11-22-2022

AcaFinder: Genome Mining for Anti-CRISPR-Associated Genes

Bowen Yang

Jinfang Zheng

Yanbin Yin

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

AcaFinder: Genome Mining for Anti-CRISPR-Associated Genes

Bowen Yang,[a] Jinfang Zheng,[a] Yanbin Yin[a]

[a]Nebraska Food for Health Center, Department of Food Science and Technology, University of Nebraska–Lincoln, Lincoln, Nebraska, USA

ABSTRACT Anti-CRISPR (Acr) proteins are encoded by (pro)viruses to inhibit their host's CRISPR-Cas systems. Genes encoding Acr and Aca (Acr associated) proteins often colocalize to form *acr-aca* operons. Here, we present AcaFinder as the first Aca genome mining tool. AcaFinder can (i) predict Acas and their associated *acr-aca* operons using guilt-by-association (GBA); (ii) identify homologs of known Acas using an HMM (Hidden Markov model) database; (iii) take input genomes for potential prophages, CRISPR-Cas systems, and self-targeting spacers (STSs); and (iv) provide a standalone program (https://github.com/boweny920/AcaFinder) and a web server (http://aca.unl.edu/Aca). AcaFinder was applied to mining over 16,000 prokaryotic and 142,000 gut phage genomes. After a multistep filtering, 36 high-confident new Aca families were identified, which is three times that of the 12 known Aca families. Seven new Aca families were from major human gut bacteria (*Bacteroidota*, *Actinobacteria*, and *Fusobacteria*) and their phages, while most known Aca families were from *Proteobacteria* and *Firmicutes*. A complex association network between Acrs and Acas was revealed by analyzing their operonic colocalizations. It appears very common in evolution that the same *aca* genes can recombine with different *acr* genes and *vice versa* to form diverse *acr-aca* operon combinations.

IMPORTANCE At least four bioinformatics programs have been published for genome mining of Acrs since 2020. In contrast, no bioinformatics tools are available for automated Aca discovery. As the self-transcriptional repressor of *acr-aca* operons, Aca can be viewed as anti-anti-CRISPRs, with great potential in the improvement of CRISPR-Cas technology. Although all the 12 known Aca proteins contain a conserved helix-turn-helix (HTH) domain, not all HTH-containing proteins are Acas. However, HTH-containing proteins with adjacent Acr homologs encoded in the same genetic operon are likely Aca proteins. AcaFinder implements this guilt-by-association idea and the idea of using HMMs of known Acas for homologs into one software package. Applying AcaFinder in screening prokaryotic and gut phage genomes reveals a complex *acr-aca* operonic colocalization network between different families of Acrs and Acas.

KEYWORDS CRISPR-Cas, anti-CRISPR, bacteriophage, bioinformatics, helix turn helix

Viruses and mobile genetic elements (MGEs) have been in a constant arms race with their prokaryote hosts for billions of years (1). To prevent viral infections, prokaryotes have evolved various antiviral defense systems, e.g., CRISPR-Cas, RM (restriction modification), TA (toxin and anti-toxin), and BREX (Bacteriophage Exclusion) systems (2). To survive, viruses also evolved various antidefense strategies (3). Among the known antidefense systems, anti-CRISPRs have received the greatest attention due to their applications in developing more controllable and safer genome editing tools, e.g., CRISPR-Cas9 (4).

Anti-CRISPR (Acr) proteins were first discovered in 2013 from *Pseudomonas* phages and prophages (5). They were found to successfully protect invading phages by inhibiting the host's CRISPR-Cas systems. These published Acrs are often encoded by (pro)viral genetic operons that also contain genes of anti-CRISPR associated (Aca) proteins in the downstream of acr genes (5–7). As of now, a total of 98 Acr proteins have been

Editor Marnix Medema, Wageningen University

Address correspondence to Yanbin Yin, yyin@unl.edu.

The authors declare no conflict of interest.

Received 27 August 2022
Accepted 3 November 2022
Published 22 November 2022

experimentally characterized; however, they do not share any significant sequence similarity and thus form 98 Acr protein families. Most Acrs do not have conserved Pfam domains and thus have few sequence homologs in the databases. In contrast, 12 Aca protein families have been defined and all of them contain a conserved helix-turn-helix (HTH) domain, commonly found in DNA-binding proteins. Because Aca is more conserved and often coexist with Acrs in operons, many of the 98 Acrs were actually identified by the guilt-by-association (GBA) idea followed by experimental characterization, i.e., sequence similarity search of aca (HTH-containing) genes first and then search for acr genes in the genomic operons of (pro)viruses.

Recent studies have shown that at least three Aca families negatively regulate Acr expression (8–11). These Acas bind to the promoter regions of the *acr-aca* operons via their HTH domain, leading to transcription repression of the operons (9, 10). This is reminiscent of the type II toxin and antitoxin (TA) systems (12), where the anti-toxin protein is also an HTH-containing transcriptional repressor of the toxin. Therefore, Aca proteins could be viewed as anti-anti-CRISPRs, with great potential in the calibration of CRISPR-Cas technologies by modulating the Acr modulators.

In the past 3 years, bioinformatics tools have been developed to aid in the discovery of novel Acrs. These include tools as automated software screening query genomes or ranking/scoring query proteins for Acr candidates, e.g., AcRanker, AcrFinder, and PaCRISPR (13–15). There are also online databases (Anti-CRISPRdb, AcrDB, AcrHub, and AcrCatalog) for experimentally verified Acrs, their homologs, and machine learning-predicted Acrs (16–19). In contrast, no automated bioinformatics tools are available for Aca discovery. This might be because Acas have HTH domains and are easier to identify. However, not all HTH-containing proteins are Acas. In fact, there are 328 HTH-related Pfam families forming the HTH clan (a higher classification level than family) due to shared distant evolutionary homology. This makes the HTH clan (Pfam clan ID: CL0123) the largest and probably also one of the most conserved clan in the Pfam database. Therefore, the HTH sequence space is much larger than the Aca sequence space, and finding HTH-contain proteins does not mean that Acas are found.

Here, we report the first software package, AcaFinder, to allow automated genome mining for reliable Acas. Two approaches are implemented to more confidently identify Acas given a query genome or metagenome-assembled genome. The first approach is based on guilt-by-association (GBA), meaning that we identify homologs of known Acrs first and then search for HTH-containing proteins in the acr gene neighborhood. The rationale is that HTH-containing proteins are more likely to be real Acas if they are located in the same genetic operons as known Acrs or their homologs. The second approach is to build an HMM (hidden Markov model) database using training data of the 12 known Aca families, and then search for Aca homologs with this AcaHMMdb instead of Pfam HTH HMMs. In addition to the two implemented approaches, AcaFinder also calls a CRISPR-Cas search tool (CRISPRCasTyper), a prophage search tool (VIBRANT), and an in-house self-targeting spacer (STS) searching process, providing users with detailed information vital to the assessment of Aca predictions (20–23).

