RNA BIOLOGY
2022, VOL. 19, NO. 1, 678–685
https://doi.org/10.1080/15476286.2022.2067714

Taylor & Francis
Taylor & Francis Group

BRIEF COMMUNICATION

OPEN ACCESS Check for updates

Identifying candidate structured RNAs in CRISPR operons

Brayon J. Fremin[a,b] and Nikos C. Kyrpides[a,b]

[a]Department of Energy, Joint Genome Institute, Berkeley, CA, USA; [b]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ABSTRACT
Noncoding RNAs with secondary structures play important roles in CRISPR-Cas systems. Many of these structures likely remain undiscovered. We used a large-scale comparative genomics approach to predict 156 novel candidate structured RNAs from 36,111 CRISPR-Cas systems. A number of these were found to overlap with coding genes, including palindromic candidates that overlapped with a variety of Cas genes in type I and III systems. Among these 156 candidates, we identified 46 new models of CRISPR direct repeats and 1 tracrRNA. This tracrRNA model occasionally overlapped with predicted *cas9* coding regions, emphasizing the importance of expanding our search windows for novel structure RNAs in coding regions. We also demonstrated that the antirepeat sequence in this tracrRNA model can be used to accurately assign thousands of predicted CRISPR arrays to type II-C systems. This study highlights the importance of unbiased identification of candidate structured RNAs across CRISPR-Cas systems.

ARTICLE HISTORY
Received 22 February 2022
Revised 8 April 2022
Accepted 11 April 2022

KEYWORDS
CRISPR; structured RNA; comparative genomics

Introduction

CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR associated) systems are utilized by bacteria and archaea to protect themselves against infectious agents. These systems use RNA-guided nucleases to target and cut specific sequences of double-stranded DNA [1]. Noncoding RNAs play central roles in CRISPR-Cas systems. CRISPR RNAs (crRNAs) are noncoding RNAs that are transcribed and processed by enzymes encoded in the CRISPR sequence array, which contains direct repeats separated by spacers. The crRNAs guide Cas nucleases to their target DNA sequences and are found in all known CRISPR-Cas systems [2]. Transactivating CRISPR RNAs (tracrRNAs) are noncoding RNAs encoded by type II and some type V CRISPR-Cas systems that aid in maturation of crRNAs and DNA cleavage by CRISPR-Cas9 [3,4]. Short complementarity untranslated RNAs (scoutRNAs) are recently discovered noncoding RNAs that assemble with Cas12c/d and crRNA to function as a DNA-targeting complex [5]. Thus, discovery of additional noncoding RNAs associated with CRISPR-Cas systems will likely be important for understanding the mechanisms and adaptation of CRISPR-Cas systems.

All of these noncoding RNAs associated with CRISPRs have been shown or predicted to form secondary structures [5,6]. Currently, there are 64 families of direct repeats and one family of tracrRNAs in Rfam [7]. More diversity exists within crRNAs and tracrRNAs that existing models do not capture. It is likely that other noncoding RNAs exist that play essential roles in CRISPR-Cas systems that have yet to be discovered, and these noncoding RNAs may also form secondary structures [5]. Additionally, there likely exists substantial diversity in structures within crRNAs and tracrRNAs that have yet to be identified. Building additional models would be beneficial both to characterize the structures as well as better search for them in genomes. There is also precedence for regulatory RNAs being embedded in bacterial coding regions [8–10]. Because most of the focus is on intergenic regions, regulatory RNAs that overlap genes tend to be overlooked. However, this is an important consideration from a genetic engineering perspective; perhaps upon codon optimization of a Cas gene, for example, the structure and function of an essential overlapping noncoding RNA is disrupted. This perspective motivated us to predict candidate structured RNAs and include coding regions in our analyses. Though comparative genomics approaches have yet to be applied to Cas operons and CRISPRs at large-scale, it has previously been a useful approach to predict candidate structured RNAs in microbiomes [11–13].

In this work, we used a comparative genomics approach to predict candidate structured RNAs across 15,144 Cas operons, 21,141 associated CRISPRs, and 20,967 orphan Cas operons from diverse ecosystems. This approach involved clustering conserved regions within CRISPRs and Cas operons, predicting possible structures, and assessing possible structures for evidence of covariation, which would indicate evolutionary constraint to preserve the structure. Overall, our pipeline predicted 156 novel candidate structured RNAs, including 1 tracrRNA. Of these 156 candidate structured RNAs, 99 overlapped coding regions, 46 were novel direct repeats,and 11 were in intergenic regions. In addition to substantially expanding upon the diversity of known RNA structures, this approach predicted palindromic candidates overlapping Cas

CONTACT Brayon J. Fremin bfremin@lbl.gov Department of Energy, Joint Genome Institute, Berkeley, CA, USA
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15476286.2022.2067714

genes and novel candidates in intergenic regions across diverse CRISPR-Cas systems. Additionally, we showed that the antirepeat region of our novel tracrRNA model can be used to accurately assign 4,661 CRISPR arrays to type II-C systems based on homology between the array repeats and tracrRNA antirepeats.

Results

CRISPRCasTyper [14] was used to predict CRISPRs and Cas operons in ~25 million contigs (>3kb) from 29,521 publicly available and published metagenomics assemblies in IMG/M [15–38]. We identified 15,144 Cas operons associated with 21,141 nearby CRISPR arrays and an additional 20,967 orphan Cas operons, which were all used to predict candidate structured RNAs. To avoid false positives, we did not consider isolated CRISPR arrays or putative Cas operons in these analyses.

