CRISPR-CAS SYSTEM CHARACTERIZATION AND ANTI-CRISPR DISCOVERY IN *STREPTOCOCCUS THERMOPHILUS* STRAINS AND PHAGES

by

CLARE EDWARDS COOPER

(Under the Direction of Michael P. Terns)

ABSTRACT

CRISPR (clustered regularly interspaced short palindromic repeat) arrays and Cas (CRISPR-associated) proteins provide bacteria and archaea with immunity against phages and other mobile genetic elements (MGEs). The immunity provided by CRISPR-Cas systems is adaptive as sequences are acquired from invaders and stored in the CRISPR array, capable of guiding sequence-specific nuclease activity during future encounters. In response, phages and other MGEs encode anti-CRISPR proteins that inhibit the defense of CRISPR-Cas immunity. This dissertation begins with characterization of CRISPR-Cas systems in *Streptococcus thermophilus* with an emphasis on Type III-A systems. It then explores anti-CRISPR protein prediction in phages of *S. thermophilus* followed by screening and identification of novel CRISPR-Cas inhibitors. The final chapter focuses on the selectivity of spacer targeting against *S. thermophilus* phage genes with potential implications for our understanding of Type III-A CRISPR-Cas system function. The results increase predictability of outcomes of phage-host encounters by expanding the repertoire of known anti-CRISPR proteins and illuminating unique features and potential roles of cooccurring CRISPR-Cas systems.

INDEX WORDS: CRISPR; Cas; Csm; Type III-A; anti-CRISPR; bacteriophage; *Streptococcus thermophilus*

CRISPR-CAS SYSTEM CHARACTERIZATION AND ANTI-CRISPR DISCOVERY IN *STREPTOCOCCUS THERMOPHILUS* STRAINS AND PHAGES

by

CLARE EDWARDS COOPER

B.S., University of Georgia, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

CRISPR-CAS SYSTEM CHARACTERIZATION AND ANTI-CRISPR DISCOVERY IN *STREPTOCOCCUS THERMOPHILUS* STRAINS AND PHAGES

by

CLARE EDWARDS COOPER

Major Professor:	Michael P. Terns
Committee:	David J. Garfinkel
	Natarajan Kannan
	Robert Sabatini

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

DEDICATION

This dissertation is dedicated to my husband Derek and our dogs (the ‘snack pack’) as well as friends in Athens that have helped us celebrate life along the way.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Terns for being a role-model in his dedication to scientific exploration and for encouraging me to chase questions during my graduate career. My committee members, Dr. David Garfinkel, Dr. Natarajan Kannan, and Dr. Bob Sabatini have been critical for my success, both asking and answering important questions. In addition, both Dr. Lance Wells and Dr. Zach Wood have been hugely helpful in their roles as graduate coordinators.

So many of my lab mates have been both inspiring and supportive. Thank you to Walter (Tom) Woodside (and by extension, Laura and Frida) for many troubleshooting and scientific conversations over the years (as well as maintenance of plants for our shared cubby). Thank you to Elizabeth Watts (as well as Dustin, Tooti, Dylan, & Harley) for your genuine kindness and caring both in and out of the lab. In addition, thank to you all past and current Terns Lab members who have helped contribute to my growth as a scientist and a person, including Yunzhou Wei, Masami Shiimori, Julie Grainy (and family!), Kawanda Foster, Jenny Kim, Sandra Garrett, Xinfu Zhang, Justin Mclean, Ryan Catchpole, Cécile Philippe, Katie Johnson, Chris Noble-Molnar, Conor Pittman, and Landon Clark.

To mentees that have taught me how to be a mentor, I am very grateful. Thank you to many rotation students and undergraduates, but especially Raven Tucker and Ela Mitchell, for your patience, excitement, and hard work.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Introduction to CRISPR-Cas Systems

In 2005, CRISPR-Cas systems were hypothesized to function in bacteria as adaptive immune systems against bacteriophage infection, and in 2013, the nuclease Cas9 was biochemically characterized and harnessed for genome editing (1-4). The use of CRISPR-Cas for genetic tractability truly exploded, and in less than a decade since its biochemical characterization (and during the time of my PhD studies), Cas9 was used to generate embryonic mutations in a set of human twins (5). This event led to a heated debate over the ethics and safety of CRISPR-Cas genome editing, making it clear that in a field that grows this rapidly, understanding the basic biology of CRISPR-Cas immunity and inhibition is paramount to safety and efficacy.

As was first hypothesized in 2005, CRISPR-Cas systems are heritable immune systems present in 95% of archaea and 48% of bacteria (6). While bacteria encode numerous defenses against phage and other mobile genetic elements, the majority are innate immune systems, incapable of invader-specific responses or memory of past encounters (7). In contrast, **c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeat (CRISPR) systems are adaptive immune systems capable of storing a sequence-specific history of encounters with foreign invaders (1, 8). There are three main stages of CRISPR defense (Figure 1.1) beginning with adaptation, where CRISPR associated (Cas) proteins

acquire and insert short nucleotide sequences (protospacers) from invading mobile genetic elements between the conserved repeats of the CRISPR Array (8). During the crRNA biogenesis stage (Figure 1.1), the CRISPR array is transcribed and processed at the repeat sequences to form crRNA (9-11). The defense phase (Figure 1.1) occurs when the same sequence is again present. Cas proteins are guided by the crRNAs to target and degrade the invader sequence (12).

There are two broad classes of CRISPR-Cas systems which are subdivided into six types with many additional subtypes (13, 14). Figure 1.2 shows these two main classes with class I systems (Types I, III, and IV) characterized by the use of multi-subunit effector complexes, while class II systems (Type II, V, and VI) have single effector proteins (15). Many bacteria encode more than one type of CRISPR-Cas system due to the utility of different defense mechanisms in one host (13, 14). The organism at the center of this dissertation, *Streptococcus thermophilus* is one such organism. *S. thermophilus*. can encode up to four CRISPR-Cas systems, including the Type I-E, Type II-A (CRISPR1 & CRISPR3), and Type III-A systems (16). These system types are highlighted in Figure 1.2. In this introduction, I will focus on these specific system types with an emphasis on Type III-A systems. Each of the systems in *S. thermophilus* function independently of one another, harboring their own CRISPR arrays, adaptation, crRNA biogenesis, and defense machinery (17).

Type II systems are the most well-known, with a single Cas9 effector protein (Figure 1.2) that targets and degrades DNA. The subtype II-A system, present in *S. thermophilus*, is distinguished by the presence of a Csn2 protein (13, 14). Type I-E systems are also DNA-targeting systems but are made up of a multi-subunit effector complex

instead of a single effector molecule (Figure 1.2). The nuclease activity of Type I-E systems is provided by Cas3 (18).

To degrade a target, the Type I and Type II systems require the presence of a **p**rotospacer **a**djacent **m**otif (PAM), which as the name describes, is a sequence adjacent to the protospacer, or the sequence that was originally adapted from the phage and integrated into the CRISPR-array (19-21). Any future target sequence must maintain this PAM. The PAM requirement of Type II-A and Type I-E systems causes a strict requirement for target sequence integrity and leaves little room for mutations (22). During infection, bacteriophages with mutations within the PAM sequence are selected for, allowing escape from CRISPR-immunity (22).

Type III-A systems are unique in carrying out both DNA and RNA degradation, with defense dependent on transcription of a target sequence and target RNA recognition (23, 24). The multi-subunit effector complex of the *S. thermophilus* type III-A system is composed of Csm1, Csm2, Csm3, Csm4 and Csm5 (Figure 1.2 and Figure 1.3) with an additional protein, Csm6, that does not associate with the complex (25-27).

Figure 1.3 demonstrates that instead of a defined PAM sequence, activity of the Type III-A or Csm complex is regulated by potential complementarity of the 5' tag of the crRNA with the 3' **p**rotospacer **f**lanking **s**equence (PFS) of the target RNA (28). The 5' tag of the crRNA is a typically 8 nucleotide long remnant of the repeat sequence left during processing of the transcribed Type III-A array (11, 25). One function of the PFS and 5' tag interaction is to prevent self-cutting of the CRISPR array as complete complementarity between the 5' tag and the repeat sequence will not activate the Type III-A complex (23, 29). The 3' PFS can differ for each target RNA and is the sequence that becomes aligned

across from the 5' crRNA tag sequence following crRNA-target RNA base-pairing (28-30). The variable defense activities of the Type III-A system, based on this crRNA tag and target RNA PFS potential interaction, are modeled in Figure 1.3. This model is based on published Cryo-EM structures of the *S. thermophilus* Csm complex bound to cognate and non-cognate target RNA (31). The Csm complex first binds crRNA with multiple Csm3 and Csm2 subunits forming the backbone of the complex and Csm5 capping off the 3' end of the crRNA. When target RNA is bound by the complex, Csm3 interacts with and cuts the target RNA at 6 nucleotide intervals, preventing pairing between the target and crRNA at these locations (31). As seen in Figure 1.3, the activity of Csm3 is not dependent on the sequence of the 3' PFS of the target RNA, but on crRNA and target RNA pairing. Csm3 can degrade the target sequence and repress target transcription even if the 3' PFS sequence of the target is complementary to the 5' tag of the crRNA (23, 31). In comparison Csm1/Cas10 is activated when the Csm complex binds to a target RNA with 3' PFS non-complementary to the 5' tag sequence of the crRNA (23, 28). Csm1 has two functional domains allosterically activated by interaction with a non-complementary 3' PFS sequence (bottom of Figure 1.3). The first is an HD nuclease domain that non-specifically degrades nearby ssDNA (28). In addition, there are two Palm domains with the Palm2 domain similar to the "Palm" of polymerases and cyclases (GGDD active site instead of GGDEF) and capable of converting ATP to cyclic oligo adenylate (cOA) (30, 32-35). Cyclic oligo A is released into the cell and acts as a second messenger, binding to Csm6 at its CRISPR-Cas Associated Rossmann Fold (CARF) domain, activating non-specific RNase activity of the Csm6 HEPN (Higher Eukaryotes and Prokaryotes Nucleotide binding) domain (30, 36-

40). Cleavage of the target RNA by Csm3 inactivates the complex and shuts down DNA cleavage and the production of cyclic oligo A (30, 31).

In Type III-A defense, base pairing between the crRNA and target RNA is forgiving, allowing multiple mutations throughout the length of the target, with the highest sequence specificity requirement adjacent to the 5' tag of the crRNA (41). The laxity in base-pairing requirements for Type III-A makes phage escape from these systems more difficult. Phages must undergo deletion of entire target sequences to escape Type III-A defense (41).

Introduction to anti-CRISPR Proteins

In opposition of CRISPR-defense systems are bacteriophages and other mobile genetic elements (7). Having endured a long evolutionary road with their hosts, phages were found to encode small protein inhibitors of CRISPR immunity (42-46). These, so-called, anti-CRISPRs (Acrs) were first identified in phages capable of inhibiting the Type I-F CRISPR-Cas system in *Pseudomonas Aeruginosa* (42). The phages were immune to CRISPR-Cas defense despite previous CRISPR-Cas immunization against them. Later studies of Acr identification noted that anti-CRISPR proteins are quite small and share no common domain or homology (42, 45). In addition, several studies found helix-turn-helix (HTH) domain (common to DNA-binding proteins and transcriptional regulators) containing proteins conserved between Acr loci and characterized them as regulators of Acr expression (47, 48). These regulatory proteins are now referred to as anti-CRISPR associated proteins (Aca) and are often more conserved than anti-CRISPR protein sequences. (47, 48). Since this initial discovery, several groups have used a guilt-by-

association approach to identify Acr encoding loci to expand the repertoire of known Acr proteins (49, 50). Figure 1.4B demonstrates guilt-by association which uses known Acr or Aca genes to identify new Acr encoding loci where co-occurring genes are considered Acr candidates. This is the strategy we employ in our identification of Acr candidates for screening (Chapter 3).

To employ guilt by association, there must be an Aca gene or a gene with known or hypothesized Acr function to begin the search for new Acrs. In our search of *Streptococcus thermophilus* phages, this initial gene was AcrIIA5 and the genes that co-occurred with it. AcrIIA5 is a Type II-A Acr identified by the Moineau lab in 2017 (51). The group noted that *S. thermophilus* phage D4276 was resistant to CRISPR-immunization and screened all phage genes to determine which gene was responsible for the activity. They identified 8 homologues of AcrIIA5 in *S. thermophilus* phages and strains (51). Since identification, the different alleles of AcrIIA5 have been shown to have varying levels of specificity against the Type II-A (CR1) system and Type II-A (CR3) system of *Streptococcus thermophilus* (52). Despite variability between alleles, AcrIIA5 has been noted for its broad-range inhibition of Cas9 orthologs in gene editing and other contexts (53-55).

At the advent of this work, there were 22 identified anti-CRISPR proteins with 14 identified against type I systems and 8 against Type II systems (56). AcrIIA5 was the only anti-CRISPR identified in phages of *Streptococcus thermophilus* (51). We sought to expand this Acr repertoire and better understand phage resistance mechanisms against the CRISPR-Cas systems in *Streptococcus thermophilus*.

Introduction to *Streptococcus thermophilus* and associated phages

Streptococcus thermophilus is a lactic acid bacterium used in large-scale fermentations for production of dairy products (57, 58). During production, contamination of a starter culture with lytic phage can lead to loss of quality, so knowledge of bacterial defense mechanisms is important to industry success and profits (59). As mentioned previously, *S. thermophilus* can encode up to four CRISPR-Cas systems. When challenged with a phage, the Type II-A (CR1) locus is the most likely to acquire a new CRISPR spacer against the phage with the Type II-A (CR3) system also shown to readily adapt but with a significantly reduced efficiency relative to the CR1 system (12, 60). Because of this, the other two CRISPR-Cas systems in these strains (Type III-A and Type I-E) are less studied. In addition, the Cas9 proteins of the Type II-A CR1 and CR3 systems have been reconstituted *in vitro* for genome editing (12, 61). For this reason, identification of anti-CRISPRs against Cas9 are of particular interest for biotechnology and biomedical applications.

Phages that infect *S. thermophilus* are of the family *Siphoviridae* and the order *Caudovirales*. They have double-stranded DNA (dsDNA) genomes and non-contractive tails (62, 63). There are two large orders of *S. thermophilus* phages based on packaging and structural gene make-up of the genome (64). Cos-type phages have cohesive ends to their genome while Pac type phages use a headful mechanism of genome packaging (64). Apart from these large orders, there have been two smaller orders identified. The 5093 group shares homology with phages infecting non-dairy strains (65) while the 987 group shares homology with *Lactococcus lactis* phages (66). These homologies likely arose from recombination events. Across all subtypes, the genome has a conserved modular architecture largely divided by gene function (Figure 1.5A) (67-70). Genetic

recombination tends to occur through modular exchange except in two regions of the genome which are the most recombinogenic (67). These regions include part of the regulatory gene module and the lysogenic or lysogenic replacement module. The lysogenic module is referred to as the lysogenic replacement module in lytic phages as it is non-functional for prophage integration, but several remnant genes remain (68, 71). This region is significant for our studies as we determined it to be the location of genes for all known *S. thermophilus* phage anti-CRISPR homologues (Chapter 3).

During infection, there is stringent regulation of phage gene transcription, with timing of expression coordinated to the stages of infection (Figure 1.5B) (72-74). Early gene expression is limited to replication genes, regulatory genes, and genes within the lysogenic replacement module (72-74). This temporal gene regulation is especially significant when considering CRISPR-Cas immunity, and it highlights the unique features of the Type III-A system in *S. thermophilus*. For the Type-III-A system, phage defense is dependent on transcription of a target RNA, so targeting of genes expressed earlier in the lytic cycle is favored (75, 76). This presumably gives the Type III-A system time to quell infection prior to extensive phage replication and lysis. For this reason, when designing Type III-A CRISPR spacers against phage 2972, we preferentially used spacers against early gene targets.

Figure 1.1 Stages of CRISPR-Cas Immunity

There are three defined stages of CRISPR-Cas immunity starting with (1) Adaptation where a sequence fragment is taken from an invader by adaptation machinery and inserted between the repeats of the CRISPR array. (2) crRNA biogenesis begins when the CRISPR array is transcribed and is processed at the repeat sequences to form short mature crRNA sequences capable of pairing with targets. (3) Defense is carried out by a crRNA-ribonucleoprotein (crRNP) complex. Defense occurs when the crRNP binds to the complementary target sequence and the target is degraded by an effector nuclease. Adapted and modified from Terns and Terns, 2014 (77).

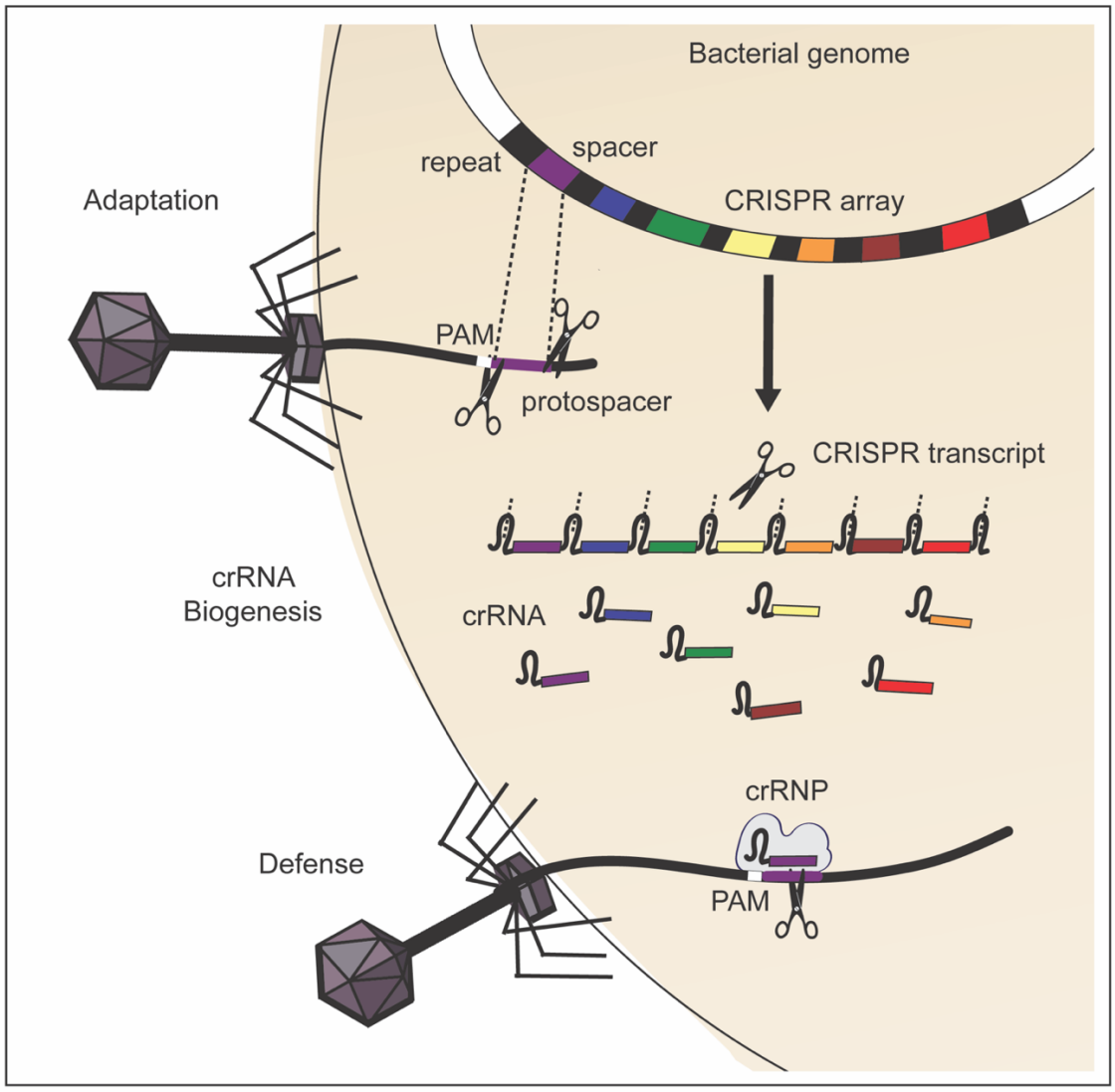

Bacterial genome
repeat
spacer
CRISPR array
Adaptation
PAM
protospacer
CRISPR transcript
crRNA
Biogenesis
crRNA
crRNP
Defense
PAM

Figure 1.2 Classes and Types of CRISPR-Cas Systems

There are two overarching classes of CRISPR-Cas systems based on the use of a multi-subunit effector complex (class 1) or a single effector protein (class 2) for defense. These two classes are further subdivided into types based on the Cas protein makeup. The three system types present in *S. thermophilus* are highlighted here with the protein names corresponding to those of the *S. thermophilus* CRISPR-Cas systems. All three systems encode a Cas1 and Cas2 protein for spacer acquisition. Type I and Type III systems encode Cas6 for crRNA processing and biogenesis while the Type II system relies on RNase III. The Type I and Type III CRISPR-Cas effector nucleases are Cas3 and Csm1/Cas10, respectively. The Type II system encodes Cas9 as the single effector protein. Type III-A systems have an ancillary nuclease (in *S. thermophilus* this is Csm6-1 and Csm6-2) with a CARF sensor domain and HEPN RNase effector domain. Additionally, the CARF domain can function as a ring nuclease to regulate the ancillary RNase function through degradation of cOA. Lastly, Type II systems can have an additional Csn2 protein for which the function is not fully understood. Adapted from Makarova et al., 2020 (13).

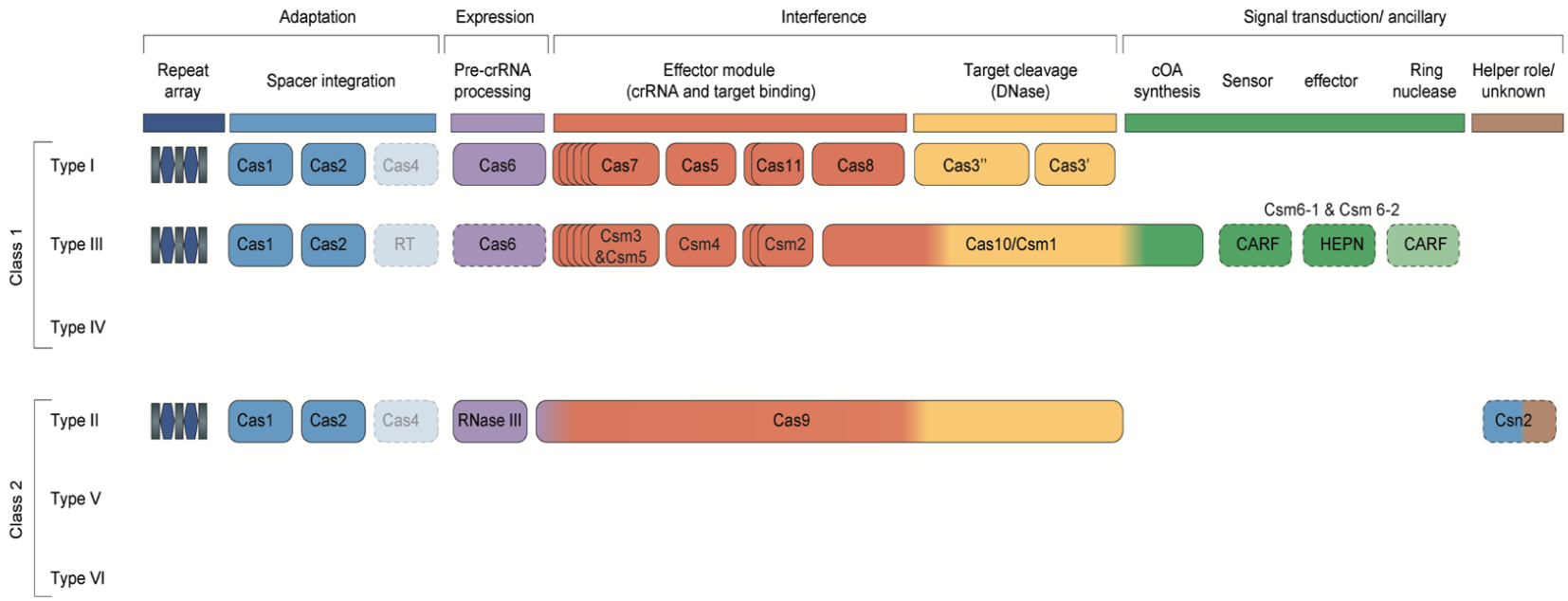

Adaptation
Expression
Interference
Signal transduction/ ancillary
Repeat array
Spacer integration
Pre-crRNA processing
Effector module (crRNA and target binding)
Target cleavage (DNase)
cOA synthesis
Sensor
effector
Ring nuclease
Helper role/ unknown
Class 1
Type I
Cas1
Cas2
Cas4
Cas6
Cas7
Cas5
Cas11
Cas8
Cas3''
Cas3'
Type III
Csm6-1 & Csm 6-2
Cas1
Cas2
RT
Cas6
Csm3 &Csm5
Csm4
Csm2
Cas10/Csm1
CARF
HEPN
CARF
Type IV
Class 2
Type II
Cas1
Cas2
Cas4
RNase III
Cas9
Csn2
Type V
Type VI

Figure 1.3 Model of Type III-A CRISPR-Cas interference

The Csm complex first binds to crRNA (red). Csm4 interacts with the 5' tag sequence of the crRNA (dark green), Csm3 and Csm2 form the backbone of the complex, and Csm5 caps off the 3' end of the crRNA and completes the complex. When the crRNP recognizes target RNA with a complementary sequence, Csm3 is able to cleave the target RNA at 6-nt intervals. Csm3 cleavage of the target RNA is independent of the 3' PFS of the target RNA (blue). However, If the 3' PFS of the target RNA is non-complementary to the 5' tag of the crRNA, then the 3' PFS flips out and interacts with Csm1, allosterically activating DNA cleavage and production of cOA. cOA goes on to activate Csm6 cleavage of RNA. Csm3 degradation of the target RNA returns the complex to an inactive state. Adapted from You et al., 2019 (31).

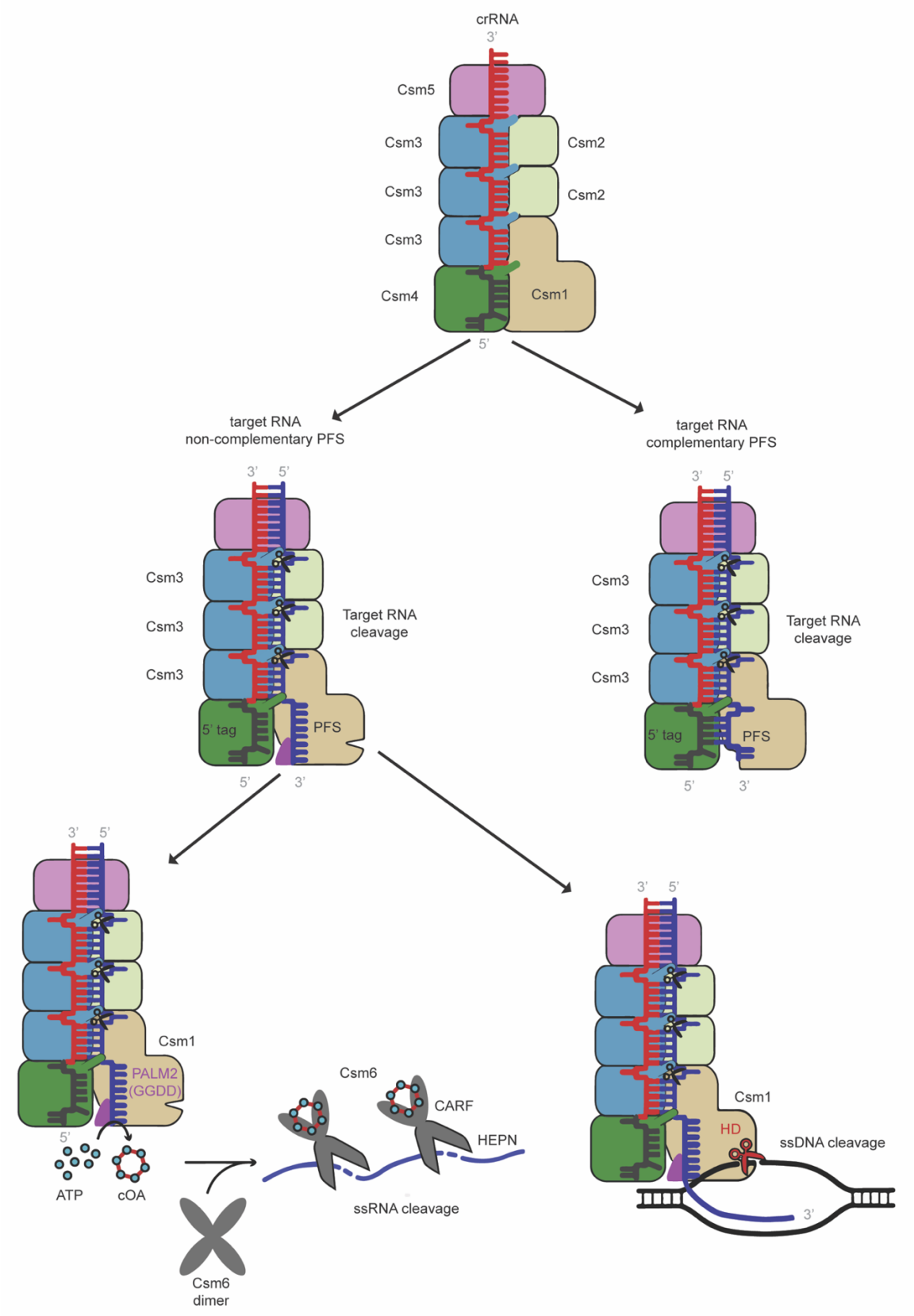
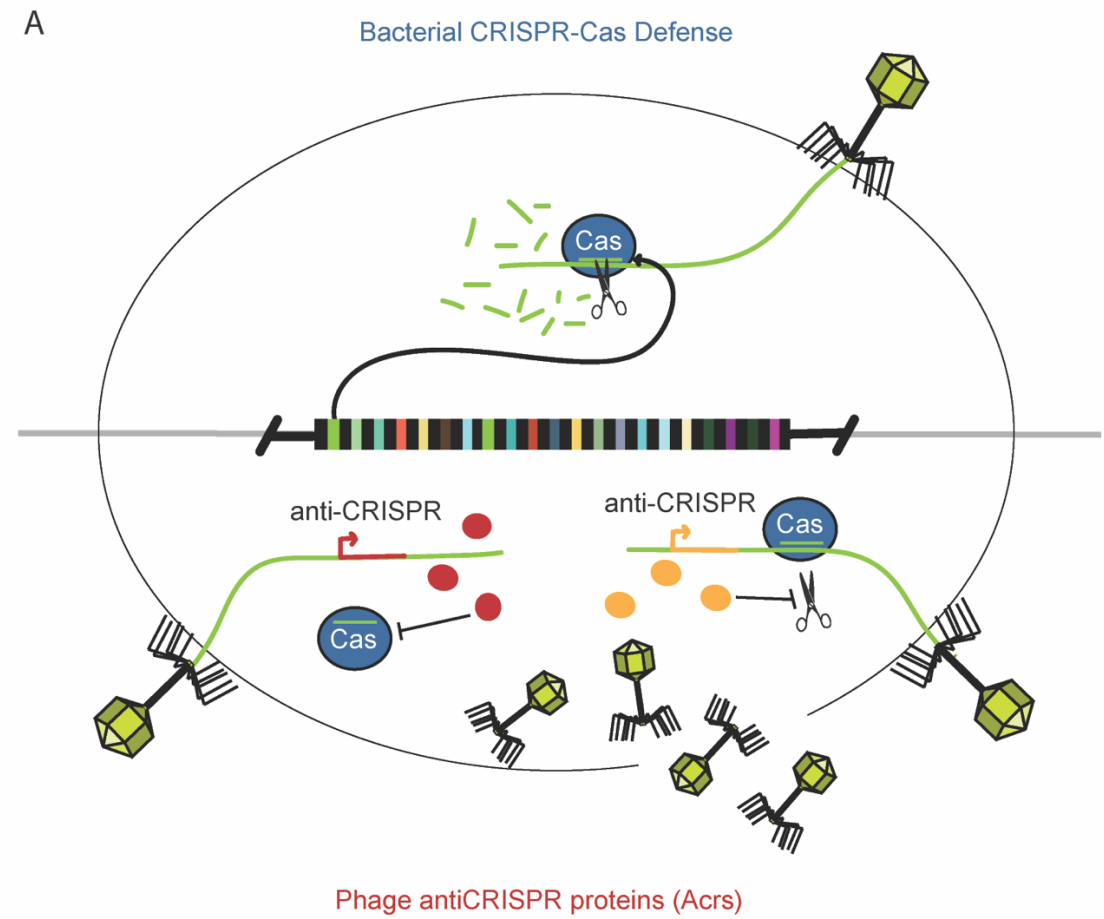

crRNA
3'
Csm5
Csm3
Csm2
Csm3
Csm2
Csm3
Csm4
Csm1
5'
target RNA
non-complementary PFS
target RNA
complementary PFS
3' 5'
Csm3
Csm3
Csm3
Target RNA
cleavage
5' tag
PFS
5'
3'
Csm1
PALM2
(GGDD)
ATP
cOA
Csm6
CARF
HEPN
ssRNA cleavage
Csm6
dimer
HD
ssDNA cleavage

Figure 1.4 Overview of phage anti-CRISPR proteins

(A) In response to bacterial adaptive immunity, phages and other mobile-genetic-elements encode anti-CRISPR proteins. The two broad mechanisms of anti-CRISPR proteins include (1) preventing the effector from binding to the target nucleic acid (red) or (2) inhibition of target nucleic acid cleavage once bound (yellow). **(B)** Identification of novel anti-CRISPR proteins is carried out via guilt-by-association where a known anti-CRISPR protein (Acr) or anti-CRISPR associated protein (Aca) is used to search for homologues in new phages. Neighboring genes are considered candidate Acrs. This is the strategy we use in our search for novel anti-CRISPRs in chapter 3.


A
Bacterial CRISPR-Cas Defense
Cas
anti-CRISPR
Cas
anti-CRISPR
Cas
Phage antiCRISPR proteins (Acrs)

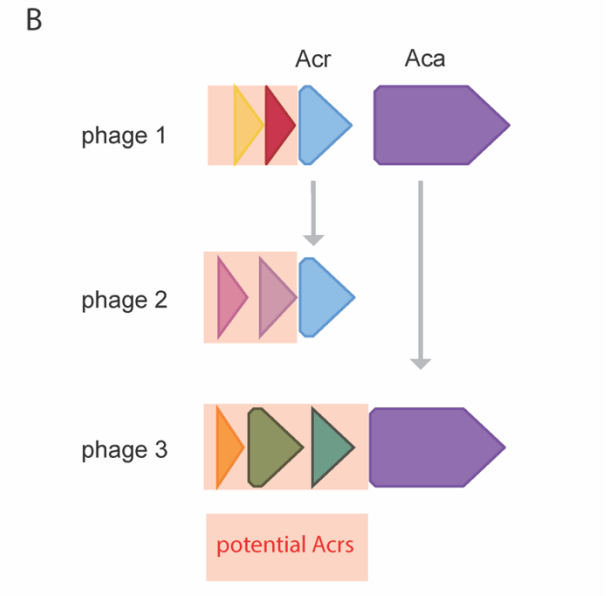
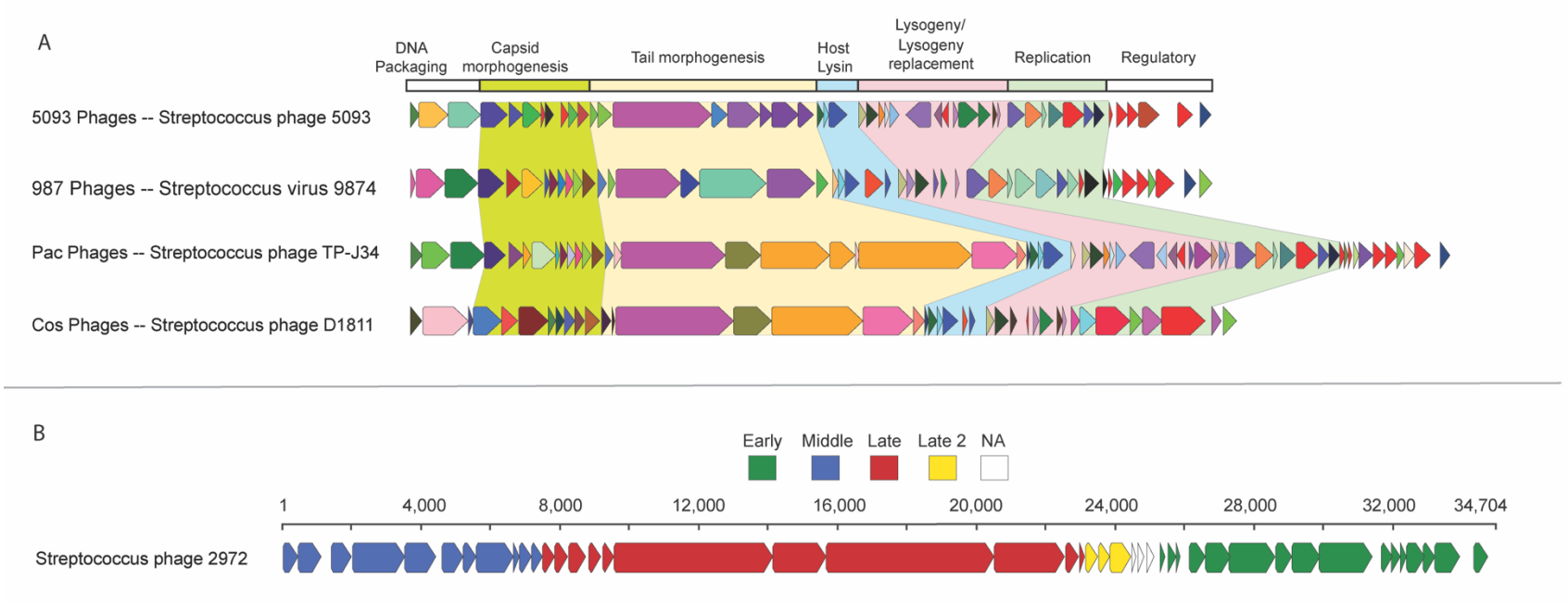

B
Acr
Aca
phage 1
phage 2
phage 3
potential Acrs

Figure 1.5 Modular architecture and temporal gene regulation of *Streptococcus thermophilus* phages

(A) The groups of *S. thermophilus* phages are cos, pac, 987, and 5093. A representative from each phage type is shown. Based on the published and annotated functions of several of the proteins within each module, the approximate module boundaries are denoted by background color. **(B)** Temporal gene expression of phage 2972 is adapted from Duplessis et al., 2005 (78).


