RNA Biology 10:5, 817–827; May 2013; © 2013 Landes Bioscience

CRISPRTarget
Bioinformatic prediction and analysis of crRNA targets

Ambarish Biswas,[1] Joshua N. Gagnon,[1] Stan J.J. Brouns,[2] Peter C. Fineran[2,3,4,*] and Chris M. Brown[1,4,*]

[1]Department of Biochemistry; University of Otago; Dunedin, New Zealand; [2]Laboratory of Microbiology; Wageningen University; Wageningen, Netherlands; [3]Department of Microbiology and Immunology; University of Otago; Dunedin, New Zealand; [4]Genetics Otago; University of Otago; New Zealand

Keywords: phage resistance, horizontal gene transfer, Cas, CRISPR, small RNA targets, crRNA, bioinformatics, R-loop

Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeats; Cas, CRISPR associated; PAM, protospacer adjacent motif; crRNA, CRISPR RNA

The bacterial and archaeal CRISPR/Cas adaptive immune system targets specific protospacer nucleotide sequences in invading organisms. This requires base pairing between processed CRISPR RNA and the target protospacer. For type I and II CRISPR/Cas systems, protospacer adjacent motifs (PAM) are essential for target recognition, and for type III, mismatches in the flanking sequences are important in the antiviral response. In this study, we examine the properties of each class of CRISPR. We use this information to provide a tool (CRISPRTarget) that predicts the most likely targets of CRISPR RNAs (http://bioanalysis.otago.ac.nz/CRISPRTarget). This can be used to discover targets in newly sequenced genomic or metagenomic data. To test its utility, we discover features and targets of well-characterized *Streptococcus thermophilus* and *Sulfolobus solfataricus* type II and III CRISPR/Cas systems. Finally, in *Pectobacterium* species, we identify new CRISPR targets and propose a model of temperate phage exposure and subsequent inhibition by the type I CRISPR/Cas systems.

Introduction

The CRISPR (clustered regularly interspaced short palindromic repeats)/Cas (CRISPR associated) system has evolved to defend microorganisms against foreign invading nucleic acids, principally DNA from bacteriophages (phages), plasmids and other mobile elements (reviewed in refs. 1–5). CRISPR/Cas systems have been identified in 47% and 86% of complete bacterial and archaeal genomes.[6] Resistance development occurs when a short sequence is acquired from the phage or plasmid genome and added, as a new spacer, to the CRISPR arrays (reviewed in ref. 7), which consist of short repeats separated by spacers. In CRISPR systems, a CRISPR RNA (crRNA) containing a "spacer" (or guide[8]) is generated from a longer precursor (pre-crRNA)[8-13] and incorporated into a ribonucleoprotein complex of one or more Cas proteins.[9,14-21] These ribonucleoprotein complexes bind to, and trigger, the destruction of complementary DNA or RNA from invading elements.[20,22,23]

Typically, organisms have several CRISPR arrays containing a range of spacers with different sequences derived from previous exposure to phages and plasmids. The largest predicted bacterial array, from *Haliangium ochraceum* DSM 14365, has 587 spacers, only two of which are identical.[6] Despite experimental proof that CRISPR/Cas systems target phages or plasmids,[22,24-26] the targets of most spacers have not been identified. For example, of 926 spacers identified for *E. coli* and *Salmonella,* Touchon and Rocha were only able to predict the likely targets of 8%;[27] similarly, a parallel study discovered the targets of 12% of spacers.[28]

There are many contributing reasons for the lack of identified crRNA targets. This is partly due to the relative paucity of studies that investigate the sequences of phages when compared with their abundance and genetic diversity.[29,30] Furthermore, many phage sequences are not easily accessible in databases such as GenBank, but many more exist in viral metagenome or virome studies.[31] Large proportions of phage sequences have no similarity to any known phage or other sequences. Therefore, most metagenomic data remains unannotated.[29,30] For example, in a recent study, the metagenomes of phages purified from thermal ocean vents were sequenced.[32] The method targeted lambdoid viruses and resulted in the sequencing of a new lambdoid virus; however, 45–55% of sequences had no database matches.[32] Another study of marine viromes identified only 10% of genes related to known phages.[33] The lack of identified crRNA targets in plasmids results from a similar dearth of sequence data relative to the their abundance and diversity. Like phages, plasmids are mobile and have mosaic sequence structures and are rapidly evolving.[34] Recent efforts have begun to sequence populations of plasmids using metagenomics, which should start to improve this plasmid data shortage.[35]

CRISPR/Cas systems are divided into three major types (I-III), and further into subtypes (e.g., types III-A and III-B).[36] Different types share similarities, yet can have differences, such

*Correspondence to: Peter C. Fineran, Email: peter.fineran@otago.ac.nz; Chris M. Brown, Email: chris.brown@otago.ac.nz
Submitted: 12/12/12; Revised: 02/15/13; Accepted: 02/19/13
http://dx.doi.org/10.4161/rna.24046

as in crRNA generation or the nature of the target (RNA or DNA). Recent studies have begun to elucidate the process of recognition of target protospacers in the major types of CRISPR/Cas systems. From early studies, it was interpreted that exact pairing along the length of the spacer RNA was required,[24,37] but recent results indicate that some mismatches are tolerated, at least for some systems.[25,38-40] For type I and II systems, protospacer adjacent motifs (PAM) are required for recognition,[23,25,37,39,41,42] and a short seed sequence within the match is required in particular subtypes.[15,20,39] For type III systems, it is unclear if a seed sequence exists and no PAMs have been identified.[41] Instead, the base-pairing potential between the 5' repeat-derived portion of the crRNA (termed the 5' handle) and the sequence flanking the protospacer target is important to enable interference, and disallow self-targeting for type III-A systems.[43]

CRISPR/Cas systems have similarity, yet differences, to RNAi in mammals, which can also provide protection from viral infection.[5,44] RNAi utilizes miRNAs of ~20–22 bases that recognize specific mRNA targets.[45] However, the key seed determinant is only six to eight bases. Therefore, predictive tools to discover functional binding sites have been developed that use the properties of known sites to predict new ones.[45,46] The critical factor is distinguishing true sites from false positives, and there are a large number of algorithms implemented for miRNA target discovery.[47-49] A number of bioinformatic tools are available for the identification of CRISPR arrays and their spacer sequences.[50-52] In contrast, few approaches have been developed to discover the targets of CRISPR.[52-54] For predicted spacers in CRISPRdb,[6] (November 2012) the mean length is 36 bases (range: 16–100), which is consistent with typical lengths for experimentally confirmed spacers of 24–37 bases. The input used when searching for targets using CRISPRFinder is a small number of these spacer sequences without adjacent repeats.[52] These spacers are used for a BLAST search of the nucleotide sequence database from GenBank using the default parameters.[55,56] The discovery of new spacer targets would be facilitated by tools that allow flexibility in these areas and enable searches of recent metagenomic data sets. Furthermore, the ability to score or visualize PAMs, basepairing in flanking sequences and define seed regions are not available in existing tools. These properties assist in biological interpretation of putative targets. In this study, we have incorporated known features of the CRISPR/Cas system into a target discovery tool. We have also allowed flexibility to enable the incorporation of new features, to generate testable hypotheses as to the targets of CRISPR systems.

Results

CRISPRTarget: Development of a tool for discovery of crRNA targets. The lack of tools for prediction of the protospacer targets of crRNAs led us to develop a web application called CRISPRTarget. We have summarized the current state of knowledge about the three major CRISPR/Cas types, their PAMs, handles and seed regions (**Table 1**) and used this information when developing CRISPRTarget. Users can provide their input as either spacers in FASTA format, or as CRISPRFinder,[52] PILER-CR[51] or CRISPR Recognition Tool (CRT)[50] output files (following CRISPR prediction via one of these methods). Putative protospacer targets can be identified, following a BLASTn search of the spacer input against a number of databases or user-uploaded sequences. These databases include ACLAME genes, GenBank-nt, GenBank-Environmental, GenBank-Phage, RefSeq-Microbial, RefSeq-Plasmid, RefSeq-Viral and parts of CAMERA. Although default setting allows the sensitive detection of potential targets, users have the ability to modify the search parameters, such as E-value, word size and penalties for gaps and match/mismatch. This flexibility enables the stringency of targets to be adjusted, depending on the user requirements. Either on the initial input screen, or following the BLAST search, targets can be displayed and scored for flanking sequences, PAMs and filtered by exact matching seed regions. These are important parameters when considering biological details about the predicted target, such as what type of system/CRISPR-type is involved. This information in **Table 1** can assist users in choosing the appropriate parameters for their particular target search. The output provided is either visual in HTML format, but can also be saved as text and opened in a spreadsheet. The target sequence is typically displayed as an R-loop, depicting a specified part of the crRNA, as well as both the target and non-target strand of the double-stranded target DNA. The target sequence R-loop can be fully reverse complemented, when users suspect that the direction of transcription of the CRISPR array starts from the downstream end instead.

A general model of the match between a spacer and protospacer target as the output from CRISRTarget is shown in **Figure 1** for types I, II and III. The differing features, such as 5' or 3' handles and the presence or absence of PAMs, can be specified, searched, sorted and displayed in CRISPRTarget. Furthermore, the parameters can be manually adjusted to incorporate new functional information (e.g., a new PAM). For clarity, we use the definition of the protospacer as the DNA strand complementary to the crRNA, and PAMs are denoted 5'-3' on the protospacer DNA (e.g., type I-E PAM is CTT, **Table 1**).[4] In addition, we refer to the flanking sequences as being 5' or 3' of this protospacer and handles as 5' or 3' of the crRNA spacer. CRISPRTarget enables detection of the most likely complements of spacers in target sequences (**Fig. 2**).