The first step for predicting candidate structured RNAs was to identify conserved regions along these Cas operons and CRISPRs. Conserved regions were identified using all versus all BLASTn [39], querying all Cas operons and associated CRISPR arrays against themselves (Fig. 1A). We filtered these blast results to exclude 100% identity matches, hits that span less than 30 bases in length, and hits with bit scores below 20. We set these filters because we ultimately wanted to align homologous regions that contained nucleotide differences to assess covariation. We clustered homologous regions into 11,546 clusters using overcluster2 with default settings. The next step was predicting structures. Using CMfinder [40,41], we generated motifs for 7,173 of these clusters. Using RNAphylo, a tool using phylogenetic models to score alignments, we found that 1,741 clusters contained motifs with an RNAphylo p score of 10 or greater. Additionally, we found that 717 of these contained at least one significant covarying base using R-scape [42]. Using cmsearch [40], we determined which of these alignments significantly (E value $< 1 \times 10^{-6}$) hit at least three regions near Cas operons and associated CRISPRs. We removed duplicates that hit any of the same regions as another candidate structured RNA, selecting the longest candidate. This resulted in a set of 159 candidate structured RNAs. Three of these 159 candidate structures were CRISPR direct repeats already found in Rfam [7]. In fact, Rfam contains 64 models of CRISPR direct repeats and 34 of these were identified in the 36,111 CRISPR-Cas systems we searched. 31 of these models did not meet the stringent phylogenetic and covariation thresholds we set, suggesting a high false negative rate using our approach.

Using this comparative genomics approach resulted in a finalized set of 156 novel candidate structured RNAs. These 156 novel structural RNAs were identified in 6,509 instances from 36,111 systems searched (Fig. 1A, File S1,

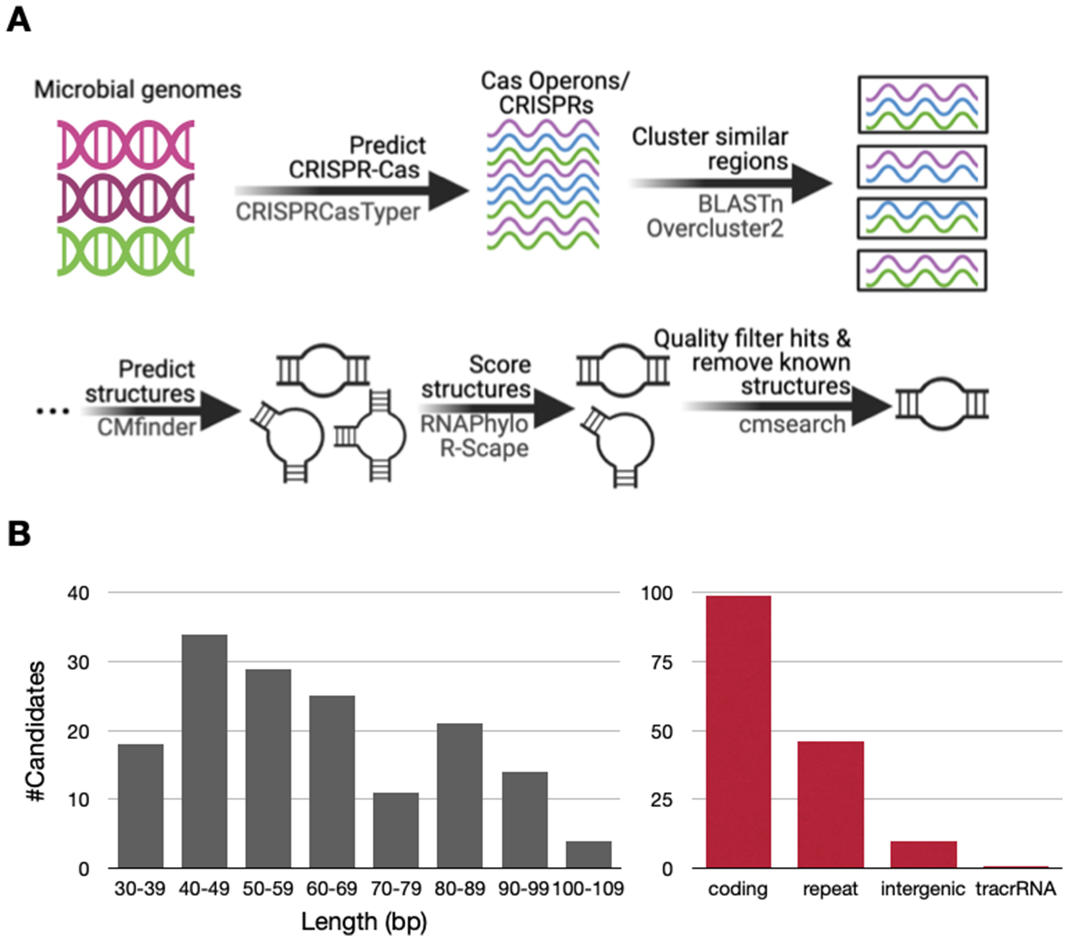

Figure 1. Prediction of candidate structured RNAs.

File S2). The average length was 62 bases (range 32 to 109 bases) (Table S1, Fig. 1B). We taxonomically classified 150 candidates to Bacteria and 6 to Archaea. Furthermore, 91 candidates were identified in *Proteobacteria*, 26 in *Firmicutes*, and 12 in *Bacteroidetes* (Table S1). These candidates also belonged to a diversity of subtypes. For example, we identified 34, 3, 15, 1, 4, and 2 candidates in I-C, II-A, III-A, IV-A1, V-A, and VI-B1 systems, respectively (Table S1). We binned these candidates into four categories: candidates that overlapped coding regions, candidates that modelled CRISPR direct repeats, candidates found exclusively in other intergenic regions, and candidates likely to be tracrRNAs (Table S1, Fig. 1B). Below we highlight interesting candidate structured RNAs for each of these categories.