Features of AcaFinder include (i) identify both potential Acas and their associated *acr-aca* operons; (ii) identify Aca homologous proteins using the AcaHMMdb; (iii) provide potential prophage regions, CRISPR-Cas systems, and STSs in the query genome; and (iv) provide a standalone software package that can run locally and a user-friendly web interface. The web server generates graphical representations of identified *acr-aca* operons with associated CRISPR-Cas, prophage, and STS information in terms of genomic context. Finally, using AcaFinder, we have screened for potential Aca proteins in the RefSeq prokaryote genomes and in the gut phage database (GPD) (24–27). We have performed a phylogenetic analysis to study the sequence diversity and taxonomic distribution of a group of highly confident Aca predictions.

RESULTS

Performance evaluation of AcaFinder. AcaFinder is the first computational tool for automated Aca identification. The AcaHMM homology approach (Fig. 1) can identify

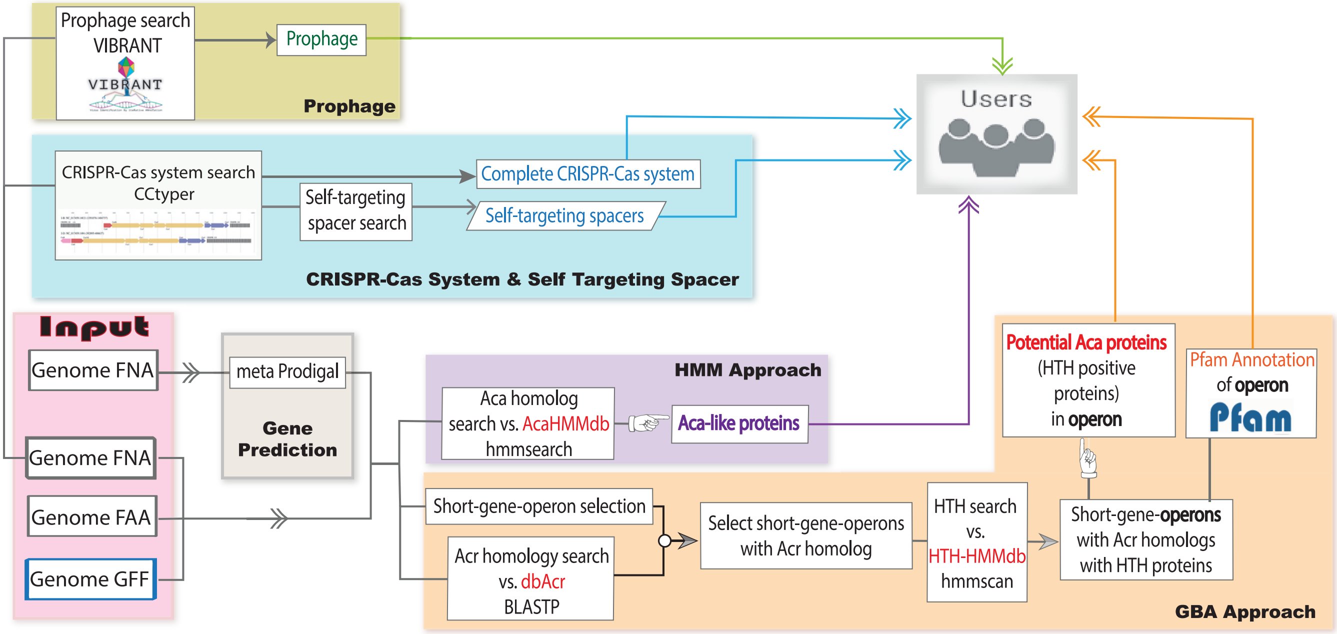

FIG 1 AcaFinder workflow. With provided input, AcaFinder proceeds to two Aca screening routes: (i) HMM approach to find Aca homologous proteins (purple box) and (ii) GBA (guilt-by-association) approach to find acr-aca operons (orange box), which must contain at least one Acr homolog and one HTH-containing gene (Aca candidate). Complete CRISPR-Cas along with STSs (blue box) and prophage regions (green box) are also searched from input genomic sequences (described in the main text). All generated results were combined and are provided to the users as tables and graphs.

Aca-like proteins sharing significant homology to the 12 known Acas. The GBA approach is more sensitive as it will identify all short HTH-containing proteins that reside in the same genetic operons with homologs of published Acrs. To evaluate the performance of AcaFinder, we have run it on the source genomes of the 12 known Acas. For each of the 12 known Acas, the genomic sequences (fna file) as well as protein annotations (faa, gff files) were downloaded from NCBI and used as input to run the AcaFinder.

Using the GBA approach (yellow box, Fig. 1), all 12 known Acas can be identified (Table 1), giving the approach a recall of 100%. With the HMM approach (purple box, Fig. 1), 10/12 (83.3%) known Acas were identified using the default parameter setting (HMM coverage >60% and E value <1e-10), which is designed to reduce false positives. As expected, when more relaxed thresholds were used (coverage >30% and E value <1e-3), the two missed known Acas (Aca6 and Aca9) were indeed found back (Table S3).

In addition to recall (the true positives divided by all positives [12]), precision is also often reported in bioinformatics tool evaluation. Precision is calculated as the true positives divided by all predictions (true positives + false positives). However, the 12 known Acas only represent a very tiny fraction of all the possible Acas that exist (28). Thus, it is not rational and acceptable to treat all other predictions as false positives, i.e., those candidate Acas from a AcaHMMdb search and *acr-aca* operons from the GBA search in the same genomes as the known Acas. In fact, it is observed that one genome can encode multiple types of Acrs (29–31) and Acas (12, 17, 31). In AcaFinder search results (Table 1 and Table S3), all the source genomes encoding the 12 known Acas had additional Aca candidates identified. Currently, we do not know whether they are real Acas or false positives, except that they have significant sequence similarity with known Acas (E value <1e-10 and coverage >60%) or are present in short gene operons (SGOs) that also encode Acr homologs. However, in most cases, the GBA gives a smaller number of Aca candidates, which may be more confident predictions given the strict SGO constraint (Table 1).

