Discovery and Characterization of Cas13b, a Differentially Regulated RNA-targeting CRISPR System

by

Aaron Andrew Smargon

A.B., Princeton University (2011)
S.M., Massachusetts Institute of Technology (2016)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

Author .....
Department of Electrical Engineering and Computer Science
January 10, 2018

Certified by .....
Feng Zhang
Associate Professor, Brain and Cognitive Sciences and Biological Engineering
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor, Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Discovery and Characterization of Cas13b, a Differentially Regulated RNA-targeting CRISPR System

by

Aaron Andrew Smargon

Submitted to the Department of Electrical Engineering and Computer Science
on January 10, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

RNA plays a significant role in human biology and disease, not only as messenger RNA encoding proteins but also as noncoding RNA regulating DNA, proteins, and other RNA species. Until recently, it has been challenging to target RNA in a simple, efficient manner. CRISPR-Cas (clustered regularly interspaced short palindromic repeats and CRISPR-associated proteins) systems, which confer adaptive immunity to prokaryotes, have revolutionized DNA targeting through the engineering of RNA-programmable Cas9-based tools. Effective RNA-programmable RNA-targeting tools would likewise transform RNA biology and biotechnology.

Class 2 CRISPR-Cas systems, which rely only on a single effector protein and programmable CRISPR RNA (crRNA) to target nucleic acids, represent the most promising tool to target RNA. Building on previous research, a biocomputational pipeline was developed to discover novel functional class 2 CRISPR systems lacking the canonical adaptive machinery of Cas1 and Cas2 at their genomic loci. Out of this pipeline emerged the class 2 CRISPR-Cas RNA-targeting system, VI-B (Cas13b with accessory Csx27/Csx28). Cas13b was characterized both biochemically and genetically, and found to be differentially regulated—inhibited by Csx27 in VI-B1 systems and enhanced by Csx28 in VI-B2 systems. RNA-targeting rules are critical to tool development, and so an *E. coli* essential gene screen was conducted and analyzed to assess the RNA sequence and structure requirements for targeting. The completion of this work advances both knowledge in the CRISPR field and possibilities in the RNA-targeting toolkit.

Thesis Supervisor: Feng Zhang
Title: Associate Professor, Brain and Cognitive Sciences and Biological Engineering

Acknowledgments

A (not so) long, strange trip−filled with excitement, uncertainty, discovery, and reflection. I wish to acknowledge all those who joined me for the journey.

First, my EECS Ph.D. Thesis Committee, the three professors who helped me reach the finish line, each guiding the evolution of the thesis text and defense: thesis committee chair Ron Weiss, for his leading support, guidance, and valuable advice in presenting; Bruce Tidor, for his careful attention, thorough comments, and concern for a clear completed text; and Collin Stultz, for his wisdom and unique perspective on the research.

Next, those MIT faculty who encouraged me along the way scientifically and personally: my academic adviser Leslie Kolodziejski, for being a source of counsel and strength in times of need; Doug Lauffenburger, for his generous time, direction, and insight; Anantha Chandrakasan, current Dean of the School of Engineering and former Chair of EECS, for his encouragement and advice throughout the years; RQE committee chair Tim Lu, for his pivotal support in my academic trajectory; RQE committee member David Gifford, for his constructive suggestions and care for my scientific career; and Thomas Heldt, a fellow physicist whom I met my first year at MIT when he became a professor, for the challenging and rewarding opportunity to TA *Cellular Neurophysiology and Computing*, my first course at MIT.

My research peers, collaborators, and mentors: "*mi amigo*" Sebastian Palacios, for sharing the maxima and minima of MIT from the first day of TQE classes, as well as research in biology and bioengineering while in EECS (and common Colombian roots); Katia Shtyrkova, for convincing me to come to MIT, even if I would later change fields; Neville Sanjana and Ophir Shalem, for early training in the Zhang Lab; David Scott, for his wet lab mentorship; every co-author on the Cas13b paper; David Cox and Neena Pyzocha, for helping turn Group 29 and 30 into a great scientific contribution; Kaijie Zheng, for being there at the beginning of the project; the rest of the Zhang Lab; Sergey Shmakov, for his collegiality and for an unforgettable tour of Moscow; Eugene Koonin, for being an inspirational computational biologist; and

my thesis supervisor Feng Zhang, for giving me the chance to enter the exciting field of CRISPR, for teaching me many facets of science and academia, and for pushing me to become an independent researcher.

My friends, at MIT and elsewhere, who have watched my progress and provided support and feedback throughout the years.

My family: my mother and academic role model Audrey, whose judgment in career and life has never been wrong; my father Dan, for his unconditional love and unlimited lessons; my brother Michael, for motivating me to tackle only problems worth solving; my grandfather and academic role model Andrew, for his intent career recommendations since I was very young; my grandmother Edith, for her care in my personal development; and my grandfather Ken and grandmother Erna, who did not get to see me complete graduate school but who would be so proud of my accomplishments.

I thank God for the miracles in my life, and for showing me that one can be a good scientist and a good person.

To all who read this: *Be grateful and be kind, for no one gets there alone.*

Contents

Chapter 1

Introduction

1.1 RNA Biology and Biotechnology

Ribonucleic acid (RNA) plays a dominant and dynamic role in human biology and disease. According to the 'central dogma' of biology, deoxyribonucleic acid (DNA) encodes for messenger RNA (mRNA) during transcription, which in turn encodes for proteins during translation. DNA also encodes for noncoding RNA (ncRNA), which can interact with DNA, proteins, and other RNA species (Figure 1-1). In the context of ncRNA, this interaction may lead to chromatin modification, RNA polymerase activity regulation, transcriptional interference, RNA spicing, RNA editing, mRNA stability, and translation initiation (Wahlestedt, 2013). Additionally, ncRNA can take on many forms, from ubiquitous 'ribosomal RNA' (rRNA), to translation adaptor 'transfer RNA' (tRNA), to post-transcriptional regulatory 'microRNA' (miRNA), to 'long non-coding RNA' (lncRNA) and other species.

Unlike DNA, of which there are two copies localized in the nucleus of diploid cells (with the exception of mitochondrial DNA), an identical RNA molecule may be present hundreds of times in a cell. Moreover, due to the relatively small size and instability of RNA, intracellular diffusion and degradation can cause its concentration and spatial distribution to vary significantly over time. RNA is also actively transported throughout cells by RNA binding proteins (RBPs), which may even load RNA into extracellular vesicles, either for waste management or potentially intercel-

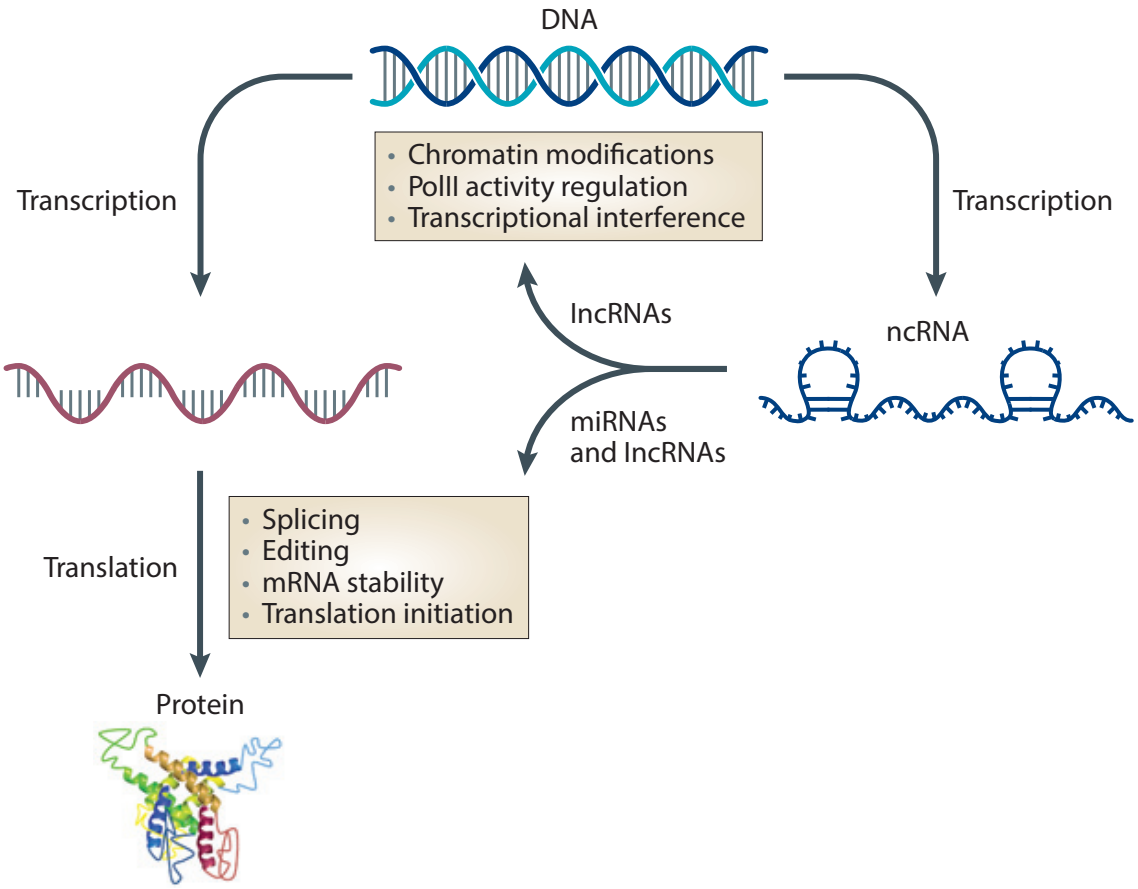


Figure 1-1: 'Central dogma' in the context of regulatory non-coding RNAs. Advances in transcriptomics have resulted in the discovery of large numbers of non-coding RNAs (ncRNAs), many of which have the capacity to regulate gene expression at transcriptional or translational levels. The concept of the 'central dogma', which is complemented in this figure with aspects of ncRNA functions, was arguably first formulated by Francis Crick in 1958: "Once information has passed into protein, it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible." lncRNA, long non-coding RNA; miRNA, microRNA. Figure and figure legend are from Wahlestedt (2013).

lular communication.

Notwithstanding technological considerations, the sheer diversity and versatility of the 'transcriptome' (sum total of RNA molecules expressed from an organism's genes) has challenged many attempts at consistent RNA manipulation. A number of applications for RNA-targeting with engineered RBPs have been proposed (Figure 1-2), including translation modulation, splicing modulation, localization detection and modulation, stabilization, degradation, and binding disruption (Mackay et al., 2011). While most readily adaptable to basic research, these applications may also be translated to human disease diagnostics (Gootenberg et al., 2017) and therapeutics (Cox et al., 2017).

RNA targeting invites multiple opportunities in synthetic biology, an emerging biological field motivated by the principles of electrical engineering and computer science. Over the last decade, efforts in programming cells, principally in bacteria and yeast, have resulted in the invention and exploitation of novel, synthetic RNA components (Isaacs et al., 2006). These components are based in design on natural ncRNA regulators, be they 'antisense' that repress translation upon binding, 'riboregulator' that repress or activate translation upon binding, 'ribozyme' that repress or activate translation upon cleavage, 'riboswitch' that repress or activate translation upon ligand binding, or 'structural scaffolds' that connect other RNA components.

By combining customized synthetic RNA components in engineered biological devices, researchers have realized applications in genetic circuitry, metabolic engineering, and biosensing (Chappell et al., 2015). Inspired by the diverse regulatory nature of ncRNA, completely novel RNA aptamers, or short sequences that bind specifically to target molecules through their 3-dimensional structures, have even been designed. With the advent of next-generation sequencing, researchers can now computationally create tens of thousands of rationally designed modular RNA aptamers and screen them against a biological or chemical input to select for optimal affinity and specificity (McKeague et al., 2016).

In the last few years, RNA synthetic biology has been translated to mammalian biology. Following viral transduction of RNA components in human cells, researchers

have implemented sophisticated logical and computational functions, such as multi-input classification or regulation of transgene expression (Wroblewska et al., 2015). These genetically precise approaches are at the frontier of treating complex diseases, and offer multiple advantages over similar DNA-targeting techniques. First, they do not require nuclear localization, often challenging in systemic delivery. Second, they circumvent direct engineering of the genome, which may have off-target effects. Finally, any RNA synthetic biology will be transient, and thus there is minimal risk for persistent cellular perturbation. More effective RNA-targeting tools would expand the advantages of RNA synthetic biology.

Until recently, only a few RNA-targeting applications had been realized. The earliest such work can be traced back to the end of the previous century, when two independent studies with the coat protein of the RNA bacteriophage MS2 led to successful localization detection and stabilization of mRNA in eukaryotic cells (Bertrand et al., 1998; Coller et al., 1998). Next, the discovery of RNA interference (RNAi) in *C. elegans* would be exploited in mammalian cells (Elbashir et al., 2001). Both approaches would set the gold standard for in vivo RNA targeting for over a decade, yet both would be severely limited–MS2 coat protein because it relies on either targeting of exogenous RNA or inefficient autonomous hybridization with endogenous RNA, and RNAi because it is an endogenous process whose associated proteins cannot be modified.

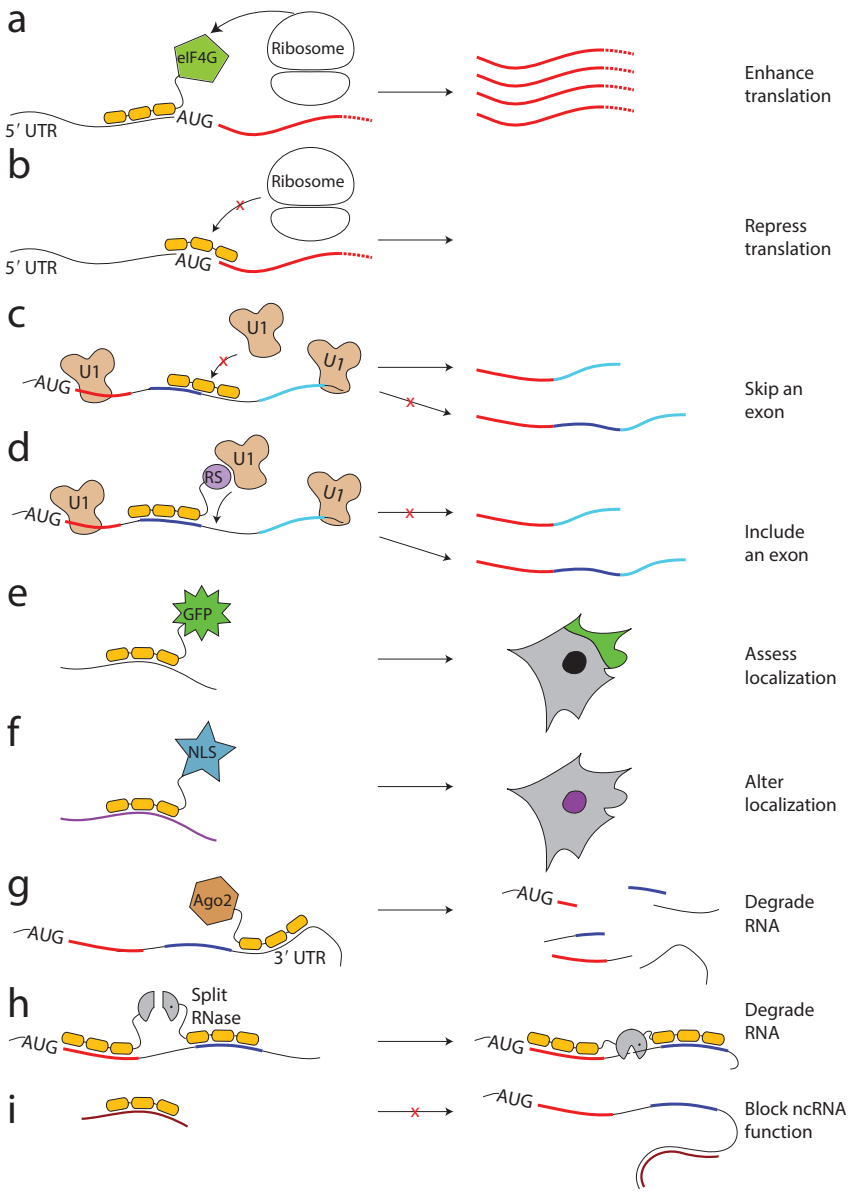

a
eIF4G
Ribosome
5′ UTR
AUG
Enhance translation
b
Ribosome
5′ UTR
AUG
Repress translation
c
U1
U1
U1
AUG
Skip an exon
d
U1
RS
U1
U1
AUG
Include an exon
e
GFP
Assess localization
f
NLS
Alter localization
g
Ago2
AUG
3′ UTR
AUG
Degrade RNA
h
Split RNase
AUG
AUG
Degrade RNA
i
AUG
Block ncRNA function

Figure 1-2: Possible uses of engineered RNA-binding proteins (RBPs). (a) Fusing the eukaryotic translation initiation factor eIF4G to an RBP targeted to the 5′ untranslated repeat (5′ UTR) of a messenger-RNA (mRNA) could drive translation. **(b)** An RBP that binds near the translational start codon could inhibit translation of an mRNA. **(c)** A 5′ splice site-binding RBP could block recruitment of the U1 component of the spliceosome, favoring the skipping of that exon. **(d)** Conversely, an RBP that recognizes a splicing enhancer site and is fused to an arginine- and serine-rich (RS) domain could favor inclusion of the associated exon. **(e)** An RBP fused to a fluorescent protein (such as GFP) could be used to track RNA localization in living cells. **(f)** An RBP fused to a nuclear localization signal (NLS) could be used to alter RNA localization. **(g)** An RBP fused to a protein such as Argonaute 2 (Ago2) and targeted to the 3′ UTR of an mRNA could promote the degradation of the message. **(h)** Fusion of an RBP to a nonspecific RNase could allow the cutting of a specific target RNA. This approach would work best using a split-RNase strategy analogous to that used for the successful zinc-finger nucleases, preventing widespread cleavage throughout the cell. **(i)** An RBP that tightly bound a specific noncoding RNA (ncRNA) could block its activity, providing a useful functional probe. Figure and figure legend are from Mackay et al. (2011).

In 2016, two unique studies shifted the paradigm in RNA targeting. In one paper, the RNA-binding protein PumHD (Pumilio homology domain) was engineered into a set of four canonical protein modules, each targeting a single RNA base (guanine, cytosine, adenosine, or uracil), in a process called Pumilio-based assembly, or 'Pumby' (Adamala et al., 2016). Such assembly would mirror the previous construction of the four canonical TALE protein modules targeting single DNA bases (Miller et al., 2011). Due to the required protein engineering for each unique target site, however, the use of Pumby in research can be costly in both time and material resources. In the other 2016 paper, the RNA-programmable (as opposed to protein-programmable) nuclease CRISPR-Cas9 that had revolutionized DNA-targeting was co-opted to target RNA, and dubbed 'RCas9' (Nelles et al., 2016). As Cas9 is intrinsically DNA-targeting, however, the efficiency of RCas9 remains to be determined. The following year, an intrinsically RNA-targeting CRISPR associated protein, Cas13, was engineered by the Zhang Lab to target and edit RNA (Abudayyeh et al., 2017; Cox et al., 2017). Two systems with unique protein architectures, Cas13a/c and Cas13b, were both exploited.

Certain CRISPR-Cas (clustered regularly interspaced short palindromic repeats

and CRISPR-associated proteins; also referred to as just CRISPR) systems have been known to target RNA for about a decade, when the Cmr complex (type III-B/C) was first found to interfere with the RNA of invading bacteriophages (Hale et al., 2009). Five years later, the Csm complex (type III-A/D), which was initially believed to target DNA, was instead described as RNA-targeting (Staals et al., 2014; Tamulaitis et al., 2014). The 2015 biocomputational discovery of C2c2/Cas13a (type VI-A) with predicted RNase activity (Shmakov et al., 2015) inspired further work to characterize RNA-targeting class 2 CRISPR systems. Following this, I set up a computational pipeline to look for additional class 2 CRISPR systems, and recruited Neena Pyzocha to join me and analyze the candidates generated by my computational pipeline. From these candidates, Neena and I identified Cas13b (type VI-B) and proceeded to characterize its function. Since both Cas13a and Cas13b have been predicted to target RNA, the study of both proteins became ongoing synergistic projects in the Zhang Lab, resulting in highly complementary studies (Abudayyeh et al., 2016; Smargon et al., 2017).

1.2 The CRISPR Field

CRISPR research originated accidentally in 1987, when the alkaline phosphate isozyme-converting gene *iap* was cloned (Ishino et al., 1987), and an intriguing bit of computational biology did not go unreported. In perhaps the most understated sentence in biology, the authors wrote, "An unusual structure was found in the 3′-end flanking region of *iap*" (Figure 1-3A). This structure, known as a 'direct repeat', would form the basis for the CRISPR acronym, and more importantly the basis for CRISPR adaptive immunity.

CRISPR-Cas systems, present in most archaea and roughly half of bacteria, help protect prokaryotes against the foreign nucleic acids of invading viruses (Makarova et al., 2006; Barrangou et al., 2007; Barrangou, 2013; Marraffini, 2015; Mohanraju et al., 2016). While a few CRISPR systems target RNA, most characterized systems target DNA (Koonin et al., 2017). CRISPR immunity is adaptive, and takes place in three

A

TGAAAATGGGAGGGAGTTCTACCGCAGAGGCGGGGGAACTCCAAGTGATATCCATCATCGCATCCAGTGCGCC (1,451)
(1,452) CGGTTTATCCCCGCTGATGCGGGGAACACCAGCGTCAGGCGTGAAATCTCACCGTCGTTGC (1,512)
(1,513) CGGTTTATCCCCTGCTGGCGCGGGGAACTCTCGGTTCAGGCGTTGCAAACCTGGCTACCGGG (1,573)
(1,574) CGGTTTATCCCCGCTAACGCGGGGAACTCGTAGTCCATCATTCCACCTATGTCTGAACTCC (1,634)
(1,635) CGGTTTATCCCCGCTGGCGCGGGGAACTCG (1,664)

consensus: CGGTTTATCCCCGCT$^{GG}_{AA}$CGCGGGGAACTC

FIG. 5. Comparison of direct-repeat sequences consisting of 61 base pairs in the 3′-end flanking region of *iap*. The 29 highly conserved nucleotides, which contain a dyad symmetry of 14 base pairs (underlined), are shown at the bottom. Homologous nucleotides found in at least two DNA segments are shown in boldface type. The second translational termination codon is boxed. The nucleotide numbers are in parentheses.

An unusual structure was found in the 3′-end flanking region of *iap* (Fig. 5). Five highly homologous sequences of 29 nucleotides were arranged as direct repeats with 32 nucleotides as spacing. The first sequence was included in the putative transcriptional termination site and had less homology than the others. Well-conserved nucleotide sequences containing a dyad symmetry, named REP sequences, have been found in *E. coli* and *Salmonella typhimurium* (28) and may act to stabilize mRNA (18). A dyad symmetry with 14 nucleotide pairs was also found in the middle of these sequences (underlining, Fig. 5), but no homology was found between these sequences and the REP sequence. So far, no sequence homologous to these has been found elsewhere in procaryotes, and the biological significance of these sequences is not known.

B

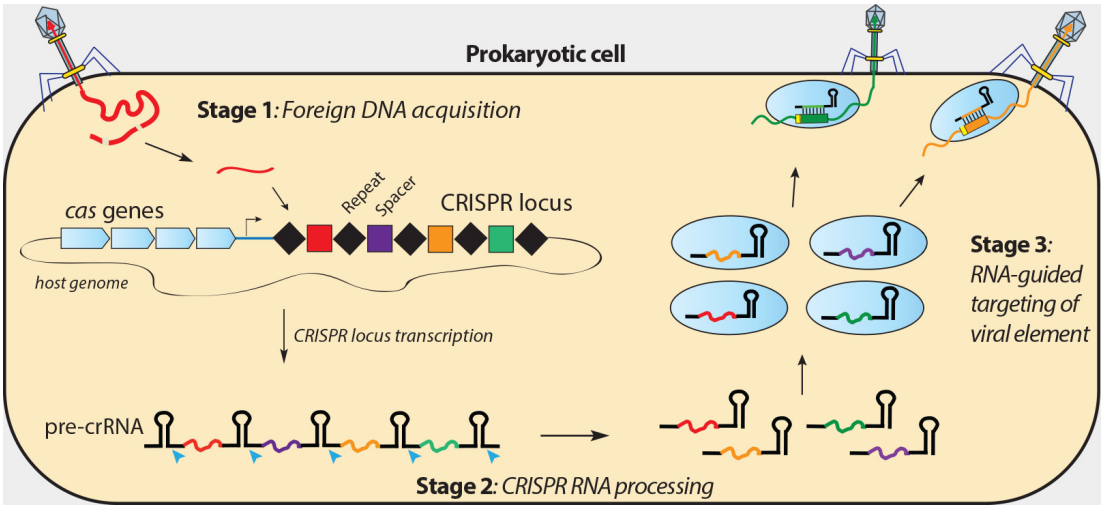


Figure 1-3: CRISPR origins and adaptive immunity mechanism. **(A)** Seminal finding of CRISPR direct-repeat sequence at 3′–end flanking region of *iap*, from Ishino et al. (1987). **(B)** CRISPR adaptive immunity mechanism, from the website of The Doudna Lab at UC Berkeley.

stages: 1) foreign DNA is acquired as a spacer between direct repeats in the CRISPR locus, which is then transcribed into pre-crRNA (pre-CRISPR RNA); 2) pre-crRNA is processed into mature crRNA and complexed with the Cas protein(s); 3) the CRISPR Cas-crRNA complex targets either DNA or RNA protospacers (sequences with reverse complementarity to spacers) of invading viruses (Figure 1-3B).