Upon first inspecting candidates that overlapped coding regions, we identified palindromic candidate structured RNAs that overlapped *cas* genes. CRISPRCas_133 was a palindromic candidate structured RNA identified in 15 CRISPR-Cas type I-B systems that overlapped *cas6* near the end of the gene (typically overlapping the stop codon). It was predominately classified to *Bacteroidetes* and was found twice in the human digestive system, five times in freshwater, and four times in endoliths (Fig. 2, Table S1). CRISPRCas_135 was found in 6 CRISPR-Cas type I-F systems and overlapped *cas3*. It was predominately classified to *Firmicutes* and was found in diverse ecosystems, including the human digestive system and hydrothermal vents. CRISPRCas_94 was found in 13 CRISPR-Cas type I-B systems and also overlapped *cas3*. It was classified as *Firmicutes* and found in clay. Three other palindromic candidate structured RNAs were also predicted to overlap *cas3*; these candidates also occurred in a similar relative position along the gene. CRISPRCas_122 was found in 19 CRISPR-Cas type III-B systems and overlapped *cas10*. It was classified as *Proteobacteria* and has so far only been found in bioreactor samples (Fig. 2). Two other palindromic candidates also overlapped *cas10* in a similar relative position. In addition to these examples, two palindromic candidate structured RNAs overlapped *cas7*, both near the middle of the gene. One palindromic candidate overlapped *cas1* closer to the 3’ end of the gene. One palindromic candidate overlapped

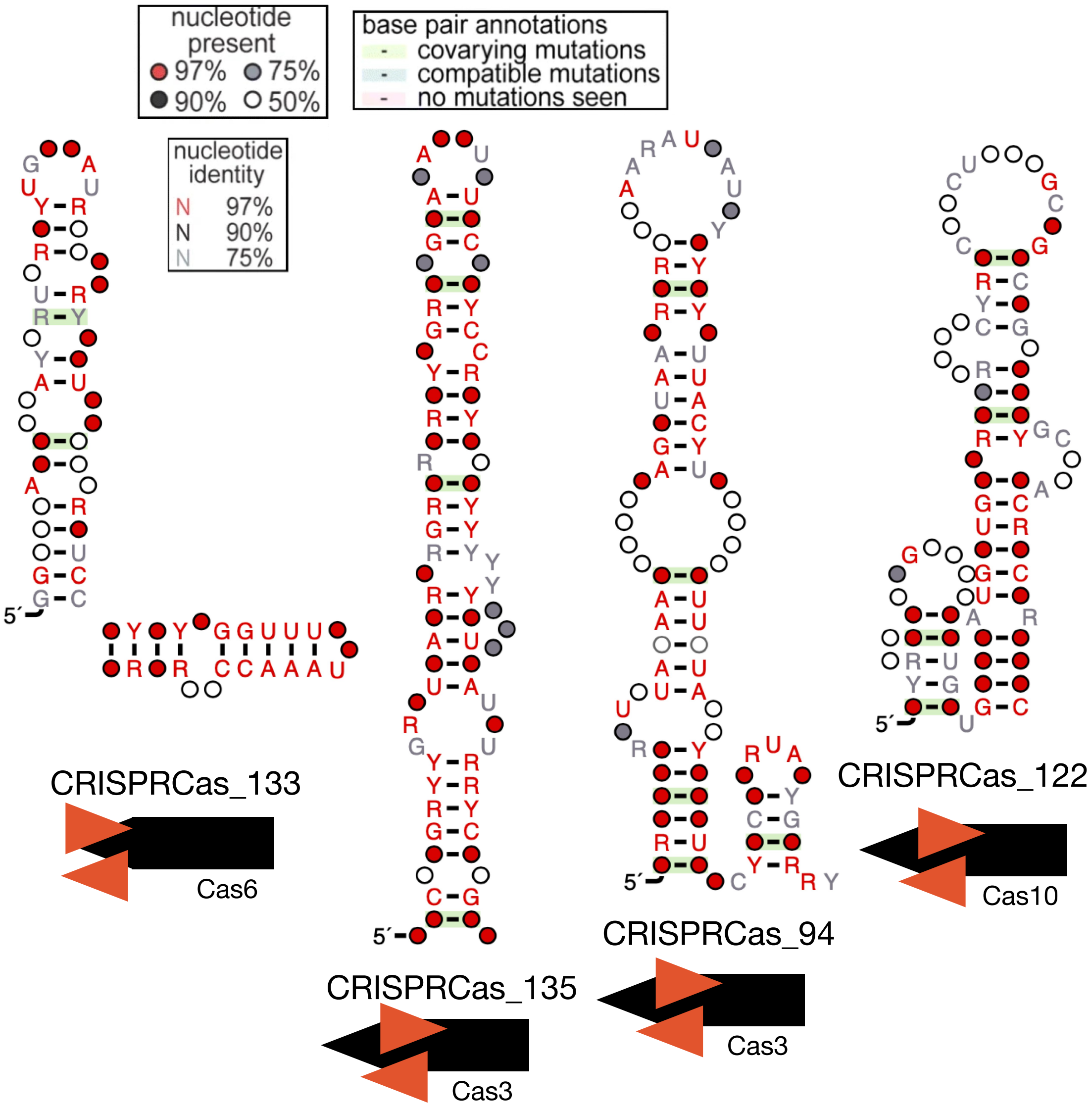

Figure 2. Palindromic candidates overlapping Cas genes.

cas8 closer to the 5' end of the gene (Table S1). One palindromic candidate overlapped *cas4* closer to the 5' end of the gene. Overall, there seemed to be an intriguing pattern of palindromic candidate structured RNAs overlapping a wide variety of *cas* genes.