A
DNA Packaging
Capsid morphogenesis
Tail morphogenesis
Host Lysin
Lysogeny/ Lysogeny replacement
Replication
Regulatory
5093 Phages -- Streptococcus phage 5093
987 Phages -- Streptococcus virus 9874
Pac Phages -- Streptococcus phage TP-J34
Cos Phages -- Streptococcus phage D1811
B
Early
Middle
Late
Late 2
NA
1
4,000
8,000
12,000
16,000
20,000
24,000
28,000
32,000
34,704
Streptococcus phage 2972

References

1. Mojica FJM, Díez-Villaseñor Cs, García-Martínez J, Soria E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. Journal of Molecular Evolution. 2005;60(2):174-82.

2. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences. 2012;109(39):E2579-E86.

3. 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(6096):816-21.

4. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nature Protocols. 2013;8(11):2281-308.

5. Cyranoski D, Ledford H. Genome-edited baby claim provokes international outcry. Nature. 2018;563(7733):607-8.

6. Jore MM, Brouns SJ, van der Oost J. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. Cold Spring Harbor perspectives in biology. 2012;4(6).

7. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nature Reviews Microbiology. 2010;8(5):317-27.

8. 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(5819):1709-12.

9. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence-and structure-specific RNA processing by a CRISPR endonuclease. Science. 2010;329(5997):1355-8.

10. 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(7340):602-7.

11. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes & development. 2008;22(24):3489-96.

12. Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;468(7320):67-71.

13. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nature Reviews Microbiology. 2020;18(2):67-83.

14. Koonin EV, Makarova KS. Origins and evolution of CRISPR-Cas systems. Philosophical Transactions of the Royal Society B: Biological Sciences. 2019;374(1772):20180087.

15. Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, van der Oost J. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science. 2016;353(6299):aad5147.

16. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, et al. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. Journal of bacteriology. 2008;190(4):1401-12.

17. Carte J, Christopher RT, Smith JT, Olson S, Barrangou R, Moineau S, et al. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in Streptococcus thermophilus. Molecular microbiology. 2014;93(1):98-112.

18. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. Embo j. 2013;32(3):385-94.

19. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009;155(Pt 3):733-40.

20. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature. 2010;463(7280):568-71.

21. Leenay RT, Maksimchuk KR, Slotkowski RA, Agrawal RN, Gomaa AA, Briner AE, et al. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol Cell. 2016;62(1):137-47.

22. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of bacteriology. 2008;190(4):1390-400.

23. Samai P, Pyenson N, Jiang W, Goldberg Gregory W, Hatoum-Aslan A, Marraffini Luciano A. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell. 2015;161(5):1164-74.

24. Goldberg GW, Jiang W, Bikard D, Marraffini LA. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature. 2014;514:633.

25. Hatoum-Aslan A, Maniv I, Samai P, Marraffini LA. Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system. Journal of bacteriology. 2014;196(2):310-7.

26. Staals RH, Zhu Y, Taylor DW, Kornfeld JE, Sharma K, Barendregt A, et al. RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Molecular cell. 2014;56(4):518-30.

27. Niewoehner O, Jinek M. Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. Rna. 2016;22(3):318-29.

28. Kazlauskiene M, Tamulaitis G, Kostiuk G, Venclovas Č, Siksnys V. Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. Molecular Cell. 2016;62(2):295-306.

29. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature. 2010;463(7280):568-71.

30. Mogila I, Kazlauskiene M, Valinskyte S, Tamulaitiene G, Tamulaitis G, Siksnys V. Genetic Dissection of the Type III-A CRISPR-Cas System Csm Complex Reveals Roles of Individual Subunits. Cell reports. 2019;26(10):2753-65 e4.

31. You L, Ma J, Wang J, Artamonova D, Wang M, Liu L, et al. Structure Studies of the CRISPR-Csm Complex Reveal Mechanism of Co-transcriptional Interference. Cell. 2019;176(1-2):239-53.e16.

32. Makarova KS, Aravind L, Wolf YI, Koonin EV. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biology direct. 2011;6(1):1-27.

33. Anantharaman V, Iyer LM, Aravind L. Presence of a classical RRM-fold palm domain in Thg1-type 3'-5'nucleic acid polymerases and the origin of the GGDEF and CRISPR polymerase domains. Biology direct. 2010;5(1):1-9.

34. Steitz TA, Yin YW. Accuracy, lesion bypass, strand displacement and translocation by DNA polymerases. Philos Trans R Soc Lond B Biol Sci. 2004;359(1441):17-23.

35. Zhang S, Li T, Huo Y, Yang J, Fleming J, Shi M, et al. Mycobacterium tuberculosis CRISPR/Cas system Csm1 holds clues to the evolutionary relationship between DNA polymerase and cyclase activity. International Journal of Biological Macromolecules. 2021;170:140-9.

36. Koonin EV, Makarova KS. Discovery of Oligonucleotide Signaling Mediated by CRISPR-Associated Polymerases Solves Two Puzzles but Leaves an Enigma. ACS Chem Biol. 2018;13(2):309-12.

37. Foster K, Kalter J, Woodside W, Terns RM, Terns MP. The ribonuclease activity of Csm6 is required for anti-plasmid immunity by Type III-A CRISPR-Cas systems. RNA Biology. 2018:null-null.

38. Jia N, Mo CY, Wang C, Eng ET, Marraffini LA, Patel DJ. Type III-A CRISPR-Cas Csm Complexes: Assembly, Periodic RNA Cleavage, DNase Activity Regulation, and Autoimmunity. Mol Cell. 2019;73(2):264-77.e5.

39. Makarova KS, Anantharaman V, Grishin NV, Koonin EV, Aravind L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. Frontiers in genetics. 2014;5:102.

40. Anantharaman V, Makarova KS, Burroughs AM, Koonin EV, Aravind L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. Biology Direct. 2013;8(1):15.

41. Pyenson NC, Gayvert K, Varble A, Elemento O, Marraffini LA. Broad Targeting Specificity during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. Cell host & microbe. 2017;22(3):343-53.e3.

42. Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature. 2013;493(7432):429-32.

43. Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, et al. Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. Nature. 2015;526(7571):136-9.

44. Shin J, Jiang F, Liu J-J, Bray NL, Rauch BJ, Baik SH, et al. Disabling Cas9 by an anti-CRISPR DNA mimic. Science advances. 2017;3(7):e1701620.

45. Pawluk A, Shah M, Mejdani M, Calmettes C, Moraes TF, Davidson AR, et al. Disabling a type IE CRISPR-Cas nuclease with a bacteriophage-encoded anti-CRISPR protein. MBio. 2017;8(6):e01751-17.

46. Harrington LB, Doxzen KW, Ma E, Liu J-J, Knott GJ, Edraki A, et al. A broad-spectrum inhibitor of CRISPR-Cas9. Cell. 2017;170(6):1224-33. e15.

47. Stanley SY, Borges AL, Chen KH, Swaney DL, Krogan NJ, Bondy-Denomy J, et al. Anti-CRISPR-Associated Proteins Are Crucial Repressors of Anti-CRISPR Transcription. Cell. 2019;178(6):1452-64.e13.

48. Birkholz N, Fagerlund RD, Smith LM, Jackson SA, Fineran PC. The autoregulator Aca2 mediates anti-CRISPR repression. Nucleic Acids Res. 2019;47(18):9658-65.

49. Borges AL, Davidson AR, Bondy-Denomy J. The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. Annual review of virology. 2017;4(1):37-59.

50. Pawluk A, Bondy-Denomy J, Cheung VHW, Maxwell KL, Davidson AR. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. mBio. 2014;5(2):e00896-e.

51. Hynes AP, Rousseau GM, Lemay ML, Horvath P, Romero DA, Fremaux C, et al. An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. Nature microbiology. 2017;2(10):1374-80.

52. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun. 2018;9(1):2919.

53. Song G, Zhang F, Zhang X, Gao X, Zhu X, Fan D, et al. AcrIIA5 Inhibits a Broad Range of Cas9 Orthologs by Preventing DNA Target Cleavage. Cell reports. 2019;29(9):2579-89 e4.

54. Liang M, Sui T, Liu Z, Chen M, Liu H, Shan H, et al. AcrIIA5 Suppresses Base Editors and Reduces Their Off-Target Effects. Cells. 2020;9(8).

55. Garcia B, Lee J, Edraki A, Hidalgo-Reyes Y, Erwood S, Mir A, et al. Anti-CRISPR AcrIIA5 Potently Inhibits All Cas9 Homologs Used for Genome Editing. Cell reports. 2019;29(7):1739-46 e5.

56. Maxwell KL. The Anti-CRISPR Story: A Battle for Survival. Mol Cell. 2017;68(1):8-14.

57. Garcia-Albiach R, Pozuelo de Felipe MJ, Angulo S, Morosini MI, Bravo D, Baquero F, et al. Molecular analysis of yogurt containing Lactobacillus delbrueckii subsp. bulgaricus and Streptococcus thermophilus in human intestinal microbiota. The American journal of clinical nutrition. 2008;87(1):91-6.

58. Achigar R, Magadán AH, Tremblay DM, Julia Pianzzola M, Moineau S. Phage-host interactions in Streptococcus thermophilus: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. Scientific Reports. 2017;7:43438.

59. Mc Grath S, Fitzgerald GF, van Sinderen D. Bacteriophages in dairy products: pros and cons. Biotechnology Journal: Healthcare Nutrition Technology. 2007;2(4):450-5.

60. Magadán AH, Dupuis M-È, Villion M, Moineau S. Cleavage of Phage DNA by the Streptococcus thermophilus CRISPR3-Cas System. PLoS One. 2012;7(7):e40913.

61. Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V. crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. RNA biology. 2013;10(5):841-51.

62. Quiberoni A, Moineau S, Rousseau GM, Reinheimer J, Ackermann H-W. Streptococcus thermophilus bacteriophages. International Dairy Journal. 2010;20(10):657-64.

63. Mahony J, Van Sinderen D. Current taxonomy of phages infecting lactic acid bacteria. Frontiers in microbiology. 2014;5:7.

64. Le Marrec C, van Sinderen D, Walsh L, Stanley E, Vlegels E, Moineau S, et al. Two groups of bacteriophages infecting Streptococcus thermophilus can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. Applied and Environmental Microbiology. 1997;63(8):3246-53.

65. Mills S, Griffin C, O'Sullivan O, Coffey A, McAuliffe O, Meijer W, et al. A new phage on the 'Mozzarella'block: bacteriophage 5093 shares a low level of homology with other Streptococcus thermophilus phages. International dairy journal. 2011;21(12):963-9.

66. McDonnell B, Mahony J, Neve H, Hanemaaijer L, Noben J-P, Kouwen T, et al. Identification and analysis of a novel group of bacteriophages infecting the lactic acid bacterium Streptococcus thermophilus. Applied and environmental microbiology. 2016;82(17):5153-65.

67. Lucchini S, Desiere F, Brüssow H. Comparative genomics of Streptococcus thermophilus phage species supports a modular evolution theory. J Virol. 1999;73(10):8647-56.

68. Neve H, Zenz KI, Desiere F, Koch A, Heller KJ, Brüssow H. Comparison of the Lysogeny Modules from the TemperateStreptococcus thermophilusBacteriophages TP-J34 and Sfi21: Implications for the Modular Theory of Phage Evolution. Virology. 1998;241(1):61-72.

69. Lavelle K, Murphy J, Fitzgerald B, Lugli GA, Zomer A, Neve H, et al. A decade of Streptococcus thermophilus phage evolution in an Irish dairy plant. Applied and environmental microbiology. 2018;84(10):e02855-17.

70. Lavelle K, Martinez I, Neve H, Lugli GA, Franz CM, Ventura M, et al. Biodiversity of Streptococcus thermophilus phages in global dairy fermentations. Viruses. 2018;10(10):577.

71. Lucchini S, Desiere F, Brüssow H. Comparative genomics of Streptococcus thermophilus phage species supports a modular evolution theory. Journal of Virology. 1999;73(10):8647-56.

72. Duplessis M, Michael Russell W, A Romero D, Moineau S. Global gene expression analysis of two Streptococcus thermophilus bacteriophages using DNA microarray2005. 192-208 p.

73. Ventura M, Foley S, Bruttin A, Chennoufi SC, Canchaya C, Brussow H. Transcription mapping as a tool in phage genomics: the case of the temperate Streptococcus thermophilus phage Sfi21. Virology. 2002;296(1):62-76.

74. Ventura M, Brüssow H. Temporal transcription map of the virulent Streptococcus thermophilus bacteriophage Sfi19. Appl Environ Microbiol. 2004;70(8):5041-6.

75. Artamonova D, Karneyeva K, Medvedeva S, Klimuk E, Kolesnik M, Yasinskaya A, et al. Spacer acquisition by Type III CRISPR–Cas system during bacteriophage infection of Thermus thermophilus. Nucleic Acids Research. 2020;48(17):9787-803.

76. Mo CY, Mathai J, Rostøl JT, Varble A, Banh DV, Marraffini LA. Type III-A CRISPR immunity promotes mutagenesis of staphylococci. Nature. 2021;592(7855):611-5.

77. Terns RM, Terns MP. CRISPR-based technologies: prokaryotic defense weapons repurposed. Trends Genet. 2014;30(3):111-8.

78. Duplessis M, Russell WM, Romero DA, Moineau S. Global gene expression analysis of two Streptococcus thermophilus bacteriophages using DNA microarray. Virology. 2005;340(2):192-208.

CHAPTER 2

COMPOSITION AND IMMUNITY OF TYPE III-A CRISPR-CAS SYSTEMS IN *STREPTOCOCCUS THERMOPHILUS*

1

[1]Clare Cooper, Walter T. Woodside, Shakela Mitchell, and Michael P. Terns. Composition and immunity of Type III-A CRISPR-Cas systems in *Streptococcus thermophilus*. *In preparation.*

ABSTRACT

Type III-A CRISPR-Cas systems provide prokaryotes with adaptive immunity against plasmids, phages, and mobile genetic elements through targeted transcript degradation as well as non-specific DNAse and RNAse activities. The lactic acid bacterium *Streptococcus thermophilus* harbors up to four distinct CRISPR-Cas systems, with each system maintaining its own CRISPR array and Cas proteins required for new spacer uptake, crRNA biogenesis, and foreign nucleic acid destruction. While much is known about the type II-A CRISPR-Cas systems of *S. thermophilus*, the type III-A systems remains more elusive despite their widespread occurrence in *S. thermophilus* strains. Here we present an up-to-date analysis of the type III-A CRISPR-Cas systems present in published genome assemblies of *S. thermophilus* and compare their properties to the other co-existing Type II-A and I-E systems. Additionally, we determine the Type III-A nuclease requirements for anti-phage and anti-plasmid immunity in the native host to better understand the defense capabilities of these systems in *Streptococcus thermophilus* and other prokaryotes.

INTRODUCTION

Streptococcus thermophilus is a lactic acid bacterium used in production of dairy products (1, 2). Lytic phages that infect *S. thermophilus* can contaminate starter cultures and negatively impact production (3). Therefore, understanding bacterial defense mechanisms against phage infection is important to industry success (3). Strains of *S. thermophilus* can encode up to four CRISPR-Cas systems, with the Type II-A (CR1) locus and Type II-A (CR3) locus being the most well-characterized due to their role in spacer acquisition and targeting during lytic phage infection (4, 5). However, the Type III-A and I-E systems

could play other roles within these strains, and the utility of these co-occurring CRISPR-Cas systems has not been extensively studied.

The first analysis of CRISPR-Cas systems in strains of *S. thermophilus* by Horvath et al. included many proprietary strains, but determined that Type III-A (or Csm) systems are widespread despite their limited number of spacer sequences (6). The conclusion of studies since is that Type III-A systems in *S. thermophilus* are largely non-functional in defense against phage and mobile genetic elements (7). This conclusion is primarily because co-occurring *S. thermophilus* Type II-A systems (CR1 and CR3) have been shown to preferentially acquire spacers during phage infection (8). However, there is an expanding number of anti-CRISPR proteins that inhibit Type II-A CRISPR-Cas systems, including two proteins identified in *S. thermophilus phages (9, 10)*. For this reason, the defense provided by Type III-A CRISPR-Cas systems may be more relevant when strains are challenged with phages carrying these anti-CRISPR proteins.

We aimed to analyze the current repertoire of *S. thermophilus* genomes to (1) verify the prevalence of Type III-A systems first reported by Horvath et al. (6), (2) better understand the repeat and spacer targeting features of Csm systems compared to co-occuring CRISPR-Cas systems in *S. thermophilus*, (3) determine the minimum nuclease requirements of Type III-A CRISPR-Cas systems in *S. thermophilus* in both phage and plasmid defense contexts, and (4) outline protein makeup and functional predictions for each of the identified *S. thermophilus* Csm systems. We hope that this work gives greater context to the role of Type III-A CRISPR-Cas systems in *S. thermophilus* and defines the current state of Csm systems in published genome assemblies.

RESULTS

Cooccurrence of CRISPR-Cas systems in *Streptococcus thermophilus*

To understand the status of type III-A CRISPR-Cas systems in strains of *Streptococcus thermophilus*, we wanted to identify all systems across *S. thermophilus* strains whether complete or fragmented. For this reason, we utilized a genome neighborhood approach to extract and characterize systems not identified by other prediction methods. To do this, we carried out initial CRISPR-Cas system predictions using CRISPRdetect (11). After mapping these predictions onto all genomes, we located conserved proteins up and downstream of each CRISPR system. We used these conserved proteins to extract defined loci from all studied genome assemblies. This approach gives a full view of the current state of CRISPR-Cas systems in *S. thermophilus* and additionally illuminates the potential degradation of CRISPR-Cas systems in these strains. Consistent with past analyses, we identified four unique CRISPR-Cas systems in *S. thermophilus* (Figure 2.1A) *(6, 7)*.

One recent analysis estimated the percentage of published *S. thermophilus* strains with type III-A systems to be 48% (13/27) (7). However, consistent with the previous work of Horvath et al., we found greater than 90% (68/73) of published *S. thermophilus* strains to contain a Type III-A array (Figure 2.1B). The majority of *S. thermophilus* strains encode more than one CRISPR-Cas system (Figure 2.1A and 2.1B), with the Type II-A (CR1) system being the most abundant (100%) followed by the Type III-A system (90%), the Type II-A (CR3) system (67%), and the Type I-E system (18%).

Array sizes by CRISPR-Cas system

Across the 68 strains with Type III-A arrays, Figure 2.1C and Figure 2.1D show that there is an average repeat count of 4.9 ±3.4 (SD). In comparison, the cooccurring Type II-A systems maintain larger CRISPR arrays with mean repeat counts of 30.0±17.0 and 18.9±9.1 for CR1 and CR3, respectively (Figure 2.1E and 2.1F). The type I-E system shows less variability with a mean repeat count of 12.6±3.4 (Figure 2.1G). This difference in repeat counts between systems indicates that the spacer acquisition and/or spacer loss profiles or perhaps roles of these system types in *S. thermophilus* may vary considerably. This is consistent with the finding that spacers in the well characterized *S. thermophilus* DGCC 7710 strain are preferentially taken up into the Type II-A (CR1) and Type II-A (CR3) arrays during challenge with phage (8, 12, 13). Notably, this preferential adaptation may vary between strains as there are 9 strains with more than double the average number of repeats in their Type III-A arrays. The individual repeat counts for the CRISPR-Cas systems of each analyzed genome assembly are shown in Table 2.1.

Spacer sequence length by CRISPR-Cas system

From an adaptation standpoint, Type II-A systems have previously been shown to have conserved spacer sequence lengths (8, 14). In Figure 2.2, we verified this for the *S. thermophilus* systems, with our results consistent with previously published *in vivo* and *in vitro* assays of spacer acquisition (8, 14). The Type II-A (Figure 2.2B and 2.2D) and I-E (Figure 2.2D) systems preferentially maintain spacers that are 30 base pairs (bp) with the Type I-E system having some spacers that are 31 bp. We wanted to know if our Type III-A dataset has a similar spacer sequence length constraint. In contrast to the other systems,

the spacers do not conform to a single length, but do center around 36 base pairs with ~40% of spacers at that length (Figure 2.2A). Compared to the other systems, there is a wider range of spacer sizes, perhaps indicating more flexibility in the spacer acquisition and array insertion process. The constraints of the protospacer adjacent motif (PAM) sequences for Type II-A and I-E systems could play a role in spacer sequence length constraints as well (15, 16).

PFS and PAM sequence Analysis

For type II-A and I-E systems, an additional measure of spacer functionality is the identity of the protospacer adjacent sequence (PAM) (15, 16). The PAM sequence helps differentiate a target sequence from the same spacer sequence encoded within the CRISPR-array to prevent cleavage of the host chromosome (17, 18). Moreover, PAM recognition by specific components of the crRNP effector complexes (e.g. Cas9 for type II-A and Cas8 for type I-E) is a key initial step required for destroying plasmid and phage invasive DNA (19, 20). While the importance of specific protospacer flanking sequences (PFS) for function has been demonstrated for some Type III systems such as the *Pyrococcus furiosus* III-B system (21, 22), no such PFS requirement for function has been reported for the III-A system in *S. thermophilus* and other Type III-A systems (23, 24). Previous studies found that activation of the Csm1 DNase (HD) and cyclase (PALM) domains occurs when the PFS does not pair with the 5' tag of the crRNA, so there is some sequence preference needed for Csm1 and Csm6 dependent defense (25).

To query for a defined PFS sequence, we carried out protospacer mapping and PFS alignment. We searched for all Type III-A spacer sequences against the Refseq viral

database and only proceeded with hits with a 100% query cover and 90% or greater identity to the spacer sequence. This produced a dataset of 131 total hits or 89 unique hits when the 10 bp up and downstream are considered. For both total (not shown) and unique hits, the aligned flanking sequence from 10 bp up and downstream of this protospacer sequence does not demonstrate significant enrichment of sequences non-complementary to the crRNA 5' tag. There is only minor preference for 'U' in these flanking regions (Figure 2.2E).

Additionally, Figure 2.2E shows the same analysis we performed for the type II-A and I-E systems. Again, a 100% query cover cut-off was used for spacer hits. In addition, A 100% identity cutoff was used for Type II-A spacer hits while a 90% identity was used for Type I-E system hits due to fewer spacers available for analysis. We identified a type II-A (CR1) PAM sequence 3' to the protospacer as 3'-NNANAAW-5'. The PAM sequence for CR1 has shown variability dependent on prediction methods, but previously has been characterized as 3'-NNAGAAW-5' (16, 26). For the type II-A (CR3) system, we identified the 3' PAM consensus sequence of 5'-NGGNG-3' which is consistent with previous analyses (16, 26). The type I-E PAM for *S. thermophilus* has previously been characterized (19) and is also supported by our analysis to be 5'-AA-3', located 5' to the protospacer sequence.

Orientation bias of protospacers

Type III systems rely on pairing between crRNA and target RNA to activate defense, so transcription is required for target cleavage (25, 27). For this reason, the orientation of a protospacer determines functionality as one orientation will be transcribed while the other

will likely not. To identify if *S. thermophilus* type III-A protospacers have an orientation bias, unique spacer sequences were mapped to coding regions within phage genomes. The orientation of the coding sequence hits were compared across all CRISPR-Cas system types and are shown in Figure 2.2F.

Protospacer mapping indicates that Type III-A hits are predominantly complementary to the coding strand of DNA and capable of pairing with target RNA (Figure 2.2F). Out of 83 unique type III-A spacer hits, 94% were complementary to the coding strand with only 6% complementary to the template strand. Type II-A and I-E systems target DNA, so we expected hits to readily map in both orientations. Consistent with this, the type I-E system showed little orientation bias with 45% of spacer hits complementary to the coding strand and 55% complementary to the template strand. The Type II-A systems of CR1 and CR3 demonstrate orientation bias opposite of the type III-A system, with 67% and 70% of spacers oriented for template strand pairing, respectively. This preference is explained by a higher proportion of PAM sequences on the coding strand than the template strand, which we demonstrate for our phage dataset in Chapter 4.

Nuclease requirements for Type III-A anti-plasmid immunity of *S. thermophilus* JIM 8232

Using the Type III-A system from the JIM 8232 strain of *S. thermophilus* expressed in E. coli, our group previously determined that Csm6 RNase activity is required for anti-plasmid immunity (28) In contrast, the Csm1 Dnase activity was not essential for robust anti-plasmid activity (28). To better predict functionality of *S. thermophilus* Type III-A systems, we wanted to understand the minimal nuclease requirements in the native host.

Notably, two copies of Csm6 are present in *S. thermophilus* JIM 8232, so we wanted to determine if Csm6-1 and Csm6-2 are redundant or play separate roles in defense.

To assay Type III-A anti-plasmid immunity in *S. thermophilus* JIM 8232, we used two plasmids for transformations. Both plasmids carry the transcribed target sequences of spacers 1-3 of the Type III-A array, but they vary in the PFS. The control plasmid (pControl) carries a PFS sequence complementary to the 5' tag of the crRNA, making it incapable of activating Csm1 and Csm6 activity. The pTarget plasmid has a PFS that is partially non-complementary to the 5' tag, and it is a sequence we previously determined to strongly activate Csm complex activity (our unpublished findings). We carried out natural transformation of pTarget and pControl plasmids and quantified defense activity as transformation efficiency.

To assay the contribution of Csm1 DNase activity to anti-plasmid immunity, we mutated the Csm1 HD DNase active site (HD to AA) (Figure 2.3A and B). In Figure 2.3C, we compare mutations of the Csm system of *S. thermophilus* JIM 8232 to wildtype as well as a Csm1-Csm6 knockout strain (this strain is indicated by Δ* in the figure). We found that DNase activity of Csm1 contributes minimally to anti-plasmid immunity with less than a log of difference in transformation efficiency between pTarget and pControl (Figure 2.3C).

To determine if both copies of Csm6 are required for anti-plasmid immunity, we generated single and double Csm6 knock-out strains as well as single and double mutations of the HEPN ribonuclease active site (H to A) (Figure 2.3A and B). Interestingly, activity provided by either Csm6 protein is sufficient to lead to anti-plasmid immunity equivalent to that of the wildtype strain (Figure 2.3C). Loss of both Csm6-1 and Csm6-2 RNase

activity (via HEPN mutation) abolishes anti-plasmid immunity with transformation efficiency equivalent to the control Csm1-6 knock-out strain (denoted by Δ *) (Figure 2.3C). This indicates that the ribonuclease activity of either Csm6-1 or Csm6-2 protein is necessary for robust anti-plasmid immunity.

Nuclease requirements for Type III-A anti-phage immunity in *S. thermophilus* JIM 8232

Next, we sought to define the type III-A nuclease requirements for anti-phage defense for the JIM 8232 strain of *S. thermophilus* (Figure 2.4). However, no phage was available that could infect this particular strain. In contrast, S. thermophilus DGCC7710 and lytic phage 2972 are a well-established host-phage system, but our unpublished results showed that the III-A system of the *S. thermophilus* DGCC 7710 strain is unable to defend against a target phage. We previously hypothesized that this was likely because both Csm6-1 and Csm6-2 genes of the *S. thermophilus* DGCC 7710 strain are predicted to be nonfunctional (Csm6-1 loss due to a premature stop codon as well as truncation of the Csm6-2 gene). To overcome this, we replaced the inactive type III-A system of *S. thermophilus* DGCC 7710 with the active III-A system from *S. thermophilus* JIM 8232, fully transplanting all adaptation and effector proteins, as well as the type III-A array. To compare anti-viral defense activity across mutant strains, we performed growth curves to track host cell growth (CRISPR defense) or lysis (lack of CRISPR defense) (Figure 2.4). Strains capable of defending against phage should continue to grow with increasing turbidity while strains incapable of defense will lyse and become clear.

We generated two mini-array vectors for our phage-based assay. The pDefend vector contains the native Type III-A leader sequence of *S. thermophilus* JIM 8232 upstream of the CRISPR array to drive expression of two spacers that target early expressed phage genes of phage 2972 (the mapped spacers are shown in Figure 2.4A). The miniarray of pControl contains native array spacers that do not target phage 2972.

S. thermophilus DGCC 7710 with and without the transplanted Type III-A system was transformed with the pDefend and pControl plasmids. We challenged both strains with phage 2972 at three different multiplicities of infection (MOI), including 0.1, 1, and 10 and recorded growth curves. The wildtype *S. thermophilus* DGCC 7710 with native Type III-A system (Figure 2.4B) lysed even when provided with a functional spacer against the phage. In contrast, the *S. thermophilus* DGCC 7710 strain with the transplanted Type III-A system (Figure 2.4C) did not lyse when carrying the pDefend mini-array vector, but did lyse with the pControl vector, indicating that the type III-A system derived from the *S. thermophilus* JIM 8283 strain is able to utilize the spacers to defend against the phage.

To determine if Csm6 activity was important for anti-phage immunity, we analyzed the effect of inactivating Csm6-1 and Csm6-2 RNase activity (HEPN active site mutation of H to A). Compared to the lack of defense against phage lysis demonstrated by the type III-A system of *S. thermophilus* DGCC 7710, the Csm6-1/Csm6-2 HEPN mutant was capable of defense at an MOI of 0.1 (Figure 2.4D). This indicates that the Csm complex, with the DNase activity and target RNase activity of Csm1 and Csm3, respectively, is capable of defense in the absence of Csm6 nuclease activity, but only at a low MOI (0.1).

To test if the Csm1 DNase activity is important for anti-phage immunity, we mutated the Csm1 HD nuclease active site in the type III-A transplant strain (Figure 2.4E).

Mutation of the HD motif weakened immunity at an MOI of 0.1 and 1, and abolished immunity at an MOI of 10. Notably, the defense is weaker than wildtype, but indicates that DNase activity is not required for anti-phage defense, at least in the context of type III-A spacers targeting phage early genes.

Prediction of Type III-A defense activity in strains of *S. thermophilus*

To better predict the functionality of the Type III-A systems in *S. thermophilus*, we wanted to characterize the Csm protein makeup of each strain. To do this, we annotated all Csm proteins identified in our type III-A neighborhood analysis. For all Type III-A genome neighborhoods, we used Prodigal to predict additional open reading frames (ORFs) that may have been missed during genome annotation (29). These are indicated in the table as 'manually annotated.'

Csm proteins predicted to be functional are highlighted green in Table 2.2. Proteins that are absent are denoted by '--' and those that are predicted to be non-functional are colored white and have a note on the reason why. Csm1 functional predictions were based on Pfam domain and motif annotations of the following: (1) HD DNase domain and motif (HD, PF01966), (2) Csm1_B, (PF18211) domain which corresponds to domain B of Thermococcus onnurineus Csm1(30) or the PALM1 domain, and (3) the GGDD (cyclase) motif which was manually annotated. Csm2 – Csm5 annotations and functional predictions were based on alignments of all protein sequences to identify variants that are truncated in addition to Pfam family predictions. Cas1, Cas2 and Cas6 were analyzed in the same way, with nucleotide sequence alignments used to verify annotated mutations in the coding sequence. The annotations of Csm6-1 and Csm6-2 were highly variable. For all strains, the

encoding region of Csm6-1 and Csm6-2 was aligned and compared to the same region in *S. thermophilus* JIM 8232. Prodigal gene predictions were used to manually annotate Csm6-1 (if not annotated) so we could determine if Csm6-1 has since been truncated and lost functionality (29).

We used the functional predictions for each protein to determine the adaptation and defense capabilities of the Type III-A system for each strain. Adaptation capability was predicted based on the presence of Cas1 and Cas2 as well as a Type III-A array. Because of the broad range of defense activities of type III-A effector proteins, we divided the overall defense functionality prediction into two categories. We predicted the defense capability of the Csm complex, made up of Csm1-Csm5 as well as Cas6 which is needed for crRNA biogenesis. We then considered the ancillary RNase activity of Csm6-1 and Csm6-2 downstream.

Of the 68 strains with Type III-A arrays, 8 (12%) of the strains have a bacteriocin module intervening in the Csm gene locus (denoted by blue in Table 2.2). This module has replaced Cas6 through to part of Csm 6-1 in these strains. For this reason, these strains were not further included in the overall functional analysis. Out of the remaining 60 strains with Type III-A arrays, 40 (67%) contain Cas1 and Cas2 proteins predicted to be capable of adaptation. In terms of predicted activity of the Csm complex, including Csm1 and Csm3 effector nucleases, 49 (82%) strains have predicted activity. *S. thermophilus* DGCC 7710 is one of the strains with predicted effector complex activity. Beyond activity of the effector complex, like the DGCC 7710 strain, most of the *S. thermophilus* strains encode truncated variants of Csm6-1 and Csm6-2, indicating that these proteins were once present but have likely lost function. Only 6 strains (10%) have both copies of Csm6 intact, including *S.*

thermophilus JIM 8232. Interestingly, 6 strains (10%) only encode a single protein in the region of Csm6-1 and Csm6-2 and are denoted by a single cell with a lighter green shade in Table 2.2.

DISCUSSION

In figure 2.4, we demonstrated that the type III-A system of *S. thermophilus* DGCC 7710 is inactive in anti-phage defense when provided with spacers against ϕ2972. In contrast, the type III-A system of *S. thermophilus* JIM 8232 is able to mount a strong anti-phage defense. Previously, we attributed a lack of anti-plasmid immunity in *S. thermophilus* DGCC 7710 to truncations of the Csm6-1 and Csm6-2 genes. However, when studying the effect of mutations of the Type III-A system of *S. thermophilus* JIM 8232, we saw weak defense against phage even when the Csm6-1 and Csm6-2 HEPN domain is mutated. Perhaps this difference is because of the potential RNA-binding activity of *S. thermophilus* JIM 8232 Csm6-1 and Csm6-2 being retained when the HEPN domain is mutated. While there was no difference in activity of the HEPN double mutant and Csm6-1 and Csm6-2 deletion mutant in the plasmid-based assay, it may be worthwhile to test phage defense activity with complete Csm6 deletions to account for this. The possibility of an anti-CRISPR or another inhibitor of the Type III-A system present in *S. thermophilus* DGCC 7710 is unlikely as the system of *S. thermophilus* JIM 8232 is functional when transplanted into this strain.

While our plasmid and phage experiments to date do not indicate that the *S. thermophilus* DGCC 7710 Csm complex is capable of phage or plasmid defense, we cannot be sure that Csm complex formation or Csm3 RNase activity are absent in *S. thermophilus* DGCC 7710. Future experiments should consider lower MOIs and phage plaque assays to

determine if there is some minimal activity of the complex. Chapter 4 further delves into the potential roles that the Type III-A complex could have in defense outside of the Figure 2.3 and 2.4 assays for complete anti-phage and anti-plasmid defense activity.

MATERIALS AND METHODS

S. thermophilus strains and CRISPR-Cas system identification

All available 'complete' and 'chromosome' genome assemblies of *Streptococcus thermophilus* were downloaded from NCBI on 10/13/20. Duplicate assemblies were discarded for a total of 73 strains analyzed (Table S1). CRISPR arrays were initially identified using a local iteration of CRISPRdetect and were edited and refined manually into subtypes based on repeat sequence and protein makeup (11). Once the genome neighborhood was identified for each CRISPR system type, manual extraction of CRISPR encoding regions was performed. Previously unidentified systems were added to the analysis.

PFS and PAM Sequence Analysis

Local BLAST iterations were completed in Geneious Prime 2019.2.3 (https://www.geneious.com) (31). All spacer sequences were searched against the Refseq Viral Database (downloaded on 10.29.20) using BLASTn with parameters adjusted to CRISPR-target recommended values of match (+1), mismatch (-1), word size (7), E-value (1), and filter (yes) (32). Due to parameters of BLAST in Geneious, gap cost to open (-5) and extend (-2) deviated from that of CRISPR-target. Results above a bit score of 40 were considered significant and were used for further analysis.

Strain Culturing

S. thermophilus cultures were grown in LM17 (Himedia). For liquid overnight growth and natural transformation, strains were grown at 37ºC. For overnight growth on agar plates, strains were grown at 42ºC. When indicated, chloramphenicol 5 ug/mL and/or erythromycin 10 ug/mLwas utilized. For assays with phage 2972, 10 mM $CaCl_2$ was supplemented. E. coli was grown in Luria broth with chloramphenicol 10 ug/mL or erythromycin 200 ug/ml at 37ºC.

