Computational and Structural Biotechnology Journal 21 (2023) 5630–5639

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Identification of over ten thousand candidate structured RNAs in viruses and phages

Brayon J. Fremin [a,b,*], Ami S. Bhatt [c,d,e], Nikos C. Kyrpides [a,b,f,*]

[a] Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[b] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[c] Blood and Marrow Transplantation) and Genetics, Stanford University, Stanford, CA, USA
[d] Department of Medicine (Hematology, Stanford University, Stanford, CA, USA
[e] Department of Genetics, Stanford University, Stanford, CA, USA
[f] Lead Contact, USA

ARTICLE INFO

Keywords:
Phage
Candidate structured RNAs
Comparative genomics

ABSTRACT

Structured RNAs play crucial roles in viruses, exerting influence over both viral and host gene expression. However, the extensive diversity of structured RNAs and their ability to act in cis or trans positions pose challenges for predicting and assigning their functions. While comparative genomics approaches have successfully predicted candidate structured RNAs in microbes on a large scale, similar efforts for viruses have been lacking. In this study, we screened over 5 million DNA and RNA viral sequences, resulting in the prediction of 10,006 novel candidate structured RNAs. These predictions are widely distributed across taxonomy and ecosystem. We found transcriptional evidence for 206 of these candidate structured RNAs in the human fecal microbiome. These candidate RNAs exhibited evidence of nucleotide covariation, indicative of selective pressure maintaining the predicted secondary structures. Our analysis revealed a diverse repertoire of candidate structured RNAs, encompassing a substantial number of putative tRNAs or tRNA-like structures, Rho-independent transcription terminators, and potentially cis-regulatory structures consistently positioned upstream of genes. In summary, our findings shed light on the extensive diversity of structured RNAs in viruses, offering a valuable resource for further investigations into their functional roles and implications in viral gene expression and pave the way for a deeper understanding of the intricate interplay between viruses and their hosts at the molecular level.

1. Introduction

Structured RNAs have diverse cis- and trans-regulatory roles across all domains of life. To predict structured RNAs in sequence data, several high-throughput comparative genomics analyses have been implemented on metagenomics data [1–4]. In summary, these strategies cluster intergenic regions from many organisms, predict motifs, and then score motifs based on nucleotide covariation. Covariation occurs when two base-paired sequences in a structured RNA vary together, suggesting there is selective pressure to preserve an underlying structure [5]. Thus, the observation of substantial covariation is useful to propose novel candidate structured RNAs (csRNAs). Viral sequences were only a small component of these metagenomic datasets used previously to predict csRNAs [1–3]; and therefore, high-throughput, large-scale analyses specifically of viral sequences have yet to be performed. Thus, many structured RNAs in viral genomes likely remain to be discovered.

Growing evidence suggests that viruses regulate expression of their genes and host genes using virus-encoded structured RNAs [6]. For example, *F-CphI* is a cis-encoded structured RNA in S-PM2 phage that infects *Synechococcus* species. *F-CphI* likely translationally regulates *psbA*, a cyanophage-encoded gene that aids the host in maintaining photosynthetic capacity [7]. *Misc_4* is a structured RNA in ΦR1–37 phage that infects *Yersinia enterocolitica*, and likely targets and translationally regulates *ptr*, *tufA*, and *ddrA* host genes [8]. *IpeX* is a cis-encoded structured RNA in PA-2 phage that infects *E. coli*. Regulating *IpeX* expression affects expression of host genes *ompC* and *ompA*. No complementary sequences exist between *IpeX*, *ompC*, and *ompA* [9,10], suggesting this regulation is indirect. These findings collectively suggest that both cis- and trans-encoded structured RNAs in phages have direct and indirect effects on phage and host gene expression. Phages also

* Corresponding authors at: Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
E-mail addresses: bfremin@lbl.gov (B.J. Fremin), nckyrpides@lbl.gov (N.C. Kyrpides).

https://doi.org/10.1016/j.csbj.2023.11.010
Received 3 August 2023; Received in revised form 3 November 2023; Accepted 3 November 2023
Available online 7 November 2023

encode their own tRNAs, perhaps to account for codons that are more prevalent in phage than host genomes or because amino acid usage differs between phage and host proteins [11,12]. Additionally, viruses can encode tRNA-like structures to control translation or perform other roles [13]. In RNA viruses, non-canonical translation has been observed [14,15]. Further, some tRNAs naturally form non-canonical structures [16], suggesting that previously unidentified tRNAs and tRNA-like structures likely exist in viruses.

In recent years, substantial work has revealed novel DNA and RNA viruses. IMG/VR [17] now contains millions of viral sequences collected from a variety of datasets [18–29]. Moreover, several metatranscriptomics datasets have substantially increased the diversity of predicted RNA viruses [30–32]. Building csRNA models across large sequence spaces is computationally expensive, especially at the initial stage of identifying homologous sequences. To overcome this computational limitation we used all-versus-all HS-BLASTN [33] as previously performed [2], to identify homologous regions within these viral datasets. Using the wealth of viral sequences now available and this approach to overcome computational limitations, we mined these viral sequences to predict over 10 thousand novel csRNAs, substantially expanding upon the diversity of structured RNAs in both DNA and RNA viruses.

2. Results

2.1. Workflows to predict csRNAs

First we downloaded IMG/VR 3.0 [17] and predicted genes along these contigs using MetaProdigal [34]. IMG/VR [17] included 2,377,994 contigs containing 48,566,528,056 base pairs (bp). From these contigs, we predicted 51,642,570 genes containing 35,182,970,109 bases. Based on the positions of these genes, we then predicted 16,551,965 intergenic regions (each region at least 30 bp) containing 2,640,080,124 bp, ignoring the ends of contigs. We performed all-versus-all