Of the 156 candidate structured RNAs, 46 (29%) were predicted to be direct repeats in CRISPR arrays. These candidates all overlapped CRISPR arrays predicted by CRISPRCasTyper, specifically repeat regions, and occurred across multiple types of arrays. These repeats were typically specific to subtypes of CRISPR-Cas systems. For example, CRISPRCas_4 and CRISPR_52 were direct repeats associated with I-C and I-G systems, respectively (Fig. 3). CRISPR_148 was a direct repeat in II-C systems. CRISPR_26 was a direct repeat in III-A systems. CRISPRCas_80 was a direct repeat in IV-A1 systems, and CRISPRCas_117 was a direct repeat in VI-B1 systems (Fig. 3). Rfam currently contains 64 families of direct repeats. This work further expands this set to 110 distinct models.

Upon inspection of predictions found exclusively in intergenic regions, we highlight three interesting candidate structured RNAs. CRISPRCas_45 was located approximately 400 bases upstream of Uma2 family endonucleases found in type V-A systems (Fig. 4, Table S1). It was found entirely in the

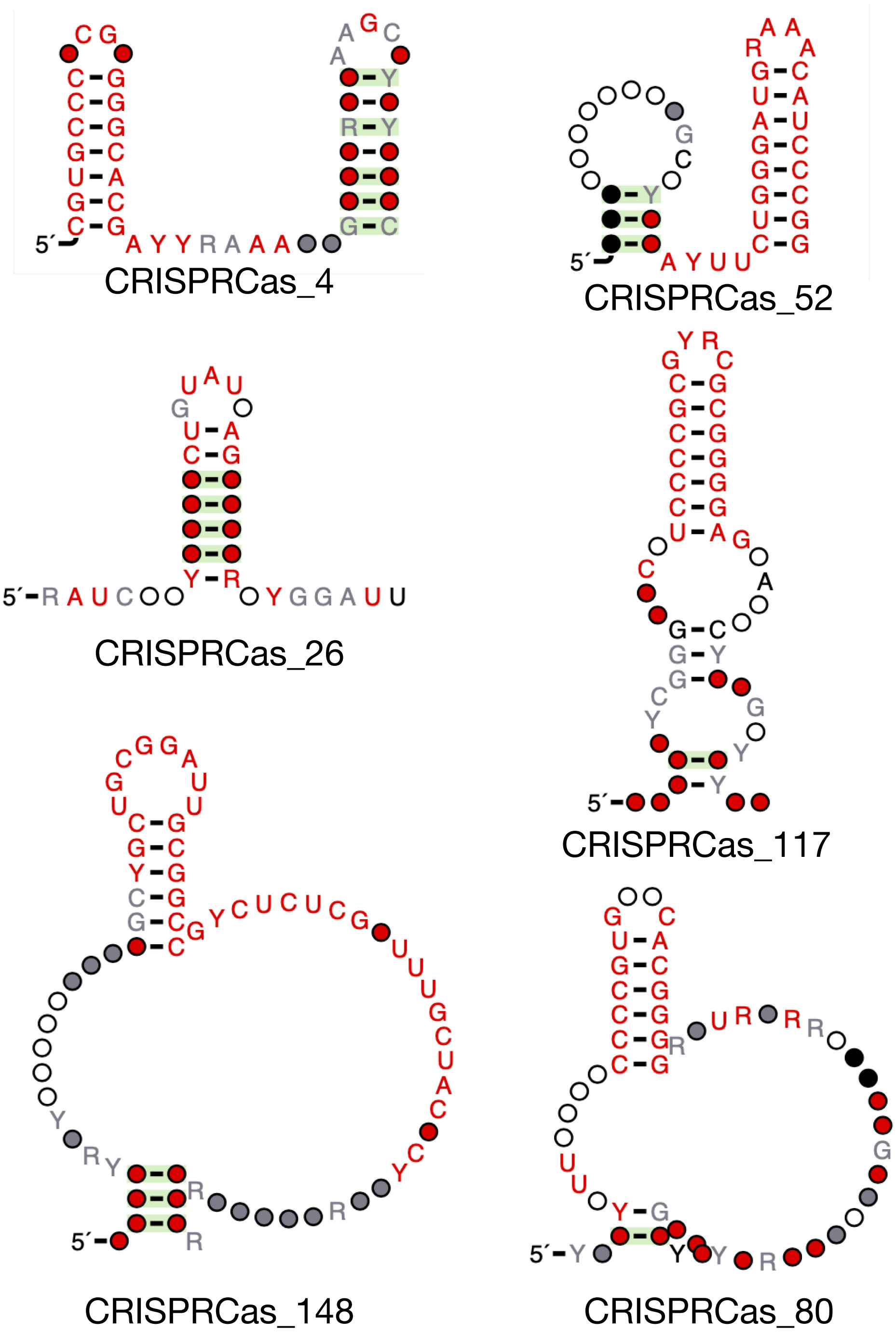

Figure 3. CRISPR direct repeat predictions.