Target vector cloning

All target sequences were cloned into a pWAR vector backbone with a chloramphenicol (Cm) resistance marker. Transcribed targets were expressed downstream of a constitutive Ppgm promoter sequence with a rho-independent terminator sequence. Target vectors were created using Gibson Assembly with Twist fragments encoding the targets with the anti-tag or strongly activating PFS sequences. Target vectors were maintained in *E. coli* Top 10.

Mini-array vector cloning

Mini-arrays were cloned into a pWAR vector backbone with a chloramphenicol (Cm) resistance marker. Type III-A spacer sequences for targeting phage 2972 were selected by BLAST of all *S. thermophilus* Type III-A spacer sequences against phage 2972. followed by selection of two spacers with the greatest homology. These spacers are characterized in

figure 4. Cloning of miniarrays with repeat sequences required use of the methods used in (33). Miniarray vectors were maintained in *E. coli* Top 10.

Phage amplification

Phage 2972 was amplified on *S. thermophilus* DGCC 7710 in LM17 supplemented with 10 mM $CaCl_2$. The amplification was carried out at 42ºC, shaking. The phage underwent two rounds of amplification. Both amplifications were filtered using a 0.45 um polyethersulfone (PES) filter. Phage was stored at 4ºC for short term storage or in 50% glycerol at -80ºC for long term storage.

Phage titration

Overnight cultures of *S. thermophilus* (OD_{600} = 0.6) were diluted in melted LM17 0.75% 'top' agar supplemented with 10 mM $CaCl_2$ (300 uL of cells to 3 mL of agar). Stocks of phage 2972 were serially diluted in phage buffer before being co-inoculated (100 uL of each dilution to 3.3 mL). The phage and strain agar mix were then poured over the top of LM17 plates supplemented with 10 mM $CaCl_2$ and antibiotics where indicated and allowed to set before being incubated at 42ºC overnight. Plaques were counted on plates with between 30 and 300 plaques to calculate the plaque forming units (PFU) of the phage stock.

Anti-Plasmid Defense Assay

Strains of *S. thermophilus* were grown overnight at 37ºC shaking. The following morning, 1mL of overnight culture was spun down (5,000 x g, 2 mins) and washed with chemically defined media (CDM) (34, 35). Two wash steps were completed and then each strain was

diluted 50% in a 96-well plate (100 uL total) for OD read. Based on OD_{600} measurement, strains were all diluted to OD600 = 0.1 in and incubated for 1 hour at 37ºC standing. Following incubation, preparations of naturally competent cells were frozen at -20ºC before use as well as for long term storage. When ready for use, competent cells were thawed at room temperature. Plasmid DNA (200 ng) was added to the wells of a 96 well-plate and naturally competent cells were combined with 1uM ComS peptide (34, 35). 100 uL of cells were then added to each well and incubated at 37ºC standing. After 3 hours, each well of the 96-well plate was resuspended prior to plating 30 uL to LM17 agar plates with indicated antibiotic selection. Plates were incubated overnight at 42ºC prior to imaging using a Biorad Gel doc XR.

Anti-Phage Defense Assay

Following overnight growth, strains of *S. thermophilus* were diluted 75% in LM17. 100 uL of each dilution was added to the wells of a 96-well plate for OD_{600} measurement on a Biotek Epoch 2 microplate reader. Based on this measurement, overnight cultures were diluted to OD_{600} = 0.2 and supplemented with 10 mM CaCl2. Based on plaque forming unit (PFU) and desired multiplicity of infection (MOI), the appropriate volume of phage 2972 (or phage buffer control) was inoculated into the wells of a fresh 96-well plate. To each well, 200 uL of diluted overnight culture was added. Growth was carried out on the plate reader at 42ºC with double orbital shaking. Measurements of OD_{600} were taken every 5 minutes for 24 hours. Growth curves were analyzed using Prism 8.4.1.

Figure 2.1. **Type III-A CRISPR-Cas systems are widespread in *S. thermophilus* and typically have a relatively low number of CRISPR repeats per array.**

(A) Depiction of the four CRISPR-Cas systems of *S. thermophilus*. The representative shown for the Type III-A system is that of *S. thermophilus* JIM 8232. The Type II-A and I-E systems are from *S. thermophilus* DGCC 77710. All systems are color coded to correspond to the following graphs. (**B**) Percent of *S. thermophilus* assemblies that carry an array for each of the four CRISPR-Cas systems (n=73). An array was defined as two conserved repeat sequences. **(C)** A Tukey box and whisker plot of the number of repeats in strains according to system type. **(D-G)** Frequency analysis of repeats for each of the CRISPR-Cas systems represented as histograms. Graphs correspond to the systems as follows: (**D**) Type III-A (4.9±4.0 repeats, n=333), (**E**) CRISPR1 Type II-A (30.0±17.0 repeats, n=2,196), (**F**) CRISPR3 Type II-A (18.9±9.1 repeats, n=927), and (**G**) Type I-E (12.6±3.4 repeats, n=164)

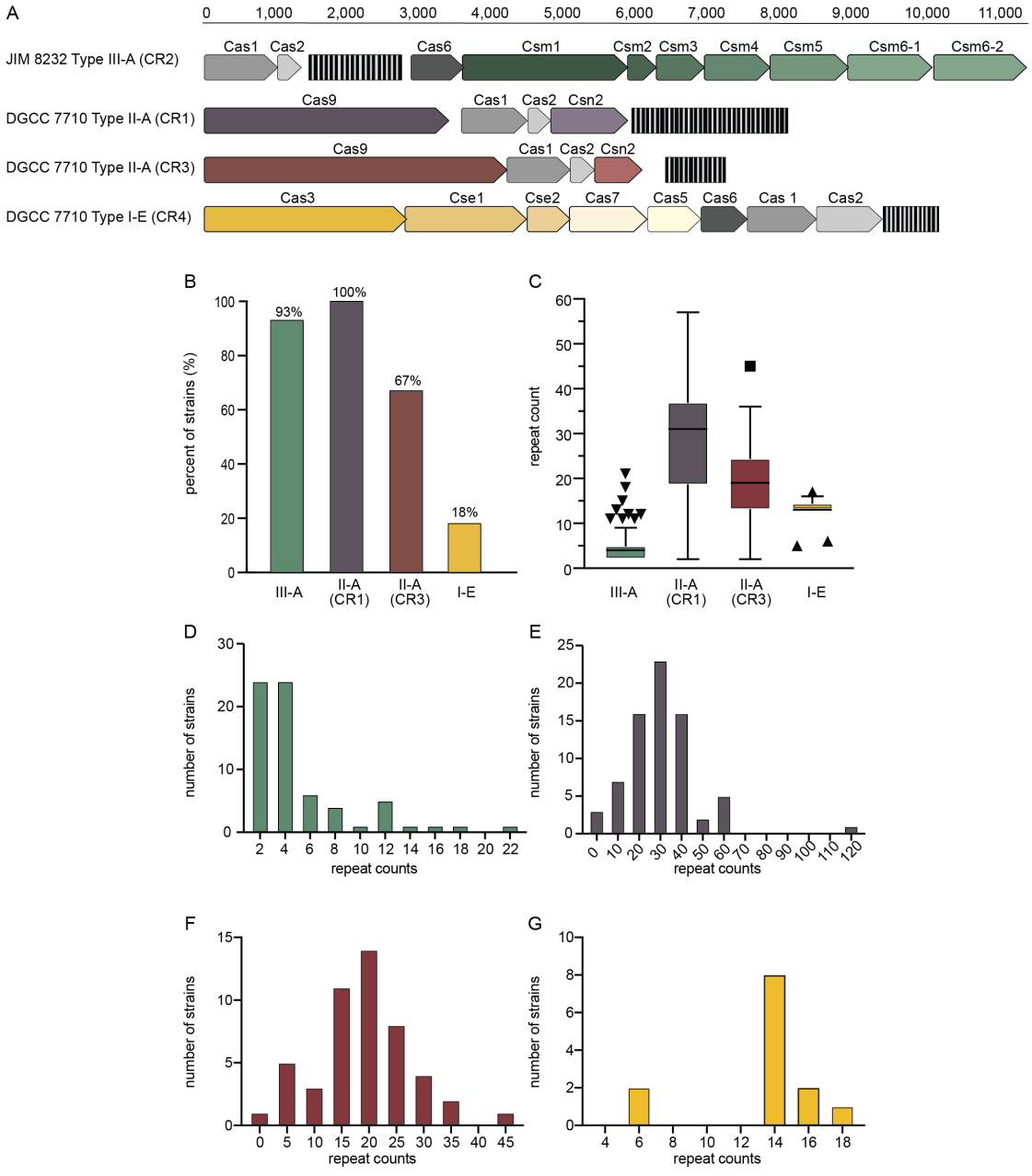

A
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
11,000
JIM 8232 Type III-A (CR2)
Cas1
Cas2
Cas6
Csm1
Csm2
Csm3
Csm4
Csm5
Csm6-1
Csm6-2
DGCC 7710 Type II-A (CR1)
Cas9
Cas1
Cas2
Csn2
DGCC 7710 Type II-A (CR3)
Cas9
Cas1
Cas2
Csn2
DGCC 7710 Type I-E (CR4)
Cas3
Cse1
Cse2
Cas7
Cas5
Cas6
Cas 1
Cas2
B
percent of strains (%)
93%
100%
67%
18%
III-A
II-A (CR1)
II-A (CR3)
I-E
C
repeat count
III-A
II-A (CR1)
II-A (CR3)
I-E
D
number of strains
repeat counts
E
number of strains
repeat counts
F
number of strains
repeat counts
G
number of strains
repeat counts

Table 2.1. *S. thermophilus* can encode up to four CRISPR-Cas systems with varying array sizes.

All *S. thermophilus* genome assemblies analyzed are listed along the left side of the table with corresponding CRISPR-Cas systems across the top. Repeat counts are listed when an array is present in the corresponding strain. Color intensity within a column corresponds to percentile of repeat count with dark to light representing highest to lowest number of repeats in a given CRISPR array.

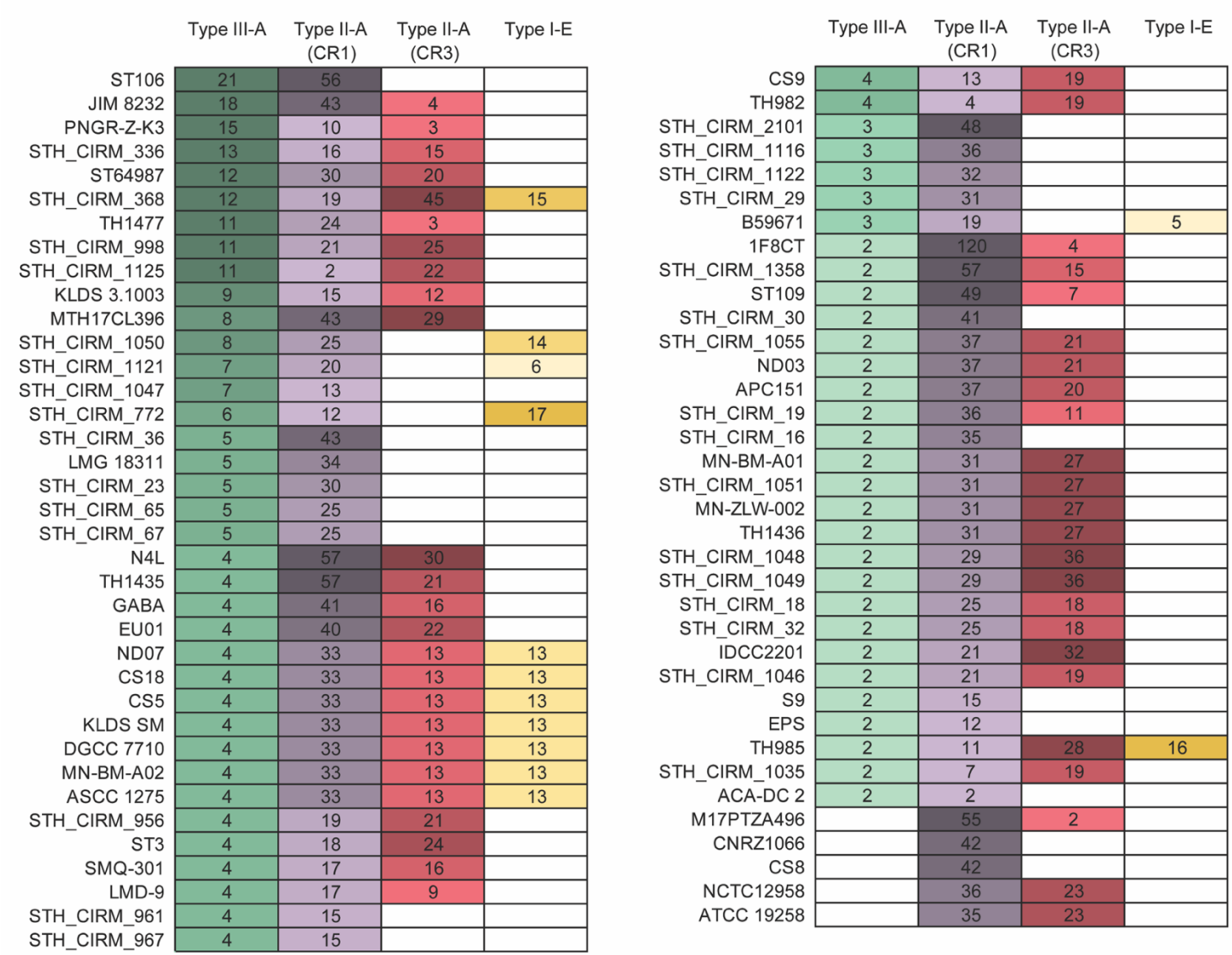

Type III-A
Type II-A (CR1)
Type II-A (CR3)
Type I-E
ST106 21 56
JIM 8232 18 43 4
PNGR-Z-K3 15 10 3
STH_CIRM_336 13 16 15
ST64987 12 30 20
STH_CIRM_368 12 19 45 15
TH1477 11 24 3
STH_CIRM_998 11 21 25
STH_CIRM_1125 11 2 22
KLDS 3.1003 9 15 12
MTH17CL396 8 43 29
STH_CIRM_1050 8 25 14
STH_CIRM_1121 7 20 6
STH_CIRM_1047 7 13
STH_CIRM_772 6 12 17
STH_CIRM_36 5 43
LMG 18311 5 34
STH_CIRM_23 5 30
STH_CIRM_65 5 25
STH_CIRM_67 5 25
N4L 4 57 30
TH1435 4 57 21
GABA 4 41 16
EU01 4 40 22
ND07 4 33 13 13
CS18 4 33 13 13
CS5 4 33 13 13
KLDS SM 4 33 13 13
DGCC 7710 4 33 13 13
MN-BM-A02 4 33 13 13
ASCC 1275 4 33 13 13
STH_CIRM_956 4 19 21
ST3 4 18 24
SMQ-301 4 17 16
LMD-9 4 17 9
STH_CIRM_961 4 15
STH_CIRM_967 4 15
Type III-A
Type II-A (CR1)
Type II-A (CR3)
Type I-E
CS9 4 13 19
TH982 4 4 19
STH_CIRM_2101 3 48
STH_CIRM_1116 3 36
STH_CIRM_1122 3 32
STH_CIRM_29 3 31
B59671 3 19 5
1F8CT 2 120 4
STH_CIRM_1358 2 57 15
ST109 2 49 7
STH_CIRM_30 2 41
STH_CIRM_1055 2 37 21
ND03 2 37 21
APC151 2 37 20
STH_CIRM_19 2 36 11
STH_CIRM_16 2 35
MN-BM-A01 2 31 27
STH_CIRM_1051 2 31 27
MN-ZLW-002 2 31 27
TH1436 2 31 27
STH_CIRM_1048 2 29 36
STH_CIRM_1049 2 29 36
STH_CIRM_18 2 25 18
STH_CIRM_32 2 25 18
IDCC2201 2 21 32
STH_CIRM_1046 2 21 19
S9 2 15
EPS 2 12
TH985 2 11 28 16
STH_CIRM_1035 2 7 19
ACA-DC 2 2 2
M17PTZA496 55 2
CNRZ1066 42
CS8 42
NCTC12958 36 23
ATCC 19258 35 23

Figure 2.2 Comparative Analysis of spacer and protospacer features of *S. thermophilus* CRISPR-Cas systems.

Individual histograms of spacer length for the total spacers of each system as follows: **(A)** Type III-A (n=270), **(B)** Type II-A (CR1) (n=2,769), **(C)** Type II-A (CR3) (n=1,316), **(D)** Type I-E (n=210). **(E)** Weblogos of the protospacer flanking sequences (PFS) adjacent to unique spacer sequence hits for each CRISPR-Cas system. Hits were required to have 100% query cover and 100% identity for Type II-A systems and 100% query cover and 90% identity for Type III-A and I-E systems due to fewer spacer sequences available for BLAST. Final PFS alignments included the following sequence counts: Type III-A n=89, Type II-A (CR1) n=335, Type II-A (CR3) n=105, Type I-E n=108. **(F)** Percentage of unique spacer hits complementary to the coding or template strand for each of the CRISPR-Cas systems. Unlike the PFS analysis, hits were required to map to coding regions of the phage genome. Using BLASTn, a bit score of 40 was used as the cut-off for a hit. Orientation was determined relative to the annotated gene. The following numbers of unique spacer sequence hits were identified for each system: Type III-A n=84, Type II-A (CR1) n=650, Type II-A (CR3) n=260, Type I-E n=50.

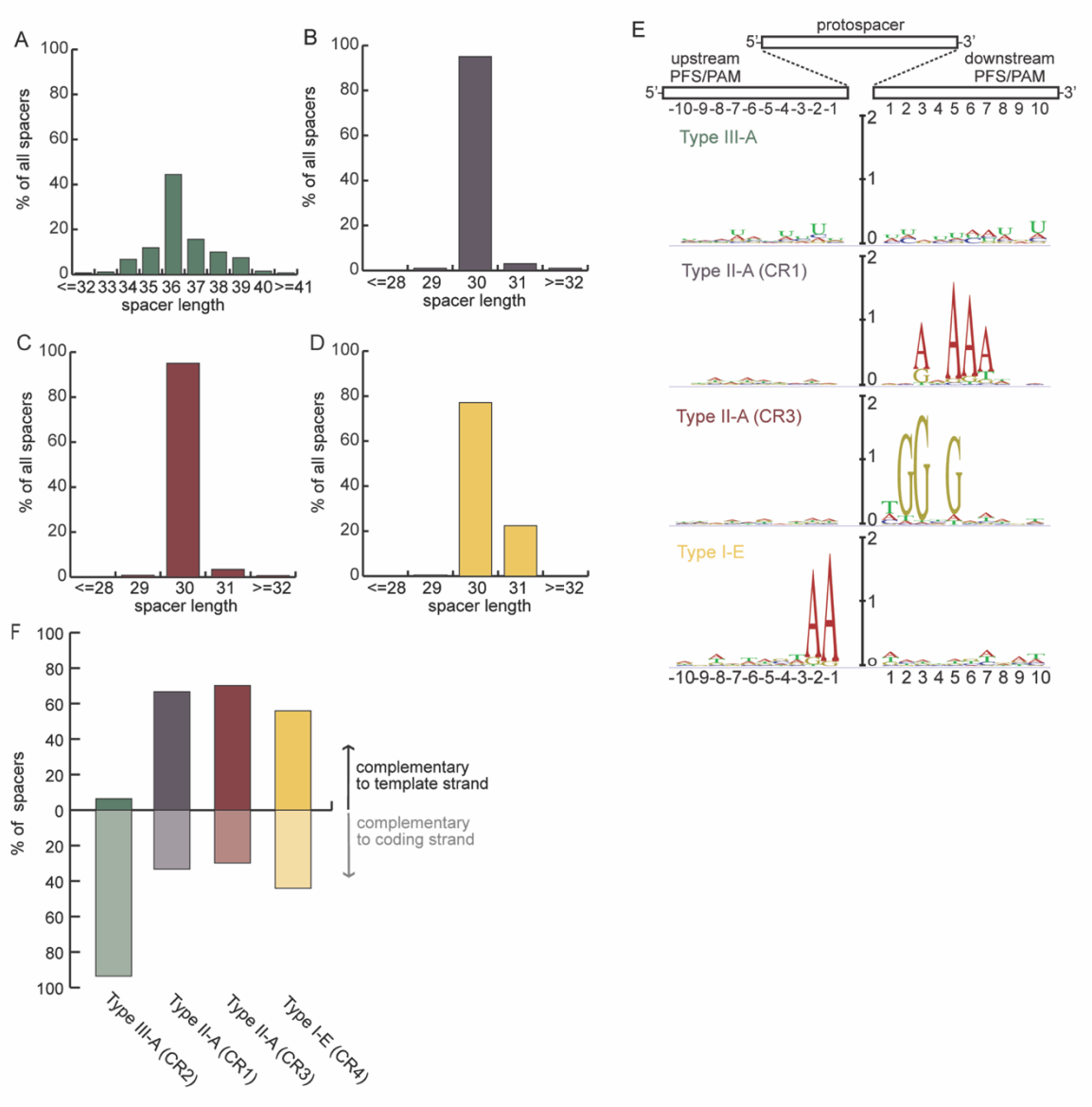

A
% of all spacers
spacer length
<=32 33 34 35 36 37 38 39 40 >=41
B
<=28 29 30 31 >=32
C
D
E
protospacer
5'
3'
upstream PFS/PAM
downstream PFS/PAM
-10-9-8-7-6-5-4-3-2-1
1 2 3 4 5 6 7 8 9 10
Type III-A
Type II-A (CR1)
Type II-A (CR3)
Type I-E
F
% of spacers
complementary to template strand
complementary to coding strand
Type III-A (CR2)
Type II-A (CR1)
Type II-A (CR3)
Type I-E (CR4)

Figure 2.3 Nuclease requirements for anti-plasmid defense of the Type III-A CRISPR-Cas system of *S. thermophilus* JIM 8232.

(A) Schematic of the Type III-A CRISPR-Cas gene locus of *S. thermophilus* JIM 8232. The functional domains of Csm1 and Csm6 are depicted. The domains we mutated are highlighted in yellow **(B)** Representation of the Type III-A Csm complex bound to target RNA (yellow) and crRNA (white) with associated 5' tag sequence (black). The interaction of the 5' tag of the crRNA with the target RNA protospacer flanking sequence determines if the HD (DNase) and PALM (cyclase) domains of Csm1 are activated. The pControl vector does not activate Csm1 while the pTarget vector does. Activation of the cyclase domain of Csm1 converts ATP into cyclic oligo A which binds to the CARF domain of Csm6. Binding of cyclic oligo A to the Csm6 CARF domain activates the Csm6 HEPN (non-specific RNase) domain. The active sites that we mutated are written in yellow. **(C)** Transformation efficiency of pTarget and pControl is indicated for each of the Type III-A system mutants. Δ* denotes the Csm1-6 knockout strain that lacks all Csm proteins. Bars represent the averages of three replicates and error bars represent standard deviation.

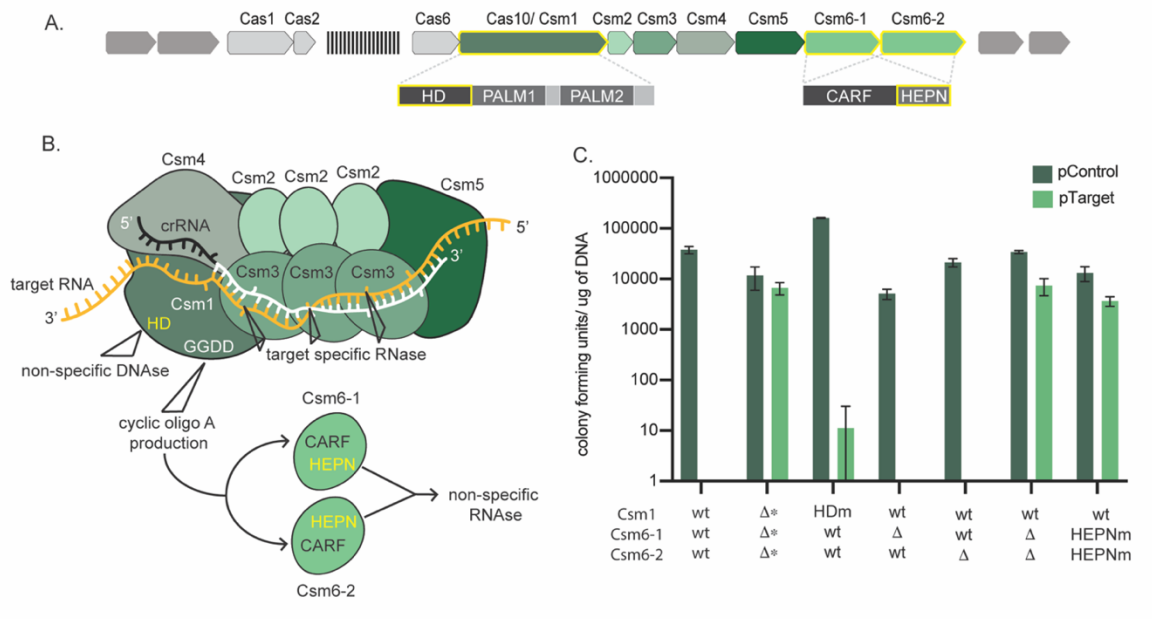

A.
Cas1
Cas2
Cas6
Cas10/ Csm1
Csm2
Csm3
Csm4
Csm5
Csm6-1
Csm6-2
HD
PALM1
PALM2
CARF
HEPN
B.
Csm4
Csm2
Csm2
Csm2
Csm5
5'
crRNA
3'
target RNA
5'
3'
Csm3
Csm3
Csm3
Csm1
HD
GGDD
non-specific DNAse
target specific RNase
cyclic oligo A
production
Csm6-1
CARF
HEPN
HEPN
CARF
Csm6-2
non-specific
RNAse
C.
pControl
pTarget
colony forming units/ ug of DNA
1000000
100000
10000
1000
100
10
1
Csm1 wt Δ* HDm wt wt wt wt
Csm6-1 wt Δ* wt Δ wt Δ HEPNm
Csm6-2 wt Δ* wt wt Δ Δ HEPNm

Figure 2.4 Nuclease requirements for anti-phage defense of the Type III-A CRISPR-Cas systems of *S. thermophilus* DGCC 7710 and JIM 8232.

(A) Phage 2972 genome transcriptional map with colors corresponding to temporal expression of genes during phage infection of ***S. thermophilus*** DGCC 7710 was adapted from Duplessis et al. (36). The two spacers used to target phage 2972 (carried on pDefense) are noted with their corresponding gene hits annotated (37). Graphs B-E are growth curves (OD600 over time) of *S. thermophilus* (*Sth*) strains transformed with pControl (pale blue) or pDefense (pale red) plasmids and challenged with phage 2972 at 3 different multiplicities of infection (MOIs). (B**)** *Sth* DGCC 7710 wt strain, **(C)** *Sth* DGCC 7710 strain with the Type III-A CRISPR-Cas system transplanted from *Sth* JIM 8232. (D) *Sth* DGCC 7710 Type III-A transplanted strain with Csm1 HD (DNase) active site mutation **(E)** *Sth* DGCC 7710 Type III-A transplanted strain with Csm6-1 and Csm6-2 HEPN (RNAse) active site mutations.

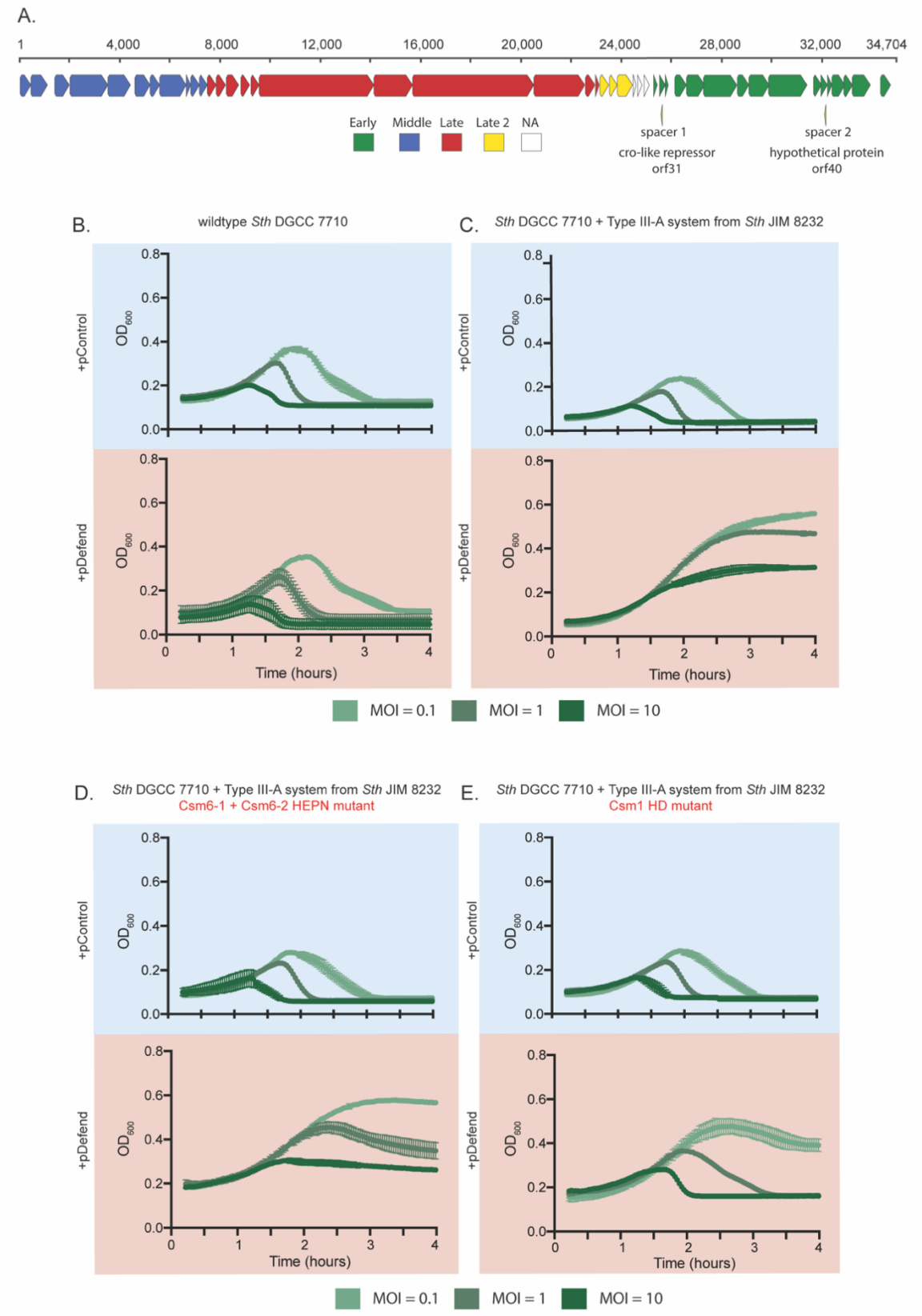

A.
1
4,000
8,000
12,000
16,000
20,000
24,000
28,000
32,000
34,704
Early
Middle
Late
Late 2
NA
spacer 1
cro-like repressor
orf31
spacer 2
hypothetical protein
orf40
B.
wildtype Sth DGCC 7710
C.
Sth DGCC 7710 + Type III-A system from Sth JIM 8232
+pControl
+pDefend
OD600
Time (hours)
MOI = 0.1
MOI = 1
MOI = 10
D.
Sth DGCC 7710 + Type III-A system from Sth JIM 8232
Csm6-1 + Csm6-2 HEPN mutant
E.
Sth DGCC 7710 + Type III-A system from Sth JIM 8232
Csm1 HD mutant

Table 2.2 S*treptococcus thermophilus* Csm protein makeup and associated functional predictions.

All analyzed S. thermophilus strains are listed across the left side of the table. The strains are ordered from most to least Type III-A repeat sequences. The remaining columns correspond to genes/proteins of the Type III-A system. For Csm1, the three required domains and motifs for predicted functionality are broken into columns. If a strain encodes the corresponding protein and our analysis predicts it to be functional, the Genbank protein accession is listed, and the box is colored green. If the gene or protein is otherwise predicted to be non-functional, the box is colored white and corresponding notes are included. '--' denotes a gene or required motif that is entirely absent. In the case that a gene was predicted by Prodigal, there is not a corresponding accession number, and the cell will indicate that it was 'manually annotated.' For the 6 strains that only had evidence of single gene in place of Csm6-1 and Csm6-2, there is only one cell in the table with the functional prediction denoted by a lighter shade of green.