CRISPR-Cas systems are divided into two classes: the more prominent class 1 whose CRISPR-Cas complex contains multiple essential effector proteins, and the less abundant but highly diverse class 2 which relies on a single effector in complex with crRNA to target nucleic acids (Makarova et al., 2015; Koonin et al., 2017). These classes are further subdivided into types, defined by the Cas proteins involved in various functions, such as expression, interference, adaptation, and ancillary roles (Figure 1-4A). In 2015, prior to the commencement of my thesis research, five types of CRISPR systems had been described: types I, III, and IV in class 1 and types II and V in class 2 (Makarova et al., 2015). By 2017, six types have been described (including class 2 type VI), with twelve class 1 subtypes and seventeen class 2 subtypes (Koonin et al., 2017). These diverse class 2 subtypes are shown in Figure 1-4B.

Due to their compact nature, Class 2 CRISPR-Cas systems have proved instrumental in genome engineering applications since their initial exploitation in 2013 (Cong et al., 2013; Mali et al., 2013). The expanding diversity of these systems in succeeding years has generated a series of potential new tools to target DNA and now RNA, rendering previously speculative biology and biotechnology an accessible reality. When I began my thesis research in 2015, however, only the class 2 DNA-targeting effectors Cas9 and Cpf1 were known to the world (Jinek et al, 2012; Zetsche et al., 2015).

1.3 Thesis Problem and Statement

CRISPR-Cas systems rely on the machinery of Cas1 and Cas2, and particularly Cas1, in the adaption phase of immunity (Marraffini, 2015). Prior to this thesis work, it was widely believed in the field that putative systems lacking this adaptive machinery

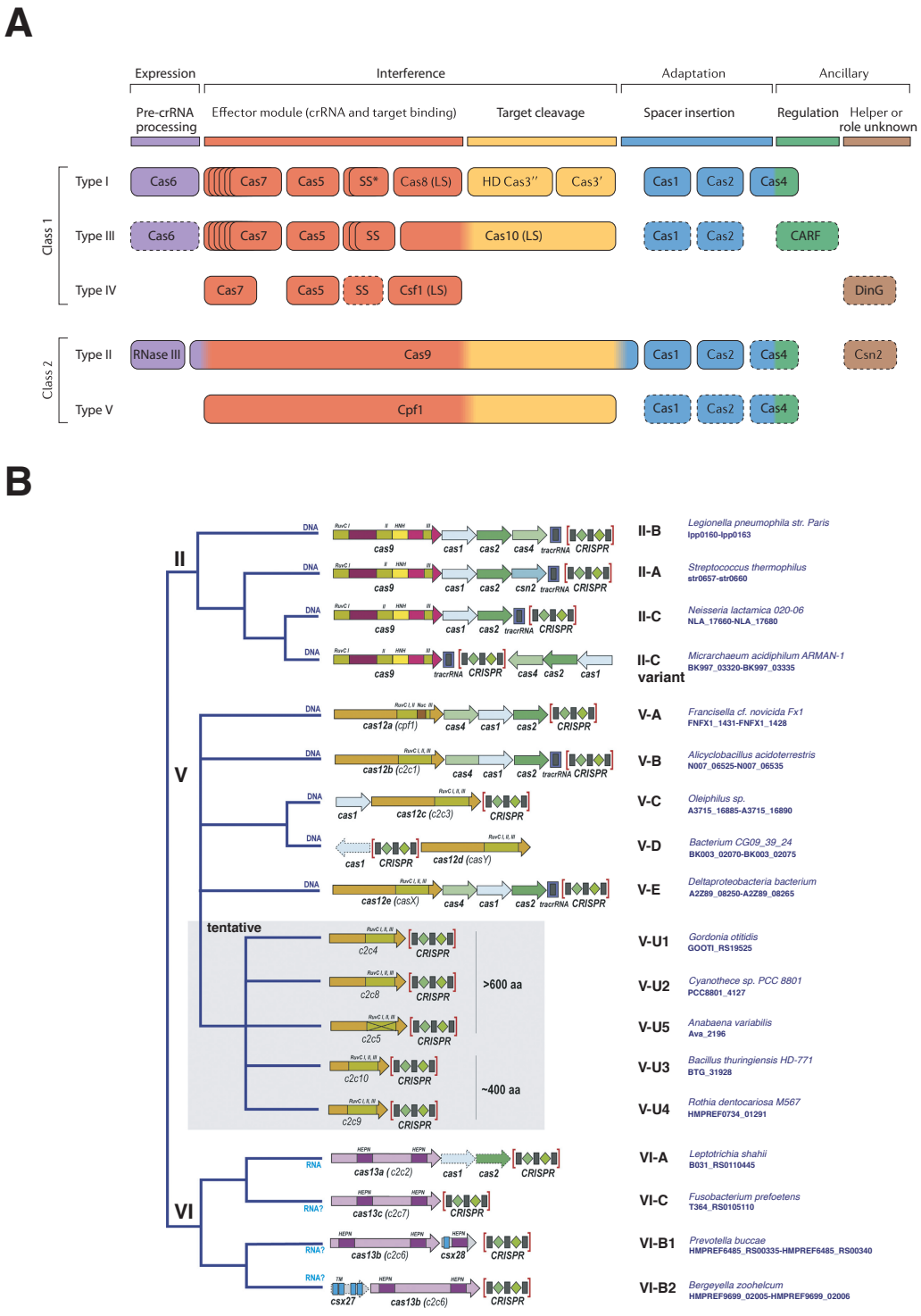


Figure 1-4: CRISPR classifications (c. 2015) and class 2 CRISPR classifications (c. 2017). **(A)** Classification of CRISPR systems by class, type and protein modules and functions, from Makarova et al. (2015). **(B)** Classification of Class 2 CRISPR systems, from Koonin et al. (2017).

would not be functional. For this reason, computational sequence database mining for CRISPR-Cas systems had been carried out using the *cas1* gene as a seed. This reasoned approach led to the discovery of Cas13a (Shamkov et al., 2015), which was uncharacterized experimentally when my thesis research commenced (but, again, the Cas13a and Cas13b initial studies would play contemporaneous synergistic roles in the Zhang Lab). Based on the absence of Cas1 and Cas2 in the proposed class 1 type IV system, my colleagues and I hypothesized that using the CRISPR array as the search seed may provide a more comprehensive census of CRISPR systems. Furthermore, if one of these systems were RNA-targeting, the research may lead to utilities beyond the CRISPR field.

My thesis statement can be put succinctly:

The aim of this research was to discover and characterize novel functional class 2 CRISPR systems. If any such systems targeted RNA, this research would help expand and improve the experimental toolkit to study RNA biology and advance RNA biotechnology.

In the remainder of the thesis, I fulfill this statement, at first through computational biology and later through biochemistry and genetics together with research colleagues in the Zhang Lab. The outcome of this thesis is the class 2 CRISPR RNA-targeting system Cas13b, which offers many opportunities for future research in RNA biology and biotechnology.

1.4 Thesis Organization

This thesis is divided into four chapters and two appendices. In Chapter 1, this introductory chapter, I provide background and motivation on RNA targeting and the CRISPR field, as well as present my thesis problem and statement. In Chapter 2, I elaborate on the search for novel function class 2 CRISPR systems, culminating in the discovery of Cas13b (type VI-B). Chapter 3 describes the characterization of Cas13b as a differentially regulated RNA-targeting system represented by two subtypes. In Chapter 4, I conclude with a summary and impact of my thesis, in

addition to a discussion of future research directions. Appendix A contains the tables relevant to Chapters 2 and 3. Appendix B comprises text from a 2017 *Nature Reviews Microbiology* analysis article of Class 2 CRISPR system diversity and evolution on which I am second author. Chapters 2 and 3 and Appendix B contain the work of multiple individuals. At the beginning of each, I provide a full citation and state my personal contributions. Altogether, these chapters and appendices encompass research spanning computational biology, biochemistry, and genetics–an exemplar of the interdisciplinary environment of today's research in the life sciences.

Chapter 2

Biocomputational Discovery of and Initial Results with Cas13b

This chapter is derived in part from the Cas13b study published in *Molecular Cell* (Smargon et al., 2017) and my Master's Thesis (Smargon, 2016). Full citation is as follows:

Smargon, A.A.*, Cox, D.B.T.*, Pyzocha, N.K.*, Zheng, K., Slaymaker, I.M., Gootenberg, J.S., Abudayyeh, O.A., Essletzbichler, P., Shmakov, S., Makarova, K.S., Koonin, E.V. and Zhang, F. (2017). *Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28.* Mol. Cell 65, 618−630.e7. (* denotes co-first authors)

Computational Contributions: With feedback from F.Z., I designed and implemented the biocomputational pipeline to discover new Class 2 CRISPR systems. K.Z. and N.K.P. assisted with a few individual components of its implementation, and N.K.P. assisted with curating the output of the pipeline. I performed the subsequent computational sequence analysis of Cas13b and Type VI-B CRISPR systems, with S.S., K.S.M., and E.V.K. providing input on annotation, classification, and naming.

Biochemical Contributions: Early on in the project, N.P., F.Z., and I working together optimized the nucleic acid preparation and nuclease assay, and designed and implemented experiments. This included carrying out the first successful single spacer BzCas13b RNA cleavage experiment and following up with spacer tiling experiments.

2.1 Summary

CRISPR-Cas adaptive immune systems defend microbes against foreign nucleic acids via RNA-guided endonucleases. Using a computational sequence database mining approach, here we identify two class 2 CRISPR-Cas systems (subtype VI-B) that lack Cas1 and Cas2 and encompass a single large effector protein, Cas13b, along with one of two previously uncharacterized associated proteins, Csx27 and Csx28. We show that Cas13b processes its own CRISPR array with short and long direct repeats, indicating that it may encapsulate a functional CRISPR system.

2.2 Introduction

CRISPR-Cas (clustered regularly interspaced short palindromic repeats and CRISPR-associated proteins) systems are divided into two classes, class 1 systems, which utilize multiple Cas proteins and CRISPR RNA (crRNA) to form an effector complex, and the more compact class 2 systems, which employ a large, single effector with crRNA to mediate interference (Makarova et al., 2015). CRISPR-Cas systems display a wide evolutionary diversity, involving distinct protein complexes and different modes of operation, including the ability to target RNA (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Hale et al., 2009; Jiang et al., 2016; Staals et al., 2013, 2014; Tamulaitis et al., 2014).

Computational sequence database mining for diverse CRISPR-Cas systems has been carried out by searching microbial genomic sequences for loci harboring the *cas1* gene, the most highly conserved *cas* gene involved in the adaptation phase of CRISPR immunity (Marraffini, 2015). Among other findings, this approach led to the discovery of the class 2 subtype VI-A system with its signature effector Cas13a (previously known as C2c2), which targets RNA (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Shmakov et al., 2015). Since distinct variants of class 1 CRISPR systems have

been discovered that lack *cas1* (Makarova et al., 2015), we sought to identify class 2 CRISPR-Cas systems lacking *cas1* by modifying the computational discovery pipeline so that it is not seeded on Cas1.

2.3 Computational Search for Novel Class 2 CRISPR Systems

We designed a computational pipeline to search specifically for putative class 2 CRISPR-Cas loci lacking Cas1 and Cas2. The pipeline consisted of two phases, a CRISPR locus discovery phase (stages 1-3) and a class 2 candidate discovery phase (stages 4-6) (Figure 2-1).

In the first phase of the pipeline, we annotated CRISPR loci from assembled microbial genomes. We downloaded ∼23K annotated assembled bacterial and archaeal genomes from the Ensembl Release 27 (June 2015) (Yates et al., 2016). For reference, Ensembl Release 37 (December 2017) contains ∼44K such genomes, effectively representing a doubling in just 2.5 years. Next, we searched for all CRISPR arrays in these assembled genomes. To achieve this, we implemented PILER-CR, open-source software for unbiased CRISPR array discovery that can scan through a "5Mb genome in around 5 seconds on a [2007] desktop computer" (Edgar, 2007). We used the default parameters, namely a minimum of 3 direct repeats with 90% sequence conservation, repeats of size 16–64 nucleotides, and spacers of size 8–64 nucleotides. With respect to previous CRISPR literature, PILER-CR with default parameters has 100% sensitivity and 94% specificity (Edgar, 2007).

From PILER-CR, we detected ∼25K CRISPR arrays in the ∼23K genomes. Fewer than half of the genomes contained CRISPR arrays, and often a single genome would contain three or more arrays, each with potentially distinct direct repeats. For each of these arrays, we determined CRISPR loci by including all annotated genes

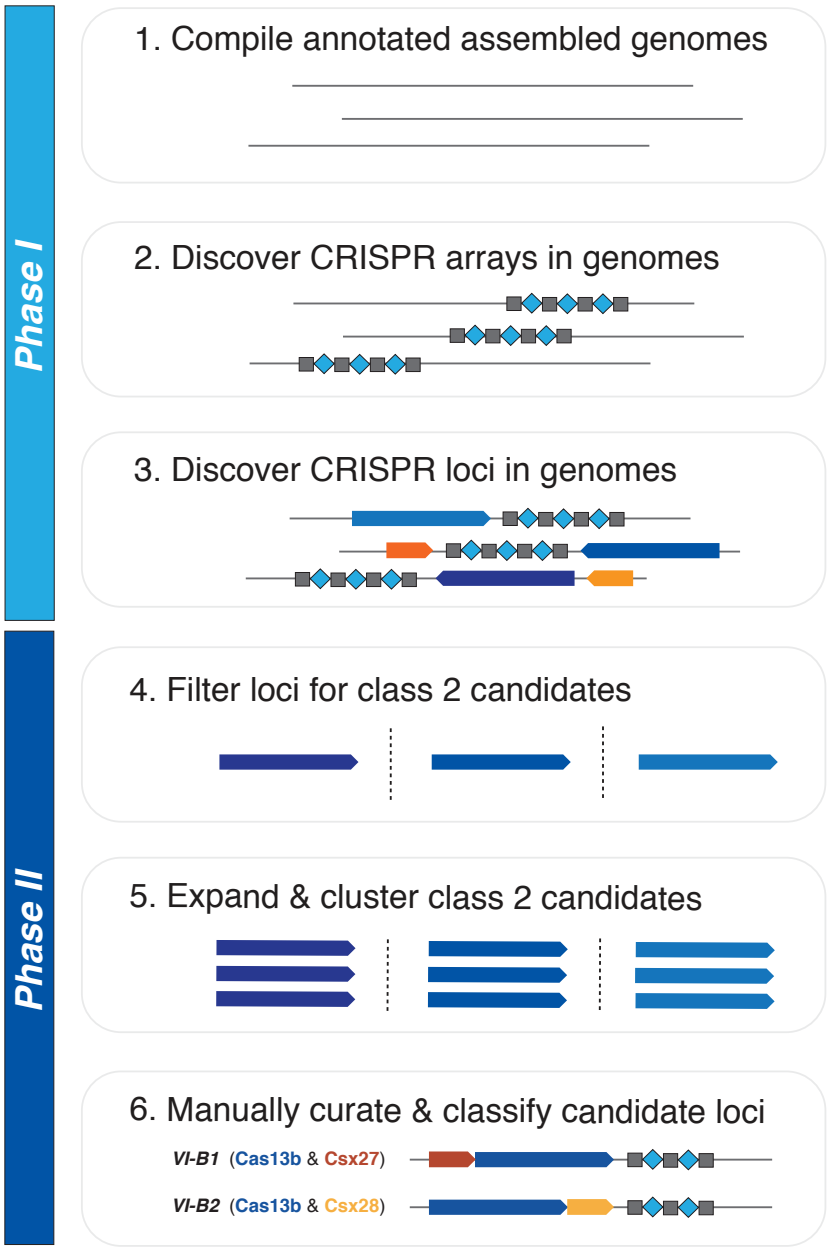


Figure 2-1: Bioinformatic pipeline to discover putative class 2 CRISPR loci lacking Cas1 and Cas2. Phase I is the CRISPR locus discovery phase, and Phase II is the class 2 candidate discovery phase.

10kb upstream and downstream of the detected arrays. Of the ∼25K defined CRISPR loci, ∼13K did not possess Cas1 or Cas2 at the locus, and ∼12K contained a single large effector, defined as only one protein greater than 700 amino acid (aa) residues in length, and thus less likely to be part of a class 1 CRISPR system.

In the second phase of the pipeline, we examined the CRISPR loci for the putative class 2 candidates most likely to be functional. The intersection of loci with no Cas1 or Cas2 and with a single large effector reduced to ∼5K loci, or about 20% of original CRISPR loci. These loci were further filtered to those most likely to be functional and compact. Informed in part by the size of previously classified class 2 effectors, we chose a lower bound for putative single effector size of 900 aa. Due to considerations of packaging in adeno-associated virus for systemic gene delivery, we chose an upper bound of 1800 aa. This further reduced the list of putative class 2 CRISPR systems to ∼1500 candidates.

To find all homologous proteins to these candidates for later analysis, we performed NCBI BLAST on the ∼1500 candidate single effectors against the NCBI non-redundant protein database (Camacho et al., 2009). Through an exhaustive homology search with an E-value cutoff of 1e-7, the list of class 2 candidates grew to ∼7200. From here, we clustered candidate loci through a nearest-neighbor E-value cutoff of 1e-7, which yielded 266 groups. After generating visualization outputs for each unique locus, we manually inspected these 266 groups containing ∼7200 loci. For the most accurate classification, this inspection occurred one locus at a time per group. Interestingly, often a CRISPR system would span multiple groups, or one group might contain multiple CRISPR systems—highlighting the imperfection of automated pipelines and confirming the necessity of manual curation one locus at a time.

Each locus was analyzed first for architecture, namely the position and orientation of CRISPR-Cas and other annotated proteins near the CRISPR array (Figure 2-

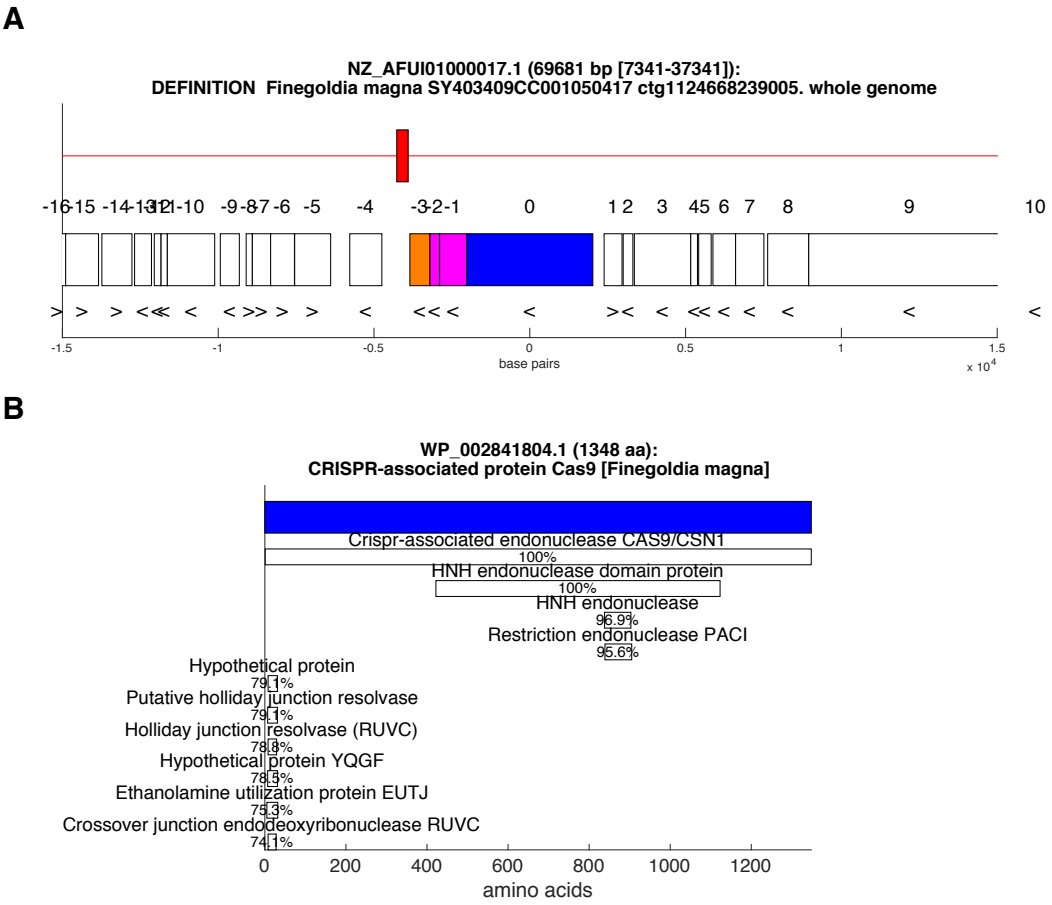


Figure 2-2: Representative pipeline visualization output for *Finegoldia magna* Cas9: candidate locus plot and protein homology. (A) Genome accession number, length, range, and name above graph of genomic locus containing CRISPR array in red and proteins in blue (selected protein), magenta (annotated Cas1, Cas 2), orange (other annotated CRISPR associated proteins), and white (all other proteins) with protein orientations depicted. **(B)** Select protein accession number, length, and name above top 10 HHpred homology hits and their respective probabilities.

2A). To check if the putative class 2 effector contained known catalytic domains, we performed HHpred, a hidden Markov homology model that predicts the likelihood of a protein containing established protein domains and subdomains (Hildebrand et al., 2009; Remmert et al., 2011). The top 10 protein domain and subdomain predictions were then plotted along the protein to aid in visualization (Figure 2-2B).

After inspecting the locus architecture and predicted single effector domains/subdomains, we were next interested in whether we could classify, and thus exclude, a candidate system as an existing CRISPR system. Here it was useful to examine the genes nearby the single effector for annotated CRISPR-Cas proteins (Figure 2-3A), and also to study the consensus direct repeat as determined by PILER-CR (Figure 2-3B). Together, we synthesized this information into a CRISPR system classification decision tree based on prior CRISPR-Cas system classification literature Figure 2-3C) (Makarova et al., 2015).

Having examined each locus in each group, we designated the 266 groups as containing one or more systems, either characterized CRISPR, characterized non-CRISPR, or uncharacterized. A number of characterized class 2 CRISPR systems containing Cas1 and Cas2 (e.g., Cas9, Cpf1, and Cas13a) slipped through the search, likely because one or more homologs of an effector lost both Cas1 and Cas2 at its locus during evolution. Outside of known class 2 CRISPR systems, we filtered out several categories of proteins from consideration. Most numerous were those well categorized by HHpred (greater than 80% homology to known non-CRISPR proteins) and of which only one or two group members were proximal to a CRISPR array. These included helicases, ATPases, kinases, proteases, and transferases, among other proteins. Second most numerous were those groups containing only one or two proteins. Third most numerous were class 1 CRISPR loci that had passed the large single effector filter, mainly Cas3 but occasionally Csm and Cmr proteins.

In the end, we were interested in uncharacterized single effectors most likely

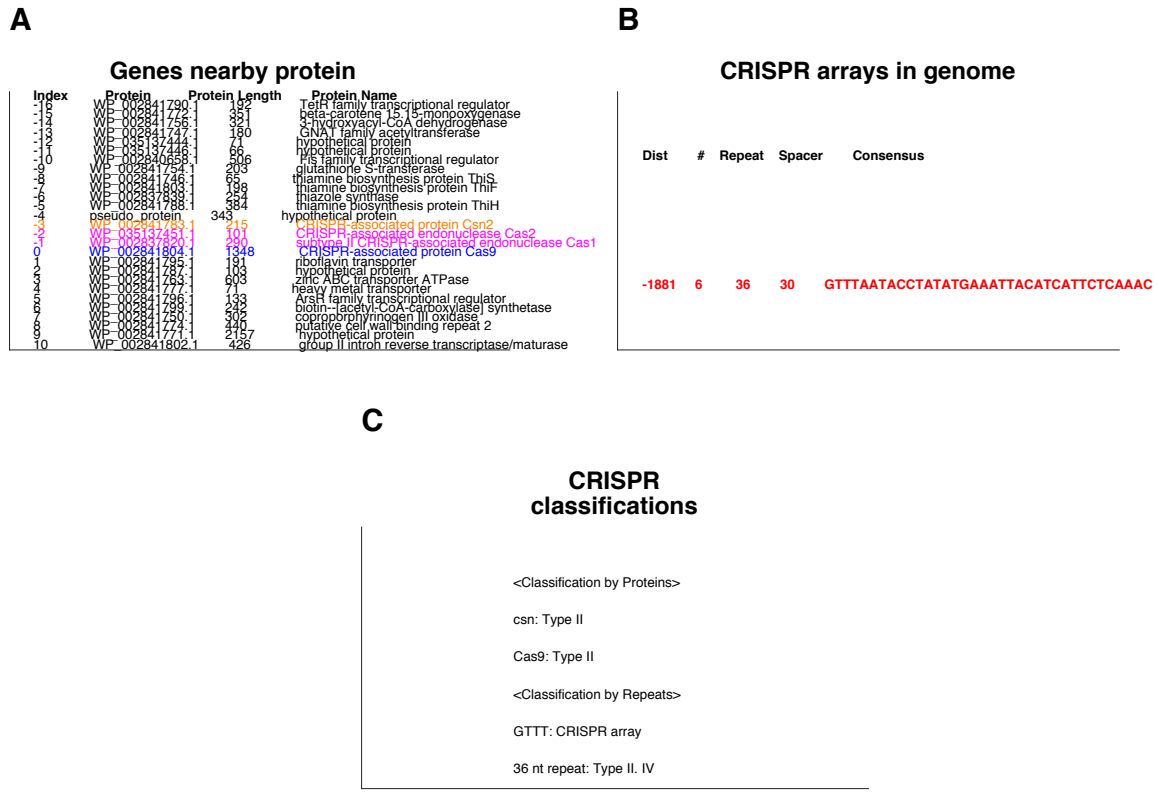


Figure 2-3: Representative pipeline visualization output for *Finegoldia magna* Cas9: candidate locus genes, CRISPR arrays, and tentative CRISPR classifications. **(A)** Key of genes nearby select proteins with index, accession number, length in amino acids, and name. **(B)** CRISPR arrays in genome with distance from select protein (negative if upstream of protein), number of repeats, length of repeats, length of spacers, and consensus repeat sequence. **(C)** Tentative CRISPR classifications from decision tree based on existing literature.

to encompass class 2 CRISPR systems. Thus, we considered only groups or merged groups of which we could be most confident—namely those with greater than 10 distinct candidates, greater than 50% of which contained CIRSPR arrays at their loci. This ruled out a number of candidates that had piqued our curiosity, including proteins annotated as methylases, transposases, integrases, and nucleases. Based on all the aforementioned considerations, we settled exclusively on the merged Groups 29 and 30.

