Nus Factors Prevent Premature Transcription Termination of Bacterial CRISPR Arrays

Anne M. Stringer[1], Gabriele Baniulyte[2], Erica Lasek-Nesselquist[1], and Joseph T. Wade[1,2]

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

A hallmark of CRISPR-Cas immunity systems is the CRISPR array, a genomic locus consisting of short, repeated sequences ("repeats") interspersed with short, variable sequences ("spacers"). CRISPR arrays are transcribed and processed into individual CRISPR RNAs (crRNAs) that each include a single spacer, and direct Cas proteins to complementary sequence in invading nucleic acid. Most bacterial CRISPR array transcripts are unusually long for untranslated RNA, suggesting the existence of mechanisms to prevent premature transcription termination by Rho, a conserved bacterial transcription termination factor that rapidly terminates untranslated RNA. We show that CRISPR arrays of *Salmonella* Typhimurium are protected from Rho by the Nus factor anti-termination complex, and we provide evidence that this is an evolutionarily ancient mechanism to facilitate complete transcription of bacterial CRISPR arrays.

CRISPR, BoxA, Nus factors, Antitermination
Correspondence: ***joseph.wade @health.ny.gov***

Introduction

CRISPR-Cas systems are adaptive immune systems found in 40% of bacteria and 90% of archaea. The hallmark of CRISPR-Cas systems is the CRISPR array, which is composed of alternating multiple, short, identical "repeat" sequences, interspersed with short, variable "spacer" sequences. A critical step in CRISPR immunity is biogenesis, (1) which involves transcription of a CRISPR array into a single, long precursor RNA that is then processed into individual CRISPR RNAs (crRNAs), with each crRNA containing a single spacer sequence. crRNAs associate with an effector Cas protein or Cas protein complex, and direct the Cas protein(s) to an invading nucleic acid sequence that is complementary to the crRNA spacer and includes a neighboring Protospacer Adjacent Motif (PAM). This leads to cleavage of the invading nucleic acid by a Cas protein nuclease, in a process known as "interference".

Rho is a broadly conserved bacterial transcription termination factor. Rho terminates transcription only when nascent RNA is untranslated (2). Hence, the primary function of Rho is to suppress the transcription of spurious, non-coding RNAs that initiate as a result of pervasive transcription (3–5). To terminate transcription, Rho must load onto nascent RNA at a "Rho utilization site" (Rut). The precise sequence and structure requirements for Rho loading are not fully understood, but Ruts typically have a high C:G ratio, limited secondary structure, and are enriched in YC dinucleotides (2, 6). However, the overall sequence/structure specificity of Ruts is believed to be low, and a large proportion of the *Salmonella* Typhimurium genome is predicted to be capable of functioning as a Rut (6). Once Rho loads onto nascent RNA, it translocates along the RNA in a 5' to 3' direction using its helicase activity. Rho typically catches the RNA polymerase (RNAP) within 60-90 nucleotides, leading to transcription termination (2).

The ability of Rho to terminate untranslated RNA with relatively low specificity effectively limits the potential length of bacterial non-coding RNAs. However, there are two classes of bacterial non-coding RNA that are notably long: ribosomal RNA (rRNA) and CRISPR array transcripts. rRNA is resistant to Rho termination, likely due to redundant mechanisms (7). The best-studied mechanism for preventing Rho termination involves the Nus factor complex, which is known to be associated with rRNA. The Nus complex consists of five proteins, NusA, NusB, NusE (ribosomal protein S10), NusG and SuhB, that bind to both nascent RNA and elongating RNAP. Nus complex formation begins with association of NusB/E with a short RNA sequence known as "BoxA". NusE has been proposed to interact with NusG (8), which is associated with elongating RNAP, thus creating a loop in the nascent RNA (7, 9). The Nus complex prevents Rho termination (10, 11) in a BoxA-dependent manner (11–13), and BoxA elements are found in phylogenetically diverse copies of rRNA (14, 15).

Results and Discussion

We identified *boxA*-like sequences a short distance upstream of both CRISPR arrays (CRISPR-I and CRISPR-II) in *S.* Typhimurium (Fig. 1A). To facilitate studies of the *S.* Typhimurium CRISPR-Cas system, which is transcriptionally silenced by H-NS (16), we introduced a strong, constitutive promoter (17) in place of *cas3* (Fig. 1A). This promoter drives transcription of the *cas8e-cse2-cas7-cas5-cas6e-cas1-cas2* operon, and our ChIP-qPCR data for RNAP indicate that transcription continues through the *boxA* into the CRISPR array (Fig. S1). We also introduced a strong, constitutive promoter upstream of the CRISPR-II array, immediately downstream of the *queE* gene (Fig. 1A). Transcription from this promoter also covers the putative *boxA*. To determine whether the putative *boxA* elements upstream of the CRISPR arrays are genuine, we measured association of TAP-tagged SuhB with elongating RNAP at the CRISPR-II array using ChIP-qPCR, which likely detects indirect association of