Although there are no other similar tools that can be compared, we sought to evaluate the performance of AcaFinder's GBA approach using leave-one-out (LOO) experiments of

TABLE 1 Test AcaFinder on genomes encoding the 12 known Acas

Genome ID	Known aca	Length (aa)	GenBank ID	Found by AcaFinder	Total aca candidates found[a]	Found in leave one out[b]
GCF_000904095.1	AcaI	79	YP_007392343	Yes	1 + 4	N + N
GCF_000381965.1	AcaII	125	WP_019933869.1	Yes	1 + 4	N + N
GCF_001066195.1	Aca3	70	WP_049360086.1	Yes	3 + 2	N + Y
GCA_002194095.1	Aca4	67	OWI97558.1	Yes	8 + 5	N + N
GCF_000754765.1	Aca5	60	WP_039494319.1	Yes	4 + 2	N + N
GCF_000518385.1	Aca6	65	WP_035450933.1	Yes	2 + 5	N + N
GCF_001662305.1	Aca7	68	WP_064702654.1	Yes	3 + 1	N + Y
GCF_001725895.1	Aca8	55	YP_009272953.1	Yes	2 + 3	N + Y
GCF_004135975.1	Aca9	69	WP_129352084.1	Yes	2 + 1	N + Y
GCA_004745455.1	Aca10	65	TGC30851.1	Yes	3 + 4	Y + Y
GCF_000314775.2	Aca11	63	WP_009730540.1	Yes	12 + 7	N + Y
GCF_000526075.1	Aca12	140	WP_231458942.1	Yes	18 + 16	Y + Y

[a] Two numbers: the first is the Aca count from the AcaHMM search, and the second is from the GBA search. Details are provided in Table S3.
[b] Two values: the first is to indicate if the known Aca was identified from the AcaHMM search (Y, yes; N, no), and the second is from the GBA search.

the 12 known Acas. The idea of LOO is that we will remove one Aca in each experiment, more specifically, by removing all published Acrs that are known to colocalize with that Aca from the database Acr (dbAcr). The colocalizations of Acr and Aca (Fig. 2 and Table S4) were manually curated from literature. For example, AcaII has been shown to colocalize with AcrIF6, AcrIF8, AcrIF9, AcrIF10, AcrIIC1, AcrIIC2, AcrIIC4, and AcrIIC5 in the literature (32–35). However, the known colocalizations are obviously incomplete, and numerous unknown *acr-aca* colocalizations remain to be characterized. In the LOO experiment of

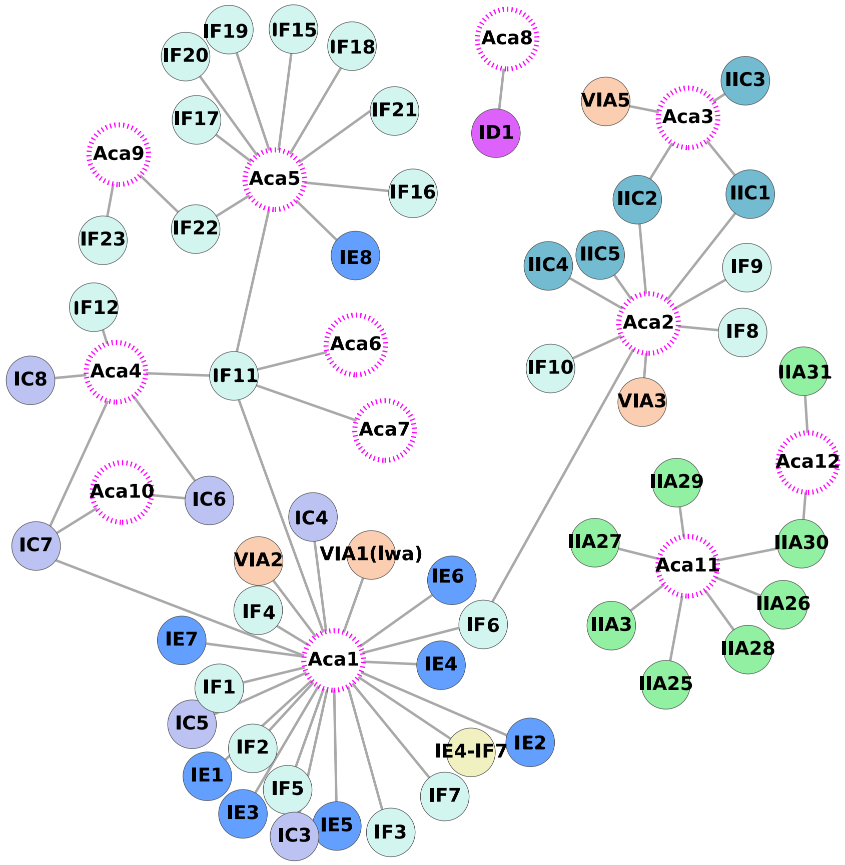

FIG 2 Colocalizations of known Acrs and Acas curated from literature. The 12 published Aca families are shown in open circles with pink dotted edges; Acr families are shown in filled circles colored based on type. The detailed data can be found in Table S4.

AcaII, we removed all known Acr associates and then ran through the GBA route of the AcaFinder to see if AcaII can be identified in the source genome of Acas.

The result (Table 1) shows that 7/12 known Acas were found in the LOO experiments (recall = 58.3%), while AcaI, 2, 4, 5, and 6 could not to be found after removing Acrs known to colocalize with them. Note that the GBA approach has the assumption that aca genes of one family can be recruited to the gene neighborhood of different acr gene families and vice versa. Therefore, if we remove all Acrs known to colocalize with an Aca, we will not find the Aca back, unless some other Acrs in dbAcr also colocalize with the Aca but such colocalizations have not been reported yet (i.e., not in Fig. 2 and Table S4). Therefore, the reason that Acas1, 2, 4, 5, and 6 failed to be found is that after removal of Acrs known to colocalize with these Acas, no other Acrs in dbAcr can find them back. However, new Acrs are constantly being characterized. It is likely that continuously adding new Acrs in dbAcr will reveal more *acr-aca* colocalizations and improve the recall of AcaFinder in LOO experiments.

We have also run the LOO experiments to evaluate AcaFinder's AcaHMM approach as the baseline. For each Aca, we removed its HMM from the AcaHMMdb and ran AcaFinder to see if the Aca can be found by the remaining 11 Aca HMMs. As expected, only two Acas were found in the LOO experiments (Table 1): Aca10 was identified by AcaI HMM and Aca12 by Aca11 HMM. Obviously, the AcaHMM approach is not good for finding novel Aca families, as between different Aca families there are very low sequence similarities.

Utilities of AcaFinder web server and standalone program. AcaFinder is provided as a standalone program and a web server. Genomic sequences in fna, faa, and gff files are accepted as input. Only providing a fna file is also allowed and recommended, in which case, gff and faa files will be generated by running Prodigal (36). Genomic data of archaea, bacteria, and viruses are all allowed. By providing the Virus flag (–Virus), viral data will not run CCtyper for CRISPR-Cas scanning (STSs search also will not run), nor VIBRANT for prophage prediction.

