ARTICLES
https://doi.org/10.1038/s41564-022-01074-3
nature microbiology

Check for updates

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

Chunyu Liao[1], Sahil Sharma[2,8], Sarah L. Svensson[2,8], Anuja Kibe[1,8], Zasha Weinberg[3,8], Omer S. Alkhnbashi[4], Thorsten Bischler[5], Rolf Backofen[4,6], Neva Caliskan[1,7], Cynthia M. Sharma[2] and Chase L. Beisel[1,7] ✉

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 most recent are known to be prioritized for immune defence. 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 defence 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 transactivating crRNA (tracrRNA) hybridization with the second of the two repeats, accelerating crRNA processing. Accordingly, disruption of this structure reduces the abundance of the associated crRNA and immune defence 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 defence against the most recently encountered invaders.

Adaptive immune systems possess the ability to remember previous 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 sit between conserved repeats in a CRISPR array, where new spacers are sequentially added at one end of the array[7–9]. To recall stored memories for immune defence, 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 induces 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 defence through the most recently acquired spacers. RNA sequencing (RNA-seq) analysis of native CRISPR–Cas systems has repeatedly revealed that the most abundant crRNAs derive from the most recent end of the CRISPR array[15–18]. Separately, defence against a high phage titre was enhanced when moving an anti-phage spacer from the fifth 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 in the array.

A leader-repeat stem-loop interferes with 'extraneous' crRNA generation. Our investigation of spacer prioritization began with the first repeat (R1), 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 antirepeat portion of a transactivating crRNA (tracrRNA)[18,22]. The hybridized repeat–antirepeat 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 the 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 defence. For the CRISPR–Cas9 system from *S. pyogenes*, RNA-seq analysis did not indicate any stable products resembling an ecrRNA[18]. However, RNA-seq analysis of different lactobacilli

[1]Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Centre for Infection Research (HZI), Würzburg, Germany. [2]Department of Molecular Infection Biology II, Institute of Molecular Infection Biology, University of Würzburg, Würzburg, Germany. [3]Bioinformatics Group, Department of Computer Science and Interdisciplinary Centre for Bioinformatics, Leipzig University, Leipzig, Germany. [4]Bioinformatics group, Department of Computer Science, University of Freiburg, Freiburg, Germany. [5]Core Unit Systems Medicine, University of Würzburg, Würzburg, Germany. [6]Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Freiburg, Germany. [7]Medical Faculty, University of Würzburg, Würzburg, Germany. [8]These authors contributed equally: Sahil Sharma, Sarah L. Svensson, Anuja Kibe, Zasha Weinberg. ✉e-mail: Chase.Beisel@helmholtz-hiri.de

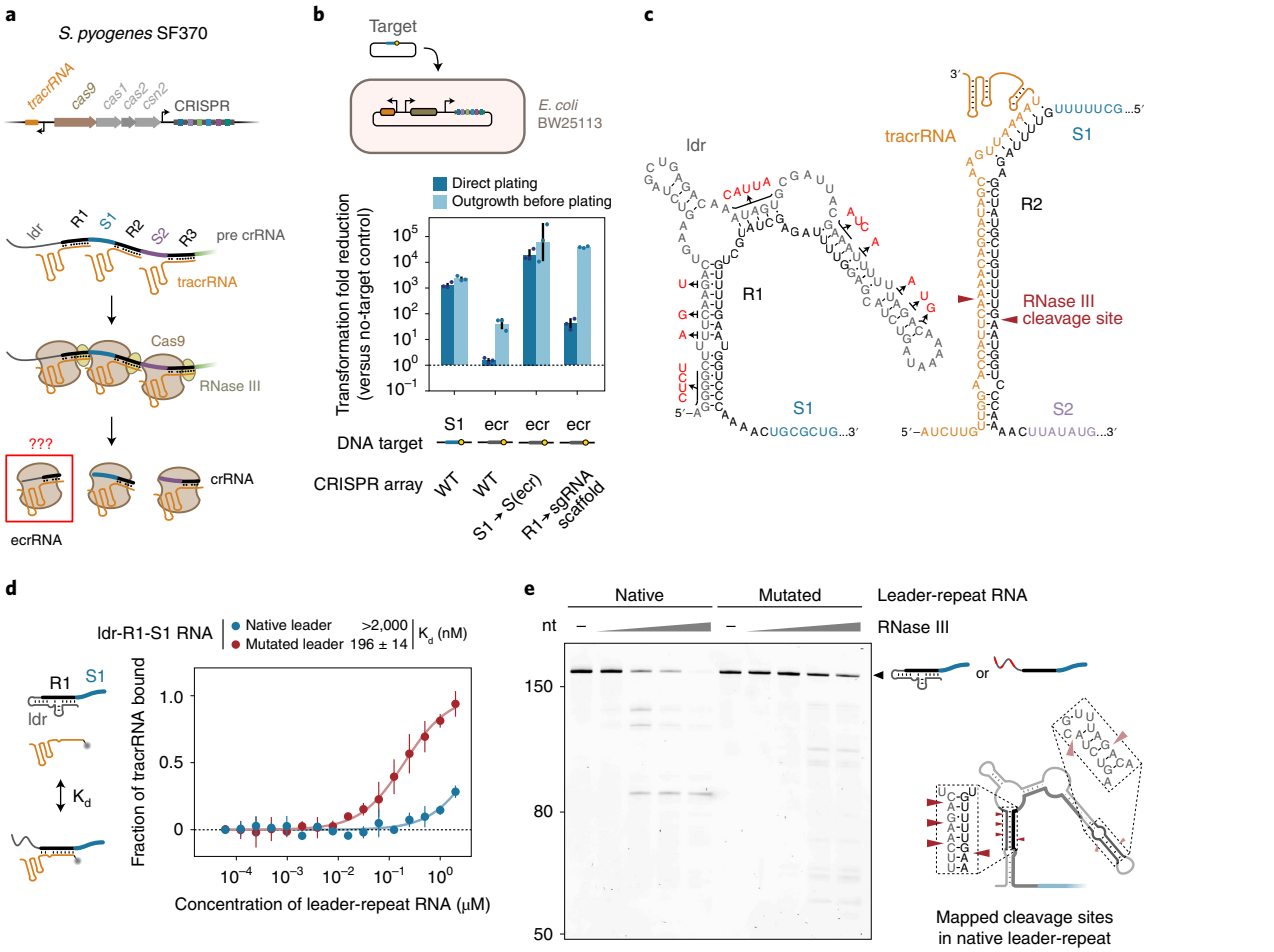


Fig. 1 | pre-crRNA from the CRISPR–Cas system in *S. pyogenes* forms a stem-loop between the leader RNA and R1 that interferes with extraneous crRNA function. a, The CRISPR–Cas system from *S. pyogenes* and the process of generating crRNAs. R1 gives rise to an ecrRNA from the pre-crRNA. ldr, leader RNA; R, conserved repeat; S, invader-derived spacer. See Extended Data Fig. 1a for the annotated sequence of the CRISPR array. **b**, Measured plasmid clearance by the leader-encoded ecrRNA in *E. coli*. Clearance was improved by inclusion of an outgrowth lacking selection for the target plasmid before 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 S1 with this spacer. Another construct replaced 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 Fig. 1c for base-pairing probabilities. Mutations indicated in red were created to disrupt stems formed between the leader RNA and R1. Pairing between second repeat (R2) and the tracrRNA is provided as a basis of comparison. Red arrowheads indicate the established RNase III cleavage site. **d**, Measured equilibrium binding affinity between the leader-repeat RNA and tracrRNA under in vitro conditions (see Extended Data Fig. 2a for additional data). We consider the difference in binding affinities to be smaller than that in vivo due to cotranscriptional folding, RNase III processing and standard RNA turnover. Values represent the mean and standard deviation of triplicate independent measurements. **e**, RNase III cleavage of native and mutated leader-repeat RNA in vitro. RNAs were stained with SYBR Green II. Right, preferred (dark red arrowheads) and less-preferred (light red arrowheads) sites of RNase III cleavage within the native leader-repeat RNA (see Extended Data Fig. 1d,e for the mapped secondary structure and RNase III cleavage sites). Results are representative of duplicate independent experiments.

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] (Fig. 1a and 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 nonselective outgrowth before plating (Fig. 1b). The outgrowth step grants more time before 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 by direct plating (1.6-fold) compared with a nontarget control. In contrast, the same plasmid encoding the crRNA1 target (that is, matching the first spacer, S1) was efficiently cleared with direct plating (1,300-fold). The ecrRNA guide sequence was not the culprit, because replacement of 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, because replacement of R1 with the single-guide RNA handle to bypass crRNA processing exhibited enhanced plasmid clearance compared with the original ecrRNA (Fig. 1b). Instead, we 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 R1 and the upstream leader (ldr) in the pre-crRNA (Fig. 1c 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 R1 and the tracrRNA, thereby inhibiting ecrRNA biogenesis. In vitro binding measurements between the tracrRNA and an RNA spanning the leader through S1 confirmed that disruption of the stem-loop through leader mutations increased binding affinity by at least tenfold (Fig. 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 R1 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 separation of S1 from its upstream repeat and trimming to a mature crRNA similar to all other spacers in the array[18].

Stem-loop disruption impairs defence by the most recent spacers. We next asked how disruption of the formed stem-loop affects plasmid interference directed by both the ecrRNA and the six encoded crRNAs. Repeating the plasmid clearance assay in *E. coli* (Fig. 2a), we found that mutation of the leader did not affect clearance by the ecrRNA with direct plating but did enhance clearance from 40- to 1,800-fold with outgrowth (comparing ecr and the DNA target of the ecrRNA mutated to match the sequence in the mutated leader RNAecr(mut); Fig. 2a). This enhancement is in line with restored access by the tracrRNA rather than specific mutations to the ecrRNA guide (Extended Data Fig. 3a). Mutation of the leader also had a positional effect on crRNA-mediated plasmid clearance: clearance was heavily disabled for crRNA1 and crRNA2, partially disabled for crRNA3 and crRNA4 and unperturbed for crRNA5 and crRNA6. This result was unexpected, because the leader RNA had been implicated only 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 R1 is critical for immune defence through the adjacent spacers.

To further evaluate the role of the leader-repeat stem-loop in immune defence, we replaced S1 with one of two spacers targeting the genome of the filamentous *E. coli* phage M13. We then evaluated defence 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 defence, indicating that the effect of the leader does not depend on the sequence of S1. These results connect the leader region to antiplasmid and antiviral defence by Cas9 and implicate the leader-repeat stem-loop in promoting defence through the most recent CRISPR spacers. Given the absence of mechanisms to explain spacer prioritization for immune defence[9], we turned our focus from the ecrRNA to the role played by the stem-loop in immune defence.

We asked whether mutation of the leader would disrupt production of crRNAs encoded near the beginning of the array. We therefore evaluated the abundance of Cas9-bound RNAs with the native or mutated leader by immunoprecipitation of Cas9 and sequencing-bound RNAs using RNA immunoprecipitation and sequencing (RIP–seq) (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. Mutation of 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 RNA-blotting analysis using total RNA (Extended Data Fig. 5a,b). The loss of plasmid clearance through crRNA1 can therefore be attributed to the marked reduction in crRNA abundance due to disruption of the leader-repeat stem-loop.

The stem-loop and R2 promote tracrRNA hybridization. The ensuing question is how the leader-repeat stem-loop accounts for enhanced crRNA production from S1. One important insight came from our RIP–seq and RNA-blotting analyses (Extended Data Figs. 4 and 5a). These revealed a stable RNA product of ~190 nt spanning the leader to the RNase III-processing site in second repeat (R2), which was also present when probing for crRNAs in the native *S. pyogenes* strain[18]. This RNA product disappeared after mutation of the leader or removal of Cas9, the tracrRNA or RNase III in an *E. coli* strain harbouring the native leader (Fig. 3b and Extended Data Fig. 5b). The leader-repeat stem-loop was therefore important for processing of R2, the exact repeat associated with crRNA1. Processing appeared to occur through R2 before R1, because the ~190-nt stable RNA product contained an intact R1 and processed R2. 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 most recent spacers (Fig. 3c and Extended Data Fig. 5c,d). However, inversion of the central stem by mutation of the leader and then R1 disrupted and then restored position-dependent plasmid clearance (Fig. 3d). The key difference between these sets of mutations is that inversion of the central stem maintained the remaining stem-loop structure, suggesting that the upper portion of the stem-loop is also important for enhanced crRNA production.