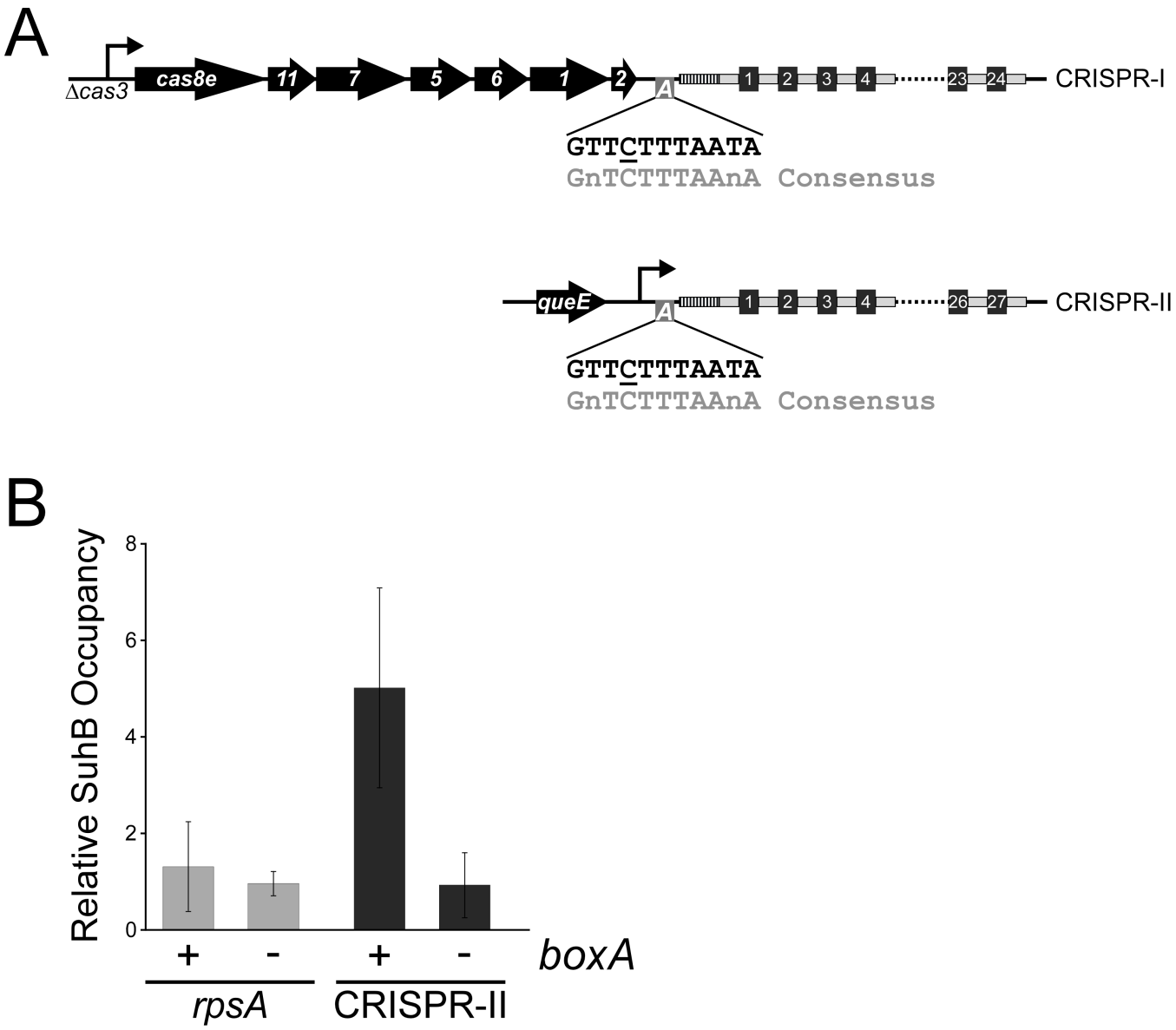

Fig. 1. *boxA* elements upstream of both CRISPR arrays in *Salmonella* Typhimurium. (A) Schematic of the two CRISPR arrays in *S.* Typhimurium. Repeat sequences are represented by gray rectangles and spacer sequences are represented by black squares. Spacers are numbered within the array, with spacer 1 being closest to the leader sequence (dashed rectangle). The CRISPR-I array is co-transcribed with the upstream cas genes. For the work presented in this study, the *cas3* gene was deleted and replaced with a constitutive promoter. Similarly, a constitutive promoter was inserted immediately downstream of *queE*, upstream of the CRISPR-II array. *boxA* elements (boxes containing an "A") are located immediately upstream of the leader sequences of both CRISPR arrays. **(B)** Relative occupancy of SuhB-TAP, determined by ChIP-qPCR, within the highly expressed *rpsA* gene (gray bars), or at the *boxA* sequence upstream of CRISPR-II (black bars). Occupancy was measured in a strain with an intact *boxA* upstream of CRISPR-II (AMD710), or a single base-pair substitution within the *boxA* (AMD711). Values plotted are the average of three independent biological replicates. Error bars represent one standard-deviation from the mean.

SuhB with the DNA (7). Our data indicate robust association of SuhB with the region immediately downstream of the putative *boxA*, but not with the highly transcribed *rpsA* gene that is not associated with a *boxA* (Fig. 1B). By contrast, we detected substantially reduced SuhB association with the same genomic region in a strain containing a single base pair substitution in the *boxA* that is expected to abrogate NusB/E association (18–20), with SuhB association being similar to that with *rpsA*. The level of SuhB association with *rpsA* was not significantly altered by the mutation in the CRISPR-II *boxA*. We conclude that the CRISPR-II array transcript includes a functional upstream BoxA. For almost 40 years, the Nus factor complex was believed to be a dedicated rRNA regulator, with no other known bacterial targets (14). We recently identified a novel function for the Nus factor complex – autoregulation of *suhB* – and we provided evidence for many additional targets (18). Nonetheless, identification of CRISPR arrays as a novel target for the Nus factor complex substantially increases the number of known targets and provides new opportunities for investigating the mechanism by which Nus factors prevent Rho termination.

We hypothesized that BoxA-mediated association of the Nus factor complex with RNAP at the S. Typhimurium CRISPR arrays prevents premature Rho-dependent transcription termination. To test this hypothesis, we constructed *lacZ* transcriptional reporter fusions that contain a constitutive pro-

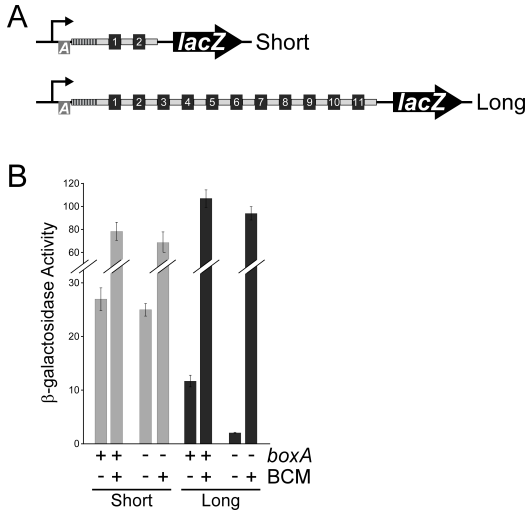

Fig. 2. BoxA-mediated antitermination of a CRISPR array. (A) Schematic of short (pGB231 and pGB237) and long (pGB250 and pGB256) *lacZ* reporter gene transcriptional fusions to the CRISPR-II array. **(B)** β-galactosidase activity of the short and long *lacZ* reporter gene fusions with either an intact (pGB231 and pGB250) or mutated (pGB237 and pGB256) *boxA* sequence, in cells grown with/without addition of the Rho inhibitor bicyclomycin (BCM). Values plotted are the average of three independent biological replicates. Error bars represent one standard-deviation from the mean.

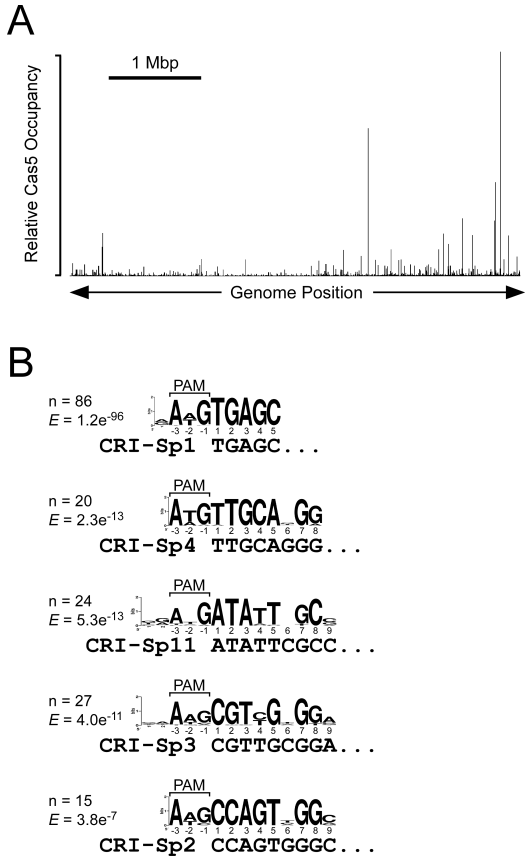

Fig. 3. Extensive off-target binding of *S.* Typhimurium Cascade. (A) Relative $FLAG_3$-Cas5 occupancy across the *S.* Tyhpimurium genome as measured by ChIP-seq in strain AMD678. **(B)** Enriched sequence motifs in Cas5-bound regions, as identified by MEME (30). Sequence matches to the AWG PAM and to the seed sequence of specific spacers are indicated. The number of Cascade-bound genomic regions containing each sequence motif is indicated, as is the enrichment score (*E*) generated by MEME (30).