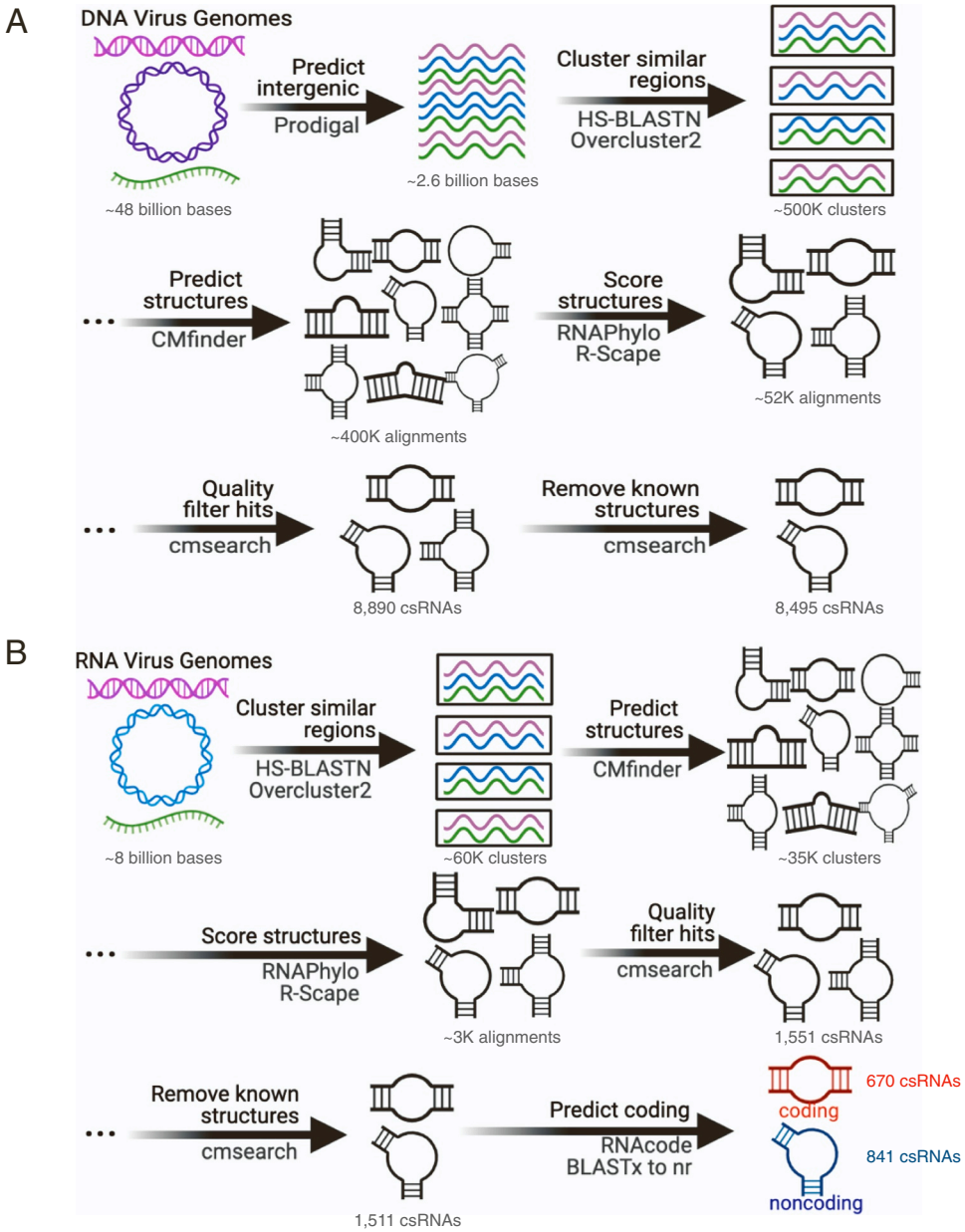

Fig. 1. Prediction of csRNAs in viruses. (A) Workflow of csRNAs from IMG/VR. HS-BLASTN is used to identify homologous intergenic regions, which are clustered and scored using RNAphylo and R-scape. Only csRNAs with unique, quality hits are retained. Candidates that are already present in Rfam are excluded.

HS-BLASTN [33] on the intergenic regions. Homologous regions that were at least 30 bp long, were less than 100% identical, and contained bit scores between 20 and 200 were retained (Fig. 1A). While accepting this range of bit scores allows for a wide range of diversity in these sequences, a tradeoff is that some sequences may be too highly divergent and align poorly to the other sequences. Using overcluster2 [1], we clustered homologous intergenic regions, resulting in 565,438 clusters. For each cluster, we predicted possible structures using CMfinder version 0.4.1, which identified 386,777 clusters containing at least one alignment of a possible structure. Using RNAphylo, a tool that uses a phylogenetic model to score alignments, we identified 110,109 clusters with an RNAphylo p score of 10 or greater. Using R-scape [5], we found that 52,780 of these contained at least one significant covarying base. We used cmsearch [35] to determine which of these alignments significantly (E value $< 1 \times 10^{-6}$) matched at least three unique regions in IMG/VR. We also removed duplicate csRNAs that matched any of the same regions as another csRNA by choosing the longest structure, resulting in 8890 csRNAs. Among these, 324 were already present in Rfam [36] and 71 were recently identified in human microbiomes [2], resulting in a finalized set of 8495 novel csRNAs (Fig. 1A). The 8495 csRNAs predicted from IMG/VR were, on average, 67 bases long (median of 65 bases), ranging from 29 to 151 bases in length (Supplementary Table 1, Supplementary File 1, Supplementary File 2). We classified 6796 csRNAs to Duplodnaviria, 351 to Varidnaviria, 71 to Monodnaviria, and 1 to Anelloviridae. Among these, 122 csRNAs were found in more than one realm. There were 1398 csRNAs identified in IMG/VR that could not be classified to a realm (Supplementary Table 1).

IMG/VR is predominantly composed of DNA viruses; however, recent work has revealed thousands of RNA viruses from various ecosystems that are not found in IMG/VR. To avoid overlooking RNA viruses, we downloaded data from several sources, including ssRNA phages [32], RNA aquatic viruses [31], all RNA virus nucleotides from NCBI Virus [30], and recent RNA virus discovery that expanded the global RNA virome by five-fold [37]. This combined set included 8,706, 281,652 bp of RNA. Because this is a computationally manageable amount of data to predict csRNAs, we chose to perform all-versus-all HS-BLASTN [33] on all sequences, not just intergenic regions. Thus, we applied an identical pipeline as performed above for the DNA viruses, except we did not run MetaProdigal and remove predictive coding regions first. We clustered homologous regions, resulting in 60,439 clusters (Fig. 1B). We found that 35,340 of these clusters contained at least one alignment of a possible structure. Among these, 4093 clusters had an