Proof of principle: Phage protospacers for *Streptococcus thermophilus* type II CRISPR/Cas. As an initial test, we used the well-characterized type II CRISPR1 array from *Streptococcus thermophilus* DGCC7710. This strain is economically important in the dairy industry and has active CRISPR/Cas systems.[24,37] The sequences of arrays with recently acquired spacers are available WTphi858phi2972+S9S10S11S12 (GenBank accession: EF434477) and, WTphi858phi2972+S13S14 (EF434478), as are many *Streptococcus thermophilus* phage sequences (114 sequences of 6,800 in the phage division of GenBank). These two strains have become resistant to ϕ858 and ϕ2972, whereas the WT strain is sensitive (EF434469).[24] We expect that spacers from the resistant strains will be predicted to target ϕ858 and ϕ2972, whereas the WT will not, but might target other mobile elements. Spacers were predicted from these CRISPR sequences

Table 1. Summary of features of CRISPR/Cas systems, including PAMs, repeats, seed regions and handles*

Type	Target	Representative species	PAM (5'-3')[§]	Typical repeat	CRISPR family[59]	Seed region	5'/3' handles (nt)
		Type I (PAMs 3' of protospacer)				Seed adjacent to PAM	
I-A	DNA	*Sulfolobus solfataricus* P2	Protospacer-NGG[40,41,78]	GATAATCTCTTATAGAATTGA AAG[¶]	CRISPR-7	Unknown	8/16–17[16]
I-B	DNA	*Clostridium thermocellum* ATCC 27405	Unknown	GTTTTTATCGTACCTATGAGG AATTGAAAC[¶]	CRISPR-6	Unknown	8/4, 10–12[79]
		C. thermocellum ATCC 27405	Unknown	GTTGAAGTGGTACTTAGT AAAACAAGGATTGAAAC[¶]	CRISPR-9		8/2–6[79]
		Haloferax volcanii H26	Protospacer-GAA, AGT, TTA, ATA, CTA, GTG[80]	GTTTCAGACGAACCCTTG TGGGDTTGAAGC[¶]	CRISPR-6†		
		Listeria monocytogenes	Protospacer-NGG[41]	GTTTTAACT ACTTATTATGAAATCTAAAT	CRISPR-1		
I-C	?	*Xanthomonas oryzae*	Protospacer-GAA[41,81]	GTC GCG TCC TCA CGG GCG CGT GGA TTG AAA C[¶]	CRISPR-3	Unknown	
		Bacillus halodurans	Protospacer-GAA[41]	GTC GCA CTC TTC ATG GGT GCG TGG ATT GAA AT	CRISPR-3		11/21[11]
I-D	?	Unknown	Unknown		Unknown	Unknown	
I-E	DNA	*Escherichia coli* K12	Protospacer-CTT, CAT, CCT, CTC[41,73,74,82]	GWG TTC CCC GCG CCA GCG GGG ATA AAC CG[¶]	CRISPR-2	1–5, 7–8[39]	8/2[19,17]
I-F	DNA	*Pseudomonas aeruginosa* PA14	Protospacer-GG[25,41]	GTTCACTGCCGTGTAGGC AGCTAAGAAA[¶]	CRISPR-4	1–8[15]	8/20[10,15]
		Pectobacterium atrosepticum SCRI1043	Protospacer-GG[41]	GTTCACTGCCGTACAGGC AGCTTAGAAA[¶]	CRISPR-4		
		Type II (PAMs 5' of protospacer)				Seed adjacent to PAM	
II-A	DNA	*Streptococcus thermophilus*	WTTCTNN-protospacer[58]	GTTTTTGTACTCTCAAGAT TTAAGTAACTGTACAAC	CRISPR-10	Unknown	
		Streptococcus thermophilus	TTTYRNNN-protospacer[83]	GTTTTTGTACTCTCAAGA TTTAAGTAACTGTACAAC	CRISPR-10		
II-B	DNA	*Streptococcus thermophilus*	CNCCN-protospacer[58,84]	GTTTTAGAGCTGTGTTGT TTCGAATGGTTCCAAAAC	CRISPR-10		
		Streptococcus pyogenes	CCN-protospacer[20,41]	GTTTTAGAGCTATGCTGT TTTGAATGGTCCCAAAAC[¶]	CRISPR-10	13[20]	None/19–22[13]
		Type III (no PAM)					
III-A	DNA	*Staphylococcus epidermidis*	No PAM[43]	GATCGATACCCACCCCGA AGAAAAGGGGACGAGAAC[¶]	CRISPR-8	Unknown	8/(37/43 entire length)[43,76]
III-B	RNA	*Pyrococcus furiosus*	No PAM[85]	GTTCCAATAAGACTAAAA TAGAATTGAAAG[¶]	CRISPR-6	Unknown	8/(39/45 entire length)[18,85]
		Sulfolobus solfataricus	No PAM[14]	GATTAATCCCAAAAGGAA TTGAAAG[¶]	CRISPR-7	Unknown	8/uncertain[14]

*Adapted from Westra et al., 2012. [§] PAM and protospacer are denoted as sequence on target strand that base pairs with crRNA. [¶] Direction of CRISPR transcription confirmed. †Described as CRISPR-9,[80] but length and sequence suggests degenerate CRISPR-6.

using CRISPRFinder,[52] CRT[50] and PILER-CR.[51] CRISPRFinder is the most cited CRISPR prediction tool; however, a combination of CRT and PILER-CR are used in the DOE-JGI standard pipeline for bacterial genome annotation.[57]

CRT and CRISPRFinder predicted the published array of 32 spacers in the WT and an additional two or four spacers in the resistant strains, whereas PILER-CR with default parameters split the array into two consisting of 22 and three spacers. The CRT predictions were used as input (**Fig. 3**), as these include information about small variations in the repeats (all inputs in this study are provided in **Fig. S1**). These spacers were searched against the phage division of GenBank and plasmid division of RefSeq. With the default settings, there were matches from 24 of 32 spacers to 84 sequences of mobile elements in the initial output, of which 81 were *Streptococcus* spp phages (supplemental file html output in **Fig. S1**, text in **Fig. S3**). This has been designated a type II-A system with a requirement for a PAM 5' of the protospacer (5'-WTTCTNN-protospacer-3');[58] 38/84 had the consensus PAM. The additional spacers in strains WTphi858phi2972+S9S10S11S12

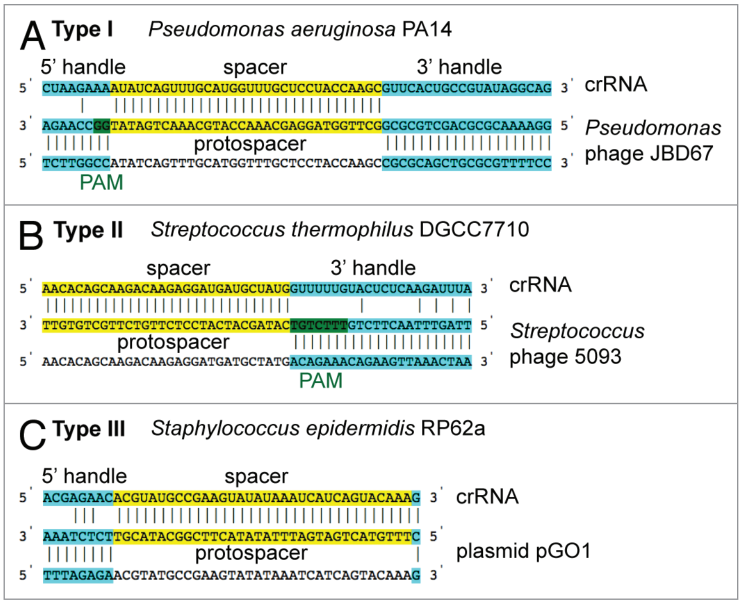

A
Type I
Pseudomonas aeruginosa PA14
5’ handle
spacer
3’ handle
5' CUAAGAAAAUAUCAGUUUGCAUGGUUUGCUCCUACCAAGCGUUCACUGCCGUAUAGGCAG 3'
crRNA
3' AGAACCGGTATAGTCAAACGTACCAAACGAGGATGGTTCGGCGCGTCGACGCGCAAAAGG 5'
protospacer
5' TCTTGGCCATATCAGTTTGCATGGTTTGCTCCTACCAAGCCGCGCAGCTGCGCGTTTTCC 3'
Pseudomonas phage JBD67
PAM
B
Type II
Streptococcus thermophilus DGCC7710
spacer
3’ handle
5' AACACAGCAAGACAAGAGGAUGAUGCUAUGGUUUUUGUACUCUCAAGAUUUA 3'
crRNA
3' TTGTGTCGTTCTGTTCTCCTACTACGATACTGTCTTTGTCTTCAATTTGATT 5'
protospacer
5' AACACAGCAAGACAAGAGGATGATGCTATGACAGAAACAGAAGTTAAACTAA 3'
Streptococcus phage 5093
PAM
C
Type III
Staphylococcus epidermidis RP62a
5’ handle
spacer
5' ACGAGAACACGUAUGCCGAAGUAUAUAAAUCAUCAGUACAAAG 3'
crRNA
3' AAATCTCTTGCATACGGCTTCATATATTTAGTAGTCATGTTTC 5'
protospacer
5' TTTAGAGAACGTATGCCGAAGTATATAAATCATCAGTACAAAG 3'
plasmid pGO1

Figure 1. Example annotated CRISPRTarget outputs of representatives of type I, II and III CRISPR/Cas systems. The protospacer is the DNA target complementary to the crRNA spacer. The crRNA is displayed as RNA 5' to 3' and the base paired protospacer is 3' to 5'. (**A**) The predicted spacer 6 crRNA from the type I-F CRISPR1 (CRISPR1_6) in *P. aeruginosa* PA14 targets *Pseudomonas* phage JBD67.[25] The output visualizes the 5'-protospacer-GG-3' PAM[41] and the crRNA with 8 and 20 nt 5' and 3' handles, respectively.[10] (**B**) The CRISPR1_15 from the type II system from *Streptococcus thermophilus* DGCC7710 WTphi858phi2972+S13S14[24] matched to *Streptococcus* phage 5093. The output shows the predicted length of the 3' handle, based on *Streptococcus pyogenes*,[13] and the 5'-WTTCTNN-protospacer-3' PAM.[58] (**C**) Spacer 1 from the type III-A system from *Staphylococcus epidermidis* RP62a targeting plasmid pGO1.[26] The output was adjusted to display the 8 nt 5' handle with an entire mature crRNA length of 43 nt and no PAMs were scored.[76] Yellow sequences include spacer and protospacer, blue indicates flanking sequences and PAMs are shown in green.

and WTphi858phi2972+S13S14 targeted ϕ2972 and ϕ858 as expected.[24] Interestingly, the WT has a spacer (CRISPR1_14, uniquely identified as EF434469_1_14 in the text output **Fig. S3**) with just one mismatch (protospacer +7) to bases 31869–31897 of ϕ2972. Additionally, the 5' region of the target differs by one base from the PAM consensus (WTcCTNN) (**Fig. 4**; **Fig. S2**). Experimentally, this strain is sensitive to ϕ2972,[24] so the system appears to have a functional requirement for the conserved consensus PAM and/or an exact match near the 5' end of the protospacer, which corresponds to the 13 nt seed region in type II systems (**Table 1**).[20] In summary, CRISPRTarget can accurately identify protospacers for crRNAs and display these with details of match/mismatch and PAMs.