moter followed by the sequence downstream of the *queE* gene (upstream of the CRISPR-II array), extending to either the 2nd ("short fusion") or the 11th ("long fusion") spacer of the array (Fig. 2A). We constructed equivalent fusions that contain a single base pair substitution in the *boxA* that is expected to abrogate NusB/E association (Fig. 1B) (18–20). We then measured β-galactosidase activity for each of the four fusions in cells grown with/without bicyclomycin (BCM), a specific inhibitor of Rho (2). In the absence of BCM, expression of the long fusion but not the short fusion was substantially reduced by mutation of the *boxA* (Fig. 2B). By contrast, expression of all fusions was similar for cells grown in the presence of BCM (Fig. 2B). Thus, our data are consistent with BoxA-mediated, Nus factor anti-termination of the CRISPR array, with Rho termination occurring between the 2nd and 11th spacer when anti-termination is disrupted. Surprisingly, expression levels of both the short and long fusions were substantially higher in cells grown with BCM, even with an intact *boxA*, suggesting that some Rho termination occurs upstream of the *boxA*. Moreover, expression of the long fusion was lower than that of the short fusion, even with an intact *boxA*, suggesting that Nus factors are unable to prevent all instances of Rho termination within the CRISPR array.

Our data indicate that CRISPR arrays in *S.* Typhimurium are protected from premature Rho termination by Nus factor association with RNAP via the BoxA sequences. However, this does not necessarily mean that CRISPR-Cas function is affected by anti-termination, since low levels of crRNA may be sufficient for Cas proteins to bind target DNA. We previously showed that the Cascade complex of Cas proteins in *E. coli* binds to DNA targets with as few as 5 bp between the crRNA and the target DNA. Consequently, the endogenous *E. coli* crRNAs direct Cascade binding to >100 chromosomal sites (21). The *E. coli* CRISPR-Cas system is a type I-E system, similar to that in *S.* Typhimurium. Given this similarity, we hypothesized that *S.* Typhimurium Cascade would also bind chromosomal targets, and that we could measure the effectiveness of different CRISPR array spacers by measuring the degree to which those spacers direct Cascade binding to chromosomal sites. We used ChIP-seq to measure $FLAG_3$-Cas5 association with the *S.* Typhimurium chromosome in cells expressing crRNAs from both CRISPR arrays. Note that we used a different constitutive promoter upstream of the CRISPR-II array to that described above for ChIP of SuhB, although the promoter was inserted at the same location (see Methods). As expected, we detected association of Cas5 with a large number (236) of chromosomal regions (Fig. 3A, Table S1). These binding sites are associated with five strongly enriched sequence motifs that correspond to a PAM sequence and the seed regions of spacers 1, 2, 3, 4, and 11 from the CRISPR-I array (Fig. 3B). These enriched motifs indicate that the optimal PAM in *S.* Typhimurium is AWG. We then attempted to identify the corresponding crRNA spacer for all 236 Cascade binding sites (see Methods). Thus, we were able to uniquely associate 152 binding sites with a single spacer from one of the two CRISPR arrays (Table S2). To test the

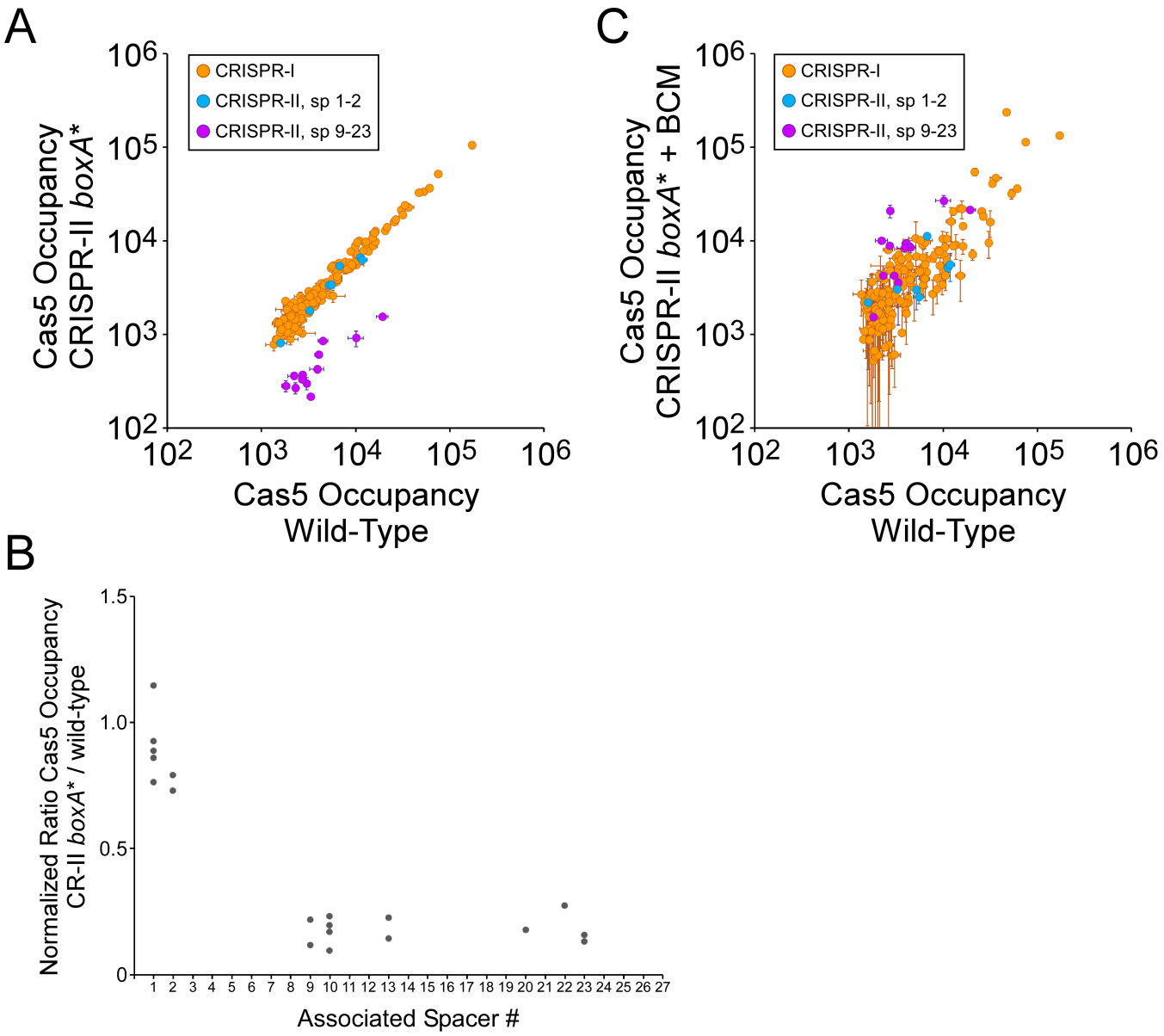

Fig. 4. The CRISPR-II *boxA* facilitates use of all spacers by preventing premature Rho termination. (A) Comparison of $FLAG_3$-Cas5 ChIP-seq occupancy at off-target chromosomal sites in cells with an intact CRISPR-II *boxA* (AMD678; *x*-axis), and cells with a single base-pair substitution in the CRISPR-II *boxA* (AMD685; *y*-axis). Cascade binding associated with spacers from CRISPR-I is indicated by orange datapoints. Cascade binding associated with spacers 1-2 from CRISPR-II is indicated by light blue datapoints. Cascade binding associated with spacers 9-23 from CRISPR-II is indicated by purple datapoints. Values plotted are the average of two independent biological replicates. Error bars represent one standard-deviation from the mean. **(B)** Normalized ratio of $FLAG_3$-Cas5 occupancy associated with spacers from CRISPR-II. Values are plotted according to the associated spacer, and are normalized to the average value for sites associated with spacers from CRISPR-I. **(C)** Comparison of $FLAG_3$-Cas5 ChIP-seq occupancy at off-target chromosomal sites in cells with an intact CRISPR-II *boxA* (AMD678; *x*-axis), and cells with a single base-pair substitution in the CRISPR-II *boxA* (AMD685) that were treated with bicyclomycin (BCM; *y*-axis).