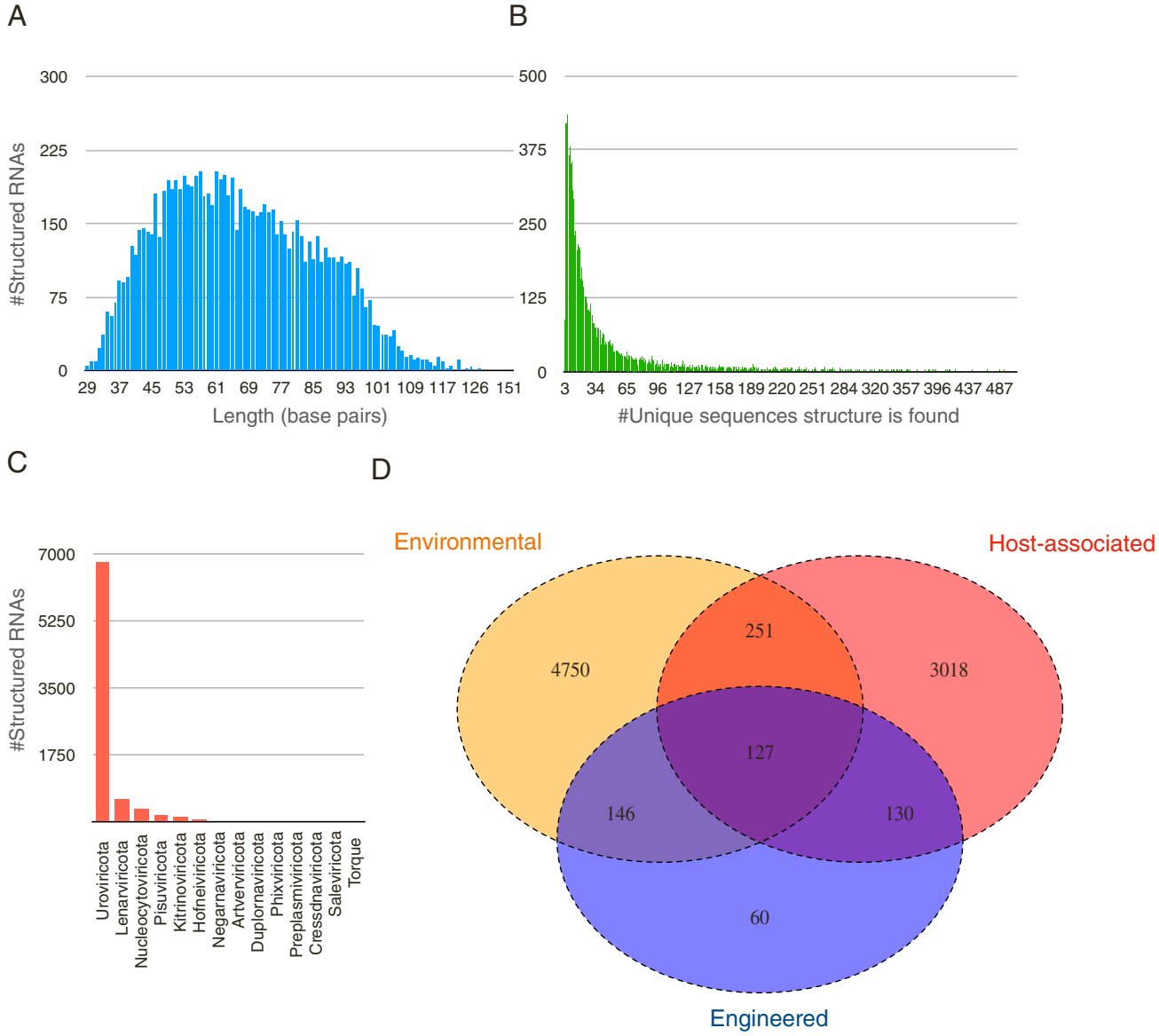

Fig. 2. csRNA statistics. (A) Histogram of the distribution of lengths of csRNAs.

RNAphylo p score of 10 or greater, and 2827 of these contained at least one significant covarying base pair using R-scape. We also used cmsearch to determine which of these alignments significantly (*E* value $< 1 \times 10^{-6}$) matched at least 3 unique regions in viral genomes. We also removed duplicate csRNAs, resulting in a finalized set of 1551 csRNAs. Among these, 40 were already present in Rfam, resulting in a finalized set of 1511 novel csRNAs from RNA viruses. We next determined whether the csRNAs were found in coding or noncoding regions. Using BLASTx [38], we queried the regions in which these csRNAs were found against the nr database. Any region with an e-value of less than 0.05 was considered to be likely coding (Supplementary Table 2). Additionally, RNAcode [39] was applied using default settings to these alignments, and regions with *p* values less than 0.05 were also classified as coding csRNAs. Overall, we predicted that 670 of these csRNAs were likely coding, while 841 were likely noncoding (Fig. 1B). The 1511 csRNAs predicted from RNA viruses were, on average, 65 bases long (median of 63 bases), ranging from 30 to 130 bases in length (Supplementary Table 1, Supplementary File 3, Supplementary File 4, Supplementary File 5).

Overall, the csRNAs predicted in this study show a wide range of diversity in length, prevalence, taxonomy, and ecosystem. These 10,006 csRNAs ranged from 30 to 151 bases in length, with an average length of 67 bases (Fig. 2 A). None of the DNA virus csRNAs were found in the RNA virus dataset; similarly, none of the RNA virus csRNAs were identified in the DNA virus dataset. The number of unique sequences representing each csRNA ranges from 3 to 41,253 sequences (Fig. 2B). These csRNAs were found across a large diversity of 14 viral phyla (Fig. 2 C). Specifically, 6796 csRNAs were identified in Uroviricota. Among the csRNAs identified in IMG/VR, 4750 were found exclusively in environmental ecosystems, 3018 were exclusively host-associated, and 378 were found in both ecosystems (Fig. 2D).

We calculated a false discovery rate (FDR) for each of these sets of csRNAs. We first shuffled Clustal W 2.056 alignments using SISSlz [40], then performed the same scoring pipeline using shuffled sequences, including CMfinder, RNAphylo (p-score > 10), and R-scape ($E < 0.05$). We found that only 18 of the 8495 shuffled alignments predicted from DNA viruses passed these filters, suggesting an FDR of 0.0021. Additionally, we found that only 13 of the 1511 shuffled alignments predicted from RNA viruses passed these filters, suggesting an FDR of 0.009. When we subset these candidate structures as likely coding or noncoding, we found that 7 of the 670 (0.01) shuffled alignments in putative coding regions passed filters, while 6 of the 841 (0.007) shuffled alignments in putative noncoding regions passed filters. Overall, this suggests the FDR is very low for all csRNA predictions. However, because we only calculated the FDR based on shuffling of the same sequences the csRNAs were predicted from instead of shuffling entire metagenomes (which would be computationally challenging), this FDR calculation is likely substantially underestimated. Reassuringly, we were also able to recreate models for 364 structures already found in Rfam (Supplementary Table 1). For example, we recreated a model for Sarbecovirus-3UTR, a structured RNA found in several species including SARS-CoV-2, and HIV_GSL3, which directs HIV-1 packaging of the genomic RNA. We named all csRNAs predicted from IMG/VR as DNAVirus_X, where X is a unique number. Similarly, we named all csRNAs predicted from RNA virus datasets as RNAVirus_Y.

2.2. csRNAs highly expressed in the gut microbiome

To identify highly expressed csRNAs, we analyzed metagenomic and small RNA-Seq data without fragment size selection performed on four human fecal samples from a previous study [41,42]. The metagenomic data was previously subjected to computational assembly, and RNA-Seq reads were previously aligned to these de novo assemblies. We first searched for the 10,006 csRNAs along these assemblies using cmsearch, identifying 544 csRNAs present in the assemblies. Using the aligned small RNA-Seq reads, we calculated the reads per kilobase million (RPKM) and found that 206 (37.9%) of these csRNAs were transcribed at a level greater than 20 RPKM (Supplementary Table 1). We highlight four of the most highly transcribed csRNAs in this dataset (Fig. 3). DNAVirus_6258 was 86 bases long and found in Caudovirales that infect Lachnospiraceae. DNAVirus_6565 was 76 bases long and found in Caudovirales that infect Alistipes. DNAVirus_6426 was 81 bases long and found in Caudovirales and Imitervirales that infect Blautia. DNAVirus_6425 was 64 bases long and found in Caudovirales that infect Blautia and Ruminococcus. All these csRNAs were exclusively found in the human gut microbiome and highly expressed in this small RNA-seq dataset.

2.3. csRNAs likely to be tRNAs or tRNA-like structures

To determine if any of these csRNAs may be tRNAs or rRNAs, we first predicted tRNAs and rRNAs across the regions in which these csRNAs were found. We used Aragorn [43] to predict tRNAs and Barrnap to predict rRNAs [44], both performed with default settings. No rRNA predictions overlapped with csRNAs. However, we identified five csRNAs that shared substantial overlap in regions that were also predicted to be tRNAs by Aragorn. DNAVirus_1423, DNAVirus_1850, DNAVirus_215, DNAVirus_2961, and DNAVirus_8881 overlapped predictions of tRNA-Tyr, tRNA-Asp, tRNA-Cys, tRNA-Arg, and tRNA-Cys, respectively (Fig. 4). DNAVirus_1850 was found 27 times in IMG/VR but only overlapped with a tRNA-Asp prediction at one of those regions (Supplementary Table 1). DNAVirus_8881 was the only csRNA among the five to overlap with tRNA-Cys in all 11 regions it was identified (Supplementary Table 1). This analysis does not prove that these csRNAs are tRNAs, but suggests these structures may be tRNA-like or tRNAs, perhaps also with non-canonical structures [13].