2.4 Discovery of Class 2 Subtype VI-B System with Cas13b

Groups 29 and 30 held a protein ('hypothethical protein') with no confident domain/subdomain predictions by HHpred (Figure 2-4A) and only a similarity to type II and IV systems by its 36 nucleotide CRISPR direct repeat (Figure 2-4B). No known CRISPR-Cas proteins were present at any of loci in immediate proximity (Figure 2-4C), indicating that this putative system was isolated in all respects except by its CRISPR array. This led us to search for any nearby proteins that might be conserved in relative position and orientation to the candidate effector. By iterating on this approach, we discovered two additional proteins, 'small protein 1' and 'small protein 2'. (After experiments described in Chapter 3, these small proteins would eventually become 'Csx27/Csx28', while the candidate effector became 'Cas13b'.)

Groups 29 and 30 formed two genetically diverse putative class 2 CRISPR-Cas systems (105 genomic loci, 81 containing a unique entry Cas13b in the non-redundant NCBI protein database, and 71 of these 81 containing an annotated CRISPR array) represented in Gram-negative bacteria (Figure 2-6A). For some genera, in particular *Porphyromonas* and *Prevotella*, Cas13b proteins are encoded in several unique sequenced loci and, occasionally, in the

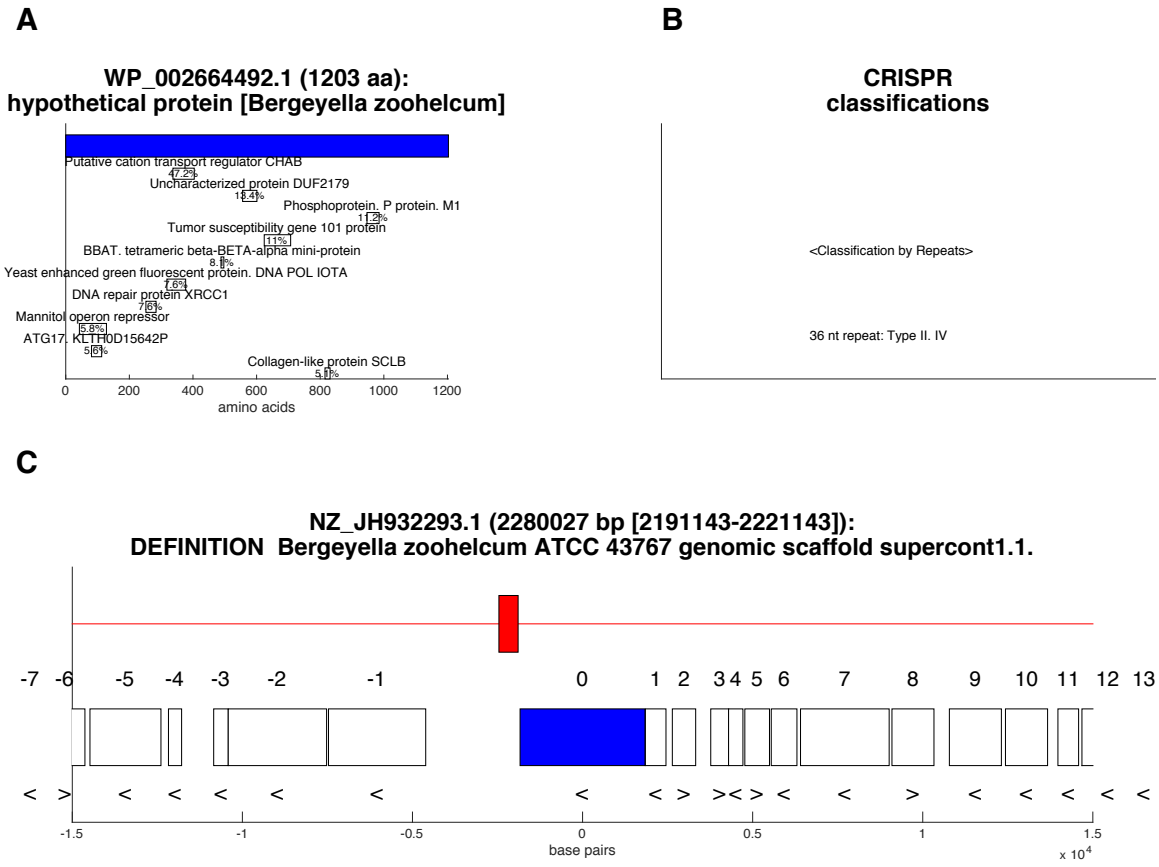


Figure 2-4: Critical elements of pipeline visualization output for manual curation of *Bergeyella zoohelcum* Cas13b (originally "hypothetical protein" in Group 29). (A) Select protein accession number, length, and name above top 10 HHpred homology hits and their respective probabilities. **(B)** CRISPR arrays in genome with distance from select protein (negative if upstream of protein), number of repeats, length of repeats, length of spacers, and consensus repeat sequence.

same sequenced genome. These systems often co-occur with other CRISPR-Cas systems. Of the 81 type VI-B loci found across complete and incomplete bacterial genomes, 62 also possess at least one other CRISPR-Cas locus that includes the key adaptation endonuclease, Cas1. However, three complete genomes carrying the type VI-B locus (Flavobacterium_branchiophilum_FL_15 _GCA_000253275.1, Paludibacter_propionicigenes_WB4_GCA_000183135.1, and Porphyromonas_gingivalis _AJW4_GCA_001274615.1) lack Cas1 altogether (Figure 2-6A).

All VI-B loci encode a large (∼1,100 aa) candidate effector protein and, in about 80% of the cases, an additional small (∼200 aa) protein (Figures 2-5 and 2-6A). The putative effector proteins contain two predicted HEPN domains (Anantharaman et al., 2013) at their N and C termini (Figure 2-6B), similar to the domain architecture of the large effector of subtype VI-A (Cas13a) (Shmakov et al., 2015). Beyond the occurrence of two HEPN domains, however, there is no significant sequence similarity between the predicted effector and Cas13a. These systems were also identified by a generalized version of the pipeline described above as part of a comprehensive analysis of class 2 CRISPR-Cas systems and were classified into subtype VI-B, with predicted effector protein Cas13b (Shmakov et al., 2017).

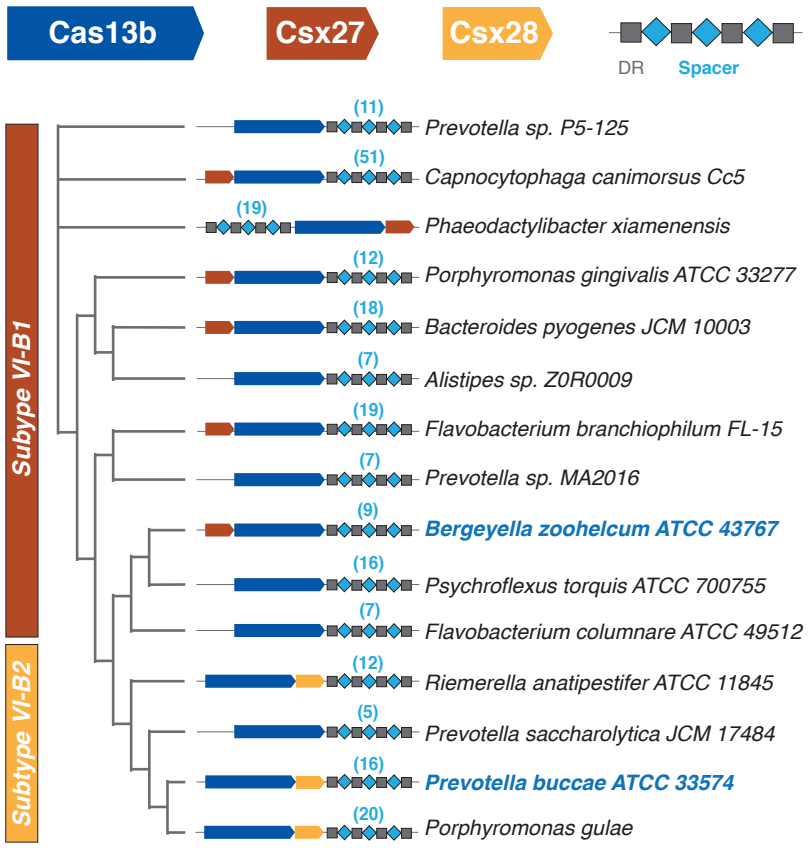
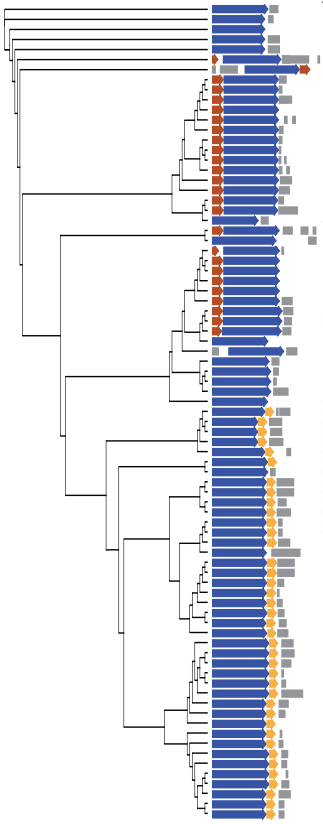


Figure 2-5: Discovery of Two Class 2 CRISPR-Cas Systems, Subtype VI-B1 and VI-B2, Containing Cas13b. A schematic phylogenetic tree of the subtype VI-B loci. Loci with Csx27 (brown) comprise variant VI-B1; loci with Csx28 (gold) comprise variant VI-B2. Strains in blue (*Bergeyella zoohelcum ATCC 43767* and *Prevotella buccae ATCC 33574*) were characterized experimentally. See also Figures 2-6, 2-7, and 3-10.

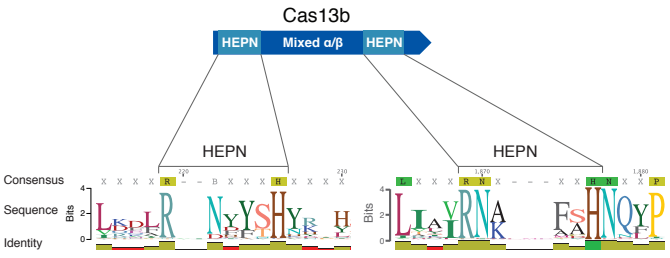
A

Phylogenetic tree of 81 non-redundant Cas13b effectors with full gene



Species (Genome Accession)	Cas13b Accession	Csx27/28 Accession	# Spacers	Cas1 in genome?	Cas13b Size (aa)
Paludibacter propionicigenes WB4 (NC_014734.1)	WP_013446107.1	NA	8	N	1155
Prevotella sp. P5-60 (NZ_JXQJ01000080.1)	WP_044074780.1	NA	5	Y	1091
Prevotella sp. P4-76 (NZ_JXQI01000021.1)	WP_044072147.1	NA	0	?	1091
Prevotella sp. P5-125 (NZ_JXQL01000055.1)	WP_044065294.1	NA	11	?	1091
Prevotella sp. P5-119 (NZ_JXQK01000043.1)	WP_042518169.1	NA	11	Y	1091
Capnocytophaga canimorsus Cc5 (NC_015846.1)	WP_013997271.1	WP_013997274.1	51	Y	1200
Phaeodactylibacter xiamenensis (NZ_JPOS01000018.1)	WP_044218239.1	WP_044218241.1	19	Y	1132
Porphyromonas gingivalis W83 (NC_002950.2)	WP_005873511.1	WP_005873518.1	7	Y	1136
Porphyromonas gingivalis F0570 (NZ_KI259168.1)	WP_021665475.1	WP_021665476.1	3	Y	1136
Porphyromonas gingivalis ATCC 33277 (NC_010729.1)	WP_012458151.1	WP_012458152.1	12	Y	1136
Porphyromonas gingivalis F0185 (AWVC01000122.1)	ERJ81987.1	ERJ81988.1	0	Y	1136
Porphyromonas gingivalis F0185 (NZ_KI259960.1)	WP_021677657.1	WP_021677658.1	6	Y	1136
Porphyromonas gingivalis SJD2 (NC_KI629875.1)	WP_023846767.1	WP_005873518.1	4	Y	1136
Porphyromonas gingivalis F0568 (AWUU01000145.1)	ERJ85637.1	ERJ85638.1	3	Y	1136
Porphyromonas gingivalis W4087 (AWVE01000130.1)	ERJ87335.1	ERJ87336.1	2	Y	1136
Porphyromonas gingivalis W4087 (NZ_KI260263.1)	WP_021680012.1	WP_005873518.1	4	Y	1136
Porphyromonas gingivalis F0568 (NZ_KI258981.1)	WP_021663197.1	WP_021663198.1	6	Y	1136
Porphyromonas gingivalis (NZ_LOEL01000010.1)	WP_061156637.1	WP_005873518.1	11	Y	1136
Porphyromonas gulae (NZ_JRAQ01000019.1)	WP_039445055.1	WP_039445052.1	10	Y	1136
Bacteroides pyogenes F0041 (KE993153.1)	ERI81700.1	ERI81699.1	5	Y	1116
Bacteroides pyogenes JCM 10003 (NZ_BAIU01000001.1)	WP_034542281.1	WP_034542279.1	18	Y	1116
Alistipes sp. ZOR0009 (NZ_JTLD01000029.1)	WP_047447901.1	NA	7	?	954
Flavobacterium branchiophilum FL-15 (NC_016001.1)	WP_014084666.1	WP_014084665.1	19	N	1151
Prevotella sp. MA2016 (NZ_JHUW01000010.1)	WP_036929175.1	NA	7	?	1323
Myroides odoratimimus CCUG 10230 (AGEC02000017.1)	EHO06562.1	EHO06560.1	2	Y	1160
Myroides odoratimimus CCUG 3837 (AGZK01000016.1)	EKB06014.1	EKB06015.1	0	?	1158
Myroides odoratimimus CCUG 3837 (NZ_JH815535.1)	WP_006265509.1	WP_006265510.1	0	?	1158
Myroides odoratimimus CCUG 12901 (NZ_JH590834.1)	WP_006261414.1	WP_006261415.1	0	?	1158
Myroides odoratimimus CCUG 12901 (AGED01000033.1)	EHO08761.1	EHO08762.1	0	?	1158
Myroides odoratimimus (NZ_CP013690.1)	WP_058700060.1	WP_006261415.1	10	Y	1160
Bergeyella zoohelcum ATCC 43767 (AGYA01000037.1)	EKB54193.1	EKB54194.1	9	Y	1225
Capnocytophaga cynodegmi (NZ_CDOD01000002.1)	WP_041989581.1	WP_041989578.1	7	?	1219
Bergeyella zoohelcum ATCC 43767 (NZ_JH932293.1)	WP_002664492.1	WP_034985946.1	8	Y	1225
Flavobacterium sp. 316 (NZ_JYGZ01000003.1)	WP_045968377.1	NA	0	Y	1156
Psychroflexus torquis ATCC 700755 (NC_018721.1)	WP_015024765.1	NA	16	Y	1146
Flavobacterium columnare ATCC 49512 (NC_016510.2)	WP_014165541.1	NA	7	Y	1180
Flavobacterium columnare (NZ_CP013992.1)	WP_060381855.1	NA	5	Y	1214
Flavobacterium columnare (NZ_CP015107.1)	WP_063744070.1	NA	3	Y	1214
Flavobacterium columnare (NZ_CP016277.1)	WP_065213424.1	NA	14	Y	1215
Chryseobacterium sp. YR477 (NZ_KN549099.1)	WP_047431796.1	NA	0	?	1146
Riemerella anatipestifer ATCC 11845 = DSM 15868 (NC_014738.1)	WP_004919755.1	WP_004919758.1	12	Y	1096
Riemerella anatipestifer RA-CH-2 (NC_020125.1)	WP_015345620.1	WP_004919758.1	12	Y	949
Riemerella anatipestifer (NZ_CP007504.1)	WP_049354263.1	WP_004919758.1	11	Y	949
Riemerella anatipestifer (NZ_LUDU01000012.1)	WP_061710138.1	WP_061710139.1	13	Y	951
Riemerella anatipestifer (NZ_LUDI01000010.1)	WP_064970887.1	WP_064970885.1	4	Y	1096
Prevotella saccharolytica F0055 (AMEP01000091.1)	EKY00089.1	EKY00090.1	0	?	1151
Prevotella saccharolytica JCM 17484 (NZ_BAKN01000001.1)	WP_051522484.1	NA	5	Y	1152
Prevotella buccae ATCC 33574 (AEPD01000005.1)	EFU31981.1	EFU31982.1	16	Y	1128
Prevotella buccae ATCC 33574 (NZ_GL586311.1)	WP_004343973.1	WP_004343974.1	16	Y	1128
Prevotella buccae D17 (NZ_GG739967.1)	WP_004343581.1	WP_004343582.1	8	?	1128
Prevotella sp. MSX73 (NZ_ALJQ01000043.1)	WP_007412163.1	WP_036927782.1	13	Y	1128
Prevotella pallens ATCC 700821 (AFPY01000052.1)	EGQ18444.1	EGQ18443.1	4	?	1126
Prevotella pallens ATCC 700821 (NZ_GL982513.1)	WP_006044833.1	WP_050795200.1	4	?	1126
Prevotella intermedia ATCC 25611 = DSM 20706 (NZ_JAEZ01000017.1)	WP_036860899.1	WP_050795200.1	11	Y	1127
Prevotella intermedia (NZ_LBGT01000010.1)	WP_061868553.1	NA	27	Y	1121
Prevotella intermedia 17 (CP003502.1)	AFJ07523.1	AFJ07898.1	16	Y	1135
Prevotella intermedia (NZ_AP014926.1)	WP_050955369.1	WP_014708440.1	16	Y	1133
Prevotella intermedia (AP014598.1)	BAU18623.1	BAU18624.1	6	Y	1134
Prevotella intermedia ZT (ATMK01000017.1)	KJJ86756.1	KJJ86755.1	2	?	1126
Prevotella aurantiaca JCM 15754 (NZ_BAKF01000019.1)	WP_025000926.1	WP_036889078.1	5	Y	1125
Prevotella pleuritidis F0068 (NZ_AWET01000045.1)	WP_021584635.1	WP_021584705.1	6	Y	1140
Prevotella pleuritidis JCM 14110 (NZ_BAJN01000005.1)	WP_036931485.1	WP_024991772.1	7	Y	1117
Prevotella falsenii DSM 22864 = JCM 15124 (NZ_BAJY01000004.1)	WP_036884929.1	WP_051527348.1	10	Y	1134
Porphyromonas gulae (NZ_JRAT01000012.1)	WP_039418912.1	WP_052073447.1	11	Y	1176
Porphyromonas sp. COT-052 OH4946 (NZ_JQZY01000014.1)	WP_039428968.1	WP_050563578.1	12	Y	1176
Porphyromonas gulae (NZ_JRFD01000046.1)	WP_039442171.1	WP_050563578.1	9	Y	1175
Porphyromonas gulae (NZ_JRAJ01000010.1)	WP_039431778.1	WP_046201041.1	2	Y	1176
Porphyromonas gulae (NZ_KQ040500.1)	WP_046201018.1	WP_046201041.1	4	Y	1176
Porphyromonas gulae (NZ_JRAL01000022.1)	WP_039434803.1	WP_039434800.1	20	Y	1176
Porphyromonas gulae (NZ_JRAI01000002.1)	WP_039419792.1	WP_052078041.1	9	Y	1120
Porphyromonas gulae (NZ_JRAK01000129.1)	WP_039426176.1	WP_039426172.1	6	Y	1120
Porphyromonas gulae (NZ_KN294104.1)	WP_039437199.1	WP_052102013.1	0	Y	1120
Porphyromonas gingivalis TDC60 (NC_015571.1)	WP_013816155.1	WP_043890185.1	2	Y	1120
Porphyromonas gingivalis ATCC 33277 (NC_010729.1)	WP_012458414.1	WP_012458413.1	4	Y	1120
Porphyromonas gingivalis A7A1-28 (NZ_CP013131.1)	WP_058019250.1	WP_043898408.1	6	Y	1176
Porphyromonas gingivalis JCVI SC001 (APMB01000175.1)	EOA10535.1	EOA10563.1	5	?	1176
Porphyromonas gingivalis W50 (NZ_AJZS01000051.1)	WP_005874195.1	WP_010955981.1	2	Y	1176
Porphyromonas gingivalis (NZ_CP011995.1)	WP_052912312.1	WP_010955981.1	7	Y	1176
Porphyromonas gingivalis AJW4 (NZ_CP011996.1)	WP_053444417.1	WP_043898408.1	11	N	1120
Porphyromonas gingivalis (NZ_CP007756.1)	WP_039417390.1	WP_021665928.1	5	Y	1120
Porphyromonas gingivalis (NZ_LOEL01000001.1)	WP_061156470.1	WP_021663076.1	5	Y	1120

B



C

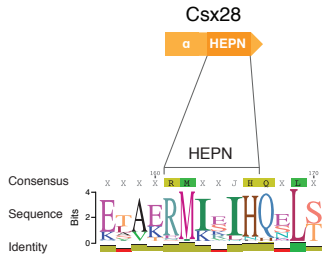


Figure 2-6: Phylogenetic tree of Cas13b bifurcates into two variants of subtype VI-B CRISPR loci. Related to Figure 2-5. (A) A phylogenetic tree (alignment generated by BLOSUM62) of non-redundant Cas13b effectors, with the full type VI-B locus depicted in every instance. Accession numbers for genome, Cas13b (blue), and Csx27 (brown)/Csx28 (gold) are included, as well as number of nearby spacers detected by PILER-CR, the presence of Cas1 in the sequenced genome, and the size of Cas13b. **(B)** Two HEPN sequences identified via multiple sequence alignment (BLOSUM62) of putative non-redundant Cas13b proteins. **(C)** Divergent HEPN sequence identified via multiple sequence alignment (BLOSUM62) of putative non-redundant Csx28 proteins.

2.5 CRISPR-Cas13b Loci Contain Small Accessory Proteins

The identity of the putative accessory protein correlates with the two distinct branches in the phylogenetic tree of Cas13b (Figures 2-5 and 2-6A) (Henikoff and Henikoff, 1992), indicative of the existence of two variant systems, which we denote VI-B1 (accessory protein referred to as Csx27) and VI-B2 (accessory protein referred to as Csx28). While subtype VI-B2 systems almost invariably contain *csx28*, *csx27* is less consistently represented in VI-B1 loci. The protein sequences of Csx27 and Csx28 show no significant similarity to any previously identified Cas proteins. (These proteins are characterized in Chapter 3.)

2.6 Cas13b-Associated CRISPR Arrays Display Unique Features

In contrast to their differing putative accessory proteins, both variants of subtype VI-B systems show distinct, conserved features in the CRISPR arrays. The direct repeats in the CRISPR arrays are conserved in size, sequence, and structure, with a length of 36 nt, a poly-U stretch in the open loop region, and complementary

sequences 5′-GUUG and CAAC-3′ at the ends of the repeat predicted to yield a defined secondary structure mediated by intramolecular base-pairing (Figures 2-7A− 2-7C) (Lorenz et al., 2011). Our analysis revealed 36 Cas13b spacers mapped with greater than 80% homology to unique protospacers in phage genomes. Twenty-seven of the identified Cas13b spacers targeted the coding strand of phage mRNA, while seven spacers targeted the noncoding strand and two spacers targeted regions of the phage genome without predicted transcripts. Although the composite of these imperfect mappings revealed no consensus flanking region sequence (Figure 2-7D) (Biswas et al., 2013), the well-conserved protospacer length of 30 nt, combined with the conserved direct repeat sequence and length, suggests that the nucleic acid targeting rules may be similar among different VI-B loci.