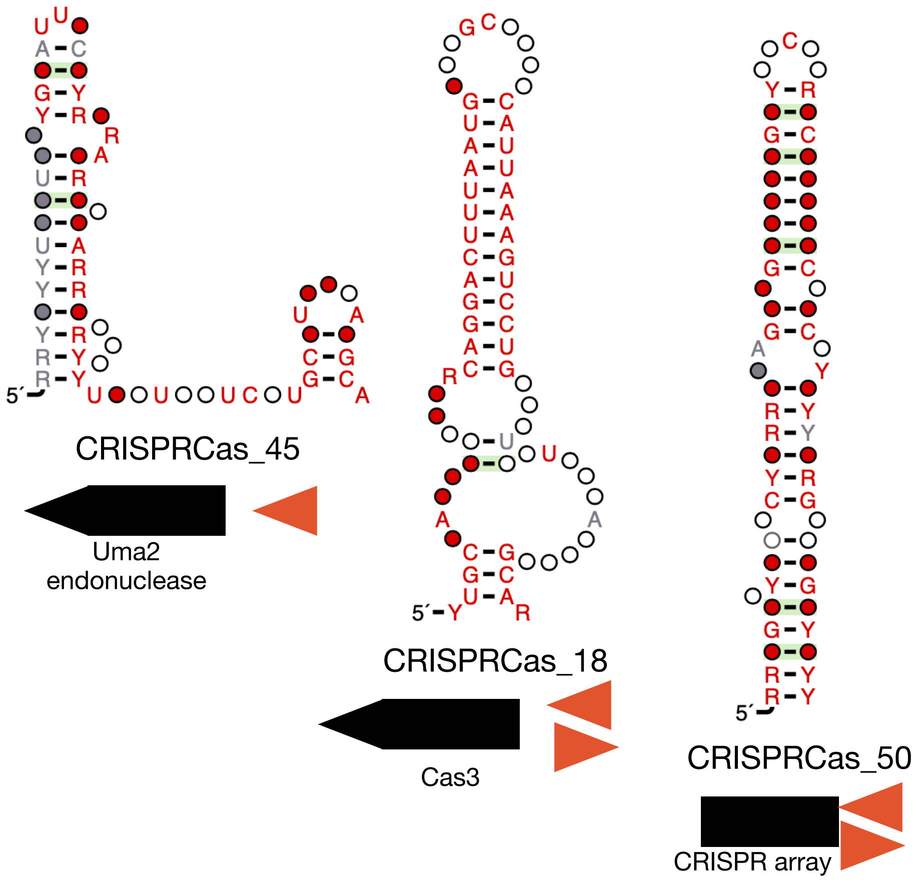

Figure 4. Candidate structured RNAs in intergenic regions.

human digestive system in *Eubacterium* species. CRISPRCas_18 was a palindromic candidate found both in *Proteobacteria* and *Firmicutes* and located approximately 300 bases upstream of *cas3* in type I-C systems. It was found in wastewater and bioreactor samples. CRISPRCas_50 was also a palindromic candidate and was found directly adjacent to CRISPR arrays (typically ~50 bases away) in type I-G systems. It was located in endoliths samples and found both in *Proteobacteria* and *Actinobacteria* (Fig. 4).

One candidate structured RNA, CRISPRCas_38, was likely a novel tracrRNA model (Fig. 5A). CRISPRCas_38 was predominantly found in *Bacteroidetes* (578/599 instances) and was broadly distributed across ecosystems, located 126 times in environmental, 340 times in host-associated, and 133 times in engineered ecosystems. It was typically found in the intergenic region within 100 bases upstream of *cas9* (Fig. 5B). However, in 50 of the 599 genomic positions in which it was identified, it partially or entirely overlapped the start of *cas9*. One likely explanation is that the start site of *cas9* has been occasionally misassigned by Prodigal. Nonetheless, this suggests that it is important to search for structured RNAs even across predicted coding regions. CRISPRCas_38 was found 599 times exclusively in type II-C systems. The antirepeat region of CRISPRCas_38 was homologous to 760 unique repeat regions identified by CRISPRCasTyper (BLASTn e-value < 0.05). These 760 regions were found in 4,661 CRISPR arrays. Interestingly, only 439 of these arrays were assigned a subtype by CRISPRCasTyper and 424 (97%) of those were assigned type II-C (Fig. 5C). For the 15 assigned a different subtype, the subtype probabilities assigned by CRISPRCasTyper ranged from 0.23 to 0.879. There were 310 CRISPR arrays with subtype probabilities greater than 0.9, and all were assigned to type II-C. Using the antirepeat region of this novel tracrRNA model, this suggests we can accurately classify thousands of these CRISPR arrays as type II-C based on their repeat sequence even if the repeat is not near type II-C Cas operons.

Discussion

As evident by the recent discovery of scoutRNAs [5], it is likely that other key noncoding RNAs that form secondary structures in CRISPR-Cas systems exist but have not been discovered. As metagenomics data becomes increasingly more available and improved tools are developed to predict CRISPR-Cas systems, it becomes possible to mine CRISPR-Cas systems at large-scale to predict novel candidate structured RNAs. In this work, we used publicly available, published datasets available through IMG/MER from diverse microbes and ecosystems and the recently developed tool, CRISPRCasTyper, to predict tens of thousands of Cas operons and CRISPRs and mine them to identify 156 candidate structured RNAs.