Identification of targets for the RNA-targeting *Sulfolobus solfataricus* type III CRISPR/Cas system. The *S. solfataricus* P2 CRISPR/Cas system has been well characterized and, recently, the structure of the type III-B ribonucleoprotein Cmr complex was published.[14] This study also demonstrated that crRNAs derived from all six CRISPR arrays are detected in the Cmr complex, which targets RNAs complementary to the crRNA spacer sequences. CRISPRdb lists a total of 255 spacers from seven detected arrays, which belong to the CRISPR-7 (and possibly CRISPR-11)[59] families. Putative protospacers were discovered using CRISPRTarget with the default settings and all predicted *S. solfatricus* CRISPRs as input (**Fig. S1**). Of the 254 unique spacers used, 517 hits were detected for 57 spacers from five of the seven arrays (**Fig. S4**; 471 hits when E-value lowered to 0.1). An earlier study identified the targets of 29 spacers.[60] The top hit was a perfect match from spacer 28 in locus A[14] (NC_002754_3_28 in output) to an *Acidianus* two-tailed virus (AJ888457). The majority of top hits are to *Sulfolobus*, *Stygiolobus* and *Acidanus* viral sequences, but there are examples of plasmid matches (e.g., *Sulfolobus* pNOB8). One spacer in locus B[14] (spacer 23 from leader end; NC_002754_4_73 in output) accounts for 393 hits, due to a very A-rich sequence. Since for Cmr no PAM has been identified and self-DNA cannot be targeted, as this system targets RNA, penalizing flanking matches or searching for PAMs was not required. If analyzing type III-A, rather than B, systems, mismatches between the 5' crRNA handle and the 3' flank of the protospacer DNA are important for interference[43] and can be scored appropriately. However, in either case, the ability to view the pairing between the handle and protospacer flanks allows matches to different CRISPR arrays to be easily distinguished.

The *P. atrosepticum* type I-F system targets a prophage in *Pectobacterium carotovorum*. Members of the genus *Pectobacterium* are economically important phytopathogens that cause a range of plant diseases.[61] CRISPR/Cas systems in plant pathogens have not been well examined to date (reviewed in ref. 62). Previously, we analyzed the type I-F system of *P. atrosepticum* SCRI1043 (previously known as *Erwinia carotovora* subsp *atrosepticum*),[12,63] which causes soft-rot and blackleg disease in potato.[64] The *cas* genes and CRISPRs are transcribed and crRNAs generated by the Cas6f endoribonuclease.[12] Furthermore, the *P. atrosepticum* Csy1, Csy2, Csy3 and Cas6f proteins form a complex, which interacts with the Cas2-Cas3 nuclease.[63] Cas1 and the Cas2-Cas3 hybrid protein also interact, suggesting a role in acquisition.[7,63] The *P. atrosepticum* SCRI1043 type I-F system contains three CRISPR arrays with a consensus repeat belonging to CRISPR-4 type (**Table 1**).[59] These arrays contain 41 spacers with 28, 10 and three spacers present in CRISPR1, 2 and 3, respectively (**Table 2**). Our previous analyses using BLAST failed to identify potential viral targets of the 41 spacers. However, spacer 6 in CRISPR2 showed 100% identity to the *eca0560* gene in its own genome.[12] To test CRISPRTarget, we searched for potential targets of all spacers. CRISPRFinder output files for each array were searched against ACLAME, GenBank-Environmental, GenBank-Phage, RefSeq-Microbial, RefSeq-Plasmid, RefSeq-Viral and a subset of the CAMERA metagenomic databases in CRISPRTarget (default settings, but -1/1 match/mismatch scores to penalize self matches with the 8 nt handles).

The CRISPR1 array identified by CRISPRFinder was in the incorrect orientation, so CRISPRTarget was adjusted for a reverse complemented output (e.g., see **Fig. 4**). CRISPRTarget gave 67 hits from 13/28 spacers from CRISPR1 (**Fig. S4**), compared with only two hits when CRISPRFinder was utilized. Selection of the I-F PAM in CRISPRTarget enabled visualization and scoring of targets that contained a consensus CRISPR-4/I-F PAM.[41] Furthermore, the site of crRNA processing by Cas6F in type I-F systems is known,[10] so 8 nt of the 5' (handle) and 20 nt of the 3' flanking regions were displayed for the crRNAs. By scoring flanks with penalties (e.g., -1/1 match/mismatch), self-targets can be penalized and moved down the output list. Usually the default cut-off score of 20 eliminates the self-matching results when default 8 nt handles are used (with -1/1 match/mismatch scores), while allowing bona fide targets. Using the same databases and increasing the E-value to 10, increased the number of hits to 406, which resulted in the identification of putative targets for 19 of the 28 spacers. A search with CRISPR1 against the GenBank-nt database with the same settings identified 21 hits for eight spacers when an E-value of 1 was used. When the E-value was increased to 10, 24 spacers gave 85 hits scoring 20 or more, but there were some false positive sequences (eukaryotic).

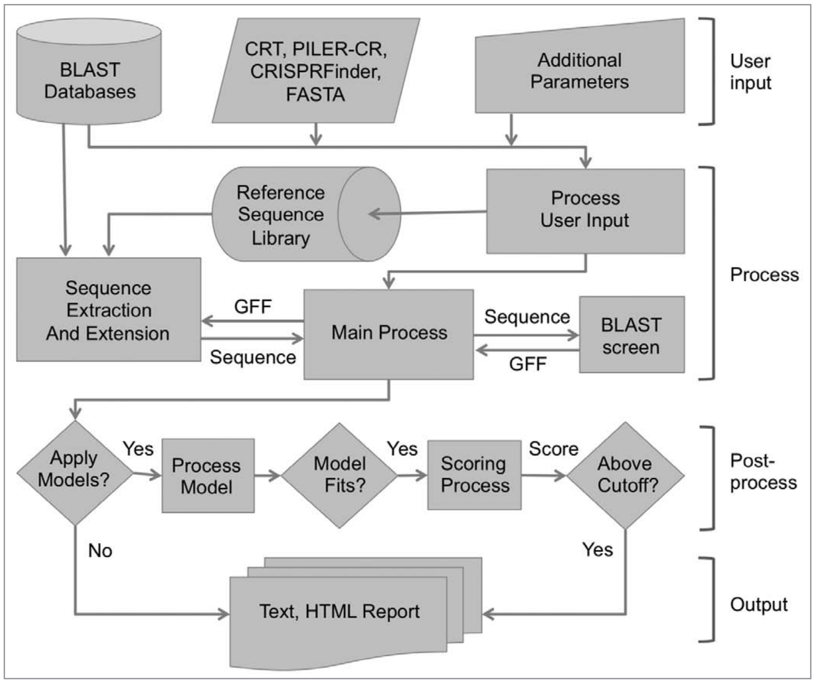


Figure 2. Flowchart of the steps in CRISPRTarget (details are in the Materials and Methods). Input is predictions of the CRISPR arrays, selected databases and initial parameters. This input is processed and the spacers screened using BLASTn for matches against the databases. The flanks of these matches are extended and PAMs and handles analyzed in an interactive manner. Output is as a text/spreadsheet format, or as a graphical display (HTML).