Strain	# of repeats	Cas1	Cas2	Cas6	Csm1 Pfam HD Domain/motif	Csm1 Pfam Csm1_B Domain	Csm1 GGDD motif	Csm1	Csm2	Csm3	Csm4	Csm5	Csm6-1	Csm6-2
ST106	21	AXT15417.1	AXT15418.1	AXT15419.1	--	--	--	D1O36_04830, frameshifted	AXT15420.1	AXT15421.1	AXT15422.1	AXT15423.1	D1O36_04855, frameshifted	AXT15424.1
JIM 8232	18	CCC19858.1	CCC19859.1	CCC19860.1	HD	Csm1_B	GGDD	CCC19861.1	CCC19862.1	CCC19863.1	CCC19864.1	CCC19865.1	CCC19866.1	CCC19867.1
PNGR-Z-K3	15	G3T60_RS04560	G3T60_RS04565	G3T60_RS04570	HD	Csm1_B	GGDD	G3T60_RS04575	G3T60_RS04580	G3T60_RS04585	G3T60_RS04590	G3T60_RS04595	G3T60_RS04600, incomplete sequencing	G3T60_RS04605, truncated basd on nucletide and protein alignment
STH_CIRM_336	13	CAD0137551.1	CAD0137553.1	CAD0137557.1	HD	Csm1_B	GGDD	CAD0137559.1	CAD0137561.1	CAD0137563.1	CAD0137565.1	CAD0137567.1	CAD0137569.1	CAD0137571.1
ST64987	12	QIS47378.1	QIS47379.1	QIS47381.1	HD	Csm1_B	GGDD	QIS47382.1	QIS47383.1	QIS47384.1	QIS47385.1	QIS47386.1	QIS47387.1	QIS47388.1
STH_CIRM_368	12	CAD0145150.1	CAD0145152.1	Bacteriocin module insertion									CAD0145200.1, incomplete	CAD0145202.1
STH_CIRM_1125	11	CAD0176710.1	CAD0176712.1	CAD0176714.1	HD	Csm1_B	GGDD	CAD0176716.1	CAD0176718.1	CAD0176720.1	CAD0176722.1	CAD0176724.1	CAD0176726.1, premature stop codon	CAD0176728.1,truncated based on nucleotide and protein alignment
STH_CIRM_998	11	CAD0152322.1	CAD0152323.1	Bacteriocin module insertion									CAD0152347.1, incomplete	CAD0152348.1
TH1477	11	EWM61883.1	EWM61884.1	EWM61885.1	HD	Csm1_B	GGDD	EWM61886.1	EWM61887.1	EWM61888.1	manually annotated	EWM61889.1	EWM61890.1	EWM61891.1
KLDS 3.1003	9	BEN15_07960, frameshift, single nucleotide insertion	AOD26994.1	AOD26995.1	HD	Csm1_B	GGDD	AOD26996.1	AOD26997.1	AOD26998.1	AOD26999.1	AOD27000.1	manually annotated, premature stop codon	AOD27001.1, truncated based on nucleotide and protein alignment
MTH17CL396	8	ETW90235.1	ETW90236.1	ETW90237.1	HD	Csm1_B	GGDD	ETW90238.1	ETW90239.1	ETW90240.1	manually annotated	ETW90241.1	ETW90242.1	ETW90243.1
STH_CIRM_1050	8	CAD0162667.1	CAD0162670.1	Bacteriocin module insertion									CAD0162742.1, incomplete	CAD0162745.1
STH_CIRM_1047	7	CAD0158862.1, non-sense mutation, premature stop codon	CAD0158863.1	CAD0158865.1	HD	Csm1_B	GGDD	CAD0158866.1	CAD0158867.1	CAD0158868.1	CAD0158869.1	CAD0158870.1	manually annotated, premature stop codon	CAD0158871.1, truncated based on nucleotide and protein alignment
STH_CIRM_1121	7	CAD0171703.1	CAD0171704.1	CAD0171705.1	HD	Csm1_B	GGDD	CAD0171706.1	CAD0171707.1	CAD0171708.1	CAD0171709.1	CAD0171710.1	CAD0171711.1, truncated based on nucleotide and protein alignment	CAD0171712.1, truncated based on nucleotide and protein alignment
STH_CIRM_772	6	CAD0141866.1	CAD0141867.1	CAD0141868.1	--	Csm1_B	GGDD	CAD0141869.1	CAD0141870.1	CAD0141871.1	CAD0141872.1	CAD0141873.1	manually annotated, premature stop codon	CAD0141874.1, truncated based on nucleotide and protein alignment
LMG 18311	5	AAV60618.1, frameshift, single nucleotide deletion	AAV60620.1	AAV60621.1	HD	Csm1_B	GGDD	AAV60622.1	AAV60623.1	AAV60624.1	AAV60625.1	AAV60626.1	manually annotated, premature stop codon	AAV60628.1, truncated based on nucleotide and protein alignment
STH_CIRM_23	5	CAD0127024.1, frameshift, single nucleotide deletion	CAD0127020.1	CAD0127016.1	HD	Csm1_B	GGDD	CAD0127014.1	CAD0127012.1	CAD0127010.1	CAD0127008.1	CAD0127006.1	manually annotated, premature stop codon	CAD0127002.1, truncated based on nucleotide and protein alignment
STH_CIRM_36	5	CAD0134462.1/CAD0134463.1, frameshift, single nucleotide deletion	CAD0134464.1	CAD0134466.1	HD	Csm1_B	GGDD	CAD0134467.1	CAD0134468.1	CAD0134469.1	CAD0134470.1	CAD0134471.1	CAD0134472.1, premature stop codon	CAD0134473.1, truncated based on nucleotide and protein alignment
STH_CIRM_65	5	CAD0137532.1, frameshift, single nucleotide deletion	CAD0137536.1	CAD0137540.1	HD	Csm1_B	GGDD	CAD0137542.1	CAD0137544.1	CAD0137546.1	CAD0137548.1	CAD0137550.1	CAD0137552.1, premature stop codon	CAD0137554.1, truncated based on nucleotide and protein alignment
STH_CIRM_67	5	CAD0193045.1, frameshift, single nucleotide deletion	CAD0193043.1	CAD0193041.1	HD	Csm1_B	GGDD	CAD0193040.1	CAD0193039.1	CAD0193038.1	CAD0193037.1	CAD0193036.1	CAD0193035.1, premature stop codon	CAD0193034.1, truncated based on nucleotide and protein alignment
DGCC 7710	4	AUF35820.1	AUF35821.1	AUF35822.1	HD	Csm1_B	GGDD	AUF35823.1	AUF35824.1	AUF35825.1	AUF35826.1	AUF35827.1	manually annotated, premature stop codon	CW339_04795, truncated based on nucleotide and protein alignment
ASCC 1275	4	AIC24412.1	AIC24413.1	AIC24414.1	HD	Csm1_B	GGDD	AIC24415.1	AIC24416.1	AIC24417.1	AIC24418.1	AIC24419.1	manually annotated, premature stop codon	manually annotated, truncated based on nucleotide and protein alignment

CS18	4	QKM74470.1	QKM74471.1	QKM74472.1	HD	Csm1_B	GGDD	QKM74473.1	QKM74474.1	QKM74475.1	QKM74476.1	QKM74477.1	manually annotated, premature stop codon	DTA40_04860,truncated based on nucleotide and protein alignment
CS5	4	QKM58662.1	QKM58663.1	QKM58664.1	HD	Csm1_B	GGDD	QKM58665.1	QKM58666.1	QKM58667.1	QKM58668.1	QKM58669.1	manually annotated, premature stop codon	C1A39_04865, truncated based on nucleotide and protein alignment
CS9	4	DR994_01385, non-sense mutation, premature stop codon	DR994_01390, frameshift, single nucleotide insertion	QKM72258.1	HD	Csm1_B	GGDD	QKM72259.1	QKM72260.1	QKM72261.1	QKM72262.1	QKM72263.1	manually annotated, premature stop codon	DR994_01425, truncated based on nucleotide and protein alignment
EU01	4	QHD72217.1	QHD72216.1	QHD72215.1	HD	Csm1_B	GGDD	QHD72214.1	QHD72213.1	QHD72212.1	QHD72211.1	QHD72210.1	manually annotated, premature stop codon	GQY29_08290, truncated based on nucleotide and protein alignment
GABA	4	CR921_RS04685	CR921_RS04690	Bacteriocin module insertion									CR921_RS04825, incomplete	CR921_RS04830
KLDS SM	4	ANJ62002.1	ANJ62003.1	ANJ62004.1	HD	Csm1_B	GGDD	ANJ62005.1	ANJ62006.1	ANJ62007.1	ANJ62008.1	ANJ62009.1	manually annotated, premature stop codon	A9497_01065, truncated based on nucleotide and protein alignment
LMD-9	4	ABJ66187.1	ABJ66188.1	ABJ66189.1	--	--	--	--	ABJ66190.1	ABJ66191.1	ABJ66192.1	ABJ66193.1	ABJ66194.1, premature stop codon	ABJ66195.1 truncated based on nucleotide and protein alignment
MN-BM-A02	4	AKH35199.1	AKH35200.1	AKH35201.1	HD	Csm1_B	GGDD	AKH35202.1	AKH35203.1	AKH35204.1	AKH35205.1	AKH35206.1	manually annotated, premature stop codon	AKH35207.1, truncated based on nucleotide and protein alignment
N4L	4	SSC62557.1, frameshift, 20nt deletion	SSC62555.1	SSC62554.1	--	Csm1_B	GGDD	SSC62553.1	SSC62552.1	SSC62551.1	SSC62550.1	SSC62549.1	manually annotated, premature stop codon	SSC62547.1, truncated based on nucleotide and protein alignment
ND07	4	AOZ59941.1	AOZ59940.1	AOZ59939.1	HD	Csm1_B	GGDD	AOZ59938.1	AOZ59937.1	AOZ59936.1	AOZ59935.1	AOZ59934.1	manually annotated, premature stop codon	AOZ59933.1, truncated based on nucleotide and protein alignment
SMQ-301	4	AKB97580.1	AKB97581.1	AKB97582.1	--	--	--	AKB97583.1/AKB97584.1	AKB97585.1	AKB97586.1	AKB97587.1	AKB97588.1	manually annotated, premature stop codon	AKB97589.1, truncated based on nucleotide and protein alignment
ST3	4	ASX19319.1	ASX19320.1	Bacteriocin module insertion									BGL51_04975, incomplete	ASX20210.1
STH_CIRM_956	4	CAD0142031.1	CAD0142028.1	Bacteriocin module insertion									CAD0141962.1, incomplete	CAD0141959.1
STH_CIRM_961	4	CAD0147752.1	CAD0147751.1	Bacteriocin module insertion									CAD0147729.1, incomplete	CAD0147728.1
STH_CIRM_967	4	CAD0149981.1	CAD0149982.1	Bacteriocin module insertion									CAD0150002.1, incomplete	CAD0150003.1
TH1435	4	ETE41510.1	ETE41511.1	ETE41512.1	HD	Csm1_B	GGDD	ETE41513.1	ETE41514.1	ETE41515.1	ETE41516.1	ETE41517.1	manually annotated, premature stop codon	manually annotated, truncated based on nucleotide and protein alignment
TH982	4	EWM58860.1	EWM58861.1	EWM58862.1	HD	Csm1_B	GGDD	EWM58863.1	EWM58864.1	EWM58865.1	EWM58866.1	EWM58867.1	EWM58868.1	EWM58869.1
B59671	3	ATH74688.1	ATH74689.1	ATH74690.1	HD	Csm1_B	GGDD	ATH74691.1	ATH74692.1	ATH74693.1	ATH74694.1	ATH74695.1	manually annotated, premature stop codon	ATH74696.1, truncated based on nucleotide and protein alignment
STH_CIRM_1116	3	CAD0173777.1/CAD0173778.1, frameshift, single nucleotide deletion	CAD0173779.1	CAD0173781.1	HD	Csm1_B	GGDD	CAD0173782.1	CAD0173783.1	CAD0173784.1	CAD0173785.1	CAD0173786.1	CAD0173787.1	
STH_CIRM_1122	3	CAD0176980.1/CAD0176982.1, frameshift, single nucleotide deletion	CAD0176984.1	CAD0176988.1	--	--	--	CAD0176990.1/CAD0176994.1, premature stop codon	CAD0176996.1	AD0176999.1	CAD0177001.1	CAD0177003.1	manually annotated, premature stop codon	CAD0177005.1, truncated based on nucleotide and protein alignment
STH_CIRM_2101	3	CAD0182072.1/CAD0182073.1, frameshift, single nucleotide deletion	CAD0182074.1	CAD0182076.1	HD	Csm1_B	GGDD	CAD0182077.1	CAD0182078.1	CAD0182079.1	CAD0182080.1	CAD0182081.1	CAD0182082.1, premature stop codon	CAD0182083.1, truncated based on nucleotide and protein alignment

STH_CIRM_29	3	CAD0127171.1/CAD0127173.1, frameshift, single nucleotide deletion	CAD0127175.1	CAD0127179.1	HD	Csm1_B	GGDD	CAD0127181.1	CAD0127183.1	CAD0127185.1	CAD0127187.1	CAD0127189.1	manually annotated, premature stop codon	CAD0127191.1, truncated based on nucleotide and protein alignment
1F8CT	2	EWM56285.1	EWM56286.1	manually annotated	--	Csm1_B	GGDD	manually annotated	manually annotated	EWM56287.1	EWM56288.1	EWM56289.1	EWM56290.1, premature stop codon	EWM56291.1, truncated based on nucleotide and protein alignment
ACA-DC 2	2	SCB63133.1	SCB63134.1	SCB63135.1	HD	Csm1_B	GGDD	SCB63136.1	SCB63137.1	SCB63138.1	SCB63139.1	SCB63140.1	manually annotated, premature stop codon	manually annotated, truncated based on nucleotide and protein alignment
APC151	2	AQW32944.1, non-sense mutation, premature stop codon	AQW32945.1	AQW32946.1	HD	Csm1_B	GGDD	AQW32947.1	AQW32948.1	AQW32949.1	AQW32950.1	AQW32951.1	manually annotated, premature stop codon	AQW32952.1, truncated based on nucleotide and protein alignment
EPS	2	CR922_RS04745, frameshift, single nucleotide deletion	WP_014608281.1	WP_011225936.1	HD	Csm1_B	GGDD	WP_059257321.1	WP_002946668.1	WP_011225938.1	WP_059257320.1	WP_011227186.1	CR922_RS04705	
IDCC2201	2	QAU28849.1	QAU28850.1	QAU28851.1	HD	Csm1_B	GGDD	QAU28852.1	QAU28853.1	QAU28854.1	QAU28855.1	QAU28856.1	manually annotated, premature stop codon	QAU28857.1, truncated based on nucleotide and protein alignment
MN-BM-A01	2	ALD16945.1	ALD16944.1	ALD16943.1	HD	Csm1_B	GGDD	ALD16942.1	ALD16941.1	ALD16940.1	ALD16939.1	ALD16938.1	manually annotated, premature stop codon	ALD16937.1, truncated based on nucleotide and protein alignment
MN-ZLW-002	2	AFJ83379.1	AFJ83378.1	AFJ83377.1	HD	Csm1_B	GGDD	AFJ83376.1	AFJ83375.1	AFJ83374.1	AFJ83373.1	AFJ83372.1	manually annotated, premature stop codon	AFJ83371.1, truncated based on nucleotide and protein alignment
ND03	2	ADQ62975.1, non-sense mutation, premature stop codon	ADQ62976.1	ADQ62977.1	HD	Csm1_B	GGDD	ADQ62978.1	ADQ62979.1	ADQ62980.1	ADQ62981.1	ADQ62982.1	ADQ62983.1, premature stop codon	ADQ62984.1, truncated based on nucleotide and protein alignment
S9	2	ALX91983.1, frameshift, single nucleotide deletion	ALX90417.1	ALX90416.1	HD	Csm1_B	GGDD	ALX90415.1	ALX90414.1	ALX90413.1	ALX90412.1	ALX90411.1	ALX91982.1	
ST109	2	DV947_04635, frameshift, single nucleotide insertion	AXN97351.1	AXN97352.1	HD	Csm1_B	GGDD	AXN97353.1	AXN97354.1	AXN97355.1	AXN97356.1	AXN97357.1	manually annotated, premature stop codon	AXN97358.1, truncated based on nucleotide and protein alignment
STH_CIRM_1035	2	CAD0155758.1	CAD0155756.1	CAD0155752.1	HD	Csm1_B	GGDD	CAD0155750.1	CAD0155748.1	CAD0155746.1	CAD0155744.1	CAD0155742.1	CAD0155740.1	
STH_CIRM_1046	2	CAD0155546.1	CAD0155548.1	CAD0155552.1	HD	Csm1_B	GGDD	CAD0155554.1	CAD0155556.1	CAD0155558.1	CAD0155560.1	CAD0155564.1	CAD0155566.1, incomplete based on nucleotide alignment	CAD0155568.1, truncated based on nucleotide and protein alignment
STH_CIRM_1048	2	CAD0163669.1	CAD0163672.1	CAD0163675.1	HD	Csm1_B	GGDD	CAD0163678.1	CAD0163681.1	CAD0163684.1	CAD0163687.1	CAD0163690.1	manually annotated, premature stop codon	CAD0163693.1, truncated based on nucleotide and protein alignment
STH_CIRM_1049	2	CAD0162864.1	CAD0162867.1	CAD0162870.1	HD	Csm1_B	GGDD	CAD0162873.1	CAD0162876.1	CAD0162879.1	CAD0162882.1	CAD0162885.1	manually annotated, premature stop codon	CAD0162888.1, truncated based on nucleotide and protein alignment
STH_CIRM_1051	2	CAD0167454.1	CAD0167453.1	CAD0167451.1	HD	Csm1_B	GGDD	CAD0167450.1	CAD0167449.1	CAD0167448.1	CAD0167447.1	CAD0167446.1	CAD0167445.1, premature stop codon	CAD0167444.1, truncated based on nucleotide and protein alignment
STH_CIRM_1055	2	CAD0169562.1, non-sense mutation, premature stop codon	CAD0169563.1	CAD0169565.1	HD	Csm1_B	GGDD	CAD0169566.1	CAD0169567.1	CAD0169568.1	CAD0169569.1	CAD0169570.1	manually annotated, premature stop codon	CAD0169571.1, truncated based on nucleotide and protein alignment
STH_CIRM_1358	2	CAD0179979.1	CAD0179980.1	CAD0179982.1	--	--	--	--	CAD0179986.1	CAD0179987.1	CAD0179988.1	CAD0179989.1	manually annotated, premature stop codon	CAD0179990.1, truncated based on nucleotide and protein alignment
STH_CIRM_16	2	CAD0119721.1, frameshift, single nucleotide deletion	CAD0119723.1	CAD0119725.1	HD	Csm1_B	GGDD	CAD0119726.1	CAD0119727.1	CAD0119728.1	CAD0119729.1	CAD0119730.1	CAD0119731.1	

STH_CIRM_18	2	CAD0121834.1	CAD0121835.1	CAD0121837.1	--	--	--	--	CAD0121840.1	CAD0121841.1	CAD0121842.1	CAD0121843.1	manually annotated, premature stop codon	CAD0121844.1, truncated based on nucleotide and protein alignment
STH_CIRM_19	2	CAD0123984.1	CAD0123985.1	CAD0123986.1	--	--	--	--	CAD0123989.1	CAD0123990.1	CAD0123991.1	CAD0123992.1	CAD0123993.1	
STH_CIRM_30	2	CAD0130226.1	CAD0130227.1	CAD0130228.1	HD	Csm1_B	GGDD	CAD0130229.1	CAD0130230.1	CAD0130231.1	CAD0130232.1	CAD0130233.1	manually annotated, premature stop codon	CAD0130234.1, truncated based on nucleotide and protein alignment
STH_CIRM_32	2	CAD0132181.1	CAD0132180.1	CAD0132179.1	--	--	--	--	CAD0132177.1	CAD0132176.1	CAD0132175.1	CAD0132174.1	manually annotated, premature stop codon	CAD0132173.1, truncated based on nucleotide and protein alignment
TH1436	2	ETE41010.1	ETE41011.1	ETE41012.1	HD	Csm1_B	GGDD	ETE41013.1	ETE41014.1	ETE41015.1	ETE41016.1	ETE41017.1	manually annotated, premature stop codon	manually annotated, truncated based on nucleotide and protein alignment
TH985	2	EWM59227.1	EWM59228.1	EWM59229.1	HD	Csm1_B	GGDD	EWM59230.1	EWM59231.1	EWM59232.1	EWM59233.1	EWM59234.1	WM59235.1, premature stop codon	EWM59236.1, truncated based on nucleotide and protein alignment
CNRZ1066	0	AAV62539.1/ AAV62540.1, frameshift, single nucleotide deletion	AAV62541.1, frameshift, single nucleotide insertion	--	--	--	--	--	--	AAV62542.1	AAV62543.1	AAV62544.1	manually annotated, premature stop codon	AAV62545.1, truncated based on nucleotide and protein alignment
CS8	0	BAY21_00240, frameshift, single nucleotide deletion	ANS60575.1, frameshift, single nucleotide insertion	--	--	--	--	--	--	ANS60574.1	ANS60573.1	ANS60572.1	manually annotated, premature stop codon	ANS60571.1, truncated based on nucleotide and protein alignment
M17PTZA496	0	--	--	--	--	--		--	--	--	--	--	--	--
NCTC12958	0	--	--	--	--	--		--	--	--	--	--	--	--
ATCC 19258	0	--	--	--	--	--		--	--	--	--	--	--	--

REFERENCES

1. Garcia-Albiach R, Pozuelo de Felipe MJ, Angulo S, Morosini MI, Bravo D, Baquero F, et al. Molecular analysis of yogurt containing Lactobacillus delbrueckii subsp. bulgaricus and Streptococcus thermophilus in human intestinal microbiota. The American journal of clinical nutrition. 2008;87(1):91-6.

2. Achigar R, Magadán AH, Tremblay DM, Julia Pianzzola M, Moineau S. Phage-host interactions in Streptococcus thermophilus: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. Scientific Reports. 2017;7:43438.

3. Mc Grath S, Fitzgerald GF, van Sinderen D. Bacteriophages in dairy products: pros and cons. Biotechnology Journal: Healthcare Nutrition Technology. 2007;2(4):450-5.

4. Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;468(7320):67-71.

5. Magadán AH, Dupuis M-È, Villion M, Moineau S. Cleavage of Phage DNA by the Streptococcus thermophilus CRISPR3-Cas System. PLoS One. 2012;7(7):e40913.

6. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, et al. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. Journal of bacteriology. 2008;190(4):1401-12.

7. Hu T, Cui Y, Qu X. Characterization and comparison of CRISPR Loci in Streptococcus thermophilus. Archives of microbiology. 2020;202(4):695-710.

8. Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev. 2015;29(4):356-61.

9. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun. 2018;9(1):2919.

10. Hynes AP, Rousseau GM, Lemay ML, Horvath P, Romero DA, Fremaux C, et al. An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. Nature microbiology. 2017;2(10):1374-80.

11. Biswas A, Staals RH, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics. 2016;17.

12. Barrangou R, Coûté-Monvoisin A-C, Stahl B, Chavichvily I, Damange F, Romero Dennis A, et al. Genomic impact of CRISPR immunization against bacteriophages. Biochemical Society Transactions. 2013;41(6):1383-91.

13. 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(5819):1709-12.

14. Kim JG, Garrett S, Wei Y, Graveley BR, Terns MP. CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR–Cas system. Nucleic Acids Research. 2019;47(16):8632-48.

15. Leenay RT, Maksimchuk KR, Slotkowski RA, Agrawal RN, Gomaa AA, Briner AE, et al. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol Cell. 2016;62(1):137-47.

16. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009;155(Pt 3):733-40.

17. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature. 2010;463(7280):568-71.

18. Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, et al. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature. 2015;519(7542):199-202.

19. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. Embo j. 2013;32(3):385-94.

20. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences. 2012;109(39):E2579-E86.

21. Johnson K, Learn BA, Estrella MA, Bailey S. Target sequence requirements of a type III-B CRISPR-Cas immune system. The Journal of biological chemistry. 2019;294(26):10290-9.

22. Foster K, Grüschow S, Bailey S, White MF, Terns MP. Regulation of the RNA and DNA nuclease activities required for Pyrococcus furiosus Type III-B CRISPR–Cas immunity. Nucleic Acids Research. 2020;48(8):4418-34.

23. Tamulaitis G, Kazlauskiene M, Manakova E, Venclovas Č, Nwokeoji AO, Dickman MJ, et al. Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. Mol Cell. 2014;56(4):506-17.

24. Jia N, Mo CY, Wang C, Eng ET, Marraffini LA, Patel DJ. Type III-A CRISPR-Cas Csm Complexes: Assembly, Periodic RNA Cleavage, DNase Activity Regulation, and Autoimmunity. Molecular Cell. 2019;73(2):264-77.e5.

25. You L, Ma J, Wang J, Artamonova D, Wang M, Liu L, et al. Structure Studies of the CRISPR-Csm Complex Reveal Mechanism of Co-transcriptional Interference. Cell. 2019;176(1-2):239-53.e16.

26. Karvelis T, Gasiunas G, Young J, Bigelyte G, Silanskas A, Cigan M, et al. Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. Genome Biol. 2015;16:253.

27. Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell. 2015;161(5):1164-74.

28. Foster K, Kalter J, Woodside W, Terns RM, Terns MP. The ribonuclease activity of Csm6 is required for anti-plasmid immunity by Type III-A CRISPR-Cas systems. RNA Biology. 2018:null-null.

29. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119-.

30. Jung TY, An Y, Park KH, Lee MH, Oh BH, Woo E. Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. Structure. 2015;23(4):782-90.

31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

32. Biswas A, Gagnon JN, Brouns SJ, Fineran PC, Brown CM. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. RNA Biol. 2013;10(5):817-27.

33. Cooper RM, Hasty J. One-Day Construction of Multiplex Arrays to Harness Natural CRISPR-Cas Systems. ACS synthetic biology. 2020;9(5):1129-37.

34. Gardan R, Besset C, Guillot A, Gitton C, Monnet V. The oligopeptide transport system is essential for the development of natural competence in Streptococcus thermophilus strain LMD-9. Journal of bacteriology. 2009;191(14):4647-55.

35. Fontaine L, Dandoy D, Boutry C, Delplace B, de Frahan MH, Fremaux C, et al. Development of a versatile procedure based on natural transformation for marker-free targeted genetic modification in Streptococcus thermophilus. Appl Environ Microbiol. 2010;76(23):7870-7.

36. Duplessis M, Michael Russell W, A Romero D, Moineau S. Global gene expression analysis of two Streptococcus thermophilus bacteriophages using DNA microarray2005. 192-208 p.

37. Duplessis M, Russell WM, Romero DA, Moineau S. Global gene expression analysis of two Streptococcus thermophilus bacteriophages using DNA microarray. Virology. 2005;340(2):192-208.

CHAPTER 3

DISCOVERY OF NOVEL ANTI-CRISPRS AGAINST THE CRISPR-CAS SYSTEMS OF *STREPTOCOCCUS THERMOPHILUS*[2]

[2] Clare Cooper and Michael P. Terns. Discovery of novel anti-CRISPRs against the CRISPR-Cas Systems of *Streptococcus thermophilus. In preparation*.

ABSTRACT

Prokaryotic CRISPR-Cas systems provide adaptive-immunity against bacteriophages and other mobile genetic elements through sequence-specific defense. During the first encounter with a phage, adaptation proteins obtain short DNA fragments from the invading genome and incorporate them as spacers between the conserved repeats of the CRISPR locus. During future encounters with the same invader, Cas proteins are guided by these spacer-encoded crRNAs to degrade the target sequence and prevent continued infection. Despite this sophistication, both lytic and lysogenic phages are able to evade CRISPR-Cas defense. One method of this evasion is attributed to anti-CRISPR (Acr) proteins. Even with an increasing number of sequenced phage and bacterial genomes, we are not always able to accurately predict the outcome of phage-host interactions. Identification of Acr proteins contained in phages increases our understanding of defense and anti-defense dynamics of bacteriophage and host. We aimed to identify Acr proteins in phages of *Streptococcus thermophilus*. To identify Acr candidates in *S. thermophilus* phages, we used a protein clustering analysis to define homologues of a known Acr, AcrIIA5, along with neighboring genes. We then mapped these clusters back onto phage genomes. This allowed us to visualize genome neighborhoods of Acrs and narrow the region for Acr and neighboring gene expression to the lysogeny module (corresponds to the lysogeny-replacement module in lytic phages). We then screened candidate Acrs using both anti-plasmid and anti-phage defense assays against the four CRISPR-Cas systems in *S. thermophilus*. Collectively, our work defined an Acr encoding locus of *S. thermophilus* phages and identified novel inhibitors of the Type I-E and Type II-A (CR3) CRISPR-Cas systems of *S. thermophilus*.

INTRODUCTION

Clustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeat-**C**RISPR-**as**sociated (CRISPR-Cas) systems are adaptive immune systems widespread in bacteria and archaea (1, 2). The adaptive nature of CRISPR-Cas systems is a consequence of their sequence-specificity in targeting phages and other mobile genetic elements. When first encountering an invader, CRISPR-Cas systems are capable of acquiring short sequences and inserting them into the CRISPR array where they are transcribed and processed into CRISPR RNA (crRNA). These crRNAs guide Cas effector nuclease to interact with and degrade the same foreign nucleic acid sequence upon future encounters.

With bacteria capable of sophisticated adaptive defense, bacteriophages evolved to inhibit CRISPR-Cas immunity through the expression of anti-CRISPR proteins (3). These proteins have diverse mechanisms of action and often lack a conserved domain or extensive homology (4). On the hunt for novel CRISPR-Cas inhibitors, previous studies used a guilt-by-association approach where an adjacent regulatory protein, an anti-CRISPR associated protein (Aca) harboring a helix-turn-helix (HTH) domain, is used to search against other phages for homologues to identify potential Acr loci (5, 6). While not all Acrs co-occur with an HTH domain containing protein, we can use the homologues of anti-CRISPR genes to identify other loci, though Acr sequences are often less conserved and lack as many homologous genes.

Understanding of CRISPR-Cas immunity and anti-CRISPRs is important in biotechnology and gene editing due to the use of Cas9 and numerous other effector nucleases and adaptation proteins in research and therapeutic settings. Even in the native host, these systems can be important for industry. Our model organism, *Streptococcus*

thermophilus, is one example of this. *S. thermophilus* is a lactic acid bacterium used in the production of cultured dairy products like yogurts and cheeses (7). Phage infections of starter cultures during production can ruin a fermentation, so understanding the defense and anti-defense interactions of phage and bacteria in these settings is paramount to productivity (8-10). *Streptococcus thermophilus* can encode up to four CRISPR-Cas systems of three different types, including Type II-A (CR1 and CR3), Type I-E and Type III-A system types. Each of these systems is maintained in an independent locus of the genome with their own adaptation, crRNA biogenesis and effector genes adjacent to a CRISPR array (8).

At the beginning of this work, there was one identified anti-CRISPR protein in *Streptococcus thermophilus* phages, AcrIIA5 (11). Alleles of AcrIIA5 are variably capable of inhibition of the Type II-A CRISPR1 and CRISPR3 systems (11). We aimed to expand this identified anti-CRISPR repertoire in an organism that encodes more than one system type. We hope that this work can illuminate the utility of the four CRISPR-Cas systems in S. thermophilus as well as mechanisms of phage evasion of CRISPR defense. This understanding will help to better predict the outcome of phage and host interactions.

RESULTS

S. thermophilus phage protein clustering by all-by-all blast homology

To track phage genome neighborhood conservation and architecture, we wanted to develop a workflow to identify homologous proteins and map them onto their coding regions in phage genomes. We began with a dataset of all Streptococcal phages. With an initial goal to identify novel type III-A Acrs, we blasted all type III-A spacer sequences against the

phage genomes and narrowed to phages hit by these spacers. We found that spacer hits were limited to *Streptococcus thermophilus* phages as well as some *S. salivarius* phages. After narrowing our phage dataset, we extracted the proteomes of the phages of interest. We clustered the protein sequences using the Enzyme Function Initiative – Enzyme Similarity Tool (EFI-EST) with the default and recommended cut-offs (12-15). EFI-EST generates a sequence similarity network (SSN) which depicts each protein as a single node and draws edges between proteins with the supplied minimum alignment threshold. A line appears in the SSN (Figure 3.1) when the proteins have an alignment score corresponding to a percent alignment of 40% or greater. This is the recommended minimum. The network for all proteins in our dataset is depicted in Figure 3.1.

To determine a predicted function for proteins within each cluster, we gathered all annotations for cluster members and extracted the two most frequent annotations for each cluster (Table 3.1). We additionally used psiBlast and HHpred to search for functional domains using a representative sequence from each cluster (16-19). Once phage proteins were assigned a cluster number and color from EFI-EST, we visualized the network of assigned clusters with Cytoscape (20, 21). Additionally, all cluster colors and numbers were annotated back onto the encoding phage genomes in Geneious.

With this approach, we can now visualize regions of conservation across phages of the same type. We can also track genome neighborhoods of Acrs and understand relationships between phage genes that may have unknown functions.

Genome neighborhoods of AcrIIA5 homologues

AcrIIA5 is a previously published Type II-A Acr with activity against the CRISPR1 system of *S. thermophilus* (11). Further analysis of the alleles of AcrIIA5 determined that they are variably capable of CRISPR3 inhibition of CRISPR-Cas9 function (11, 22, 23). Using the clustering analysis, we mapped all homologues of AcrIIA5 back onto phage genomes (Figure 3.2) to determine which proteins co-occur with AcrIIA5 most frequently. In our dataset, all eight homologues of AcrIIA5 are contained within cluster 127. Consistent with the modular architecture of *S. thermophilus* phage genomes, the AcrIIA5 homologues all mapped to the same region of the genome: the lysogenic module in lysogenic phages or the lysogenic-replacement module in lytic phages (10). To define protein clusters that cooccur with AcrIIA5 in this module, we set module boundaries at the end of the lytic module (upstream) and the beginning of the replication module (downstream). The protein clusters that define the upstream border are clusters 2 and 78 (a lysin and amidase, respectively). The protein cluster that defines the downstream border is cluster 19 (Cro repressor). We extracted all proteins that co-occurred with AcrIIA5 between these borders and considered them candidate Acrs (Figure 3.2). Once we defined protein clusters that are associated with AcrIIA5, we applied another degree of guilt-by-association to identify clusters that are present alongside of these AcrIIA5 associated-clusters. We mapped all homologues within these clusters back onto the phage genomes and found that again, all clusters mapped to the lysogeny region (Figure 3.3).

Acr candidates are limited to the lysogenic / lysogenic-replacement module

The lysogenic module of *S. thermophilus* phage genomes is considered a 'variable region' that falls between conserved modules (10). There is a second region of variable expression downstream in the regulatory module (24). We wanted to know if proteins that occurred within the lysogenic module (now candidates in our analysis) ever mapped to the variable region of the downstream regulatory module. We analyzed the genome locations of all clusters of the lysogenic module (Figure 3.3) and found that none of the cluster members ever moved outside of the bounds we set for this module. This gave us additional confidence in our use of the lysogenic module as the sole region for candidate selection. We chose to consider all lysogenic module protein clusters as candidate clusters (Figure 3.3).

AcrIIA6 genome neighborhoods

During candidate selection and screening, protein D1811_026 from candidate cluster 44 was published as AcrIIA6, a Type II-A Acr against CRISPR1 (25, 26). This gave us additional confidence in the candidate predictions and greatly expanded the genome neighborhoods that contained a homologue of a known Acr. All members of cluster 44 map to the lysogenic or lysogenic replacement module, consistent to what we saw for AcrIIA5 and all co-occurring clusters (Figure 3.3).

Plasmid-based assay of CRISPR defense and anti-CRISPR activity

To screen candidate Acrs for activity against the four CRISPR-Cas systems in *Streptococcus thermophilus*, we utilized target plasmids unique to each CRISPR-loci

(Figure 3.4-A and 3.4B). For the type I-E and II-A systems, the complementary sequence of the first spacer of the CRISPR locus was cloned into a pWAR plasmid with chloramphenicol selection. In addition, the appropriate protospacer adjacent motif (PAM) was included adjacent to the target (Figure 3.4A). For the Type III-A system, expression of the target RNA sequence is required, so we used a constitutive promoter, Ppgm, to express the sequence as well as a rho-independent terminator from the Type II-A CRISPR1 tracrRNA gene. Also, for Type III-A, we found a single target sequence to be too weak in activating defense and thus expressed the complementary sequence of the first three spacers of the type III-A locus with a strongly activating protospacer flanking sequence (PFS) adjacent to each (Figure 3.4A).

Acr genes were expressed on pTRK882 under a constitutive promoter, Ppgm, with erythromycin selection. *S. thermophilus DGCC* 7710 was initially transformed with each Acr vector prior to being made naturally competent and transformed with target vectors (Figure 3.4B). We then plated the transformants to double selective plates for both pTRK882 and pWAR. As this was an initial screen, we looked for presence of colonies on the plate as an indicator of Acr activity (Figure 3.4B).

Phage-based assay of CRISPR defense and anti-CRISPR activity

To assay our anti-CRISPR candidates in inhibiting CRISPR-Cas defense against an invading phage, we chose to use the well-studied lytic phage 2972 and host, *DGCC* 7710 (Figure 3.4-C and 3.4-D). Lysis of the host leads to total clearing of the culture and is a simple readout of phage-host dynamics. In the case that a bacteriophage insensitive mutant (BIM) strain that we isolated as being resistant to the phage through harboring a single

spacer matching the phage is inoculated with phage, the culture will not lyse and will remain turbid. To test anti-CRISPR activity against each of the CRISPR-Cas systems, we utilized four different BIMs, each with a spacer acquired in one of the four CRISPR-Cas loci (Figure 3.4C). Survival of the host and turbidity of the culture is thus dependent on the activity of the CRISPR-loci that has undergone spacer acquisition against phage 2972.

We again expressed anti-CRISPR candidates constitutively from pTRK882 and transformed them into each of the CRISPR-Cas system BIMs (Figure 3.4D). We inoculated a 96-well plate with the Acr encoding BIMS plus and minus phage 2972 at a multiplicity of infection (MOI) of 1. To track strain growth and lysis, we took OD600 measurements on a plate reader every 5 minutes for 24 hours. As this is a screen, we use the read-out of lysis or growth as a positive or negative Acr result, respectively (Figure 3.4D).

Initial screening candidates

For our initial screening, we selected AcrIIA5 (cluster 127) as our positive control and the putative HTH-containing Aca protein, cluster 76, as our negative control. Additionally, cluster 44 was published as AcrIIA6 and is not characterized here (22). Two of the candidates are within subgroups of a single cluster, cluster 69a and cluster 69b, as the network for this cluster showed two sub-groups of nodes. (Figure 3.1). The remaining candidates include candidate clusters 65, 93, 119, 144, and 152. The positive screening results are shown in the remaining figures. We are currently moving forward with screening the remaining candidates from the clustering analysis for Acr activity aginst the four CRISPR-Cas systems of *S. thermophilus*.

Identification of a widespread Type II-A anti-CRISPR, Cluster 93, AcrIIA25

The plasmid and phage screening results for candidate 93 are shown in figure 3.5A and 3.5B. The candidate we selected to screen from cluster 93 is phage gene P5641_25. Here we demonstrate that this gene is capable of selectively inhibiting the Type II-A system of CRISPR3 without inhibiting the Type II-A system of CRISPR1. This is a novel finding as alleles of previously identified AcrIIA5 and AcrIIA6 are either CRISPR1 specific or inhibit both CR1 and CR3. Compared to the Cas9 of CR1, Cas9 of CR3 shares higher homology and PAM specificity with the Cas9 protein of *S. pyogenes* which is used most frequently in gene editing (23, 27, 28). Thus, this newly identified Acr could have utility in inhibition of *S. pyogenes* Cas9 for biotechnological purposes.

The genome neighborhoods of all homologues of candidate 93 are shown in figure 3.5-C. Within these other genomes, the same architecture of the genome neighborhood is maintained with a lysin or related phage gene upstream.

Identification of a phage-specific Type II-A anti-CRISPR, Cluster 119, AcrIIA26

The plasmid and phage screening results for candidate cluster 119 are shown in figure 3.6-A and 3.6-B. For candidate cluster 119, we selected gene Sfi11_gp83 to screen. We found this gene to have no activity in our plasmid based anti-CRISPR assay due to no transformation of the target vectors. However, it is a clear inhibitor of CRISPR3 in our phage-based assay with complete lysis of the Type II-A CR3 BIM. Despite a late lag in growth of the CR1 BIM in our phage-based assay, candidate 119 appears to be specific for the Type II-A CR3 system when compared to AcrIIA5 (Figure 3.6B). It also has no off-

target effects against the Type I-E or Type III-A CRISPR-Cas systems, indicating that it is likely a specific Type II-A anti-CRISPR in functionality.

Within our *S. thermophilus* phage dataset, all cluster 119 members co-occur with AcrIIA6 (Figure 3.6-C). When cluster 119 proteins occur with an integrase (cluster 125), indicating a lysogenic lifestyle for the phage, they additionally maintain a putative Aca downstream (cluster 76). For genomes outside of our phage dataset, the same genome neighborhood conservation is maintained within prophages that have integrated into the bacterial genomes (Figure 3.6-C). Potential Acr candidates not found in our phage clustering dataset are noted in these expanded genome neighborhoods.

Identification of a novel Type I-E anti-CRISPR, Cluster 152, AcrIE10

The plasmid and phage screening results from candidate 152 are shown in figure 3.7-A and 3.7-B. For candidate cluster 152, we selected gene D1811_027 to screen. We found this gene to have clear Type I-E anti-CRISPR activity in both plasmid and phage-based assays. While there is some colony formation on the plate for the Type III-A target plasmid transformation, there is not transformation of the target plasmid equivalent to the empty vector, and there is no activity against the Type III-A BIM in the phage-lysis assay. This is the first Type I-E anti-CRISPR identified in gram-positive bacteria and the only identified anti-CRISPR against a Class I system in *Streptococcus thermophilus*.

Genome neighborhoods for all candidate cluster 152 members are shown in Figure 3.6C. Unique to candidate 152, the genome neighborhoods outside of our dataset do not have the same conservation that we saw for phage and prophage genomes. One homologue is present in a genome adjacent to a transposase, indicating potential transmission through

a non-phage mobile-genetic-element. Additionally, in other contexts, there are large genes present between cluster 152 and AcrIIA5 homologues. These additional genes are indicated as Acr candidates that can be tested in the future.