2.4. Functional analyses of csRNAs

It is challenging to functionally characterize csRNAs; the diversity of possible functions among structured RNAs is immense and their regulatory functions can be cis- or trans-acting. csRNAs that typically occur immediately upstream of genes, potentially in the 5′ UTR, may be cis-regulatory. We inspected csRNAs that are directly upstream (within 25 bases) of the start codon of genes. We assigned protein domains [38,45] to the encoded proteins of these downstream genes to link specific csRNAs to possible functions. We found that 879 csRNAs were found within 25 bp upstream of genes encoding proteins with known domains (Supplementary Table 3). Further validation is necessary to determine if these csRNAs play cis-regulatory roles of these downstream genes.Using RNIE [46], we predicted that 1241 of the 8495 csRNAs found in IMG/VR potentially contained a Rho-independent transcription terminator (Supplementary Table 1). Additionally, 5 of the 1511 csRNAs predicted from RNA virus datasets potentially contained Rho-independent transcription terminators (Supplementary Table 1), which may be false positives. Of note, IMG/VR is enriched in phages; therefore, it makes sense that csRNAs in IMG/VR would be more enriched in Rho-independent termination, which only occurs in prokaryotes. We additionally found that 598 pairs of csRNAs predicted from IMG/VR and 270 pairs of candidates predicted from RNA viruses overlapped each other in opposing strand orientations; for example, DNAVirus_4 and DNAVirus_5 overlapped but were on opposing strands. Previously characterized systems of overlapping csRNAs, such as RyeA and SdsR systems, have been previously identified [47]. Additionally, 1592 csRNAs were predicted to overlap with themselves at least partially in opposing orientations. For example, DNAVirus_7368 is a palindromic csRNA with transcriptional evidence to support that both strands are expressed (Supplementary Fig. 1). DNAVirus_7368 in the forward orientation is 5 bases longer (112 bases in total) than DNAVirus_7368 in the reverse orientation, which is missing parts of the stem in some regions.

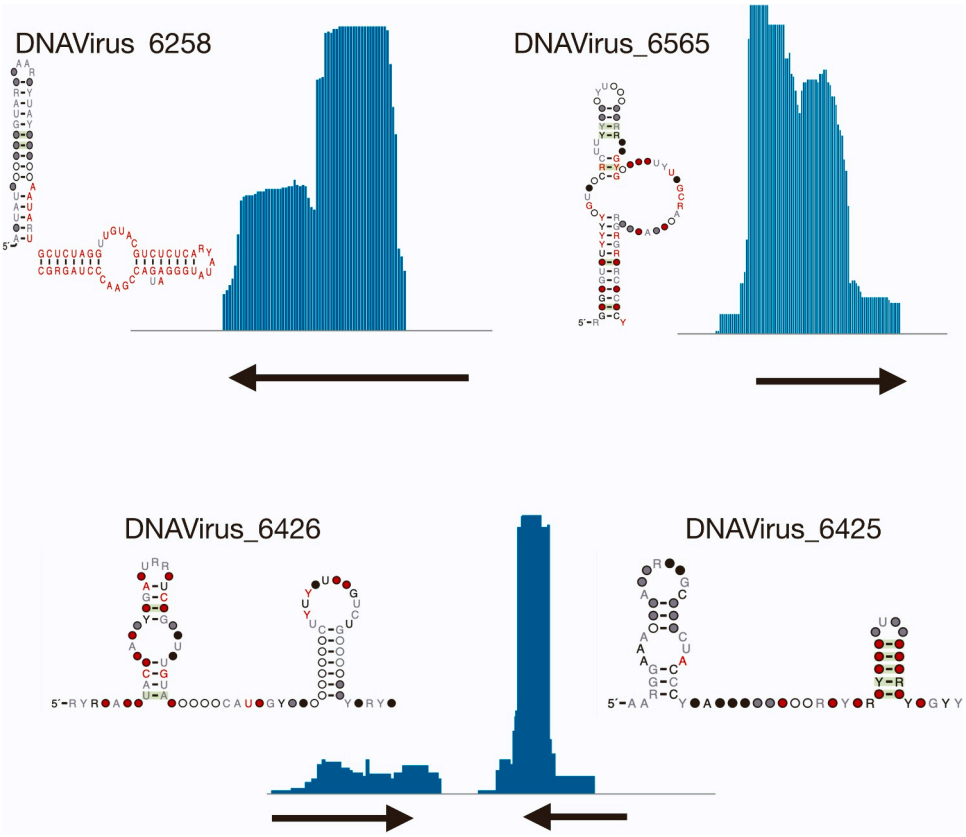

Fig. 3. Highly expressed csRNAs. Structure diagrams and IGV visualizations of csRNAs with high RNA-Seq expression. This includes DNAVirus_6258, DNAVirus_6565, DNAVirus_6426, and DNAVirus_6425.

2.5. Cis-regulatory csRNAs that likely regulate GP20

To highlight potential cis-regulatory csRNAs of interest, we observed six csRNAs consistently found directly upstream of GP20 genes (Fig. 5, Supplementary Table 3). Though GP20 is a gene of unknown function, two motifs, GP20-a (107 bp) and GP20-b (72 bp) have previously been predicted to occur upstream of GP20 [1]. Both GP20-a (RF02991) and GP20-b (RF03003) were predicted by our pipeline (and thus removed from our 8495 set). These six novel csRNAs, which were all found in *Caudovirales*, were found in a wide variety of ecosystems and hosts. For example, DNAVirus_3400 was primarily found in human-associated microbiomes, mostly in the digestive system but occasionally in the reproductive system and oral microbiome and was found mostly in the host bacterium *Veillonella* (Fig. 5, Supplementary Table 1). DNAVirus_56 was found in human digestive system-associated microbiomes, mostly associated with host bacterium *Lachnoclostridium*. DNAVirus_5891 was found in soil, mostly associated with the host bacterium *Lysinibacillus*. DNAVirus_6837 was found in human digestive system-associated microbiomes. DNAVirus_7012 was found in the human digestive system-associated microbiomes, mostly associated with the host bacterium *Gemmiger*. DNAVirus_8070 was found in wastewater (Fig. 5, Supplementary Table 1). All six csRNAs were found within 25 bases directly upstream of GP20 (Supplementary Table 3); thus, we identified both the two known GP20 motifs as well as six additional csRNAs that likely regulate GP20 using our pipeline.

2.6. Cis-regulatory csRNAs that likely regulate GP32

We identified 12 csRNAs consistently found directly upstream of GP32 genes (Fig. 6, Supplementary Table 3). GP32 encodes a single-stranded DNA binding protein that plays essential roles in viral replication, recombination, and T4 DNA repair. It has been shown that GP32 regulates its own translation through binding to a pseudoknot RNA [48] structure located directly 5′ upstream of the GP32 gene [49,50]. However, no models of this RNA structure are present in Rfam. These 12 csRNAs may regulate GP32 translation and represent a diverse set of candidate structures found in a variety of ecosystems and hosts (Fig. 6, Supplementary Table 1). To highlight some examples, DNAVirus_6622 was found in marine and freshwater and found in *Caudovirales*; DNAVirus_1944 was also found in *Caudovirales* in the host *Planctomycetes*. DNAVirus_7999 was found in aquatic sediment, marine, and freshwater and was found in both *Caudovirales* and *Imitervirales*. DNAVirus_8758 was found in marine and freshwater and associated with the host *Gammaproteobacteria* (Fig. 6, Supplementary Table 1). These findings likely represent a diverse set of structures that regulate GP32.