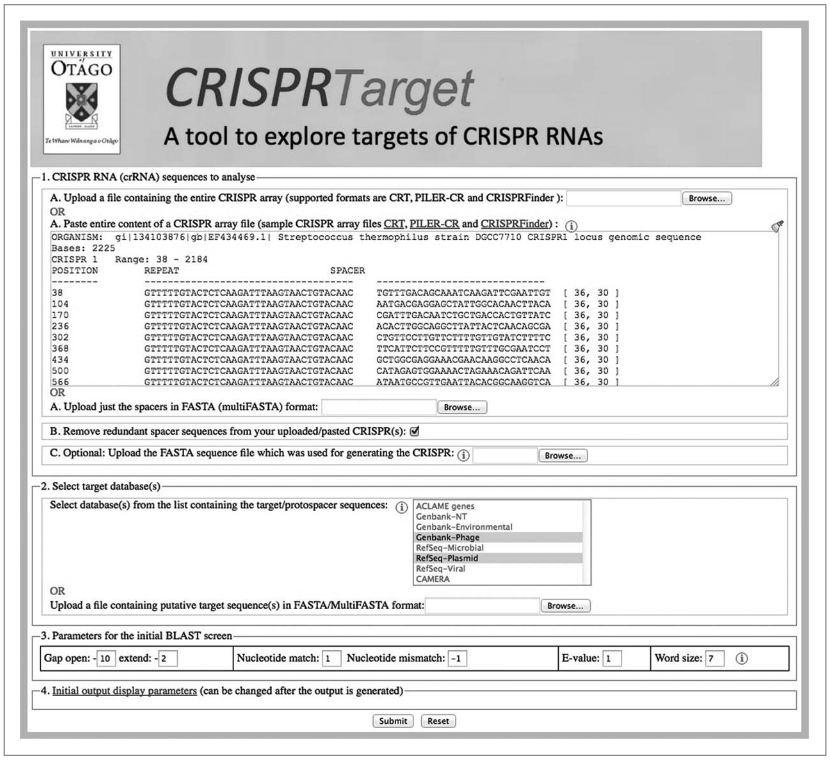
CRISPR1_19 matched a putative phage gene in *Pectobacterium carotovorum* subsp *carotovorum* PCC21. Note that we denote spacer 1 as the leader-proximal spacer, but the spacer numbers in the CRISPRTarget output are numbered according to the input file. For example, since CRISPR1 was reversed in the output, spacer 19 of 28 (relative to the leader) is numbered spacer 10 in the CRISPRFinder input file. Comparing *P. carotovorum* PCC21[65] and *P. atrosepticum* SCRI1043[64,66] revealed that the spacer 19 target is within a 45 kb prophage containing 54 predicted coding sequences (here designated ΦPCC21_1; **Fig. 5A and B**). ΦPCC21_1 is inserted in *ryeAB*, but is absent in *P. atrosepticum* SCRI1043. The *ryeAB* genes are two overlapping small non-coding RNAs. In *Salmonella*, this locus is an important insertion site for prophages that have influenced this pathogen's evolution.[67] Interestingly, CRISPR1 spacer 2 also matched ΦPCC21_1, albeit ~32 kb from the spacer 19 target (**Fig. 5A**). Mismatches in the predicted RNA-DNA hybrid suggest that these spacers might no longer target this particular prophage, but it is also possible that they derived from a related phage. We propose that *P. atrosepticum* has been exposed to this, or a related, phage in the past, but lysogenization has been inhibited by CRISPR/Cas.

The remaining spacers had matches to a variety of phage, prophage, microbial genome and metagenome samples (**Fig. S4**). For example, a protospacer target for spacer 11 was identified in *Salmonella enterica* epsilon 15 serotype-converting phage.[68]

***Pectobacterium carotovorum* crRNAs match prophages in *P. atrosepticum* and *P. carotovorum*.** As *P. atrosepticum* spacers matched a prophage in a related strain, we examined CRISPR targets in other representative *Pectobacterium* genomes. First, we uploaded the genome of *P. carotovorum* subsp *carotovorum* PCC21[65] into CRISPRFinder and identified five arrays; three CRISPR-4/type I-F arrays containing 38, 3 and 3 spacers and two CRISPR-2/type I-E arrays with 14 spacers each (output in **Fig. S4**).

Two spacers in CRISPR1 (type I-F with 38 spacers) matched different regions of *eca2627* in the *P. atrosepticum* SCRI1043 ΦECA29 prophage[69] (also termed HAI9;[64] **Fig. 5C**). Comparison of *P. carotovorum* subsp *carotovorum* PCC21 and *P. atrosepticum* SCRI1043 demonstrated the absence of a ΦECA29 prophage in PCC21 (**Fig. 5D**). Spacer 34 also matched a putative prophage (here designated ΦPC1_1) in *P. carotovorum* subsp *carotovorum* PC1 (**Fig. 5E and F**). The two type I-E arrays are separated by 76 bp, so it is possible that these are one large array with 29 spacers. Spacer 8 within CRISPR4 was self-matching to its own ΦPCC21_1 prophage, but this will be non-targeting due to a position 2 seed mutation.[39] Spacer 3 in CRISPR4 matches a transposase gene in *Pectobacterium wasabiae* WPP163 (Pecwa_0911), which is not predicted to be part of an island.[70]

***P. wasabiae* CRISPRs have targets against multiple prophages.** Next, the CRISPRs of *P. wasabiae* WPP163 were analyzed (**Fig. S4**). *P. wasabiae* has four CRISPRs, two CRISPR-4/type I-F with 17 and 25 spacers and two CRISPR-2/type I-E containing 16 and six spacers (**Table 2**). Spacers 2 and 10 from CRISPR1 (I-F array with 17 spacers) match ϕPCC21_1 (**Fig. 5G and H**), which is also targeted by the *P. atrosepticum* type I-F system (**Fig. 5A and B**). ΦPCC21_1 is absent in *P. wasabiae*, but



CRISPRTarget

A tool to explore targets of CRISPR RNAs

1. CRISPR RNA (crRNA) sequences to analyse

A. Upload a file containing the entire CRISPR array (supported formats are CRT, PILER-CR and CRISPRFinder):

OR

A. Paste entire content of a CRISPR array file (sample CRISPR array files CRT, PILER-CR and CRISPRFinder) :

```
ORGANISM:  gi|134103876|gb|EF434469.1| Streptococcus thermophilus strain DGCC7710 CRISPR1 locus genomic sequence
Bases: 2225
CRISPR 1   Range: 38 - 2184
POSITION        REPEAT                              SPACER
--------        ------------------------------------    ------------------------------
38              GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    TGTTTGACAGCAAATCAAGATTCGAATTGT  [ 36, 30 ]
104             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    AATGACGAGGAGCTATTGGCACAACTTACA  [ 36, 30 ]
170             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    CGATTTGACAATCTGCTGACCACTGTTATC  [ 36, 30 ]
236             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    ACACTTGGCAGGCTTATTACTCAACAGCGA  [ 36, 30 ]
302             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    CTGTTCCTTGTTCTTTTGTTGTATCTTTTC  [ 36, 30 ]
368             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    TTCATTCTTCCGTTTTTGTTTGCGAATCCT  [ 36, 30 ]
434             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    GCTGGCGAGGAAACGAACAAGGCCTCAACA  [ 36, 30 ]
500             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    CATAGAGTGGAAAACTAGAAACAGATTCAA  [ 36, 30 ]
566             GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC    ATAATGCCGTTGAATTACACGGCAAGGTCA  [ 36, 30 ]
```

OR

A. Upload just the spacers in FASTA (multiFASTA) format:

B. Remove redundant spacer sequences from your uploaded/pasted CRISPR(s): ☑

C. Optional: Upload the FASTA sequence file which was used for generating the CRISPR:

2. Select target database(s)

Select database(s) from the list containing the target/protospacer sequences: Genbank-Phage, RefSeq-Plasmid

OR

Upload a file containing putative target sequence(s) in FASTA/MultiFASTA format:

3. Parameters for the initial BLAST screen

Gap open: -10 extend: -2 | Nucleotide match: 1 Nucleotide mismatch: -1 | E-value: 1 | Word size: 7

4. Initial output display parameters (can be changed after the output is generated)

Figure 3. CRISPRTarget input. Several formats are accepted. The BLASTn parameters for the initial screen are defined at this step. They default to values that favor a gapless match, but some mismatches. The output may be refined and reordered (**Fig. 4**) after it is obtained.

in this location is Pecwa_2124 (a pseudogene homologous to the ΦPCC21_1 integrase) and Pecwa_2125-9. Remarkably, spacers 3, 4, 5 and 6, from the CRISPR2 (I-F array with 25 spacers), targeted genes PC1_3175, PC1_3187, PC1_3191 and PC1_3182, respectively, in a putative prophage in *P. carotovorum* subsp *carotovorum* PC1 (here designated ϕPC1_2) that is absent in *P. wasabiae* (**Fig. 5I and J**). In addition, spacer 5 matches to the P2-type tail fiber protein H, *eca2608*, in ΦECA29 (**Fig. 5K**) and spacer 20 targeted ΦPC1_1 (**Fig. 5F and L**), which is also absent in *P. wasabiae* (data not shown). Therefore, *P. wasabiae* appears to have previously encountered phages similar to ΦPCC21_1, ΦECA29, ΦPC1_1 and ΦPC1_2, and has developed CRISPR/Cas immunity to these elements.

Overall, this analysis indicated that CRISPRTarget can reveal new targets of spacers in CRISPR arrays and demonstrates, with the example of *Pectobacterium*, that novel biologically relevant information can be obtained. Specifically, inter-species prophage exclusion by *Pectobacterium* type I CRISPR/Cas systems was suggested.