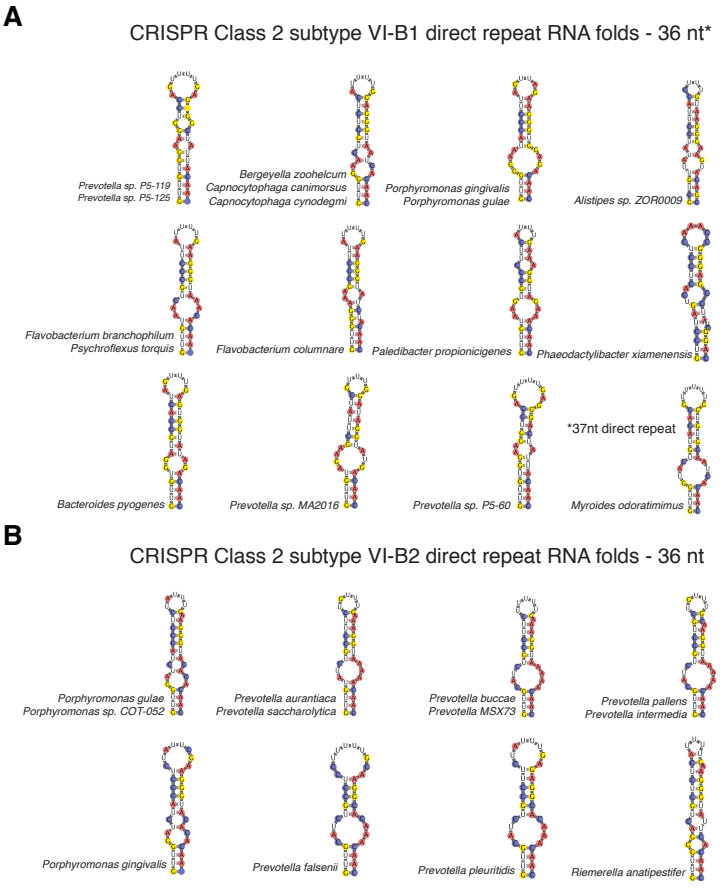

A
CRISPR Class 2 subtype VI-B1 direct repeat RNA folds - 36 nt*
Prevotella sp. P5-119
Prevotella sp. P5-125
Bergeyella zoohelcum
Capnocytophaga canimorsus
Capnocytophaga cynodegmi
Porphyromonas gingivalis
Porphyromonas gulae
Alistipes sp. ZOR0009
Flavobacterium branchophilum
Psychroflexus torquis
Flavobacterium columnare
Paledibacter propionicigenes
Phaeodactylibacter xiamenensis
Bacteroides pyogenes
Prevotella sp. MA2016
Prevotella sp. P5-60
*37nt direct repeat
Myroides odoratimimus
B
CRISPR Class 2 subtype VI-B2 direct repeat RNA folds - 36 nt
Porphyromonas gulae
Porphyromonas sp. COT-052
Prevotella aurantiaca
Prevotella saccharolytica
Prevotella buccae
Prevotella MSX73
Prevotella pallens
Prevotella intermedia
Porphyromonas gingivalis
Prevotella falsenii
Prevotella pleuritidis
Riemerella anatipestifer
C
Conservation of 36 nt direct repeats (24 unique)
Bits
Direct repeat
D
Flanking regions of protospacers mapped to phage genomes (36 unique)
Bits
Protospacer

Figure 2-7: Predicted sequence and secondary structure of type VI-B direct repeats; predicted protospacer flanking sequences. Related to Figure 2-5. **(A)** Predicted secondary structure folds of structurally unique CRISPR class 2 type VI-B1 direct repeats (Vienna RNAfold). **(B)** Predicted secondary structure folds of structurally unique CRISPR Class 2 type VI-B2 direct repeats. **(C)** Weblogo of all unique VI-B direct repeat sequences of length 36 nt, taken as the same transcriptional orientation as Cas13b. **(D)** Weblogo of all unique VI-B protospacer flanking sequences from CRISPRTarget mapping of protospacers to phage databases.

2.7 Cas13b Processes Its Associated CRISPR Array

With the VI-B system classified computationally, it was time to turn to experimental validation. A critical proof-of-concept experiment for characterizing any CRISPR system is testing for RNA processing of its associated CRISPR array, for, if processing does not occur to produce mature crRNA from pre-crRNA, the system is unlikely to be functional in subsequent programmable targeting of nucleic acids. To test CRISPR array processing for putative class 2 CRISPR-Cas13b, we conducted complementary experiments. Genetically, we performed RNA sequencing of native *B. zoohelcum* to determine whether pre-crRNA processing occurs in vivo. Following up biochemically, we performed an in vitro cleavage assay with purified Cas13b from *B. zoohelcum* and synthesized pre-crRNA associated with the same system.

RNA sequencing of the total RNA from *B. zoohelcum* (subtype VI-B1) showed processing of the pre-crRNA into a 66 nt mature crRNA, with the full 30 nt 5′ spacer followed by the 3′ direct repeat (Figure 2-8A) (Heidrich et al., 2015; Li and Durbin, 2009; Shmakov et al., 2015). A longer 118 nt crRNA, distal to the 36 nt crRNAs in the CRISPR array and with a direct repeat consisting of 5′ and 3′ fragments of the 36 nt direct repeat sequence interrupted by an intervening repeat sequence, was also processed. This phenomenon was computationally predicted to occur in additional VI-B loci, such as those from *Capnocytophaga canimorsus*, *Myroides odoratimimus*, and *Riemerella anatipestifer*.

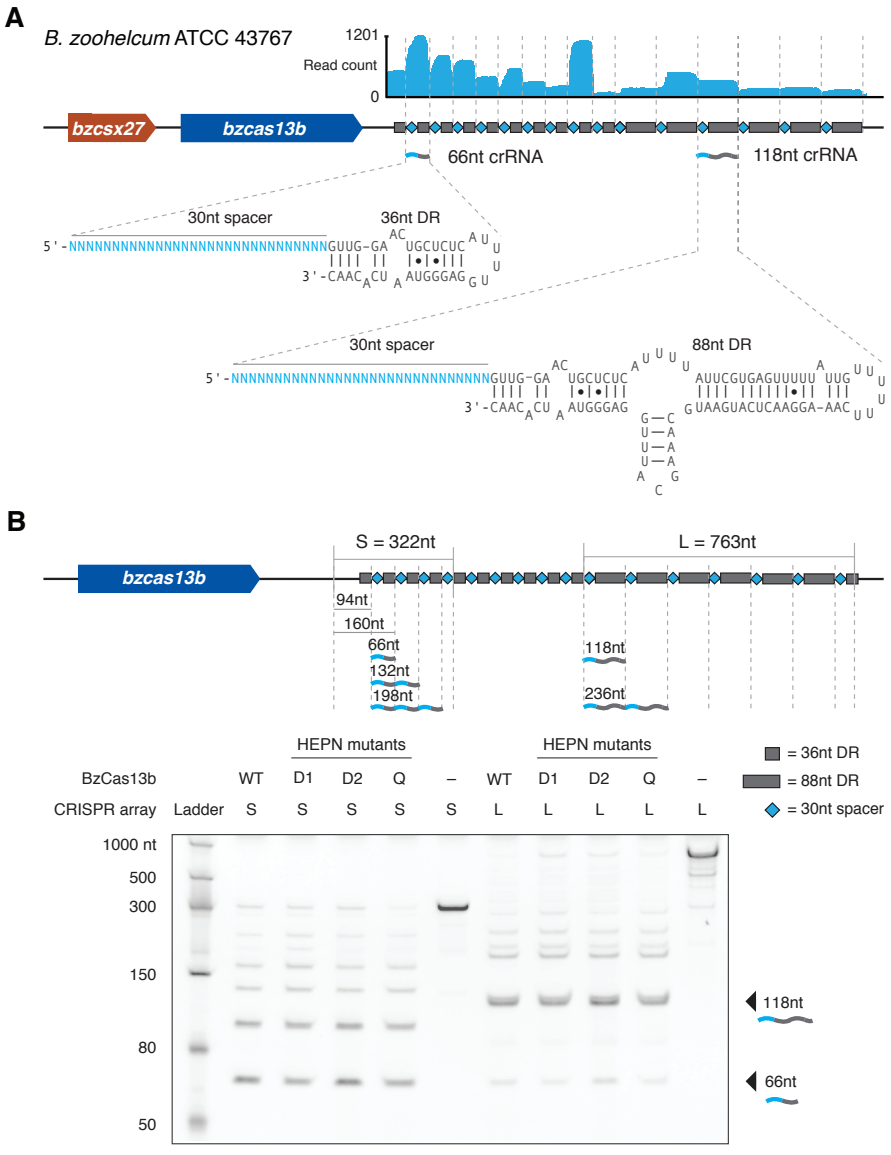


Figure 2-8: Cas13b from the VI-B1 Locus Processes a CRISPR Array with Two Direct Repeat Variants. (A) RNA sequencing of the native VI-B1 locus from *Bergeyella zoohelcum* ATCC 43767. **(B)** Denaturing gel showing cleavage products of in vitro-synthesized short-DR-containing or long-DR-containing CRISPR arrays from the *B. zoohelcum* genome by either wild-type or HEPN mutant BzCas13b (D1, R116A/H121A; D2, R1177A/H1182A; and Q, R116A/H121A/R1177A/H1182A). The schematic shows fragment lengths of a cleaved CRISPR array. See also Figure 2-9 and Table S1.

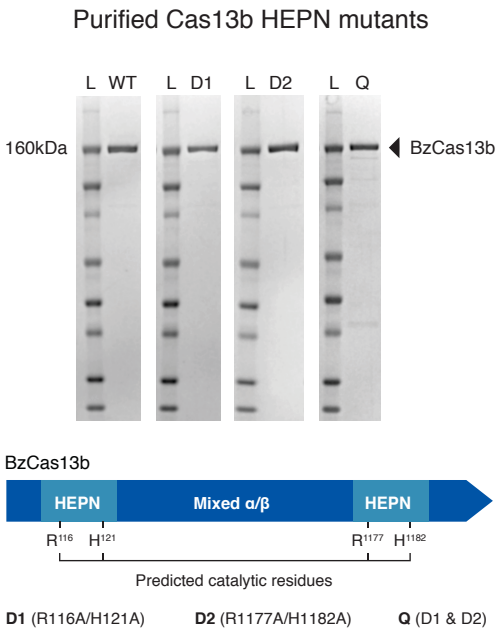


Figure 2-9: Protein gels of purified WT BzCas13b and three mutant BzCas13b proteins. Related to Figures 2-8, 3-4, and 3-8. Denaturing protein gels of *B. zoohelcum* wildtype, D1 (R116A/H121A mutant), D2 (R1177A/H1182A) mutant, and Q (R116A/H121A/R1177A/H1182A) mutant Cas13b.

Other CRISPR class 2 effectors are known to process their arrays without involvement of additional RNases (East-Seletsky et al., 2016; Zetsche et al., 2015). Similarly, we find that purified BzCas13b is capable of cleaving its associated CRISPR array, generating mature crRNAs with short or long direct repeats, and spacers which are not further processed beyond 30 nt, an activity which is not affected by mutation of the predicted catalytic residues of the HEPN domain (Figures 2-8B and 2-9; Table S1 and Table S2). Both genetically and biochemically we experimentally validated a computational prediction of functional CRISPR systems, namely that Cas13b processes its associated CRISPR array.

2.8 Discussion

The biocomputational pipeline described in this chapter led to the discovery of the first putative class 2 CRISPR system lacking the canonical adaptive machinery of Cas1 and Cas2. Given its success, this computational approach could be generalized to discover other interesting–CRISPR or non-CRISPR–systems in genomes. A more abstract algorithm might look something like this (with my specific implementation in parentheses):

1. Collect genomes of species of interest (Ensembl Release 27 compiled prokaryotic genomes);

2. For each genome, run a sub-algorithm (PILER-CR) to detect landmarks (CRISPR arrays) of systems of interest (CRISPR-Cas systems);

3. For each landmark in each genome, reconstruct a locus from proximal (10kb upstream and downstream of CRISPR arrays) annotated genes and transcripts (genes only);

4. Filter all discovered loci for constraining features (putative class 2 CRISPR

systems lacking Cas1 and Cas2) of interesting systems;

5. Expand loci by searching for all features homologous to constraining features (NCBI BLAST of class 2 candidate effectors), then reconstructing any additional loci as in Steps 1-3;

6. Hierarchically cluster (nearest-neighbor NCBI BLAST E-value) loci by constraining features, and manually classify putative novel systems from visual output (based on existing CRISPR classifications, and observing the presence or absence of noteworthy conserved genes, transcripts, sequences, and domains within the systems).

Such an algorithm could be applied to a host of biocomputational problems, for instance discovering novel immune systems in prokaryotes, or searching for transposable elements or endogenous retroviruses in eukaryotes, or tracking variable genomic, epigenomic, and/or transcriptomic sequences in cancer, immunology, neuroscience, or development. With the growing amount of data in biology, and with the imperfection of any purely artificial intelligence algorithms, such 'landmark-proximal feature-clustering' approaches may be necessary for certain future projects in the field. Indeed, this algorithm proved instrumental in identifying Cas13b.

After the initial biocomputational discovery, the in vivo and in vitro CRISPR array processing assays were essential go-no go experiments to any subsequent genetic and biochemical characterization of Cas13b. Putative CRISPR systems without the ability to process their associated CRISPR arrays cannot form mature crRNAs in complex, and thus cannot defend against invading bacteriophages. Had these experiments been unsuccessful, we likely would have attempted the same with other Cas13b orthologs. In this manner, we could have tested whether the CRISPR inactivity were due to the evolutionary degeneration of a few orthologs or whether this inactivity generalized to the entire VI-B system. If processing occurred in vivo but not in vitro,

we would have performed the equivalent biochemical experiments either with other purified CRISPR proteins (for example, Cas1, Cas2, and Csx27/Csx28) or bacterial lysates of other CRISPR proteins, with the appropriate controls.

Thankfully these initial CRISPR array processing experiments were successful, prompting us to investigate the nature of Cas13b functionality. At this point in the research, many exciting questions remained. For instance, does Cas13b indeed target RNA? Are both VI-B direct repeat variants functional? And what are the roles of Csx27 and Csx28? These questions and more are answered in Chapter 3.

2.9 Methods

2.9.1 Experimental Model and Subject Details

E. coli *E. coli* was grown in LB at 37°C at 250 rpm overnight.

One Shot Stbl3 *E. coli* *E. coli* was grown in LB at 37°C at 250 rpm overnight.

B. zoohelcum *B. zoohelcum* ATCC 43767 was grown in ATCC medium 44 (Brain Heart Infusion broth) at 37°C at 250 rpm overnight.

One Shot BL21(DE3)pLysE Chemically Competent *E. coli* The BzCas13b expression construct (Table S2) was transformed into One Shot BL21(DE3)pLysE (Invitrogen) cells. 25 mL of 6hr growing culture were inoculated into 2 l of Terrific Broth 4 growth media (12 g/L tryptone, 24 g/L yeast extract, 9.4 g/L K_2HPO_4, 2.2 g/L KH_2PO_4, Sigma). Cells were then grown at 37°C to a cell density of 0.6 OD_{600}, and then SUMO-BzCas13b expression was induced by supplementing with IPTG to a final concentration of 500 mM. Induced culture was grown for 16-18 hr before harvesting cell paste, which was stored at −80°C until subsequent purification. For each BzCas13b mutant, 1 L of Terrific Broth was used to generate cell paste and

all other reagents were scaled down accordingly. Protein purification was performed using the same protocol as wild-type Cas13b. PbCas13b was cloned into the same pET based vector and purified using a similar protocol as BzCas13b with the following differences: cells were grown at 21°C for 18 hr.

2.9.2 Method Details

Computational Sequence Analysis From complete compiled Ensembl Release 27 genomes (Yates et al., 2016), CRISPR repeats were identified using PILER-CR (Edgar, 2007). Proteins within 10kb of identified CRISPR arrays were clustered into loci, with loci rejected if more than one protein of size 700 amino acids or larger or if either Cas1 or Cas2 were present. For candidate Class 2 effectors, only proteins in these remaining loci of size 900aa to 1800aa were selected. These candidate effectors were subjected to the BLASTP (Camacho et al., 2009) search against the NCBI non-redundant (NR) protein sequence database with an E-value cutoff of 1e-7. All discovered proteins were then grouped into putative families via a nearest-neighbor grouping with the same E-value cutoff. Only putative families with at least ten candidate effectors and more than 50% of candidate effectors within 10kb of CRISPR arrays were considered. HHpred (Remmert et al., 2011) and existing CRISPR locus classification rules (Makarova et al., 2015) were used to classify each family, leaving Cas13b as the only unclassified family. Additional Cas13b proteins in the family were found through a nearest-neighbor search of previously discovered Csx27/Csx28 against the NCBI non-redundant (NR) protein sequence database with an E-value cutoff of 1e-7, and then by searching in genomes within 1kb of any newly discovered Csx27/Csx28. Within this Cas13b family, truncated or suspected partially sequenced effectors were discarded, leaving 105 loci, and 81 with a unique protein accession number in the NCBI non-redundant (NR) protein sequence database. Multiple sequence alignments on these 81 proteins (as well as the accessory Csx27 and Csx28

proteins) were performed using BLOSUM62 (Henikoff and Henikoff, 1992) to identify the HEPN domains and to sort the loci into phylogenetic trees. Loci represented in the tree of 81 non-redundant proteins were selected first for annotated Csx27/Csx28 within 1kb of Cas13b, and next for annotated CRISPR array within 10kb of Cas13b. Vienna RNAfold (Lorenz et al., 2011) was used to predict the secondary structure of each direct repeat, whose transcriptional orientation was chosen as identical to that of Cas13b in its locus. CRISPRTarget (Biswas et al., 2013) was used to search the spacers in each locus against NCBI phage and plasmid genomes. Weblogos were generated for all unique direct repeats and protospacer flanking sequences (Crooks et al., 2004). TMHMM Server v. 2.0 (Möller et al., 2001) was used to predict the transmembrane helices in Csx27 and Csx28.

Bacterial RNA-Sequencing RNA was isolated and prepared for sequencing using a modification of a previously described protocol (Heidrich et al., 2015; Shmakov et al., 2015). RNA was isolated from 5 mL of stationary phase of bacterial cultures by resuspending pelleted cells in 1mL of TRIzol (ThermoFisher Scientific) and then homogenizing with 300 uL zirconia/silica beads (BioSpec Products) in a BeadBeater (BioSpec Products) for 7 1 min cycles. 200 uL of chloroform was added to the homogenized sample and then samples were centrifuged for 15 min. (12000xg, 4°C). The aqueous phase was then used for input into the Direct-Zol RNA miniprep kit (Zymo). Purified RNA was DNase treated with TURBO DNase (Life Technologies) and 3′ dephosphorylated/5′ phosphorylated with T4 Polynucleotide Kinase (New England Biolabs). rRNA was eliminated using the bacterial Ribo-Zero rRNA removal kit (Illumina). Next, RNA was treated with RNA 5′ polyphosphatase (Epicenter Bio) to convert 5′-triphosphates to 5′-monophosphates for adaptor ligation. Samples were then polyA tailed with *E. coli* Poly(A) polymerase (New England Biolabs), and a 5′ RNA Illumina sequencing adaptor ligated to cellular RNA using T4 RNA Ligase 1 (ssRNA

ligase) (New England Biolabs). RNA was reverse transcribed using AffinityScript cDNA synthesis kit (Agilent Technologies) and an oligo-dT primer. cDNA was amplified with Herculase II polymerase (Agilent Technologies) and barcoded primers. The prepared cDNA libraries were sequenced on a MiSeq (Illumina).

For RNA sequencing of native *B. zoohelcum* ATCC 43767, we repeated the experiment with a modified protocol, omitting RNA 5′ polyphosphatase prior to 5′ adaptor ligation, to promote enrichment of processed transcripts originating from the CRISPR array. For heterologous *P. buccae* ATCC 33574 RNA sequencing in *E. coli*, we cloned the locus into pACYC184 (Table S1). Reads from each sample were identified on the basis of their associated barcode and aligned to the appropriate RefSeq reference genome using BWA (Li and Durbin, 2009). Paired-end alignments were used to extract entire transcript sequences using Galaxy (`https://usegalaxy.org`), and these sequences were analyzed using Geneious 8.1.8.

Nucleic Acid Preparation For in vitro synthesis of RNA, a T7 DNA fragment must be generated. To create T7 DNA fragments for crRNAs, top and bottom strand DNA oligos were synthesized by IDT. The top DNA oligo consisted of the T7 promoter, followed by the bases GGG to promote transcription, the 30 nt target and then direct repeat. Oligos were annealed together using annealing buffer (30 mM HEPES pH 7.4, 100 mM potassium acetate, and 2 mM magnesium acetate). Annealing was performed by incubating the mixture for 1 min at 95°C followed by a −1°C/minute ramp down to 23°C. To create ssRNA targets, short targets (Trunc2, 3, 4) were synthesized as top and bottom strand oligos containing the T7 promoter. For long ssRNA targets (E1, E2, S and L CRISPR Arrays), DNA primers (Table S1) with a T7 handle on the forward primer were ordered and the DNA fragment was amplified using PCR. T7 DNA constructs for RNA generation without body labeling were incubated with T7 polymerase overnight (10-14 hr) at 30°C using the HiScribe

T7 Quick High Yield RNA Synthesis kit (New England Biolabs). Body-labeled constructs were incubated with Cyanine 5-UTP (Perkin Elmer) and incubated with T7 polymerase overnight at 30°C using the HiScribe T7 High Yield RNA Synthesis kit (New England Biolabs). For a complete list of crRNAs and target ssRNAs used in this study see Table S1. 5′ end labeling was accomplished using the 5′ oligonucleotide kit (VectorLabs) and with a maleimide-IR800 probe (LI-COR Biosciences). 3′ end labeling was performed using a 3′ oligonucleotide labeling kit (Roche) and Cyanine 5-ddUTP (Perkin Elmer). RNAs were purified using RNA Clean and Concentrator columns^{TM}-5 (Zymo Research). Body-labeled dsRNA substrates were prepared by T7 DNA fragments for the bottom and top RNA strand. After synthesis, 1.3-fold excess of non-labeled bottom strand ssRNA was added and re-annealed to ensure the top strand would be annealed to a bottom strand by incubating the mixture for 1 min at 95°C followed by a −1°C/minute ramp down to 23°C.

BzCas13b Protein Purification The mammalian codon-optimized gene for Cas13b (*B. zoohelcum*) was synthesized (GenScript) and inserted into a bacterial expression vector (6x His/Twin Strep SUMO, a pET based vector received as a gift from Ilya Finkelstein) after cleaving the plasmid with the BamHI and NotI restriction enzymes and cloning in the gene using Gibson Assembly Master Mix (New England Biolabs). The BzCas13b expression construct (Table S2) was transformed into One Shot BL21(DE3)pLysE (Invitrogen) cells. 25 mL of 6hr growing culture were inoculated into 2 l of Terrific Broth 4 growth media (12 g/L tryptone, 24 g/L yeast extract, 9.4 g/L K_2HPO_4, 2.2 g/L KH_2PO_4, Sigma). Cells were then grown at 37°C to a cell density of 0.6 OD_{600}, and then SUMO-BzCas13b expression was induced by supplementing with IPTG to a final concentration of 500 uM. Induced culture was grown for 16-18 hr before harvesting cell paste, which was stored at −80°C until subsequent purification. Frozen cell paste was crushed and resuspended via stirring at 4°C in

500 mL of Lysis Buffer (50 mM NaH2PO4 pH 7.8, 400 mM NaCl) supplemented with protease inhibitors (cOmplete, EDTA-free, Roche Diagnostics Corporation) and 1250 U of benzonase (Invitrogen). The resuspended cell paste was lysed by a LM20 microfluidizer at 18,000 psi (Microfluidics). Lysate was cleared by centrifugation at 10,000 g for 1 hr. Filtered lysate was incubated with StrepTactin Sepharose High Performance (GE Healthcare Life Sciences) at 4°C for 1 hr with gentle agitation, and then applied to an Econo-column chromatography column (Bio-Rad Laboratories). Resin was washed with Lysis Buffer for 10 column volumes. One column volume of fresh Lysis Buffer was added to the column and mixed with 10 units of SUMO protease (Invitrogen) and incubated overnight. The eluate was removed from the column, SUMO cleavage was confirmed by SDS-PAGE and BlueFast protein staining (Eton Bioscience), and the sample was concentrated via Centrifugal Filter Unit to 2 mL. Concentrated sample was loaded onto a HiTrap Heparin HP column (GE Healthcare Life Sciences) via FPLC (AKTA Pure, GE Healthcare Life Sciences) and eluted over a gradient with an elution buffer with salt concentration of 1.2 M. The resulting fractions were tested for presence of BzCas13b protein by SDS-PAGE; fractions containing BzCas13b were pooled, and concentrated via Centrifugal Filter Unit to 1 mL. Concentrated sample was loaded a gel filtration column (HiLoad 16/600 Superdex 200, GE Healthcare Life Sciences) via FPLC (AKTA Pure, GE Healthcare Life Sciences) with buffer 500 mM NaCl, 50 mM Tris-HCl pH 7.5, 1 mM DTT.

BzCas13b HEPN Mutant Protein Purification Alanine mutants (Table S2) at each of the HEPN catalytic residues were generated using the Q5 site-directed mutagenesis kit (New England Biolabs) and transformed into One Shot BL21(DE3)pLysE cells (Invitrogen). For each mutant, 1 L of Terrific Broth was used to generate cell paste and all other reagents were scaled down accordingly. Protein purification was performed using the same protocol as wild-type Cas13b.

Nuclease Assay Nuclease assays were performed with equimolar amounts of end-labeled or body-labeled ssRNA target, purified protein, and crRNA, for targeted ssRNA cleavage. For CRISPR array cleavage, protein was supplied in a four times molar excess of the CRISPR array. Reactions were incubated in nuclease assay buffer (10 mM TrisHCl pH 7.5, 50 mM NaCl, 0.5 mM $MgCl_2$, 20 U SUPERase Inhibitor (ThermoFisher Scientific), 0.1% BSA). Reactions were allowed to proceed at 37°C for times specified in the figure legends. After incubation, samples were then quenched with 0.8U of Proteinase K (New England Biolabs) for 15 min at 25°C. The reactions were mixed with equal parts of RNA loading dye (New England Biolabs) and denatured at 95°C for 5 min and then cooled on ice for 2 min. Samples were analyzed by denaturing gel electrophoresis on 10% PAGE TBE-Urea (Invitrogen) run at 45°C. Gels were imaged using an Odyssey scanner (LI-COR Biosciences).