The AcaFinder website is powered by SQLite + Django + JavaScript + Apache + HTML. A help page is available to provide users with step-by-step instructions on the usage of the webserver, along with the interpretation of the outputs. Users can submit FASTA sequences of their genomes (Fig. S1A in the supplemental material). A prokaryotic genome on average takes 10 min of runtime, whereas a virus genome/contig takes 1 min. Users can skip prophage and CRISPR-Cas searches, and the job-predicting Acas can finish in 1 to 2 min on average. The result page has tables showing information on predicted Aca proteins and associated genes within the same *acr-aca* operons of the GBA approach and the Aca-like proteins from the AcaHMM approach (Fig. S1B). Prophage and complete CRISPR-Cas system information will be provided if identified; in addition, their relationship with the predicted Acas and *acr-aca* operons will be indicated (Fig. S1C). Graphical representations will be generated if the input is prokaryotic genome, showing where the aca genes/operons are located on the input genome/contig(s) and how far they are from prophage regions and CRISPR-Cas systems. If any STSs were found, a link will connect the CRISPR spacer and the target region on the genome/contig(s) (Fig. S1D).

Genome mining for new Aca families. A total of 15,201 complete bacterial genome, 961 archaea genomes of the RefSeq database (37), and 142,809 viral contigs of the human gut phage database (GPD) (38) were scanned with AcaFinder using the GBA approach. We used the "–HTH-HMM_strictdb" flag for potential Aca proteins/operons. To further increase the confidence level of predictions, we limited the predicted Acas from RefSeq to genomes with complete CRISPR-Cas systems and within prophage regions. To cluster predicted Acas into potential families, Cd-Hit (39) was used with a 40% sequence identity threshold, on the basis of proteins above this threshold are more likely to share structure and function similarities (15). Cd-Hit clusters/families with a size ≥3 were selected to filter out singletons and smaller-size Aca families.

After all these processes, a total of 1,422 Aca families were found (Table S5 and Fig. S2). We further selected Aca families that have at least 1 Aca member located next

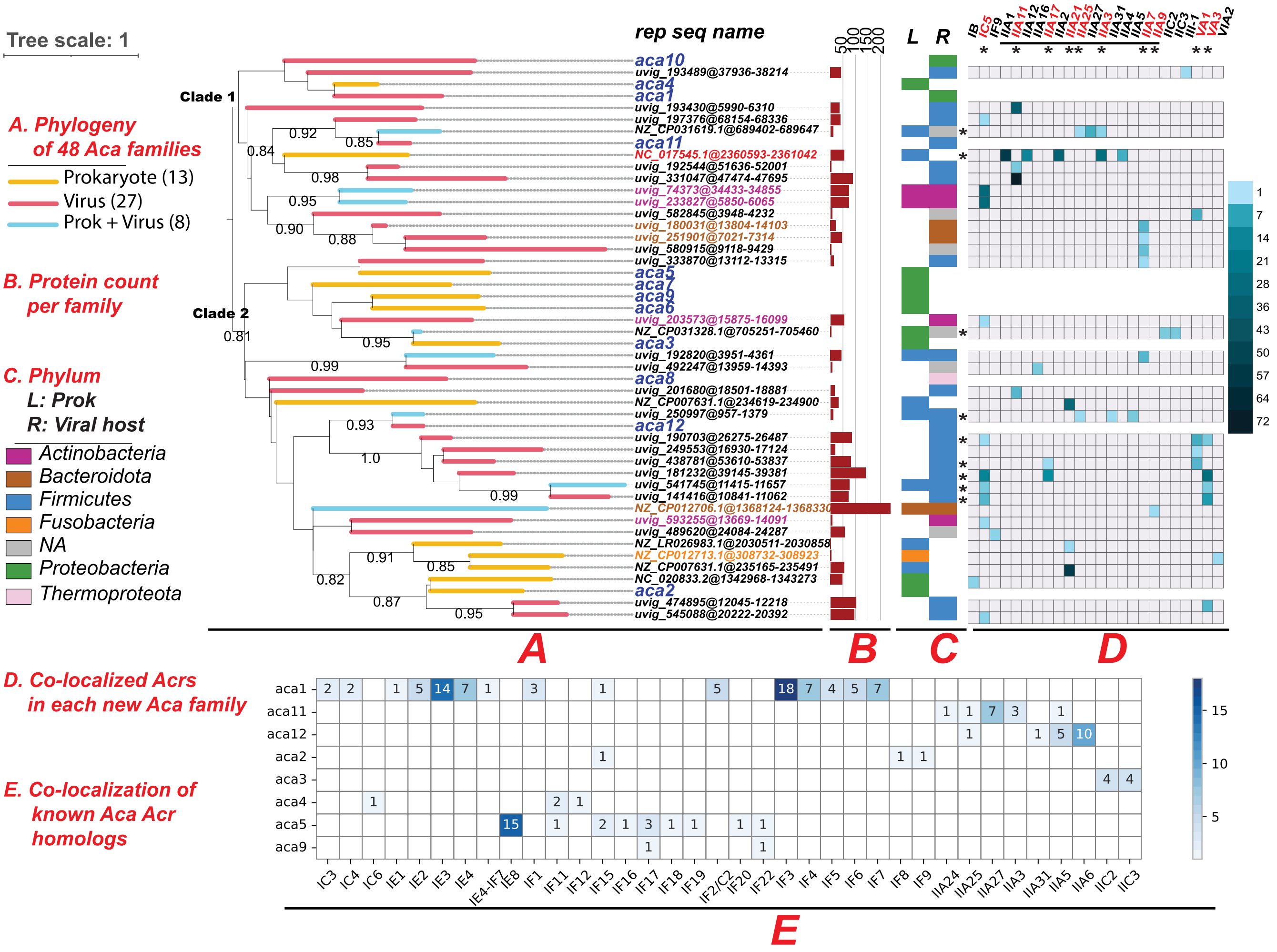

FIG 3 Phylogenetic and taxonomic distribution of 48 Aca families. (A) Branch colors represent Prokaryotic (yellow), Viral (red), or Prokaryotic + Viral (blue) protein content of each Aca family. Bootstrap values >0.8 are shown beside the nodes; seq ID of each node are located to the right of the tree, with 12 published Aca highlighted in blue and the predicted Acas printed in "Contig"@"genomic location" format (e.g., NZ_LR026983.1). (B) Protein counts of each family are presented as barplot (dark red); (C) Phyla of Aca proteins are displayed in two stacked bars, with one representing prokaryotic proteins (left) and the other viral proteins (right). For Viral proteins, the hosts' phyla are displayed. The legend of the stacked barplot resides at the leftmost of the figure. (D) The heatmap displays the member counts of the 36 Aca families that are colocalized with homologs of the 89 known Acr families (columns) in genomic operons. Rows and columns are labeled with "*" if they have multiple filled cells, meaning that Aca family is colocalized with more than one Acr families and vice versa. (E) The heatmap shows the colocalization of protein homolog counts of known Aca families and homologs of the 89 known Acr families in genomic operons of RefSeq and GPD genomes.

to homologs of the 98 experimentally characterized Acr proteins. Only 36 Aca families remained but were considered to have a very high probability of being true Aca families. Combining the representative sequences of the 36 families and the 12 published Acas, a phylogenetic tree was constructed using the multiple sequence alignment of the 48 protein sequences (40, 41) (Fig. 3A). These proteins form 2 major clades, and the 12 published Acas are spread across different clades of the constructed tree (Fig. 3A and Fig. S2 for a larger tree with all 1,422 Aca families). This indicates that the 12 published Aca proteins are of rather distant families and also strengthens the point that a large sequence diversity of Acas in nature awaits to be discovered. Some of the new Aca families have a large size (e.g., over 200 members, Fig. 3B), and the average family size is 49. Among the 48 families, 8 contain members from both RefSeq prokaryotes and GPD phages, 27 are exclusively from GPD phages, and 13 are exclusively from RefSeq prokaryotes.