DISCUSSION

Here we demonstrate a methodology for tracking genome neighborhood relationships in related phages using all-by-all BLAST homology to assign clusters (Figure 3.1 and Table 3.1) followed by mapping of the clusters back onto phage genomes (Figures 3.2 and 3.3). In addition, we define a functional screen for anti-CRISPR activity against both anti-plasmid and anti-phage CRISPR-Cas defense (Figure 3.4). We use these approaches to identify a conserved genome neighborhood of anti-CRISPR proteins and demonstrate activity of three novel anti-CRISPRs (Figures 3.5-3.7). Compared to previous anti-CRISPR analyses, this is a comprehensive method for identifying anti-CRISPRs against four unique CRISPR-Cas systems of three unique system types. Future screens of the anti-CRISPR candidates shown in Figure 3.3 should further define any additional anti-CRISPR proteins present in this genome locus.

The current work more than doubles the identified Acr proteins in Streptococcal phages. Using this knowledge, we should be capable of more accurately defining outcomes of phage-host interaction by predicting if a phage is resistant to CRISPR immunity. It also highlights the potential importance of carrying more than one CRISPR-Cas system in *Streptococcus thermophilus*. Perhaps the multiple systems evolved to combat a continued evolution of anti-CRISPR proteins.

Future studies should consider whether anti-CRISPRs or CRISPR-Cas systems eventually win in long-standing phage infections. These studies could consider if CRISPR-Cas systems are lost in entirety or if mutations in effector proteins allow escape from anti-CRISPR proteins. Additionally, there is the potential for identification of "anti-anti-CRISPR proteins" or other mechanisms that favor bacterial defense.

An additional consideration could be that if the dominant Type II-A systems in *Streptococcus thermophilus* are inhibited, perhaps there are non-CRISPR defense mechanisms that take over, or preferential adaptation in the Type III-A or Type I-E systems. With the widespread nature of AcrIIA6 and the newly identified AcrIIA25, it is likely that both Type II-A systems are inhibited by many *S. thermophilus* phages. Future studies should consider how *S. thermophilus* strains overcome anti-CRISPRs and the implication that Type II-A CRISPR-Cas systems are still widespread despite this apparent inhibition.

MATERIALS AND METHODS

Phage amplification

Phage 2972 was amplified on *S. thermophilus* DGCC 7710 in LM17 supplemented with 10 mM $CaCl_2$. The amplification was carried out at 42ºC with shaking (give rpms). The phage underwent two rounds of amplification. Both amplifications were filtered using a 0.45 um polyethersulfone (PES) filter. Phage was stored at 4ºC for short term storage or in 50% glycerol at -80ºC for long term storage.

Phage titration

Overnight cultures of *S. thermophilus* (OD_{600} = 0.6) were diluted in melted LM17 0.75% 'top' agar supplemented with 10 mM $CaCl_2$ (300 uL of cells to 3 mL of agar). Stocks of phage 2972 were serially diluted in phage buffer before being co-inoculated (100 uL of each dilution to 3.3 mL). The phage and strain agar mix were then poured over the top of LM17 plates supplemented with 10 mM $CaCl_2$ and antibiotics where indicated and allowed to set before being incubated at 42ºC overnight. Plaques were counted on plates with between 30 and 300 plaques to calculate the plaque forming units (PFU) of the phage stock.

Genome neighborhoods analysis outside of Phage dataset

To identify gene homologues outside of *S. thermophilus* phages, the Uniprot IDs of all cluster members were mapped to Uniref50 cluster IDs. The corresponding Uniref50 IDs were entered into EFI-EST to generate a sequence similarity network (SSN) which depicts each protein as a node in a connected network. This SSN was then submitted for genome neighborhood analysis. This analysis outputs the genome neighborhoods of homologues present in bacterial genomes. We further characterized the genome neighborhoods and added annotations for proteins from phage protein clustering as well as those listed in Uniprot.

Clustering of phage proteins

Genomes of Streptococcal phages were downloaded from NCBI Assembly on 04/10/19. Proteomes of the associated genomes were downloaded from Uniprot and uploaded to EFI-EST for creation of a sequence similarity network (SSN) (12-15). Standard

recommendations for SSN generation were used with alignment score selection based on a percent alignment of 40%. Individual nodes were created for each protein. Clusters assigned by the SSN were visualized using Cytoscape 3.7.2 (20, 21) and annotated back onto phage genomes in Geneious Prime 2019.2.3 (https://www.geneious.com).

Target vector cloning

All target sequences were cloned into a pWAR vector backbone (ref) with a chloramphenicol (Cm) resistance marker. Transcribed targets were expressed downstream of a constitutive Ppgm promoter sequence while non-transcribed targets were placed downstream of a rho-independent terminator sequence from the Type II-A CRISPR tracrRNA gene. Target vectors were created using blunt end ligation with ssDNA oligos or inverse PCR using primers containing the target sequence as 5' overhangs. Target vectors were maintained in *E. coli* Top 10.

Anti-CRISPR vector cloning

Initial cloning of Acrs was carried out in pMEU5a using Golden Gate assembly with BbsI restriction enzyme. Acr sequences were ordered as gblocks from Integrated DNA Technologies with appropriate overhangs for cloning. Later, Acrs were moved to pTRK882 using the same overhangs or a combination of PCR and Gibson assembly using HiFi DNA assembly master mix. In both plasmid contexts, Acrs were expressed under a constitutive Ppgm promoter with a rho independent terminator sequence from the Type II-A CRISPR1 tracrRNA gene.

Strain Culturing

S. thermophilus cultures were grown in LM17 (HiMedia). For liquid overnight growth and natural transformation, strains were grown at 37ºC. For overnight growth on agar plates, strains were grown at 42ºC. When indicated, chloramphenicol 5 ug/mL and/or erythromycin 10 ug/mL was utilized. For assays with phage 2972, 10 mM $CaCl_2$ was supplemented.

E. coli was grown in Luria broth with chloramphenicol 10 ug/mL or erythromycin 200 ug/ml at 37ºC.

Plasmid preparation and detection

To assess presence and concentration of target plasmids for transformation assays, plasmids were purified from overnight culture by Zymopure Plasmid Midiprep Kit (Zymo Research). 50 ng and 200 ng of each target DNA were analyzed by agarose gel electrophoresis.

Plasmid transformation

Strains of *S. thermophilus* were grown overnight at 37ºC with shaking at 200rpm. The following morning, 1 mL of overnight culture was spun down (5,000 x g, 2 mins) and washed with chemically defined media (CDM) (29).Two wash steps were completed and then each strain was diluted 50% in a 96-well plate (100 uL total) for OD read. Based on OD_{600} measurement, strains were all diluted to OD_{600} = 0.1 in and incubated for 1 hour at 37ºC standing. Following incubation, preparations of naturally competent cells were frozen at -20ºC before use as well as for long term storage. When ready for use, competent

cells were thawed at room temperature. Plasmid DNA was added to the wells of a 96 well-plate (200 ng) and naturally competent cells were combined with ComS peptide at a concentration of 1uM. 100 uL of cells were then added to each well and incubated at 37ºC without shaking. After 3 hours, each well of the 96-well plate was resuspended prior to plating 30 uL to LM17 agar plates with indicated antibiotic selection. Plates were incubated overnight at 42ºC prior to imaging. Images were made using the Biorad Gel Doc XR.

Phage lysis

Following overnight growth at 37ºC shaking, strains of *S. thermophilus* were diluted 75% in LM17. 100 uL of each dilution was added to the wells of a 96-well plate for OD_{600} measurement on the plate reader. Based on this measurement, overnight cultures were diluted to OD_{600} = 0.2 and supplemented with 10 mM $CaCl_2$. Based on desired MOI, the appropriate volume of phage 2972 (or phage buffer control) was inoculated into the wells of a fresh 96-well plate. To each well, 200 uL of diluted overnight culture was added. Growth was carried out on the plate reader at 42ºC with double orbital shaking at 1500rpm. Measurements of OD_{600} were taken every 5 minutes for 24 hours. Growth curves were analyzed using Prism 8.4.1.

Figure 3.1 Sequence Similarity Network (SSN) of *Streptococcus thermophilus* phage proteins.

All proteins from the phage dataset underwent all-by-all BLAST analysis to generate protein clusters based on homology. Nodes represent a single protein. Edges connect proteins with an alignment score corresponding to greater than or equal to 40% alignment. The edge length between two nodes corresponds to percent alignment with shorter edge length corresponding to a higher degree of similarity. Cluster numbers were assigned in ascending order according to cluster size.

1
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
17 19 18 26 23 21 24 20 25 22 28 27 29 30 31 32 34 33 35 37 36 39 38 40 41 43 42 44 47 50 46
49 45 48 51 52 53 54 55 56 57 60 59 58 61 62 63 64 65 67 66 69 68 70 72 71 73 74 75 76 77
78 79 90 85 87 82 89 86 88 83 80 81 84 92 93 91 94 100 98 99 95 96 97 101 103 102 119 111 108 107 113 109 115 114 118 110 105 106 104 120
116 117 112 123 122 121 126 124 128 127 125 131 132 129 130 136 135 133 134 137 145 144 142 141 138 143 146 139 140 153 150 157 152 148 154 156 149 151 147 155 164 158 166 162 165
161 159 163 160 175 182 185 177 178171 184 183 180 173 168 169 167 170 172 186 179 176 174 181

Table 3.1 Cluster annotations corresponding to the phage protein SSN clusters.

Corresponding annotations for each cluster number are listed in the table. The top two most frequent CDS annotations for the proteins within the cluster are listed with their corresponding count.

cluster #	Count of proteins in the Cluster	Annotation	Count of annotation
1	948	Uncharacterized protein	640
		DNA binding protein	94
2	252	Lysin	151
		Choline binding protein A	43
3	218	Antireceptor	64
		Capsid and scaffold protein	39
4	206	Uncharacterized protein	131
		ORF6C domain-containing protein	43
5	189	Tape measure protein	101
		Peptidase C51 domain-containing protein	24
6	179	Holin	165
		Uncharacterized protein	5
7	171	Terminase large subunit	105
		TerL	39
8	163	Baseplate component	35
		Uncharacterized protein	31
9	162	Distal tail protein	79
		Uncharacterized protein	38
10	161	Uncharacterized protein	153
		ORF23	1
		Orf117b gp	1
		Structural protein	1
		Orf131 gp	1
		Phage protein	1
		Gp149	1
		ORF47	1
		ORF40	1
11	159	Uncharacterized protein	155
		ORF3	1
		Gp40	1
		ORF7	1
		ORF34	1
12	147	Uncharacterized protein	142
		Holin	5

13	142	Uncharacterized protein	137
		ORF17	1
		ORF21	1
		Orf98 gp	1
		Gp93	1
		ORF45	1
14	135	Uncharacterized protein	53
		Apaf-1 related killer DARK	44
15	135	Uncharacterized protein	133
		ORF41	1
		ORF48	1
16	126	Uncharacterized protein	90
		Recombinational DNA repair protein	12
17	124	Cro-like protein	104
		Transcriptional Cro repressor	13
18	115	Major tail protein	108
		Uncharacterized protein	4
19	115	Cro-like repressor	89
		Transcriptional regulator	11
20	114	Major capsid protein	63
		Capsid protein	40
21	114	Portal protein	109
		Orf384 gp	1
		Orf387 gp	1
		ORF24	1
		Putative portal protein	1
		Portal (Connector) protein	1
22	114	Clp protease	42
		Clp protease-like protein	41
23	114	HNH endonuclease	93
		Homing endonuclease	17
24	114	Tail assembly protein	39
		Tail component	26
25	114	Head-tail connector protein	45
		Capsid and scaffold protein	39
26	114	Tail chaperone protein	42
		Tail assembly protein	35

27	113	Terminase small subunit	55
		TerS	26
28	113	Tail assembly protein	39
		Tail component	26
29	113	DNA packaging protein	103
		Putative DNA packaging protein	5
30	106	Head-tail connector protein	42
		Capsid and scaffold protein	36
31	104	DNA binding protein	41
		DNA-binding protein	22
32	102	Primase	48
		Replication protein	29
33	101	Helicase	64
		Uncharacterized protein	17
34	101	Uncharacterized protein	70
		Single-stranded DNA-binding protein	26
35	99	VRR-NUC domain-containing protein	95
		Orf106 gp	1
		Gp106	1
		ORF41	1
		ORF15	1
36	83	Helicase loader	27
		DNA replication protein	18
37	83	Uncharacterized protein	78
		Gp157	2
38	82	Replication initiation protein A	15
		DNA replication protein	14
		Replication initiation	14
39	82	Terminase large subunit	56
		TerL	18
40	79	Single-stranded DNA-binding protein	58
		SsDNA-binding protein	9
41	76	Endonuclease	48
		HNHc domain-containing protein	12
42	66	Holliday junction resolvase	27
		Endodeoxyribonuclease	21
43	66	Uncharacterized protein	66
44	65	Uncharacterized protein	48

		Acr-like protein	13
45	60	Portal protein	53
		Phage portal protein	1
		Putative portal protein	1
		ORF6	1
		ORF27	1
		Gp502	1
		Uncharacterized protein	1
		Portal (Connector) protein	1
46	60	Minor capsid protein	31
		Phage_Mu_F domain-containing protein	21
47	60	Major tail protein	26
		Tail protein	26
48	60	Uncharacterized protein	23
		Capsid and scaffold protein	21
49	60	Uncharacterized protein	30
		Tail chaperone protein	27
50	60	Uncharacterized protein	53
		Capsid and scaffold protein	2
51	59	Head-tail connector protein	28
		Uncharacterized protein	28
52	54	Terminase small subunit	32
		TerS	14
53	51	Uncharacterized protein	17
		ArpU late transcriptional regulator	14
54	49	Uncharacterized protein	45
		ORF17	1
		Gp57	1
		Orf57 gp	1
		ORF42	1
55	48	Erf protein	28
		RecT recombinase	4
56	46	Capsid and scaffold protein	16
		Scaffolding protein	14
57	46	Uncharacterized protein	42
		ORF18	1
		Gp105	1

		Tail chaperone protein	1
		ORF39	1
58	45	Major capsid protein	30
		Major capsid protein E	7
59	45	Uncharacterized protein	42
		ORF13	1
		Gp104	1
		ORF34	1
60	45	Uncharacterized protein	41
		Gp53	1
		Capsid	1
		ORF32	1
		ORF11	1
61	42	Antirepressor	21
		Antirepressor protein	16
62	35	Uncharacterized protein	35
63	33	Nuclease	19
		DUF3799 domain-containing protein	13
64	32	HTH cro/C1-type domain-containing protein	10
		Cro repressor	9
		Cro-like repressor	9
65	31	Uncharacterized protein	26
		Orf11 gp	1
		Orf110 gp	1
		ORF54	1
		Gp111	1
		Orf111	1
66	30	HNH endonuclease	19
		HNHc domain-containing protein	5
67	30	Uncharacterized protein	30
68	29	Uncharacterized protein	25
		TcdA-E operon negative regulator	3
69	29	Uncharacterized protein	24
		Gp71	1
		ORF52	1
		Orf88 gp	1

		Gp145	1
		ORF53	1
70	27	Uncharacterized protein	23
		Gp143	2
71	20	Uncharacterized protein	20
72	20	Uncharacterized protein	19
		ORF24	1
73	19	Uncharacterized protein	19
74	17	Uncharacterized protein	17
75	17	Uncharacterized protein	17
76	16	Uncharacterized protein	9
		HTH cro/C1-type domain-containing protein	5
77	16	Terminase large subunit	10
		TerL	4
78	15	Lysin	11
		N-acetylmuramoyl-L-alanine amidase	3
79	15	Site-specific DNA methyltransferase	8
		Adenine-specific methyltransferase	6
80	14	Tail associated lysin	8
		Tail-associated lysin	4
81	14	Antireceptor	12
		Upper baseplate protein	2
82	14	Major capsid protein	14
83	14	Distal tail protein	12
		Capsid and scaffold protein	2
84	14	Scaffolding protein	12
		Capsid and scaffold protein	2
85	14	Uncharacterized protein	13
		ORF14	1
86	14	Uncharacterized protein	14
87	14	HTH cro/C1-type domain-containing protein	7
		Uncharacterized protein	4
88	14	Uncharacterized protein	14
89	14	Ig domain containing protein	5
		BIG2 domain-containing protein	4
90	14	Uncharacterized protein	13

		ORF29	1
91	13	Transferase	8
		Serine acetyltransferase	5
92	13	Endodeoxyribonuclease RusA	7
		Uncharacterized protein	5
93	13	Uncharacterized protein	13
94	13	Holin	13
95	12	Antireceptor	8
		Uncharacterized protein	2
		SGNH_hydro domain-containing protein	2
96	12	DNA methyltransferase	11
		Methylase	1
97	12	HTH_Tnp_1_2 domain-containing protein	12
98	12	Uncharacterized protein	12
99	12	Uncharacterized protein	11
		ORF25	1
100	12	Uncharacterized protein	12
101	11	N6_Mtase domain-containing protein	3
		Uncharacterized protein	3
102	11	YopX domain-containing protein	8
		Uncharacterized protein	2
103	11	RHH_3 domain-containing protein	10
		Ribbon-helix-helix domain-containing protein	1
104	10	Portal protein	8
		Minor capsid protein	2
105	10	Tail associated lysin	4
		Tail-associated lysin	4
106	10	Terminase large subunit	5
		TerL	4
107	10	Minor capsid protein	9
		Putative minor capsid protein	1
108	10	Cytosine-specific methyltransferase	9
		DNA-cytosine methyltransferase	1
109	10	Major capsid protein	9
		Major head protein	1
110	10	Distal tail protein	8
		Uncharacterized protein	2

111	10	GIY-YIG domain-containing protein	10
112	10	Uncharacterized protein	10
113	10	Scaffolding protein	8
		Capsid and scaffold protein	1
		Putative scaffolding protein	1
114	10	Major tail protein	5
		Tail protein	5
115	10	Minor capsid protein	9
		Putative minor capsid protein	1
116	10	Uncharacterized protein	7
		Putative head-tail connector protein	3
117	10	Minor capsid protein	9
		Putative minor capsid protein	1
118	10	Uncharacterized protein	10
119	10	Uncharacterized protein	7
		Orf54 gp	1
		Gp83	1
		Orf83	1
120	10	Putative termination factor	4
		Rho termination factor	4
121	9	Minor capsid protein	8
		Uncharacterized protein	1
122	9	Uncharacterized protein	4
		XRE family transcriptional regulator	4
123	9	Uncharacterized protein	9
124	8	Transposase	6
		Mobile element protein	2
125	8	Integrase	5
		ORF1	1
		Uncharacterized protein	1
		Phage integrase	1
126	8	Transposase	6
		Y1_Tnp domain-containing protein	1
		Transposase IS200-family protein	1
127	8	Uncharacterized protein	5
		Orf140a gp	1
		AcrIIA5	1

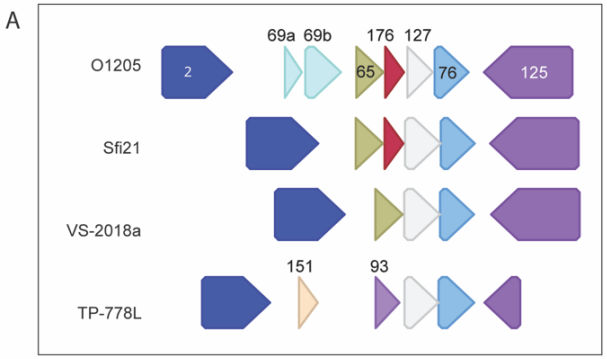
		ORF56	1
128	8	Uncharacterized protein	5
		Orf80 (92) gp	1
		ORF6	1
		Orf93	1
129	7	Peptidase_M78 domain-containing protein	2
		CI-like repressor	1
		Orf122 gp	1
		CI-like repressor metal proteinase motif protein	1
		Putative metallo-proteinase	1
		ORF3	1
130	7	Uncharacterized protein	6
		ORF22	1
131	7	Uncharacterized protein	7
132	7	Uncharacterized protein	7
133	6	Orf141 gp	2
		Putative holin	1
		Holin	1
		ORF49	1
		ORF42	1
134	6	Uncharacterized protein	6
135	6	Uncharacterized protein	6
136	6	Uncharacterized protein	6
137	6	Uncharacterized protein	6
138	5	IbrA	4
		Putative phosphoadenosine phosphosulfate reductase	1
139	5	IbrB	4
		Putative chromosome partitioning protein	1
140	5	DNA methyltransferase	5
141	5	Uncharacterized protein	5
142	5	Uncharacterized protein	3
		Putative host cell surface-exposed lipoprotein	1
		Superinfection exclusion lipoprotein	1
143	5	Uncharacterized protein	2
		ORF37	1

		Gp57	1
		ORF44	1
144	5	Uncharacterized protein	5
145	5	Uncharacterized protein	5
146	5	Uncharacterized protein	4
		ORF32	1
147	4	Uncharacterized protein	4
148	4	Uncharacterized protein	4
149	4	Uncharacterized protein	4
150	4	Holin	4
151	4	Uncharacterized protein	3
		Orf73	1
152	4	Uncharacterized protein	4
153	4	Uncharacterized protein	4
	4	Uncharacterized protein	4
155	4	Uncharacterized protein	4
156	4	Uncharacterized protein	3
		ORF26	1
157	4	Uncharacterized protein	4
158	3	DNA methyltransferase	2
		Adenine specific methyltransferase	1
159	3	Uncharacterized protein	3
160	3	Uncharacterized protein	3
161	3	Uncharacterized protein	2
		DNA binding protein	1
162	3	Uncharacterized protein	3
163	3	Uncharacterized protein	1
		Gp68	1
		ORF1	1
164	3	Uncharacterized protein	3
165	3	Uncharacterized protein	3
166	3	Uncharacterized protein	3
167	2	Helicase	2
168	2	Type III restriction endonuclease subunit R	2
169	2	NTP-binding protein	2
170	2	Replication protein	2
171	2	Orf229 gp	1
		Uncharacterized protein	1

172	2	Uncharacterized protein	2
173	2	Uncharacterized protein	2
174	2	Uncharacterized protein	2
175	2	Uncharacterized protein	2
176	2	ORF55	1
		Orf74a gp	1
177	2	Uncharacterized protein	2
178	2	HTH cro/C1-type domain-containing protein	2
179	2	Uncharacterized protein	2
180	2	RHH_3 domain-containing protein	1
		Ribbon-helix-helix domain-containing protein	1
181	2	Uncharacterized protein	2
182	2	Uncharacterized protein	2
183	2	Uncharacterized protein	2
184	2	Putative excisionase	1
		Uncharacterized protein	1
185	2	Uncharacterized protein	2
186	2	Uncharacterized protein	2

Figure 3.2 AcrIIA5 genome neighborhoods.

Protein clusters were mapped back onto their encoding genes to identify candidate anti-CRISPR clusters based on guilt-by-association with AcrIIA5. **(A)** corresponds to genome neighborhoods of lysogenic phages while **(B)** corresponds to lytic phage genome neighborhoods. The boundaries for the genome neighborhood are set at the lysogenic or lysogenic replacement module boundaries. Proteins that define the modules and boundaries are listed in **(C)**. **(D)** shows the sequence similarity network for each neighboring protein cluster with the top annotation listed.



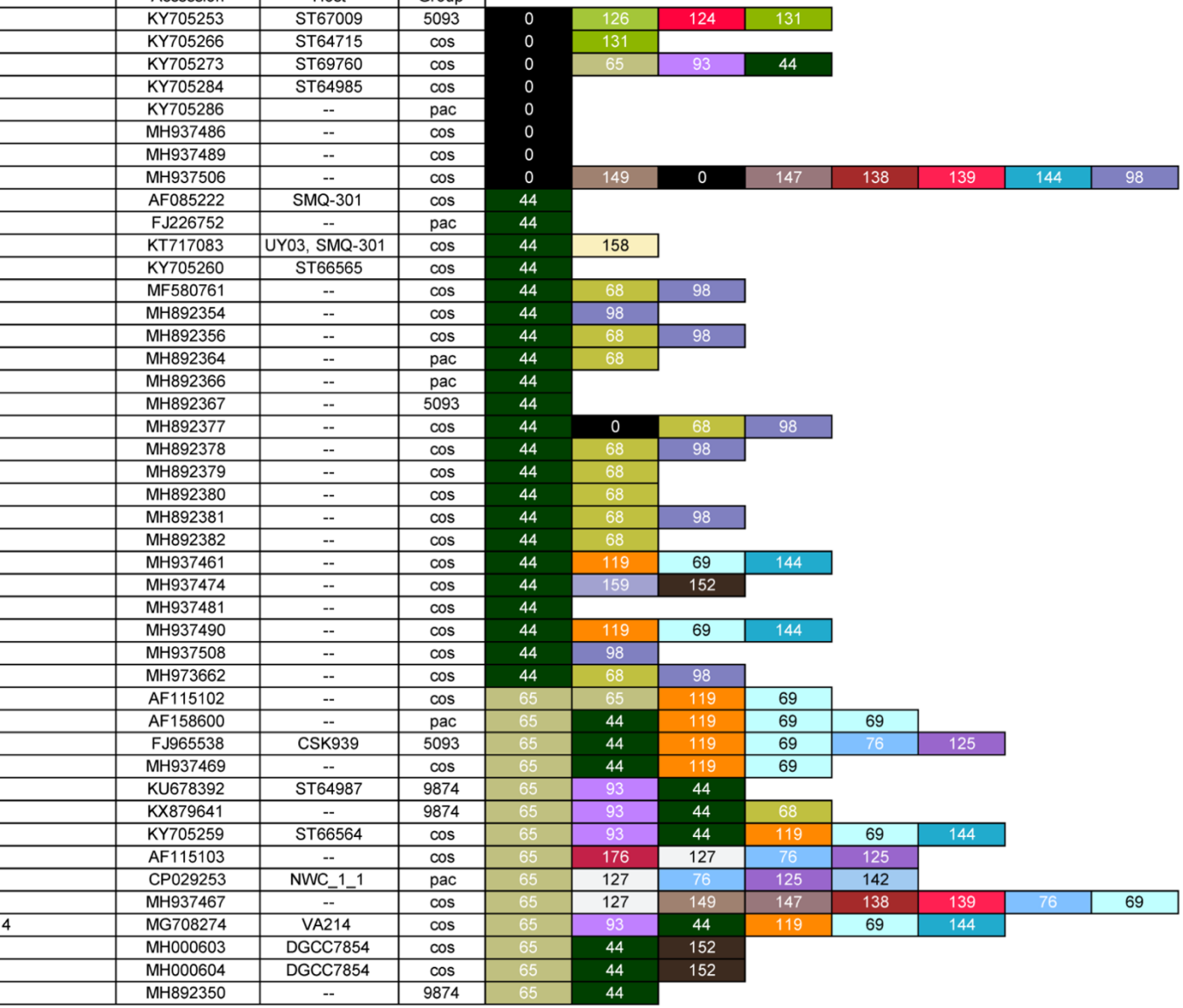
D

Cluster #	Cluster Network	Annotation
65		Uncharacterized
69	69a 69b	Uncharacterized
76		HTH cro/C1-type domain-containing
93		Uncharacterized
127		AcrIIA5
137		Uncharacterized
138		IbrA Putative phosphoadenosine phosphosulfate reductase
139		IbrB Putative chromosome partitioning protein
142		Putative host cell surface-exposed lipoprotein/ Superinfection exclusion lipoprotein
147		Uncharacterized
149		Uncharacterized
151		Uncharacterized
176		Uncharacterized

Figure 3.3 AcrIIA5 and AcrIIA6 homologues are limited to the lysogenic or lysogenic replacement module.

For phages with proteins within the bounds of the lysogenic or lysogenic replacement module, all candidate clusters are shown in the order that they appear in the phage genome. Previously reported AcrIIA5 (cluster 127) and AcrIIA6 (cluster 44) cluster members are noted. Cluster numbers and top annotations correspond to those in Table 3.1.

Name	Accession	Host	Group
Streptococcus phage P0093	KY705253	ST67009	5093
Streptococcus phage P7152	KY705266	ST64715	cos
Streptococcus phage P7602	KY705273	ST69760	cos
Streptococcus phage P9851	KY705284	ST64985	cos
Streptococcus phage P9853	KY705286	--	pac
Streptococcus phage CHPC1027	MH937486	--	cos
Streptococcus phage CHPC1034	MH937489	--	cos
Streptococcus phage CHPC1148	MH937506	--	cos
Streptococcus virus DT1	AF085222	SMQ-301	cos
Streptococcus virus ALQ132	FJ226752	--	pac
Streptococcus phage 73	KT717083	UY03, SMQ-301	cos
Streptococcus phage P5651	KY705260	ST66565	cos
Streptococcus phage 9A	MF580761	--	cos
Streptococcus phage SW3	MH892354	--	cos
Streptococcus phage SW21	MH892356	--	cos
Streptococcus phage SW15	MH892364	--	pac
Streptococcus phage SW18	MH892366	--	pac
Streptococcus phage SW19	MH892367	--	5093
Streptococcus phage SWK1	MH892377	--	cos
Streptococcus phage SWK2	MH892378	--	cos
Streptococcus phage SWK3	MH892379	--	cos
Streptococcus phage SWK4	MH892380	--	cos
Streptococcus phage SWK5	MH892381	--	cos
Streptococcus phage SWK6	MH892382	--	cos
Streptococcus phage CHPC642	MH937461	--	cos
Streptococcus phage CHPC930	MH937474	--	cos
Streptococcus phage CHPC979	MH937481	--	cos
Streptococcus phage CHPC1036	MH937490	--	cos
Streptococcus phage CHPC1156	MH937508	--	cos
Streptococcus phage SW7	MH973662	--	cos
Streptococcus virus Sfi19	AF115102	--	cos
Streptococcus virus Sfi11	AF158600	--	pac
Streptococcus phage 5093	FJ965538	CSK939	5093
Streptococcus phage CHPC919	MH937469	--	cos
Streptococcus virus 9874	KU678392	ST64987	9874
Streptococcus phage CHPC577	KX879641	--	9874
Streptococcus phage P5641	KY705259	ST66564	cos
Streptococcus virus Sfi21	AF115103	--	cos
Streptococcus phage VS-2018a	CP029253	NWC_1_1	pac
Streptococcus phage CHPC877	MH937467	--	cos
Streptococcus phage vB_SthS_VA214	MG708274	VA214	cos
Streptococcus phage D1024	MH000603	DGCC7854	cos
Streptococcus phage D1811	MH000604	DGCC7854	cos
Streptococcus phage SW16	MH892350	--	9874



Streptococcus phage SW11	MH892360	--	cos	65	44	69							
Streptococcus phage CHPC572	MH937458	--	cos	65	93	44							
Streptococcus phage CHPC640	MH937460	--	pac	65	93	44	148	76	69				
Streptococcus phage CHPC676	MH937463	--	pac	65	93	44	148	76	69				
Streptococcus phage CHPC873	MH937465	--	cos	65	93	44	148	76					
Streptococcus phage CHPC875	MH937466	--	cos	65	93	44							
Streptococcus phage CHPC954	MH937480	--	cos	65	93	44							
Streptococcus phage CHPC1042	MH937494	--	pac	65	93	44	69	69					
Streptococcus phage CHPC1091	MH937504	--	cos	65	44	69	69						
Streptococcus phage CHPC1246	MH937510	--	pac	65	44	159	69	69					
Streptococcus virus O1205	U88974	--	pac	69	69	65	176	127	76	125	0		
Streptococcus phage 53	KT717084	Uy01	cos	123	137								
Streptococcus virus 9871	KU678389	ST64987	9874	123									
Streptococcus virus 9872	KU678390	ST64987	9874	123									
Streptococcus virus 9873	KU678391	ST64987	9874	123									
Streptococcus phage P0095	KY705255	ST67009	5093	123	65	44							
Streptococcus phage P7571	KY705268	ST68757	pac	123	137								
Streptococcus phage P7574	KY705271	ST68757	cos	123	126	124							
Streptococcus phage P7952	KY705278	ST47795	pac	123	126	124	131						
Streptococcus phage SW17	MH892365	--	9874	123									
Streptococcus phage SW4	MH892355	--	5093	126	124								
Streptococcus phage SW27	MH892373	--	5093	126	124								
Streptococcus phage CHPC1248	MH937457	--	pac	126	124								
Streptococcus phage P0094	KY705254	ST67009	5093	131									
Streptococcus phage P7954	KY705280	ST47795	pac	131									
Streptococcus phage CHPC1073	MH937501	--	cos	131									
Streptococcus phage P5652	KY705261	ST66565	cos	132	44								
Streptococcus phage P7151	KY705265	ST64715	cos	132	44								
Streptococcus phage P7154	KY705267	ST64715	cos	132	44								
Streptococcus phage P7601	KY705272	ST69760	cos	132	0								
Streptococcus phage P7633	KY705276	ST69763	cos	132	44	68	0	98					
Streptococcus phage P8921	KY705282	ST64892	cos	132	44								
Streptococcus phage P8922	KY705283	ST64892	cos	132	44								
Streptococcus phage D4276	MF161328	DGCC7854	cos	135	65	127	149	147	138	139	76	69	
Streptococcus phage P7953	KY705279	ST47795	pac	135	44								
Streptococcus phage CHPC1046	MH937496	--	cos	135									
Streptococcus phage CHPC1048	MH937497	--	cos	135									
Streptococcus phage P4761	KY705258	ST64476	pac	137	127	76	0						
Streptococcus phage P7132	KY705262	ST64713	cos	137									
Streptococcus phage P7572	KY705269	ST68757	pac	137									
Streptococcus phage P7573	KY705270	ST68757	cos	137	175								
Streptococcus phage CHPC1230	MH937509	--	pac	138	139	76	69	69	126				
Streptococcus phage SW12	MH892361	--	cos	141	44								
Streptococcus phage 20167	HG424323	DSM 20167	pac	141	44	119	69	76	0	125	142		
Streptococcus phage vB_SthS_VA460	MG708275	VA460	cos	141	44	69							

Streptococcus phage CHPC663	MH937462	--	cos	141	44	98	76	1				
Streptococcus phage Javan389 (*oralis* phage)	MK448735	ATCC_49296	--	141	44	0	0	125	0	0		
Streptococcus virus 2972	AY699705	--	pac	146								
Streptococcus virus 858	EF529515	--	pac	146								
Streptococcus phage CHPC1008	MH937484	--	pac	146								
Streptococcus phage CHPC1057	MH937498	--	pac	146								
Streptococcus phage CHPC1041	MH937493	--	cos	148	44	152	69	69				
Streptococcus phage TP-J34	HE861935	--	pac	151	65	44	119	69	76	125	142	
Streptococcus phage TP-778L	HG380752	SK778	pac	151	93	127	76	125	142			
Streptococcus phage CHPC929	MH937473	--	pac	151	0	65	44	146				
Streptococcus phage CHPC1062	MH937499	--	cos	159	127	149	147	138	139	76		
Streptococcus phage P7631	KY705274	ST69763	cos	164	44	68	0	98				
Streptococcus phage P7632	KY705275	ST69763	cos	164	126	124						
Streptococcus phage P7951	KY705277	ST47795	pac	164	131							
Streptococcus phage P9901	KY705288	ST62990	cos	166	44							
Streptococcus phage CHPC595	MH937459	--	cos	166	65	44						
Streptococcus phage CHPC1045	MH937495	--	cos	166								
Streptococcus phage P9902	KY705289	ST62990	cos	182	44	158						
Streptococcus phage P9903	KY705290	ST62990	cos	182	44	0	158					
Streptococcus phage P7955	KY705281	ST47795	pac	185								
Streptococcus phage L5A1	MF580769	--	cos	185								

Figure 3.4 Plasmid and phage defense assays for anti-CRISPR screening.

(A) CRISPR-Cas systems in the *S. thermophilus* DGCC 7710 strain with the Type III-A CRISPR-Cas system transplanted from *S. thermophilus* JIM 8232. The boundaries of the transplant are from the upstream gene's stop codon to the downstream gene's start codon. **(B)** Overview of the plasmid-based defense assay. Anti-CRISPR candidates are encoded on pTRK882. Target sequences for each of the CRISPR-Cas systems correspond to the spacers colored in (A) with PAM or PFS sequences noted. The readout for anti-CRISPR activity is colony formation following target plasmid transformation. **(C)** Spacers corresponding to each of the CRISPR-Cas system BIMs are mapped to Phage 2972 genes with color of the gene corresponding to temporal expression during phage infection of *S. thermophilus* DGCC 7710. The annotation of the phage gene hit by each spacer as well as the PAM or PFS sequences are shown below. **(D)** Overview of the Phage-based anti-CRISPR assay. BIMs for the corresponding CRISPR-Cas system are challenged with lytic phage 2972. The screening outputs are growth curves based on optical density (OD_{600}). Culture lysis corresponds to a positive anti-CRISPR screening result.