2.9.3 Data and Software Availability

Data Resources Data have been deposited in the following resources:

Next-Generation Sequencing for bacterial RNA-sequencing: `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA358111`

Chapter 3

Biochemical and Genetic Characterization of Cas13b

This chapter is derived in part from the Cas13b study published in *Molecular Cell* (Smargon et al., 2017). Full citation is as follows:

Smargon, A.A.*, Cox, D.B.T.*, Pyzocha, N.K.*, Zheng, K., Slaymaker, I.M., Gootenberg, J.S., Abudayyeh, O.A., Essletzbichler, P., Shmakov, S., Makarova, K.S., Koonin, E.V. and Zhang, F. (2017). *Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28.* Mol. Cell 65, 618−630.e7. (* denotes co-first authors)

Biochemical Contributions: Early on in the project, N.K.P., F.Z., and I working together optimized the nucleic acid preparation and nuclease assay, and designed and implemented experiments. This included carrying out the first successful single spacer BzCas13b RNA cleavage experiment and following up with spacer tiling experiments.

Genetic Contributions: With feedback from F.Z., I designed, implemented, and analyzed the *E. coli* essential gene screen. D.B.T.C. and I working together designed and implemented the kanamycin validation screen experiment, which I analyzed. F.Z. and I working together designed the RFP-tagged protein fluorescent imaging experiments, which I implemented.

3.1 Summary

Here we report the characterization of a class 2 sub-type, VI-B, which was discovered through the computational approach described in Chapter 2, and demonstrate

that the VI-B effector, Cas13b, is an RNA-guided RNase. Through a combination of biochemical and genetic experiments, we show that Cas13b cleaves target RNA and exhibits collateral RNase activity. Using an *E. coli* essential gene screen, we demonstrate that Cas13b has a double-sided protospacer-flanking sequence and elucidate RNA secondary structure requirements for targeting. We also find that Csx27 represses, whereas Csx28 enhances, Cas13b-mediated RNA interference. Characterization of these CRISPR systems creates opportunities to develop tools to manipulate and monitor cellular transcripts.

3.2 Introduction

Functional class 2 CRISPR-Cas systems have been found to target both DNA and RNA (Koonin et al., 2017). In order to achieve this nucleic acid targeting, they rely on the complexing of their single effector with a programmable crRNA that includes a direct repeat and spacer with reverse complementarity to its target protospacer. In Chapter 2, through RNA sequencing we demonstrated that Cas13b processes its associated pre-crRNA into a 66 nt mature crRNA, with a 30 nt 5′ spacer followed by a 36 nt 3′ direct repeat.

After this proof-of-concept experiment, we set out to determine the ability of Cas13b to target nucleic acids in an RNA-programmable fashion. In order to achieve this goal, we designed and implemented a series of experiments, both genetic and biochemical. In addition to many of the experimental techniques employed in Cas13a studies (Abudayyeh et al., 2016; East-Seletsky et al., 2016), in the Cas13b study we developed a few new assays—most notably the *E. coli* essential gene screen.

3.3 An *E. coli* Essential Gene Screen Reveals Targeting Rules for Cas13b

To validate the expected interference activity of the VI-B system and to determine the targeting rules for the VI-B1 locus from *B. zoohelcum*, we developed an *E. coli* essential gene screen (Figure 3-1A). For this negative selection screen, we generated a library of 54,600 unique spacers tiled with single-nucleotide resolution over the coding region of 45 monocistronic essential genes (Baba et al., 2006; Gerdes et al., 2003), plus 60 nt into the 5′ and 3′ UTRs. We also included 1,100 randomly generated non-targeting spacers to establish baseline activity (Table S3 and Table S4). We then transformed this library with plasmids carrying *bzcas13b* (*cas13b* gene from *B. zoohelcum*) and *bzcsx27*, just *bzcas13b*, or a control empty vector. After quality-control filtering of all screened spacers, we found a statistically significant depletion of targeting spacers over non-targeting spacers, indicating that Cas13b, alone or with Csx27, can achieve nucleic acid interference (Figure 3-1B).

To assess the targeting rules for Cas13b, we established two spacer depletion levels: strongly depleted (top 1% of depleted spacers) and safely depleted (spacers depleted 5σ above the mean depletion of the filtered non-targeting spacers). From spacers passing the strongly depleted cutoff we derived sequence motifs qualitatively identifying a double-sided protospacer flanking sequence (PFS) (Figure 3-1C) (Crooks et al., 2004). Because each position in a sequence motif is assumed to be independent, we developed a more quantitative, base-dependent PFS score defined as the ratio of the number of safely depleted spacers to the number of all spacers with a given PFS, normalized across all PFS scores (Figure 3-1D).

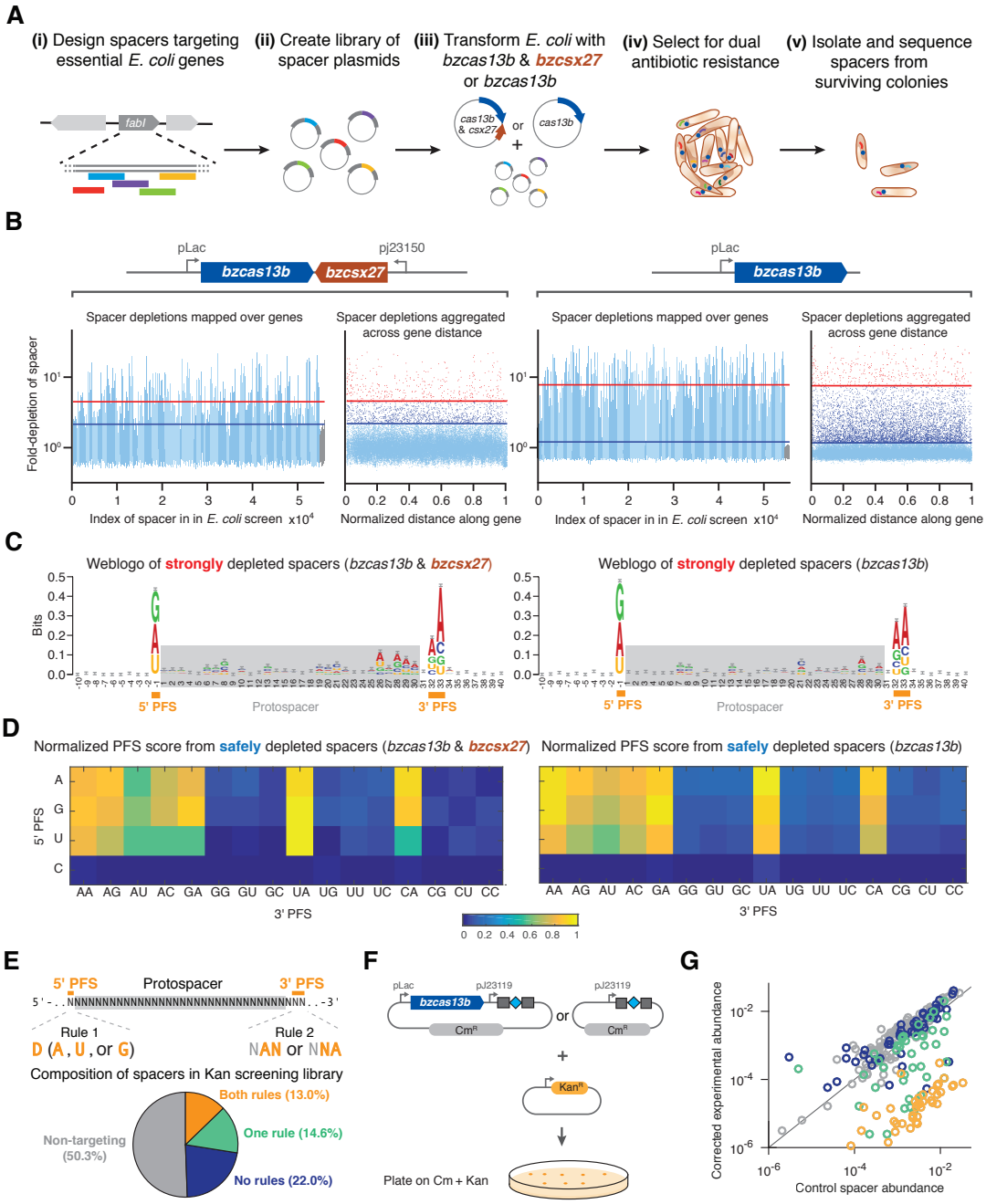

A
(i) Design spacers targeting essential E. coli genes
(ii) Create library of spacer plasmids
(iii) Transform E. coli with bzcas13b & bzcsx27 or bzcas13b
(iv) Select for dual antibiotic resistance
(v) Isolate and sequence spacers from surviving colonies
fabI
cas13b & csx27
or
cas13b
B
pLac
bzcas13b
bzcsx27
pj23150
pLac
bzcas13b
Spacer depletions mapped over genes
Spacer depletions aggregated across gene distance
Fold-depletion of spacer
Index of spacer in in E. coli screen x10^4
Normalized distance along gene
C
Weblogo of strongly depleted spacers (bzcas13b & bzcsx27)
Weblogo of strongly depleted spacers (bzcas13b)
Bits
5' PFS
Protospacer
3' PFS
D
Normalized PFS score from safely depleted spacers (bzcas13b & bzcsx27)
Normalized PFS score from safely depleted spacers (bzcas13b)
5' PFS
A
G
U
C
AA AG AU AC GA GG GU GC UA UG UU UC CA CG CU CC
3' PFS
E
5' PFS
Protospacer
3' PFS
5'-...NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN...-3'
Rule 1
D (A, U, or G)
Rule 2
NAN or NNA
Composition of spacers in Kan screening library
Both rules (13.0%)
One rule (14.6%)
No rules (22.0%)
Non-targeting (50.3%)
F
pLac
pJ23119
bzcas13b
Cm^R
or
pJ23119
Cm^R
+
Kan^R
Plate on Cm + Kan
G
Corrected experimental abundance
Control spacer abundance

Figure 3-1: Heterologous Expression of Cas13b Mediates Knockdown of *E. coli* Essential Genes by a Double-Sided PFS. (A) Design of *E. coli* essential gene screen to determine targeting rules of nucleic acid interference. **(B)** Manhattan plots of mean spacer depletions mapped over 45 genes and aggregated across normalized gene distance for either the full *B. zoohelcum* VI-B1 locus (left) or *cas13b* alone (right), with non-targeting spacers in gray, safely depleted spacers ($>5\sigma$ above mean depletion of non-targeting spacers) above blue line, and strongly depleted spacers (top 1% depleted) above red line. For the full locus, 36,142 targeting spacers and 630 non-targeting spacers passed QC filter. Of the targeting, 367 are strongly depleted, and 1,672 are safely depleted. For *cas13b* alone, 35,272 targeting spacers and 633 non-targeting spacers passed QC filter. Of the targeting, 359 are strongly depleted, and 6,374 are safely depleted. **(C)** Weblogo of sequence motifs of strongly depleted *B. zoohelcum* spacers. **(D)** Normalized PFS score matrix, where each score is the ratio of number of safely depleted *B. zoohelcum* spacers to total number of spacers for a given PFS, scaled so that maximum PFS score is 1. The 3′ PFS letters represent the RNA bases at the second and third 3′ PFS position. **(E)** Spacers targeting kanamycin to validate PFS targeting rules of 5′ PFS **(D)** and 3′ PFS (NAN or NNA). **(F)** Schematic of kanamycin validation screen for *B. zoohelcum* cas13b in *E. coli*. **(G)** Results from kanamycin validation screen; spacer abundances versus control for individual *B. zoohelcum* spacers, with abundances colored by type of spacer. See also Figure 3-2, Table S2, Table S3, Table S4, and Table S5.

The normalized PFS scores revealed a 5′ PFS of D (A, U, or G) and 3′ PFS of NAN or NNA, consistent for Cas13b with Csx27, as well as for Cas13b alone. To validate these sequence-targeting rules, we performed an orthogonal depletion screen with Cas13b alone, targeting the Kanamycin resistance gene (Figures 3-1E and 3-1F). Four classes of spacers were created: non-targeting, targeting with both 5′ and 3′ PFS rules, targeting with only the 5′ or 3′ PFS rule, and targeting with neither rule. Consistent with our findings from the *E. coli* essential gene screen, the combined 5′ and 3′ PFS spacers resulted in the highest Kanamycin sensitivity (Figures 3-1G and 3-2A; Table S5).

In addition to experimenting with Cas13b from *Bergeyella zoohelcum*, we also chose *Prevotella buccae* and *Porphyromonas gingivalis* due to our ability to order genomic DNA for each organism readily from ATCC. We ended up exclusively selecting *B. zoohelcum* and *P. buccae*, which conveniently corresponded to subtypes

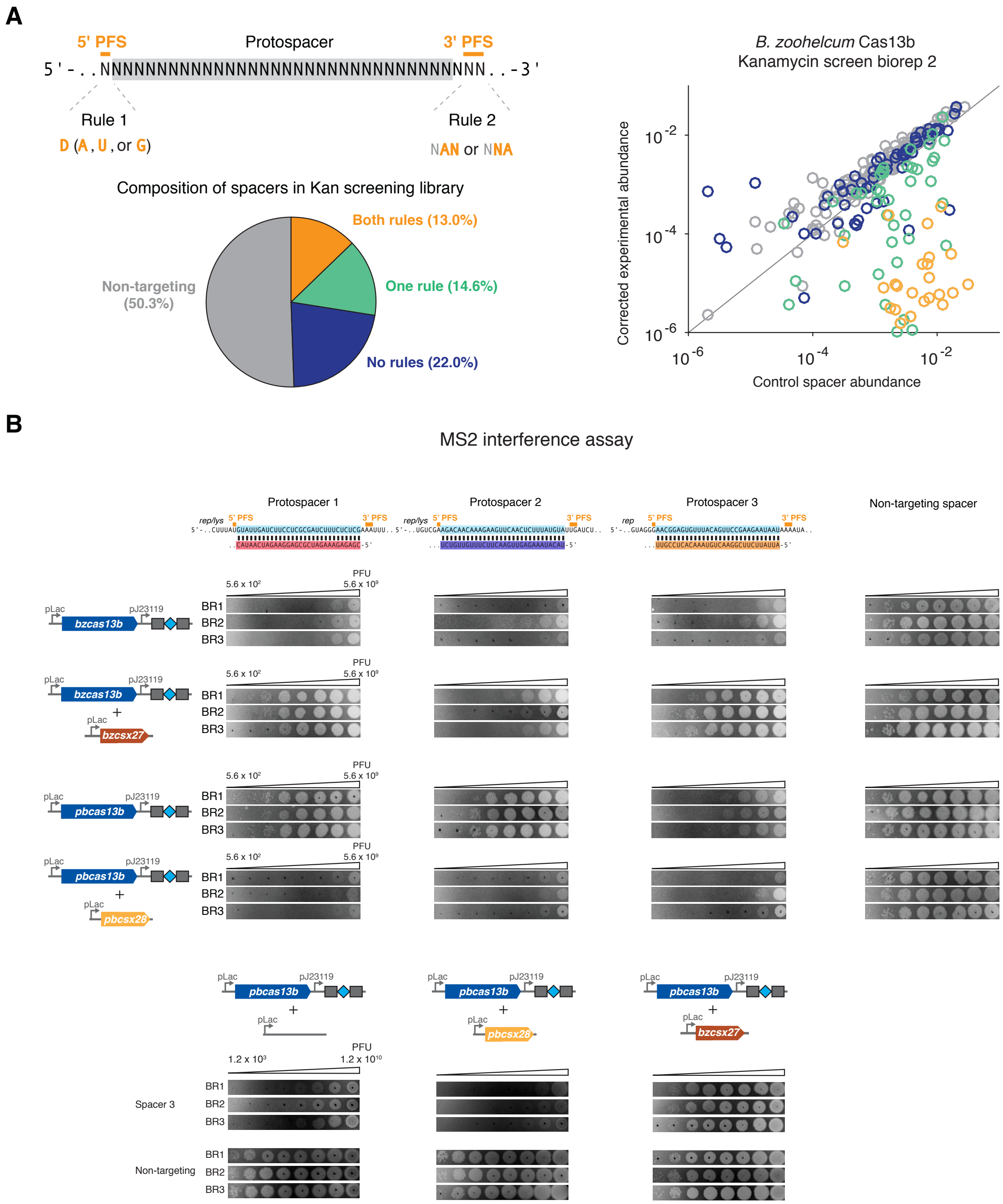


Figure 3-2: Targeting rule validation of BzCas13b and MS2 interference assay of BzCas13b and PbCas13b. Related to Figures 3-1, 3-8, and 3-11. **(A)** Spacers targeting kanamycin to validate PFS targeting rules of 5′PFS **(D)** and 3′ PFS (NAN or NNA) (left). Second kanamycin validation screen bioreplicate of spacer abundances versus control for individual *B. zoohelcum* spacers, with abundances colored by type of spacer (right). **(B)** Plaque drop assay with bioreplicates for *B. zoohelcum* VI-B1 locus and *cas13b*, for *P. buccae* VI-B2 locus and *cas13b*, and for *P. buccae cas13b* with pUC19, *B. zoohelcum csx27*, and *P. buccae csx28*.

VI-B1 and VI-B2, respectively. When the *E. coli* essential gene screen was performed with PgCas13b (Cas13b from *P. gingivalis*), targeting spacer depletions were indistinguishable from non-targeting spacer depletions (Figure 3-3). As was confirmed by additional genetic assays (unpublished data), PgCas13b has highly non-specific and potentially toxic activity; even in non-specific crRNA conditions, it interfered with any RNA targeted in our experiments. This corroborated result motivated us to focus on *B. zoohelcum* and *P. buccae*.

BzCas13b proved to be easier to purify from bacteria than PbCas13b, and so most Cas13b biochemical experiments were conducted with BzCas13b. We had also hoped to observe how Csx27 and Csx28 might impact Cas13b targeting of RNA. Neither BzCsx27 nor PbCsx28, however, yielded any substantial purification, perhaps due to toxicity in overexpression (unpublished data). Under these limiting circumstances, we explored the biochemical properties of Cas13b.

3.4 Cas13b Cleaves Single-Stranded RNA and Exhibits Collateral Activity In Vitro

Based on the presence of the computationally predicted HEPN domains that function as RNases in other CRISPR-Cas systems, including VI-A and some class 1 systems (Abudayyeh et al., 2016; Kim et al., 2013; Sheppard et al., 2016; Staals et al., 2014), we anticipated that Cas13b interferes with RNA. We confirmed this by demonstrating that purified Cas13b exclusively cleaves single-stranded RNA with both direct repeat architectures (Figures 3-4A and 3-5A). We then validated the PFS targeting rules biochemically, showing that a 5′ PFS of C greatly inhibits single-stranded RNA cleavage (Figure 3-4B), whereas a 3′ PFS of NAN or NNA enhances this activity (Figure 3-4C).

Other HEPN domain-containing CRISPR-Cas RNA-targeting systems, such as

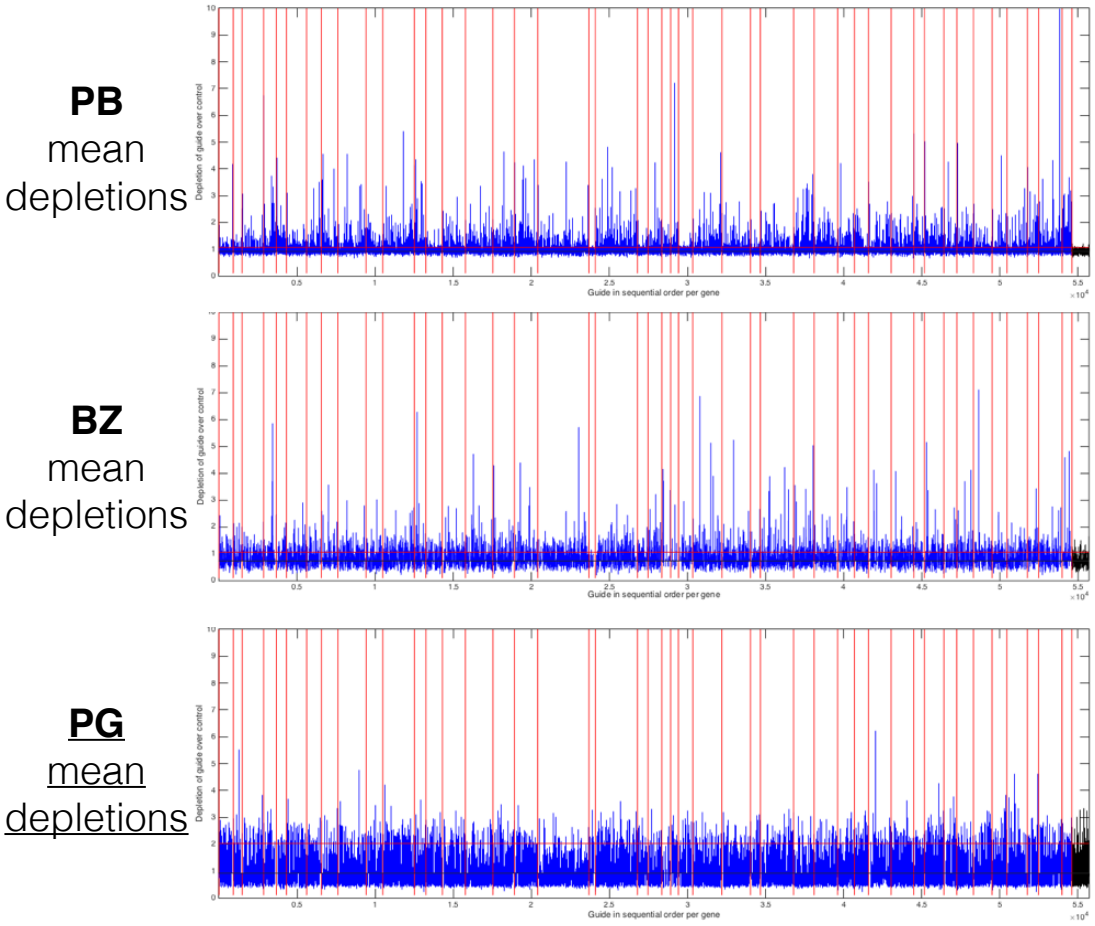


Figure 3-3: Raw *E. coli* essential gene screen data from three Cas13b orthologs. Screen was conducted with Cas13b from *Prevotella buccae ATCC 33574* (PB), *Bergeyella zoohelcum ATCC 43767* (BZ), and *Porphyromonas gingivalis ATCC 33277* (PG). Axes are the same as in Figure 3-1B (targeting spacers in blue, non-targeting spacers in black).

Csx1 from the type III-B CRISPR-Cas systems, preferentially cleave targets containing specific single-stranded nucleotides (Sheppard et al., 2016). To determine if Cas13b exhibits such a preference, we tested an RNA substrate with a variable homopolymer loop outside of the spacer:protospacer duplex region (Figure 3-4D). A heteropolymer loop consisting of alternating A then U was also tested (Figure 3-5B). We observed cleavage at pyrimidine residues, with a strong preference for uracil. This activity is abolished in the presence of EDTA (Figure 3-5C), suggesting a divalent metal ion-dependent mechanism for RNA cleavage akin to that of a similar HEPN-containing, class 2 effector protein, Cas13a (Abudayyeh et al., 2016; East-Seletsky et al., 2016).

Given that Cas13a has also been reported to cleave RNA non-specifically once activated by interaction with the target ("collateral effect") (Abudayyeh et al., 2016; East-Seletsky et al., 2016), we sought to test the ability of Cas13b to cleave a second, non-specific substrate following target cleavage. Using an in vitro assay similar to the one we previously used with Cas13a (Abudayyeh et al., 2016), we incubated Cas13b-crRNA complexes with both a target and non-target RNA substrate. We observed collateral cleavage of the non-targeted RNA, but only in the presence of the target RNA (Figure 3-4E).