We have examined the taxonomic origin of the member proteins of each Aca family at the phylum level. The phylum of member proteins of RefSeq prokaryotes was plotted (L or left of Fig. 3C). However, most viral contigs of GPD were not assigned to a viral taxonomy group. Hence, we plotted the phylum of their hosts (R or right stripe of Fig. 3C), which were determined in the original paper of GPD (38). The 12 known Acas are from two phyla: *Proteobacteria* and *Firmicutes* (except Aca8 from viruses of *Thermoproteota archaea*), while the 36 new Aca families are distributed in three additional bacterial phyla and their phages (*Bacteroidota*, *Actinobacteria*, and *Fusobacteria*) known to be important human gut bacteria. Note that the four Aca families of *Actinobacteria* are all colocalized with AcrIC5, and the three Aca families of *Bacteroidota* origin are all colocalized with AcrIIA7 and IIA9 (Fig. 3D).

We have studied the colocalizations of Acas and Acrs in the same genetic operons. First, Fig. 2 depicts all the known *acr-aca* colocalizations extracted from literature (Table S4), showing that 9 (75%) of 12 known Aca families are colocalized with multiple known Acr families. AcaI, AcaII, Aca5, and Aca11 have 22, 9, 10, and 7 colocalized Acr families, respectively. Some Acas are even colocalized with different Acr types (e.g., AcaI with Acr type IE, IF, IC, and VIA). Similarly, 5 Acr families are colocalized with multiple Acas. For example, AcrIC7 is colocalized with AcaI, Aca4, Aca10, and AcrIF11 is with AcaI and Acas 4 to 7 (Fig. 2). Although there are many Acrs colocalizing with just one Aca, only three Acas colocalize with exclusively one Acr (Aca6-AcrIF11, Aca7-AcrIF11, and Aca8-AcrID1).

In the attempt to expand these observations, a homology search against the RefSeq and GPD genomes with known Aca and Acr proteins as a query was performed using blastp (coverage >80%, E value <1e-10). Only Acr and Aca homologs present in the same SGOs were considered colocalizations, and their instances were recorded (Fig. 3E and Table S6). Most of the colocalizations are already known (Fig. 2), but there are a few new ones (e.g., AcaI-AcrIF15, AcaII-AcrIF15, Aca9-AcrIF17, Aca11-AcrIIA24, Aca12-AcrIIA5, and Aca12-AcrIIA6, Fig. 3E). Data in Fig. 2 and Fig. 3E suggest that it is very common in evolution that the same aca genes recombine with different acr genes and vice versa to form diverse *acr-aca* operon combinations.

We further examined the colocalizations of the 36 new Acas and known Acr homologs. We found that 9 (25%) of the 36 new Aca families are colocalized with multiple Acr families, which is different from known *acr-aca* colocalizations (Fig. 2). This is probably due to our stringent filtering process for the 36 new Aca families. The new Aca family "NC_017545.1" of *Firmicutes* has the most types of colocalized Acrs (AcrIIA1 + AcrIIA12 + AcrIIA2 + AcrIIA3 + AcrIIA4, Fig. 3D). This suggests this Aca family has evolved to regulate various subtype IIA Acr subfamilies in *Firmicutes*. We also found that 10 Acrs are colocalized with multiple new Aca families, compared to only five Acrs colocalizing with multiple known Acas (Fig. 2). AcrIC5 is known to colocalize with AcaI (Fig. 2) but colocalize with 10 new Aca families from distant taxonomic groups (Fig. 3D). Similarly, AcrIIA11 colocalize with four, AcrIIA7 with five, AcrVA1 with four, AcrVA3 with four new Aca families. These diverse Aca family colocalizations indicate the flexibility of these Acrs in accepting different Aca families as transcriptional regulators.

DISCUSSION

Anti-CRISPRs (Acrs) have been under extensive studies since their discovery in 2013. These studies include various biotechnical and biomedical applications of Acrs, e.g., in genome editing, with very promising success (4, 42). Compared to Acrs, Acas are underresearched, although as anti-anti-CRISPRs, Acas also have great potential in the development of genome editing biotechnology. In addition, due to Aca's association with Acrs, the study of Acas can assist the continuous discovery of novel Acrs (5, 8).

AcaFinder is the first automated Aca prediction tool. A common practice in the discovery of the 12 published Acas is to search the Acr gene neighborhood for Pfam HTH-containing proteins (i.e., the GBA approach). AcaFinder implements this GBA approach and an Aca HMM approach (Fig. 1) as an automated computer program and webserver and thus will assist anti-CRISPR researchers to perform more rapid genome screening for Acas. In addition, AcaFinder also searches for inverted repeats (IR) in the promoter

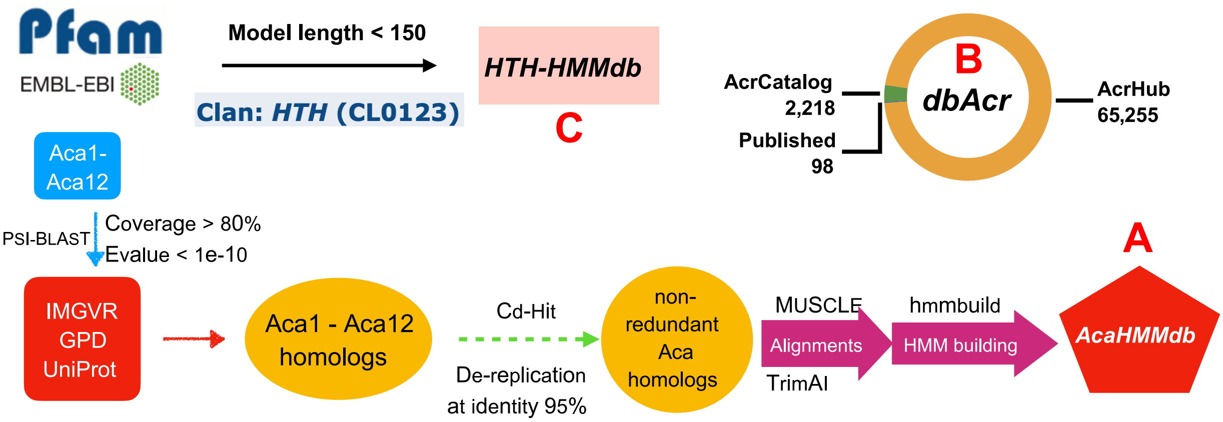

FIG 4 Three databases are built within AcaFinder. (A) The workflow to build the AcaHMMdb. Homologs of 12 known Acas were used in the construction of 12 HMMs. (B) The data composition of the Acr database (dbAcr). (C) The making of HTH-HMM database. The Pfam database was downloaded and filtered for HTH HMMs that fit the set criteria (described in the main text).

regions of predicted Acr-Aca operons, accounting for the recent finding that Aca proteins bind to IR regions for the regulation of Acr transcription (11, 43).