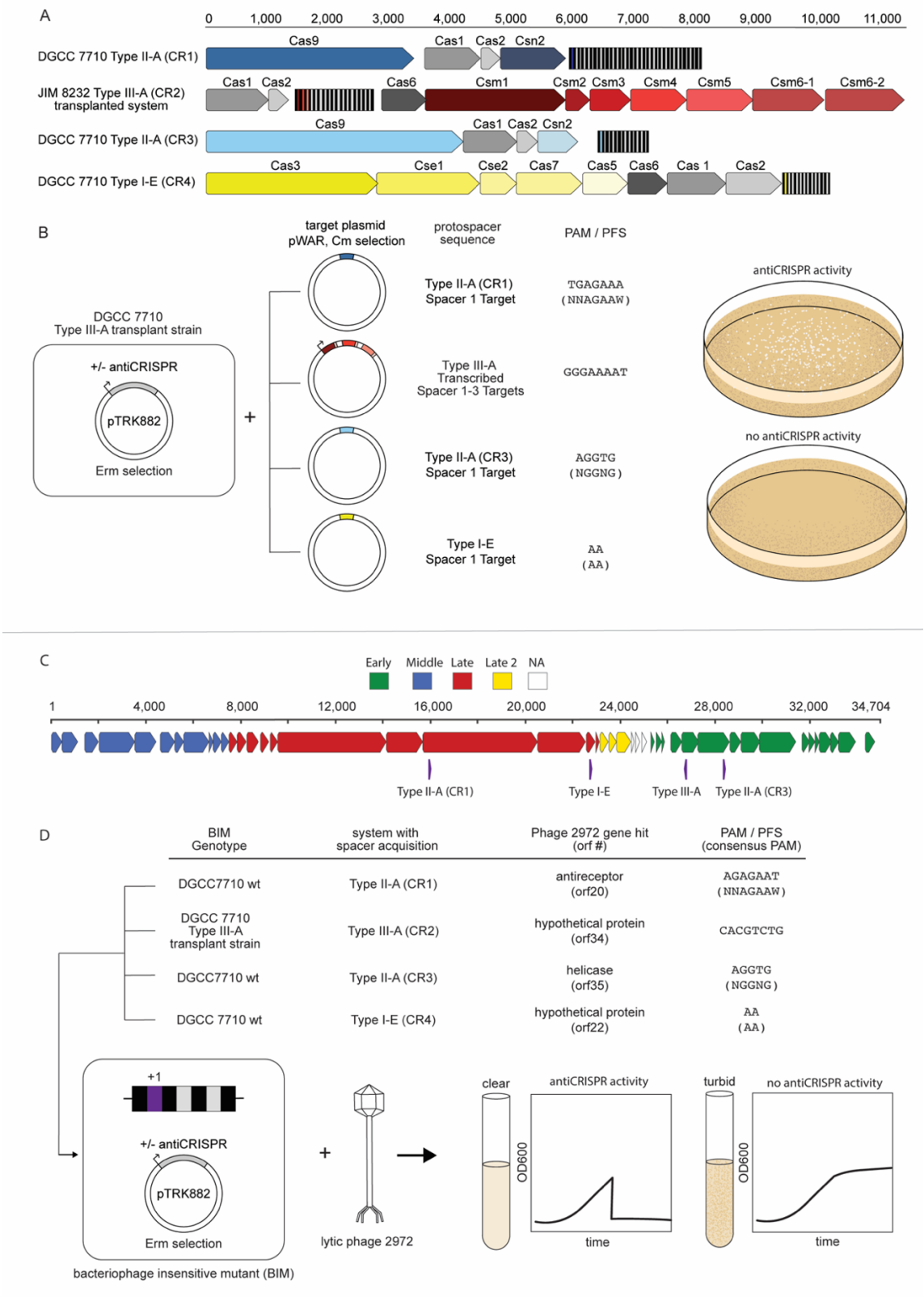

A
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
11,000
DGCC 7710 Type II-A (CR1)
Cas9
Cas1
Cas2
Csn2
JIM 8232 Type III-A (CR2) transplanted system
Cas1
Cas2
Cas6
Csm1
Csm2
Csm3
Csm4
Csm5
Csm6-1
Csm6-2
DGCC 7710 Type II-A (CR3)
Cas9
Cas1
Cas2
Csn2
DGCC 7710 Type I-E (CR4)
Cas3
Cse1
Cse2
Cas7
Cas5
Cas6
Cas 1
Cas2
B
target plasmid
pWAR, Cm selection
protospacer sequence
PAM / PFS
DGCC 7710
Type III-A transplant strain
+/- antiCRISPR
pTRK882
Erm selection
+
Type II-A (CR1)
Spacer 1 Target
TGAGAAA
(NNAGAAW)
Type III-A
Transcribed
Spacer 1-3 Targets
GGGAAAAT
Type II-A (CR3)
Spacer 1 Target
AGGTG
(NGGNG)
Type I-E
Spacer 1 Target
AA
(AA)
antiCRISPR activity
no antiCRISPR activity
C
Early
Middle
Late
Late 2
NA
1
4,000
8,000
12,000
16,000
20,000
24,000
28,000
32,000
34,704
Type II-A (CR1)
Type I-E
Type III-A
Type II-A (CR3)
D
BIM Genotype
system with spacer acquisition
Phage 2972 gene hit (orf #)
PAM / PFS (consensus PAM)
DGCC7710 wt
Type II-A (CR1)
antireceptor (orf20)
AGAGAAT
(NNAGAAW)
DGCC 7710 Type III-A transplant strain
Type III-A (CR2)
hypothetical protein (orf34)
CACGTCTG
DGCC7710 wt
Type II-A (CR3)
helicase (orf35)
AGGTG
(NGGNG)
DGCC 7710 wt
Type I-E (CR4)
hypothetical protein (orf22)
AA
(AA)
+1
+/- antiCRISPR
pTRK882
Erm selection
bacteriophage insensitive mutant (BIM)
+
lytic phage 2972
clear
OD600
antiCRISPR activity
time
turbid
OD600
no antiCRISPR activity
time

Figure 3.5 Cluster 93 contains a Type II-A anti-CRISPR.

(A) Plasmid-based anti-CRISPR assay results are shown for cluster 93 member P5641_25 compared to an empty pTRK882 vector and positive control AcrIIA5. AcrIIA5 specifically inhibits the Type II-A CRISPR1 system while P5641_25 specifically inhibits the Type II-A CRISPR3 system. **(B)** Phage-based anti-CRISPR assay results at an MOI of 0.1 indicate that P5641_25 is a Type II-A CRISPR3 specific anti-CRISPR protein. Positive control AcrIIA5 shows inhibition of both Type II-A CRISPR1 and CRISPR3 systems in the phage-based assay. **(C)** Genome neighborhoods for all cluster 93 members in our phage dataset are shown with co-occurring clusters annotated. On the right are genome neighborhoods for P5641_25 BLAST hits outside of our phage dataset. The cutoff for a hit was considered a BIT score of greater than or equal to 40.

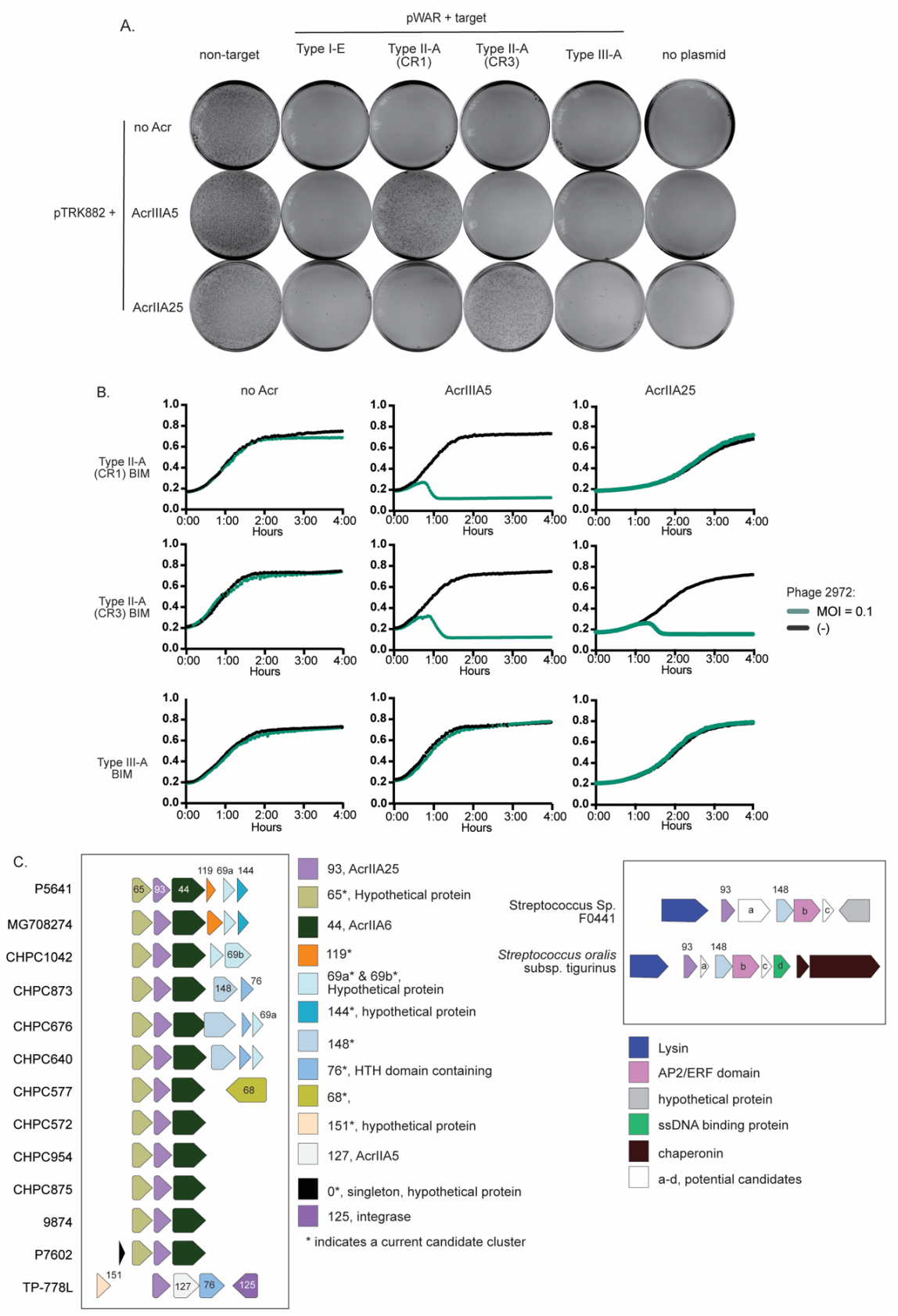

A.
pWAR + target
non-target
Type I-E
Type II-A (CR1)
Type II-A (CR3)
Type III-A
no plasmid
pTRK882 +
no Acr
AcrIIIA5
AcrIIA25
B.
no Acr
AcrIIIA5
AcrIIA25
Type II-A (CR1) BIM
Type II-A (CR3) BIM
Type III-A BIM
Hours
Phage 2972:
MOI = 0.1
(-)
C.
P5641
MG708274
CHPC1042
CHPC873
CHPC676
CHPC640
CHPC577
CHPC572
CHPC954
CHPC875
9874
P7602
TP-778L
93, AcrIIA25
65*, Hypothetical protein
44, AcrIIA6
119*
69a* & 69b*, Hypothetical protein
144*, hypothetical protein
148*
76*, HTH domain containing
68*,
151*, hypothetical protein
127, AcrIIA5
0*, singleton, hypothetical protein
125, integrase
* indicates a current candidate cluster
Streptococcus Sp. F0441
Streptococcus oralis subsp. tigurinus
Lysin
AP2/ERF domain
hypothetical protein
ssDNA binding protein
chaperonin
a-d, potential candidates

Figure 3.6 Cluster 119 contains a Type II-A anti-CRISPR with functionality specific to inhibition of anti-phage defense activity.

(A) Plasmid-based anti-CRISPR assay results are shown for cluster 119 member Sfi11_gp83 compared to an empty pTRK882 vector and positive control AcrIIA5. AcrIIA5 specifically inhibits the Type II-A CRISPR1 system while demonstrates no activity in the plasmid-based assay. **(B)** Phage-based anti-CRISPR assay results at an MOI of 0.1 indicate that Sfi11_gp83 is a Type II-A CRISPR3 specific anti-CRISPR protein. The lag in growth seen for CRISPR1 does not lead to lysis during the 24-hour assay. Positive control AcrIIA5 shows inhibition of both Type II-A CRISPR1 and CRISPR3 systems in the phage-based assay. **(C)** Genome neighborhoods for all cluster 119 members in our phage dataset are shown with co-occurring clusters annotated. On the right are genome neighborhoods for Sfi11_gp83 BLAST hits outside of our phage dataset. The cutoff for a hit was considered a BIT score of greater than or equal to 40.

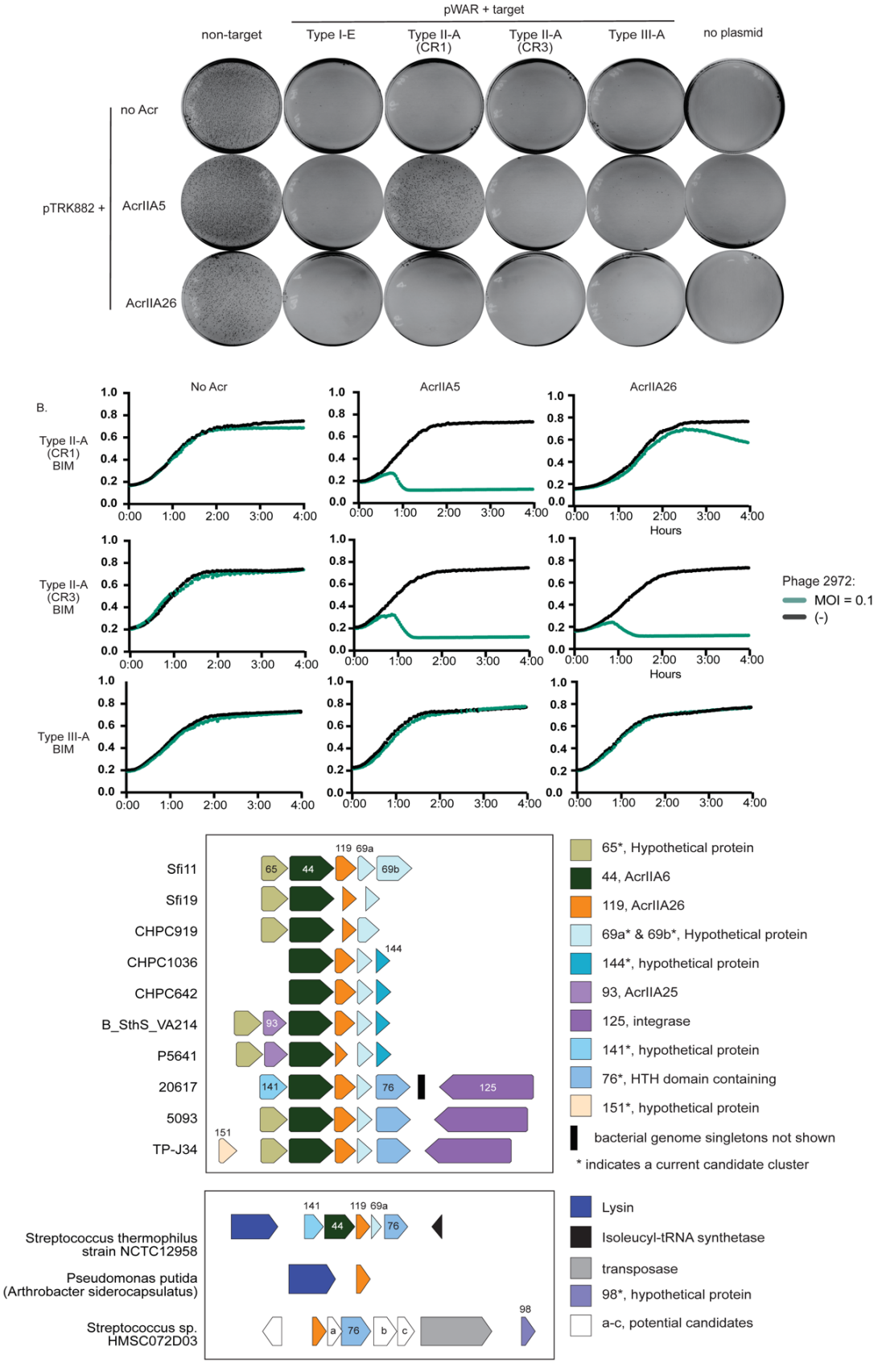
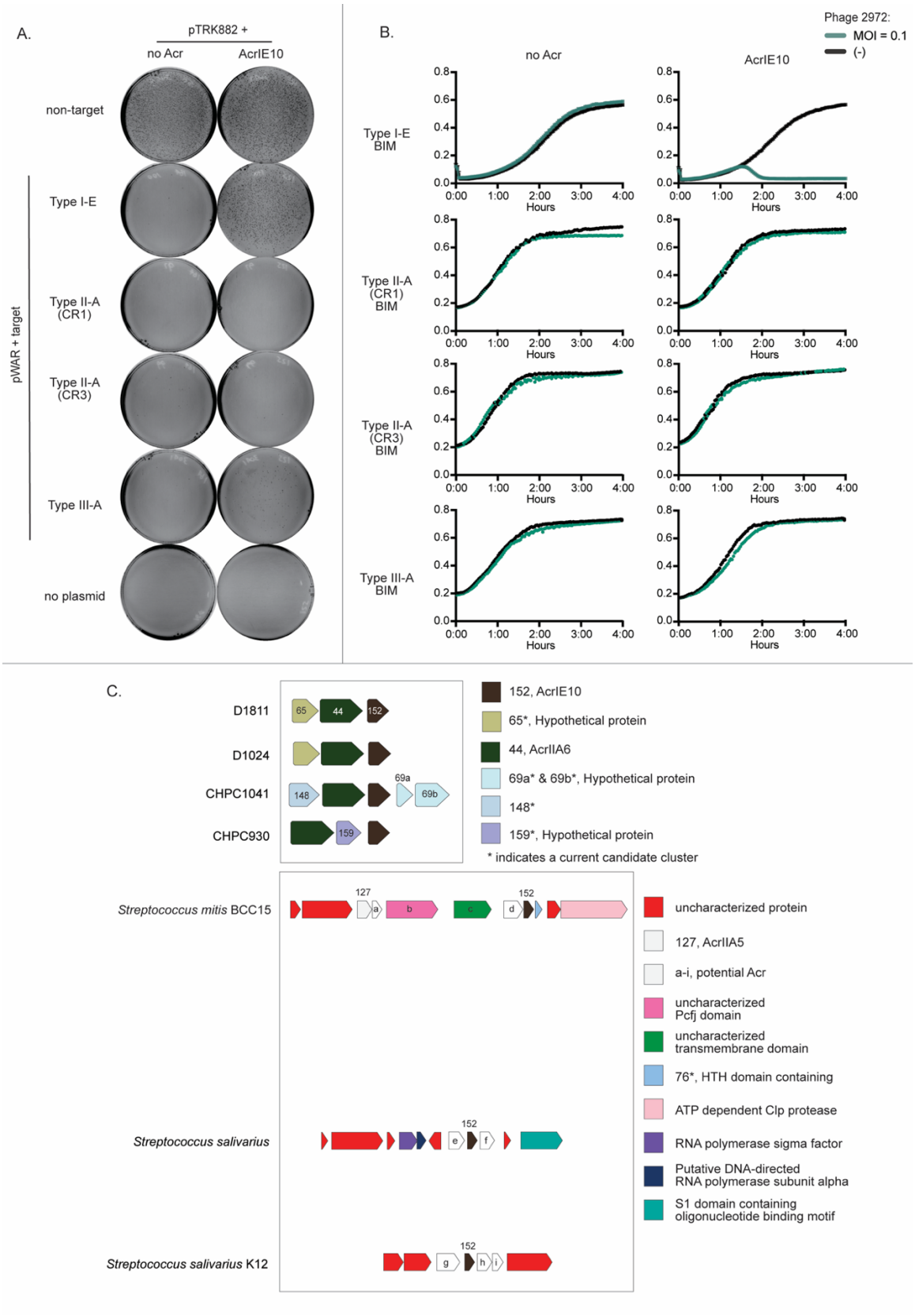

pWAR + target
non-target
Type I-E
Type II-A (CR1)
Type II-A (CR3)
Type III-A
no plasmid
pTRK882 +
no Acr
AcrIIA5
AcrIIA26
B.
No Acr
AcrIIA5
AcrIIA26
Type II-A (CR1) BIM
Type II-A (CR3) BIM
Type III-A BIM
Hours
Phage 2972:
MOI = 0.1
(-)
Sfi11
Sfi19
CHPC919
CHPC1036
CHPC642
B_SthS_VA214
P5641
20617
5093
TP-J34
65*, Hypothetical protein
44, AcrIIA6
119, AcrIIA26
69a* & 69b*, Hypothetical protein
144*, hypothetical protein
93, AcrIIA25
125, integrase
141*, hypothetical protein
76*, HTH domain containing
151*, hypothetical protein
bacterial genome singletons not shown
* indicates a current candidate cluster
Streptococcus thermophilus strain NCTC12958
Pseudomonas putida (Arthrobacter siderocapsulatus)
Streptococcus sp. HMSC072D03
Lysin
Isoleucyl-tRNA synthetase
transposase
98*, hypothetical protein
a-c, potential candidates

Figure 3.7 Cluster 152 contains a Type I-E anti-CRISPR.

(A) Plasmid-based anti-CRISPR assay results are shown for cluster 152 member D1811_027 compared to an empty pTRK882 vector. The strain harboring D1811_027 transforms the Type I-E target with comparable efficiency to the empty vector, indicating Type I-E anti-CRISPR activity. There is some transformation of the Type III-A target compared to the empty Acr vector control strain. However, transformation of the Type III-A target is not equivalent to the empty vector control, indicating potential weak activity. **(B)** Phage-based, anti-CRISPR assay results at an MOI of 0.1 indicate that D1811_027 is a Type I-E specific anti-CRISPR. No Type III-A activity is seen in this phage-based context. **(C)** Genome neighborhoods for all cluster 152 members in our phage dataset are shown with cooccurring clusters annotated. Below are genome neighborhoods for D1811_027 BLAST hits outside of our phage dataset. The cutoff for a hit was considered a BIT score of greater than or equal to 40.


A.
pTRK882 +
no Acr
AcrIE10
non-target
Type I-E
Type II-A (CR1)
Type II-A (CR3)
Type III-A
no plasmid
pWAR + target
B.
Phage 2972:
MOI = 0.1
(-)
no Acr
AcrIE10
Type I-E BIM
Type II-A (CR1) BIM
Type II-A (CR3) BIM
Type III-A BIM
Hours
C.
D1811
D1024
CHPC1041
CHPC930
65
44
152
148
69a
69b
159
152, AcrIE10
65*, Hypothetical protein
44, AcrIIA6
69a* & 69b*, Hypothetical protein
148*
159*, Hypothetical protein
* indicates a current candidate cluster
Streptococcus mitis BCC15
Streptococcus salivarius
Streptococcus salivarius K12
127
a
b
c
d
152
e
f
g
h
i
uncharacterized protein
127, AcrIIA5
a-i, potential Acr
uncharacterized Pcfj domain
uncharacterized transmembrane domain
76*, HTH domain containing
ATP dependent Clp protease
RNA polymerase sigma factor
Putative DNA-directed RNA polymerase subunit alpha
S1 domain containing oligonucleotide binding motif

REFERENCES

1. Mojica FJM, Díez-Villaseñor Cs, García-Martínez J, Soria E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. Journal of Molecular Evolution. 2005;60(2):174-82.

2. 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(5819):1709-12.

3. Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature. 2013;493(7432):429-32.

4. Li Y, Bondy-Denomy J. Anti-CRISPRs go viral: the infection biology of CRISPR-Cas inhibitors. Cell host & microbe. 2021.

5. Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature. 2013;493(7432):429-32.

6. Pawluk A, Bondy-Denomy J, Cheung VHW, Maxwell KL, Davidson AR. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. mBio. 2014;5(2):e00896-e.

7. Achigar R, Magadán AH, Tremblay DM, Julia Pianzzola M, Moineau S. Phage-host interactions in Streptococcus thermophilus: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. Scientific Reports. 2017;7:43438.

8. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, et al. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. Journal of bacteriology. 2008;190(4):1401-12.

9. Le Marrec C, van Sinderen D, Walsh L, Stanley E, Vlegels E, Moineau S, et al. Two groups of bacteriophages infecting Streptococcus thermophilus can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. Applied and Environmental Microbiology. 1997;63(8):3246-53.

10. Lucchini S, Desiere F, Brüssow H. Comparative genomics of Streptococcus thermophilus phage species supports a modular evolution theory. J Virol. 1999;73(10):8647-56.

11. Hynes AP, Rousseau GM, Lemay ML, Horvath P, Romero DA, Fremaux C, et al. An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. Nature microbiology. 2017;2(10):1374-80.

12. Zallot R, Oberg NO, Gerlt JA. 'Democratized' genomic enzymology web tools for functional assignment. Current Opinion in Chemical Biology. 2018;47:77-85.

13. Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry. 2019;58(41):4169-82.

14. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2015;1854(8):1019-37.

15. Gerlt JA. Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence–Function Space and Genome Context to Discover Novel Functions. Biochemistry. 2017;56(33):4293-308.

16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

17. Madeira F, Park Ym, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Research. 2019;47(W1):W636-W41.

18. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology. 2018;430(15):2237-43.

19. Gabler F, Nam S-Z, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Current Protocols in Bioinformatics. 2020;72(1):e108.

20. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.

21. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. Genome Biol. 2019;20(1):185.

22. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun. 2018;9(1):2919.

23. Song G, Zhang F, Zhang X, Gao X, Zhu X, Fan D, et al. AcrIIA5 Inhibits a Broad Range of Cas9 Orthologs by Preventing DNA Target Cleavage. Cell reports. 2019;29(9):2579-89 e4.

24. Foley S, Lucchini S, Zwahlen MC, Brussow H. A short noncoding viral DNA element showing characteristics of a replication origin confers bacteriophage resistance to Streptococcus thermophilus. Virology. 1998;250(2):377-87.

25. Hynes AP, Moineau S. Phagebook: The Social Network. Mol Cell. 2017;65(6):963-4.

26. Fuchsbauer O, Swuec P, Zimberger C, Amigues B, Levesque S, Agudelo D, et al. Cas9 Allosteric Inhibition by the Anti-CRISPR Protein AcrIIA6. Mol Cell. 2019;76(6):922-37 e7.

27. Liang M, Sui T, Liu Z, Chen M, Liu H, Shan H, et al. AcrIIA5 Suppresses Base Editors and Reduces Their Off-Target Effects. Cells. 2020;9(8).

28. Garcia B, Lee J, Edraki A, Hidalgo-Reyes Y, Erwood S, Mir A, et al. Anti-CRISPR AcrIIA5 Potently Inhibits All Cas9 Homologs Used for Genome Editing. Cell reports. 2019;29(7):1739-46 e5.

29. Fontaine L, Dandoy D, Boutry C, Delplace B, de Frahan MH, Fremaux C, et al. Development of a versatile procedure based on natural transformation for marker-free targeted genetic modification in Streptococcus thermophilus. Appl Environ Microbiol. 2010;76(23):7870-7.

CHAPTER 4

FEATURES AND SELECTIVITY OF *STREPTOCOCCUS THERMOPHILUS* CRISPR-CAS SYSTEM TARGETING OF PHAGE GENES[3]

[3] Clare Cooper and Michael P. Terns. Features and selectivity of *Streptococcus thermophilus* CRISPR-Cas system targeting of phage genes. *In preparation*.

ABSTRACT

On average, *Streptococcus thermophilus* Type III-A CRISPR-Cas systems maintain very few spacer sequences, but are conserved across many strains. Because of their unique ability to bind and degrade target RNA sequences, we wanted to know if Type III-A systems preferentially target and maintain spacers against certain phage genes. We chose to compare the targeting specificity of Type III-A CRISPR-Cas systems to co-occurring system types in *S. thermophilus*. To do this, we extracted spacer sequences for each of the systems in *S. thermophilus* and mapped them to phage gene clusters. We tested the correlation of spacer hits to relative abundance and length of phage genes (for Type III-A spacers hits) and to PAM sequence abundance (for Type I-E and Type II-A spacer hits). For Type III-A systems, there was no significant linear correlation between length and abundance of genes within a cluster to the number of spacer hits that map to the cluster, even when considering the subset of genes with mapped spacers. The spacer hits were especially enriched against the type II-A anti-CRISPR protein, AcrIIA6. Overall, increase knowledge of the specificity of spacer sequences that are maintained by Type III-A systems and implicate Type III-A systems as inhibitors of anti-CRISPR activity through phage gene transcript degradation.

INTRODUCTION

In strains of *S. thermophilus*, there can be up to four co-occurring CRISPR-Cas systems (1). The Type II-A systems are predominant in their acquisition of spacers against invaders with the CRISPR1 locus preferentially acquiring spacers during phage infection (2). Despite this, as we demonstrated in Chapter 2, Type III-A systems are highly conserved in these strains. However, their actual relevance to defense in strains with dominant Type II-

A systems is still not well understood. Identification of widespread anti-CRISPR (Acr) proteins against the Type II-A systems of *S. thermophilus* brings the potential importance of the co-occurring Type III-A system back into focus (3).

One of the unique functionalities of Type III-A systems is their sequence-specific RNase activity (4). This activity can regulate gene expression through transcript degradation and is relevant to defense against phage infection (5). We wanted to know if there is selection for Type III-A spacers against certain phage genes or if spacers are maintained against genes according to the general abundance and availability of protospacer sequences for acquisition. We hypothesized that DNA targeting systems (Type I-E and Type II-A) would likely show little bias for targeted phage genes while Type III-A systems, capable of target transcript degradation, would show preferential acquisition of certain phage genes with less regard to total gene length and abundance of phage genes. To test this, we brought together and expanded upon analyses from previous chapters.

In Chapter 3 we clustered *S. thermophilus* phage proteins based on homology and used the most common annotations for each cluster to predict a function for the proteins within the cluster (Figure 3.1 and Table 3.1). In this chapter, we assigned the same cluster numbers to the representative gene for each phage protein. Additionally, we extracted all spacer sequences from our *S. thermophilus* genome dataset (Chapter 2) to BLAST against the phage genes from the clustering analysis (Chapter 3). We then extracted the top hits and determined the orientation of the hits relative to the directionality of the gene. For Type II-A and I-E systems, we compare total or unique hits to abundance of PAM sequences within the corresponding phage gene clusters. For the *S. thermophilus* Type III-A systems, there is no consensus protospacer flanking sequence (PFS) adjacent to mapped

spacers (Figure 2.3), so we compared the total and unique hits to total length and count of sequences within the phage gene clusters. We show that Type III-A spacer hits against Type II-A anti-CRISPRs are common in strains of *Streptococcus thermophilus*, and we lay out an experimental plan to test this hypothesis.

RESULTS

Type II-A spacer hits correlate to relative PAM sequence abundance

We wanted to determine if Type II-A systems preferentially acquire and maintain spacers against certain phage genes or if hits correspond to relative PAM abundance. *S. thermophilus* Type II-A systems have a well-characterized PAM sequences that are required for both novel spacer acquisition and crRNA-guided Cas9 nuclease destruction (6-8), so we needed to account for the relative abundance of available PAM sequences within each gene of the phage dataset. To do this, we used the published PAM sequences of 'NNAGAA' for CR1 and 'NGGNG' for CR3 (6-8). We annotated all PAMs and enumerated the number of PAMs in each orientation for each gene in the dataset. We then totaled the number of PAMs in both orientations for genes that correspond to the same assigned cluster to give the total available PAM sequences for a phage gene cluster. This total PAM count is a product of both the sequence of the cluster members as well as the total number of sequences within a gene cluster. The assigned cluster number is inversely proportional to the number of genes within a cluster, so cluster 1 has the most genes and cluster 186 has the least.

The graphs in figures 4.1 (CR1) and 4.2 (CR3) show the total number of Type II-A spacers that map to each phage gene cluster as bars that correspond to the left y-axis.

The forward oriented hits are in red, and the reverse oriented hits are in blue. The line graphs correspond to the right y-axis and are a measure of the total number of PAM sequences within each gene cluster. The forward oriented PAM counts are in red, and the reverse oriented PAMs are in blue.

To quantify if the total PAM sequence counts correlate to the pattern we see for total spacer hits against each cluster, we calculated a Pearson correlation coefficient. The CR1 spacer hits and PAM counts for all clusters have a correlation coefficient of 0.837 (n=372, t-statistic=29.37, p=1.01E-98). This indicates a significant linear correlation exists between CR1 total PAM counts and total spacer hits. The correlation coefficient for CR3 is 0.807 (n=372, t-statistic=26.26, p=1.47E-86), also indicating a significant linear correlation between 'NGGNG' PAM count and total spacer hits from the CR3 locus. Of note, the total Type II-A availability of PAM sequences on the forward strand is greater than the total availability of PAM sequences on the reverse strand, which is consistent with the strand bias we saw previously for Type II-A targeting in Figure 2.2.

Type I-E unique spacer hits correlate to relative PAM sequence abundance

For the Type I-E System, there is less variability in the spacer composition between arrays. Out of 15 Type I-E arrays, 9 have the same spacer composition (Figure 4.3A). For this reason, we considered unique and total spacer hits and excluded the clusters with 0 hits. For studying unique spacer hits, duplicate spacer sequences were removed from the dataset prior to BLAST analysis. In addition, the Type I-E system has a known PAM sequence, 'AA,' which we verified in Figure 2.2 (9). We again mapped the total number of reverse and forward strand PAM sequences for each cluster and calculated a correlation coefficient

to understand if the total PAM counts correlate to unique acquisition events. The correlation coefficient is 0.744 (n=31, t-statistic=5.99, p=1.6E-6) for unique spacer hits and 0.4759 (n=31, t-statistic=2.91, p=0.007) for total spacer hits. The correlation coefficient for unique spacer hits indicates a significant high-positive linear correlation between total Type I-E PAM sequence counts and spacer acquisition events against the phage sequence clusters, while the total spacer hits indicate a low-positive correlation. This low positive correlation for total hits could be due to 9 arrays having the same overall spacer composition, indicating that these were not likely unique spacer acquisition events (Figure 4.3B).

Type III-A spacer hits do not strongly-correlate to relative gene abundance even in the subset of acquired genes

For the Type III-A systems we completed analyses of unique and total spacer hits. The mapping of unique spacer hits to phage gene clusters is shown in figure 4.4. Of note, because the crRNA of Type II-A and Type I-E systems pairs with a target DNA sequence, the directionality of the spacer hit relative to directionality of the gene should have little bearing on targeting. However, for Type III-A systems, the crRNA pairs with a target RNA, so defense activation is dependent on transcription of the targeted gene (10). The spacer must map in the reverse orientation of the gene to pair with the transcript and lead to function. We see this directionality bias in the number of reverse oriented (blue) spacer hits.

Because the Type III-A system does not have a consensus PAM sequence, we wanted to know if the trend we see for spacer acquisition is due to a combination of the

variability in number of genes within a cluster as well as the length of the genes within the cluster. To account for this, we totaled the length of all genes within each cluster as the total sequence length available for spacer acquisition. We again calculated a Pearson correlation coefficient to understand if the trend in spacer acquisition correlates to variability in total sequence length of the phage gene clusters. Considering all potential genes that could be acquired in the population, the correlation coefficient is 0.441 (n=186, t-statistic=6.66, p=3.05E-10) for unique spacer hits and 0.349 (n=186, t-statistic=5.05, p=1.07E-6) for total spacer hits. Therefore, there is a significant low-positive correlation between total cluster sequence length and unique and total acquisition events. This indicates that there are likely other factors contributing to Type III-A spacer acquisition or maintenance outside of the relative abundance and size of genes.

Previous studies demonstrate that Type III-A systems are better able to defend against phage infection when targeting genes expressed early in infection (11, 12). This could be a simple explanation for why we see spacer hits that do not strongly correlate to the relative sequence lengths and abundance of phage gene clusters, if only early phage genes are targeted. For instance, the late-expressed scaffolding genes are some of the largest in the phage genome while the early-expressed regulatory genes are some of the smallest Figure 1.5 (13-15). We do not have sufficient data on temporal expression for all of the gene clusters in our dataset. So, to try to account for this, we calculated the correlation coefficient without consideration of clusters with 0 acquisition events. If there is selective preference for some subset of phage genes, we hypothesized that within that subset of genes, spacer hits would correlate to relative abundance of nucleotide sequences available for acquisition. In other words, if there is no additional selective pressure, we

expected that within the subgroup of clusters that are targeted, there should be a linear trend based on relative abundance and sequence length of the cluster members. For this analysis, the correlation coefficient is 0.332 (n=37, t-statistic=2.09, p=0.044) for unique spacer hits and 0.226 (n=37, t-statistic=2.373, p=0.178) for total spacer hits. For unique spacer hits this again indicates a significant low-positive relationship between spacer hits and total length of genes within a cluster. For total spacer hits there is no significant correlation. Moving forward, we hypothesized that the spacers that are maintained serve some selective function.

S. thermophilus Type III-A arrays target Type II-A anti-CRISPRs, AcrIIA5 and AcrIIA6

Table 4.1 lists the gene clusters from greatest to least total Type III-A spacer hits. Based on the analysis of Type III-A spacer hits (figure 4.4 and Table 4.1), out of the 40 phage gene clusters that are targeted by the Type III-A system, AcrIIA6 homologues (cluster 44) have the second highest number of unique and total spacer hits (3, 16). 6 unique spacer sequences and 25 total spacer sequences map to AcrIIA6, making up 7% and 13% of unique and total spacers, respectively. AcrIIA5 homologues (cluster 127) have the ninth highest number of unique hits, with 3 unique spacer sequences (3% of all unique spacers) and 4 total spacer hits (2.1% of total spacers) mapping to AcrIIA5.

Alignments of AcrIIA6 spacer hits are shown in figure 4.5. Because Type III-A systems rely on crRNA pairing with target RNA, G-U base pairing has been added to the alignment. Additionally, there is variability in the homologues of AcrIIA6 along the length of the transcript, so dependent on where a spacer hits, it may have varying specificity for

different allele combinations. We have listed representative alleles of AcrIIA6 that have a BIT score cut-off above 40 for each unique sequence in the region of spacer binding. At the top of figure 4.5, the spacers are mapped along the length of the AcrIIA6 transcript with percentages indicating the percent of AcrIIA6 alleles targeted above the BIT score cut-off. Similarly, AcrIIA5 spacer hits are shown in figure 4.6 (17-19).