There are several other interesting structured RNAs that may be cis-regulatory (Supplementary Table 3). For example, we found 14 csRNAs consistently upstream of psiM2_ORF9, a putative large terminase. There were 11 candidates typically found upstream of resolvase genes. Six csRNAs were consistently found upstream of Podovirus_gp16, which controls the genome-encapsidation reaction. Six candidates were consistently found upstream of RusA, which resolves Holliday intermediates. Five candidates were typically found upstream of PHA00201, a major capsid protein. Four candidates were typically found upstream of DNA N-6-adenine-methyltransferase. These are among many examples of csRNAs that may play cis-regulatory roles.

3. Discussion

To date, no comparative genomics approaches to predict csRNAs have specifically focused on viruses at large scale. Our pipelines analyzed billions of bases from viral genomes to predict 10,006 novel csRNAs from diverse viruses across 14 phyla that infect various hosts and inhabit multiple ecosystems. We further validated transcription for

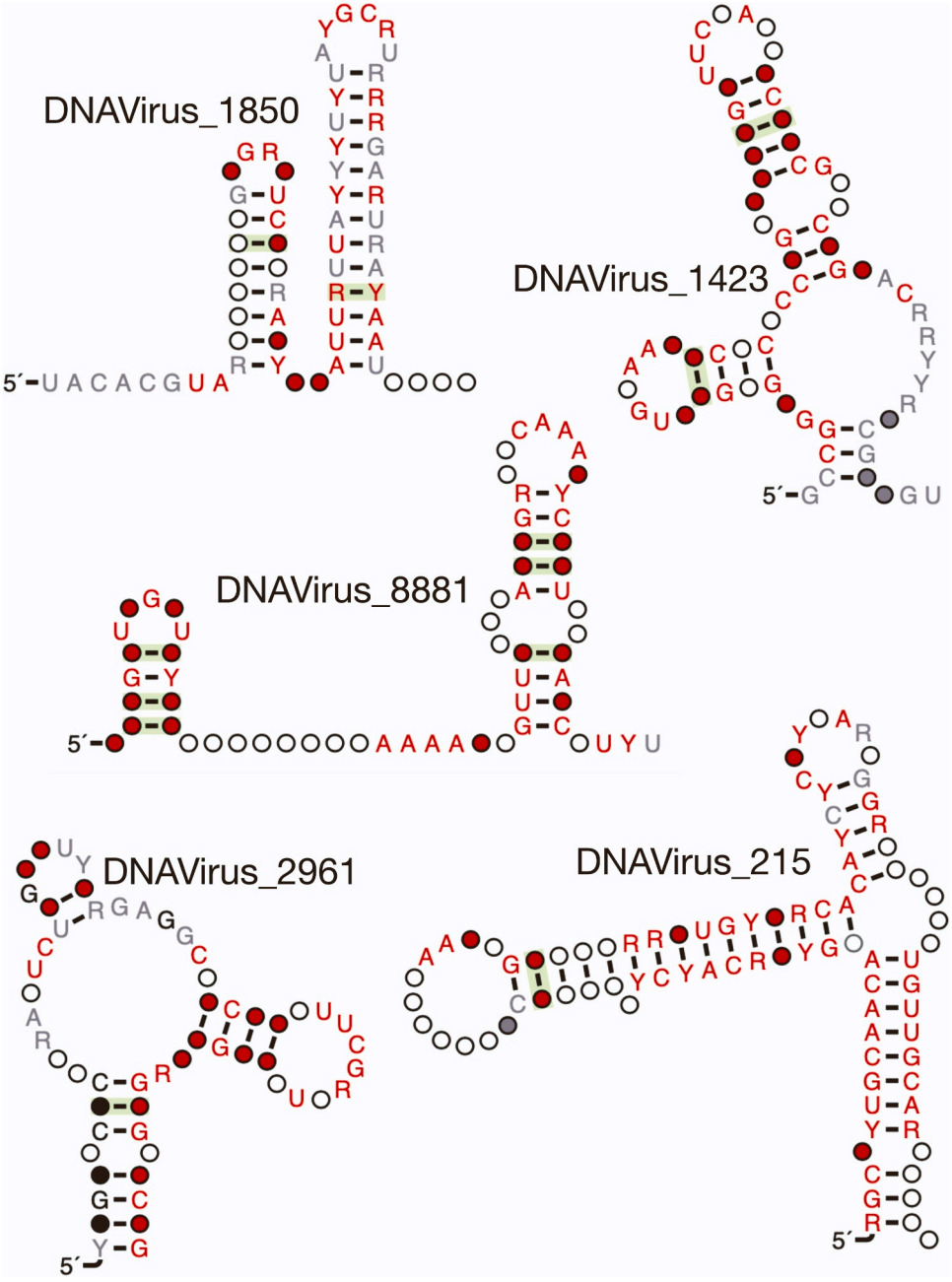

Fig. 4. csRNAs likely to be tRNA or tRNA-like. Consensus diagrams of five csRNAs that share substantial overlap with tRNAs predicted by Aragorn. DNAVirus_8881 consistently shared overlap with tRNA in all regions predicted, while DNAVirus_1850, DNAVirus_1423, DNAVirus_2961, and DNAVirus_215 only sometimes overlapped tRNA predictions.

206 of these csRNAs in the human fecal microbiome. We highlighted some interesting candidate structures from these 10,006 candidates, including tRNA or tRNA-like structures, potentially cis-regulatory csRNAs directly upstream of genes, such as GP20 and GP32, playing essential roles in virus physiology.

There are several limitations to our comparative genomics approach, consistent with similar approaches performed previously [1–3]. First, this method does not validate RNA structures experimentally using methods like SHAPE-Seq or FragSeq for the majority of predictions [51, 52]. Second, this approach has a high false-negative rate considering that we employ stringent scoring metrics based on phylogeny and covariation. If a structured RNA is strongly conserved or relatively rare, it will likely not be identified by these analyses. Third, the false-positive rate of our analysis is likely higher than our calculations. Random shuffling of alignments followed by rescoring the structures is unlikely to model all biological features that might increase false positive rates and does not represent the full metagenomic search space from which the structures were predicted. Fourth, we cannot truly assign functions to csRNAs without experimental validation and can only propose likely roles given genomic contexts.