quality of these spacer assignments, we repeated the ChIP-seq experiment in a strain where the CRISPR-I array had been deleted. Consistent with our spacer assignments, Cas5 association decreased specifically at all DNA sites associated with spacers from the CRISPR-I array (Fig. S2). By comparing the ChIP-seq data from wild-type and CRISPR-I deleted cells, we were able to unambiguously associate an additional 32 Cascade binding sites with a single spacer from one of the two CRISPR arrays (these binding sites had previously been associated with multiple possible spacers; Table S2).

We then measured $FLAG_3$-Cas5 association with the *S.* Typhimurium chromosome in cells containing a single base pair substitution in the *boxA* upstream of CRISPR-II that is expected to abrogate NusB/E association (Fig. 1B) (18–20). We observed no difference in Cas5 binding between the wild-type and mutant cells for sites associated with spacers from the CRISPR-I array, or sites associated with spacers 1-2 from the CRISPR-II array (Fig. 4A; Table S2). By contrast, mutation of the CRISPR-II *boxA* led to a large decrease in Cas5 binding to sites associated with spacers 9-23 from the CRISPR-II array (Fig. 4A-B; Table S2). This effect was reversed by addition of BCM to the *boxA* mutant cells (Fig. 4C; Table S2), indicating that reduced Cascade binding associated with spacers 9-23 of CRISPR-II is due to premature

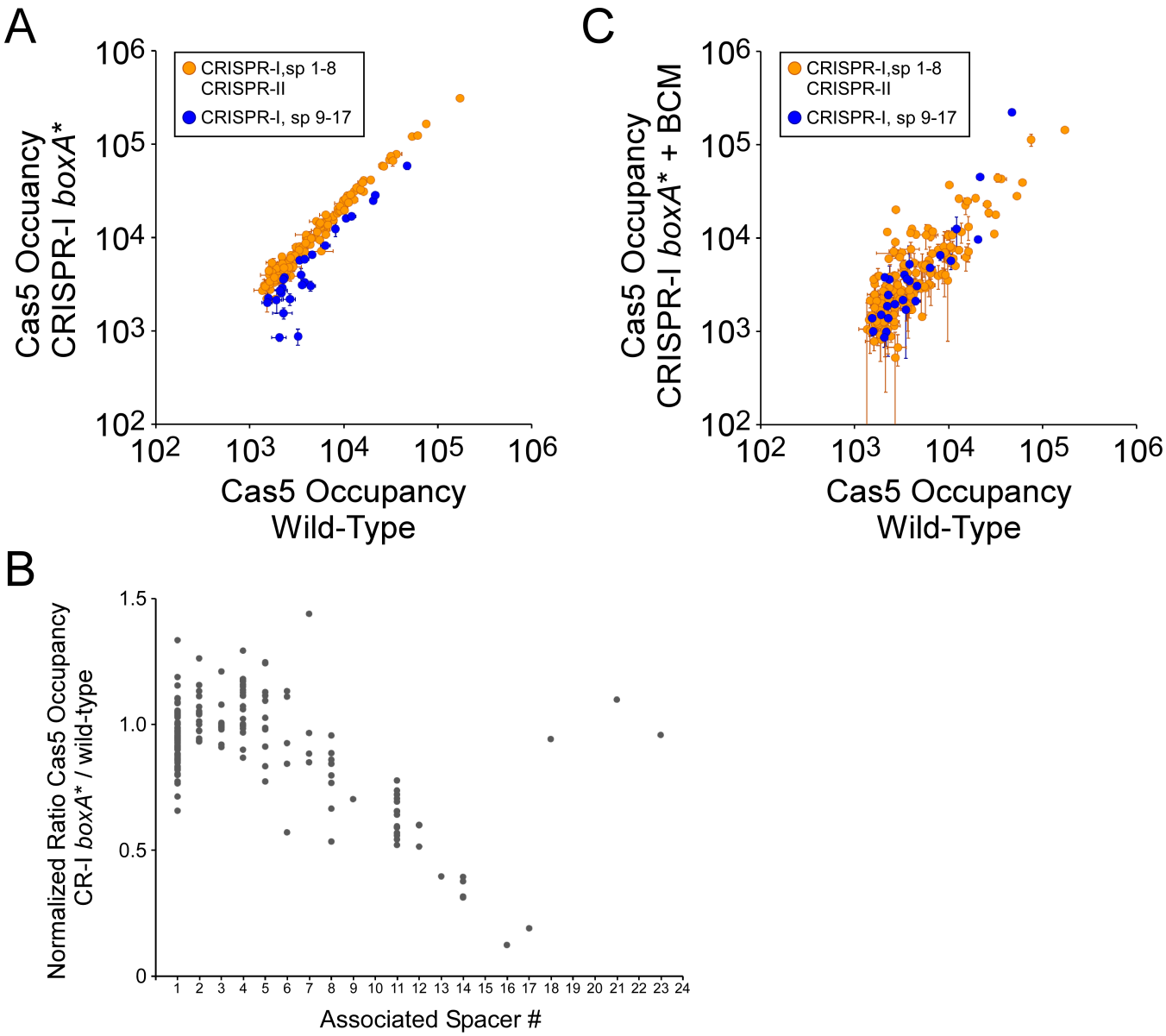

Fig. 5. The CRISPR-I *boxA* facilitates use of all spacers by preventing premature Rho termination. (A) Comparison of $FLAG_3$-Cas5 ChIP-seq occupancy at off-target chromosomal sites in cells with an intact CRISPR-I *boxA* (AMD678; *x*-axis), and cells with a single base-pair substitution in the CRISPR-I *boxA* (AMD684; *y*-axis). Cascade binding associated with spacers from CRISPR-II, and spacers 1-8 from CRISPR-I, is indicated by orange datapoints. Cascade binding associated with spacers 9-17 from CRISPR-I is indicated by dark blue datapoints. Values plotted are the average of two independent biological replicates. Error bars represent one standard-deviation from the mean. **(B)** Normalized ratio of $FLAG_3$-Cas5 occupancy associated with spacers from CRISPR-I. Values are plotted according to the associated spacer, and are normalized to the average value for sites associated with spacers from CRISPR-II. **(C)** Comparison of $FLAG_3$-Cas5 ChIP-seq occupancy at off-target chromosomal sites in cells with an intact CRISPR-I *boxA* (AMD678; *x*-axis), and cells with a single base-pair substitution in the CRISPR-I *boxA* (AMD684) that were treated with bicyclomycin (BCM; *y*-axis).