Because there are no other similar tools that we can compare AcaFinder with, to evaluate its performance, we have conducted LOO (leave one out) experiments using the 12 known Acas. The recall of the GBA approach in LOO experiments was 58.3%. This low recall is not surprising because of the way LOO worked: removing all Acrs associated/colocalized with an Aca from dbAcr and using other Acrs to find the Aca back. GBA requires the search for Acr homologs as the first step. If the remaining Acrs in dbAcr do not have homologs in the gene neighborhood of the tested Aca, then the Aca will not be found. Therefore, the GBA approach relies on the assumption that the same Acr family can recombine with different Aca families to form *acr-aca* operons in different genomes and vice versa. This assumption is supported by the observations made in Fig. 2, Fig. 3D, and Fig. 3E, which only represent a small fraction of possible *acr-aca* associations in nature. Future characterization of more Acrs, Acas, and their colocalizations to form genetic operons will undoubtedly reveal a much higher diversity of Acr and Aca combinations and help improve the power of the GBA approach in AcaFinder.

AcaFinder has the following limitations: (i) the GBA approach relies on Acr references to locate short-gene operons; thus, certain Acas that do not have Acrs in proximity will be missed and Acas predicted may be biased toward the Acrs used for reference; and (ii) the Aca HMM approach relies on HMMs built from sequence alignments of known Aca homologs; thus, novel Acas that share little to none sequence similarity to any known Acas will likely be missed. The drawbacks mentioned are due to the workflow design, which was intended to minimize false positives. With the continuous characterization of novel Acrs and Acas, we plan to update AcaFinder once a year to update the AcaHMMdb and dbAcr. In addition, we also plan to apply machine learning methods such as RNN and SVM for Aca discovery to be included into AcaFinder in the future.

MATERIALS AND METHODS

Build the AcaHMMdb. For Aca homology search, we built 12 HMMs corresponding to the 12 published Aca protein families with the following steps (Fig. 4A):

(i) We downloaded the UniProt (44), IMGVR (25), and GPD protein databases;
(ii) We downloaded the 12 published Aca protein (Aca1-Aca12) sequences from https://tinyurl.com/anti-CRISPR (45). Aca13 is excluded as it has no HTH domain and is not validated for Aca functions. The 12 Aca sequences were used as the query to PSI-BLAST search the three protein databases. Acr protein homologs with query coverage >80% and E value <1e-10 and protein length <150 amino acids (aa) were considered as Aca(1-12)-like proteins;
(iii) Cd-Hit (39) was used to dereplicate the Aca(1-12)-like proteins with an identity threshold <95%;

(iv) Using MUSCLE (46), 12 multiple sequence alignments were created from the dereplicated Aca(1-12)-like proteins along with the corresponding known Acas; the alignments were then trimmed with TrimAl (47). The details of the alignment trimming are provided in Table S1; and

(v) hmmbuild (48) was used to create 12 HMMs based on the 12 trimmed alignments; the 12 HMMs were combined into one file as the AcaHMMdb.

Collect Acr sequences to form the database of Acr. For GBA search of Acas, we prepared a highly confident Acr database from two sources. The first source contains protein sequences of 98 experimental characterized Acrs downloaded from https://tinyurl.com/anti-CRISPR (45). The other source includes machine learning-predicted Acr sequences that are less than 200 aa in length and share no sequence similarity with the 98 known Acrs. Those include 2,218 Acr sequences from AcrCatalog (17) and 65,255 Acr sequences from AcrHub (18). More detailed information can be found on the AcrCatalog and AcrHub websites. Therefore, the dbAcr contains 98 + 2,218 + 65,255 Acr protein sequences (Fig. 4B).

Build the HTH-HMM database (HTH-HMMdb and its subset HTH-HMM_strictdb). For the GBA search of Acas, we also built an HTH HMM database. An HMMER search of the 12 known Acas against the Pfam database found that all of them have their best Pfam HMM match from the HTH clan (clan ID: CL0123) (Table S2). Among them, nine Acas matched HTH HMMs with "HTH" in their Pfam family descriptions (e.g., HTH_24, HTH_3, and HTH_29), and three Acas matched HTH HMMs without "HTH" in their Pfam family descriptions (e.g., KORA, DUF1870, and YdaS_antitoxin).

Knowing that all known Acas match the HTH HMMs, we downloaded HMMs of the Pfam HTH clan (CL0123). Only keeping HMMs with length <150 aa (all 12 known Acas are shorter than 150 aa), we obtained in total 328 HMMs to form the HTH-HMM database (Fig. 4C). Another more conservative database, the HTH-HMM_strictdb, was also built with only 89 Pfam HTH HMMs that must have "HTH" in their family descriptions (e.g., HTH_24, HTH_3). Users have the option to choose either the HTH-HMMdb or the HTH-HMM_strictdb for searching the gene neighborhood of Acr homologs (see below).

AcaFinder workflow. Given a nucleotide sequence fna file, AcaFinder will call gene prediction tool Prodigal (36) to generate a protein sequence faa file and an associated gff (General Feature Format) file. In addition, AcaFinder also allows users to submit their own annotation files (fna, faa, and gff files, pink box, Fig. 1). After initial input, the workflow splits into two routes.

One route (purple box, Fig. 1) uses hmmsearch with the built-in AcaHMMdb as query and user input faa file as the database. Any HMM hits that passed set filters (HMM coverage >60% and E value <1e-10) will be extracted as Aca-like proteins, which will be provided as outputs to the user.

The other route uses the GBA approach to find HTH-containing Aca candidates in the Acr gene neighborhood (orange box, Fig. 1). Compared to the simple search for HTH-containing proteins, this GBA approach considers the cooccurrence of *acr* and *aca* genes in the same genetic operons. Thus, it adds a strong constraint in the identification of highly confident novel Aca proteins. There are three consecutive steps:

Step 1: The input faa file will be used as a query to DIAMOND blastp against the built-in dbAcr for Acr homologs (coverage >60% and E value <1e-3). Once Acr homologs are determined, the input fna file will be scanned for short-gene-operons (SGOs) by the following criteria: (i) at least one acr homologous genes in the SGO; (ii) all genes on the same strand; (iii) all intergenic distances <250 bp; and (iv) all genes have protein sequence length <200 aa (except that when the Acr homologs are homologous to known Acrs that are longer than 200 aa, e.g., AcrIIIB1 [249 aa], AcrVA2 [322 aa]).