Overall, we report 8495 csRNAs predicted from IMG/VR and 1511 csRNAs predicted from RNA virus datasets. Many different classes of csRNAs were identified, including likely tRNA or tRNA-like structures and potentially cis-regulatory structures. Predicted structure alone is not enough to confidently assign function to most csRNAs, thus future work is necessary. Computational analyses support that 879 candidates were found directly upstream of genes with assigned protein domains and 1246 candidates were predicted to contain Rho-independent

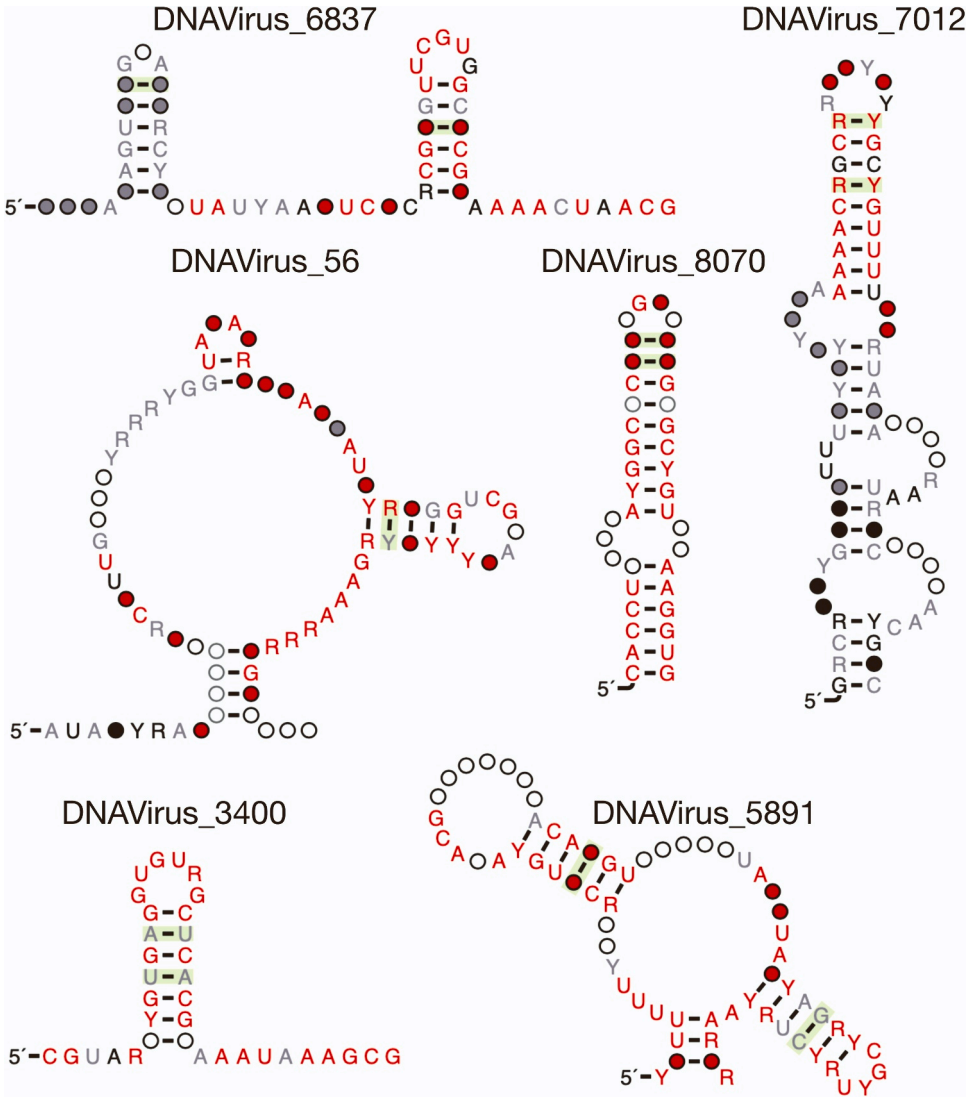

Fig. 5. csRNAs that likely regulate GP20. Consensus diagrams of six candidates that are found within 25 bases directly upstream of GP20 genes. These six models, DNAVirus_6837, DNAVirus_7012, DNAVirus_56, DNAVirus_80, DNAVirus_3400, and DNAVirus_5891, are structurally distinct from two existing models in Rfam also upstream of GP20 genes.

transcription terminators. CsRNAs can be prioritized based on expression levels in samples of interest or proximity to known genes of interest. Some examples of csRNAs likely to be cis-regulatory include six candidates directly upstream of GP20 genes and 12 candidates directly upstream of GP30 genes. These 10,006 csRNAs represent a first step at comprehensively identifying the diversity of structured RNAs in viruses at large scale. We anticipate this resource will lead to functional characterizations and reveal new biological mechanisms in viruses.

4. Methods

4.1. Data download

All DNA virus contigs used in this study were downloaded from IMG/VR [17]. The RNA virus data used in this manuscript can be accessed from four sources: ssRNA phages (from Supplemental files) [32], RNA aquatic viruses (from Supplemental files) [31], recent work mining diverse metatranscriptomes at large-scale [37], and all RNA virus nucleotides from NCBI Virus (as of March 2021: 3,367,662 nucleotide sequences) [30]. The samples used for the RNA-Seq analysis can be found under BioProject accession no. PRJNA510123.

4.2. Comparative genomics workflow

We predicted genes along IMG/VR [17] contigs using MetaProdigal [34] with default settings. We extracted sequences that were predicted to be genes. Additionally, we used BEDTools complement [53] to extract sequences that were predicted to be intergenic and were greater than 30 bp. In this case, we ignored the edges of contigs and required that these regions be between two genes to be intergenic. We performed all-versus-all HS-BLASTN [33] across intergenic regions using default settings. We removed homologous regions that were shorter than 30 base pairs, were 100% identical to each other, were assigned bit scores less than 20, or were assigned bit-scores greater than 200. Subsequences were clustered from the HS-BLASTN results using a single-linkage clustering algorithm called overcluster2 (Weinberg, Z., unpublished open-source software, available at http://weinberg-overcluster2.sourceforge.io) using default settings. For each cluster, we extracted the underlying RNA sequences using BEDTools [53] and structurally aligned the clusters using CMfinder version 0.4.1 [35]. We scored motifs using RNAPhylo with a p-score cutoff of at least 10. Additionally, motifs were scored for significant covariation ($E < 0.05$) using R-scape [5] with default settings. Using the models that met the above thresholds, we performed cmsearch [35] of candidate motifs against all IMG/VR [17]