There are several limitations to our approach, many of which are similar to limitations of previous approaches [11–13]. First, very few of these candidates were found in metatranscriptomics-associated samples, and we were unable to

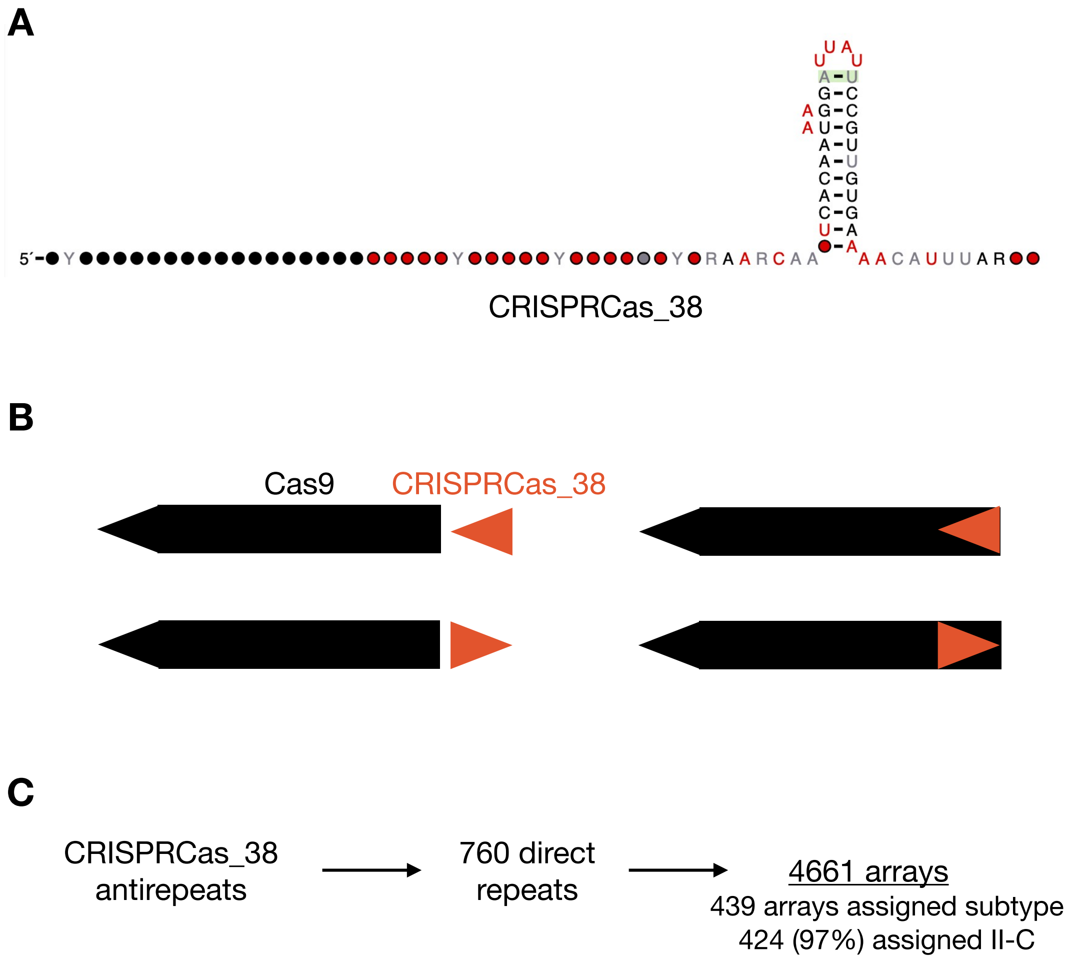

Figure 5. tracrRNA prediction.

quantify expression of these candidate structured RNAs. We also did not validate RNA structures experimentally, which would involve using methods like SHAPE-Seq or FragSeq [43–45]. Certain regions of the predicted candidate structures, especially those with less covariation evidence, may not be accurately depicted by this analysis. Second, it was difficult to accurately calculate a false-positive rate for our analyses. Thus, these predictions should be treated as candidates until further followup is performed. Third, we expected a high false-negative rate in these predictions given the scoring metrics and covariation requirements set. For example, we were unlikely to predict rare or highly conserved candidates, which would be difficult to assess for covariation. For example, 34 of the 64 direct repeats present in Rfam were identified in the CRISPR arrays we searched; however, we only rebuild models for 3 of these with our pipeline. There was not enough sequence divergence within our set to build models with significant covariation for the remaining 31 Rfam structures. In fact, most direct repeat models in Rfam do not display significant covariation and would not be retained by our pipeline. Fourth, we could not assign functions to these candidates.

Overall, we provided 156 candidate structured RNAs predicted from Cas operons and CRISPRs. Though follow up work is necessary to validate these candidates, we confidently identified 46 new direct repeats and 1 tracrRNA. We show that the discovery of this tracrRNA model can be useful to improve assignment of CRISPR arrays to type II systems. We also propose some especially interesting candidates, including palindromic candidate structured RNAs that overlap *cas1*, *cas3*, *cas4*, *cas6*, *cas7*, *cas8*, and *cas10*. Perhaps these candidates play roles in crRNA maturation or regulation of gene expression, for example, though more work is needed to assign such functions. Nonetheless, if any of these candidates play essential roles in these CRISPR-Cas systems, they will require consideration upon codon optimization of Cas and associated genes and meeting the system requirements from a genetic engineering perspective. We anticipate this resource will prompt experimental characterization, improve searchability of structured RNAs in CRISPR-Cas systems, and may have broader implications in adapting diverse CRISPR-Cas systems for genetic engineering purposes.

Methods

Data download and processing

All publicly available assembled metagenomic data with associated publications in IMG/MER were downloaded. We only considered contigs greater than 3 kb for analysis. This resulted in 25,658,797 contigs containing a total of 212,328,312,212 bases. We predicted Cas operons and CRISPRs along these contigs with CRISPRCasTyper version 1.6.1 [14] using default settings. Only regions predicted to be Cas operons or CRISPRs associated with Cas operons were considered for further analysis. We extended these Cas operon and CRISPR regions by 500 bp upstream and downstream using BEDTools slop [46] to capture regions between Cas operons and CRISPR arrays as well as upstream or downstream regions that may be within the operons. These regions were merged together with BEDTools merge, and the sequences corresponding to these regions were isolated using BEDTools getfasta. This resulted in 38,202 regions containing 285,229,248 bases, which we used to search for candidate structured RNAs.