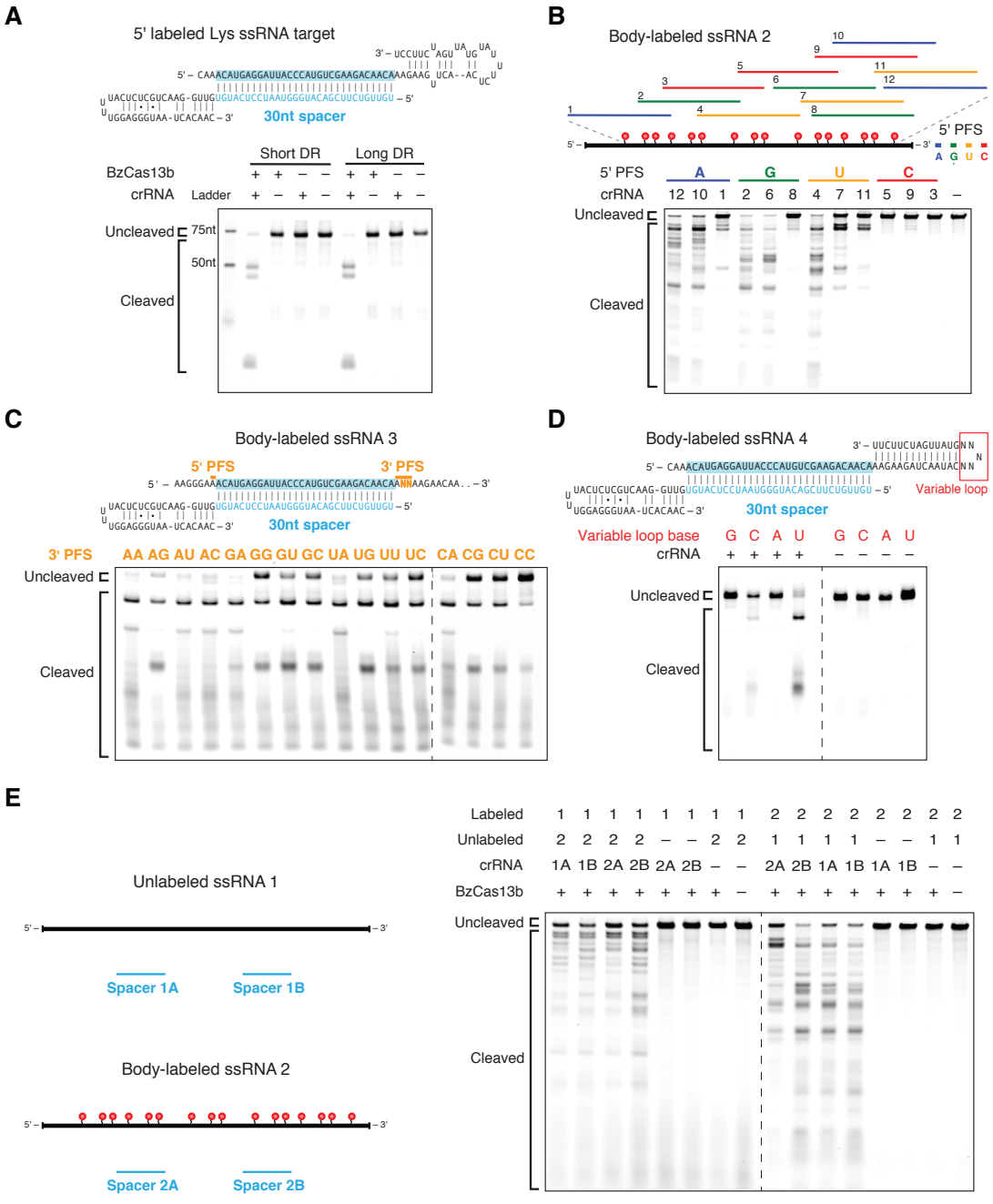

A
5' labeled Lys ssRNA target
30nt spacer
Short DR
Long DR
BzCas13b
crRNA
Ladder
Uncleaved
Cleaved
75nt
50nt
B
Body-labeled ssRNA 2
5' PFS
A G U C
12 10 1 2 6 8 4 7 11 5 9 3
C
Body-labeled ssRNA 3
5' PFS
3' PFS
30nt spacer
AA AG AU AC GA GG GU GC UA UG UU UC CA CG CU CC
D
Body-labeled ssRNA 4
Variable loop
30nt spacer
Variable loop base
G C A U G C A U
E
Unlabeled ssRNA 1
Spacer 1A
Spacer 1B
Body-labeled ssRNA 2
Spacer 2A
Spacer 2B
Labeled
Unlabeled
crRNA
BzCas13b
1A 1B 2A 2B 2A 2B – – 2A 2B 1A 1B 1A 1B – –

Figure 3-4: Cas13b Is a Programmable Single-Stranded RNase with Collateral Activity. **(A)** Schematic showing the RNA secondary structure of the cleavage target in complex with a targeting 30 nt spacer connected to short direct repeat (top). Denaturing gel demonstrating short direct repeat and long direct repeat crRNA-mediated ssRNA cleavage (bottom). Reactions were incubated for 10 min. The ssRNA target is 5′ labeled with IRDye 800. Three cleavage sites are observed. **(B)** Schematic showing three numbered protospacers for each colored 5′ PFS on a body-labeled ssRNA target (top); denaturing gel showing crRNA-guided ssRNA cleavage activity demonstrating the requirement for a D 5′ PFS (not C) (bottom). Reactions were incubated for 60 min. crRNAs correspond to protospacer numbered from the 5′ to the 3′ end of the target. Gel lane containing RNA ladder is not shown. **(C)** Schematic of a body-labeled ssRNA substrate being targeted by a crRNA (top). The protospacer region is highlighted in blue, and the orange bars indicate the 5′ PFS and 3′ PFS sequences. The orange letters represent the altered sequences in the experiment. Denaturing gel showing crRNA-guided ssRNA cleavage activity after 60 min of incubation, with the 5′ PFS tested as A and the 3′ PFS tested as ANN (bottom). The orange 3′ PFS letters represent the RNA bases at the second and third 3′ PFS position within each target ssRNA. Gel lane containing RNA ladder is not shown. Dashed line indicates two separate gels shown side by side. **(D)** Schematic showing the secondary structure of the body-labeled ssRNA targets used in the denaturing gel. The variable loop of the schematic (represented as N^5) is substituted with five monomers of the variable loop base in the gel (top). Denaturing gel showing cleavage bands of the homopolymer variable loop base (bottom). The targets were incubated for 30 min. Dashed line indicates seprate gel images (shown in Figure 3-12B). Gel lane containing RNA ladder is not shown. **(E)** Denaturing gel showing BzCas13b collateral cleavage activity after 30 min of incubation, with schematic of cleavage experiment to the right. Two crRNAs (A and B) target substrate 1 (1A and 1B) or substrate 2 (2A and 2B). Gel lane containing RNA ladder is not shown. Dashed line indicates two separate gels shown side by side. See also Figures 2-9 and 3-5 and Table S1.

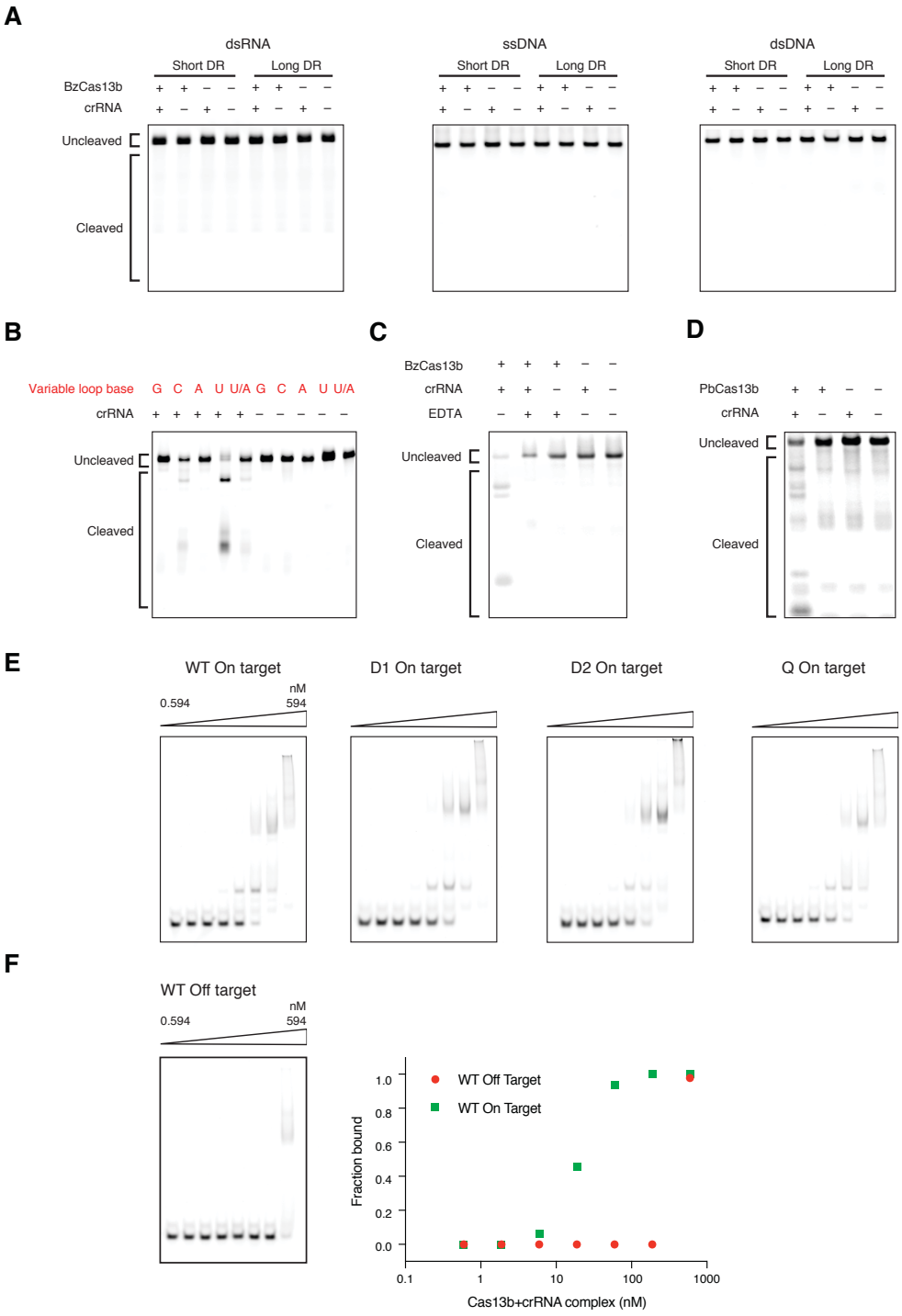

A
dsRNA
ssDNA
dsDNA
Short DR
Long DR
BzCas13b
crRNA
Uncleaved
Cleaved
B
Variable loop base
G C A U U/A G C A U U/A
crRNA
C
BzCas13b
crRNA
EDTA
D
PbCas13b
crRNA
E
WT On target
D1 On target
D2 On target
Q On target
nM
0.594
594
F
WT Off target
WT Off Target
WT On Target
Fraction bound
Cas13b+crRNA complex (nM)

Figure 3-5: Cas13b cleaves and binds to single-stranded RNA. Related to Figures 3-4 and 3-8. (A) Denaturing gels demonstrating no cleavage of dsRNA, ssDNA, or dsDNA by BzCas13b with either the short DR or long DR. Reactions were incubated for 10 minutes, the same amount of time which results in robust ssRNA cleavage for this target and crRNA pair. The ssDNA and top strand of the dsDNA target is 5′ labeled with IRDye 800. The dsRNA target is body labeled. Gel lane containing RNA ladder not shown. **(B)** Denaturing gel showing cleavage bands from the variable loop target as shown in Figure 3-4D. The U/A heteropolymer consists of the N5 variable loop of alternating U and A residues (5′ AUAUA 3′). **(C)** ssRNA cleavage requires BzCas13b and a targeting crRNA, and this cleavage activity is abolished by addition of EDTA. Gel lane containing RNA ladder not shown. **(D)** Denaturing gel showing PbCas13b cleavage activity of an ssRNA targeted substrate. The ssRNA is 5′ labeled with IRDye 800 and incubated for 30 minutes. Gel lane containing RNA ladder not shown. **(E)** EMSA gels that were used to quantify the K_D of the WT and mutant BzCas13b proteins, using an on-target crRNA complementary to the targeted ssRNA. **(F)** EMSA gel of WT BzCas13b with an off-target crRNA. The off-target crRNA is non-complementary to the targeted ssRNA.

Obtaining convincing Cas13b biochemical data took months of optimization. Initially our in vitro cleavage RNA gels looked inefficient, blurry, and wholly uninterpretable (unpublished data). Then, upon adding certain reagents to our chemical reactions during a combined pH-[NaCl] optimization, suddenly the gels came into focus (Figure 3-6). In this critical chemical reaction, we added DTT to prevent disulfide bond formation, BSA to prevent Cas13b binding to the test tube, and RNase inhibitor to reduce RNA degradation from contaminating RNases. While all reagents improved the quality of the gel, RNase inhibitor had the highest marginal benefit.

Before the *E. coli* essential gene screen had revealed the targeting rules of Cas13b, we attempted to deduce them biochemically. Due to the collateral effect, such biochemical screens were inconclusive upon sequencing (all protospacer flanking sequences cut with equal probability), and we were left with single spacer-variable protospacer RNA cleavage assays (Figure 3-6). To overcome this impasse somewhat, we performed randomized single spacer-single protospacer tiling cleavage assays (Figure 3-7A). Regrettably, data at such low sample size did not reveal any clear PFS. Upon later inspection, data from spacers that successfully cleaved did corroborate the

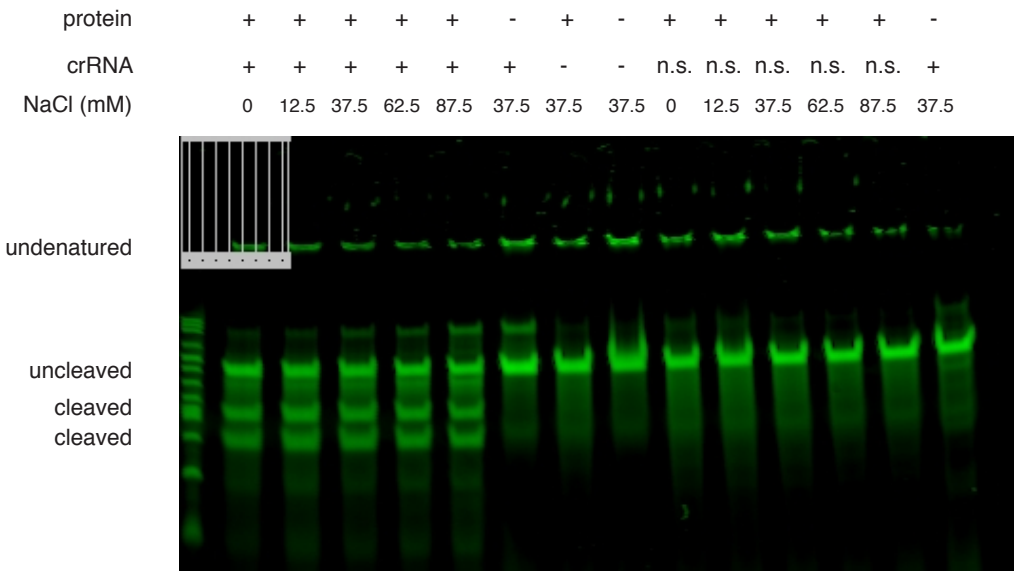


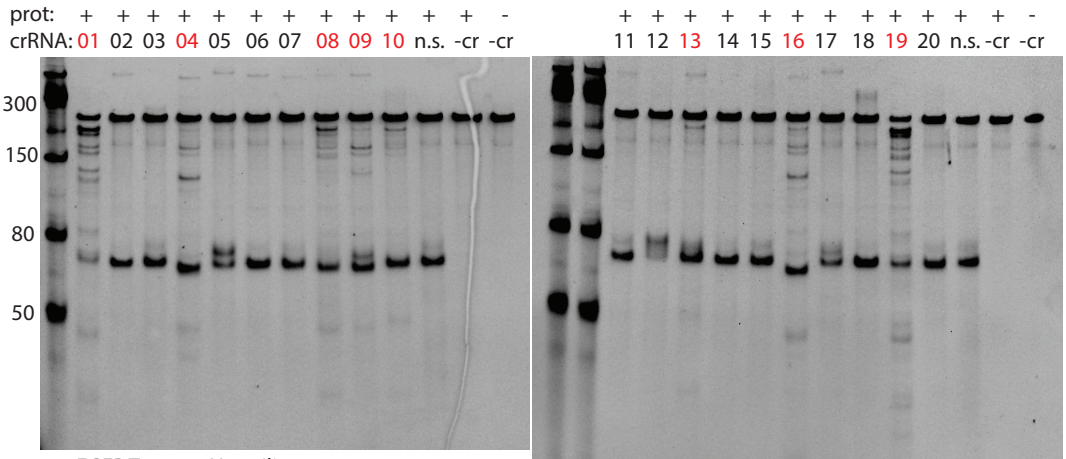
Figure 3-6: Earliest conclusive evidence of single spacer RNA cleavage by Cas13b. BzCas13b in vitro cleavage assay with single spacer against body-labeled target with a protospacer of N^7 complexity on the 5′ PFS. This optimization over various concentrations of NaCl was the first time DTT, BSA, and RNase inhibitor were used in biochemical reactions. Experimental conditions are as stated in the figure (n.s.: non-specific crRNA).

double-sided PFS as ascertained by the *E. coli* essential gene screen (Figure 3-7B).

3.5 Cas13b Shows Robust HEPN-Dependent Interference and Is Repressed by Csx27 Activity

To validate RNA interference in vivo, we assayed interference against the lytic, single-stranded RNA bacteriophage MS2, whose life cycle contains no DNA intermediates. We performed an MS2 drop plaque assay at serial dilutions of phage for both *bzcas13b* with *bzcsx27* and for *bzcas13b* alone with three spacers targeting the MS2 genome, two

A



EGFP Target 1 N^2 tiling:
100ng target,
1:1:1 molar ratio target: crRNA: protein
60 min incubation @37C, 50 mM NaCl, 1mM DTT, 10mM Tris-HCl pH 7.5, 0.1% BSA, 12.5% RNAse inhibitor

B

Guides that cut:

Guide Index	Left Flank	Target	Right Flank	Target Position
1	CCTGGGGCACAAGCTGGAGTACAACTACAA	CAGCCACAACGTCTATATCATGGCCGACAA	GCAGAAGAACGGCATCAAGGTGAACTTCAA	114
4	ATCGAGCTGAAGGGCATCGACTTCAAGGAG	GACGGCAACATCCTGGGGCACAAGCTGGAG	TACAACTACAACAGCCACAACGTCTATATC	73
8	AACTACAACAGCCACAACGTCTATATCATG	GCCGACAAGCAGAAGAACGGCATCAAGGTG	AACTTCAAGATCCGCCACAACATCGAGGAC	136
9	TGAAGGGCATCGACTTCAAGGAGGACGGCA	ACATCCTGGGGCACAAGCTGGAGTACAACT	ACAACAGCCACAACGTCTATATCATGGCCG	80
10	ACATCCTGGGGCACAAGCTGGAGTACAACT	ACAACAGCCACAACGTCTATATCATGGCCG	ACAAGCAGAAGAACGGCATCAAGGTGAACT	110
13	ACGTCTATATCATGGCCGACAAGCAGAAGA	ACGGCATCAAGGTGAACTTCAAGATCCGCC	ACAACATCGAGGACGGCAGCGTGCAGCTCG	152
16	GCATCGAGCTGAAGGGCATCGACTTCAAGG	AGGACGGCAACATCCTGGGGCACAAGCTGG	AGTACAACTACAACAGCCACAACGTCTATA	71
19	CAACTACAACAGCCACAACGTCTATATCAT	GGCCGACAAGCAGAAGAACGGCATCAAGGT	GAACTTCAAGATCCGCCACAACATCGAGGA	135

Guides that didn't visibly cut:

Guide Index	Left Flank	Target	Right Flank	Target Position
2	GTACAACTACAACAGCCACAACGTCTATAT	CATGGCCGACAAGCAGAAGAACGGCATCAA	GGTGAACTTCAAGATCCGCCACAACATCGA	132
3	GGCAACATCCTGGGGCACAAGCTGGAGTAC	AACTACAACAGCCACAACGTCTATATCATG	GCCGACAAGCAGAAGAACGGCATCAAGGTG	106
5	GGAGTACAACTACAACAGCCACAACGTCTA	TATCATGGCCGACAAGCAGAAGAACGGCAT	CAAGGTGAACTTCAAGATCCGCCACAACAT	129
6	CAAGCAGAAGAACGGCATCAAGGTGAACTT	CAAGATCCGCCACAACATCGAGGACGGCAG	CGTGCAGCTCGCCGACCACTACCAGCAGAA	171
7	AAGAACGGCATCAAGGTGAACTTCAAGATC	CGCCACAACATCGAGGACGGCAGCGTGCAG	CTCGCCGACCACTACCAGCAGAACACCCCC	178
11	TACAACAGCCACAACGTCTATATCATGGCC	GACAAGCAGAAGAACGGCATCAAGGTGAAC	TTCAAGATCCGCCACAACATCGAGGACGGC	139
12	AGCTGGAGTACAACTACAACAGCCACAACG	TCTATATCATGGCCGACAAGCAGAAGAACG	GCATCAAGGTGAACTTCAAGATCCGCCACA	125
14	ACGGCAACATCCTGGGGCACAAGCTGGAGT	ACAACTACAACAGCCACAACGTCTATATCA	TGGCCGACAAGCAGAAGAACGGCATCAAGG	104
15	TATATCATGGCCGACAAGCAGAAGAACGGC	ATCAAGGTGAACTTCAAGATCCGCCACAAC	ATCGAGGACGGCAGCGTGCAGCTCGCCGAC	157
17	TGGAGTACAACTACAACAGCCACAACGTCT	ATATCATGGCCGACAAGCAGAAGAACGGCA	TCAAGGTGAACTTCAAGATCCGCCACAACA	128
18	GCAACATCCTGGGGCACAAGCTGGAGTACA	ACTACAACAGCCACAACGTCTATATCATGG	CCGACAAGCAGAAGAACGGCATCAAGGTGA	107
20	TTCAAGGAGGACGGCAACATCCTGGGGCAC	AAGCTGGAGTACAACTACAACAGCCACAAC	GTCTATATCATGGCCGACAAGCAGAAGAAC	94

Figure 3-7: Earliest conclusive, comprehensive ($N > 2$) evidence of single spacer-single protospacer RNA cleavage by Cas13b. (A) BzCas13b in vitro cleavage assay with randomized spacer tiling along body-labeled target RNA from EGFP, which would form the basis of Figure 3-4B. Experimental conditions are as stated in the figure (n.s.: non-specific crRNA). **(B)** Corresponding sequences of protospacer ('target') and 5′ ('left') and 3′ ('right') flanking sequences, along with their indices and positions along the target RNA. In retrospect, these data (cutting vs. non-cutting by index) are compatible with the complex double-sided PFS of Cas13b.

at the *lys-rep* interface and one in *rep*, as well as one non-targeting spacer (Figure 3-8A). We observed substantial reduction in plaque formation for all targeting spacers compared to the non-targeting spacer, confirming sequence-specific RNA targeting by VI-B1 systems. (Figures 3-8A and 3-2B; Table S6). Notably, the presence of *bzcsx27* weakened RNA interference by *bzcas13b* for all three targeting spacers.

To confirm the lack of DNA interference in vivo, we adapted a previously established plasmid interference assay (Zetsche et al., 2015) with a protospacer placed either in-frame at the 5′ end of the *bla* ampicillin-resistance gene (transcribed target) or upstream of the *bla* gene promoter on the opposite strand (non-transcribed target). Bacteria co-transformed with *bzcas13b* and spacer as well as the non-transcribed target plasmid survived at a rate comparable to that of co-transformation of the same target with the empty vector on dual antibiotic selection. For bacteria co-transformed with the transcribed target, the colony-forming unit rate under dual antibiotic selection was reduced by approximately two orders of magnitude in the presence of *bzcas13b*, corroborating that Cas13b exclusively targets RNA in vivo (Figure 3-8B).

We next tested if predicted catalytic residues in the HEPN domains were responsible for RNA cleavage by Cas13b. Three HEPN mutants were generated by replacing the conserved catalytic arginines and histidines in the two HEPN domains with alanines (R116A/H121A, termed domain 1 [D1]; R1177A/H1182A, termed domain 2 [D2]; and R116A/H121A/R1177A/H1182A, termed quadruple [Q]) (Figure 2-9). All mutants lacked observable cleavage activity (Figure 3-8C), yet retained RNA binding capacity in vitro (Figures 3-8D and 3-5E). The wild-type and all three HEPN mutant Cas13b proteins showed comparable binding affinities for a single-stranded target RNA substrate, with K_D values ranging from 27 nM to 42 nM (Figures 3-8D and 3-5E; Table S7). The K_D for off-target binding was found to be greater than 188nM (Figure 3-5F).