The importance of the upper portion of the leader-repeat stem-loop for efficient processing of R2 could reflect a direct interaction between these physically separate parts of the pre-crRNA. If cotranscriptional folding forms the leader-repeat stem-loop before R2 is transcribed and before tracrRNA can hybridize with R1, then the protruding loops of the stem-loop would be most readily available to interact with R2. Following this logic, in silico folding predicted that the two main protruding loops of the leader-repeat stem-loop can extensively base pair with R2 (Fig. 4a). To test these predictions, we created compensatory mutations in the loops and R2 to disrupt and then reform this interaction while preserving the predicted secondary structure of the leader-repeat stem-loop (Fig. 4a). When mutating R2, the tracrRNA antirepeat was also mutated to maintain the repeat–antirepeat duplex for processing and utilization by Cas9 (ref. [37]). Mutation of the protruding loops disrupted plasmid clearance by crRNA1 by 130-fold under direct plating (Fig. 4b), although clearance was also high with outgrowth. Similarly, mutation of R2 and the tracrRNA fully eliminated any measurable

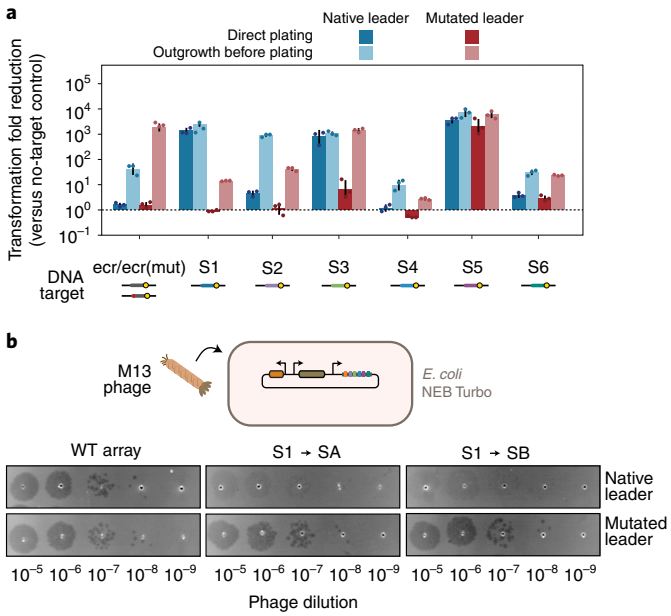


Fig. 2 | Disruption of the leader-repeat stem-loop impairs immune defence through the most recent CRISPR spacers. a, Impact of disrupting the stem-loop in the *S. pyogenes* CRISPR–Cas9 system on plasmid clearance in *E. coli*. Clearance assays were conducted with or without nonselective outgrowth, where outgrowth enhances clearance. The mutated leader is the same as shown in Fig. 1c, as are data for the native leader with the ecr and S1 targets. 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 (Extended Data Fig. 3a). Values represent the geometric mean and standard deviation from independent experiments starting from three separate colonies. **b**, Impact of mutating the leader on defence against the M13 phage. S1 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.

clearance, even with nonselective outgrowth (Fig. 4c). Importantly, re-establishment of the predicted interactions by mutation of the loops and R2 restored plasmid clearance partially with direct plating (tenfold), and fully with nonselective outgrowth (2,900-fold) (Fig. 4c). The leader-repeat stem-loop therefore appears to interact with R2, which promotes immune defence through the most recent spacer.

The interaction between the leader-repeat stem-loop and R2 raises this question: how could this interaction promote processing of R2? If anything, the interaction would interfere with tracrRNA binding by sequestration of at least a portion of R2. 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). Because the predicted interaction between the leader-repeat stem-loop and R2 and the predicted internal hairpin of R2 are mutually exclusive (Fig. 4a), this interaction could disrupt the internal hairpin and promote hybridization with 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 (S2) (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 R2. Mutation of the two protruding loops of the leader-repeat stem-loop reduced binding between R2 and tracrRNA, by 4.2-fold. From these results, we conclude that the interaction between the leader-repeat and R2 promotes preferential hybridization of the tracrRNA to R2, thereby prioritizing biogenesis of the crRNA derived from the most recent spacer.

Leader-repeat stem-loops found across CRISPR–Cas9 systems. Given the role of the leader-repeat stem-loop in prioritizing immune defence 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 leader-repeat stem-loop and R2 for the *S. pyogenes* system are probably weaker, transient and dependent on cotranscriptional folding and thus difficult to predict[38]. However, the extensive stem formed between the leader RNA and R1 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 R1 and the upstream 180 nt. We found numerous arrays with extensive predicted base pairing between R1 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 ($P = 3 \times 10^{-6}$, 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 R1, 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. Because 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 in the promotion of immune defence through the most recent 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. Apart from 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 R1 and the 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 R1 and the upstream sequence (Fig. 5b and Extended Data Fig. 10). A stem-loop between R1 and the upstream sequence can therefore 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 and I-F) within the abundant type I CRISPR–Cas

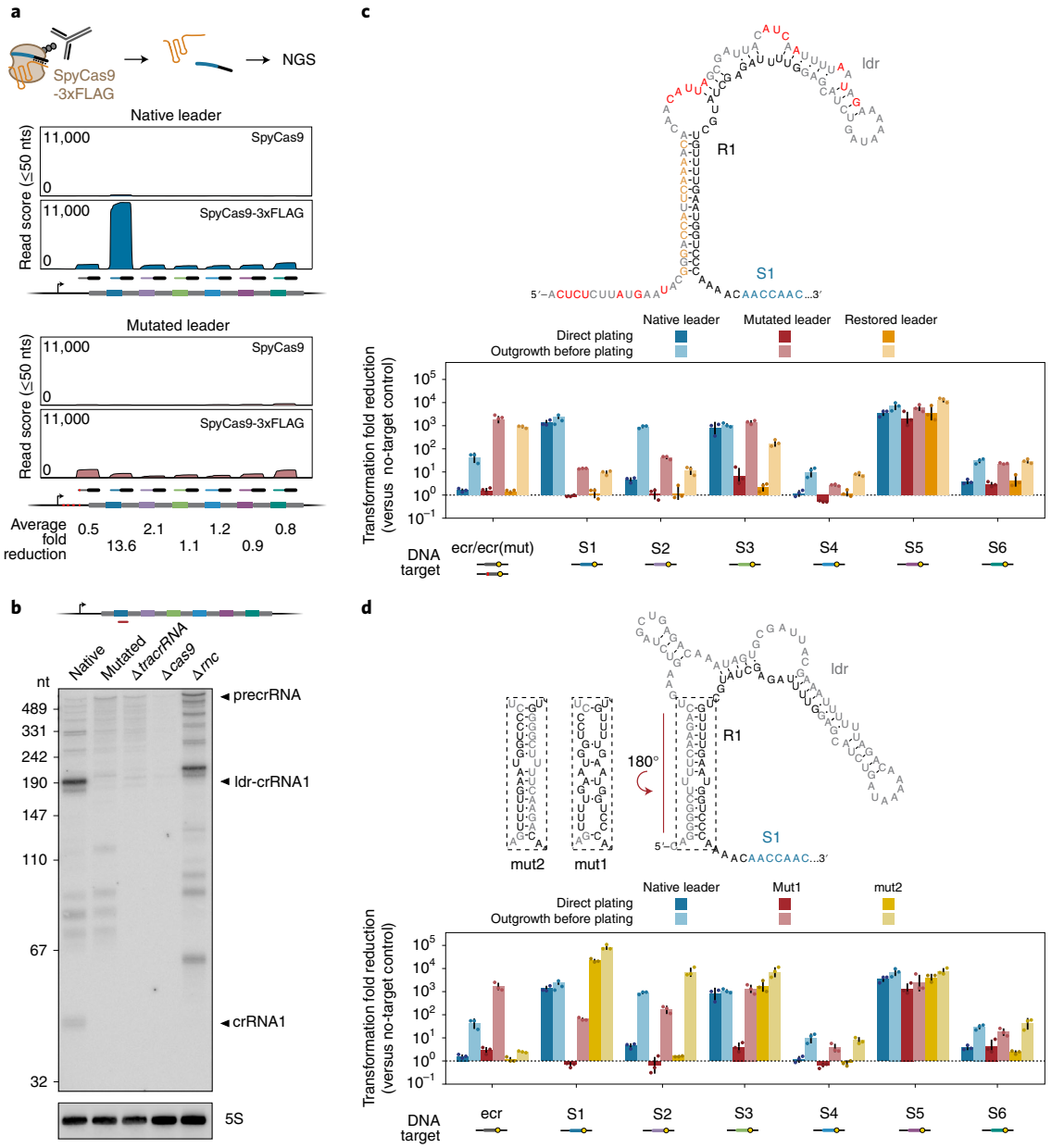


Fig. 3 | The leader-repeat stem-loop is important for the increased abundance and enhanced processing of the crRNA derived from the most recent spacer. a, RIP–seq analysis from the *S. pyogenes* CRISPR–Cas9 system expressed in *E. coli*. Adapter-trimmed reads representing processed products for the ecrRNA and crRNAs were mapped. The ratio of normalized read counts between the mutated leader and native leader for the ecrRNA and each crRNA is indicated below the plots (see Extended Data Fig. 4 for additional analyses). Extended Data Fig. 4d shows a rescaled version of the plot, while Extended Data Fig. 4c plots the same data without size filtering. Results are representative of duplicate independent experiments. **b**, RNA-blotting analysis of pre-crRNA from the *S. pyogenes* CRISPR–Cas9 system expressed in *E. coli*. All gene deletion mutants utilize the native leader. RNA was probed through S1. ldr-crRNA1, RNA corresponding to the leader through the processed R2. A similar RNA species was observed as part of RIP–seq analyses (Extended Data Fig. 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 denotes mutated nucleotides in the mutated leader in Fig. 1c; yellow denotes nucleotides that were mutated to restore the central stem. Bottom: impact of mutations on plasmid clearance. Mutations in red correspond to the mutated leader, while those in red and yellow correspond to the restored leader. Results for two additional sets of restoring mutations can be found in Extended Data Fig. 5c,d. **d**, Impact of inversion of the central stem of the leader-repeat stem-loop. Top: inversion of the central stem. mut1, inversion of the leader only; mut2, inversion of both leader and repeat. Bottom: impact of inversion of the central stem on plasmid clearance. **c**,**d**, Values represent the geometric mean and standard deviation from independent experiments starting from three separate colonies.

systems (Fig. 5a). These systems also acquire spacers through R1 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 R1 contributes the 5′ end of crRNA1 required for effector complex formation[10,43]. Therefore, a leader-repeat stem-loop would also be counterproductive to

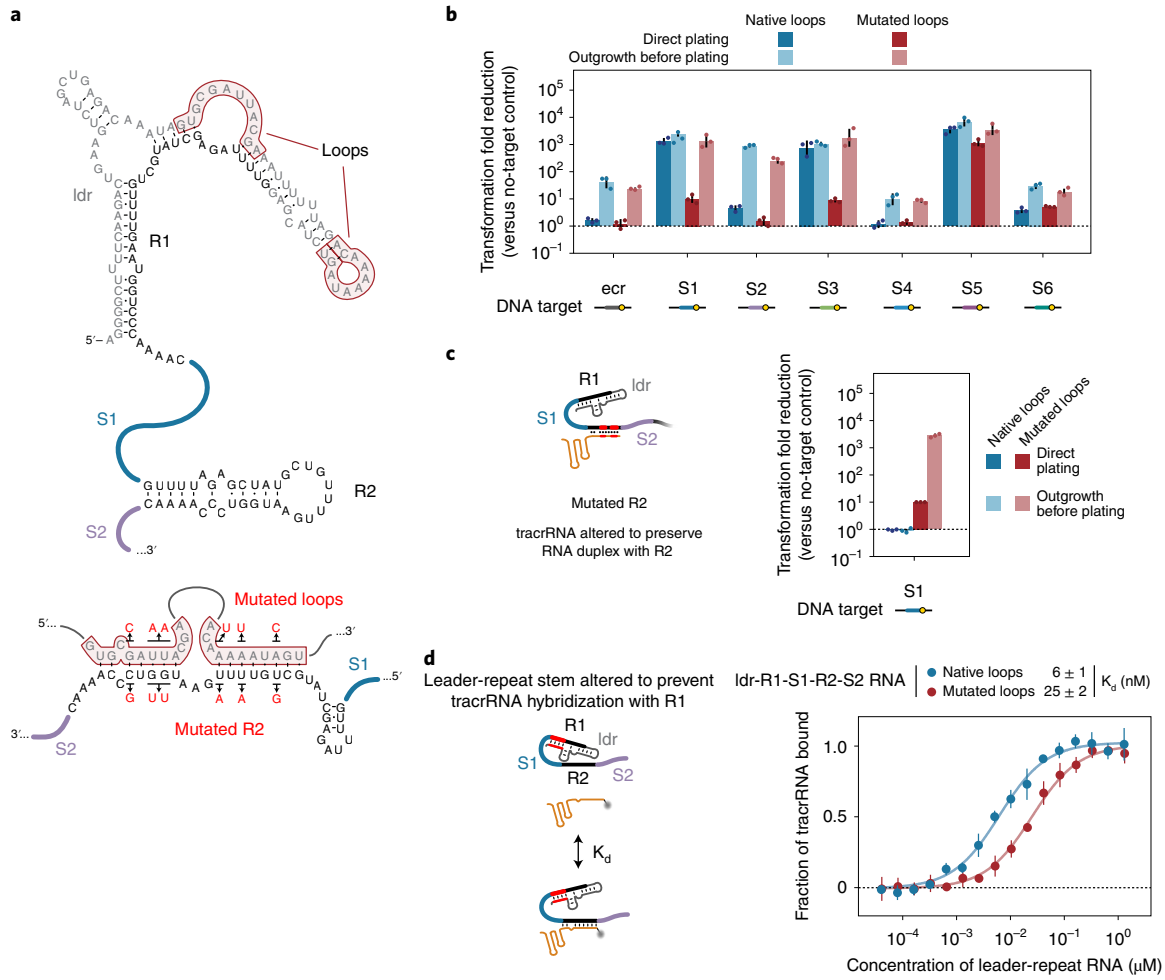


Fig. 4 | Interaction between the leader-repeat stem-loop and R2 promotes tracrRNA hybridization to R2. a, Predicted interaction between the leader-repeat stem-loop and R2 in the pre-crRNA. R2 is predicted to fold into a hairpin, while the predicted interactions with the two loops disrupt this hairpin. Red base pairs indicate mutations made in **b–d** to mutate either the loops, R2 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 Fig. 2a. **c**, Impact of mutating R2 to restore predicted interactions with the mutated loops on plasmid clearance through S1. The tracrRNA was mutated to maintain the crRNA–tracrRNA duplex. Other crRNAs were not tested because mutations in the antirepeat portion of the tracrRNA would prevent efficient hybridization to the corresponding repeats. **d**, Measured equilibrium binding affinity between tracrRNA and the RNA spanning the leader through the beginning of S2 in vitro. The leader-repeat stem was mutated to prevent hybridization between R1 and tracrRNA. **c**,**d**, Values represent the geometric mean and standard deviation from independent experiments starting from three separate colonies. See Extended Data Fig. 2b for additional data. Results are representative of triplicate independent measurements.

