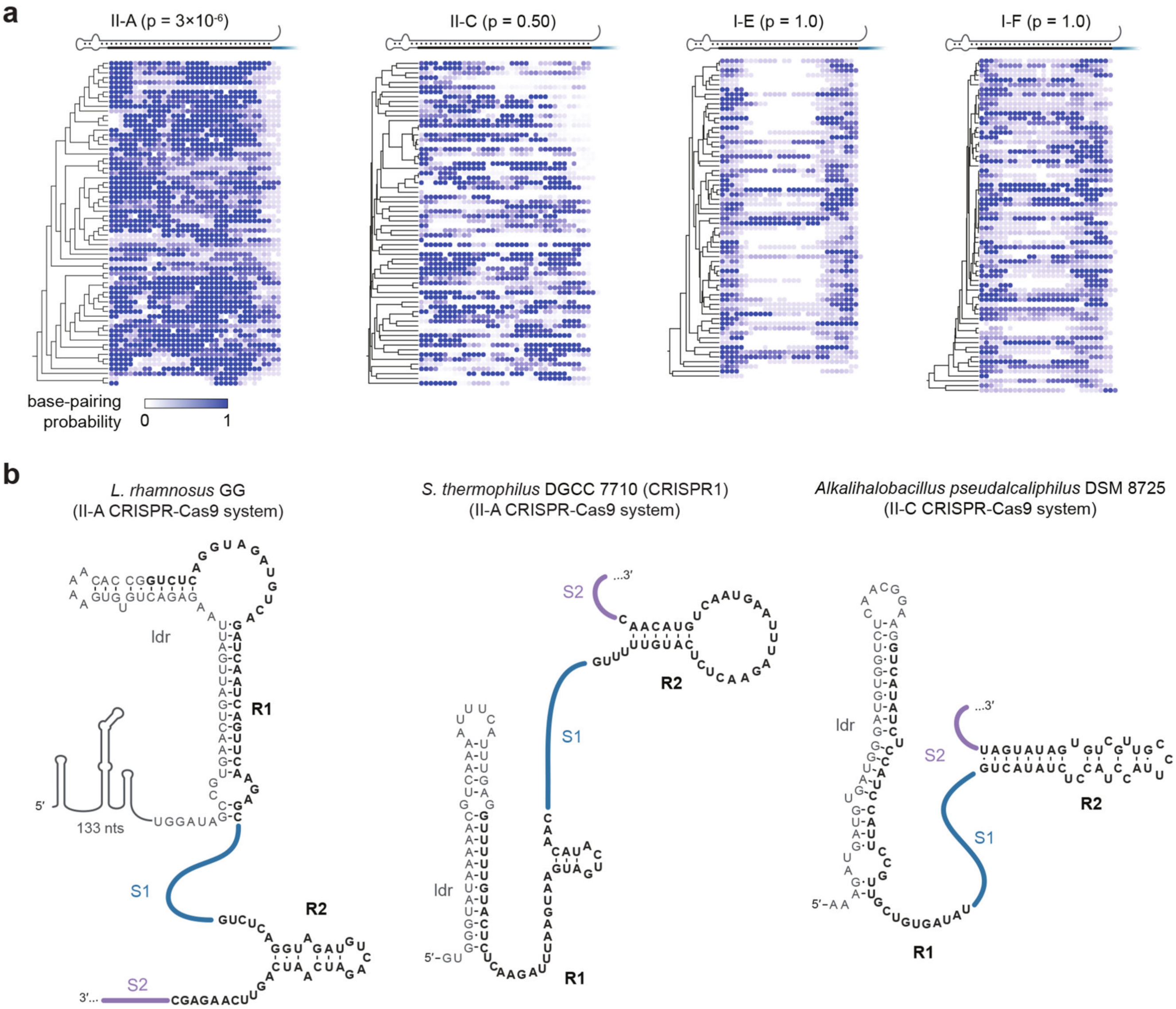

a
II-A (p = 3×10-6)
II-C (p = 0.50)
I-E (p = 1.0)
I-F (p = 1.0)
base-pairing probability
0
1
b
L. rhamnosus GG
(II-A CRISPR-Cas9 system)
S. thermophilus DGCC 7710 (CRISPR1)
(II-A CRISPR-Cas9 system)
Alkalihalobacillus pseudalcaliphilus DSM 8725
(II-C CRISPR-Cas9 system)
ldr
R1
R2
S1
S2
5′
3′...
...3′
133 nts
5′-GU
5′-AA

Figure 5. A stem-loop formed between the leader RNA and first repeat is found across CRISPR-Cas9 systems.

a, Interactions predicted between the first repeat and upstream sequence within different CRISPR-Cas subtypes. Trees depict similarity between repeat sequences. Blue circles represent the extent of base pairing by the corresponding nucleotide in the first repeat with the leader RNA at thermodynamic equilibrium. Stated p-values reflect the statistical significance of a stem-loop formed between each first repeat and upstream sequence in a subset of CRISPR-Cas systems that we had not previously analyzed. The aggregate p-values for the I-E and I-F systems were both 1.0 because many I-E and I-F CRISPR-Cas systems exhibited weak stem-loops. Empirical p-values were calculated using randomly shuffled leader sequences (n = 1,000) and then combined into a single p-value using Fisher's Method. **b**, Predicted structures of the leader-repeat stem-loop and the second repeat from representative II-A and II-C systems. The structures were predicted using NUPACK. In the case of *L. rhamnosus* GG and *S. thermophilus* DGCC 7710 (CRISPR1), the leader-repeat structures were confirmed by *in vitro* structural probing and shown to block tracrRNA binding and undergo processing by RNase III (Extended Data Figure 8).