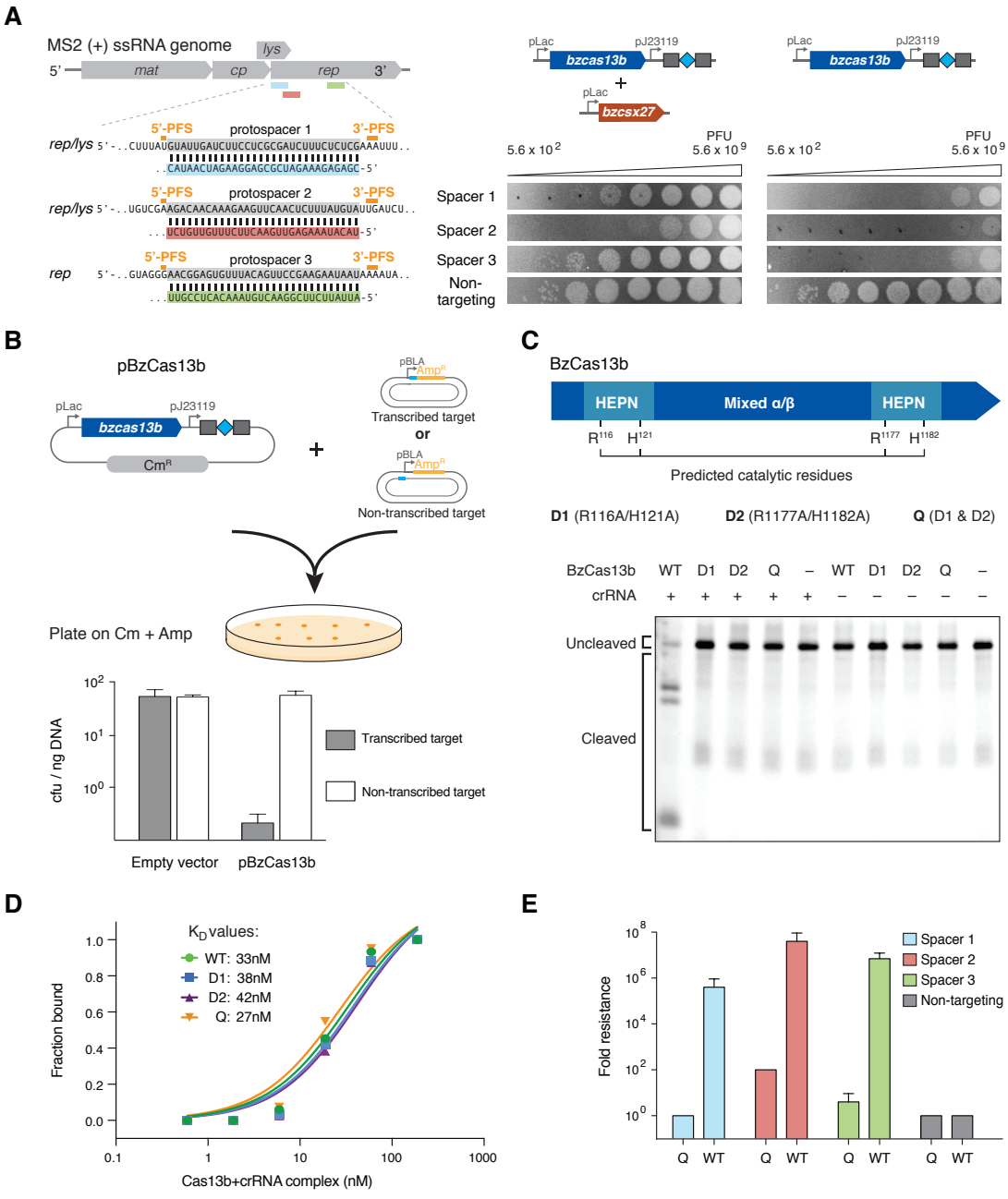

A
MS2 (+) ssRNA genome
lys
5'
mat
cp
rep
3'
rep/lys
protospacer 1
5'-PFS
3'-PFS
protospacer 2
protospacer 3
rep
pLac
pJ23119
bzcas13b
bzcsx27
PFU
Spacer 1
Spacer 2
Spacer 3
Non-targeting
B
pBzCas13b
Cm^R
pBLA
Amp^R
Transcribed target
or
Non-transcribed target
Plate on Cm + Amp
cfu / ng DNA
Empty vector
pBzCas13b
C
BzCas13b
HEPN
Mixed α/β
Predicted catalytic residues
D1 (R116A/H121A)
D2 (R1177A/H1182A)
Q (D1 & D2)
crRNA
Uncleaved
Cleaved
D
K_D values:
WT: 33nM
D1: 38nM
D2: 42nM
Q: 27nM
Fraction bound
Cas13b+crRNA complex (nM)
E
Fold resistance

Figure 3-8: HEPN Domains Mediate RNA Cleavage by Cas13b, the Activity of Which Is Repressed by Csx27. (A) Protospacer design for MS2 phage drop plaque assay to test RNA interference (left); drop plaque assay for full *B. zoohelcum* VI-B1 locus (center) and *bzcas13b* (right). **(B)** DNA interference assay schematic (top) and results (bottom). A target sequence is placed in-frame at the start of the transcribed bla gene that confers ampicillin resistance or in a non-transcribed region on the opposite strand of the same target plasmid. Target plasmids were co-transformed with *bzcas13b* plasmid or empty vectors conferring chloramphenicol resistance and plated on double selection antibiotic plates. **(C)** Schematic (top) and denaturing gel (bottom) showing ssRNA cleavage activity of WT and HEPN mutant BzCas13b. The protein and targeting crRNA complexes were incubated for 10 min. Gel lane containing RNA ladder is not shown. **(D)** Electrophoretic mobility shift assay (EMSA) graph showing the affinity of BzCas13b proteins and targeting crRNA complex to a 5′ end-labeled ssRNA. EMSA was performed with supplemental EDTA to reduce any cleavage activity. **(E)** Quantification of MS2 phage drop plaque assay with *B. zoohelcum* wild-type and Q (R116A/H121A/R1177A/H1182A) mutant Cas13b. See also Figures 2-9− 3-5, Table S1, Table S2, Table S6, and Table S7.

We confirmed the involvement of the HEPN domains in RNA interference in vivo, finding ∼5.5 orders of magnitude decrease in resistance to MS2 phage in the quadruple HEPN mutants versus wild-type Cas13b (Figures 3-8E and 3-2B). Interestingly, quadruple mutant Cas13b with spacers 2 and 3 still showed weak phage resistance, potentially due to catalytically inactive Cas13b binding to phage genomic RNA, leading to reduced phage replication.

3.6 Computational Modeling Predicts Additional Targeting Rules Governing Cas13b

Our sequence-based targeting results from the *E. coli* essential gene screen implied the existence of additional RNA-targeting rules beyond the PFS (only 18% of spacers were safely depleted for *bzcas13b*; from the PFS rules alone, the expected value would be 33%). Given that RNA targets contain a variety of secondary structures, we sought to determine how RNA accessibility impacts targeting. Using the Vienna

RNAplfold method (Bernhart et al., 2006), which has been successfully employed to predict RNAi efficiency (Tafer et al., 2008) (Figure 3-9A), we trained and tested an RNA accessibility model for spacer efficiency on our screen data and found that RNA accessibility matters the most in the protospacer region most distal to the direct repeat of the crRNA (Figures 3-9B and 3-9C).

Given the collateral activity observed in vitro, we examined our screen data for indications of non-specific RNA cleavage by Cas13b. To this end, we calculated the empirical cumulative distribution functions of safely depleted spacers aggregated across all essential genes from the 5′ UTR into the gene and from the 3′ UTR into the gene (Figure 3-9D). Because cleavage closer to the 5′ UTR is more likely to disrupt gene function, without non-specific RNase activity we would expect an overrepresentation of spacers in the 5′ UTR and an underrepresentation in the 3′ UTR. By contrast, in the presence of collateral activity a nearly uniform distribution would be expected. From our screen data, we observed a marginal underrepresentation of spacers in the 3′ UTR compared to a uniform distribution, suggesting that collateral activity may occur in vivo.

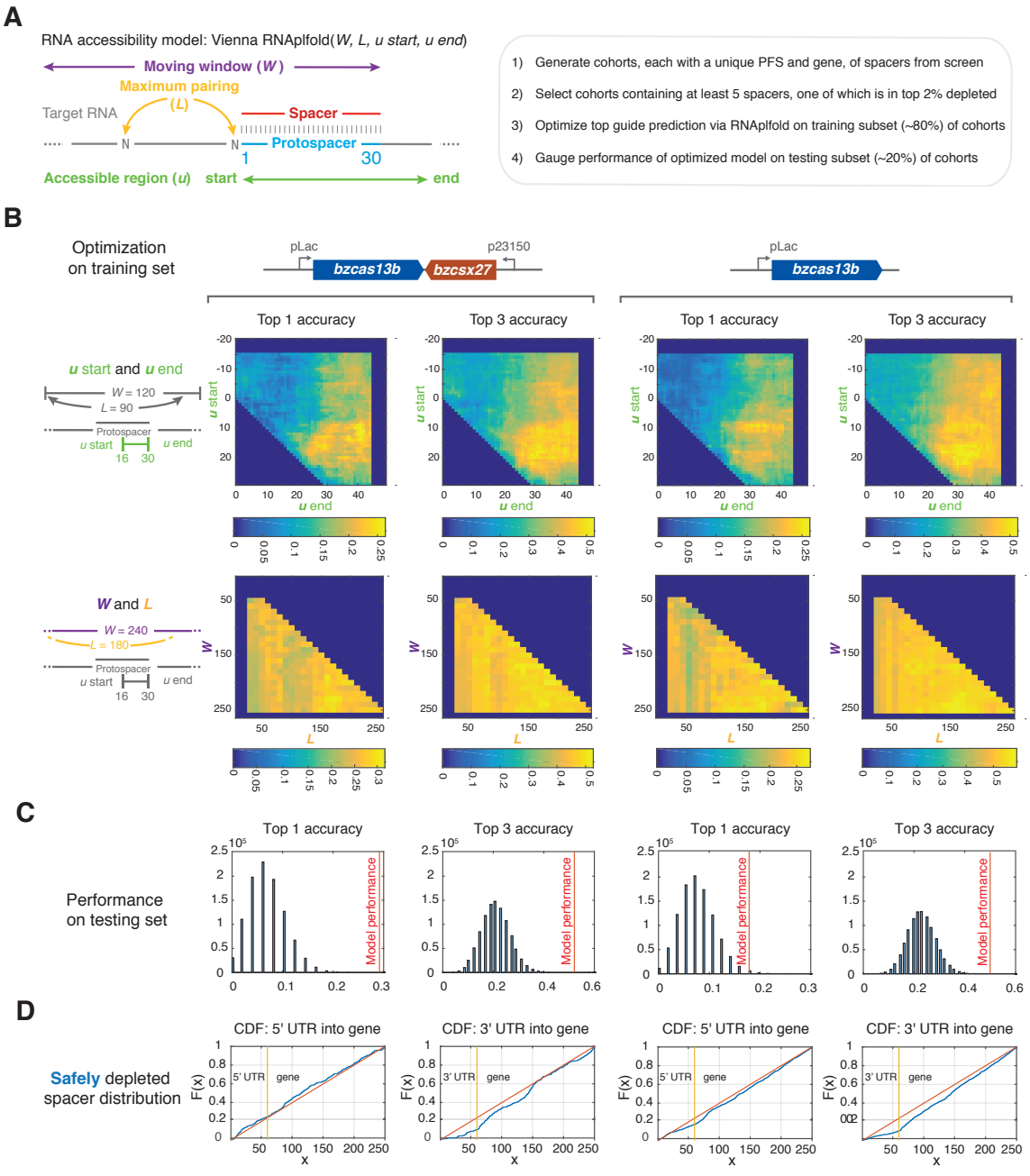

A
RNA accessibility model: Vienna RNAplfold(W, L, u start, u end)
Moving window (W)
Maximum pairing
(L)
Target RNA
Spacer
N
N
Protospacer
1
30
Accessible region (u)
start
end
1) Generate cohorts, each with a unique PFS and gene, of spacers from screen
2) Select cohorts containing at least 5 spacers, one of which is in top 2% depleted
3) Optimize top guide prediction via RNAplfold on training subset (~80%) of cohorts
4) Gauge performance of optimized model on testing subset (~20%) of cohorts
B
Optimization on training set
pLac
bzcas13b
bzcsx27
p23150
pLac
bzcas13b
Top 1 accuracy
Top 3 accuracy
Top 1 accuracy
Top 3 accuracy
u start and u end
W = 120
L = 90
Protospacer
u start
u end
16
30
u start
u end
W and L
W = 240
L = 180
Protospacer
u start
u end
16
30
W
L
C
Performance on testing set
Top 1 accuracy
Top 3 accuracy
Top 1 accuracy
Top 3 accuracy
Model performance
D
Safely depleted spacer distribution
CDF: 5' UTR into gene
CDF: 3' UTR into gene
CDF: 5' UTR into gene
CDF: 3' UTR into gene
5' UTR
gene
3' UTR
gene
F(x)
x

Figure 3-9: Efficient RNA Targeting by Cas13b Is Correlated with Local RNA Accessibility. (A) Methodology of secondary structure-mediated spacer efficiency analysis of *E. coli* essential gene screen data with Vienna RNAplfold. **(B)** Optimization of *top 1 accuracy* (computationally predicted most accessible spacer matches the top experimentally depleted spacer) and *top 3 accuracy* (computationally predicted top spacer falls in top three experimentally depleted spacers) on randomly selected *B. zoohelcum* training dataset using RNAplfold, first with *u start* and *u end*, and then with *W* and *L*. **(C)** Performance of optimized RNAplfold model on randomly selected *B. zoohelcum* testing dataset (48 cohorts for full *B. zoohelcum* VI-B1 locus, 56 cohorts for *bzcas13b*) against 10^6 Monte Carlo simulations: empirical p values from left to right of 3e-6, 1e-6, 8.7e-3, and 6e-6. **(D)** Empirical cumulative distribution function of safely depleted *B. zoohelcum* spacers over all genes from 5′ UTR into gene and from 3′ UTR into gene. Yellow line separates UTR and gene, red line is theoretical cumulative distribution function of uniformly distributed spacers, and blue line is empirical cumulative distribution of safely depleted *B. zoohelcum* spacers. See also Table S3 and Table S4.

3.7 CRISPR-Cas13b Effectors Are Differentially Regulated by Csx27 and Csx28

Having characterized Cas13b alone (VI-B), we sought to characterize the VI-B1 and V-B2 subtypes in the presence of Csx27 and Csx28, respectively. Both putative accessory proteins were predicted to contain one or more transmembrane segments (Figure 3-10A) (Möller et al., 2001). However, Csx27 of *Bergeyella zoohelcum* and Csx28 of *Prevotella buccae* tagged with RFP at either the N or the C terminus did not show membrane localization when expressed in *E. coli* (Figure 3-10B). In addition to the predicted hydrophobic domains, analysis of the multiple sequence alignment of Csx28 proteins indicated the presence of a divergent HEPN domain (Figure 2-6C). This property in particular led us to believe that Csx27 and/or Csx28 might affect the RNA-targeting capability of Cas13b.

To determine how the established RNA targeting rules generalize across the subtype VI-B systems from diverse bacteria, we characterized the subtype VI-B2

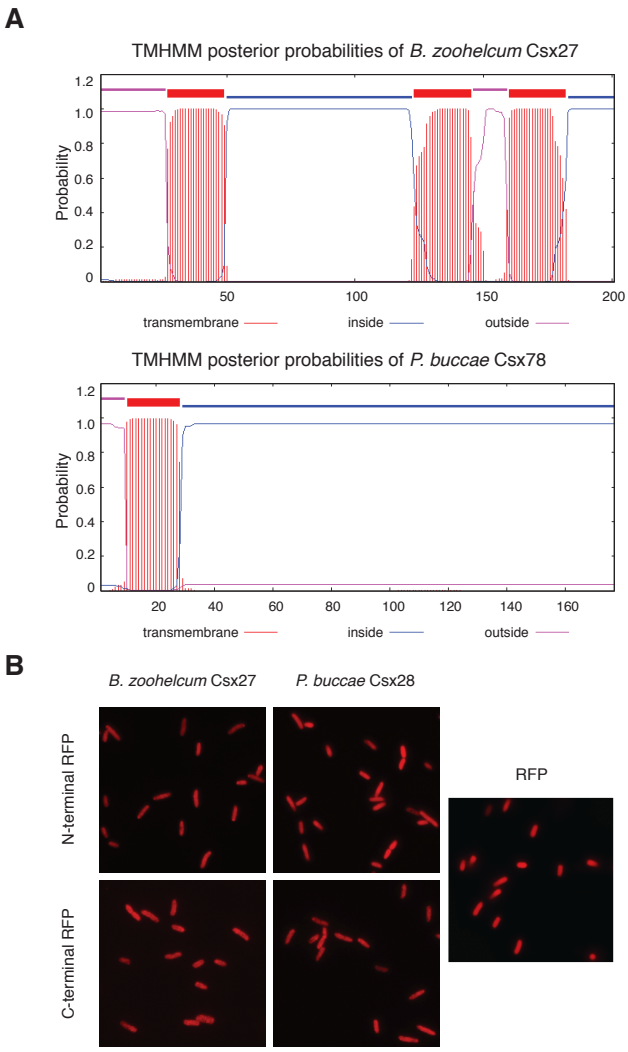


Figure 3-10: Predicted transmembrane domains of Csx27 and Csx28 not validated experimentally. Related to Figure 2-5. **(A)** Transmembrane domain prediction in Csx27 of *B. zoohelcum* and Csx28 of *P. buccae* using TMHMM v2. **(B)** N- and C-terminally fused RFP imaging of Csx27 of *B. zoohelcum* and Csx28 of *P. buccae*.

locus from *P. buccae*. RNA sequencing of the CRISPR array revealed processing effectively identical to that of *B. zoohelcum*, excluding the long crRNA (Figure 3-12A). The *E. coli* essential gene screen with *pbcas13b* and *pbcsx28* or *pbcas13* alone led to the identification of a PFS matrix similar to that of *B. zoohelcum*, with certain PFSs disfavored (Figures 3-11A, 3-12B, and 3-12C). Similar to BzCas13b, PbCas13b was found to cleave targeted single-stranded RNA in vitro (Figure 3-5D). As with *bzcsx27*, the presence of *pbcsx28* did not appreciably alter the PFS. We also repeated the secondary structure analysis with *pbcas13b*, and a comparable RNAplfold model applied (Figure 3-12D). Strikingly, in these experiments the safely depleted spacers for *pbcas13b* alone were highly biased to the beginning of the 5′ UTR of genes, suggestive of inhibited or more spatially localized RNase activity in the absence of *pbcsx28* (Figure 3-12E). We further explored the apparent reduced activity of *pbcas13b* alone relative to the respective full CRISPR-Cas locus using the MS2 phage drop plaque assay and found that *pbcsx28* enhances MS2 phage interference by up to four orders of magnitude (Figures 3-11B and 3-2B). The differential ability of *csx27* to repress and *csx28* to enhance *cas13b* activity generalizes across thousands of spacers in the *E. coli* essential gene screen (Figure 3-11C), highlighting the distinctive regulatory modes of the two variants of subtype VI-B CRISPR-Cas systems.

To investigate the ability of the small accessory proteins to modulate Cas13b activity further, we tested if Csx27 can also repress PbCas13b using the MS2 drop plaque assay. Cells co-transformed with *pbcas13b* and *bzcsx27* expression plasmids exhibited a 10^5-fold reduction in interference activity relative to *pbcas13b* expression plasmid and pUC19 empty vector, indicating that Csx27 exerts an inhibitory effect on PbCas13b (Figures 3-11D and 3-2B). The ability of Csx27 to modulate the interference activity of BzCas13b and PbCas13b suggests that it is a modular protein that can function across multiple VI-B loci.

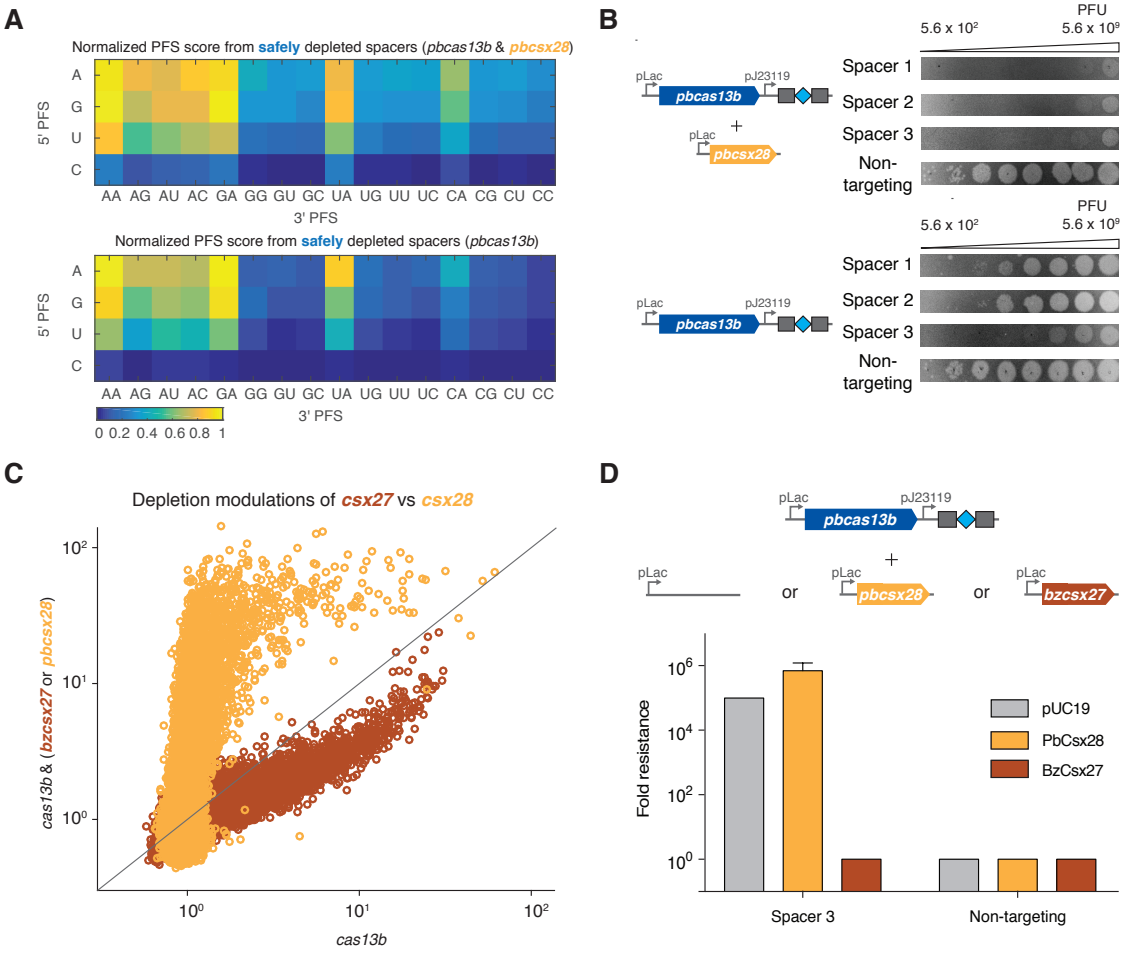


Figure 3-11: Class 2 Type VI-B Systems Are Differentially Regulated across Two Loci by Csx27 and Csx28. (A) Normalized PFS matrix for *P. buccae* VI-B2 locus (top) and *pbcas13b* (bottom). The 3′ PFS letters represent the RNA bases at the second and third 3′ PFS position. **(B)** MS2 drop plaque assay for full *P. buccae* VI-B2 locus (top) and *pbcas13b* (bottom). **(C)** Spacer depletions of *bzcas13b* with and without *bzcsx27* (brown), as compared to *pbcas13b* with and without *pbcsx28* (gold). **(D)** Fold resistance to MS2 infection for cells co-transformed with *pbcas13b* and the indicated *csx* expression plasmid. See also Figures 3-2 and 3-12, Table S2, Table S3, Table S4, and Table S6.

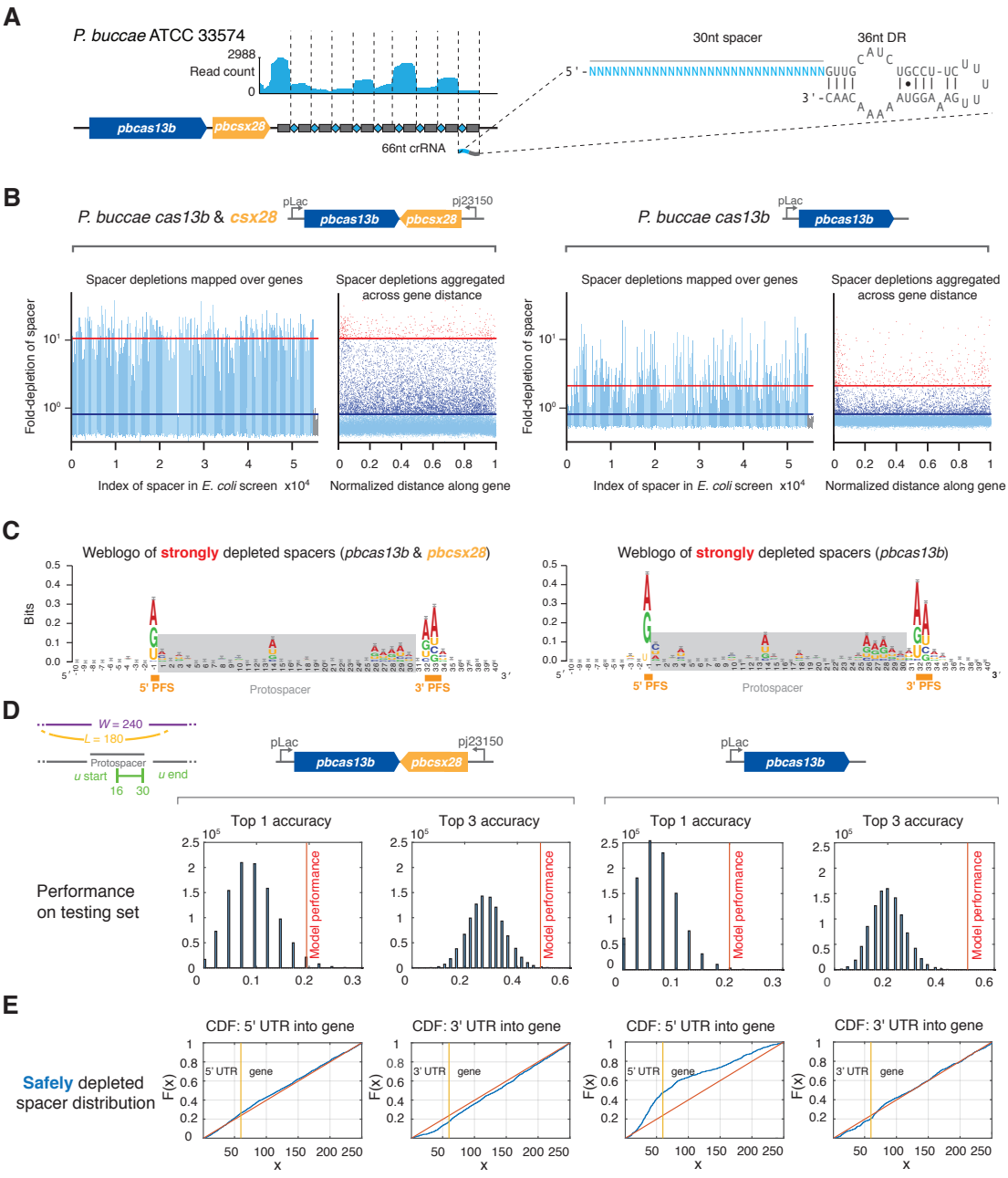

A
P. buccae ATCC 33574
2988
Read count
0
pbcas13b
pbcsx28
66nt crRNA
30nt spacer
36nt DR
5'-NNNNNNNNNNNNNNNNNNNNNNNNNNNNNGUUG
3'-CAAC
UGCCU-UC
AUGGA AG
B
P. buccae cas13b & csx28
pLac
pj23150
pbcas13b
pbcsx28
P. buccae cas13b
Spacer depletions mapped over genes
Spacer depletions aggregated across gene distance
Fold-depletion of spacer
Index of spacer in E. coli screen x10^4
Normalized distance along gene
C
Weblogo of strongly depleted spacers (pbcas13b & pbcsx28)
Weblogo of strongly depleted spacers (pbcas13b)
Bits
5' PFS
Protospacer
3' PFS
D
W = 240
L = 180
Protospacer
u start
u end
16
30
Performance on testing set
Top 1 accuracy
Top 3 accuracy
Model performance
E
Safely depleted spacer distribution
CDF: 5' UTR into gene
CDF: 3' UTR into gene
5' UTR
3' UTR
gene
F(x)
x

Figure 3-12: RNA-targeting of *P. buccae* VI-B2 CRISPR locus. Related to Figure 3-11. (A) RNA-Sequencing of heterologously expressed VI-B2 locus from *P. buccae* ATCC 33574 in *E. coli*. **(B)** Manhattan plots of spacer depletions mapped over 45 genes and aggregated across normalized gene distance for full *P. buccae* VI-B2 locus (left) and *cas13b* (right), with non-targeting spacers in gray, safely depleted ($>5\sigma$ above mean depletion of non-targeting spacers) spacers above blue line, and strongly depleted (top 1% depleted) spacers above red line. For the full locus, 36,141 targeting spacers and 859 non-targeting spacers passed QC filter. Of the targeting, 370 are strongly depleted and 8065 are safely depleted. For *cas13b* alone, 41,126 targeting spacers and 824 non-targeting spacers passed QC filter. Of the targeting, 419 are strongly depleted and 3295 are safely depleted. **(C)** Sequence weblogos of strongly depleted *P. buccae* spacers, revealing double-sided PFS (protospacer flanking sequence). **(D)** Performance of optimized RNAplfold model (W=240, L=180, *u start*=16, *u end*=30) on randomly selected *P. buccae* testing dataset (41 cohorts for full *P. buccae* VI-B2 locus, 40 cohorts for pbcas13b) against 10^6 Monte Carlo simulations: empirical P-values from left to right of 3.3e-2, 2.7e-3, 3.9e-3, 1.5e-5. **(E)** Empirical cumulative distribution function of safely depleted *P. buccae* spacers over all genes from 5′ UTR into gene and from 3′ UTR into gene. Yellow line separates UTR and gene, red line is theoretical cumulative distribution function of uniformly distributed spacers, and blue line is empirical cumulative distribution of safely depleted *P. buccae* spacers.