crRNA1 production and immune defence through the most recent spacer. Accordingly, both subtypes were not predicted to exhibit helix formation between the leader region and R1 more than that expected by chance ($P = 1.0$, Fisher's method; Fig. 5a). Other mechanisms thus may exist to prioritize the most recent spacers for immune defence across CRISPR–Cas immune systems.

Discussion

Through this work, we discovered an RNA-based mechanism that allows some CRISPR–Cas systems to prioritize immune defence against the most recently encountered invaders. As part of the proposed mechanism (Fig. 6), the leader RNA base pairs with R1 to form a stem-loop through cotranscriptional folding. The upper portion of the stem-loop then interacts with R2, 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, this interaction allows the tracrRNA to more readily hybridize with R2, leading to accelerated processing by RNase III and binding by Cas9. Because crRNAs bound to Cas9 are shielded from RNase attack, these appear much more abundant than other crRNAs in the array. After R2 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 harbouring 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 or type I systems, suggesting that other mechanisms underlying spacer prioritization await discovery. Elucidation of 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 R2, a more complex

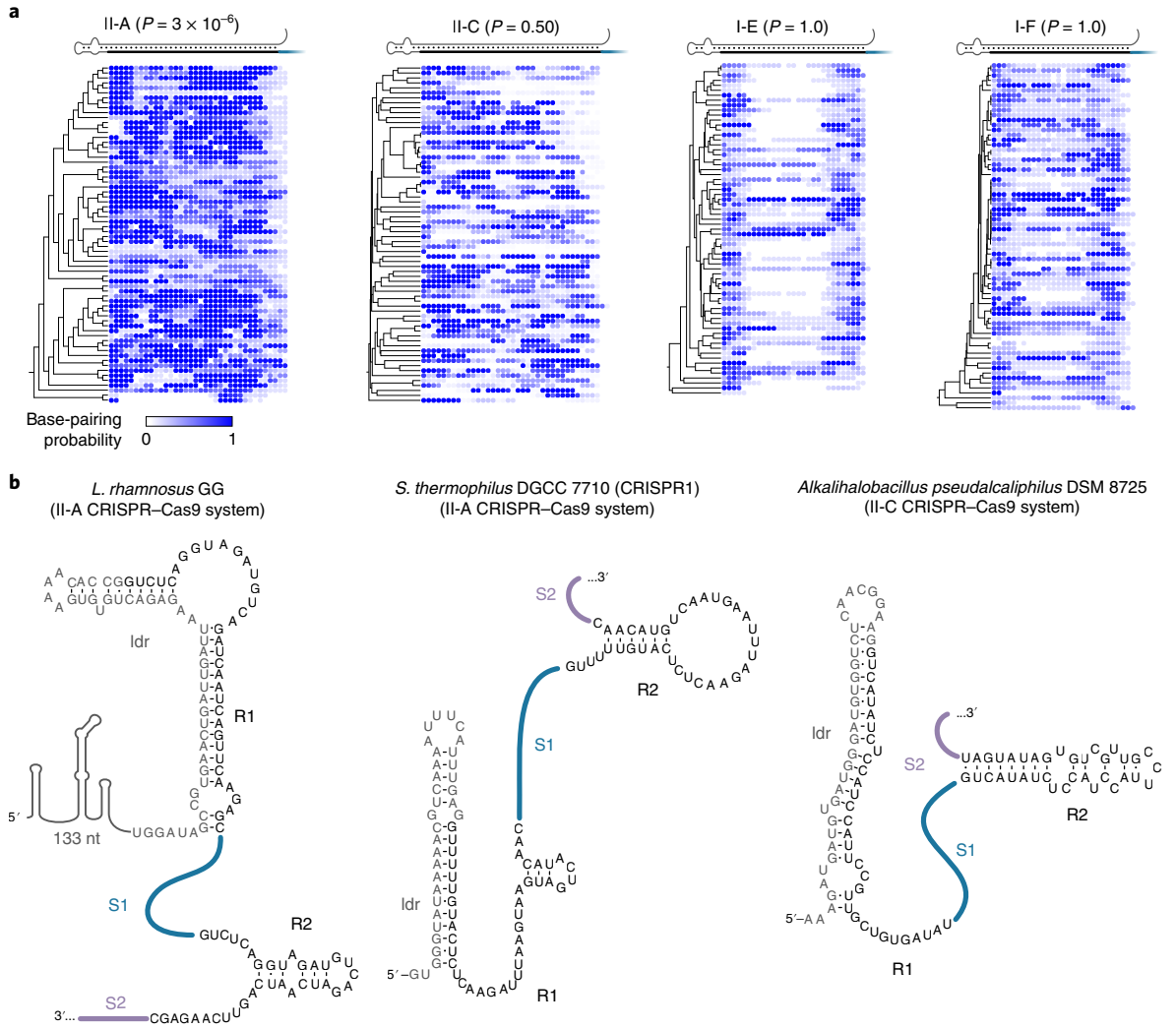


Fig. 5 | A stem-loop formed between the leader RNA and R1 is found across CRISPR–Cas9 systems. a, Interactions predicted between R1 and upstream sequence within different CRISPR–Cas subtypes. Trees depict similarity between repeat sequences. Blue circles represent the extent of base pairing by the corresponding nucleotide in R1 with the leader RNA at thermodynamic equilibrium. Stated *P* values reflect the statistical significance of a stem-loop formed between each R1 and upstream sequence in a subset of CRISPR–Cas systems that we had not previously analysed. Aggregate *P* values for systems I-E and I-F are both 1.0, because many I-E and I-F CRISPR–Cas systems exhibit weak stem-loops within R1. 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 R2 from representative systems II-A and II-C. The structures were predicted using NUPACK. In the case of *L. rhamnosus* GG and *S. thermophilus* DGCC 7710 (CRISPR1) (Extended Data Fig. 6), leader-repeat structures were confirmed by in vitro structural probing and shown to block tracrRNA binding and undergo processing by RNase III (Extended Data Fig. 7 and 8).

structure probably exists and could be the focus of future work. That structure would be expected to depend on the dynamics of transcriptional cofolding in the cellular cytoplasm, where observation of such dynamic structures would be less amenable to approaches such as crystallography or cryogenic electron microscopy. Instead, methods such as time-resolved microscopy using integrated fluorescent probes, single-molecule studies with optical tweezers or in-cell selective 2′-hydroxyl acylation analysed by primer extension and sequencing could help resolve dynamic structures[46,47]. Regardless of the exact structure, interactions between the leader-repeat stem-loop and R2 have multiple implications for spacer prioritization. One implication is that the leader-repeat stem-loop could also interact with repeats downstream of R2—particularly after the tracrRNA hybridizes to this repeat. These longer-range interactions possibly help explain why downstream crRNAs are also negatively impacted by disruption of the leader-repeat stem-loop. Another implication is that the interaction could prevent the most recent spacer from base pairing with R2, thereby removing potential secondary structures 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 that 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 R1 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 to interact with R2. However, it is intriguing that the central stem also blocks ecrRNA formation. Beyond CRISPR–Cas9 systems,

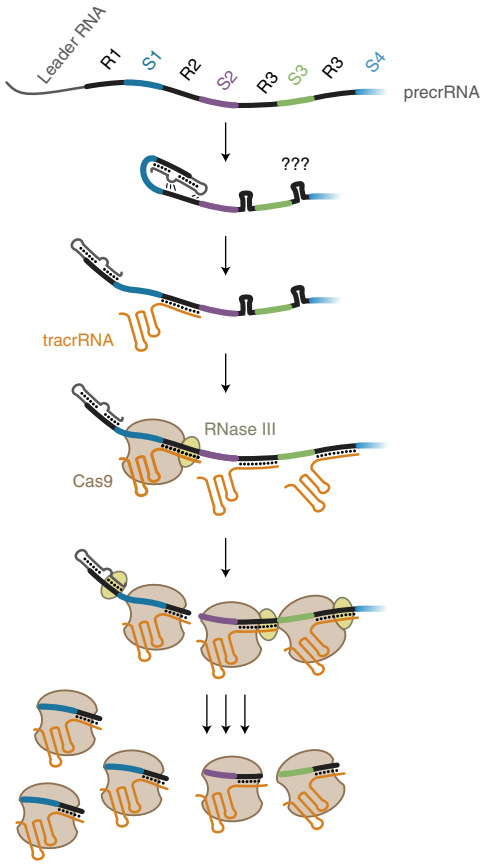


Fig. 6 | Proposed model for the role of the leader region in prioritization of crRNA biogenesis associated with the most recent spacer for CRISPR–Cas9 systems. The transcribed leader RNA forms a stem-loop with R1 that interacts with R2. The transient structure promotes hybridization of the tracrRNA to R2, 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 R1 later undergoes tracrRNA-independent processing by RNase III to yield a mature crRNA derived from the most recent spacer (S1). The tracrRNA eventually hybridizes with the other repeats, leading to mature crRNA derived from the other spacers. R3, R4, third and fourth repeats, respectively; S3, S4, third and fourth spacers, respectively.

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 R1. Because new spacers are acquired through this repeat, mutations that would disrupt ecrRNA formation would also disrupt defence 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 could elucidate the fate of ecrRNAs across CRISPR–Cas systems and whether they provide a hindrance to immune defence 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, and Supplementary Table 4 lists all strains, plasmids, oligonucleotides and gBlocks.

Escherichia coli cells were grown at 37 °C in Luria Bertani (LB) broth (5 g l^{-1} NaCl, 5 g l^{-1} yeast extract, 10 g l^{-1} tryptone) with shaking at 250 r.p.m. or on LB agar plates (LB broth, 18 g l^{-1} agar). Ampicillin and/or kanamycin was added at 50 μg ml^{-1} 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). Chloramphenicol and erythromycin were added at 10 μg ml^{-1} as necessary to maintain any plasmids.

Plasmid pCBS2225 expressing the tracrRNA, SpyCas9 and associated native CRISPR array was constructed by insertion of the corresponding cassette, amplified from the genomic DNA of *S. pyogenes* SF370 using Gibson assembly (New England Biolabs), into backbone plasmid pCB902 following the manufacturer's instructions. Mutations in the leader and/or R1 were introduced through Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions (pCBS2226). R1 was replaced by a sgRNA scaffold using Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions (pCBS2247). For immunoblotting and RIP–seq analyses, the 3×FLAG-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 first by PCR amplification of the gene encoding LrhCas9, along with the 437 upstream base pairs 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 backbone plasmid pCB591 by Gibson Assembly (New England Biolabs) following the manufacturer's instructions. 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 plasmids pCB858 and pCB591 to insert the protospacer and PAM. *E. coli* TOP10 was utilized for the construction of plasmids used in *E. coli*. *L. plantarum* WCFS1 was used as the cloning strain for those plasmids that can be propagated in *L. rhamnosus* but not in *E. coli*.

The plasmids used for interrogation of whether the mutating leader affects transcription of array (nos. pCBS2243 and pCBS2244) were constructed first by PCR amplification of the fragments encoding the native promoter–native/mutated leader from plasmid pCBS2225 or pCBS2226. The resulting PCR product was inserted into 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 S1 on plasmid pCBS2225 or pCBS2226 with the corresponding spacers targeting gene VIII in the genome of the M13 phage through Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions.

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.

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 was disrupted by replacing the portion of leader base pairing with R1 by the corresponding sequence of R1 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 R1 with the corresponding sequence of the leader.