Because of the enrichment of AcrIIA6 targeting spacers in the population, we wanted to know if Type III-A systems are capable of degrading transcripts of Acr proteins and inhibiting their functionality.

Future experimental approach to study the effect of Type III-A spacer sequences on Type II-A anti-CRISPR activity

Figure 4.7 outlines future directions to test the hypothesis that Type III-A CRISPR-Cas systems repress Type II-A Acr activity. As shown in Figure 4.5, *S. thermophilus* DGCC 7710 and *S. thermophilus* JIM 8232 both encode Type III-A CRISPR-Cas spacers against AcrIIA6. The Type III-A system of *S. thermophilus* JIM 8232 is capable of inhibiting transformation of plasmid encoded targets (20) as well as defending against phage (our unpublished results). For this reason, we want to use the *S. thermophilus* DGCC 7710 strain. We show in Chapter 2 that the *S. thermophilus* DGCC 7710 strain is unable to defend against phage lysis or prevent plasmid transformation. However, Csm3 and other Csm proteins are maintained, indicating potential functionality of the complex in target RNA binding and degradation (Table 2.2). Because of this, phage defense will hinge on the activity of the Type II-A CRISPR-Cas system, despite a Type III-A spacer against the phage-encoded AcrIIA6. We hypothesize that the Type III-A system will still have intact

Csm3 function (i.e. sequence-specific RNase activity) to degrade the AcrIIA6 transcript, allowing the Type II-A system to carry out defense. To have levels of anti-CRISPR expression comparable to native expression, we decided to use a phage-encoded AcrIIA6. This requires phage genome editing to insert AcrIIA6 into phage 2972 under the native promoter (21, 22). In the case that the Type III-A system inhibits AcrIIA6, we expect the culture to grow to turbidity. In the case that the Type III-A system is unable to inhibit AcrIIA6, we expect the culture to lyse. If the *S. thermophilus* DGCC 7710 system is negative for activity against AcrIIA6, we will use a transplanted Type III-A system from JIM 8232 with mutation of Csm1 HD active site and Csm6-1 and Csm6-2 HEPN active sites to study the function of Csm3.

DISCUSSION

We began this analysis by analyzing the targeting specificity of the Type II-A and Type I-E systems in *Streptococcus thermophilus* to use as a comparison for the Type III-A system gene targeting specificity. Our Type II-A (CR1 and CR3) spacer targeting results somewhat contradict the results of Paez-Espino et al. (23). They tested *S. thermophilus* DGCC 7710 adaptation against phage 2972 at MOI of 2 and 10 over the course of 15 days. They tracked the spacer acquisition and adaptation against the phage in the Type II-A CR1 and CR3 loci and found enrichment of certain 'super spacers' (23). There are many differences between their analysis and ours. One being that they were not looking at a strain under intense selective pressure at high MOIs of a single phage. Their spacer mapping along the length of the phage genome does appear to correlate to relative PAM sequence abundance (consistent with our results), but they did not calculate the correlation

coefficient or further analyze these trends. In addition, they focused on adaption events at a single point (PAM location) in a gene while we are looking at adaptation events along the entire length of all genes in a phage cluster. It could be worthwhile to look at total spacer events for each cluster to see if there are enrichment of certain spacer sequences within the cluster being targeted.

Regarding the Type III-A system of S. thermophilus (Figures 4.1, 4.2, and 4.3), we found a specificity of spacer acquisition and maintenance against phage gene clusters that does not strongly correlate to relative sequence abundance, even when considering only the subset of genes that are targeted. Instead, we see multiple unique spacer sequences against genes that are relatively underrepresented in the population (Figure 4.4 and Table 4.1). Our analysis of Type III-A spacers suggests that there could be a selectivity for spacers that target Acrs and other adaptive phage genes.

To test this hypothesis, we proposed experiments in Figure 4.7. We have considered and completed other approaches to test this hypothesis. (1) First, we used a Type II-A (CR1) BIM (the same setup shown in Figure 4.7) with the anti-CRISPR constitutively expressed from the same plasmid used in Chapter 3 (pTRK882 plasmid with Ppgm promoter). We then challenged the strain with phage 2972. However, we saw no effect of the Type III-A system targeting the plasmid and worried that we may have pre-loaded the cell with anti-CRISPR and overwhelmed the Csm3 activity with the strong constitutive expression of the Acr. Upon first encounter with a lytic phage, we would expect the anti-CRISPR to be expressed as the phage infects and not to be pre-loaded in the cell. (2) Because of this, we decided to move forward with a pTRK882 plasmid containing an inducible Ptet promoter for anti-CRISPR expression. After continued troubleshooting, we

saw no anti-CRISPR activity with this promoter and were concerned about the functionality of the promoter for protein expression.

We have made some progress towards the experimental goal outlined in figure 4.7. Concurrently with attempted use of the Ptet promoter, we decided to try phage editing to insert an anti-CRISPR into phage 2972. Editing of a lytic phage is completed using CRISPR-Cas9 to target a region of the phage genome while providing a repair template with the gene of interest flanked by homologous arms for recombination (21, 22, 24). We used a Type II-A (CR1) BIM to complete replacement of orf33 (22) with AcrIIA6 in the genome of phage 2972. The boundaries of the insertion were from start codon to the promoter region for the downstream gene, putting AcrIIA6 under the native promoter for gp33, an early expressed gene in the replication region (22). We verified the insertion using PCR, but could not purify the edited phage away from wildtype phage 2972. Each time we used a BIM to target phage 2972 gp33 for purification of the mutant phage, we would see no phage plaquing. We concluded that the mutant phage likely needed wt phage 2972 to coinfect the host due to fitness costs of the insertion.

Finally, we decided to insert the anti-CRISPR proteins into the lysogenic replacement module that we studied in Chapter 2, hoping this would have less of a fitness cost. In Chapter 2, we saw that this region commonly expands and contracts with variable gene expression, so it made sense to insert genes here. We created repair templates for anti-CRISPR insertion into the lysogenic replacement module including insertion of native promoters from this region. In addition, we created mini-array plasmids to target a small ORF maintained in the lysogenic replacement module by phage 2972. For editing, we decided to target the phage using the Type III-A system, which is much more difficult for

the phage to escape using simple point mutations and could potentially select for large insertions and deletions. This is where we left off prior to completing the second round of editing, so we have yet to see if this will be a valid approach.

Notably, the experimental approach outlined in Figure 4.7 is centered on testing the potential functionality of Csm3 in Type III-A systems that do not maintain activity of Csm1 and Csm6 for plasmid and phage degradation. The experimental plan does not answer the question of what happens when the strain maintains all Type III-A activity and is challenged with an anti-CRISPR targeted by the Type III-A system. For both the Type III-A system in the native S. thermophilus JIM 8232 strain as well as the transplanted Type III-A system in S. thermophilus DGCC 7710 (discussed in previous chapters), we are unable to transform a plasmid containing AcrIIA6 when it is constitutively promoted under Ppgm (our unpublished results). When either strain contains Csm6-1 and Csm6-2 HEPN active site mutations, the AcrIIA6 plasmid transforms comparable to the empty vector control. This is consistent with transforming a Type III-A target plasmid into these strains, thus lacking novelty other than the gene being targeted. However, we can question what the significance of this is compared to anti-CRISPR gene targeting carried out by the Type II-A or Type I-E systems. First, it is significant because of the additional level of Type III-A transcript degradation that could prevent anti-CRISPR activity while the Type III-A or other system is completing defense, especially if the anti-CRISPR had capability to shut down the Type III-A system. Second, the Type III-A system has been shown to apply significant selective pressure to phages during targeting (25). Because of the laxity in the PFS sequence requirements and target RNA base-pairing, the Type III-A system is largely resistant to point mutations in the target sequence (25). In comparison, the Type I-E and

Type II-A systems have strict PAM sequence requirements, allowing phage to acquire single point mutations in the PAM or seed sequence of the target to escape defense (25, 26). Point mutations induced by the Type I-E and Type II-A systems may not impact the functionality of the anti-CRISPR protein compared the larger mutations needed to escape Type III-A defense (25, 26). From this standpoint, an entirely active Type III-A system is a strong inhibitor of phage anti-CRISPRs ever entering the cell unless they have undergone mutagenesis to escape defense, which could impact their activity. Therefore, this activity is significant even when not novel per se, but it does not address the unique functionality of Type III-A systems that may no longer maintain defense capability.

Previous work by Landsberger et al. ((27)) and Borges et al. ((28)) looked at population level effects of phages carrying anti-CRISPR proteins. One take away from both is that phages cooperate to overcome CRISPR-Cas immunity. Landsberger et al. demonstrated that initial viral titer, number of phage-targeting spacers carried by the host, and relative strength of the anti-CRISPR protein are all factors that determine at the outcome of infection (27). Phages carrying anti-CRISPR proteins were not always successful at infecting a host with multiple phage-targeting spacers, especially at low viral titers. However, they found that initial infection of a BIM with a low MOI of an anti-CRISPR carrying phage would immunosuppress the host, even if the phage did not replicate and lyse the cell (27). This initial immune suppression was also demonstrated by Borges et al. and both groups demonstrated that successive infections of an immunosuppressed host could eventually lead to an epidemic (27, 28). Our results are perhaps more relevant on this population-based scale and would be an additional consideration to the factors demonstrated by Lansdberger et al. that determine the outcome

of phage infection on this larger scale. Similar analyses could be completed to study the impact of Acr targeting spacers at lower viral titers over larger windows of time.

Lastly, we also identified additional spacer sequences against our novel Type II-A anti-CRISPRs (from Chapter 3), but these hits were below the BIT score cutoff used in Figures 4.1, 4.2, 4.3 and 4.4. Because of the laxity of base pairing for Type III-A systems, these spacers could still be relevant to the function of these other anti-CRISPR proteins (29). After experimentally testing the current hypothesis, additional testing of Type III-A inhibition of anti-CRISPRs against the Type II-A (CR3) and Type I-E systems could strengthen our understanding of Type III-A co-occurrence with other systems in S. thermophilus and beyond.

In conclusion, we've demonstrated the potential specificity of the Csm system of S. thermophilus in targeting phage genes. We show that compared to the other CRISPR-Cas systems of S. thermophilus, the Csm system spacer hits do not correlate to features of the genes, such as the total length of all phage gene cluster members (Figure 4.4). This potential specificity for phage gene targeting could play a role in degrading gene transcripts during phage infection, especially given the target RNase activity of Csm3 within the Csm complex. We show that phage anti-CRISPR genes are preferentially targeted by the Type III-A system (Table 4.1, Figures 4.5 and 4.6). In the context of widespread Type II-A anti-CRISPR proteins, this phage gene specific targeting could abrogate the effects of Type II-A anti-CRISPRs and allow the Type II-A system to carry out defense. We proposed an experiment to test this hypothesis and discussed our progress on this thus far (Figure 4.7).

MATERIALS AND METHODS

Clustering of phage genes

Genomes of Streptococcal phages were downloaded from NCBI Assembly on 04/10/19. Proteomes of the associated genomes were downloaded from Uniprot and uploaded to EFI-EST for creation of a sequence similarity network (SSN) (30-33). Standard recommendations for SSN generation were used with alignment score selection based on a percent alignment of 40%. Individual nodes were created for each protein. Clusters assigned by the SSN were visualized using Cytoscape 3.7.2 (34, 35) and annotated back onto their corresponding gene in Geneious Prime 2019.2.3 (https://www.geneious.com). In this chapter, we study the corresponding gene for the assigned protein clusters and map all spacer hits back onto these genes.

Spacer Targeting Analyses

Local BLAST iterations were completed in Geneious Prime 2019.2.3 (https://www.geneious.com) (36). All spacer sequences were searched against the phage gene clusters using BLASTn with parameters adjusted to CRISPR-target recommended values of match (+1), mismatch (-1), word size (7), E-value (1), and filter (yes) (37). Due to parameters of BLAST in Geneious, gap cost to open (-5) and extend (-2) deviated from that of CRISPR-target. Results at or above a bit score of 40 were considered significant and were used for further analysis. We extracted spacer hits and used them for forward analyses in Excel and Prism.

Figure 4.1 Total spacer hits for the type II-A (CR1) system correlate to PAM frequency within phage gene clusters.

All Type II-A (CR1) spacer sequences were blasted against our phage dataset. We considered only top hits with a BIT score cutoff of 40. Phage cluster numbers are listed on the x-axis with clusters 1-93 on the top graph and clusters 94-186 on the bottom graph. The number of hits that map to each phage gene cluster is represented by bars (corresponding to the left y-axis) with forward oriented hits in red and reverse oriented hits in blue. We searched all phage proteins for the consensus Type II-A (CR1) 'NNAGAA' PAM sequence. The line graph is the total number of PAM sequences present in all phage cluster members in the forward (red) and reverse (blue) orientation and corresponds to the right y-axis.

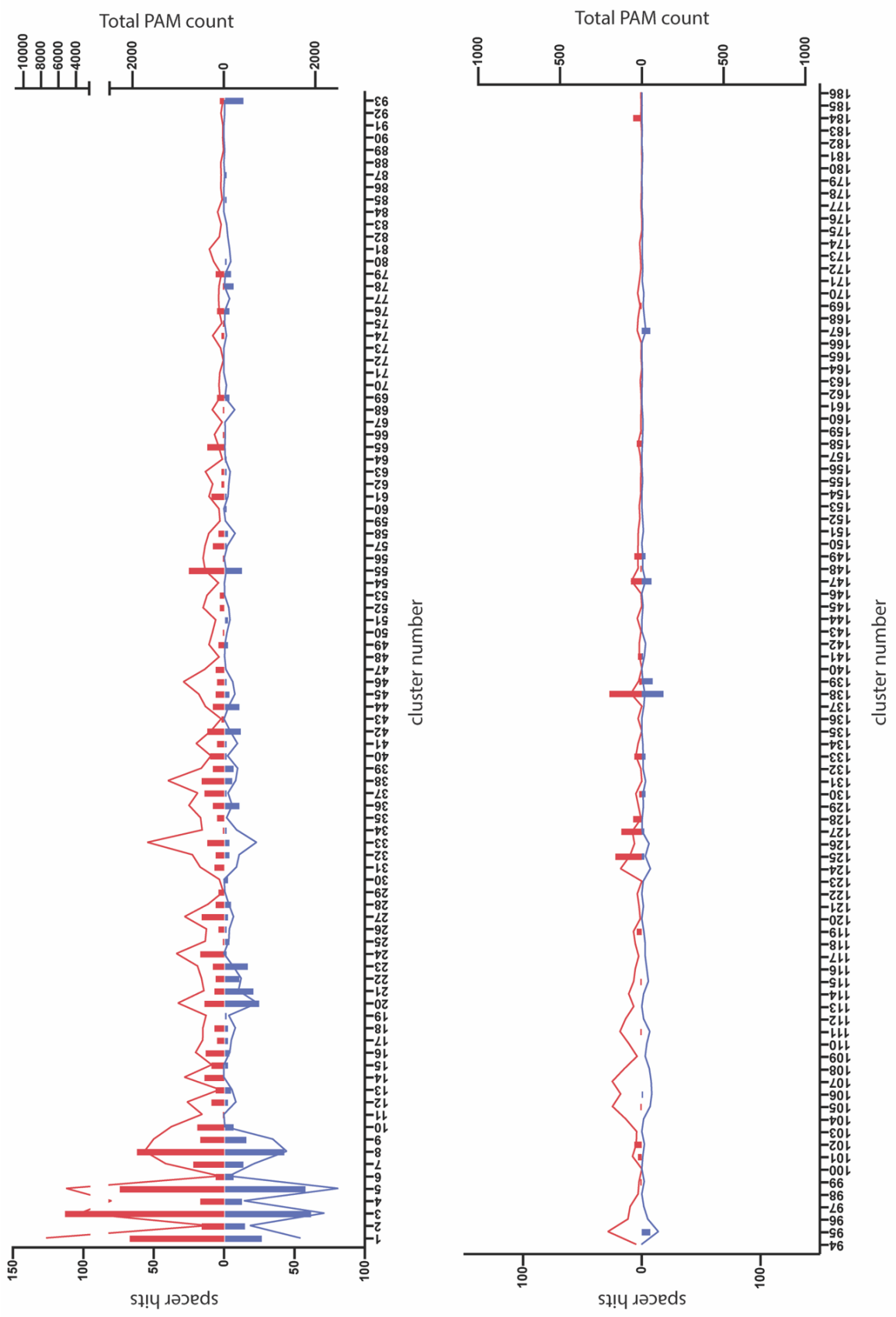

Total PAM count
10000
8000
6000
4000
2000
0
2000
cluster number
spacer hits
150
100
50
0
50
100
Total PAM count
1000
500
0
500
1000
cluster number
spacer hits
100
0
100

Figure 4.2 Total spacer hits for the type II-A (CR3) system correlate to PAM frequency of phage genes.

All Type II-A (CR3) spacer sequences were blasted against our phage dataset. We considered only top hits with a BIT score cutoff of 40. Phage cluster numbers are listed on the x-axis with clusters 1-93 on the top graph and clusters 94-186 on the bottom graph. The number of hits that map to each phage gene cluster is represented by bars (corresponding to the left y-axis) with forward oriented hits in red and reverse oriented hits in blue. We searched all phage proteins for the consensus Type II-A (CR3) 'NGGNG' PAM sequence. The line graph represents the total number of PAM sequences present in all phage cluster members in the forward (red) and reverse (blue) orientation and corresponds to the right y-axis.

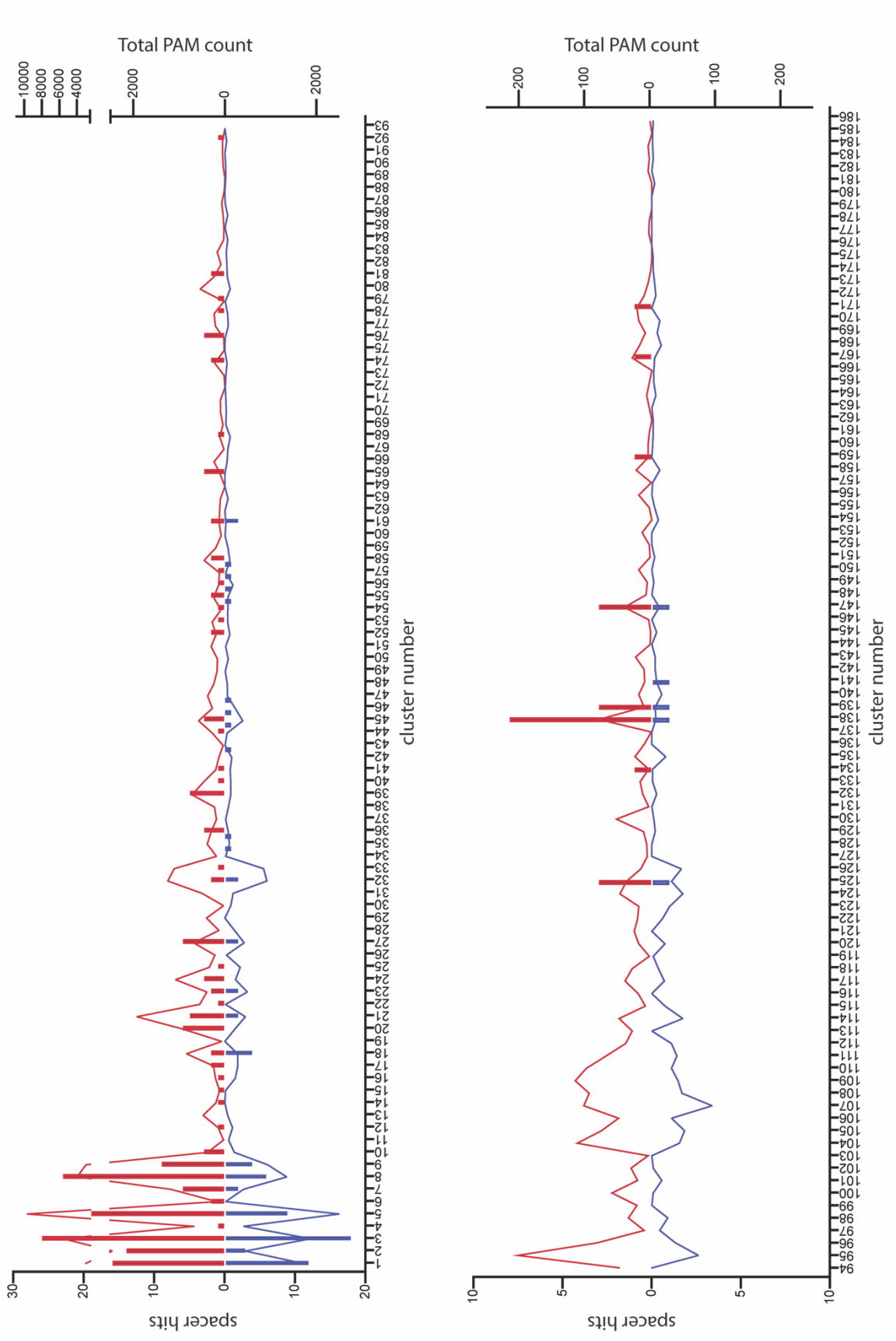

Total PAM count
10000
8000
6000
4000
2000
0
2000
cluster number
spacer hits
30
20
10
0
10
20
Total PAM count
200
100
0
100
200
cluster number
spacer hits
10
5
0
5
10

Figure 4.3 Unique spacer hits for the type I-E system correlate with PAM frequency

Unique Type I-E spacer sequences were blasted against our phage dataset. We considered only top hits with a BIT score cutoff of 40. Phage cluster numbers are listed on the x-axis with clusters 1-93 on the top graph and clusters 94-186 on the bottom graph. The number of hits that map to each phage gene cluster is represented by bars (corresponding to the left y-axis) with forward oriented hits in red and reverse oriented hits in blue. We searched all phage proteins for the consensus Type I-E 'AA' PAM sequence. The line graph is the total number of PAM sequences present in all phage cluster members in the forward (red) and reverse (blue) orientation and corresponds to the right y-axis.

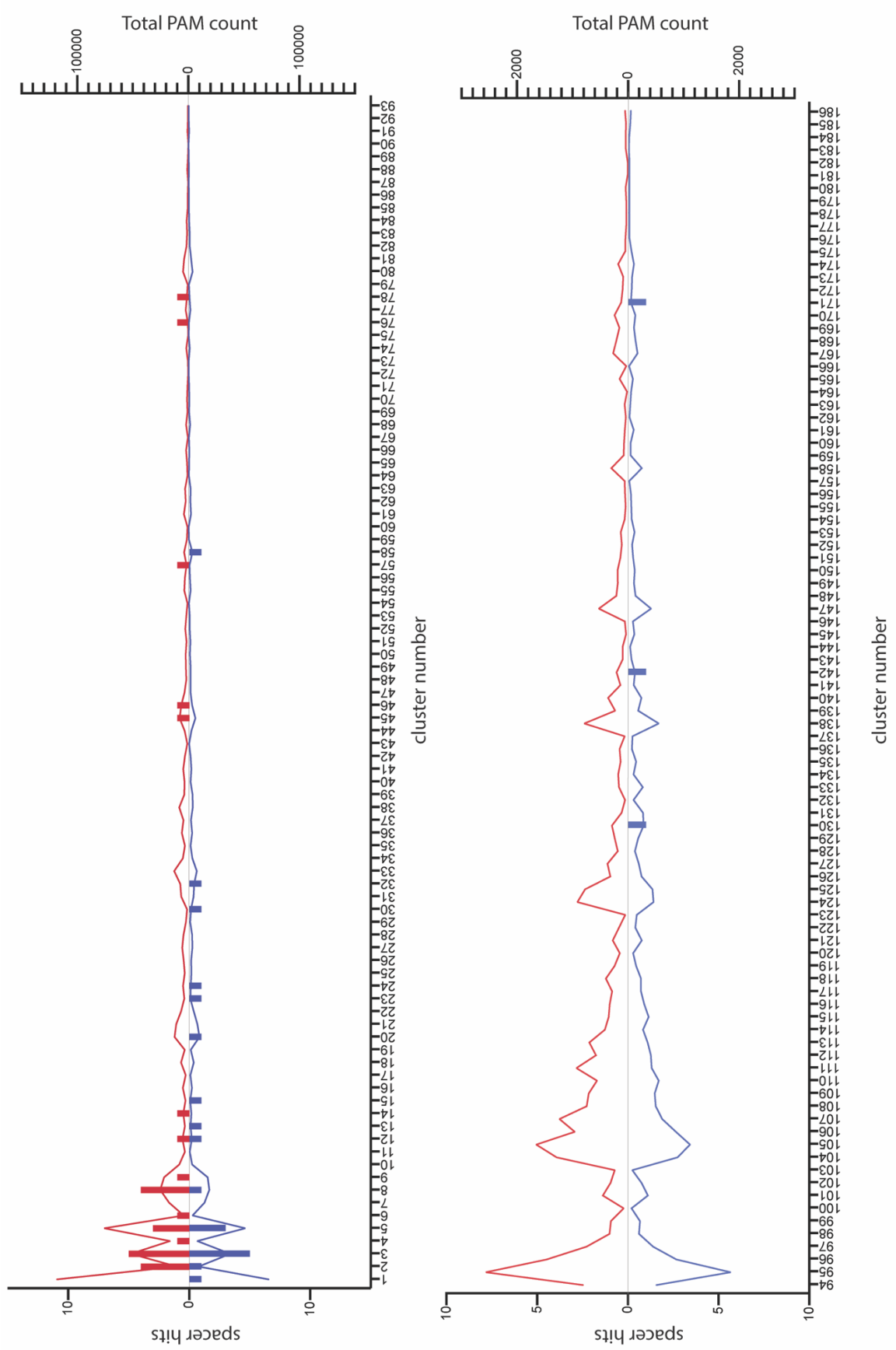

Total PAM count
100000
0
100000
cluster number
spacer hits
10
0
10
Total PAM count
2000
0
2000
cluster number
spacer hits
10
5
0
5
10

Figure 4.4 Spacer hits for the Type III-A system do not strongly correlate to relative size and abundance of phage gene clusters

Unique Type III-A spacer sequences were blasted against our phage dataset. We considered only top hits with a BIT score cutoff of 40. Phage cluster numbers are listed on the x-axis with clusters 1-93 on the top graph and clusters 94-186 on the bottom graph. The number of hits that map to each phage gene cluster is represented by bars (corresponding to the left y-axis) with forward oriented hits in red and reverse oriented hits in blue. The line graph is the total nucleotide length (corresponding to the right y-axis) of all phage gene cluster members.

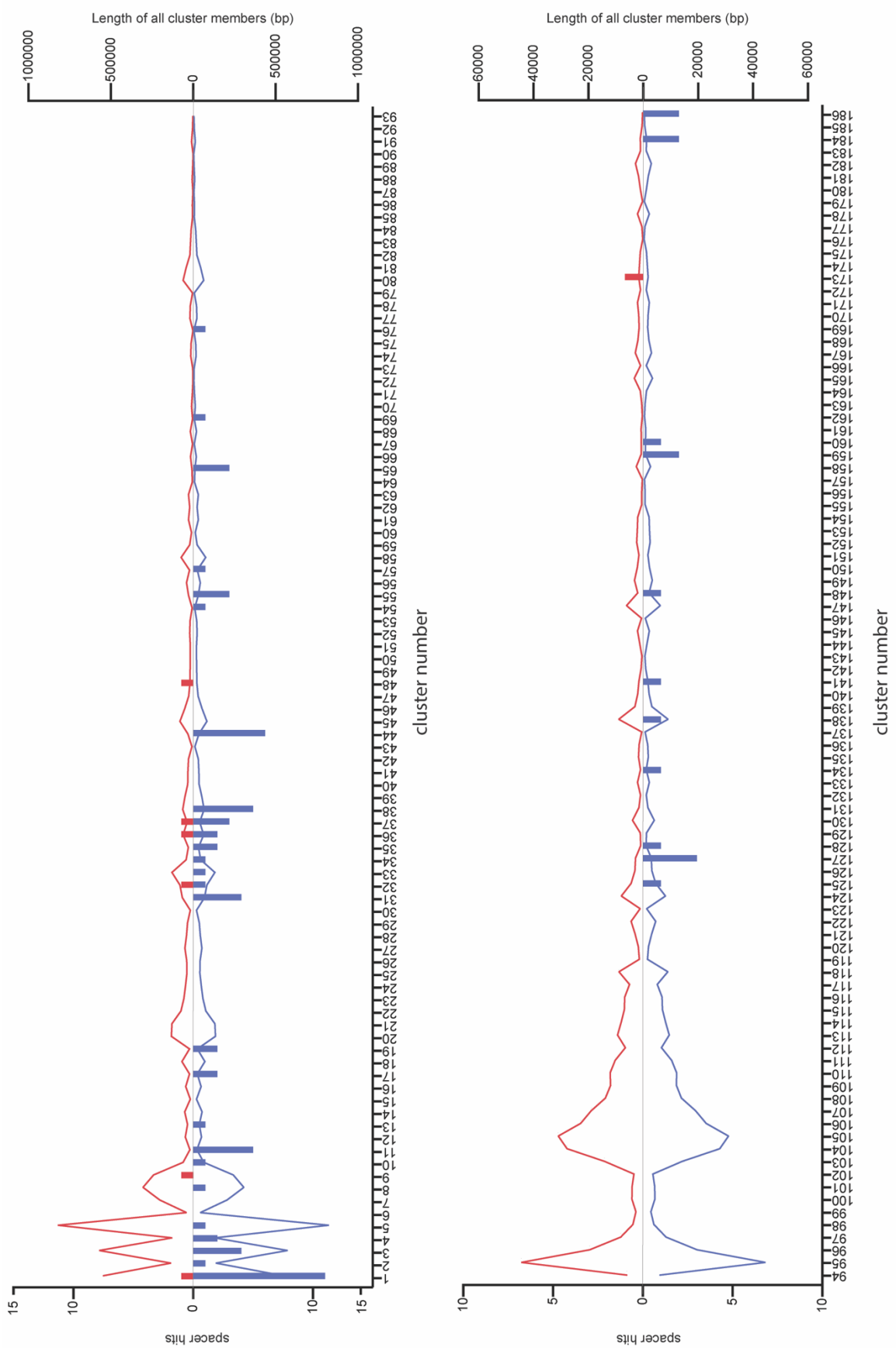

Length of all cluster members (bp)
1000000
500000
0
500000
1000000
cluster number
spacer hits
15
10
0
10
15
Length of all cluster members (bp)
60000
40000
20000
0
20000
40000
60000
cluster number
spacer hits
10
5
0
5
10

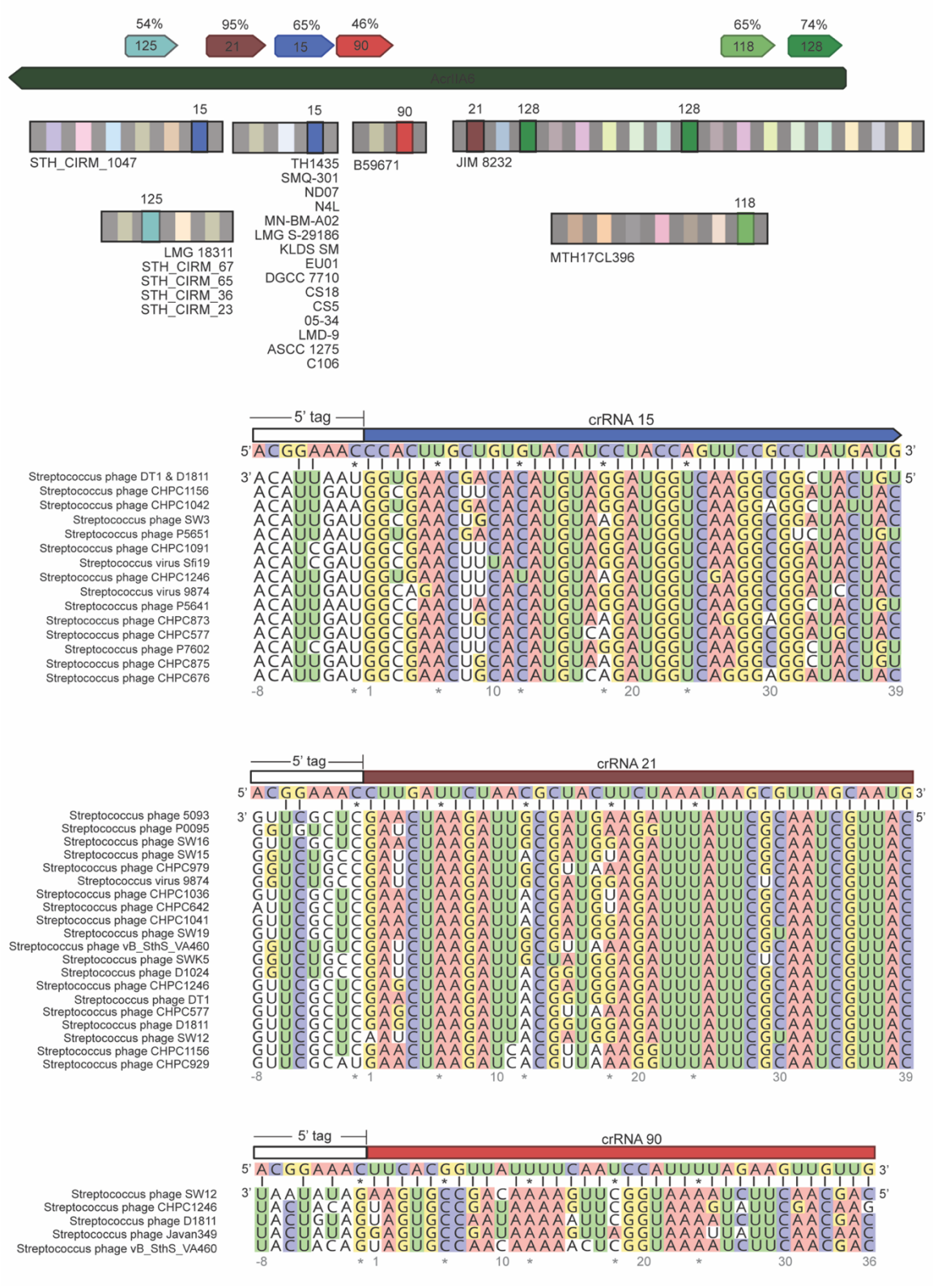
Table 4.1 Gene Clusters Hit by Type III-A Spacers

Table 4.1 ranks gene clusters hit by Type III-A spacer sequences. The order of clusters in the table is from most to least total spacer hits. Additionally, the number of genes in each sequence cluster as well as the top annotation are listed. The cluster numbers also correspond to those in Figure 3.1 and Table 3.1.

Phage cluster number	Count of genes in cluster	Annotation	Unique hits	Percentage of all unique hits	Total hits	Percentage of all total hits
1	948	Uncharacterized protein/DNA binding protein	12	13%	27	14%
44	65	AcrIIA6	6	7%	25	13%
31	104	DNA binding protein	4	4%	15	8%
54	49	Uncharacterized protein	1	1%	14	7%
37	83	Uncharacterized protein	4	4%	13	7%
11	159	Uncharacterized protein	5	6%	8	4%
184	2	Putative excisionase	2	2%	8	4%
3	218	Antireceptor	4	4%	7	4%
8	163	Baseplate component	1	1%	7	4%
38	82	Replication initiation protein A	5	6%	7	4%
55	48	Erf protein	3	3%	6	3%
65	31	Uncharacterized protein	3	3%	5	3%
127	8	AcrIIA5	3	3%	4	2%
13	142	Uncharacterized protein	1	1%	3	2%
36	83	Helicase loader	3	3%	3	2%
69	29	Uncharacterized protein	1	1%	3	2%
159	3	Uncharacterized protein	2	2%	3	2%
186	2	Uncharacterized protein	2	2%	3	2%
4	206	Uncharacterized protein	2	2%	2	1%
10	161	Uncharacterized protein	1	1%	2	1%
17	124	Cro-like protein	2	2%	2	1%
19	115	Cro-like repressor	2	2%	2	1%
32	102	Primase	2	2%	2	1%
35	99	VRR-NUC domain-containing protein	2	2%	2	1%
48	60	Uncharacterized protein	1	1%	2	1%
160	3	Uncharacterized protein	1	1%	2	1%
2	252	Lysin	1	1%	1	1%
5	189	Tape measure protein	1	1%	1	1%
9	162	Distal tail protein	1	1%	1	1%
33	101	Helicase	1	1%	1	1%
34	101	Uncharacterized protein	1	1%	1	1%
57	46	Uncharacterized protein	1	1%	1	1%
76	16	Uncharacterized protein	1	1%	1	1%
125	8	Integrase	1	1%	1	1%
128	8	Uncharacterized protein	1	1%	1	1%
134	6	Uncharacterized protein	1	1%	1	1%
138	5	IbrA	1	1%	1	1%
141	5	Uncharacterized protein	1	1%	1	1%
148	4	Uncharacterized protein	1	1%	1	1%
173	2	Uncharacterized protein	1	1%	1	1%

Figure 4.5 *S. thermophilus* Type III-A spacers target AcrIIA6

Spacers with a top BLAST hit (BIT score cutoff of 40) against AcrIIA6 are shown mapped along the length of the gene. The percentage listed above each spacer corresponds to the percentage of AcrIIA6 alleles that map to the spacer sequence above the BIT score cutoff. Below are Type III-A arrays that contain the corresponding spacer sequences, showing the relative position within the CRISPR-arrays. For arrays with the same spacer composition, all strains are listed below the array. Lastly, the RNA alignments for each of the spacer sequence hits are shown. For AcrIIA6 alleles that are identical in the region hit, a representative phage is listed. Because this is an RNA alignment, alignments were manually corrected to allow for G-U pairing. A 5' extension of the crRNA did not have a bearing on BIT score cutoffs but is shown to represent potential interaction between the 5' tag of the crRNA and the 3' PFS of the target.