Step 2: Each SGO will be scanned for HTH-containing genes using hmmscan (HMM coverage >40% and E value <1e-3) with the HTH-HMMdb or its subset HTH-HMM_strictdb as database.

Step 3: HTH-containing proteins (<150 aa) from SGOs will be output as candidate Acas. All non-Aca and non-Acr genes in SGOs will be further annotated with Pfam database using PfamScan. Information regarding each SGO, the contained Acr homolog, and candidate Acas will be provided to the user.

In addition to the identification of Acas and *acr-aca* operons, AcaFinder will also scan the input fna file for prophages, CRISPR-Cas systems, and STSs. Bacterial genomes carrying CRISPR-Cas systems and STSs are more likely to encode Acr/Aca proteins to prevent genome self-destruction (49). Prokaryotic genome input will be scanned for complete CRISPR-Cas systems (presence of both CRISPR arrays and Cas operons) with CRISPRCasTyper (20). With the presence of complete CRISPR-Cas, blastn will be then performed with all associated CRISPR spacers as query and user's input fna file as database (Fig. 1) for STSs. CRISPR-Cas and STSs information will be also provided to users independently, together with *acr-aca* operons from the previous step.

Acr/Aca genes are more likely to be found in prophages (50). To integrate prophage information, VIBRANT (21) will be used in the search for prophage regions within the user's prokaryotic genomic input. All discovered prophage regions will be provided to the user as a table, as well as VIBRANT's original outputs. *acr-aca* operons or Aca-like proteins that reside in any of the identified prophages will be indicated.

AcaFinder does not report a prediction score. However, the CRISPR-Cas, STSs, and prophage output from AcaFinder will allow users to further filter the predicted Acas. We recommend high-quality Acas fit the following categories: (i) predicted by AcaHMMdb, i.e., high-sequence homology to the 12 known Acas; (ii) located within a predicted prophage region; (iii) and from genomes that contain complete CRISPR-Cas systems and STSs. In addition, AcaFinder searches the 400 bp upstream potential promoter region of predicted operons using Palindrome of EMBOSS (51), providing user extra information on Aca putative binding sites.

Phylogenetic analysis. MAFFT (40) was used for sequence alignment, and FastTree (41) was used for phylogenetic tree construction. The phylogeny was then visualized in the iTOL (Interactive Tree Of Life, https://itol.embl.de/) web server (52).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.4 MB.
FIG S2, PDF file, 1.7 MB.
TABLE S1, XLSX file, 0.01 MB.
TABLE S2, XLSX file, 0.01 MB.
TABLE S3, XLSX file, 0.03 MB.
TABLE S4, XLSX file, 0.01 MB.
TABLE S5, XLSX file, 0.6 MB.
TABLE S6, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH) awards R21AI17195 and R01GM140370, the U.S. Department of Agriculture (USDA) award 58-8042-7-089, and the Nebraska Tobacco Settlement Biomedical Research Enhancement Funds as part of a start-up grant of University of Nebraska–Lincoln (2019-YIN to Y.Y).

We thank the lab members for helpful discussions. This work was partially completed utilizing the Holland Computing Center of the University of Nebraska–Lincoln, which receives support from the Nebraska Research Initiative.

REFERENCES

1. Stern A, Sorek R. 2011. The phage-host arms race: shaping the evolution of microbes. Bioessays 33:43–51. https://doi.org/10.1002/bies.201000071.
2. Bernheim A, Sorek R. 2020. The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol 18:113–119. https://doi.org/10.1038/s41579-019-0278-2.
3. Samson JE, Magadán AH, Sabri M, Moineau S. 2013. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol 11:675–687. https://doi.org/10.1038/nrmicro3096.
4. Marino ND, Pinilla-Redondo R, Csörgő B, Bondy-Denomy J. 2020. Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. Nat Methods 17:471–479. https://doi.org/10.1038/s41592-020-0771-6.
5. Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. 2013. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature 493:429–432. https://doi.org/10.1038/nature11723.
6. Pawluk A, Bondy-Denomy J, Cheung VH, Maxwell KL, Davidson AR, Hendrix R. 2014. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. mBio 5:e00896-14. https://doi.org/10.1128/mBio.00896-14.
7. Mahendra C, Christie KA, Osuna BA, Pinilla-Redondo R, Kleinstiver BP, Bondy-Denomy J. 2020. Author Correction: broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer. Nat Microbiol 5:872–872. https://doi.org/10.1038/s41564-020-0726-9.
8. Stanley SY, Borges AL, Chen K-H, Swaney DL, Krogan NJ, Bondy-Denomy J, Davidson AR. 2019. Anti-CRISPR-associated proteins are crucial repressors of anti-CRISPR transcription. Cell 178:1452–1464.e13. https://doi.org/10.1016/j.cell.2019.07.046.
9. Birkholz N, Fagerlund RD, Smith LM, Jackson SA, Fineran PC. 2019. The autoregulator Aca2 mediates anti-CRISPR repression. Nucleic Acids Res 47:9658–9665. https://doi.org/10.1093/nar/gkz721.
10. Liu Y, Zhang L, Guo M, Chen L, Wu B, Huang H. 2021. Structural basis for anti-CRISPR repression mediated by bacterial operon proteins Aca1 and Aca2. J Biol Chem 297:101357. https://doi.org/10.1016/j.jbc.2021.101357.
11. Shehreen S, Birkholz N, Fineran PC, Brown CM. 2022. Widespread repression of anti-CRISPR production by anti-CRISPR-associated proteins. Nucleic Acids Res 50:8615–8625. https://doi.org/10.1093/nar/gkac674.
12. Yin Y, Yang B, Entwistle S. 2019. Bioinformatics identification of anti-CRISPR loci by using homology, guilt-by-association, and CRISPR self-targeting spacer approaches. mSystems 4:e00455-19. https://doi.org/10.1128/mSystems.00455-19.
13. Yi H, Huang L, Yang B, Gomez J, Zhang H, Yin Y. 2020. AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses. Nucleic Acids Res 48:W358–W365. https://doi.org/10.1093/nar/gkaa351.
14. Wang J, Dai W, Li J, Xie R, Dunstan RA, Stubenrauch C, Zhang Y, Lithgow T. 2020. PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins. Nucleic Acids Res 48:W348–W357. https://doi.org/10.1093/nar/gkaa432.
15. Eitzinger S, Asif A, Watters KE, Iavarone AT, Knott GJ, Doudna JA, Minhas Fayyaz Ul Amir A. 2020. Machine learning predicts new anti-CRISPR proteins. Nucleic Acids Res 48:4698–4708. https://doi.org/10.1093/nar/gkaa219.
16. Huang L, Yang B, Yi H, Asif A, Wang J, Lithgow T, Zhang H, Minhas FuAA, Yin Y. 2021. AcrDB: a database of anti-CRISPR operons in prokaryotes and viruses. Nucleic Acids Res 49:D622–D629. https://doi.org/10.1093/nar/gkaa857.
17. Gussow AB, Park AE, Borges AL, Shmakov SA, Makarova KS, Wolf YI, Bondy-Denomy J, Koonin EV. 2020. Machine-learning approach expands the repertoire of anti-CRISPR protein families. Nat Commun 11:3784. https://doi.org/10.1038/s41467-020-17652-0.
18. Wang J, Dai W, Li J, Li Q, Xie R, Zhang Y, Stubenrauch C, Lithgow T. 2021. AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins. Nucleic Acids Res 49:D630–D638. https://doi.org/10.1093/nar/gkaa951.
19. Dong C, Hao G-F, Hua H-L, Liu S, Labena AA, Chai G, Huang J, Rao N, Guo F-B. 2018. Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins. Nucleic Acids Res 46:D393–D398. https://doi.org/10.1093/nar/gkx835.
20. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ. 2020. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. Crispr J 3:462–469. https://doi.org/10.1089/crispr.2020.0059.
21. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8:90. https://doi.org/10.1186/s40168-020-00867-0.
22. Watters KE, Fellmann C, Bai HB, Ren SM, Doudna JA. 2018. Systematic discovery of natural CRISPR-Cas12a inhibitors. Science 362:236–239. https://doi.org/10.1126/science.aau5138.
23. Jiang F, Doudna JA. 2015. The structural biology of CRISPR-Cas systems. Curr Opin Struct Biol 30:100–111. https://doi.org/10.1016/j.sbi.2015.02.002.
24. Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504. https://doi.org/10.1093/nar/gki025.
25. Roux S, Páez-Espino D, Chen IMA, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC. 2021. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res 49:D764–D775. https://doi.org/10.1093/nar/gkaa946.
26. Unterer M, Khan Mirzaei M, Deng L. 2021. Gut Phage Database: phage mining in the cave of wonders. Signal Transduct Target Ther 6:193. https://doi.org/10.1038/s41392-021-00615-2.
27. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD.