Plasmids with mutated loops and/or R2 were constructed by Q5 mutagenesis (New England Biolabs) following the manufacturer's instructions. The resulting plasmids were used as templates for PCR and Q5 mutagenesis for mutation of 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 first propagated in EC135 before transfer 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^{-1} lysozyme (Carl Roth) in lysozyme buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.1 M NaCl, 5% Triton X-100). After incubation at 37 °C with shaking at 250 r.p.m. for 40 min, cells were pelleted by centrifugation and washed once with water. 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 absorbance reading at 600 nm (ABS_{600}) = ~0.8 in MRS broth with 2% glycine were collected by centrifugation, washed with 10 mM $MgCl_2$ 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 was 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 for *L. plantarum*, except that ampicillin was added to the culture to a final concentration of 10 μg ml^{-1} when ABS_{600} = ~0.2, then cells were pelleted by centrifuging at 4 °C at 5,000*g* for 15 min when 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 the addition of 100 μl of electrocompetent cells and at least 5 μg of plasmid (no more than 5 μl) to a 2-mm-gap cuvette and electroporation 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 either *L. rhamnosus* or *E. coli* harbouring 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 kit (Zymo Research) following the manufacturer's instructions. The resulting RNA was treated with 5′ terminator exonuclease (Epicentre) to degrade processed RNAs following the manufacturer's instructions, purified using the RNA Clean & Concentrator kit and subjected to 5′ rapid amplification of cDNA ends 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 harbouring the plasmid encoding the tracrRNA, SpyCas9 and the array with the native or mutated leader. After recovery for 1 h in super optimal broth with catabolite repression (SOC) (20 $g\,l^{-1}$ tryptone, 5 $g\,l^{-1}$ yeast extract, 3.6 $g\,l^{-1}$ glucose, 0.5 $g\,l^{-1}$ NaCl, 0.186 $g\,l^{-1}$ KCl, 0.952 $g\,l^{-1}$ $MgCl_2$, pH 7.0) at 37 °C with shaking at 250 r.p.m., 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 r.p.m. 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 recovery 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 leader-repeat RNAs and repeat–tracrRNA duplexes was predicted using the online NUPACK algorithm[55,56] (http://www.nupack.org/partition/new). Default parameters were used in addition to the following for folding of 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 inter- and intramolecular base pairing. Interactions between the two protruding loops of the stem-loop and R2 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.

Immunoblotting analysis. As a quality control for coimmunoprecipitation (coIP), a volume of cell culture equivalent to $ABS_{600} = 1.0$ was collected during different stages of 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 immunoblot analysis. Overnight culture of CB414 *E. coli* cells harbouring plasmid pCBS2225, pCBS2226, pCBS2240 or pCBS2241 was back diluted to $ABS_{600} = \sim 0.05$ in LB medium with kanamycin and shaken at 250 r.p.m. at 37 °C to $ABS_{600} = \sim 0.8$. Pelleted cells equivalent to $ABS_{600} = 1.44$ were resuspended in 144 μl of protein loading buffer, boiled at 95 °C for 8 min and stored at −20 °C for immunoblot analysis. Immunoblot analyses were conducted as described previously[30]. Briefly, the resulting samples corresponding to cell $ABS_{600} \sim 0.8$ were resolved on a 10% SDS-polyacrylamide gel, transferred to nitrocellulose 0.45-μM NC membrane (Amersham Protan), blotted using a semidry blotter (VWR), washed with Tris-buffered saline (20 mM Tris, 150 mM NaCl) with 0.1% Tween 20 and visualized on an 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 S2, 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. RNA was resolved on an 8% polyacrylamide gel (20 × 20 cm^2) containing 7 M urea at 300 V for 240 min, stained with SYBR Green II (Biozym), excised and extracted using a 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 respective tracrRNAs were measured by microscale thermophoresis (MST). TracrRNAs 3′-labelled 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 to room temperature on a bench for 10 min, RNAs were serially diluted twofold 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-labelled 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 analysed (MO.Affinity Analysis software v.2.3, NanoTemper Technologies) using the signal from an MST-on time of 20 s for Sth1Cas9-related RNA, 5 s for LrhCas9- and SpyCas9-related RNAs for testing R1 and 1.5 s for SpyCas9-related RNA for testing R2.

In vitro RNase III cleavage assay. In vitro transcribed and purified RNAs were boiled in a thermocycler at 95 °C for 10 min, cooled to room temperature on a bench for 10 min and kept on ice. Cleavage reactions were prepared by the addition of 40 ng of RNA; 1.0, 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 the addition of 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 was then boiled in a thermocycler at 95 °C for 10 min, resolved on an 8% polyacrylamide gel (20 × 20 cm^2) 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 assays with the leader-repeat-spacer RNA for LrhCas9, RNA was truncated within the leader and spacer to avoid cleavage of irrelevant secondary structures formed internally within either domain.

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

RNA structural probing and RNase III cleavage site mapping. In 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-labelled 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× colourless loading buffer (10 M urea and 1.5 mM EDTA, pH 8.0). For RNase III cleavage assays, the same 5′-end-labelled 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 transfer RNA (Ambion) to a final concentration of 0.1 $mg\,ml^{-1}$. 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 of RNase III (Invitrogen) and further incubation at 37 °C for 5 min. Reactions were stopped by the addition of an equal volume of Gel-loading buffer II (95% (v/v) formamide, 18 mM EDTA and 0.025% (w/v) 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, and were dried and exposed to a PhosphorImager screen. RNA ladders were prepared using either 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 harbouring plasmids encoding the green fluorescent protein (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 ABS_{600} = ~0.05 in LB medium supplemented with kanamycin, and shaken at 250 r.p.m. at 37 °C to ABS_{600} = ~0.8. GFP fluorescence of single cells was then measured as described previously[48]. Briefly, cultures were diluted 1:100 in 1x PBS and analysed 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 (cutoff, 11,500) and side scatter (cutoff, 600) were used to eliminate noncellular events. The mean fluorescein isothiocyanate-A value of 30,000 events within a gate set for live *E. coli* cells was used for data analysis after subtraction of cell autofluorescence.

Phage sensitivity assay. Overnight cultures of NEB Turbo cells harbouring the CRISPR cassette plasmid with either the native leader (pCBS2225) or mutated leader (pCBS2226) were back diluted to ABS_{600} = ~0.05 in LB medium supplemented with kanamycin, and shaken at 250 r.p.m. at 37 °C to ABS_{600} = ~0.5. Cells were then collected by centrifugation and resuspended in a 1/10 volume of LB with kanamycin. Petri dishes (Ø = 90 × 16.2 mm^2) with 24 ml of LB agar supplemented with kanamycin were overlaid with 4 ml of soft LB agar (7.5 $g\,l^{-1}$) and kanamycin containing 0.75 ml of the cell suspension. After solidification for 10 min, 3 µl of tenfold serial dilutions of phage lysates was 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. Plaques were visualized using an ImageQuant LAS 4000 imaging system (GE Healthcare).

RNA immunoprecipitation for sequencing. Cas9-3×FLAG coIP combined with RIP–seq was performed on *E. coli* and *L. rhamnosus* as described previously, with minor modifications[30]. Briefly, overnight cultures of CB414 harbouring plasmid pCBS2225, pCBS2226, pCBS2240 or pCBS2241 were back diluted to ABS_{600} = ~0.05 in LB medium with kanamycin, and shaken at 250 r.p.m. at 37 °C to ABS_{600} = ~0.8. Overnight cultures of *L. rhamnosus* with or without the plasmid encoding 3×FLAG-tagged LrhCas9 pCBS2227 were back diluted to ABS_{600} = ~0.05 in MRS medium with or without chloramphenicol and incubated at 37 °C without agitation to ABS_{600} = ~0.5. The equivalent of 37–40 ABS_{600} of cells was 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*g*. 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 of buffer A, 1 µl of 1 mM DTT, 10 µl of 0.1 M PMSF, 2 µl of triton X-100, 20 µl of DNase I, 10 µl of Superase-In RNase Inhibitor) and distributed between two precooled fast-prep tubes for lysis (750 µl each). Rapid lysis was performed twice with the FastPrep homogenizer (6.5 $M\,s^{-1}$; 1 min), and the resulting lysate from both tubes was centrifuged at 4 °C for 10 min at 13,000 r.p.m.. Following centrifugation, the supernatant (that is, 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 a further 90 min at 4 °C (supernatant 2). After centrifugation the supernatant was removed and 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. For each coIP, RNA was recovered from the aqueous phase, precipitated overnight using a 30:1 mix of ethanol and 3 M 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 of ice-cold acetone was added with incubation overnight at −20 °C. Samples were centrifuged at 15,000 r.p.m. for 1 h to precipitate the protein, and washed twice with 1 ml of 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 coIP was successful, protein samples equivalent to 1.0 ABS_{600} of cells were collected during different stages of the procedure (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 with boiling for 8 min. Protein samples corresponding to ABS_{600} = 0.2 (lysate, supernatant 1, supernatant 2 and wash fraction) and ABS_{600} = 10 (eluate fraction) were used for immunoblotting analysis.

cDNA library preparation and deep sequencing. The extracted RNA was treated with DNase I (Thermo Scientific, no. EN0525) following the manufacturer's instructions. cDNA libraries for Illumina sequencing were constructed by Vertis Biotechnologie (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 an Agencourt AMPure XP kit (Beckman Coulter Genomics) following the manufacturer's instructions and analysed 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] v.2.5 using a cutoff Phred score of 20 in NextSeq mode, and reads with no remaining bases were discarded (command line parameters: –nextseq-trim=20 -m 1 -a AGATCGGAAGAGCACACGTCTGA ACTCCAGTCAC). Next, we applied the pipeline READemption[57] v.0.4.5 to align all reads longer than 11 nt (-l 12) to the respective reference sequences using segemehl[58] v.0.2.0 with an accuracy cutoff 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 nt (-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 (for example, 3×FLAG-tagged *cas9*). Links to plasmid sequences and annotations can be found in 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 of reference length >50 nt 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] v.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 both filtered and unfiltered read counts and coverage. 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] v.3.5 after conversion of normalized coverage files to BigWig format[61] using wigToBigWig v.4.

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 only from genomes 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. Supplementary Table 2 shows all leader-repeat sequences.

Bioinformatic assessment of leader-repeat structure formation. To gain insight into potential mechanisms for the inactivation of R1 in type II-A CRISPR arrays, we initially interrogated four CRISPR arrays in *S. pyogenes* M1 GAS, *L. rhamnosus* GG, *Streptococcus thermophilus* CNRZ1066 and *S. 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 nt 5′ to R1, 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 v.4.8.1 to cluster sequences by percentage identity[66]. Within these two groups, we removed sequences so that they were <70% similar to one another. We performed the same procedure for type II-C examples 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 another in the leader. We also considered the probability of forming helices in R1 with different numbers of base pairs and different numbers of mismatches or bulges. Base-pairing probabilities were calculated using v.2.4.14 of the ViennaRNA library for Python. Since an efficient algorithm for determination of the probability of helix formation has not yet been published, we used ViennaRNA to sample random structures from 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 those fully contained within the leader did not contribute to either base-pairing probabilities or helix-formation probabilities. To estimate the statistical significance of base-pairing or helix-formation probabilities, we generated random samples by randomly permuting nucleotides within the leader sequence. 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]: http://clavius.bc.edu/~clotelab/RNAdinucleotideShuffle/ShuffleCodeParts/altschulEriksonDinuclShuffle.txt. We used 1,000 random samples to estimate empirical *P* values.For each of the II-A or II-C leader-repeat pairs, and for each statistic (for example, helix-formation probability), we calculated a corresponding *P* value. We combined the *P* values for both II-A and II-C examples using Fisher's method, as implemented by the scipy.stats.combine_pvalues function in Python3 (using v.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 individual *P* values. Because our empirical *P* values are based on 1,000 samples, some estimated values will be zero (in cases of a very strong helix), leading to logarithms of negative infinity. To address this issue, we replaced empirical $P=0$ with $P=1/1{,}000$. This value is slightly higher 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 *P* values are independent but ours are based on sequences that presumably are evolutionarily related. We hoped that elimination of sequences >70% identical would eliminate this problem. It was not practical to more aggressively eliminate similar sequences (for example, 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 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 R1 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 helix-formation probability. However, we ultimately decided that variant methods did not noticeably change the overall statistics and we continued to use the original formulation. We utilized 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 may be pairing propensity in some type II-C leaders. For the diagram in Fig. 5a we used all 68 type II-A leader-repeat examples <70% identical to one another. So that all II-C examples would be of 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, followed by removal of systems >70% identical. We thus arrived at 379 I-E systems and 151 I-F systems in the initial set. We used this set to analyse our results, and quickly found that our previously applied procedure did not lead to statistically significant aggregate *P* values. We then analysed 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. *P* values for the I-E and I-F systems were both 1.0 because, in a high proportion of I-E and I-F CRISPR–Cas systems, the probability that eight consecutive base pairs would form is 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) are similar to the 68 systems used for the II-A and II-C depictions.

Statistical analyses. Statistical comparisons of experimental data were performed using 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 analyse 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.