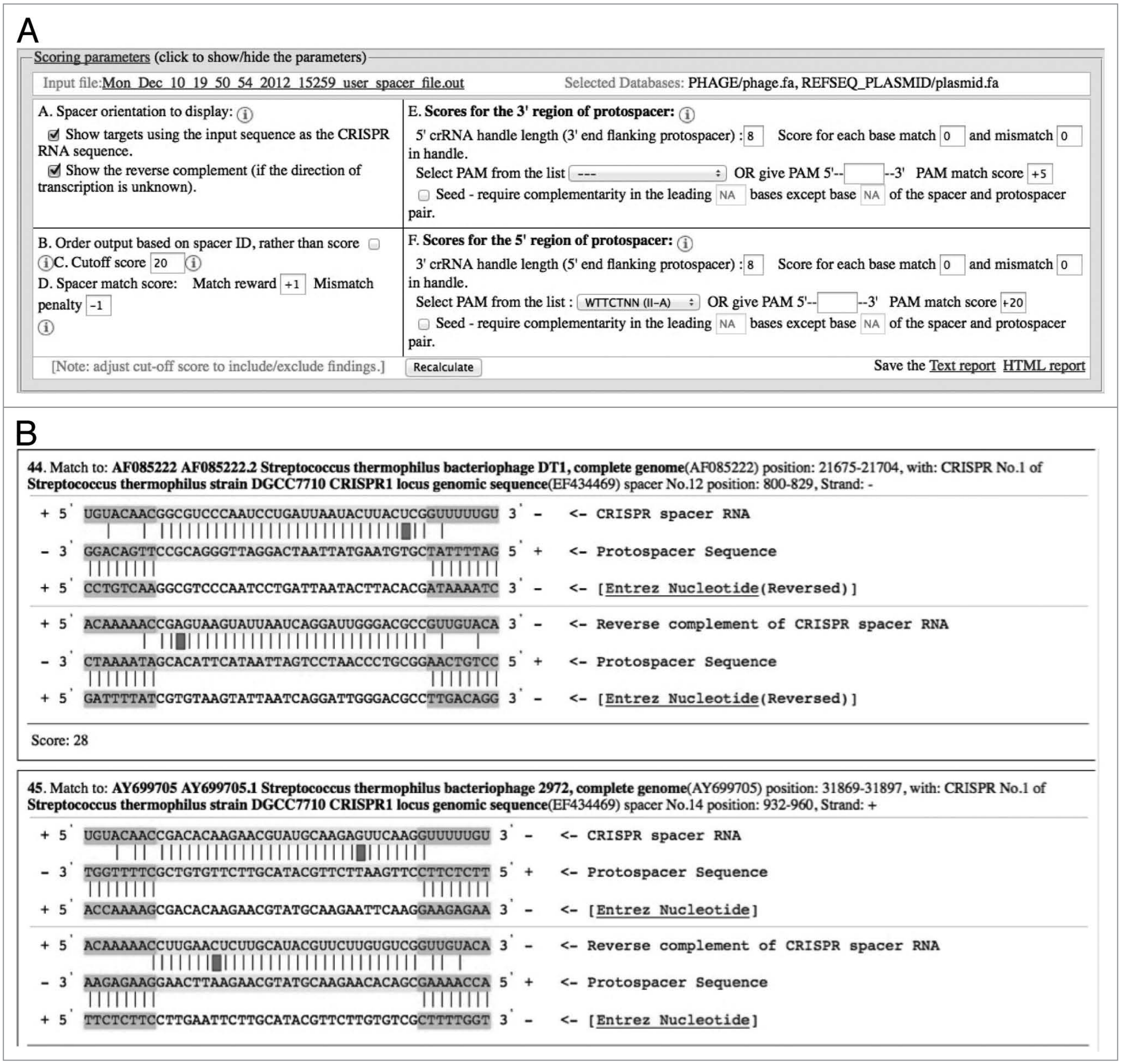
Discussion

We have developed a tool designed to detect, and interactively explore, the targets of CRISPR RNA spacers. This is the first tool of this kind designed for this purpose. The inputs into CRISPRTarget are predicted CRISPR arrays or spacer sequences. These CRISPR and spacer prediction methods were initially developed in 2007–2009[50-52] and, thus, do not incorporate recent refinements. These current CRISPR predictions do not take into account the direction of CRISPR transcription and errors that can occur when defining spacer and repeat boundaries. CRISPRTarget enables the user to search for matches in either or both orientations of a given input and display adjacent PAM and flanking sequences. These features provide the flexibility to discover targets with PAMs and also any adjacent pairing potential, ensuring greater power in predicting biologically relevant protospacer targets.

The initial screen for database matches in CRISPRTarget is done by BLASTn, with a range of parameters able to be defined. The defaults chosen penalize gaps with -10. We know of no publications that indicate that insertions/deletions are permitted in the RNA/DNA hybrid, although in some systems, mismatches are tolerated.[25,38-40] The use of BLASTn allows for a smaller exact hit match of wordsize 7, compared with MegaBLAST (minimum word size of 28). However, BLASTn is slower.[71] Specific databases are provided; the use of databases of mobile elements (e.g., phage, plasmid, ACLAME) reduces the execution time and increases the number of biologically relevant positives. Hits that might have high expect (E) values (e.g., > 1) in larger databases will be shown as significant at the same E-value in a smaller database. Not using the "nt" database as the default also avoids the showing of high-scoring self-matches in the source or related genomes. Selected parts of the CAMERA databases, enriched in phage sequences, are provided,[72] and the user can upload custom data e.g., new genomic or metagenomic data for searching.

Following the initial BLAST screen, the user can interactively refine and reduce the putative targets shown. In some systems, PAMs are required, or seed sequences. These can be weighted so that only those with this feature are displayed. In the case of *S. thermophilus* DGCC7710 WT spacer 14, there is a one base mismatch to ϕ2972 and a T to C substitution in the PAM. The consensus PAM for this *S. thermophilus* type II system is WT**T**TCTNN (or NNAGAAW on the other strand). This T was conserved in experimentally confirmed protospacers. Recent reports have demonstrated that pre-existing spacers that match to a target, but can have subtle mutations that abolish interference, increase the acquisition of new spacers in a process termed priming.[73,74] It is tempting to speculate that this spacer might increase the spacer acquisition activity of this CRISPR array against ϕ2972 and related phages.[24] The ability to detect potential targets for the type III-B system of *S. solfataricus* P2 was also demonstrated and resulted in putative targets for ~20% of the > 250 spacers. Most of these were matches to archaeal viruses and plasmids, demonstrating potentially relevant crRNA targets.

To demonstrate the utility and functionality of CRISPRTarget, we investigated possible protospacer targets in



Scoring parameters (click to show/hide the parameters)

Input file:Mon_Dec_10_19_50_54_2012_15259_user_spacer_file.out   Selected Databases: PHAGE/phage.fa, REFSEQ_PLASMID/plasmid.fa

A. Spacer orientation to display:
☑ Show targets using the input sequence as the CRISPR RNA sequence.
☑ Show the reverse complement (if the direction of transcription is unknown).

E. **Scores for the 3' region of protospacer:**
5' crRNA handle length (3' end flanking protospacer) : 8 Score for each base match 0 and mismatch 0 in handle.
Select PAM from the list --- OR give PAM 5'-- --3' PAM match score +5
□ Seed - require complementarity in the leading NA bases except base NA of the spacer and protospacer pair.

B. Order output based on spacer ID, rather than score □
C. Cutoff score 20
D. Spacer match score: Match reward +1 Mismatch penalty -1

F. **Scores for the 5' region of protospacer:**
3' crRNA handle length (5' end flanking protospacer) : 8 Score for each base match 0 and mismatch 0 in handle.
Select PAM from the list : WTTCTNN (II-A) OR give PAM 5'-- --3' PAM match score +20
□ Seed - require complementarity in the leading NA bases except base NA of the spacer and protospacer pair.

[Note: adjust cut-off score to include/exclude findings.] Save the Text report HTML report

44. Match to: **AF085222 AF085222.2 Streptococcus thermophilus bacteriophage DT1, complete genome**(AF085222) position: 21675-21704, with: CRISPR No.1 of **Streptococcus thermophilus strain DGCC7710 CRISPR1 locus genomic sequence**(EF434469) spacer No.12 position: 800-829, Strand: -

```
+ 5' UGUACAACGGCGUCCCAAUCCUGAUUAAUACUUACUCGGUUUUUGU 3' -  <- CRISPR spacer RNA
- 3' GGACAGTTCCGCAGGGTTAGGACTAATTATGAATGTGCTATTTTAG 5' +  <- Protospacer Sequence
+ 5' CCTGTCAAGGCGTCCCAATCCTGATTAATACTTACACGATAAAATC 3' -  <- [Entrez Nucleotide(Reversed)]

+ 5' ACAAAAACCGAGUAAGUAUUAAUCAGGAUUGGGACGCCGUUGUACA 3' -  <- Reverse complement of CRISPR spacer RNA
- 3' CTAAAATAGCACATTCATAATTAGTCCTAACCCTGCGGAACTGTCC 5' +  <- Protospacer Sequence
+ 5' GATTTTATCGTGTAAGTATTAATCAGGATTGGGACGCCTTGACAGG 3' -  <- [Entrez Nucleotide(Reversed)]
```

Score: 28

45. Match to: **AY699705 AY699705.1 Streptococcus thermophilus bacteriophage 2972, complete genome**(AY699705) position: 31869-31897, with: CRISPR No.1 of **Streptococcus thermophilus strain DGCC7710 CRISPR1 locus genomic sequence**(EF434469) spacer No.14 position: 932-960, Strand: +

```
+ 5' UGUACAACCGACACAAGAACGUAUGCAAGAGUUCAAGGUUUUUGU 3' -  <- CRISPR spacer RNA
- 3' TGGTTTTCGCTGTGTTCTTGCATACGTTCTTAAGTTCCTTCTCTT 5' +  <- Protospacer Sequence
+ 5' ACCAAAAGCGACACAAGAACGTATGCAAGAATTCAAGGAAGAGAA 3' -  <- [Entrez Nucleotide]

+ 5' ACAAAAACCUUGAACUCUUGCAUACGUUCUUGUGUCGGUUGUACA 3' -  <- Reverse complement of CRISPR spacer RNA
- 3' AAGAGAAGGAACTTAAGAACGTATGCAAGAACACAGCGAAAACCA 5' +  <- Protospacer Sequence
+ 5' TTCTCTTCCTTGAATTCTTGCATACGTTCTTGTGTCGCTTTTGGT 3' -  <- [Entrez Nucleotide]
```

Figure 4. Graphical output of CRISPRTarget. The output of a search for the targets of the *Streptomyces thermophilus* DGCC7710 CRISPR array. The direction of transcription is known; however, both strands are shown in diagram, as if the direction of transcription was unknown. Two relatively low-scoring matches using these interactive settings are shown (rank 44–45). They have good spacer-protospacer base pairing but lack a WTTCTNN PAM. Match 45 is to a phage to which this strain is sensitive (Φ2972). Pale gray indicates spacer/protospacer, dark gray shows flanking sequences and mismatches between the crRNA and the target DNA protospacer are indicated as gray boxes.