Predicting candidate structured RNAs

We used BLASTn 2.5.0 +[39] with default settings to identify homologous regions within these Cas operons and CRISPRs. We retained matches with nucleotide differences, alignment lengths of at least 30, and bit scores of at least 20. Regions were clustered together using a single-linkage clustering algorithm, overcluster2, with default settings (Weinberg, Z., unpublished open-source software, available at http://weinberg-overcluster2.sourceforge.io), resulting in 11,546 clusters. We extracted sequences for these clusters using BEDTools getfasta. These clusters were structurally aligned using CMfinder version 0.4.1 [40], resulting in alignments for 7,173 cluster^s. We scored motifs using RNAPhylo, requiring a p-score of at least 10, filtering to 1,741 clusters. Motifs were also scored for significant covariation ($E < 0.05$) using R-scape [42] with default settings, further filtering to 717 clusters. Using motifs that passed above thresholds, we performed cmsearch [40] of candidate motifs against Cas operons and CRISPRs, retaining those models that uniquely and significantly (E value $< 1 \times 10^{-6}$) hit at least three unique regions across the regions. This ensured that the models were searchable and unique. Only 159 alignment files were retained from these analyses. We performed cmsearch [40] of Rfam 14.7 [7] against Cas operons and CRISPRs, considering those that meet the GA cut-off. Using BEDTools [46] intersect, we discarded the candidate structured RNAs that overlapped with any regions that were also predicted to be structures in Rfam, resulting in a total of 156 new candidate structured RNAs. RNA structure renderings were drawn using R2R [47]. The highlighted covariation in the renderings indicate bases with significant covariation predicted by R-scape [42]. Additionally, we used R-scape with – fold option to improve covariation among these alignments [48]. To determine if the candidate structured RNAs were found in coding or noncoding regions, we assessed which RNAs overlapped genes predicted by Prodigal [49]. We annotated genes using BLASTp to the nr database. Taxonomy of each contig was assigned using One Codex [50].

Identifying repeats with homology to tracrRNA antirepeats

Using BLASTp, we queried all predicted repeats identified with CRISPRCasTyper against all the tested Cas operons and CRISPRs. We removed significant hits (e value < 0.05) to CRISPR arrays (self matches) using BEDTools intersect. We then used BEDTools intersect to determine if any candidate structured RNAs were identified in regions homologous to the direct repeats. The only overlap identified was to the candidate structured RNA CRISPRCas_38, which was found in regions that were homologous to 760 distinct direct repeat sequences that could be traced back to 4,661 CRISPR arrays.

Acknowledgments

We would like to thank Heather Maughan for critical reading of the manuscript. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Office of Science of the U.S. Department of Energy [Contract No. DE-AC02-05CH11231].

Data and code availability

This article generated no new sequencing data and included all results of analyses performed. Models are provided as supplemental files. Code used can be found on github: https://github.com/bfremin-lbl/Candidate-Structures-CRISPR-Operons.

References

[1] Barrangou R, Fremaux C, Deveau H, *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315(5819):1709–1712.

[2] Brouns SJJ, Jore MM, Lundgren M, *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008;321(5891):960–964.

[3] Deltcheva E, Chylinski K, Sharma CM, *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011;471(7340):602–607.

[4] Jinek M, Chylinski K, Fonfara I, *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–821.

[5] Harrington LB, Ma E, Chen JS, *et al.* A scoutRNA is required for some type V CRISPR-Cas systems. Mol Cell. 2020;79(3):416–424.e5.

[6] Wang R, Zheng H, Preamplume G, et al. The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex. Protein Sci. 2012;21:405–417.

[7] Kalvari I, Nawrocki EP, Ontiveros-Palacios N, *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 2021;49(D1):D192–D200.

[8] Dar D, Sorek R. Bacterial noncoding RNAs excised from within protein-coding transcripts. MBio. 2018;9. DOI:10.1128/mBio.01730-18

[9] Adams PP, Storz G. Prevalence of small base-pairing RNAs derived from diverse genomic loci. Biochim. Biophys. Acta Gene Regul. Mech. 2020;1863:194524.

[10] Adams PP, Baniulyte G, Esnault C, *et al.* Regulatory roles of 5' UTR and ORF-internal RNAs detected by 3' end mapping. Elife. 2021;10. DOI:10.7554/eLife.62438.

[11] Weinberg Z, Lünse CE, Corbino KA, *et al.* Detection of 224 candidate structured RNAs by comparative analysis of specific

subsets of intergenic regions. Nucleic Acids Res. 2017;45(18):10811–10823.

[12] Fremin BJ, Bhatt AS. Comparative genomics identifies thousands of candidate structured RNAs in human microbiomes. Genome Biol. 2021;22(1):100.

[13] Weinberg Z, Wang JX, Bogue J, *et al*. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol. 2010;11(3):R31.

[14] Russel J, Pinilla-Redondo R, Mayo-Muñoz D, et al. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas Loci. CRISPR J. 2020;3(6):462–469.

[15] Chen I-MA, Chu K, Palaniappan K, *et al*. The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res. 2021;49:D751–D763.

[16] Schulz F, Yutin N, Ivanova NN, *et al*. Giant viruses with an expanded complement of translation system components. Science. 2017;356(6333):82–85.