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

Data availability

Next-generation sequencing data for RNA immunoprecipitation sequencing are accessible through NCBI Gene Expression Omnibus accession no. GSE158637 using the link https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158637 (Supplementary Table 4). Source data for Figs. 1b,d,e, 2a,b, 3b–d and 4b–d and Extended Data Figs. 1b,d, 3a,c,d, 4a–d, 5a,b,d, 6a, 7a,c, 8b,c,e,f and 9a,d are included in the Source Data files. Source data are provided with this paper.

Code availability

Custom scripts analysing 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.

Received: 7 November 2021; Accepted: 1 February 2022;
Published online: 21 March 2022

References

1. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
2. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **12**, 479–492 (2014).
3. Jackson, S. A. et al. CRISPR-Cas: adapting to change. *Science* **356**, eaal5056 (2017).
4. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
5. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).
6. Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* **6**, 181–186 (2008).
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 Res.* **42**, 7884–7893 (2014).
8. Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137–141 (2017).
9. McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat. Rev. Microbiol.* **17**, 7–12 (2019).
10. Brouns, S. J. J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
11. Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol. Rev.* **39**, 428–441 (2015).
12. Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
13. Meeske, A. J., Nakandakari-Higa, S. & Marraffini, L. A. Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241–245 (2019).
14. Rostøl, J. T. et al. The Card1 nuclease provides defence during type III CRISPR immunity. *Nature* **590**, 624–629 (2021).
15. Elmore, J. R. et al. Programmable plasmid interference by the CRISPR-Cas system in *Thermococcus kodakarensis*. *RNA Biol.* **10**, 828–840 (2013).
16. Carte, J. et al. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol. Microbiol.* **93**, 98–112 (2014).
17. Crawley, A. B., Henriksen, E. D., Stout, E., Brandt, K. & Barrangou, R. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Sci. Rep.* **8**, 11544 (2018).
18. Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
19. McGinn, J. & Marraffini, L. A. CRISPR-Cas systems optimize their immune response by specifying the site of spacer integration. *Mol. Cell* **64**, 616–623 (2016).
20. Martynov, A., Severinov, K. & Ispolatov, I. Optimal number of spacers in CRISPR arrays. *PLoS Comput. Biol.* **13**, e1005891 (2017).
21. Rao, C., Chin, D. & Ensminger, A. W. Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. *RNA* **23**, 1525–1538 (2017).
22. Liao, C. & Beisel, C. L. The tracrRNA in CRISPR biology and technologies. *Annu. Rev. Genet.* **55**, 161–181 (2021).
23. Karvelis, T. et al. crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. *RNA Biol.* **10**, 841–851 (2013).
24. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
25. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).
26. Bikard, D. et al. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).
27. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).

28. Citorik, R. J., Mimee, M. & Lu, T. K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nat. Biotechnol.* **32**, 1141–1145 (2014).
29. Leenay, R. T. & Beisel, C. L. Deciphering, communicating, and engineering the CRISPR PAM. *J. Mol. Biol.* **429**, 177–191 (2017).
30. Dugar, G. et al. CRISPR RNA-dependent binding and cleavage of endogenous RNAs by the *Campylobacter jejuni* Cas9. *Mol. Cell* **69**, 893–905 (2018).
31. Xue, C. et al. CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Res.* **43**, 10831–10847 (2015).
32. Collias, D. et al. A positive, growth-based PAM screen identifies noncanonical motifs recognized by the Cas9. *Sci. Adv.* **6**, eabb4054 (2020).
33. Altuvia, Y. et al. *In vivo* cleavage rules and target repertoire of RNase III in *Escherichia coli*. *Nucleic Acids Res.* **46**, 10530–10531 (2018).
34. Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res.* **43**, 1749–1758 (2015).
35. Pougach, K. et al. Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol. Microbiol.* **77**, 1367–1379 (2010).
36. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
37. Jiao, C. et al. Noncanonical crRNAs derived from host transcripts enable multiplexable RNA detection by Cas9. *Science* **372**, 941–948 (2021).
38. Jabbari, H., Wark, I. & Montemagno, C. RNA secondary structure prediction with pseudoknots: contribution of algorithm versus energy model. *PLoS ONE* **13**, e0194583 (2018).
39. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev.* **29**, 356–361 (2015).
40. Laanto, E., Hoikkala, V., Ravantti, J. & Sundberg, L.-R. Long-term genomic coevolution of host-parasite interaction in the natural environment. *Nat. Commun.* **8**, 111 (2017).
41. Zhang, Y. et al. Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Mol. Cell* **50**, 488–503 (2013).
42. Dugar, G. et al. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet.* **9**, e1003495 (2013).
43. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).
44. Li, R. & Bowerman, B. Symmetry breaking in biology. *Cold Spring Harb. Perspect. Biol.* **2**, a003475 (2010).
45. McCarty, N. S., Graham, A. E., Studená, L. & Ledesma-Amaro, R. Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nat. Commun.* **11**, 1281 (2020).
46. Al-Hashimi, H. M. & Walter, N. G. RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.* **18**, 321–329 (2008).
47. Watters, K. E., Strobel, E. J., Yu, A. M., Lis, J. T. & Lucks, J. B. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nat. Struct. Mol. Biol.* **23**, 1124–1131 (2016).
48. Liao, C. et al. Modular one-pot assembly of CRISPR arrays enables library generation and reveals factors influencing crRNA biogenesis. *Nat. Commun.* **10**, 2948 (2019).
49. Wimmer, F. & Beisel, C. L. CRISPR-Cas systems and the paradox of self-targeting spacers. *Front. Microbiol.* **10**, 3078 (2019).
50. Leenay, R. T. et al. Genome editing with CRISPR-Cas9 in *Lactobacillus plantarum* revealed that editing outcomes can vary across strains and between methods. *Biotechnol. J.* **14**, e1700583 (2019).
51. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA Websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
52. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
53. Sharma, C. M. et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).
54. Papenfort, K. et al. σ^E-Dependent small RNAs of *Salmonella* respond to membrane stress by accelerating global *omp* mRNA decay. *Mol. Microbiol.* **62**, 1674–1688 (2006).
55. Pernitzsch, S. R., Tirier, S. M., Beier, D. & Sharma, C. M. A variable homopolymeric G-repeat defines small RNA-mediated posttranscriptional regulation of a chemotaxis receptor in *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA* **111**, E501–E510 (2014).
56. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 1 (2011).
57. Förstner, K. U., Vogel, J. & Sharma, C. M. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics* **30**, 3421–3423 (2014).
58. Hoffmann, S. et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* **5**, e1000502 (2009).
59. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Lopez-Delisle, L. et al. pyGenomeTracks: reproducible plots for multivariate genomic data sets. *Bioinformatics* **37**, 422–423 (2020).
61. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
62. Padilha, V. A., Alkhnbashi, O. S., Shah, S. A., de Carvalho, A. C. P. L. F. & Backofen, R. CRISPRcasIdentifier: machine learning for accurate identification and classification of CRISPR-Cas systems. *Gigascience* **9**, giaa062 (2020).
63. Padilha, V. A. et al. Casboundary: automated definition of integral Cas cassettes. *Bioinformatics* **37**, 1352–1359 (2020).
64. Mitrofanov, A. et al. CRISPRidentify: identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res.* **49**, e20 (2021).
65. Alkhnbashi, O. S. et al. CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489–i496 (2014).
66. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
67. Ding, Y. & Lawrence, C. E. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* **31**, 7280–7301 (2003).
68. Altschul, S. F. & Erickson, B. W. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.* **2**, 526–538 (1985).

Acknowledgements

We thank T. Achmedov for extensive assistance with RNA preparation and RNA-blotting, F. Tippel from NanoTemper Technologies (Munich) for technical support and J. Vogel and G. Storz for critical feedback on the manuscript. This work was supported by funding through the European Research Council Consolidator Award (no. 865973 to C.L.B.), Deutsche Forschungsgemeinschaft SPP 2141 (nos. BE 6703/1-1 to C.L.B., SH 580/9-1 to C.M.S. and BA 2168/23-1 to R.B.) and the Interdisciplinary Center for Clinical Research Würzburg project Z-6.

Author contributions

C.L. and C.L.B. conceived this study. C.L. and C.L.B. designed the experiments. C.L. performed plasmid cloning, in vivo assays in *E. coli* and *L. rhamnosus*, in vitro RNA transcription and purification, RNase III cleavage assays and RNA-blotting. C.L. and C.L.B. analysed the associated data. S.S. conducted immunoblotting and RNA immunoprecipitation for RIP–seq and helped analyse the data. S.L.S. conducted RNA structural probing and RNase III cleavage site mapping and helped analyse data. C.M.S. supervised the work performed by S.S. and S.L.S. A.K. designed and performed the in vitro assay for RNA–RNA binding affinity and analysed 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. analysed RIP–seq data. 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

C.L.B. is a cofounder 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.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-022-01074-3.

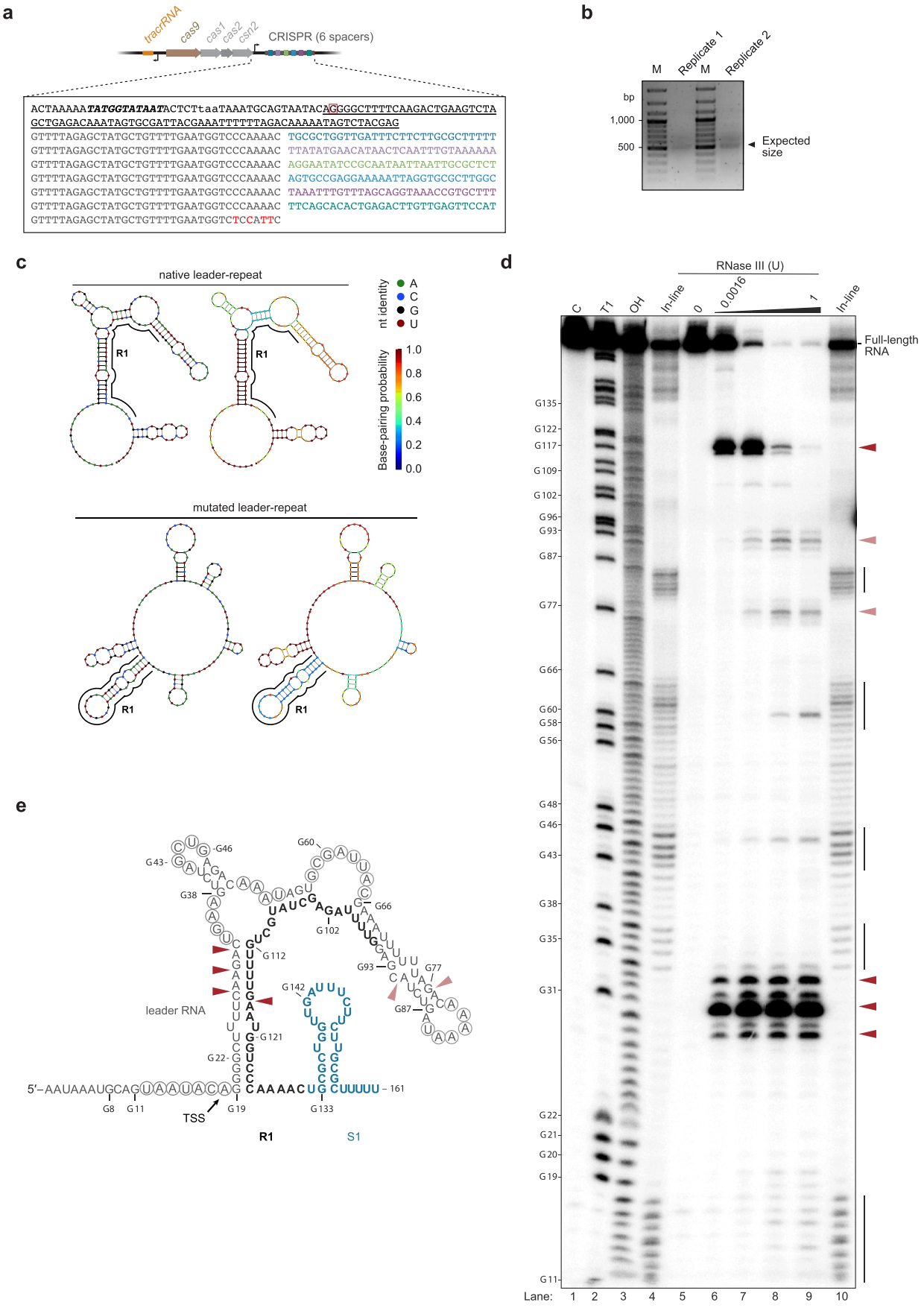
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01074-3.

Correspondence and requests for materials should be addressed to Chase L. Beisel.