Rho termination of the array. Consistent with our reporter gene fusion data (Fig. 2B), addition of BCM to cells with an intact *boxA* led to an increase in Cascade binding associated with spacers 9-23 of the CRISPR-II array (Fig. S3; Table S2), supporting the notion that the BoxA prevents only a subset of premature Rho termination events.

We also measured $FLAG_3$-Cas5 association with the S. Typhimurium chromosome in cells containing a single base-pair substitution in the *boxA* upstream of the CRISPR-I array. Mutation of the CRISPR-I *boxA* had no impact on Cas5 binding to sites associated with spacers from the CRISPR-II array or spacers 1-5 from the CRISPR-II array (Fig. 5A; Table S2). By contrast, mutation of the CRISPR-I *boxA* led to a decrease in Cas5 binding to sites associated with spacers 9-17 from the CRISPR-I array (Fig. 5A), with the magnitude of the effect increasing as a function of the position of the spacer in the array (Fig. 5B). This effect was reversed by addition of BCM to the *boxA* mutant cells (Fig. 5C; Table S2), indicating that reduced Cascade binding using spacers 9-17 of CRISPR-I is due to premature Rho termination of the array. Unexpectedly, Cas5 binding associated with CRISPR-I spacers 18-23 was unaffected by the *boxA* mutation (Fig. 5B), suggesting the existence of an additional promoter within spacer 16 or 17. Overall, the effect of mutating the CRISPR-I *boxA* was smaller than that of mutating the CRISPR-II *boxA*, suggesting that more RNAP can evade Rho termination at CRISPR-I.

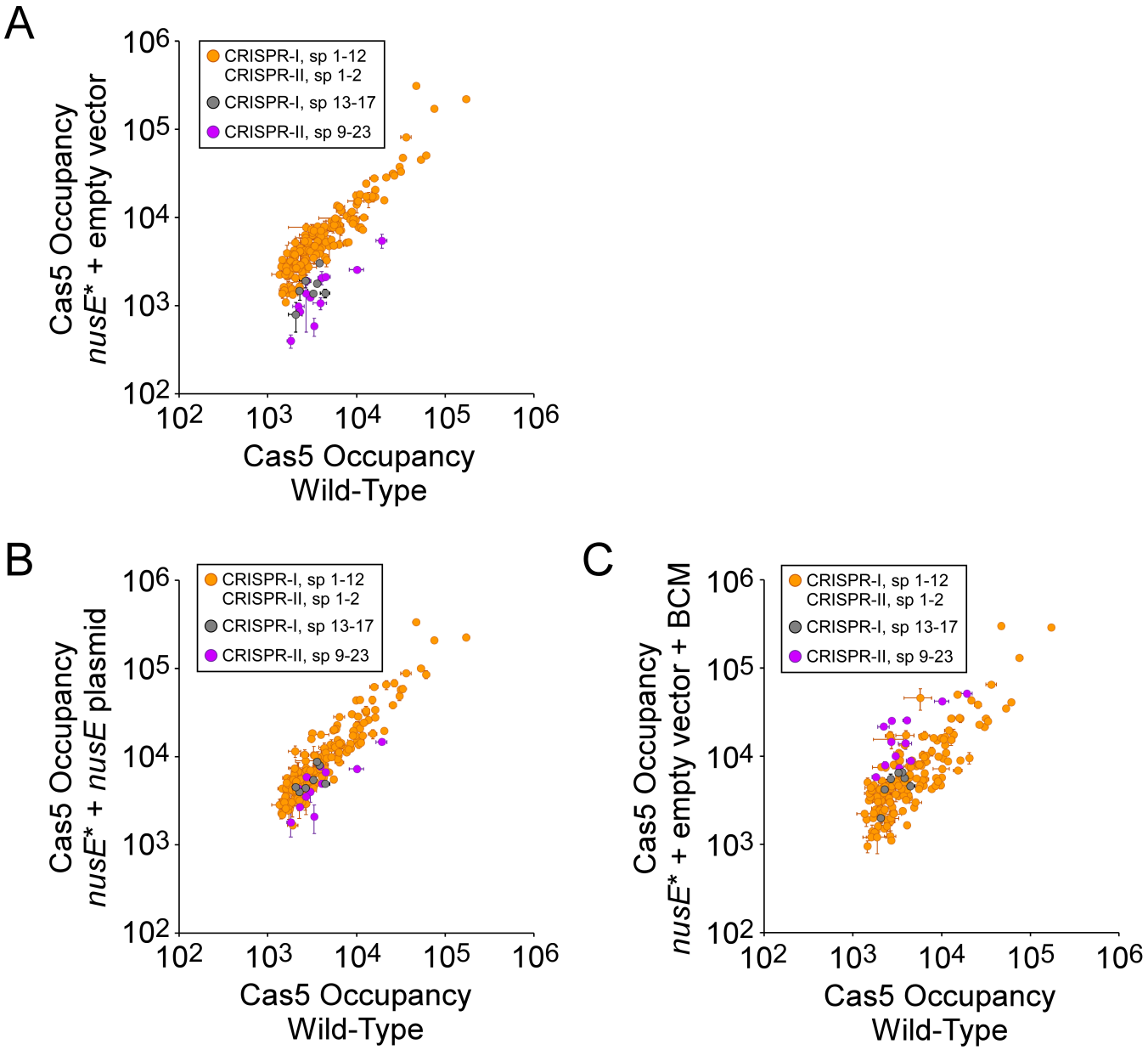

Fig. 6. Nus factors facilitate use of all spacers by preventing premature Rho termination. (A) Comparison of $FLAG_3$-Cas5 ChIP-seq occupancy at off-target chromosomal sites in *nusE*$^+$ (AMD678; *x*-axis), and *nusE* mutant cells (AMD698) containing an empty vector (pBAD24-amp; *y*-axis). Cascade binding associated with spacers 1-12 from CRISPR-I, and spacers 1-2 from CRISPR-II, is indicated by orange datapoints. Cascade binding associated with spacers 13-17 from CRISPR-I is indicated by gray datapoints. Cascade binding associated with spacers 9-23 from CRISPR-II is indicated by purple datapoints. Values plotted are the average of two independent biological replicates. Error bars represent one standard-deviation from the mean. **(B)** Comparison of $FLAG_3$-Cas5 ChIP-seq occupancy at off-target chromosomal sites in *nusE*$^+$ (AMD678; *x*-axis) or *nusE* mutant cells (AMD698) expressing a plasmid-encoded (pAMD239) copy of wild-type *nusE* (*y*-axis). **(C)** Comparison of $FLAG_3$-Cas5 ChIP-seq occupancy at off-target chromosomal sites in *nusE*$^+$ cells (AMD678; *x*-axis), and *nusE* mutant cells containing an empty vector (AMD698) and treated with bicyclomycin (BCM; *y*-axis).