Pectobacterium species. This analysis revealed that there appears to be a history of prophage exposure and CRISPR content, indicative of an adaptive immunity against prophages. In other words, the presence of CRISPR arrays containing spacers matching prophages in other *Pectobacterium* genomes correlated with the absence of these mobile elements. The current role, if any, of these prophages is not clear. However, in the case of ΦECA29 in *P. atrosepticum* SCRI1043, this prophage was shown to excise from the chromosome and circularize.[69] Furthermore, deletion of this entire prophage led to a reduction in motility and phytopathogenicity[69] and, hence, CRISPR/Cas might limit the acquisition or retention of prophage-encoded virulence determinants. In our study, the detection of protospacer targets also led to the identification of new putative prophages (ΦPCC21_1, ΦPC1_1 and ΦPC1_2) in recently sequenced genomes. Thus, these CRISPRTarget hits enable confidence in the prediction of mobile regions of bacterial genomes, which are often poorly annotated. *Pectobacterium* strains PCC21 and WPP163 also contained spacers that matched phage ZF40 (JQ177065), a "dwarf" Myoviridae,[75] suggesting previous exposure to this, or a related, temperate phage. Given the phage and prophage interactions detected, it is of interest that strains WPP163,

Table 2. Predicted CRISPR arrays in *Pectobacterium* species

Name§	Type	*P. atrosepticum* SCRI1043 (NC_004547)	*P. carotovorum* subsp *carotovorum* PCC21 (NC_018525)	*P. wasabiae* (NC_013421)
CRISPR1	I-F	28	38	17
CRISPR2	I-F	10	3	25
CRISPR3	I-F	3	3	
CRISPR4	I-E		14*	16
CRISPR5	I-E		14*	6

§Names do not indicate CRISPR relationship between strains. *Likely to be one array of 29 spacers, with a 76 base spacer in the middle.

PCC21 and SCRI1043 were isolated from the USA, Korea and Scotland, respectively, over 20 y apart.

In conclusion, we have developed and tested CRISPRTarget, a flexible, interactive tool for the discovery of the targets of crRNAs in diverse databases. There is currently no comparable webserver available and, thus, CRISPRTarget will provide a valuable resource for the growing CRISPR research community.

Materials and Methods

Target databases. Selected databases are provided in CRISPRTarget. GenBank databases: BLAST Nucleotide databases (1) The nr/nt collection ~43 billion bases (15/10/2012, GenBank 192). This database contains "All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS or phase 0, 1 or 2 HTGS sequences)." (2) env_nt, 8.5 billion bases (15/10/2012). This contains "Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples. The largest single source is Sargasso Sea project. This does not overlap with nucleotide nr." This is part of the whole genome shotgun (wgs), but these sequences have no taxonomic classification other than metagenome. (3) Phage division (phg). This is one of the smallest GenBank divisions containing 6,800 sequences of 88 million bases. RefSeq databases: Several relevant divisions of the NCBI Reference Sequence databases are available, which contain better annotated (by NCBI) versions of GenBank sequences. (1) RefSeq-Plasmid. 3,707 sequences, 282 million bases. (2) RefSeq-Viral. 4,279 sequences, 95 million bases. (3) RefSeq-Microbial. 5,234 complete microbial genomes, 7 billion bases. We also included parts of the CAMERA databases. 913,9883 sequences, 1 billion bases. ACLAME. 125,190 sequences, 96 million bases. (4) User defined. Users can upload sequences of up to 50 Mb.

CRISPR array sequences. CRISPR arrays were used from published studies or CRISPRdb. They were also predicted with CRISPRFinder, PILER-CR or CRT using the default parameters. The current tools for prediction have some limitations, notably, the lack of prediction of the transcribed strand, the imprecise definition of the DR/Spacer junctions or splitting into several sub arrays.

Algorithm. *Input data.* Spacer sequences are extracted from the input CRISPR arrays using the locations specified and converted to FASTA format. Alternatively, spacer sequences can be uploaded directly, without repeat sequences, however this limits subsequent processing.

BLAST screen. Each spacer sequence is used to query the selected databases. Multiple databases can be selected, except where there are identical accession identifiers (nt + phg). The default values used by NCBI BLASTn for short sequences, < 30 bases (defaults for long sequences are in brackets), are Gap open -5(-5), gap extend -2(-2), match +1(+1), mismatch -3(-3), word size 7(11), Expect (E): 1,000 (10). Filter: No (Yes). The initial CRISPRTarget defaults are the same except that a gap is penalized more highly (-10), the mismatch penalty is -1 and the E filter is 1. In addition, there is also no filter or masking for low complexity. The CRISPRTarget BLASTn parameters favor gapless matches but allow a number of mismatches at this screening stage. BLAST calculates the scores over the length of the match, and only shows this match. For example, a spacer of 32 bases that matches to a target in 17 of 20 bases would score 20 - 3 = 17 and 20 bases would be output. The expected (E) values of the match will be more likely to pass the filter if smaller databases are used (e.g., the default phg and plasmid). The hits are converted into GFF format.

Extension of the BLAST match. The full spacer and handles are extracted from the input sequences. In the case of CRISPRFinder input, only a single repeat is in the input and this is used for all spacer handles. Both CRT and PILER-CR outputs enable small differences in the repeat to be used. If the user wishes to extract more sequence than provided in the array files, e.g., the sequence following the final repeat, this can be extracted from a FASTA file (if provided by the user). Extension of the spacer is not possible if only spacer sequences are in the input. The protospacer target is extended by extracting the user-specified length of sequence from the BLAST database.

CRISPRTarget interactive scoring. All putative spacer/protospacer targets passing the BLAST screen are displayed in an interactive manner. An initial score is calculated by scoring matches (+1) and mismatches (-1) across the whole length of the spacer without gaps. Specific user defined 'seed' regions can be required to match at either or both ends of the protospacer. A match to pre-defined, or novel user-defined, PAM sequences can increase the score. In order to penalize self-matches that would match 100% in both spacers and flanking handles (e.g., to the original genomic array sequence), a score can be used that penalizes matches (e.g., -1) in the flanking handles. Mismatch penalties can also be used to identify targeting that is facilitated by mismatches in the handles (e.g., type III-A).[43] Finally, a cutoff score can be applied to display only those matches with the best scores.