Peer review information *Nature Microbiology* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

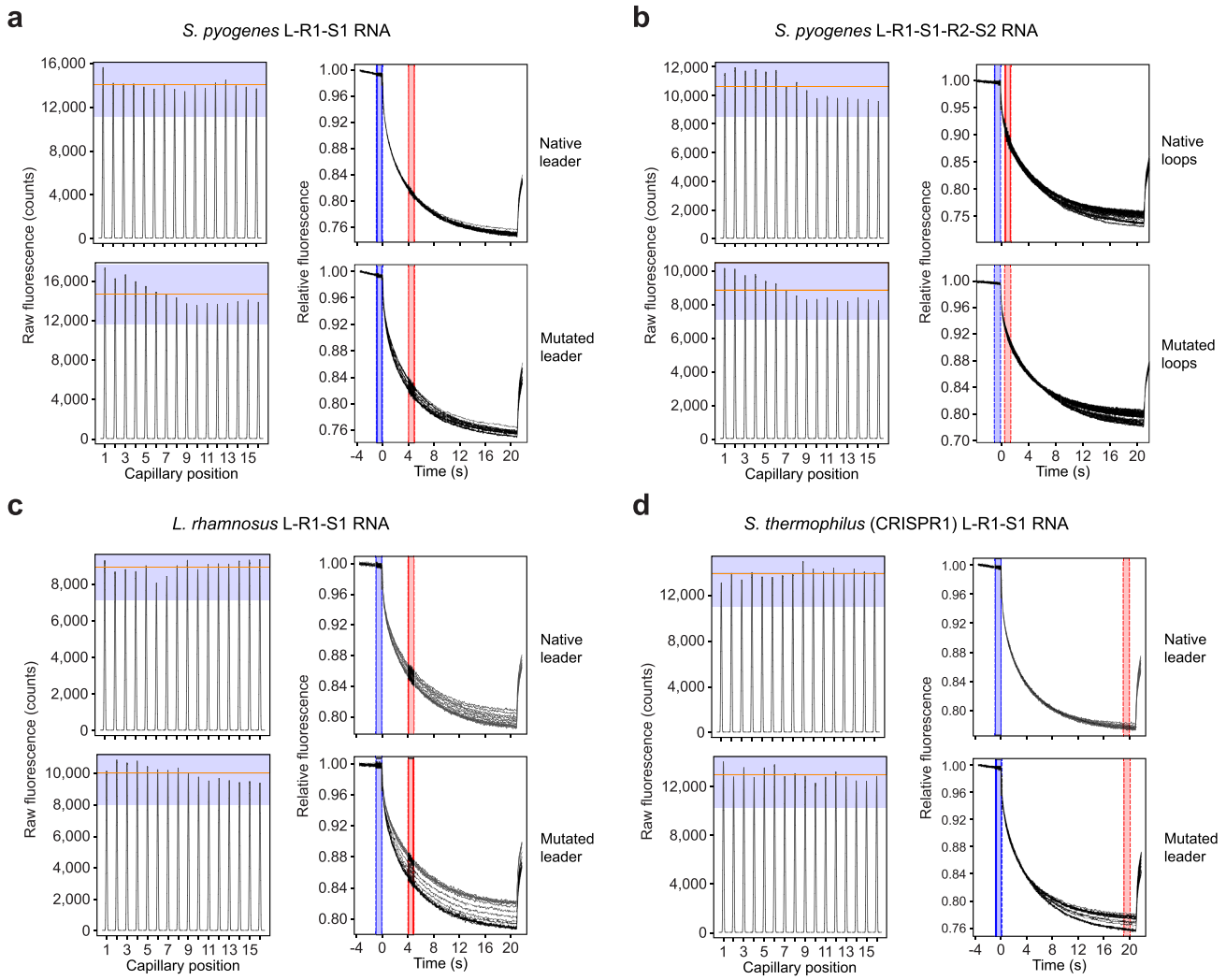
Reprints and permissions information is available at www.nature.com/reprints.

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

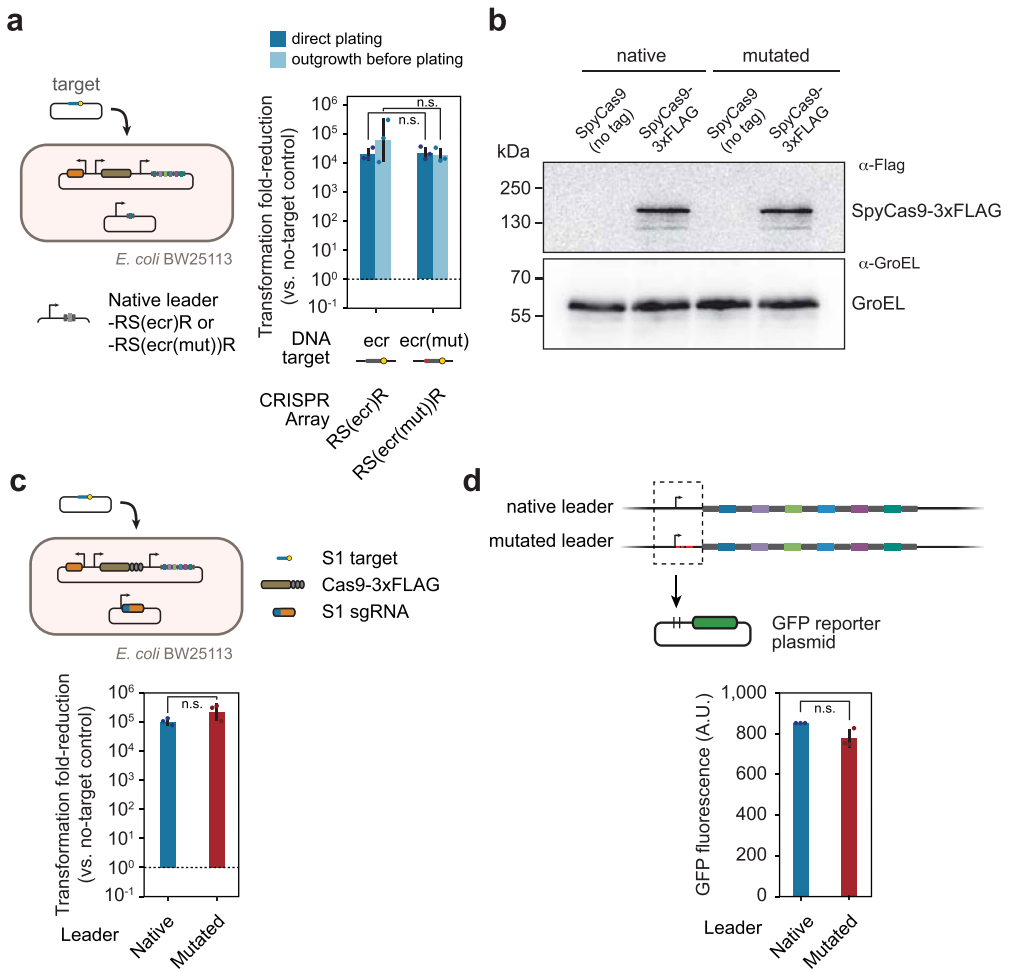


Extended Data Fig. 1 | See next page for caption.

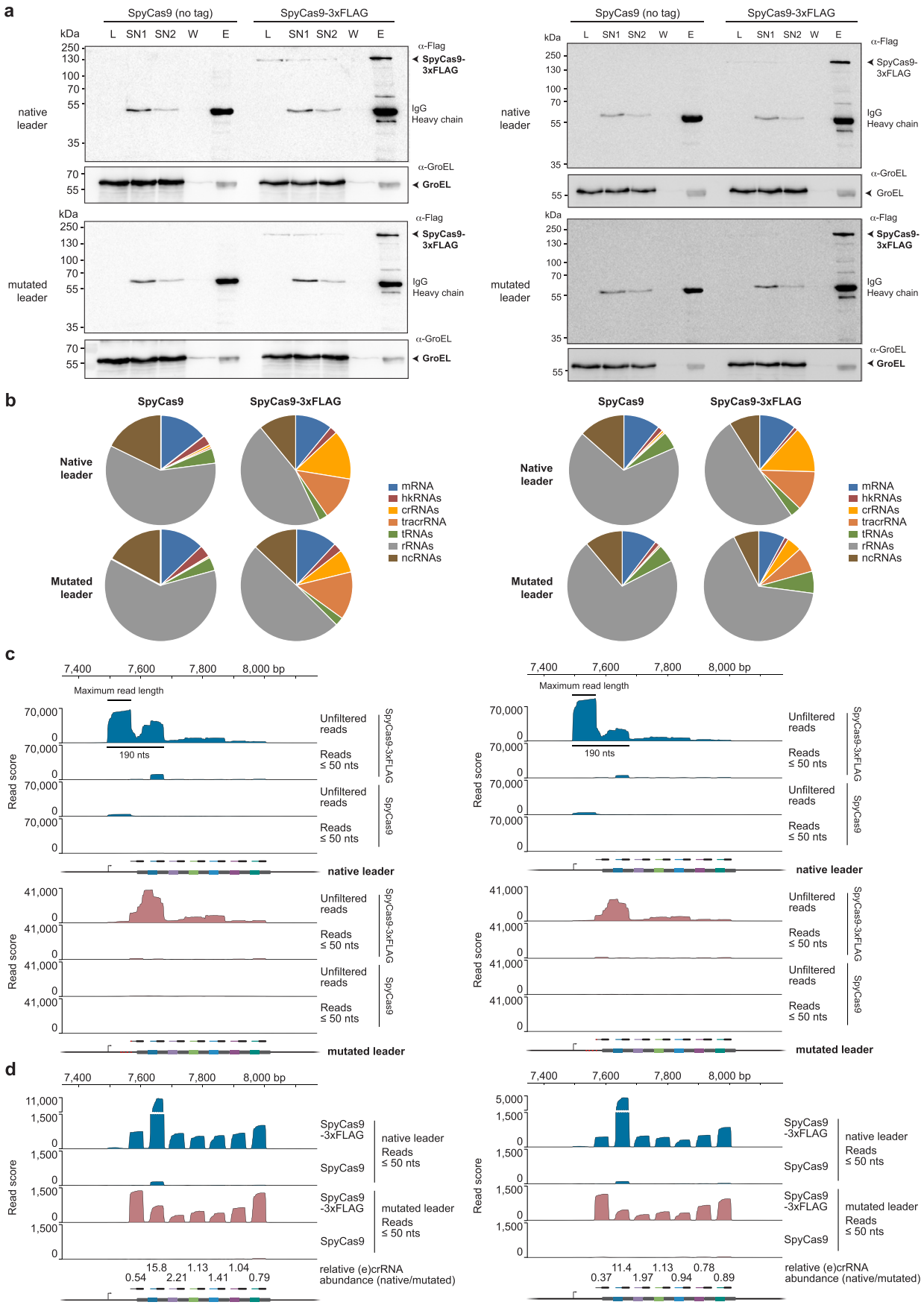
Extended Data Fig. 1 | The leader-repeat stem-loop from the CRISPR-Cas9 system native to *Streptococcus pyogenes* SF370. Accession #: NC_002737.2. **a**, Array sequence and context within the CRISPR-Cas system. Repeats are in gray, spacers match the corresponding color in the cartoon, and mutations to 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 Fig. 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 (CRISPR1) 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 Figs. 1d and 4d and Extended Data Figs. 8b and 8e for the resulting binding curves. Values in a-d represent the mean and standard deviation of triplicate independent measurements.