2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39:105–114. https://doi.org/10.1038/s41587-020-0603-3.

28. Pawluk A, Davidson AR, Maxwell KL. 2018. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol 16:12–17. https://doi.org/10.1038/nrmicro.2017.120.
29. Pinilla-Redondo R, Shehreen S, Marino ND, Fagerlund RD, Brown CM, Sorensen SJ, Fineran PC, Bondy-Denomy J. 2020. Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic elements. Nat Commun 11:5652. https://doi.org/10.1038/s41467-020-19415-3.
30. Marino ND, Zhang JY, Borges AL, Sousa AA, Leon LM, Rauch BJ, Walton RT, Berry JD, Joung JK, Kleinstiver BP, Bondy-Denomy J. 2018. Discovery of widespread type I and type V CRISPR-Cas inhibitors. Science 362:240–242. https://doi.org/10.1126/science.aau5174.
31. Song G, Zhang F, Tian C, Gao X, Zhu X, Fan D, Tian Y. 2022. Discovery of potent and versatile CRISPR-Cas9 inhibitors engineered for chemically controllable genome editing. Nucleic Acids Res 50:2836–2853. https://doi.org/10.1093/nar/gkac099.
32. Pawluk A, Staals RHJ, Taylor C, Watson BNJ, Saha S, Fineran PC, Maxwell KL, Davidson AR. 2016. Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. Nat Microbiol 1:16085. https://doi.org/10.1038/nmicrobiol.2016.85.
33. Pawluk A, Amrani N, Zhang Y, Garcia B, Hidalgo-Reyes Y, Lee J, Edraki A, Shah M, Sontheimer EJ, Maxwell KL, Davidson AR. 2016. Naturally occurring off-switches for CRISPR-Cas9. Cell 167:1829–1838.e9. https://doi.org/10.1016/j.cell.2016.11.017.
34. Lee J, Mir A, Edraki A, Garcia B, Amrani N, Lou HE, Gainetdinov I, Pawluk A, Ibraheim R, Gao XD, Liu P, Davidson AR, Maxwell KL, Sontheimer EJ. 2018. Potent Cas9 inhibition in bacterial and human cells by AcrIIC4 and AcrIIC5 anti-CRISPR proteins. mBio 9:e02321-18. https://doi.org/10.1128/mBio.02321-18.
35. Lin P, Qin S, Pu Q, Wang Z, Wu Q, Gao P, Schettler J, Guo K, Li R, Li G, Huang C, Wei Y, Gao GF, Jiang J, Wu M. 2020. CRISPR-Cas13 inhibitors block RNA editing in bacteria and mammalian cells. Mol Cell 78:850–861.e5. https://doi.org/10.1016/j.molcel.2020.03.033.
36. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
37. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
38. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. Cell 184:1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.
39. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
40. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
41. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
42. Nakamura M, Srinivasan P, Chavez M, Carter MA, Dominguez AA, La Russa M, Lau MB, Abbott TR, Xu X, Zhao D, Gao Y, Kipniss NH, Smolke CD, Bondy-Denomy J, Qi LS. 2019. Anti-CRISPR-mediated control of gene editing and synthetic circuits in eukaryotic cells. Nat Commun 10:194. https://doi.org/10.1038/s41467-018-08158-x.
43. Usher B, Birkholz N, Beck IN, Fagerlund RD, Jackson SA, Fineran PC, Blower TR. 2021. Crystal structure of the anti-CRISPR repressor Aca2. J Struct Biol 213:107752. https://doi.org/10.1016/j.jsb.2021.107752.
44. UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100.
45. Bondy-Denomy J, Davidson AR, Doudna JA, Fineran PC, Maxwell KL, Moineau S, Peng X, Sontheimer EJ, Wiedenheft B. 2018. A unified resource for tracking anti-CRISPR names. Crispr J 1:304–305. https://doi.org/10.1089/crispr.2018.0043.
46. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113.
47. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
48. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37. https://doi.org/10.1093/nar/gkr367.
49. Stanley SY, Maxwell KL. 2018. Phage-encoded anti-CRISPR defenses. Annu Rev Genet 52:445–464. https://doi.org/10.1146/annurev-genet-120417-031321.
50. Borges AL, Davidson AR, Bondy-Denomy J. 2017. The discovery, mechanisms, and evolutionary impact of anti-CRISPRs. Annu Rev Virol 4:37–59. https://doi.org/10.1146/annurev-virology-101416-041616.
51. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.
52. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245. https://doi.org/10.1093/nar/gkw290.