54%
125
95%
21
65%
15
46%
90
65%
118
74%
128
AcrIIA8
15
STH_CIRM_1047
15
TH1435
SMQ-301
ND07
N4L
MN-BM-A02
LMG S-29186
KLDS SM
EU01
DGCC 7710
CS18
CS5
05-34
LMD-9
ASCC 1275
C106
90
B59671
21
128
128
JIM 8232
125
LMG 18311
STH_CIRM_67
STH_CIRM_65
STH_CIRM_36
STH_CIRM_23
118
MTH17CL396
5' tag
crRNA 15
5' ACGGAAACCCACUUGCUGUGUACAUCCUACCAGUUCCGCCUAUGAUG 3'
3' ACAUUAAUGGUGAACGACACAUGUAGGAUGGUCAAGGCGGCUACUGU 5'
Streptococcus phage DT1 & D1811
Streptococcus phage CHPC1156
Streptococcus phage CHPC1042
Streptococcus phage SW3
Streptococcus phage P5651
Streptococcus phage CHPC1091
Streptococcus virus Sfi19
Streptococcus phage CHPC1246
Streptococcus virus 9874
Streptococcus phage P5641
Streptococcus phage CHPC873
Streptococcus phage CHPC577
Streptococcus phage P7602
Streptococcus phage CHPC875
Streptococcus phage CHPC676
-8
1
10
20
30
39
5' tag
crRNA 21
5' ACGGAAACCUUGAUUCUAACGCUACUUCUAAAUAAGCGUUAGCAAUG 3'
3' GUUCGCUCGAACUAAGAUUGCGAUGAAGAUUUAUUCGCAAUCGUUAC 5'
Streptococcus phage 5093
Streptococcus phage P0095
Streptococcus phage SW16
Streptococcus phage SW15
Streptococcus phage CHPC979
Streptococcus virus 9874
Streptococcus phage CHPC1036
Streptococcus phage CHPC642
Streptococcus phage CHPC1041
Streptococcus phage SW19
Streptococcus phage vB_SthS_VA460
Streptococcus phage SWK5
Streptococcus phage D1024
Streptococcus phage CHPC1246
Streptococcus phage DT1
Streptococcus phage CHPC577
Streptococcus phage D1811
Streptococcus phage SW12
Streptococcus phage CHPC1156
Streptococcus phage CHPC929
-8
1
10
20
30
39
5' tag
crRNA 90
5' ACGGAAACUUCACGGUUAUUUUCAAUCCAUUUUAGAAGUUGUUG 3'
3' UAAUAUAGAAGUGCCGACAAAAGUUCGGUAAAAUCUUCAACGAC 5'
Streptococcus phage SW12
Streptococcus phage CHPC1246
Streptococcus phage D1811
Streptococcus phage Javan349
Streptococcus phage vB_SthS_VA460
-8
1
10
20
30
36

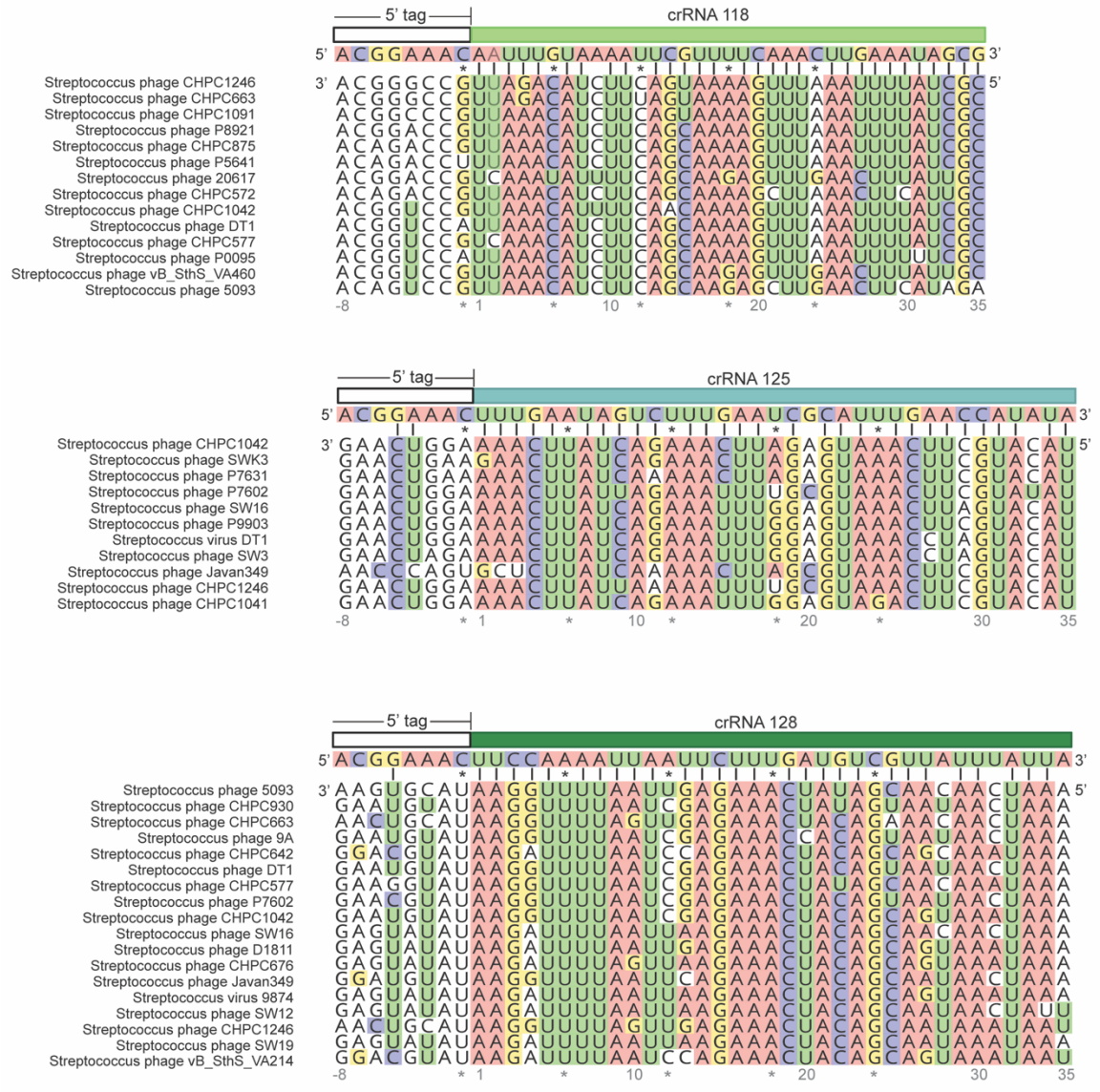

5' tag
crRNA 118
5' ACGGAAACAAUUUGUAAAAUUCGUUUUCAAACUUGAAAUAGCG 3'
Streptococcus phage CHPC1246
Streptococcus phage CHPC663
Streptococcus phage CHPC1091
Streptococcus phage P8921
Streptococcus phage CHPC875
Streptococcus phage P5641
Streptococcus phage 20617
Streptococcus phage CHPC572
Streptococcus phage CHPC1042
Streptococcus phage DT1
Streptococcus phage CHPC577
Streptococcus phage P0095
Streptococcus phage vB_SthS_VA460
Streptococcus phage 5093
-8 1 10 20 30 35
5' tag
crRNA 125
5' ACGGAAACUUUGAAUAGUCUUUGAAUCGCAUUUGAACCAUAUA 3'
Streptococcus phage CHPC1042
Streptococcus phage SWK3
Streptococcus phage P7631
Streptococcus phage P7602
Streptococcus phage SW16
Streptococcus phage P9903
Streptococcus virus DT1
Streptococcus phage SW3
Streptococcus phage Javan349
Streptococcus phage CHPC1246
Streptococcus phage CHPC1041
-8 1 10 20 30 35
5' tag
crRNA 128
5' ACGGAAACUUCCAAAAUUAAUUCUUUGAUGUCGUUAUUUAUUA 3'
Streptococcus phage 5093
Streptococcus phage CHPC930
Streptococcus phage CHPC663
Streptococcus phage 9A
Streptococcus phage CHPC642
Streptococcus phage DT1
Streptococcus phage CHPC577
Streptococcus phage P7602
Streptococcus phage CHPC1042
Streptococcus phage SW16
Streptococcus phage D1811
Streptococcus phage CHPC676
Streptococcus phage Javan349
Streptococcus virus 9874
Streptococcus phage SW12
Streptococcus phage CHPC1246
Streptococcus phage SW19
Streptococcus phage vB_SthS_VA214
-8 1 10 20 30 35

Figure 4.6 *S. thermophilus* Type III-A spacers target AcrIIA5

Spacers with a top BLAST hit (BIT score cutoff of 40) against AcrIIA5 are shown mapped along the length of the gene. The percentage listed above each spacer corresponds to the percentage of AcrIIA5 alleles that map to the spacer sequence above the BIT score cutoff. Below are Type III-A arrays that contain the corresponding spacer sequences, showing the relative position within the CRISPR-arrays. RNA alignments for each of the spacer sequence hits are shown. Because this is an RNA alignment, alignments were manually corrected to allow for G-U pairing. A 5' extension of the crRNA did not have a bearing on BIT score cutoffs but is shown to represent potential interaction between the 5' tag of the crRNA and the 3' PFS of the target.

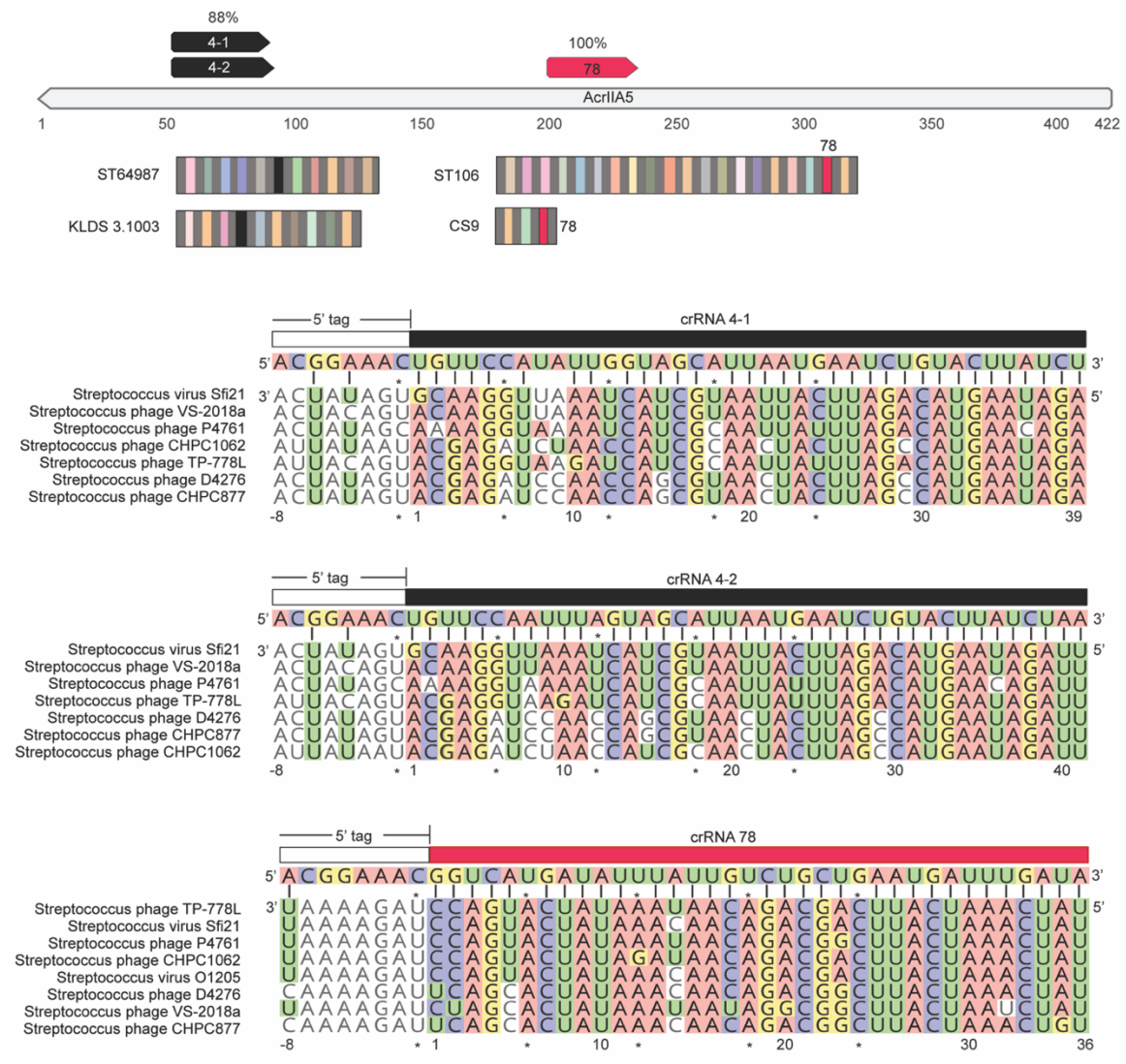
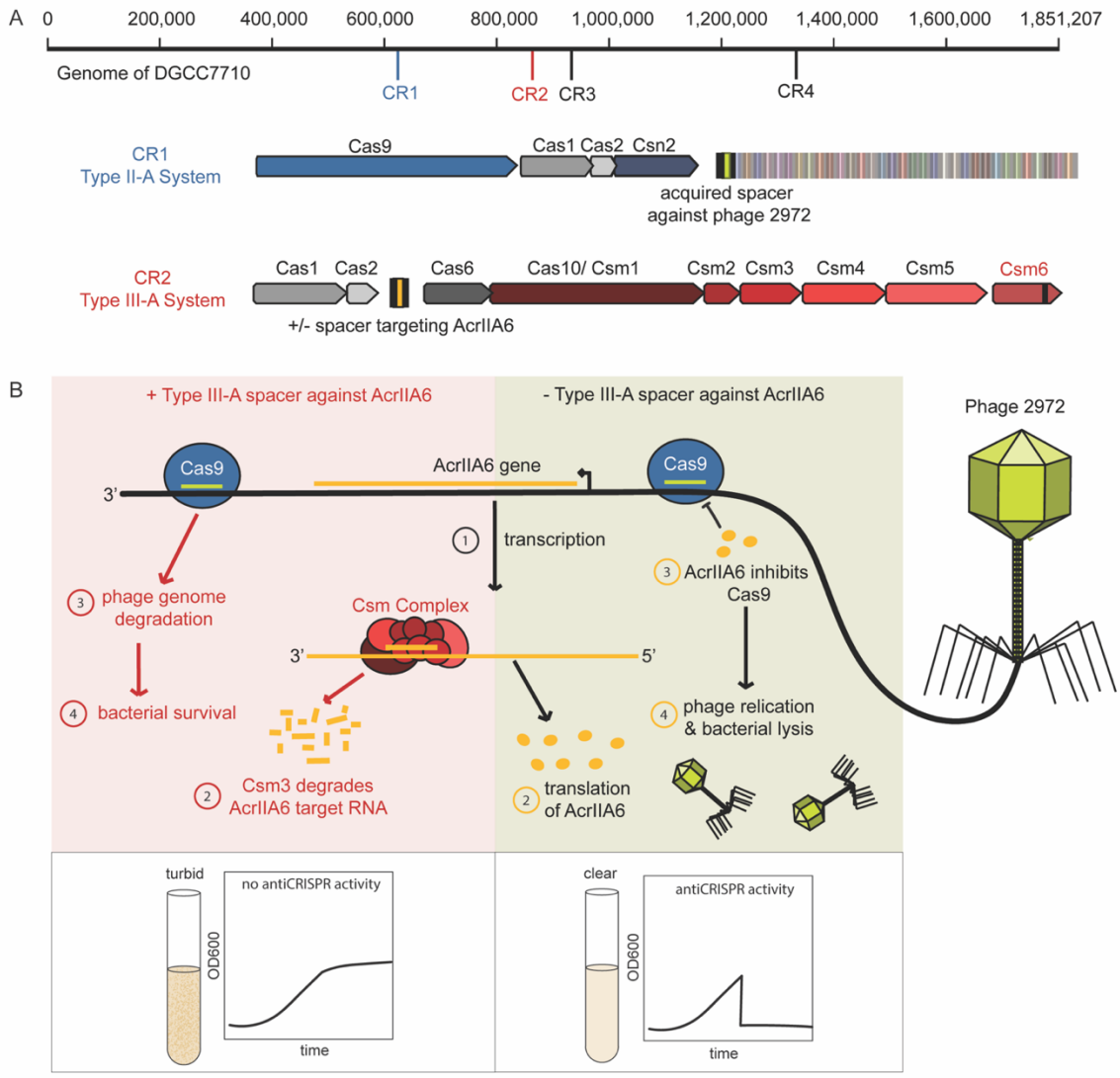

88%
4-1
4-2
100%
78
AcrIIA5
1
50
100
150
200
250
300
350
400
422
ST64987
ST106
78
KLDS 3.1003
CS9
78
5' tag
crRNA 4-1
Streptococcus virus Sfi21
Streptococcus phage VS-2018a
Streptococcus phage P4761
Streptococcus phage CHPC1062
Streptococcus phage TP-778L
Streptococcus phage D4276
Streptococcus phage CHPC877
5' tag
crRNA 4-2
Streptococcus virus Sfi21
Streptococcus phage VS-2018a
Streptococcus phage P4761
Streptococcus phage TP-778L
Streptococcus phage D4276
Streptococcus phage CHPC877
Streptococcus phage CHPC1062
5' tag
crRNA 78
Streptococcus phage TP-778L
Streptococcus virus Sfi21
Streptococcus phage P4761
Streptococcus phage CHPC1062
Streptococcus virus O1205
Streptococcus phage D4276
Streptococcus phage VS-2018a
Streptococcus phage CHPC877

Figure 4.7 Experimental approach to test AcrIIA6 inhibition by the Type III-A CRISPR-Cas system

(A) Genome overview of the strains used for this assay. We plan to use *S. thermophilus* DGCC 7710 which encodes four CRISPR-Cas systems. We have a bacteriophage insensitive mutant (BIM) with a spacer acquisition in the CR1 Type II-A array against phage 2972. The control strain has the type III-A array deleted and does not encode a native spacer against AcrIIA6. The experimental strain has a native Type III-A array targeting AcrIIA6. Of note, the Type III-A system of *S. thermophilus* DGCC 7710 does not have measurable defense against phage lysis when provided with a spacer against phage 2972. **(B)** The strains +/- an AcrIIA6 targeting spacer will be challenged with phage 2972 edited to express AcrIIA6. We depict the hypothesized mechanism of inhibition of AcrIIA6 by the Type III-A system when provided with a spacer against AcrIIA6 (pink background) versus expected AcrIIA6 inhibition of Cas9 when the Type III-A system does not encode a spacer against AcrIIA6 (green background). If AcrIIA6 is inhibited, we expect Cas9 to degrade the phage genome and the strain to grow to turbidity. If AcrIIA6 is not inhibited, we expect AcrIIA6 to prevent Cas9 action so that the phage can replicate and lyse the cell (clearing the culture).


A
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,851,207
Genome of DGCC7710
CR1
CR2
CR3
CR4
CR1
Type II-A System
Cas9
Cas1 Cas2 Csn2
acquired spacer
against phage 2972
CR2
Type III-A System
Cas1
Cas2
Cas6
Cas10/ Csm1
Csm2
Csm3
Csm4
Csm5
Csm6
+/- spacer targeting AcrIIA6
B
+ Type III-A spacer against AcrIIA6
- Type III-A spacer against AcrIIA6
Phage 2972
Cas9
AcrIIA6 gene
Cas9
3'
1 transcription
3 AcrIIA6 inhibits
Cas9
3 phage genome
degradation
Csm Complex
3'
5'
4 bacterial survival
4 phage relication
& bacterial lysis
2 Csm3 degrades
AcrIIA6 target RNA
2 translation
of AcrIIA6
turbid
no antiCRISPR activity
OD600
time
clear
antiCRISPR activity
OD600
time

REFERENCES

1. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, et al. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. Journal of bacteriology. 2008;190(4):1401-12.

2. Magadán AH, Dupuis M-È, Villion M, Moineau S. Cleavage of Phage DNA by the Streptococcus thermophilus CRISPR3-Cas System. PLoS One. 2012;7(7):e40913.

3. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun. 2018;9(1):2919.

4. You L, Ma J, Wang J, Artamonova D, Wang M, Liu L, et al. Structure Studies of the CRISPR-Csm Complex Reveal Mechanism of Co-transcriptional Interference. Cell. 2019;176(1-2):239-53.e16.

5. Jiang W, Samai P, Marraffini Luciano A. Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell. 2016;164(4):710-21.

6. Agudelo D, Carter S, Velimirovic M, Duringer A, Levesque S, Rivest J-F, et al. Versatile and robust genome editing with Streptococcus thermophilus CRISPR1-Cas9. bioRxiv. 2019:321208.

7. Leenay RT, Maksimchuk KR, Slotkowski RA, Agrawal RN, Gomaa AA, Briner AE, et al. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol Cell. 2016;62(1):137-47.

8. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences. 2012;109(39):E2579-E86.

9. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. Embo j. 2013;32(3):385-94.

10. Samai P, Pyenson N, Jiang W, Goldberg Gregory W, Hatoum-Aslan A, Marraffini Luciano A. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell. 2015;161(5):1164-74.

11. Artamonova D, Karneyeva K, Medvedeva S, Klimuk E, Kolesnik M, Yasinskaya A, et al. Spacer acquisition by Type III CRISPR-Cas system during bacteriophage infection of Thermus thermophilus. Nucleic Acids Res. 2020;48(17):9787-803.

12. Mo CY, Mathai J, Rostøl JT, Varble A, Banh DV, Marraffini LA. Type III-A CRISPR immunity promotes mutagenesis of staphylococci. Nature. 2021;592(7855):611-5.

13. Duplessis M, Michael Russell W, A Romero D, Moineau S. Global gene expression analysis of two Streptococcus thermophilus bacteriophages using DNA microarray2005. 192-208 p.

14. Ventura M, Brüssow H. Temporal transcription map of the virulent Streptococcus thermophilus bacteriophage Sfi19. Appl Environ Microbiol. 2004;70(8):5041-6.

15. Ventura M, Foley S, Bruttin A, Chennoufi SC, Canchaya C, Brussow H. Transcription mapping as a tool in phage genomics: the case of the temperate Streptococcus thermophilus phage Sfi21. Virology. 2002;296(1):62-76.

16. Fuchsbauer O, Swuec P, Zimberger C, Amigues B, Levesque S, Agudelo D, et al. Cas9 Allosteric Inhibition by the Anti-CRISPR Protein AcrIIA6. Mol Cell. 2019;76(6):922-37 e7.

17. Garcia B, Lee J, Edraki A, Hidalgo-Reyes Y, Erwood S, Mir A, et al. Anti-CRISPR AcrIIA5 Potently Inhibits All Cas9 Homologs Used for Genome Editing. Cell reports. 2019;29(7):1739-46 e5.

18. Hynes AP, Rousseau GM, Lemay ML, Horvath P, Romero DA, Fremaux C, et al. An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. Nature microbiology. 2017;2(10):1374-80.

19. Song G, Zhang F, Zhang X, Gao X, Zhu X, Fan D, et al. AcrIIA5 Inhibits a Broad Range of Cas9 Orthologs by Preventing DNA Target Cleavage. Cell reports. 2019;29(9):2579-89 e4.

20. Foster K, Kalter J, Woodside W, Terns RM, Terns MP. The ribonuclease activity of Csm6 is required for anti-plasmid immunity by Type III-A CRISPR-Cas systems. RNA Biology. 2018:null-null.

21. Lemay M-L, Renaud A, Rousseau GM, Moineau S. Targeted Genome Editing of Virulent Phages Using CRISPR-Cas9. Bio-protocol. 2018;8(1):e2674.

22. Martel B, Moineau S. CRISPR-Cas: an efficient tool for genome engineering of virulent bacteriophages. Nucleic Acids Research. 2014;42(14):9504-13.

23. Paez-Espino D, Morovic W, Sun CL, Thomas BC, Ueda K-i, Stahl B, et al. Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nature Communications. 2013;4(1):1430.

24. Lemay ML, Tremblay DM, Moineau S. Genome Engineering of Virulent Lactococcal Phages Using CRISPR-Cas9. ACS synthetic biology. 2017;6(7):1351-8.

25. Pyenson NC, Gayvert K, Varble A, Elemento O, Marraffini LA. Broad Targeting Specificity during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. Cell host & microbe. 2017;22(3):343-53.e3.

26. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of bacteriology. 2008;190(4):1390-400.

27. Landsberger M, Gandon S, Meaden S, Rollie C, Chevallereau A, Chabas H, et al. Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity. Cell. 2018;174(4):908-16.e12.

28. Borges AL, Zhang JY, Rollins MF, Osuna BA, Wiedenheft B, Bondy-Denomy J. Bacteriophage Cooperation Suppresses CRISPR-Cas3 and Cas9 Immunity. Cell. 2018;174(4):917-25.e10.

29. Pyenson NC, Marraffini LA. Type III CRISPR-Cas systems: when DNA cleavage just isn't enough. Curr Opin Microbiol. 2017;37:150-4.

30. Zallot R, Oberg NO, Gerlt JA. ‘Democratized’ genomic enzymology web tools for functional assignment. Current Opinion in Chemical Biology. 2018;47:77-85.

31. Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry. 2019;58(41):4169-82.

32. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for

generating protein sequence similarity networks. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2015;1854(8):1019-37.

33. Gerlt JA. Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence–Function Space and Genome Context to Discover Novel Functions. Biochemistry. 2017;56(33):4293-308.

34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.

35. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. Genome Biol. 2019;20(1):185.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

37. Biswas A, Gagnon JN, Brouns SJ, Fineran PC, Brown CM. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. RNA Biol. 2013;10(5):817-27.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

Composition and Immunity of Type III-A CRISPR-Cas systems in *Streptococcus thermophilus*

The overall aim of Chapter 2 was to characterize the state of Type III-A (Csm) CRISPR-Cas systems in *S. thermophilus*. We first compared repeat and spacer features to those of co-occurring CRISPR-Cas systems. We then aimed to determine minimal nuclease requirements of the Csm system in both plasmid and phage-based defense contexts so we could use this knowledge to define the functional capacity of each identified Type III-A system in *S. thermophilus.*

We determined that Csm6-1 and Csm6-2 in *S. thermophilus* JIM 8232 play redundant roles in Type III-A defense against plasmids. In addition, we verified previous results from an assay of the Csm system expressed in *E. coli*, that Csm6 activity is required for robust plasmid-based defense (1). We further determined that the *S. thermophilus* JIM 8232 Type III-A system is capable of defense against phage when provided with a spacer sequence against an early expressed phage gene. From mutation analysis, we found that mutation of the HEPN active sites of Csm6-1 and 6-2 did not abrogate defense against phage at low MOIs, indicating that the Csm complex alone may be capable of low-level phage defense activity. However, we did not see defense activity of the *S. thermophilus* DGCC 7710 Csm system when provided with a spacer against phage 2972, despite

containing Csm complex proteins (Csm1-Csm5) with predicted functionality. Perhaps this difference is due to the Csm6 proteins of *S. thermophilus* JIM 8232 strain retaining some functionality despite HEPN active site mutation, so a Csm6 knock-out strain should be tested in this phage-based context.

Within the scope of the analysis of Csm systems with Csm6 activity, we need to include early, middle, and late expressed spacer targets in our phage-based defense assays to determine Csm gene necessity for targeting during different phases of phage infection. Overall, while we were able to make forward progress on this aim, we also opened doors to further questions that need to be answered to fully define the state of Csm systems in *S. thermophilus*. Our analysis determined that 80% of Type III-A systems in *S. thermophilus strains*, including the DGCC 7710 strain are predicted to have functional Csm complexes. However, only 10% of *S. thermophilus* Csm systems (including JIM 8232 strain) have both Csm6-1 and Csm6-2 proteins. Despite this, our analysis largely focused on the functionality of the *S. thermophilus* JIM 8232 Type III-A system.

In future, we need to assay other potential functions of the Csm systems in strains with protein composition (and spacer composition) similar to *S. thermophilus* DGCC 7710. The use of essential gene targeting (i.e. targeting antibiotic resistance or plasmid replication genes) may better assay the utility of Csm3 target specific RNase activity in defense. In addition, simple assays, including RT-PCR can be employed to determine if Csm3 mediated target transcript degradation is occurring.

Additional musings about Type III-A CRISPR-Cas systems in *S. thermophilus*

Previously our lab demonstrated the utility of Type III-A systems to degrade specific target transcripts when expressed in *E. coli* (2). This work included multiplexed targeting of transcripts as well as targeting of a non-coding RNA with complex folds (rnpB RNA) (2). The application of Type III-A systems in editing the transcriptome of prokaryotes opens the door for gene function and network analysis of complex cellular pathways. By characterizing the protein makeup and functionality of Type III-A systems in *S. thermophilus*, we are also highlighting natively expressed tools for transcriptome editing and pathway analysis of this important host strain.

A previous publication identified genes linked to CRISPR-Cas systems and found that Type III systems have the widest array of gene linkages (3). These linked genes range from having unknown functions to transmembrane domains or CARF domains, indicating potential activity in cellular signaling pathways (3, 4). This analysis did not indicate any genes linked to Type III-A systems in strains of *Streptococcus thermophilus*, but the cutoffs used for Type III-A system prediction may have been too restrictive to identify Type III-A systems that are not well annotated. Their supplemental data indicates that they looked at 24 *Streptococcus thermophilus* strains with Type III-A systems, while we identified more than double this number in our genome neighborhood analysis (3). It may be worth repeating their search to determine if there are linked genes in *S. thermophilus* based on our genome neighborhood analysis. During neighborhood extraction for Chapter 2, we determined that the Type III-A system is located within the operon for pyrimidine metabolism between PyrD and PyrF genes indicating potential co-regulation of these two systems.

Lastly, in 8 of the analyzed Type III-A systems in *S. thermophilus* we saw insertion of genes with annotated bacteriocin function in place of Cas6 through Csm5. Bacteriocins are bacterial genes that are often detrimental to other strains in co-culture, acting as natively expressed antimicrobials (5). In the context of S. thermophilus, these bacteriocin genes can increase safety of fermented products by inhibiting growth of pathogenic bacteria, but can also be detrimental to quality if they inhibit growth of other starter culture strains (5). It is interesting to see the Type III-A defense system replaced by bacteriocin insertion, and perhaps there was a selection event that led to defense loss in this case. I was unable to directly link the entire operon to a currently characterized bacteriocin module, so further characterization of the genes could be carried out in future.

Discovery of novel anti-CRISPRs against the CRISPR-Cas systems of *Streptococcus thermophilus*

In Chapter 3 we identified 3 novel inhibitors of CRISPR-Cas systems in *S. thermophilus*. Two of the identified anti-CRISPR proteins, AcrIIA25 and AcrIIA26, inhibited the Type II-A (CR3) system while AcrIE10 inhibited the Type I-E system. We are also continuing to screen candidate proteins from the clustering analysis in phage and plasmid based anti-CRISPR assays. Future studies can focus on biochemical characterization of these anti-CRISPRs as well as the potential Type II-A anti-CRISPR inhibition of Cas9 proteins used for genome editing.

Interestingly, we found that AcrIIA26 (candidate cluster 119) is unable to inhibit CRISPR3 Cas9 in a plasmid-based defense context but has anti-CRISPR activity in a phage-based assay. We notice a similar pattern with AcrIIA5, seeing inhibition of only the

Type II-A (CR1) CRISPR-Cas system in the plasmid-based assay, but inhibition of both CR1 and CR3 (Type II-A systems) in the phage-based assay. My overall assessment of this differential activity is that it is due to specificity of these anti-CRISPR proteins for inhibiting certain Cas9 alleles. The phage-based assay and immediacy of phage lysis may put enough pressure on the CRISPR-defense systems, allowing us to see more minor anti-CRISPR activity. This additionally highlights the importance of analyzing multiple Acr candidate alleles in our screen in case there is specificity of Acr inhibition that leads to a false negative.

Additional musings about anti-CRISPRs and phage applications

There are many exciting applications of our findings in Chapter 3 in the context of biotechnology, but also within the native organism. Bacteriophages have such limited space for gene expression, with S. thermophilus phages having roughly 40 kb to fit many essential genes. One can imagine that genes would not be conserved unless they serve some novel or essential functionality. Beyond identifying anti-CRISPRs, we defined a function for genes otherwise considered 'viral dark matter.' The more we can understand the genomic architecture and gene functionality of phages, the closer we are to designing phages for therapeutic uses. In a landscape of increasing antibiotic resistance, the concept of a targeted therapy against a single species or strain is much more ideal. While still far from fruition, increasing our understanding of the dynamics between anti-CRISPRs and CRISPR-Cas systems will increase our toolkit for phage editing to combat host defense systems.

Features and selectivity of *Streptococcus thermophilus* CRISPR-Cas system targeting of phage genes

Chapter 4 defines our current work and future directions focused on the potential specificity of Type III-A targeting of phage genes. While still incomplete, this work is the culmination of data from the previous chapters, including analyses of the Type III-A systems in *S. thermophilus* as well as phage gene clustering for identification of anti-CRISPRs.

We wanted to begin our analysis by again comparing the features of the Csm system to those of the cooccurring systems in *S. thermophilus*. The implications of the targeting data for the Type II-A and I-E systems on their own should not be overlooked. While there are correlations between targeting and PAM sequence abundance, there are instances of enrichment against genes that don't entirely follow this trend (cluster 138 for Type II-A hits, corresponds to IbrA (6)). Perhaps these genes are more enriched in the mobile-genetic-element population than in our phage dataset. One interesting analysis could be to compare recent spacer acquisition events (i.e. the first three spacers) compared to 'older' acquisition events (i.e the last spacers in the array) to determine if the subset of genes with acquisition events changes over time.

Several weaknesses of our analysis of Type III-A spacer hits include **(1)** that there is no defined PFS sequence for the *S. thermophilus* Csm system, so we used gene length and abundance to determine if spacer targeting should be enriched against certain genes. A better understanding of the sequence requirements for Type III-A targeting (if there are any) could improve our ability to understand if certain sequences should be enriched for reasons other than our hypothesized activity that targeting of certain genes is more advantageous. **(2)** Another weakness is that we do not have a defined set of early, middle,

and late expressed genes. Other groups have demonstrated a selective advantage to Csm system targeting of early expressed genes (7, 8). We tried to abrogate this by studying spacer targeting features in the subset of targeted genes, but this is not as ideal as studying trends within each temporal expression group. **(3)** Finally, we can't be certain that all of the genes in our dataset are present in phages that infect hosts with Type III-A systems. We could be looking at enriched hits based on phage genes present in phages that infect hosts with more highly adaptive Type III-A systems. While I'm not certain that this is the case, as the other systems don't appear to follow this trend, it is a potential data bias that we should consider.

Another potential (and perhaps not conflicting) hypothesis for the Type III-A spacer enrichment against certain genes could be that when the Type II-A CRISPR-Cas system is inhibited (i.e. when anti-CRISPR proteins such as AcrIIA6 are present), the Type III-A system is relied on for phage defense. While AcrIIA6 is not enriched in the general population of phages, it would be enriched in these instances, so we would expect to see preferential adaptation events against it and other Type II-A (and potentially I-E) inhibitors. This hypothesis has exciting implications for anti-CRISPR discovery (as hits could correspond to likely anti-CRISPRs) and may not entirely contradict our current hypothesized activity, as Type III-A systems could still play some role in anti-CRISPR silencing through Csm3 functionality.

REFERENCES

1. Foster K, Kalter J, Woodside W, Terns RM, Terns MP. The ribonuclease activity of Csm6 is required for anti-plasmid immunity by Type III-A CRISPR-Cas systems. RNA Biology. 2018:null-null.

2. Woodside WT, Vantsev N, Terns MP. Type III-A CRISPR systems as a versatile gene knockdown technology. bioRxiv. 2020:2020.09.25.310060.

3. Shmakov SA, Makarova KS, Wolf YI, Severinov KV, Koonin EV. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proceedings of the National Academy of Sciences. 2018;115(23):E5307-E16.

4. Makarova KS, Timinskas A, Wolf YI, Gussow AB, Siksnys V, Venclovas Č, et al. Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. Nucleic Acids Research. 2020;48(16):8828-47.

5. Rossi F, Marzotto M, Cremonese S, Rizzotti L, Torriani S. Diversity of Streptococcus thermophilus in bacteriocin production; inhibitory spectrum and occurrence of thermophilin genes. Food microbiology. 2013;35(1):27-33.

6. Sandt CH, Hopper JE, Hill CW. Activation of prophage eib genes for immunoglobulin-binding proteins by genes from the IbrAB genetic island of Escherichia coli ECOR-9. Journal of bacteriology. 2002;184(13):3640-8.

7. Artamonova D, Karneyeva K, Medvedeva S, Klimuk E, Kolesnik M, Yasinskaya A, et al. Spacer acquisition by Type III CRISPR–Cas system during bacteriophage infection of Thermus thermophilus. Nucleic Acids Research. 2020;48(17):9787-803.

8. Jiang W, Samai P, Marraffini LA. Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell. 2016;164(4):710-21.