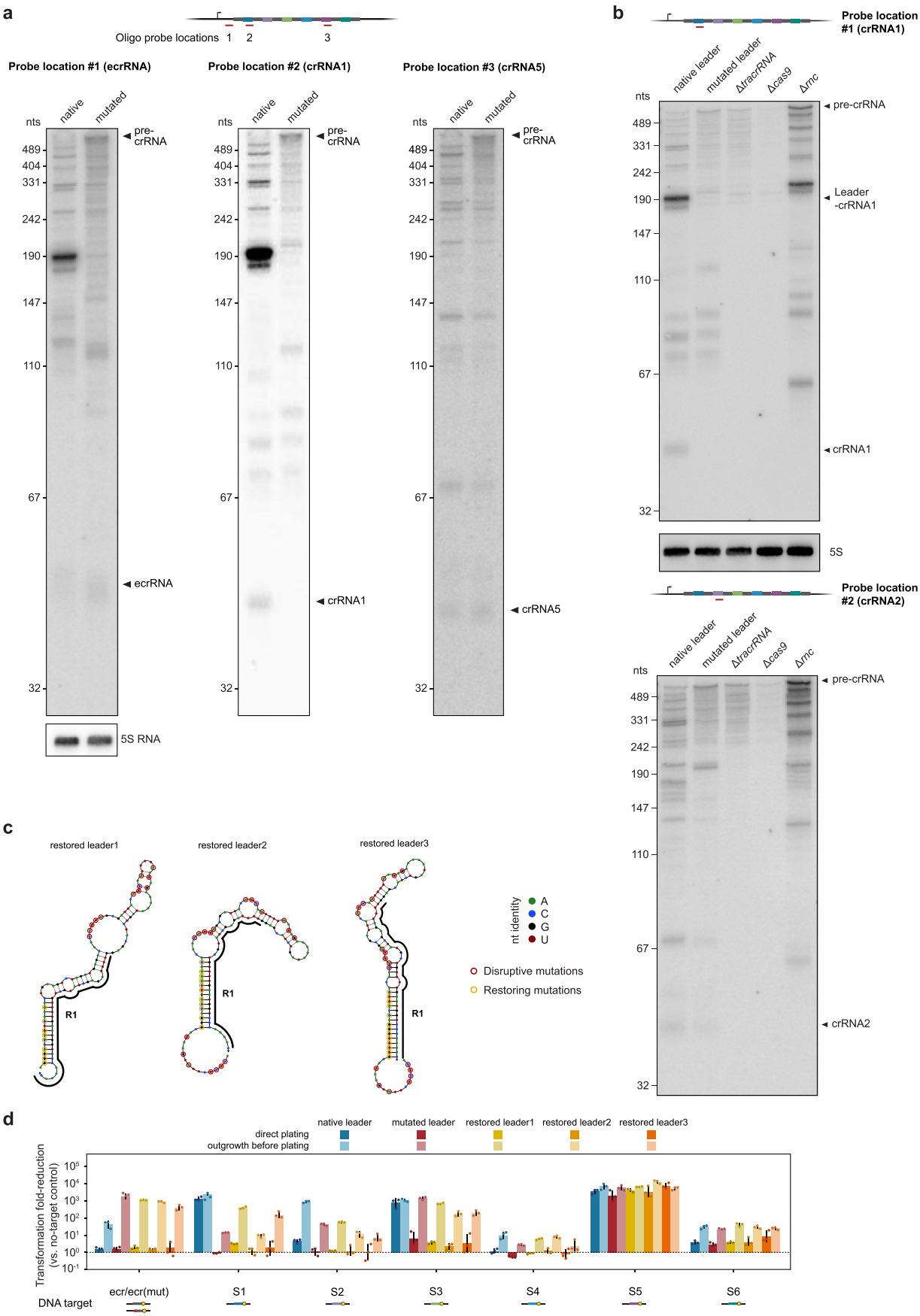


Extended Data Fig. 3 | Data rejecting alternative explanations for the impact of mutating the leader region associated with SpyCas9. a, Assessing targeting by the mutated ecrRNA guide by plasmid clearance in *E. coli*. The native and mutated ecrRNAs were encoded as single-spacer arrays with the native 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.



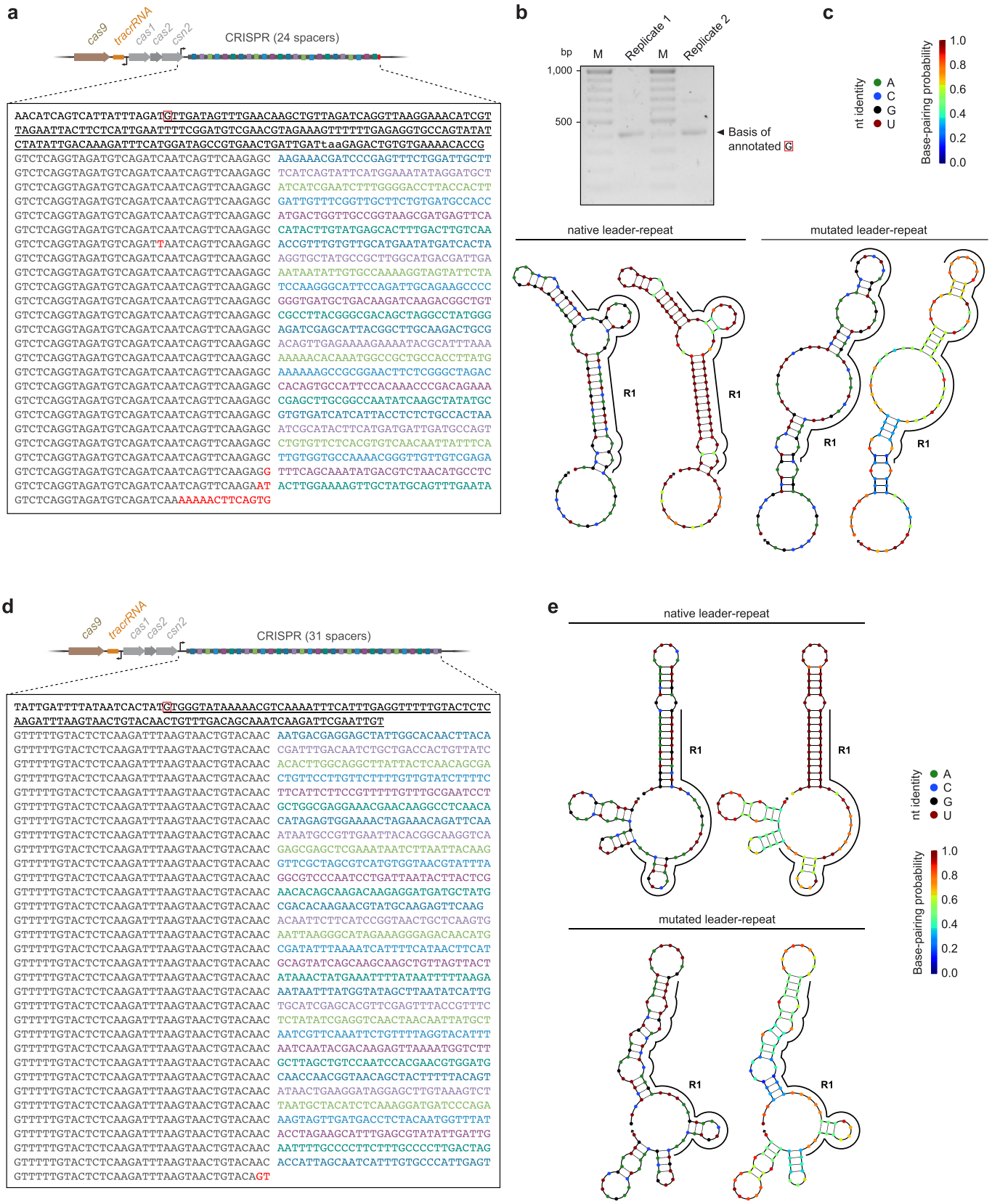
Extended Data Fig. 4 | See next page for caption.

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 Fig. 3b and Extended Data Fig. 5a-b). We also note that the reads begin ~12 nts upstream of the transcriptional start site mapped by 5′ RACE (see Extended Data Fig. 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 Fig. 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.



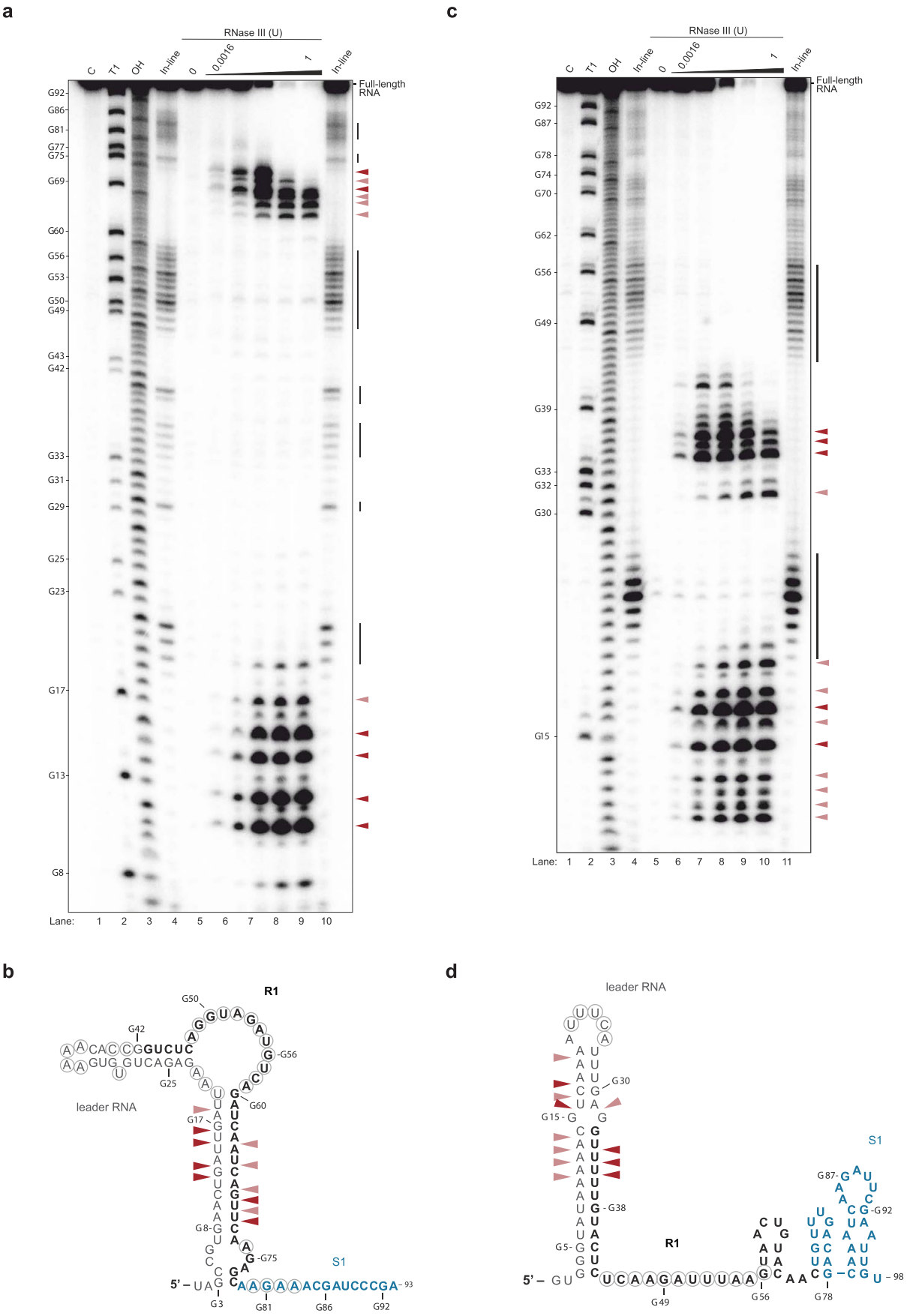
Extended Data Fig. 5 | See next page for caption.

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 Fig. 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 Fig. 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 Fig. 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.



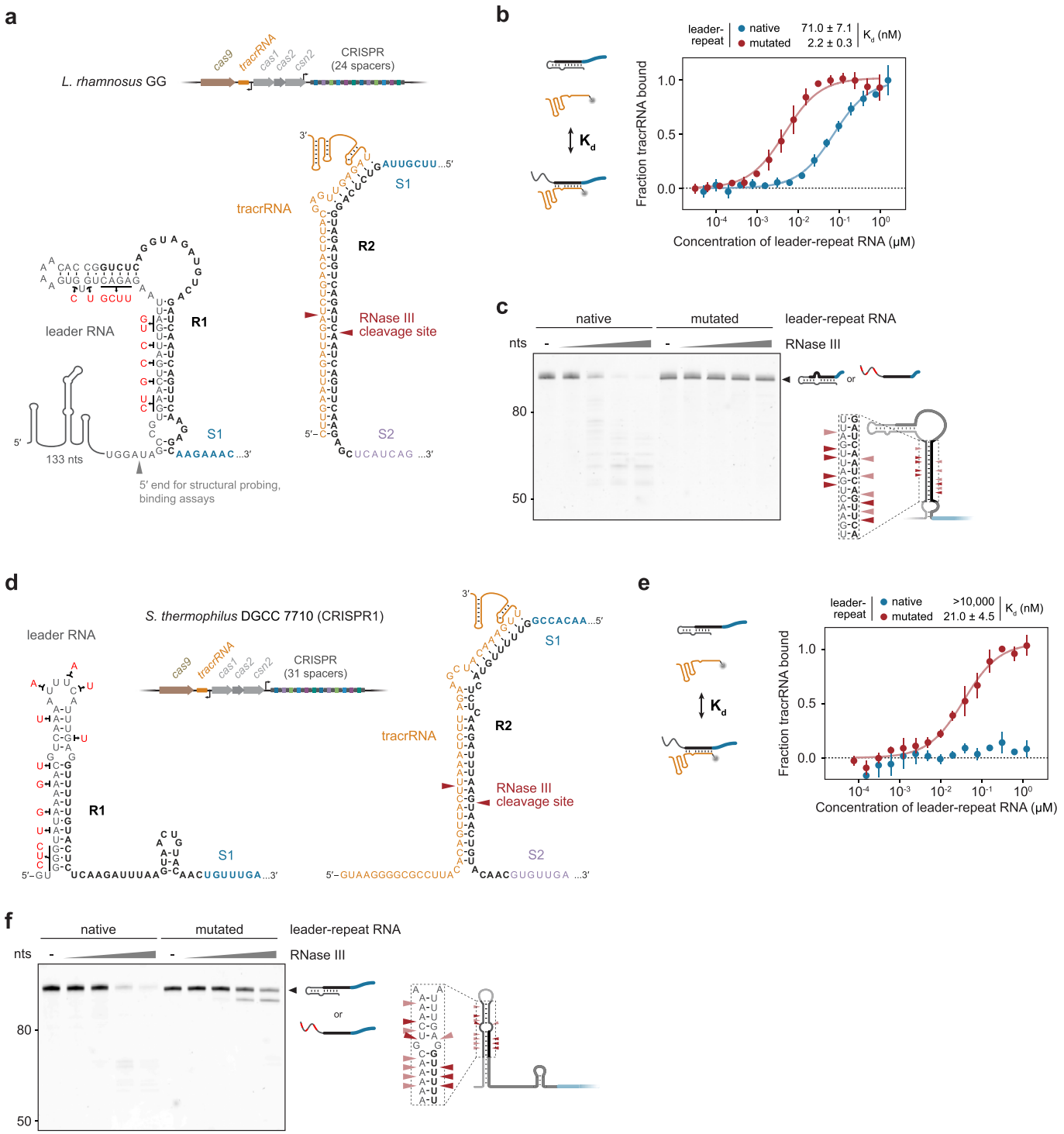
Extended Data Fig. 6 | See next page for caption.