[17] Ziels RM, Sousa DZ, Stensel HD, et al. DNA-SIP based genome-centric metagenomics identifies key long-chain fatty acid-degrading populations in anaerobic digesters with different feeding frequencies. ISME J. 2018;12:112–123.

[18] Thiel V, Wood JM, Olsen MT, *et al*. The dark side of the mushroom spring microbial mat: life in the shadow of chlorophototrophs. I. Microbial diversity based on 16S rRNA gene amplicons and metagenomic sequencing. Front Microbiol. 2016;7:919.

[19] Meyer JL, Jaekel U, Tully BJ, *et al*. A distinct and active bacterial community in cold oxygenated fluids circulating beneath the western flank of the Mid-Atlantic ridge. Sci Rep. 2016;6(1):22541.

[20] Stolze Y, Bremges A, Rumming M, *et al*. Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants. Biotechnol Biofuels. 2016;9(1):156.

[21] Reiss RA, Guerra P, Makhnin O. Metagenome phylogenetic profiling of microbial community evolution in a tetrachloroethene-contaminated aquifer responding to enhanced reductive dechlorination protocols. Stand Genomic Sci. 2016;11(1). DOI:10.1186/s40793-016-0209-z

[22] Sergeant MJ, Constantinidou C, Cogan TA, *et al*. Extensive microbial and functional diversity within the chicken cecal microbiome. PLoS One. 2014;9:e91941.

[23] Fortunato CS, Crump BC. Microbial Gene Abundance and Expression Patterns across a River to Ocean Salinity Gradient. PLoS One. 2015;10:e0140578.

[24] Wu Y-W, Tang Y-H, Tringe SG, et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome. 2014;2(1):26.

[25] An D, Caffrey SM, Soh J, *et al*. Metagenomics of hydrocarbon resource environments indicates aerobic taxa and genes to be unexpectedly common. Environ Sci Technol. 2013;47(18):10708–10717.

[26] Antunes LP, Martins LF, Pereira RV, *et al*. Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. Sci Rep. 2016;6(1):38915.

[27] Rossmassler K, Dietrich C, Thompson C, *et al*. Metagenomic analysis of the microbiota in the highly compartmented hindguts of six wood- or soil-feeding higher termites. Microbiome. 2015;3(1):56.

[28] Daly RA, Roux S, Borton MA, *et al*. Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. Nat Microbiol. 2019;4(2):352–361.

[29] Afshinnekoo E, Meydan C, Chowdhury S, *et al*. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Syst. 2015;1(1):97–97.e3.

[30] Tschitschko B, Erdmann S, DeMaere MZ, *et al*. Genomic variation and biogeography of Antarctic haloarchaea. Microbiome. 2018;6(1):113.

[31] Liang X, Whitham JM, Holwerda EK, *et al*. Development and characterization of stable anaerobic thermophilic methanogenic microbiomes fermenting switchgrass at decreasing residence times. Biotechnol Biofuels. 2018;11(1):243.

[32] He S, Kunin V, Haynes M, *et al*. Metatranscriptomic array analysis of 'Candidatus Accumulibacter phosphatis'-enriched enhanced biological phosphorus removal sludge. Environ Microbiol. 2010;12(5):1205–1217.

[33] Bäckhed F, Roswall J, Peng Y, *et al*. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. 2015;17(6):852.

[34] Coleine C, Albanese D, Onofri S, *et al*. Metagenomes in the borderline ecosystems of the Antarctic cryptoendolithic communities. Microbiol Resour Announc. 2020;9(10). DOI:10.1128/MRA.01599-19.

[35] Woodcroft BJ, Singleton CM, Boyd JA, *et al*. Genome-centric view of carbon processing in thawing permafrost. Nature. 2018;560(7716):49–54.

[36] Sorensen JW, Dunivin TK, Tobin TC, et al. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. Nat Microbiol. 2019;4(1):55–61.

[37] Hawley ER, Piao H, Scott NM, *et al*. Metagenomic analysis of microbial consortium from natural crude oil that seeps into the marine ecosystem offshore Southern California. Stand. Genomic Sci. 2014;9(3):1259–1274.

[38] Hernsdorf AW, Amano Y, Miyakawa K, *et al*. Potential for microbial H2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. ISME J. 2017;11(8):1915–1929.

[39] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990;215:403–410.

[40] Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–2935.

[41] Weinberg Z, Barrick JE, Yao Z, *et al*. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res. 2007;35(14):4809–4819.

[42] Rivas E, Clements J, Eddy SR. Estimating the power of sequence covariation for detecting conserved RNA structure. Bioinformatics. 2020;36:3072–3076.

[43] Uzilov AV, Underwood JG. High-Throughput nuclease probing of RNA structures using FragSeq. Methods Mol Biol. 2016;1490:105–134.

[44] Fremin BJ, Bhatt AS. Structured RNA contaminants in bacterial Ribo-Seq. mSphere. 2020;5. DOI:10.1128/mSphere.00855-20

[45] Takahashi MK, Watters KE, Gasper PM, *et al*. Using in-cell SHAPE-Seq and simulations to probe structure-function design principles of RNA transcriptional regulators. RNA. 2016;22(6):920–933.

[46] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–842.

[47] Weinberg Z, Breaker RR. R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinformatics. 2011;12(3). DOI:10.1186/1471-2105-12-3

[48] Rivas E. RNA structure prediction using positive and negative evolutionary information. PLOS Comput Biol. 2020;16:e1008387.

[49] Hyatt D, Chen G-L, LoCascio PF, *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119.

[50] Minot SS, Krumm N, Greenfield NB. One codex: a sensitive and accurate data platform for genomic microbial identification. bioRxiv. 2015;027607. DOI:10.1101/027607.