To confirm that the effects of mutating the *boxA* sequences are due to the inability to recruit the Nus factor complex, we measured $FLAG_3$-Cas5 association with the *S*. Typhimurium chromosome in cells containing a defective Nus factor. We were unable to delete either *nusB* or *suhB*, suggesting that all Nus factors are essential in *S*. Typhimurium. Hence, we introduced a single base substitution into the chromosomal copy of the *nusE* gene, resulting in the N3H amino acid substitution. The equivalent change in the *E. coli* NusE leads to a defect in Nus factor complex function (18). Note that mutation of *nusE* also resulted in a ∼147 kb deletion of the chromosome at an unlinked site. Mutation of *nusE* led to a decrease in Cascade binding to sites associated with spacers 9-23 from the CRISPR-II array and spacers 13-17 from the CRISPR-I array (Fig. 6A; Table S2). The effect of the *nusE* mutation on Cascade binding was smaller than that of the *boxA* mutations; however, the Nus factor complex is likely to retain partial function in the *nusE* mutant strain. To rule out the possibility that the chromosomal deletion in the *nusE* mutant strain was responsible for the effect on Cascade binding, we complemented the strain with plasmid-expressed wild-type NusE. Complementation led to a significant increase in the

binding of Cas5 to sites associated with spacers 9-23 from the CRISPR-II array and spacers 13-17 from the CRISPR-I array (Fig. 6B; Table S2), consistent with a specific effect of mutating *nusE*. Similarly, addition of BCM to the *nusE* mutant cells resulted in a significant increase in the binding of Cas5 to sites associated with spacers 9-23 from the CRISPR-II array and spacers 13-17 from the CRISPR-I array (Fig. 6C; Table S2), indicating that reduced Cascade binding using spacers 9-23 of CRISPR-II and spacers 13-17 from the CRISPR-I array in the *nusE* mutant is due to premature Rho termination of the arrays. Based on the effects of mutating the *boxA* sequences and the effect of mutating nusE, we conclude that Nus factors prevent premature transcription termination of both *S*. Typhimurium CRISPR arrays, and that this has a direct impact on the ability of *S*. Typhimurium to use the majority of spacers in the CRISPR arrays.

Given that Rho is found in 90% of bacterial species (22), and Nus factors are broadly conserved, we reasoned that species other than *S*. Typhimurium may use the Nus factor complex to facilitate expression of their CRISPR arrays. Hence, we performed a phylogenetic analysis of sequences associated with CRISPR arrays. Among sequences found between a *cas2* gene and a downstream CRISPR array in 187 bacterial genera (each genus represented only once; see Methods for details of sequence selection), there was a strongly enriched sequence motif that is a striking match to the known *boxA* consensus (Fig. 7; Table S3) (15, 18). This motif was detected in 52 of the 187 genera examined, predominantly for genera in the *Proteobacteria*, *Bacteroidetes* and *Cyanobacteria* phyla. The *boxA* consensus is known to vary between species, diverging from the *E. coli* sequence with increasing evolutionary distance (15). Hence, it is possible that *boxA* sequences were missed upstream of CRISPR arrays in bacterial genera that are less closely related to *E. coli*. Given that the *Proteobacteria* likely share a more recent common ancestor with phyla other than the *Bacteroidetes* and *Cyanobacteria* (23), we propose that *boxA* sequences upstream of CRISPR arrays either (i) evolved in a very early bacterial ancestor but were not detected in our analysis due to *boxA* sequence divergence, or (ii) evolved independently in multiple bacterial lineages. Regardless of its evolutionary history, it is clear that this phenomenon is distributed broadly across the bacterial kingdom.

Conclusions

In summary, we have identified a mechanism to prevent premature Rho-dependent transcription termination of CRISPR arrays that is conserved across a broad range of bacterial species. Our data indicate that anti-termination of CRISPR arrays by the Nus factor complex can be essential to allow use of all spacers, and thus highlight the importance of transcription anti-termination in the process of CRISPR biogenesis. While BoxA-mediated anti-termination is likely used to prevent premature termination of CRISPR array transcription in hundreds of species, many CRISPR arrays do not appear to be associated with a boxA. By contrast, Rho homologues are found in 90% of bacterial species (22). An alternative strategy to circumvent the problems associated with Rho would be to avoid sequences in CRISPR arrays that can function as Rho loading sites. However, this would limit the sequence space for spacers that can be added to a CRISPR array, since any newly acquired spacer containing a strong Rho loading site would likely inactivate downstream spacers in the array. Similarly, limiting CRISPR array length would be an imperfect strategy to circumvent the problem of Rho termination, since this would limit the potential for acquisition of additional spacers in the array. Hence, we conclude that additional anti-termination mechanisms likely exist for bacterial CRISPR arrays.

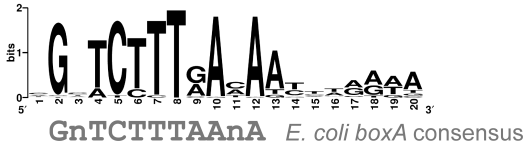

Fig. 7. Widespread conservation of *boxA* sequences upstream of CRISPR arrays in diverse bacterial genera. Sequence motif found by MEME (30) to be significantly enriched upstream of 52 CRISPR arrays spread across the bacterial kingdom (from 187 tested; MEME *E*-value = $7.1e^{-69}$).

Materials and Methods

Strains and plasmids. All strains, plasmids and oligonucleotides used in this study are listed in Tables S4 and S5, respectively. All strains are derivatives of *Salmonella* enterica subspecies *enterica* serovar Typhimurium 14028s (24). Strains AMD678, AMD679, AMD684, AMD685, AMD698, AMD710, and AMD711 were generated using the FRUIT recombineering method (25). To construct AMD678, oligonucleotides JW8576 and JW8577 were used to N-terminally $FLAG_3$-tag *cas5*. Then, oligonucleotides JW8610 + JW8611 and JW8797 + JW8798 were used for insertion of a constitutive $P_{KAB\text{-}TG}$ promoter (26) in place of *cas3*, and upstream of the CRISPR-II array, respectively. AMD679, AMD684, AMD685, and AMD698 are derivatives of AMD678 and were generated using oligonucleotides (i) JW8913 and JW8914 to amplify *thyA* and replace the CRISPR-I array, (ii) JW8904-JW8907 for introduction of a CRISPR-I *boxA* mutation (C4A), (iii) JW8568-JW8571 for introduction of a CRISPR-II *boxA* mutation (C4A), and (iv) JW9441-JW9444 for introduction of a *nusE* mutation (N3H). AMD710 and AMD711 were derived from AMD678 and AMD685, respectively. They were constructed using oligonucleotides JW9355 + JW9356 to C-terminally TAP-tag *suhB*, with pVS030 (see below) as a PCR template. The constitutive $P_{KAB\text{-}TG}$ promoter (26) upstream of the CRISPR-II array was replaced by P_{J23119}-*thyA*, using oligonucleotides JW9627 + JW9628 and a strain containing thyA driven by the P_{J23119} promoter as a template for colony PCR.

The plasmid pAMD239 was constructed by using oligonucleotides JW9674 and JW9675 to amplify *nusE* from wild-type 14028s. The resulting DNA fragment was cloned into pBAD24 (27) digested with *Hin*dIII and *Nhe*I. For the construction of pVS030, duplicate sets of TAP tags were colony PCR-amplified from a TAP-tagged strain of *E. coli* (28) using oligonucleotides JW6401 + JW6445 and JW6448 +

JW6406. Oligonucleotides JW6446 + JW6447 were used to colony PCR-amplify *thyA*, as described previously (25). All three DNA fragments were cloned into the pGEM-T plasmid (Promega) digested with *Sal*I and *Nco*I.