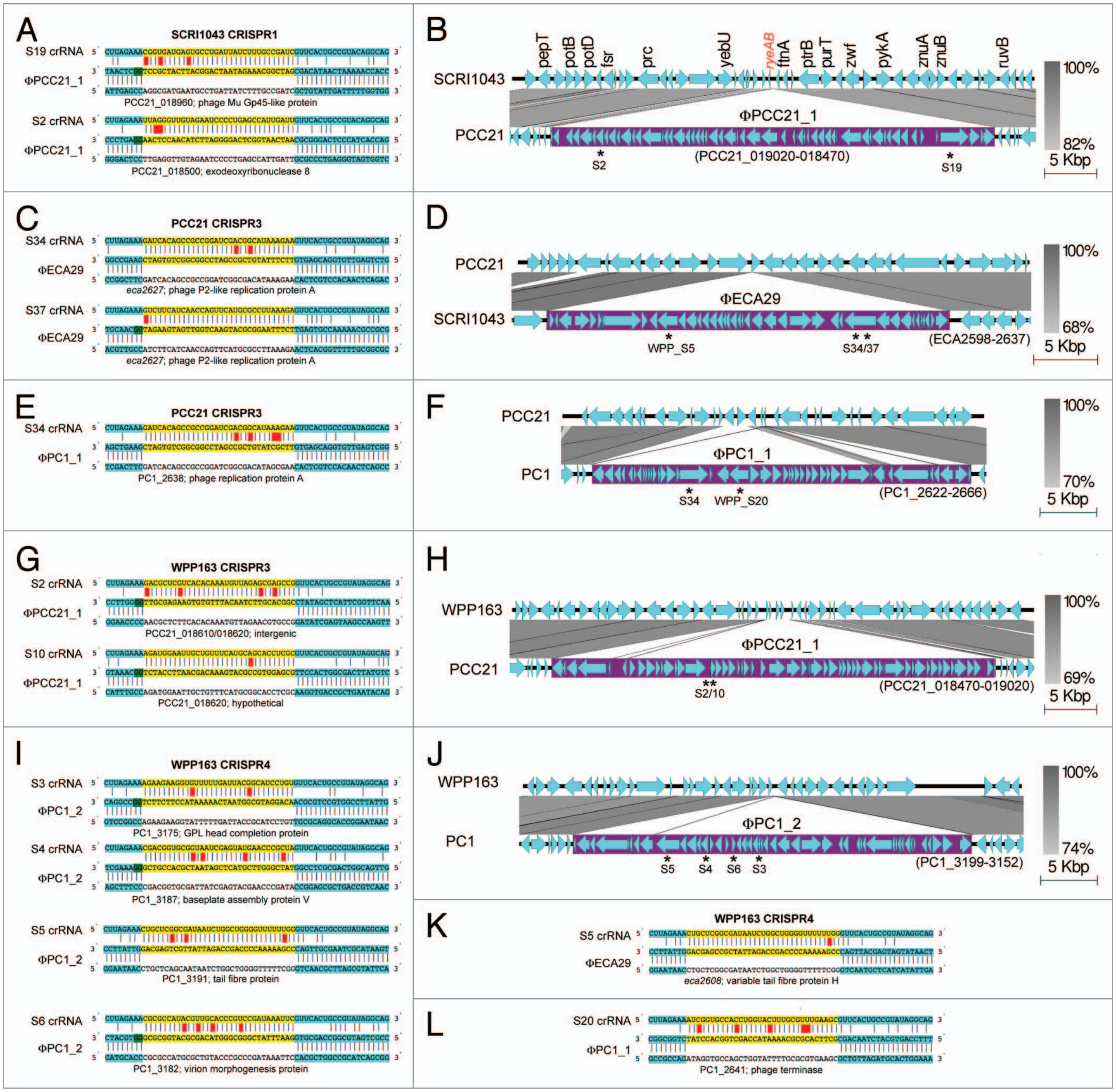


Figure 5. *Pectobacterium* prophages are targeted by CRISPR/Cas. (**A**) Prophage ϕPCC21_1 is targeted by spacers in *P. atrosepticum*. (**B**) *P. atrosepticum* SCRI1043 (top, 2761697–2811697) compared with ϕPCC21_1 in *P. carotovorum* subsp *carotovorum* PCC21 (bottom, phage coordinates: PCC21_018470–019020 from 2092807–2135244. PCC21 is reversed for clarity). (**C**) Prophage ϕECA29 is targeted by spacers in *P. carotovorum* subsp *carotovorum* PCC21. (**D**) *P. carotovorum* subsp *carotovorum* PCC21 (top, PCC21_017190–017500 from 1936500–1976500. PCC21 is reversed) compared with ϕECA29 (HAI9) in *P. atrosepticum* SCRI1043 (bottom, ECA2598-ECA2637 from 2935264–2966783). (**E**) Prophage ϕPC1_1 is targeted by a spacer in *P. carotovorum* subsp *carotovorum* PCC21. (**F**) *P. carotovorum* subsp *carotovorum* PCC21 (top, PCC21_027150–027460 from 3058299–3095299) compared with ϕPC1_1 in *P. carotovorum* subsp *carotovorum* PC1 (bottom, PC1_2622–2666 from 2989228–3022511). (**G**) Prophage ϕPCC21_1 is targeted by spacers in *P. wasabiae*. (**H**) *P. wasabiae* WPP163 (top, 2291600–2341600) compared with ϕPCC21_1 in *P. carotovorum* subsp *carotovorum* PCC21 (bottom, phage coordinates: PCC21_018470–019020 from 2092807–2135244). (**I**) Prophage ϕPC1_2 is targeted by spacers in *P. wasabiae*. (**J**) *P. wasabiae* WPP163 (top, 1192372–1236372) compared with ϕPC1_2 in *P. carotovorum* subsp *carotovorum* PC1 (bottom, phage coordinates: PC1_3152–3199 from 3573374–3608557. PC1 is reversed). Prophages (**K**) ϕECA29 and (**L**) ϕPC1_2 are targeted by *P. wasabiae* spacers. Genome comparisons were generated using Easyfig;[77] genes are cyan arrows, putative prophage regions are purple and spacer target locations indicated with asterisks. Homologous regions by BLASTn are shown in shades of gray.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This work was supported by a Rutherford Discovery Fellowship from the Royal Society of NZ to P.C.F., the VLAG (Voeding, Levensmiddelentechnologie, Agrobiotechnologie en Gezondheid) to P.C.F., a NWO-VIDI grant to S.J.J.B. (864.11.005) and by a Human Frontier Science Program Grant to Ian Macara, Anne Spang and C.M.B. A.B. is a recipient of a University of Otago Postgraduate Scholarship. We thank John van der Oost, Raymond Staals and Edze Westra for useful discussions.

Supplemental Material

Supplemental material may be found:
www.landesbioscience.com/journals/rnabiology/article/24046

References

1. Bhaya D, Davison M, Barrangou R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 2011; 45:273-97; PMID:22060043; http://dx.doi.org/10.1146/annurev-genet-110410-132430.
2. Richter C, Chang JT, Fineran PC. The function and regulation of CRISPR/Cas systems. Viruses 2012; 4:2291-311; PMID:23202464; http://dx.doi.org/10.3390/v4102291.
3. Terns MP, Terns RM. CRISPR-based adaptive immune systems. Curr Opin Microbiol 2011; 14:321-7; PMID:21531607; http://dx.doi.org/10.1016/j.mib.2011.03.005.
4. Westra ER, Swarts DC, Staals RH, Jore MM, Brouns SJ, van der Oost J. The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet 2012; 46:311-39; PMID:23145983; http://dx.doi.org/10.1146/annurev-genet-110711-155447.
5. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature 2012; 482:331-8; PMID:22337052; http://dx.doi.org/10.1038/nature10886.
6. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 2007; 8:172; PMID:17521438; http://dx.doi.org/10.1186/1471-2105-8-172.
7. Fineran PC, Charpentier E. Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. Virology 2012; 434:202-9; PMID:23123013; http://dx.doi.org/10.1016/j.virol.2012.10.003.
8. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev 2008; 22:3489-96; PMID:19141480; http://dx.doi.org/10.1101/gad.1742908.
9. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 2008; 321:960-4; PMID:18703739; http://dx.doi.org/10.1126/science.1159689.
10. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science 2010; 329:1355-8; PMID:20829488; http://dx.doi.org/10.1126/science.1192272.
11. Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, et al. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure 2012; 20:1574-84; PMID:22841292; http://dx.doi.org/10.1016/j.str.2012.06.016.
12. Przybilski R, Richter C, Gristwood T, Clulow JS, Vercoe RB, Fineran PC. Csy4 is responsible for CRISPR RNA processing in *Pectobacterium atrosepticum.* RNA Biol 2011; 8:517-28; PMID:21519197; http://dx.doi.org/10.4161/rna.8.3.15190.
13. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 2011; 471:602-7; PMID:21455174; http://dx.doi.org/10.1038/nature09886.
14. Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, et al. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell 2012; 45:303-13; PMID:22227115; http://dx.doi.org/10.1016/j.molcel.2011.12.013.
15. Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci USA 2011; 108:10092-7; PMID:21536913; http://dx.doi.org/10.1073/pnas.1102716108.
16. Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, et al. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J Biol Chem 2011; 286:21643-56; PMID:21507944; http://dx.doi.org/10.1074/jbc.M111.238485.
17. Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, et al. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol 2011; 18:529-36; PMID:21460843; http://dx.doi.org/10.1038/nsmb.2019.
18. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, et al. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell 2009; 139:945-56; PMID:19945378; http://dx.doi.org/10.1016/j.cell.2009.07.040.
19. Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, et al. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature 2011; 477:486-9; PMID:21938068; http://dx.doi.org/10.1038/nature10402.
20. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 2012; 337:816-21; PMID:22745249; http://dx.doi.org/10.1126/science.1225829.
21. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci USA 2012; 109:E2579-86; PMID:22949671; http://dx.doi.org/10.1073/pnas.1208507109.
22. Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 2010; 468:67-71; PMID:21048762; http://dx.doi.org/10.1038/nature09523.
23. Westra ER, van Erp PB, Künne T, Wong SP, Staals RH, Seegers CL, et al. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell 2012; 46:595-605; PMID:22521689; http://dx.doi.org/10.1016/j.molcel.2012.03.018.
24. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007; 315:1709-12; PMID:17379808; http://dx.doi.org/10.1126/science.1138140.
25. Cady KC, Bondy-Denomy J, Heussler GE, Davidson AR, O'Toole GA. The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. J Bacteriol 2012; 194:5728-38; PMID:22885297; http://dx.doi.org/10.1128/JB.01184-12.
26. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 2008; 322:1843-5; PMID:19095942; http://dx.doi.org/10.1126/science.1165771.
27. Touchon M, Rocha EP. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella.* PLoS One 2010; 5:e11126; PMID:20559554; http://dx.doi.org/10.1371/journal.pone.0011126.
28. Díez-Villaseñor C, Almendros C, García-Martínez J, Mojica FJ. Diversity of CRISPR loci in *Escherichia coli.* Microbiology 2010; 156:1351-61; PMID:20133361; http://dx.doi.org/10.1099/mic.0.036046-0.
29. Hatfull GF, Hendrix RW. Bacteriophages and their genomes. Curr Opin Virol 2011; 1:298-303; PMID:22034588; http://dx.doi.org/10.1016/j.coviro.2011.06.009.
30. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev 2011; 75:610-35; PMID:22126996; http://dx.doi.org/10.1128/MMBR.00011-11.
31. Culley AI. Virophages to viromes: a report from the frontier of viral oceanography. Curr Opin Virol 2011; 1:52-7; PMID:22440567; http://dx.doi.org/10.1016/j.coviro.2011.05.003.
32. Ray J, Dondrup M, Modha S, Steen IH, Sandaa RA, Clokie M. Finding a needle in the virus metagenome haystack--micro-metagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. PLoS One 2012; 7:e34238; PMID:22509283; http://dx.doi.org/10.1371/journal.pone.0034238.
33. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, et al. The marine viromes of four oceanic regions. PLoS Biol 2006; 4:e368; PMID:17090214; http://dx.doi.org/10.1371/journal.pbio.0040368.
34. Leplae R, Lima-Mendez G, Toussaint A. A first global analysis of plasmid encoded proteins in the ACLAME database. FEMS Microbiol Rev 2006; 30:980-94; PMID:17064288; http://dx.doi.org/10.1111/j.1574-6976.2006.00044.x.
35. Li LL, Norman A, Hansen LH, Sørensen SJ. Metamobilomics--expanding our knowledge on the pool of plasmid encoded traits in natural environments using high-throughput sequencing. Clin Microbiol Infect 2012; 18(Suppl 4):5-7; PMID:22647039; http://dx.doi.org/10.1111/j.1469-0691.2012.03862.x.
36. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 2011; 9:467-77; PMID:21552286; http://dx.doi.org/10.1038/nrmicro2577.
37. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus.* J Bacteriol 2008; 190:1390-400; PMID:18065545; http://dx.doi.org/10.1128/JB.01412-07.
38. Manica A, Zebec Z, Teichmann D, Schleper C. In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. Mol Microbiol 2011; 80:481-91; PMID:21385233; http://dx.doi.org/10.1111/j.1365-2958.2011.07586.x.

39. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci USA 2011; 108:10098-103; PMID:21646539; http://dx.doi.org/10.1073/pnas.1104144108.