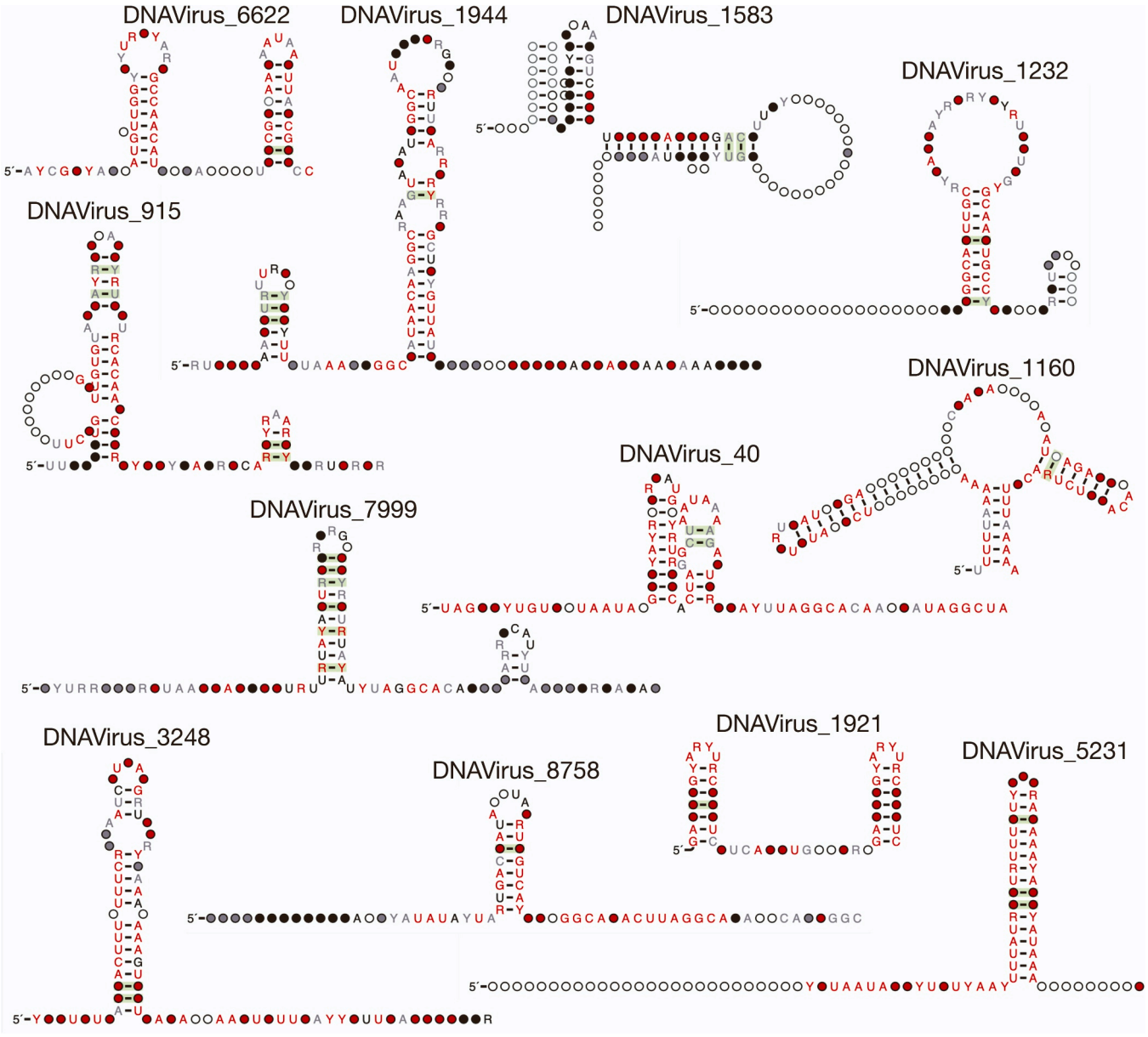

Fig. 6. csRNAs that likely regulated GP32. Consensus diagrams of 12 candidates that are found within 25 bases directly upstream of GP32 genes, including DNAVirus_6622, DNAVirus_1944, DNAVirus_1583, DNAVirus_1232, DNAVirus_915, DNAVirus_7999, DNAVirus_40, DNAVirus_1160, DNAVirus_3248, DNAVirus_8758, DNAVirus_1921, and DNAVirus_5231.

contigs and retained those models that uniquely and significantly (*E* value $< 1 \times 10^{-6}$) matched at least three unique regions across the contigs. In other words, we ensured that the models strongly matched at least three of the unique regions they were created from and were thus unique and searchable in the way we intended. In cases in which CMfinder [35] proposed multiple alignments for the same region that met p-score and covariation cutoffs, the alignment with the highest p-score was chosen. We performed all-versus-all HS-BLASTN [33] across all RNA virus contigs using default settings, filtering out homologous regions that were shorter than 30 base pairs, were 100% identical to each other, were assigned bit scores less than 20, or were assigned bit-scores greater than 200. All downstream steps for these RNA viruses were performed identical to the workflow for DNA viruses.

Using BEDTools [53] intersect, if any region contained an overlap with an Rfam structure and a csRNA, we discarded the candidates. Transfer RNAs were predicted by Aragorn [43], and rRNAs were predicted by Barrnap. RNA structure renderings were drawn using R2R [54], which was previously an output of R-scape [5]. The green highlighted covariation in the renderings only include covariation predicted to be significant by R-scape [5].

To determine if the csRNAs in RNA viruses were found in coding or noncoding regions, we assessed which structures overlapped predicted genes. Using BLASTx [38], we queried the regions in which these csRNAs were found against the nr database. Any region with an e-value of less than 0.05 was considered to be likely coding. Additionally, these regions were aligned with Clustal W 2.0 [55]. RNAcode [39] was applied using default settings to these alignments and regions with *p* values less than 0.05 were also classified as coding csRNAs. The steps to follow this workflow can be found here: https://github.com/bfremin-lbl/csRNAs-in-Viruses-and-Phage/tree/main.

4.3. False discovery rate (FDR) estimates

Clustal W 2.0 alignments were shuffled using SISSlz[40]we performed

the same pipeline, including CMfinder [35], RNAphylo [1], and R-scape [5] with the same thresholds as above. To calculate the FDR we divided the number of shuffled alignments that met the thresholds of a csRNA by the number of alignments considered.

4.4. Taxonomic classification of RNA viruses

We taxonomically classified the contigs on which these csRNAs were found using kaiju [56] with default settings. We did not classify contigs from IMG/VR because their taxonomic classifications were already provided within the database, as was information on ecosystems and predicted hosts.

4.5. RNA-Seq analysis

CsRNAs were identified along previously created metagenomic assemblies using cmsearch. RNA-seq reads were trimmed using trim galore version 0.4.0 and cutadapt 1.8[58,59]1 with parameters -q 30 and -illumia. These reads were mapped to their associated metagenomic assemblies using bowtie version 1.1.1 [57]. The number of reads mapping to each csRNA were identified using BEDTools, and RPKM values for each structure were calculated. IGV [58] was used to visualize coverage.

4.6. Characterizing csRNAs near genes

We determined which csRNAs occurred within 25 bp upstream of genes. Using RPSblast [38] against the CDD [45], we assigned protein domains to the encoding proteins of these downstream genes. We considered protein domains with assigned e-values of 0.01 or less across 80% of the query length. Additionally, we performed RNIE [46] with default settings to predict which csRNAs overlapped with predicted Rho-independent terminators. We assessed which csRNAs were palindromic or overlapping using BEDTools.

(B) Workflow of csRNAs from RNA virus datasets. HS-BLASTN is used to identify homologous regions, including coding regions. Regions are clustered and scored using RNAphylo and R-scape. csRNAs with unique, quality hits are retained, and those already present in Rfam are excluded.

(B) Histogram of the number of unique sequences representing the csRNAs.

(C) Histogram of the number of csRNAs identified in each viral phylum.

(D) Venn diagram of ecosystems in which csRNAs are found within IMG/VR.

Author statement

Thank you for considering our manuscript, now titled Identification of Over Ten Thousand Candidate Structured RNAs in Viruses and Phage, for publication. We have made substantial revisions, including the addition of an RNA-Seq analysis both to strengthen select predictions as well as better highlight limitations throughout the manuscript. We also improved transparency by providing code to walk through each step on GitHub. Finally, we made substantial edits throughout the paper based on reviewer comments and in general to improve readability and flow. We thank you again for this opportunity.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We 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. Computing costs were also supported via NIH S10 Shared Instrumentation Grant (1S10OD02014101), NIH R01 #AI148623–01, A Sloan Foundation Fellowship, and Damon Runyon Clinical Investigator Award to ASB.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.11.010.