Plasmids pGB231, pGB237, pGB250, pGB256 were constructed by PCR-amplifying and cloning a truncated *S*. Typhimurium 14028s CRISPR-II array (to the 2nd or 11th spacer) and 292 bp of upstream sequence into the *Nsi*I and *Nhe*I sites of the pJTW064 plasmid (29) with either a wild-type or mutant *boxA* sequence. A constitutive promoter was introduced upstream of the CRISPR-II sequence, creating a transcriptional fusion to *lacZ*. The following primers were used to PCR-amplify each CRISPR-II truncation: JW9381 + JW9383 (pGB231 and pGB237), JW9381 + JW9605 (pGB250 and pGB256). Fusions carrying *boxA* mutations were made by amplifying CRII from AMD678 template, which contains a *boxA*(C4A) mutation.

ChIP-qPCR. Strains 14028s, AMD678, AMD710 and AMD711 were subcultured 1:100 in LB and grown to an OD_{600} of 0.5-0.8. ChIP and input samples were prepared and analyzed as described previously (29). For ChIP of β, 1 µl anti-β (RNA polymerase subunit) antibody was used. For ChIP of SuhB-TAP, IgG Sepharose was used in place of Protein A Sepharose. Enrichment of ChIP samples was determined using quantitative real-time PCR with an ABI 7500 Fast instrument, as described previously (30). Enrichment was calculated relative to a control region, within the *sseJ* gene (PCR-amplified using oligonucleotides JW4477 + JW4478), which is expected to be free of SuhB and RNA polymerase. Oligonucleotides used for qPCR amplification of the region within the CRISPR-I array were JW9305 + JW9306. Oligonucleotides used for qPCR amplification of the region surrounding the CRISPR-II *boxA* were JW9329 + JW9330. Oligonucleotides used for qPCR amplification of the region within *rpsA* were JW9660 + JW9661. Occupancy values represent background-subtracted enrichment relative to the control region.

ChIP-seq. All ChIP-seq experiments were performed in duplicate. Cultures were inoculated 1:100 in LB with fresh overnight cultures of AMD678, AMD679, AMD684, and AMD685. Cultures of AMD678, AMD684 and AMD685 were split into two cultures at an OD_{600} of 0.1, and bicyclomycin was added to one of the two cultures to a concentration of 20 µg/mL. At an OD_{600} of 0.5-0.8, cells were processed for ChIP-seq of $FLAG_3$-Cas5, following a protocol described previously (29). For ChIP-seq using derivatives of the *nusE* mutant strain (AMD698), AMD698 containing either empty pBAD24 or pAMD239, was subcultured 1:100 in LB supplemented with 100 µg/mL ampicillin and 0.2% arabinose. AMD698 + pBAD24 cultures were split into two cultures at an OD_{600} of 0.1. Bicyclomycin was added to one of the two cultures to a concentration of 20 µg/mL. At an OD_{600} of 0.5-0.8, cells were processed for ChIP-seq of $FLAG_3$-Cas5, following a protocol described previously (29).

ChIP-seq data analysis. Peak calling from ChIP-seq data was performed as previously described (31). To assign specific crRNA spacer sequences to Cascade binding sites, we extracted 101 bp regions centered on each ChIP-seq peak for ChIP-seq data generated from AMD678. Overlapping regions were merged and the central position was used as a reference point for downstream analysis. We refer to this position as the "peak center". We then searched each ChIP-seq peak region for a perfect match to positions 1-5 of each spacer from the CRISPR-I and CRISPR-II arrays, in addition to an immediately adjacent AAG or ATG sequence (the expected PAM sequence). Additionally, we searched each ChIP-seq peak region for a perfect match to positions 1-5 and positions 7-8 of each spacer from the CRISPR-I and CRISPR-II arrays. Spacers were only assigned to a ChIP-seq peak if they had a unique match to a spacer sequence. This yielded 152 uniquely assigned peak-spacer combinations from the 236 ChIP-seq peaks. Enriched sequence motifs within the 236 peak regions (Figure 3B) were identified using MEME (v5.0.1, default parameters) (32).

To determine relative sequence read coverage at each ChIP-seq peak center, we used Rockhopper (33) to determine relative sequence coverage at every genomic position on both strands for each ChIP-seq dataset. We then summed the relative sequence read coverage values on both strands for each peak center position to give peak center coverage values (Table S2). To refine the assignment of spacers to ChIP-seq peaks, we compared peak center coverage values for each peak in ChIP-seq datasets from AMD678 (CRISPR-I^+) and AMD679 (ΔCRISPR-I) strains. We calculated ratio of peak center coverage values in the first replicates of AMD679:AMD678 data, and repeated this for the second replicates, generating two ratio values. Based on ratio values assigned to peak centers that were already uniquely assigned to a spacer, we conservatively assumed that peak centers with both ratios <0.2 should be assigned a CRISPR-I spacer, whereas peak centers with both ratios >1.0 should be assigned a CRISPR-II spacer. Thus, we were able to uniquely assign spacers to an additional 32 peak centers that had previously been assigned multiple spacers. For example, if a peak center had previously been assigned two spacers from CRISPR-I and one spacer from CRISPR-II, we could uniquely assign the spacer from CRISPR-II of both rations were >1.0.

For data plotted in Figures 3C-D, S2, S3, S4A-C and S5A-C, values for peak center coverage were normalized in one replicate (two biological replicates were performed for all ChIP-seq experiments) by summing the values at all peak centers to be analyzed (i.e. peak centers that could be uniquely assigned to a spacer) and multiplying values in the second replicate by a constant such that the summed values for each replicate were the same.

β-galactosidase assays. S. Typhimurium 14028s containing pGB231, pGB237, pGB250, or pGB256 was grown at 37 °C in LB medium to an OD_{600} of 0.4-0.6. 810 µL of culture was pelleted and resuspended in 810 µL of Z buffer (0.06 M Na_2HPO_4, 0.04 M NaH_2PO_4, 0.01 M KCl, 0.001 M $MgSO_4$) + 50 mM β-mercaptoethanol + 0.001% SDS + 20 µL chloroform. Cells were lysed by brief vortexing. β-galactosidase reactions were started by adding 160 µL 2-Nitrophenyl β-D-

galactopyranoside (4 mg/mL) and stopped by adding 400 µL of 1 M Na_2CO_3. The duration of the reaction and OD_{420} readings were recorded. β-galactosidase activity was calculated as 1000 * (A_{420}/(A_{600})(time [min])).

Phylogenetic analysis of CRISPR array boxA *conservation.* Cas2 protein sequences were extracted from the PF09707 and PF09827 PFAM families (34, 35). These sequences were then used to search a local collection of bacterial genome sequences using tBLASTn (36). We then selected the 612 sequences with perfect matches to full-length Cas2 sequences. We further refined this set of sequences by arbitrarily selecting only one sequence per genus. We then extracted 300 bp downstream of each *cas2* gene. We trimmed the 3' end of each sequence at the first instance of a 15 bp repeat, which is likely to be the beginning of a CRISPR array, since *cas2* is often found upstream of a CRISPR array. Any sequences that lacked a 15 bp repeat, or were trimmed to <20 bp, were discarded. We used MEME (v5.0.1, default parameters) (32) to identify enriched sequence motifs in the remaining 187 sequences (Table S3).