40. Gudbergsdottir S, Deng L, Chen Z, Jensen JV, Jensen LR, She Q, et al. Dynamic properties of the *Sulfolobus* CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Mol Microbiol 2011; 79:35-49; PMID:21166892; http://dx.doi.org/10.1111/j.1365-2958.2010.07452.x.

41. Mojica FJ, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 2009; 155:733-40; PMID:19246744; http://dx.doi.org/10.1099/mic.0.023960-0.

42. Sashital DG, Wiedenheft B, Doudna JA. Mechanism of foreign DNA selection in a bacterial adaptive immune system. Mol Cell 2012; 46:606-15; PMID:22521690; http://dx.doi.org/10.1016/j.molcel.2012.03.020.

43. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 2010; 463:568-71; PMID:20072129; http://dx.doi.org/10.1038/nature08703.

44. van Rij RP, Andino R. The silent treatment: RNAi as a defense against virus infection in mammals. Trends Biotechnol 2006; 24:186-93; PMID:16503061; http://dx.doi.org/10.1016/j.tibtech.2006.02.006.

45. Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. Nucleic Acids Res 2011; 39:6845-53; PMID:21652644; http://dx.doi.org/10.1093/nar/gkr330.

46. Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. Nat Struct Mol Biol 2010; 17:1169-74; PMID:20924405; http://dx.doi.org/10.1038/nsmb.1921.

47. Saito T, Sætrom P. Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments. Silence 2012; 3:3; PMID:22325809; http://dx.doi.org/10.1186/1758-907X-3-3.

48. Tan Gana NH, Victoriano AF, Okamoto T. Evaluation of online miRNA resources for biomedical applications. Genes Cells 2012; 17:11-27; PMID:22077698; http://dx.doi.org/10.1111/j.1365-2443.2011.01564.x.

49. Witkos TM, Koscianska E, Krzyzosiak WJ. Practical Aspects of microRNA Target Prediction. Curr Mol Med 2011; 11:93-109; PMID:21342132; http://dx.doi.org/10.2174/156652411794859250.

50. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 2007; 8:209; PMID:17577412; http://dx.doi.org/10.1186/1471-2105-8-209.

51. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics 2007; 8:18; PMID:17239253; http://dx.doi.org/10.1186/1471-2105-8-18.

52. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 2007; 35(Web Server issue):W52-7; PMID:17537822; http://dx.doi.org/10.1093/nar/gkm360.

53. Cady KC, White AS, Hammond JH, Abendroth MD, Karthikeyan RS, Lalitha P, et al. Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. Microbiology 2011; 157:430-7; PMID:21081758; http://dx.doi.org/10.1099/mic.0.045732-0.

54. Rousseau C, Gonnet M, Le Romancer M, Nicolas J. CRISPI: a CRISPR interactive database. Bioinformatics 2009; 25:3317-8; PMID:19846435; http://dx.doi.org/10.1093/bioinformatics/btp586.

55. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res 2008; 36(Web Server issue):W5-9; PMID:18440982; http://dx.doi.org/10.1093/nar/gkn201.

56. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics 2008; 24:1757-64; PMID:18567917; http://dx.doi.org/10.1093/bioinformatics/btn322.

57. Mavromatis K, Ivanova NN, Chen IM, Szeto E, Markowitz VM, Kyrpides NC. The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. Stand Genomic Sci 2009; 1:63-7; PMID:21304638; http://dx.doi.org/10.4056/sigs.632.

58. Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M, Deveau H, Moineau S, et al. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus.* J Bacteriol 2008; 190:1401-12; PMID:18065539; http://dx.doi.org/10.1128/JB.01415-07.

59. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 2007; 8:R61; PMID:17442114; http://dx.doi.org/10.1186/gb-2007-8-4-r61.

60. Lillestøl RK, Redder P, Garrett RA, Brügger K. A putative viral defence mechanism in archaeal cells. Archaea 2006; 2:59-72; PMID:16877322; http://dx.doi.org/10.1155/2006/542818.

61. Toth IK, Birch PR. Rotting softly and stealthily. Curr Opin Plant Biol 2005; 8:424-9; PMID:15970273; http://dx.doi.org/10.1016/j.pbi.2005.04.001.

62. Frampton RA, Pitman AR, Fineran PC. Advances in bacteriophage-mediated control of plant pathogens. Int J Microbiol 2012; 2012:326452; PMID:22934116; http://dx.doi.org/10.1155/2012/326452.

63. Richter C, Gristwood T, Clulow JS, Fineran PC. *In vivo* protein interactions and complex formation in the *Pectobacterium atrosepticum* subtype I-F CRISPR/Cas System. PLoS One 2012; 7:e49549; PMID:23226499; http://dx.doi.org/10.1371/journal.pone.0049549.

64. Bell KS, Sebaihia M, Pritchard L, Holden MT, Hyman LJ, Holeva MC, et al. Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. Proc Natl Acad Sci USA 2004; 101:11105-10; PMID:15263089; http://dx.doi.org/10.1073/pnas.0402424101.

65. Park TH, Choi BS, Choi AY, Choi IY, Heu S, Park BS. Genome sequence of *Pectobacterium carotovorum* subsp. *carotovorum* strain PCC21, a pathogen causing soft rot in Chinese cabbage. J Bacteriol 2012; 194:6345-6; PMID:23105077; http://dx.doi.org/10.1128/JB.01583-12.

66. Abbott JC, Aanensen DM, Rutherford K, Butcher S, Spratt BG. WebACT--an online companion for the Artemis Comparison Tool. Bioinformatics 2005; 21:3665-6; PMID:16076890; http://dx.doi.org/10.1093/bioinformatics/bti601.

67. Balbontín R, Figueroa-Bossi N, Casadesús J, Bossi L. Insertion hot spot for horizontally acquired DNA within a bidirectional small-RNA locus in *Salmonella enterica.* J Bacteriol 2008; 190:4075-8; PMID:18390661; http://dx.doi.org/10.1128/JB.00220-08.

68. Kropinski AM, Kovalyova IV, Billington SJ, Patrick AN, Butts BD, Guichard JA, et al. The genome of epsilon15, a serotype-converting, Group E1 *Salmonella enterica*-specific bacteriophage. Virology 2007; 369:234-44; PMID:17825342; http://dx.doi.org/10.1016/j.virol.2007.07.027.

69. Evans TJ, Coulthurst SJ, Komitopoulou E, Salmond GP. Two mobile *Pectobacterium atrosepticum* prophages modulate virulence. FEMS Microbiol Lett 2010; 304:195-202; PMID:20146746; http://dx.doi.org/10.1111/j.1574-6968.2010.01901.x.

70. Langille MG, Brinkman FS. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 2009; 25:664-5; PMID:19151094; http://dx.doi.org/10.1093/bioinformatics/btp030.

71. Gotea V, Veeramachaneni V, Makałowski W. Mastering seeds for genomic size nucleotide BLAST searches. Nucleic Acids Res 2003; 31:6935-41; PMID:14627826; http://dx.doi.org/10.1093/nar/gkg886.

72. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. CAMERA: a community resource for metagenomics. PLoS Biol 2007; 5:e75; PMID:17355175; http://dx.doi.org/10.1371/journal.pbio.0050075.

73. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun 2012; 3:945; PMID:22781758; http://dx.doi.org/10.1038/ncomms1937.

74. Swarts DC, Mosterd C, van Passel MW, Brouns SJ. CRISPR interference directs strand specific spacer acquisition. PLoS One 2012; 7:e35888; PMID:22558257; http://dx.doi.org/10.1371/journal.pone.0035888.

75. Comeau AM, Tremblay D, Moineau S, Rattei T, Kushkina AI, Tovkach FI, et al. Phage morphology recapitulates phylogeny: the comparative genomics of a new group of myoviruses. PLoS One 2012; 7:e40102; PMID:22792219; http://dx.doi.org/10.1371/journal.pone.0040102.

76. Hatoum-Aslan A, Maniv I, Marraffini LA. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. Proc Natl Acad Sci USA 2011; 108:21218-22; PMID:22160698; http://dx.doi.org/10.1073/pnas.1112832108.

77. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics 2011; 27:1009-10; PMID:21278367; http://dx.doi.org/10.1093/bioinformatics/btr039.

78. Lillestøl RK, Shah SA, Brügger K, Redder P, Phan H, Christiansen J, et al. CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. Mol Microbiol 2009; 72:259-72; PMID:19239620; http://dx.doi.org/10.1111/j.1365-2958.2009.06641.x.

79. Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, Randau L. Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis.* Nucleic Acids Res 2012; 40:9887-96; PMID:22879377; http://dx.doi.org/10.1093/nar/gks737.

80. Fischer S, Maier LK, Stoll B, Brendel J, Fischer E, Pfeiffer F, et al. An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. J Biol Chem 2012; 287:33351-63; PMID:22767603; http://dx.doi.org/10.1074/jbc.M112.377002.

81. Semenova E, Nagornykh M, Pyatnitskiy M, Artamonova II, Severinov K. Analysis of CRISPR system function in plant pathogen *Xanthomonas oryzae.* FEMS Microbiol Lett 2009; 296:110-6; PMID:19459963; http://dx.doi.org/10.1111/j.1574-6968.2009.01626.x.

82. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli.* Nucleic Acids Res 2012; 40:5569-76; PMID:22402487; http://dx.doi.org/10.1093/nar/gks216.

83. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 2005; 151:2551-61; PMID:16079334; http://dx.doi.org/10.1099/mic.0.28048-0.

84. Magadán AH, Dupuis ME, Villion M, Moineau S. Cleavage of phage DNA by the *Streptococcus thermophilus* CRISPR3-Cas system. PLoS One 2012; 7:e40913; PMID:22911717; http://dx.doi.org/10.1371/journal.pone.0040913.

85. Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, et al. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Mol Cell 2012; 45:292-302; PMID:22227116; http://dx.doi.org/10.1016/j.molcel.2011.10.023.