References

1 Weinberg Z, et al. Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. Nucleic Acids Res 2017;45:10811–23.
2 Fremin BJ, Bhatt AS. Comparative genomics identifies thousands of candidate structured RNAs in human microbiomes. Genome Biol 2021;22:100.
3 Weinberg Z, et al. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol 2010;11:R31.
4 Fremin BJ, Kyrpides NC. Identifying candidate structured RNAs in CRISPR operons. RNA Biol 2022;19:678–85.
5 Rivas E, Clements J, Eddy SR. Estimating the power of sequence covariation for detecting conserved RNA structure. Bioinformatics 2020;36:3072–6.
6 Bloch S, Lewandowska N, Węgrzyn G, Nejman-Faleńczyk B. Bacteriophages as sources of small non-coding RNA molecules. Plasmid 2021;113:102527.
7 Millard AD, et al. An antisense RNA in a lytic cyanophage links psbA to a gene encoding a homing endonuclease (Preprint at) ISME J Vol 2010;4:1121–35. https://doi.org/10.1038/ismej.2010.43.
8 Leskinen K, Blasdel BG, Lavigne R, Skurnik M. RNA-Sequencing reveals the progression of phage-host interactions between φR1-37 and Yersinia enterocolitica. Viruses 2016;8:111.
9 Castillo-Keller M, Vuong P, Misra R. Novel mechanism of Escherichia coli porin regulation. J Bacteriol 2006;188:576–86.
10 Altuvia S, Storz G, Papenfort K. Cross-regulation between bacteria and phages at a posttranscriptional level (Preprint at) Regul RNA Bact Archaea 2018:499–514. https://doi.org/10.1128/9781683670247.ch29.
11 Delesalle VA, Tanke NT, Vill AC, Krukonis GP. Testing hypotheses for the presence of tRNA genes in mycobacteriophage genomes. Bacteriophage 2016;6:e1219441.
12 Ivanova NN, et al. Stop codon reassignments in the wild. Science 2014;344:909–13.
13 Katz A, Elgamal S, Rajkovic A, Ibba M. Non-canonical roles of tRNAs and tRNA mimics in bacterial cell biology. Mol Microbiol 2016;101:545.
14 Miras M, Miller WA, Truniger V, Aranda MA. Non-canonical translation in plant RNA viruses. Front Plant Sci 2017;8:494.
15 Firth AE, Brierley I. Non-canonical translation in RNA viruses (Preprint at) J Gen Virol Vol 2012;93:1385–409. https://doi.org/10.1099/vir.0.042499-0.
16 Krahn N, Fischer JT, Söll D. Naturally occurring tRNAs with non-canonical structures. Front Microbiol 2020;11:596914.
17 Roux S, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res 2021;49:D764–75.
18 Gregory AC, et al. Marine DNA viral macro- and microdiversity from pole to pole. e14 Cell 2019;177:1109–23. e14.
19 Gregory AC, et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. e8 Cell Host Microbe 2020;28:724–40. e8.
20 Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. Nat Protoc 2017;12:1673–82.
21 Roux S, et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat Microbiol 2019;4:1895–906.
22 Schulz F, et al. Giant virus diversity and host interactions through global metagenomics. Nature 2020;578:432–6.
23 Paez-Espino D, et al. Diversity, evolution, and classification of virophages uncovered through global metagenomics. Microbiome 2019;7:157.
24 Nayfach S, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* Prepr 2021. https://doi.org/10.1038/s41564-021-00928-6.
25 Mehrshad, M. et al. Energy efficiency and biological interactions define the core microbiome of deep oligotrophic groundwater. Preprint at https://doi.org/10.1101/2020.05.24.111179.
26 Bushman TJ, et al. Draft genome sequence of Mn(II)-oxidizing bacterium sp. strain AB_14. Microbiol Resour Announc 2019;8.
27 Garcia MO, et al. Soil microbes trade-off biogeochemical cycling for stress tolerance traits in response to year-round climate change. Front Microbiol 2020;11:616.
28 Mobilian C, et al. Differential effects of press vs. pulse seawater intrusion on microbial communities of a tidal freshwater marsh. Limnol Oceanogr Lett Prepr 2020. https://doi.org/10.1002/lol2.10171.
29 Espínola F, et al. Metagenomic analysis of subtidal sediments from polar and subpolar coastal environments highlights the relevance of anaerobic hydrocarbon degradation processes. Microb Ecol 2018;75:123–39.

30 Hatcher EL, et al. Virus variation resource - improved response to emergent viral outbreaks. Nucleic Acids Res 2017;45:D482–90.
31 Wolf YI, et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol 2020;5:1262–70.
32 Callanan J, et al. Expansion of known ssRNA phage genomes: from tens to over a thousand. Sci Adv 2020;6:eaay5981.
33 Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 2015;43:7762–8.
34 Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 2012;28:2223–30.
35 Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013;29:2933–5.
36 Kalvari I, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res 2021;49:D192–200.
37 Neri, U. et al. A five-fold expansion of the global RNA virome reveals multiple new clades of RNA bacteriophages. Preprint at https://doi.org/10.1101/2022.02.15.480533.
38 McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004;32:W20–5.
39 Washietl S, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 2011;17:578–94.
40 Gesell T, Washietl S. Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinforma 2008;9:248.
41 Fremin BJ, Sberro H, Bhatt AS. MetaRibo-Seq measures translation in microbiomes. Nat Commun 2020;11:3268.
42 Fremin BJ, Nicolaou C, Bhatt AS. Simultaneous ribosome profiling of hundreds of microbes from the human microbiome. Nat Protoc 2021;16:4676–91.
43 Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 2004;32:11–6.
44 Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9.
45 Marchler-Bauer A. CDD: a conserved domain database for protein classification (Preprint at) *Nucleic Acids Res* Vol 2004;33:D192–6. https://doi.org/10.1093/nar/gki069.
46 Gardner PP, Barquist L, Bateman A, Nawrocki EP, Weinberg Z. RNIE: genome-wide prediction of bacterial intrinsic terminators. Nucleic Acids Res 2011;39:5845–52.
47 Choi JS, Park H, Kim W, Lee Y. Coordinate regulation of the expression of SdsR toxin and its downstream pphA gene by RyeA antitoxin in Escherichia coli. Sci Rep 2019;9:9627.
48 Brierley I, Pennell S, Gilbert RJC. Viral RNA pseudoknots: versatile motifs in gene expression and replication. Nat Rev Microbiol 2007;5:598–610.
49 Uzan M, Miller ES. Post-transcriptional control by bacteriophage T4: mRNA decay and inhibition of translation initiation. Virol J 2010;7:360.
50 Hall MN, Gabay J, Débarbouillé M, Schwartz M. A role for mRNA secondary structure in the control of translation initiation (Preprint at) *Nat* Vol 1982;295:616–8. https://doi.org/10.1038/295616a0.
51 Takahashi MK, et al. Using in-cell SHAPE-Seq and simulations to probe structure-function design principles of RNA transcriptional regulators. RNA 2016;22:920–33.
52 Fremin BJ, Bhatt AS. Structured RNA contaminants in bacterial Ribo-Seq. mSphere 2020;5.
53 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features (Preprint at) *Bioinforma* Vol 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
54 Weinberg Z, Breaker RR. R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinforma 2011;12:3.
55 Larkin MA, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–8.
56 Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 2016;7:11257.
57 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.
58 Robinson JT, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24–6.