ACKNOWLEDGEMENTS

We thank the Wadsworth Center Applied Genomic Technologies Core Facility for sequencing. We thank the Wadsworth Center Bioinformatics Core Facility for bioinformatic support. We thank the Wadsworth Center Media and Tissue Culture and Glassware Core Facilities. We thank Shailab Shrestha, Todd Gray, Keith Derbyshire, and Randy Morse for helpful discussions. We thank Vimi Singh for creating the TAP-tagging construct used in this study. This study was supported by NIH Grants AI126416 and GM122836 (to J.T.W.)

Raw ChIP-seq data are available from EBI ArrayExpress/ENA

References

1. A. V. Wright, J. K. Nuñez, J. A. Doudna, Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell. 164, 29–44 (2016).
2. P. Mitra, G. Ghosh, M. Hafeezunnisa, R. Sen, Rho Protein: Roles and Mechanisms. Annu. Rev. Microbiol. (2017), doi:10.1146/annurev-micro-030117-020432.
3. J. M. Peters et al., Rho and NusG suppress pervasive antisense transcription in Escherichia coli. Genes. Dev. 26, 2621–2633 (2012).
4. J. T. Wade, D. C. Grainger, Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Micro. 12, 647–653 (2014).
5. M. Lybecker, I. Bilusic, R. Raghavan, Pervasive transcription: detecting functional RNAs in bacteria. Transcription. 5, e944039 (2014).
6. C. Nadiras, E. Eveno, A. Schwartz, N. Figueroa-Bossi, M. Boudvillain, A multivariate prediction model for Rho-dependent termination of transcription. Nucleic Acids Res. (2018), doi:10.1093/nar/gky563.
7. N. Singh et al., SuhB Associates with Nus Factors To Facilitate 30S Ribosome Biogenesis in Escherichia coli. mBio. 7, e00114-16 (2016).
8. B. M. Burmann et al., A NusE:NusG complex links transcription and translation. Science. 328, 501–504 (2010).
9. M. Bubunenko et al., Nus transcription elongation factors and RNase III modulate small ribosome subunit biogenesis in Escherichia coli. Mol. Microbiol. 87, 382–393 (2012).
10. M. Torres, J. M. Balada, M. Zellars, C. Squires, C. L. Squires, In vivo effect of NusB and NusG on rRNA transcription antitermination. J. Bacteriol. 186, 1304–1310 (2004).
11. C. L. Squires, J. Greenblatt, J. Li, C. Condon, C. L. Squires, Ribosomal RNA antitermination in vitro: requirement for Nus factors and one or more unidentified cellular components. Proc Natl Acad Sci U S A. 90, 970–4 (1993).
12. S. Aksoy, C. L. Squires, C. Squires, Evidence for antitermination in Escherichia coli rRNA transcription. J. Bacteriol. 159, 260–264 (1984).
13. S. C. Li, C. L. Squires, C. Squires, Antitermination of E. coli rRNA transcription is caused by a control region segment containing lambda nut-like sequences. Cell. 38, 851–860 (1984).
14. R. Sen, J. Chalissery, G. Muteeb, E. P. 2008 Sep;3(1), Nus Factors of Escherichia coli. EcoSal Plus. 3 (2008).
15. K. B. Arnvig et al., Evolutionary comparison of ribosomal operon antitermination function. J. Bacteriol. 190, 7251–7257 (2008).
16. W. W. Navarre et al., Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science. 313, 236–238 (2006).
17. M. L. Luo, A. S. Mullis, R. T. Leenay, C. L. Beisel, Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. Nucleic Acids Res. 43, 674–681 (2014).
18. G. Baniulyte et al., Identification of regulatory targets for the bacterial Nus factor complex. Nat Commun. 8, 2027 (2017).
19. K. L. Berg, C. Squires, C. L. Squires, Ribosomal RNA operon antitermination. Function of leader and spacer region box B-box A sequences and their conservation in diverse micro-organisms. J. Mol. Biol. 209, 345–358 (1989).
20. J. R. Nodwell, J. Greenblatt, Recognition of boxA antiterminator RNA by the E. coli antitermination factors NusB and ribosomal protein S10. Cell. 72, 261–268 (1993).
21. L. A. Cooper, A. M. Stringer, J. T. Wade, Determining the Specificity of Cascade Binding, Interference, and Primed Adaptation In Vivo in the Escherichia coli Type I-E CRISPR-Cas System. MBio. 9 (2018), doi:10.1128/mBio.02100-17.
22. F. D'Heygère, M. Rabhi, M. Boudvillain, Phyletic distribution and conservation of the bacterial transcription termination factor Rho. Microbiology (Reading, Engl.). 159, 1423–1436 (2013).
23. M. Bern, D. Goldberg, Automatic selection of representative proteins for bacterial phylogeny. BMC Evol. Biol. 5, 34 (2005).
24. T. Jarvik, C. Smillie, E. A. Groisman, H. Ochman, Short-term signatures of evolutionary change in the Salmonella enterica serovar Typhimurium 14028 genome. J. Bacteriol. 192, 560–567 (2010).
25. A. M. Stringer et al., FRUIT, a scar-free system for targeted chromosomal mutagenesis, epitope tagging, and promoter replacement in Escherichia coli and Salmonella enterica. PLoS One. 7, e44841 (2012).
26. T. Burr, J. Mitchell, A. Kolb, S. Minchin, S. Busby, DNA sequence elements located immediately upstream of the -10 hexamer in Escherichia coli promoters: a systematic study. Nucleic Acids Res. 28, 1864–1870 (2000).
27. L. M. Guzman, D. Belin, M. J. Carson, J. Beckwith, Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. J. Bacteriol. 177, 4121–4130 (1995).
28. G. Butland et al., Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature. 433, 531–537 (2005).
29. A. M. Stringer et al., Genome-Scale Analyses of Escherichia coli and Salmonella enterica AraC Reveal Non-Canonical Targets and an Expanded Core Regulon. J. Bacteriol. 196, 660–671 (2014).
30. O. Aparicio et al., in Current Protocols in Molecular Biology, F. M. Ausubel et al., Eds. (John Wiley & Sons, Inc., Hoboken, NJ, 2005), pp. 21.3.1-21.3.33.
31. D. M. Fitzgerald, R. P. Bonocora, J. T. Wade, Comprehensive Mapping of the Escherichia coli Flagellar Regulatory Network. PLOS Genet. 10, e1004649 (2014).
32. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 2, 28–36 (1994).
33. R. McClure et al., Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res. 41, e140 (2013).
34. E. L. Sonnhammer, S. R. Eddy, R. Durbin, Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 28, 405–420 (1997).
35. R. D. Finn et al., The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279-285 (2016).
36. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

Supplementary Information

Tables S1-5 are provided as a single Excel file.

Table S1. List of ChIP-seq peaks
Table S2. Assignment of ChIP-seq peak centers to spacers, and relative Cas5 occupancy at each peak center
Table S3. Sequences used for phylogenetic analysis of CRISPR array upstream regions
Table S4. List of strains and plasmids used in this study
Table S5. List of oligonucleotides used in this study

Figures S1-3 are provided as a single pdf.

Fig. S1. The CRISPR-I array is co-transcribed with the upstream cas gene operon
Fig. S2. Confirmation of spacer assignments to Cas5 binding sites
Fig. S3. Inhibition of Rho in cells with an intact *boxA* increases the use of spacers 9-23 of CRISPR-II