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 Fig. 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 Fig. 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 leader-repeat RNA predicted by NUPACK. See Extended Data Fig. 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 Fig. 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 Fig. 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 leader-repeat RNA predicted by NUPACK. See Extended Data Fig. 1c for details.

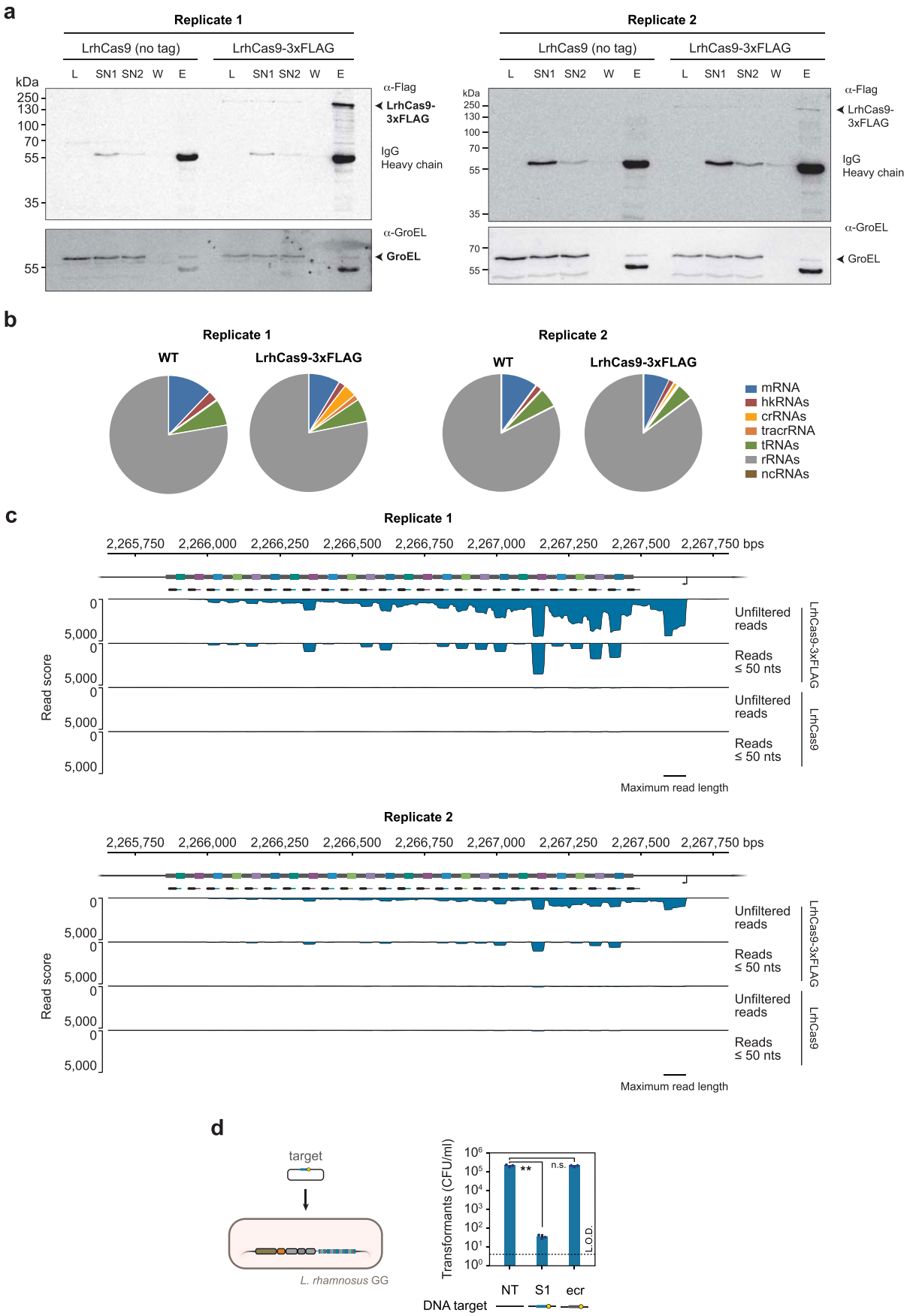


Extended Data Fig. 7 | See next page for caption.

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 Fig. 8a). R1: first repeat. S1: first spacer. **c**, *In vitro* determination of the secondary structure and RNase III cleavage sites for the leader-repeat 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 in-line probing. The preferred site of RNase III cleavage lies above the equivalent site within the crRNA:tracrRNA duplex (see Extended Data Fig. 8d). R1: first repeat. S1: first spacer.

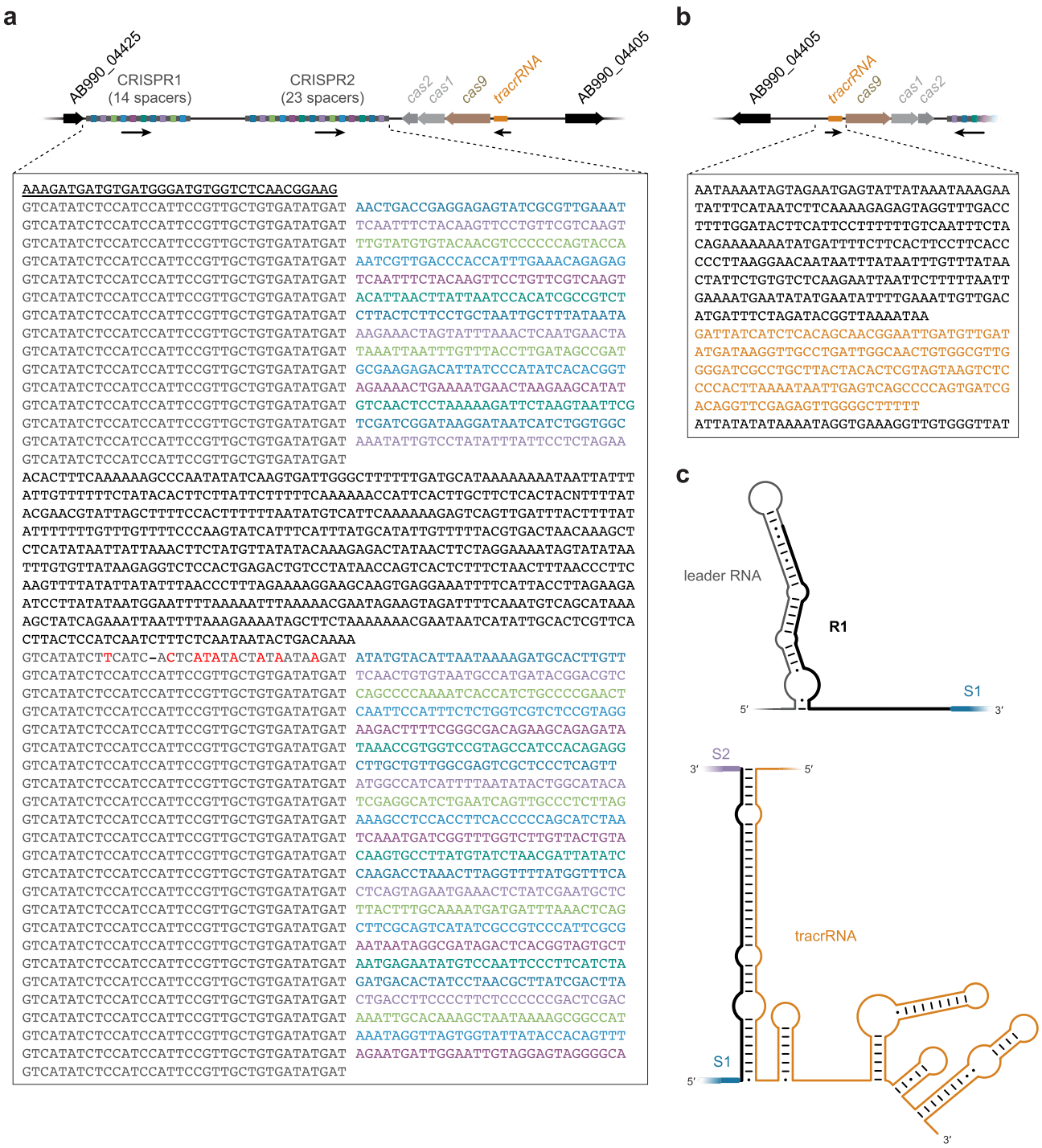


Extended Data Fig. 8 | II-A CRISPR-Cas9 systems form distinct leader-repeat stem-loops. a, The CRISPR-Cas system from *L. rhamnosus* GG and the secondary structure of the leader-repeat RNA. The structure was predicted by NUPACK and confirmed *in vitro* (see Extended Data Fig. 6c and 7a-b). Mutations 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 Fig. 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 Fig. 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 Fig. 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 Fig. 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 | See next page for caption.

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.

nature research

Corresponding author(s): Chase L. Beisel

Last updated by author(s): 1/24/2022

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NextSeq 500 (Illumina)
Data analysis	Open source codes (Peter Clote's implementation) were used for statistical analysis of leader-repeat pairing across CRISPR-Cas systems. Available from http://clavius.bc.edu/~clotelab/RNAdinucleotideShuffle/ShuffleCodeParts/altschulEriksonDinuclShuffle.txt Online NUPACK algorithm (2007 - 2022 Caltech) was used for predicting equilibrium folding of RNA. Available from http://www.nupack.org/partition/new. MO.Affinity Analysis software version 2.3 was used for analyzing binding affinity data. Cutadapt version 2.5, READemption version 0.4.5, Segemehl version 0.2.0, pysam version 0.16.0.1, pyGenomeTracks version 3.5, and wigToBigWig v4 are used for analyzing of RIP-seq data. cd-hit version 4.8.1 was used to cluster sequences by percent identity during bioinformatic assessment of leader-repeat structure formation. Microsoft excel 6.16.27 A custom script (https://github.com/zashaweinberglab/type-II-A-leader-repeat) was used to predict folding of leader-repeat RNAs and perform statistical tests using 1,000 randomly generated leader sequences.

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

Data

Policy information about availability of data

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

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

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. 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.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	No statistical methods were used to predetermine sample size. At least three random colonies were picked as biological replication for each growth related assay. Sample sizes were determined based on our previous experiences and what were described in similar experiments in published papers.
Data exclusions	Experiments were done with at least three biological replications. All reported data were reproducible. Data that were not reproducible because of misconducting were excluded.
Replication	Values in the figures represent the average of at least three independent experiments starting from separate colonies or structure probing conducted on separate days. Each set of RIP-seq was conducted twice as starting from two random colonies.
Randomization	Randomization was not relevant to this study because no objects of study were assigned to different experimental groups. Colonies were picked randomly for subjecting to experiments.
Blinding	Blinding was not performed in this study. In most cases, E. coli harboring different plasmids used for study were assigned irrelevant names e.g. A, B, C, D...) during conducting of experiments. Image acquisitions and numeric data analyses were automated in most cases.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Antibodies

Antibodies used	Monoclonal ANTI-FLAG M2 antibody (Sigma,#F 1804, 1:1,000 in PBS), anti-GroEL primary antibody (Sigma, cat. # G6532, 1:1,000 in PBS), horseradish peroxidase-coupled anti-mouse IgG secondary antibody (Thermo Fisher, cat. #31430, 1:10,000 in PBS), anti-rabbit IgG secondary antibody (GE-Healthcare, cat. #NA934V, 1:10,000 in PBS)

Validation	We used commercial antibody reagents for Western-blot and immunoprecipitation. Validation data are available on the manufacturer's websites and data sheets.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	E. coli cells were grown in liquid medium to specific OD and diluted for reaching appropriate number of events per ml.
Instrument	Accuri C6 Plus flow cytometer with BD CSampler Plus (Becton Dickinson)
Software	BD Accuri™ C6 Software version 264.21
Cell population abundance	Flow Cytometry was used for measuring the GFP expression of E. coli, no sorting was done.
Gating strategy	For E. coli, cells stained with DRAQ5 was used to set the specific gate to ensure that no debris appeared within the gate. The gating figure is provided as Source data for Figure 1b.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.