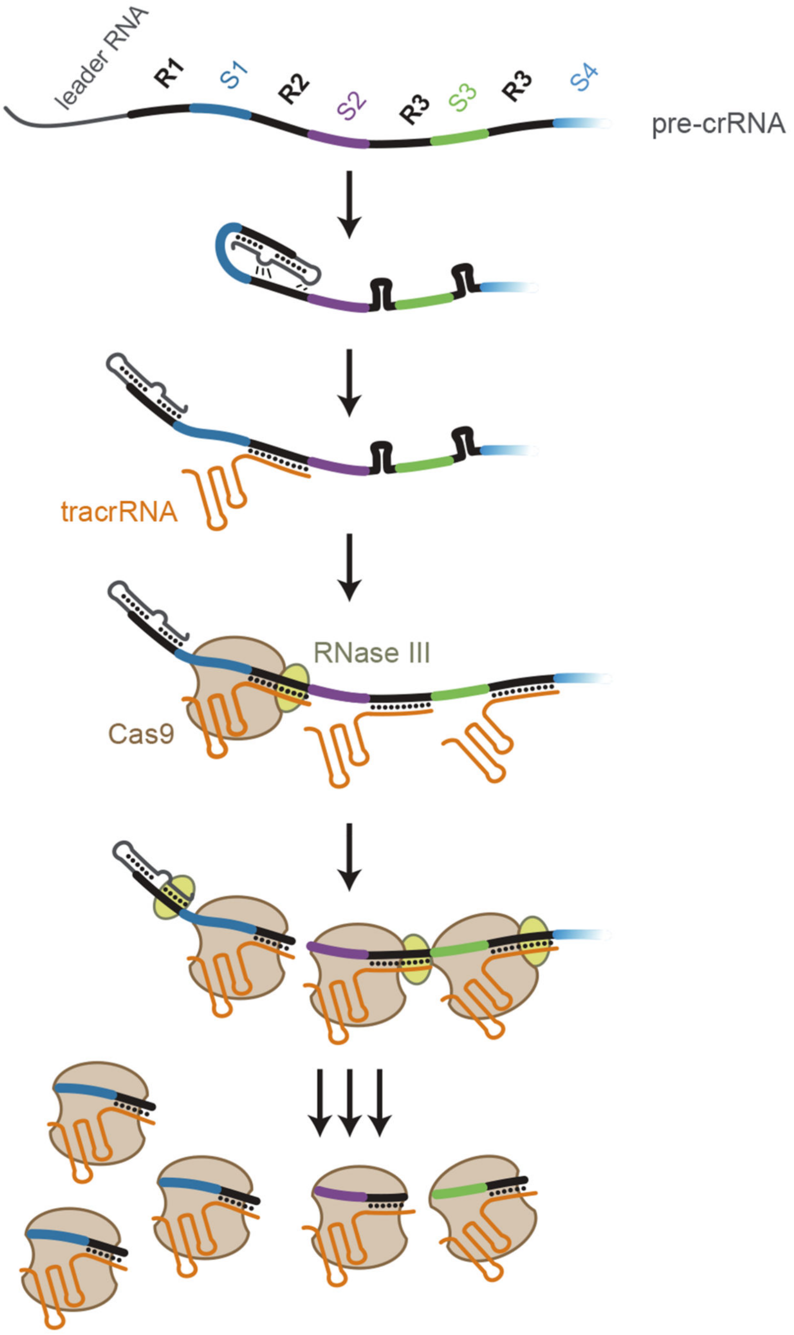


Figure 6. Proposed model for the role of the leader region in prioritizing crRNA biogenesis associated with the newest spacer for CRISPR-Cas9 systems.

The transcribed leader RNA forms a stem-loop with the first repeat (R1) that interacts with the second repeat (R2). The transient structure promotes hybridization of the tracrRNA to the second repeat, potentially by disrupting a predicted hairpin formed by each repeat. The repeat:tracrRNA duplex then undergoes processing by RNase III and binding by Cas9. The stem-loop formed between the leader and first repeat later undergoes tracrRNA-independent processing by RNase III to yield a mature crRNA derived from the newest spacer (S1). The tracrRNA eventually hybridizes with the other repeats, leading to mature crRNA derived from the other spacers.