3.8 Discussion

Here we describe two RNA-targeting CRISPR class 2 systems of subtype VI-B (VI-B1 and VI-B2), containing the computationally discovered RNA-guided RNase Cas13b. Type VI-B systems show several notable similarities to the recently characterized VI-A system. The single protein effectors of both systems cleave single-stranded RNA via HEPN domains, process their CRISPR arrays independent of the HEPN domains, and exhibit collateral RNase activity. Cas13b proteins, however, show only limited sequence similarity to Cas13a, and the common ancestry of the two type VI subtypes remains uncertain. Furthermore, the type VI-B systems differ from VI-A in several other novel ways, including the absence of both *cas1* and *cas2*, which are involved in spacer acquisition in other CRISPR-Cas systems (Mohanraju et al., 2016). The VI-B CRISPR arrays contain multiple spacers that differ among closely related bacterial

strains, suggesting that acquisition does occur, either autonomously or possibly *in trans*, by recruiting Cas1 and Cas2 encoded in other CRISPR-Cas loci from the same genome. *In trans* utilization of adaptation modules of other CRISPR-Cas systems is compatible with the finding that the great majority of type VI-B systems co-occur in the same bacterial genome as other CRISPR-Cas loci that include cas1 and cas2 genes; conceivably, the three VI-B-carrying genomes that lack adaptation modules have lost them recently. Additionally, VI-B systems differ from VI-A systems by the presence of the small accessory proteins Csx27 (VI-B1 systems) and Csx28 (VI-B2 systems), which exert opposing regulatory effects on Cas13b activity. The near ubiquity and mutual exclusivity of these accessory proteins in CRISPR-Cas13b loci is notable and quite intriguing.

Repression of Cas13b by Csx27 in VI-B1 systems could be part of an important regulatory mechanism of phage interference. The ability of Csx27 to repress Cas13b activity may be a general property, as we found that it can also repress PbCas13b (subtype VI-B2). In the case of type VI-B2 systems, Csx28 might enhance the collateral activity of Cas13b to inactivate numerous transcripts of invading bacteriophages or to promote programmed cell death. Both Csx27 and Csx28 contain predicted long, hydrophobic α helices that might enable them to interact physically with Cas13b, but this remains to be determined. We did not find homologs of Csx27 or Csx28 encoded in any CRISPR-Cas loci other than type VI-B loci, suggesting that these proteins might function in tight association with Cas13b.

As with previously characterized class 2 CRISPR-Cas effectors, such as Cas9 and Cpf1, there is enormous potential to harness Cas13b for use as a molecular tool (Cong et al., 2013; Mali et al., 2013; Wright et al., 2016). A comprehensive understanding of the factors that affect target selection is essential to the success of any such tools, particularly those that target RNA, where secondary structure will likely impact activity. We therefore developed a novel *E. coli* essential gene screen

to explore the targeting rules of Cas13b more fully. This *E. coli* screen offers several advantages by increasing the number of guides testable in a single experiment to explore how diverse spacer and flanking sequences may affect Cas13b activity. This screen revealed a double-sided PFS in VI-B systems, which may give insight into Cas13b protein-RNA interactions and could help improve specificity by expanding sequence targeting constraints (Ran et al., 2015).

The characterization of Cas13b and other RNA-targeting CRISPR systems raises the prospect of a suite of precise and robust in vivo RNA manipulation tools for studying a wide range of biological processes (Abil and Zhao, 2015; Filipovska and Rackham, 2011; Mackay et al., 2011). The ability of Cas13b to process its own CRISPR array could be extended to multiplex transcriptome engineering. In addition, the VI-B functional long direct repeats could be altered to incorporate stem loops akin to the Cas9-SAM system (Konermann et al., 2015). Like Cas9 and Cpf1, Cas13a and Cas13b may be utilized for complementary applications in science and technology.

3.9 Methods

3.9.1 Experimental Model and Subject Details

E. coli *E. coli* was grown in LB at 37°C at 250 rpm overnight.

One Shot Stbl3 *E. coli* *E. coli* was grown in LB at 37°C at 250 rpm overnight.

NEB 10-Beta Competent *E. coli* (High Efficiency) NEB 10-beta Competent *E. coli* was transformed on LB agar at 37°C overnight.

MegaX DH10B T1R Electrocompetent Cells MegaX DH10B T1R Electrocompetent *E. coli* was transformed on LB agar at 37°C overnight.

3.9.2 Method Details

***E. coli* Essential Gene Screen Experiment** The intersection of two *E. coli* DH10B strain essential gene studies (Baba et al., 2006; Gerdes et al., 2003) was taken, and further pared down to 45 genes by only selecting genes exclusive to their respective operons (Table S3). Over these 45 genes 54,600 spacers were designed to tile at single resolution across the coding region, as well as to extend 60 nt into the 5′ UTR and 3′ UTR. In addition, 1100 non-targeting, pseudorandomly generated spacers with no precise match to the *E. coli* DH10B strain genome were added to the library as a non-targeting negative control. The library of spacers (Table S4) was cloned into a *B. zoohelcum* or *P. buccae* direct repeat-spacer-direct repeat backbone containing a chloramphenicol resistance gene using Golden Gate Assembly (NEB) with 100 cycles, and then transformed over five 22.7cm x 22.7cm chloramphenicol LB Agar plates. Libraries of transformants were scraped from plates and DNA was extracted using the Macherey-Nagel Nucleobond Xtra Midiprep Kit (Macherey-Nagel). 50 ng of library plasmid and equimolar gene plasmid containing an ampicillin resistance gene (*bzcas13b*, *bzcas13b* & *bzcsx27*, *pbcas13b*, *pbcas13b* & *pbcsx28*, empty vector pBR322) (Table S2) were transformed into MegaX DH10B T1R Electrocomp Cells (ThermoFisher) according to manufacturer's protocol, with four separate 22.7cm x 22.7cm carbenicillin-chloramphenicol LB Agar plates per bioreplicate, and three bioreplicates per condition (twelve transformations total per condition). Eleven hours post-transformation, libraries of transformants were scraped from plates and DNA extracted using the Macherey-Nagel Nucleobond Xtra Maxiprep Kit (Macherey-Nagel).

***E. coli* Essential Gene Screen Analysis** Prepared DNA libraries were sequenced on a NextSeq (Illumina), with reads mapped to the input library of spacers. Spacer depletions were calculated as the read abundance of a spacer in the empty vector condition divided by read abundance in each gene plasmid condition. Mean deple-

tions over three bioreplicates were calculated. We imposed a two-step quality-control filter on the data: a maximum coefficient of variation of 0.2 for depletion over three bioreplicates, and a minimum spacer read abundance of $1/3N$ in each bioreplicate, where $N = 55{,}700$. Weblogos of the strongly depleted (top 1% depleted) spacers were generated (Crooks et al., 2004), and from each identified PFS, heatmaps of the ratio of safely depleted ($> 5\sigma$ above mean depletion of non-targeting spacers) spacers to all spacers in the screen were generated. For spatial analysis via empirical cumulative distribution functions, safely depleted spacers were aggregated across the first or last 250 nt of genes.

For secondary structure analysis, we utilized the RNA accessibility model from Vienna RNAplfold (Bernhart et al., 2006). RNAplfold calculates through a moving average of RNA folds the probability that a region u of RNA is unpaired given its cis sequence context in a four-parameter model, where W is the moving average window length in nucleotides, L is the maximum permissible pairing distance between nucleotides in the window, and u_{start} and u_{end} are the start and end of the region u, respectively. To apply this model to our data, we separated spacers from our *E. coli* essential gene screen into training/testing cohorts of five or more, each represented by a unique permissible PFS and gene and containing at least one spacer in the top 2% of depleted spacers from the screen (to enhance predictive signal). We then randomly divided these cohorts into a training set (80%) and a testing set (20%). For optimizing a secondary structure-mediated model of efficient spacer design we selected as objective functions *top 1* or *top 3 accuracy*, the percent of cohorts for which the top spacer is accurately predicted or falls in the top 3 depleted spacers in a cohort, respectively. We optimized the two objective functions on the training dataset, first by fixing W and L while varying u_{start} and u_{end}, then by fixing u_{start} and u_{end} and varying W and L (Figure 3-4B). In the case of *bzcas13b* with *bzcsx27*, as well as that of *bzcas13b* alone, the optimized parameters were found to be approximately $W = 240$,

$L = 180$, $u_{start} = 16$, and $u_{end} = 30$. We gauged the performance of this RNAplfold model relative to 10^6 Monte Carlo simulations performed on the testing dataset and found empirical P-values of less than 1e-2 for *top 1 accuracy*, and less than 1e-5 for *top 3 accuracy*. Similar predictive power applied to *pbcas13b* with *pbcsx28*, as well as to *pbcas13b* alone.

Kanamycin Validation Screen Experiment A total of 160 kanamycin-targeting spacers was selected, 42 of which contain both PFS rules, 47 of which contain one rule, and 71 of which contain no rules, to which 162 non-targeting control spacers were added (Table S5). The library of spacers was cloned into either a *bzcas13b* and *B. zoohelcum* direct repeat-spacer-direct repeat backbone or simply a *B. zoohelcum* direct repeat-spacer-direct repeat backbone containing a chloramphenicol resistance gene using Golden Gate Assembly (NEB) with 100 cycles, and then transformed over one 22.7cm x 22.7cm carbenicillin LB Agar plate. The two cloned library plasmids were then re-transformed with over a 22.7cm x 22.7cm chloramphenicol LB Agar plate or a 22.7cm x 22.7cm kanamycin-chloramphenicol LB Agar plate. Libraries of transformants were scraped from plates and DNA extracted using the QIAGEN Plasmid Plus Maxi Kit (QIAGEN). 100 ng of library DNA and 100 ng of pMAX-GFP (Lonza), containing a kanamycin resistance gene were added to 50 uL of chemically competent 10-beta cells (NEB) and transformed according to the manufacturer's protocol.

Kanamycin Validation Screen Analysis Prepared DNA libraries were sequenced on a NextSeq (Illumina), with reads mapped to the input library of spacers. For normalizing the abundance of spacers of two separate clonings, the corrected experimental read abundance of a given spacer was calculated as the read abundance of that spacer in the *bzcas13b* plasmid (kanamycin-chloramphenicol transformation) multiplied by the ratio of the read abundance ratio of that spacer in the non-*bzcas13b* plasmid (chloramphenicol-only transformation) to the read abundance ratio of that

spacer in the *bzcas13b* plasmid (chloramphenicol-only transformation).

Nucleic Acid Preparation For in vitro synthesis of RNA, a T7 DNA fragment must be generated. To create T7 DNA fragments for crRNAs, top and bottom strand DNA oligos were synthesized by IDT. The top DNA oligo consisted of the T7 promoter, followed by the bases GGG to promote transcription, the 30 nt target and then direct repeat. Oligos were annealed together using annealing buffer (30 mM HEPES pH 7.4, 100 mM potassium acetate, and 2 mM magnesium acetate). Annealing was performed by incubating the mixture for 1 min at 95°C followed by a −1°C/minute ramp down to 23°C. To create ssRNA targets, short targets (Trunc2, 3, 4) were synthesized as top and bottom strand oligos containing the T7 promoter. For long ssRNA targets (E1, E2, S and L CRISPR Arrays), DNA primers (Table S1) with a T7 handle on the forward primer were ordered and the DNA fragment was amplified using PCR. T7 DNA constructs for RNA generation without body labeling were incubated with T7 polymerase overnight (10-14 hr) at 30°C using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). Body-labeled constructs were incubated with Cyanine 5-UTP (Perkin Elmer) and incubated with T7 polymerase overnight at 30°C using the HiScribe T7 High Yield RNA Synthesis kit (New England Biolabs). For a complete list of crRNAs and target ssRNAs used in this study see Table S1. 5′ end labeling was accomplished using the 5′ oligonucleotide kit (VectorLabs) and with a maleimide-IR800 probe (LI-COR Biosciences). 3′ end labeling was performed using a 3′ oligonucleotide labeling kit (Roche) and Cyanine 5-ddUTP (Perkin Elmer). RNAs were purified using RNA Clean and Concentrator columns^{TM}-5 (Zymo Research). Body-labeled dsRNA substrates were prepared by T7 DNA fragments for the bottom and top RNA strand. After synthesis, 1.3-fold excess of non-labeled bottom strand ssRNA was added and re-annealed to ensure the top strand would be annealed to a bottom strand by incubating the mixture for 1

min at 95°C followed by a −1°C/minute ramp down to 23°C.

Nuclease Assay Nuclease assays were performed with equimolar amounts of end-labeled or body-labeled ssRNA target, purified protein, and crRNA, for targeted ssRNA cleavage. For CRISPR array cleavage, protein was supplied in a four times molar excess of the CRISPR array. Reactions were incubated in nuclease assay buffer (10 mM TrisHCl pH 7.5, 50 mM NaCl, 0.5 mM $MgCl_2$, 20 U SUPERase Inhibitor (ThermoFisher Scientific), 0.1% BSA). Reactions were allowed to proceed at 37°C for times specified in the figure legends. After incubation, samples were then quenched with 0.8U of Proteinase K (New England Biolabs) for 15 min at 25°C. The reactions were mixed with equal parts of RNA loading dye (New England Biolabs) and denatured at 95°C for 5 min and then cooled on ice for 2 min. Samples were analyzed by denaturing gel electrophoresis on 10% PAGE TBE-Urea (Invitrogen) run at 45°C. Gels were imaged using an Odyssey scanner (LI-COR Biosciences).

EMSA Assay For the Electrophoretic Mobility Shift Assay (EMSA), binding experiments were performed with a series of half-log complex dilutions (crRNA and BzCas13b) from 0.594 to 594 nM. Binding assays were performed in nuclease assay buffer (without $MgCl_2$) supplemented with 10 mM EDTA to prevent cutting, 5% glycerol, and 5 mg/mL heparin in order to avoid non-specific interactions of the complex with target RNA. Protein was supplied at two times the molar amount of crRNA. Protein and crRNA were preincubated at 37°C for 15 min, after which the 5′-labeled target was added. Reactions were then incubated at 37°C for 10 min and then resolved on 6% PAGE TBE gels (Invitrogen) at 4°C (using 0.5X TBE buffer). Gels were imaged using an Odyssey scanner (LI-COR Biosciences). Gel shift of the RNA targets was quantified from an EMSA gel using ImageJ (Wayne Rasband, NIH) and plotted in GraphPad Prism version 7 (GraphPad Software, La Jolla California USA). Line regression was performed in Prism 7 using nonlinear fit with one-site binding

hyperbola. K_D values are calculated by GraphPad Prism based on regression analysis of data (Table S7).

PbCas13b Protein Purification PbCas13b (*Prevotella buccae*) was cloned into the same pET based vector and purified using a similar protocol as BzCas13b with the following differences: cells were grown at 21°C for 18 hr. Frozen cell paste was resuspended into 500 mM NaCl, 50 mM HEPES 7.5 and 2 mM DTT prior to breaking cells in the microfluidizer. The Superdex 200 column was run in 500 mM NaCl, 10 mM HEPES 7.0, and 2 mM DTT.

MS2 Phage Drop Plaque Assay Individual spacers for bacteriophage MS2 interference were ordered as complementary oligonucleotides containing overhangs allowing for directional cloning in between two direct repeat sequences in vectors containing *cas13b* (Table S2 and Table S6). 10 uM of each complementary oligo were annealed in 10X PNK Buffer (NEB), supplemented with 10 mM ATP and 5 units of T4PNK (NEB). Oligos were incubated at 37°C for 30 min., followed by heating to 95°C for 5 min. and then annealed by cooling to 4°C. Annealed oligos were then diluted 1:100 and incubated with 25 ng of Eco31I digested cas13b vector in the presence of Rapid Ligation Buffer and T7 DNA ligase (Enzymatics). Individual plasmids were prepared using the QIAprep Spin Miniprep Kit (QIAGEN), sequence confirmed and then transformed into C3000 (ATCC 15597) cells made competent using the Mix & Go *E. coli* Transformation Kit (Zymo). In the case of experiments using *csx27* or *csx28*, C3000 cells harboring *csx* plasmids were made competent and then transformed with *cas13b* direct repeat-spacer-direct repeat plasmids. Following transformation, individual clones were picked and grown overnight at 37°C in LB containing the appropriate antibiotics. The following morning, cultures were diluted 1:100 and grown to an OD_{600} of 2.0 by shaking at 37°C with 5% CO_2 at 250 rpm, then mixed with 4mL of antibiotic containing Top Agar (10 g/L tryptone, 5 g/L yeast extract, 10

g/L sodium chloride, 5 g/L agar) and poured on to LB-antibiotic base plates. 10-fold serial-dilutions of MS2 phage (ATCC 15597-B1) were made in LB and then spotted onto hardened top agar with a multi-channel pipette. Plaque formation was assessed after overnight incubation of the spotted plates at 37 C. For assessing interference levels in Figures 3-8E and 3-11D, samples were blinded using a key and the lowest dilution of phage at which plaque formation occurred was compared to a pACYC condition by eye, where the lowest dilution of MS2 that formed plaques on pACYC was set to 1. The lowest dilution of phage used for Figure 3-8E was $1.05*10^8$ pfu.

DNA Interference Assay A 34 nt target sequence consisting of a 30 nt protospacer and a permissive PFS (5′-G, 3′-AAA) was cloned into pUC19 in two locations (Table S2 and Table S6). For the transcribed target, the target sequence was cloned into the coding strand of the bla gene, in frame immediately after the start codon, with the G of the start codon serving as the 5′ PFS. For the non-transcribed target the identical target sequence (protospacer and PFS) were cloned into the AatII site of pUC19, so that the protospacer appears on the non-transcribed strand with respect to the pBla and pLac promoters. To determine interference, 25 ng of the ampicillin resistant target plasmid and 25 ng of the chloramphenicol resistant bzcas13b or empty vector (pACYC) were added to 5 uL of NovaBlue GigaSingle cells (Novagen). The cells were incubated for 30 min on ice, heatshocked for 30 s at 42°C and incubated on ice for 2 min. Then, 95 uL of SOC was added to cells and they were incubated with shaking at 37°C for 90 min, before plating the entire outgrowth (100 uL) on plates containing both chloramphenicol and ampicillin.

RFP-Tagged Protein Fluorescent Imaging One Shot Stbl3 Chemically Competent *E. coli* were transformed with plasmids containing RFP (negative control) or RFP fused to the N- or C- terminus of Csx27 of *B. zoohelcum* or Csx28 of *P. buccae* (Table S2). Clones were cultured up in 5 mL of antibiotic LB overnight, then spun

down at 5000 g and resuspended in PBS with 1% methanol-free formaldehyde. After 30 min fixation, cells were washed once with PBS and then diluted 1:2 in PBS. 5uL of sample was pipetted onto a silane-coated slide, which was covered with a coverslip. Fluorescent imaging was performed in a 63x objective microscope with oil immersion.

3.9.3 Quantification and Statistical Analysis

MS2 Interference Assay-HEPN Mutants Three bioreplicates of the MS2 interference assay were performed for the fold resistance quantification in Figures 3-8E and 3-11D. For assessing interference levels in Figures 3-8E and 3-11D, samples were blinded using a key and the highest dilution of phage at which plaque formation occurred was compared to a vector only condition by eye, where the highest dilution of MS2 that formed plaques on pACYC was set to 1. The error bars are the standard deviation of the fold-resistance for each condition.

DNA Interference Assay Three bioreplicates of the DNA interference assay were performed for the colony forming unit quantification. The mean values were taken from the mean of number of colony forming units from a standard colony forming unit count, and the standard deviation values accordingly from the same standard count.

***E. coli* Essential Gene Screen** Spacer depletions from the screen were calculated as the read abundance of a spacer in the empty vector condition divided by read abundance in each gene plasmid condition. Mean depletions over three bioreplicates were calculated. We imposed a two-step quality-control filter on the data: a maximum coefficient of variation of 0.2 for depletion over three bioreplicates, and a minimum spacer read abundance of $1/3N$ in each bioreplicate, where $N = 55{,}700$. This reduced the number of guides represented from N to approximately 30,000-40,000.

For secondary structure analysis, we utilized the RNA accessibility model from

Vienna RNAplfold (Bernhart et al., 2006). To apply this model to our data, we separated spacers from our *E. coli* essential gene screen into training/testing cohorts of five or more, each represented by a unique permissible PFS and gene and containing at least one spacer in the top 2% of depleted spacers from the screen (to enhance predictive signal). We then randomly divided these cohorts into a training set (80%) and a testing set (20%), with the size of a testing set ranging from approximately $n = 40$ to $n = 60$, depending on the screen. For optimizing a secondary structure-mediated model of efficient spacer design we selected as objective functions *top 1* or *top 3 accuracy*, the percent of cohorts for which the top spacer is accurately predicted or falls in the top 3 depleted spacers in a cohort, respectively. We gauged the performance of this RNAplfold model relative to 10^6 Monte Carlo simulations performed on the testing dataset and found empirical P-values of less than 1e-2 for *top 1 accuracy*, and less than 1e-5 for *top 3 accuracy*. Similar predictive power applied to *pbcas13b* with *pbcsx28*, as well as to *pbcas13b* alone.

K_D Calculations Gel shift of the RNA targets was quantified from an EMSA gel using ImageJ (Wayne Rasband, NIH) and plotted in GraphPad Prism version 7 (GraphPad Software, La Jolla California USA). Line regression was performed in Prism 7 using nonlinear fit with one-site binding hyperbola. K_D values are calculated by GraphPad Prism based on analysis of regression data (Table S7).

3.9.4 Data and Software Availability

Data Resources Data have been deposited in the following resources:

Next-Generation Sequencing for *E. coli* essential gene screen, kanamycin validation screen: `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA358111`

Chapter 4

Conclusion

4.1 Thesis Summary and Impact

The discovery and characterization of the class 2 CRISPR-Cas system Cas13b advanced the CRISPR field in two ways. First, it demonstrated that functional CRISPR systems could exist without the adaptive machinery of Cas1 and Cas2 at their loci. Second, it found that one particular class 2 RNA-targeting system possessed a natural "on/off" switch. These two properties are captured in the Graphical Abstract from Smargon et al. (2017) (Figure 4-1). Although my research was predicated on the notion that certain functional class 2 CRISPR systems might lack the canonical adaptive machinery, nevertheless the result is still surprising. If the systems are functional yet cannot adapt, how did natural selection permit them to remain in the efficient genomes of bacteria? Csx27 and Csx28 may hold the answer. It is perhaps even more surprising that these auxiliary CRISPR proteins are nearly always present at type VI-B loci. Undoubtedly there is more to the story.

Cas13b differs from the previously described Cas13a in several important respects. For one, with the exception of the HEPN domains, the two effector proteins have homologously distinct architectures (Shmakov et al., 2017). Next, while Cas13a

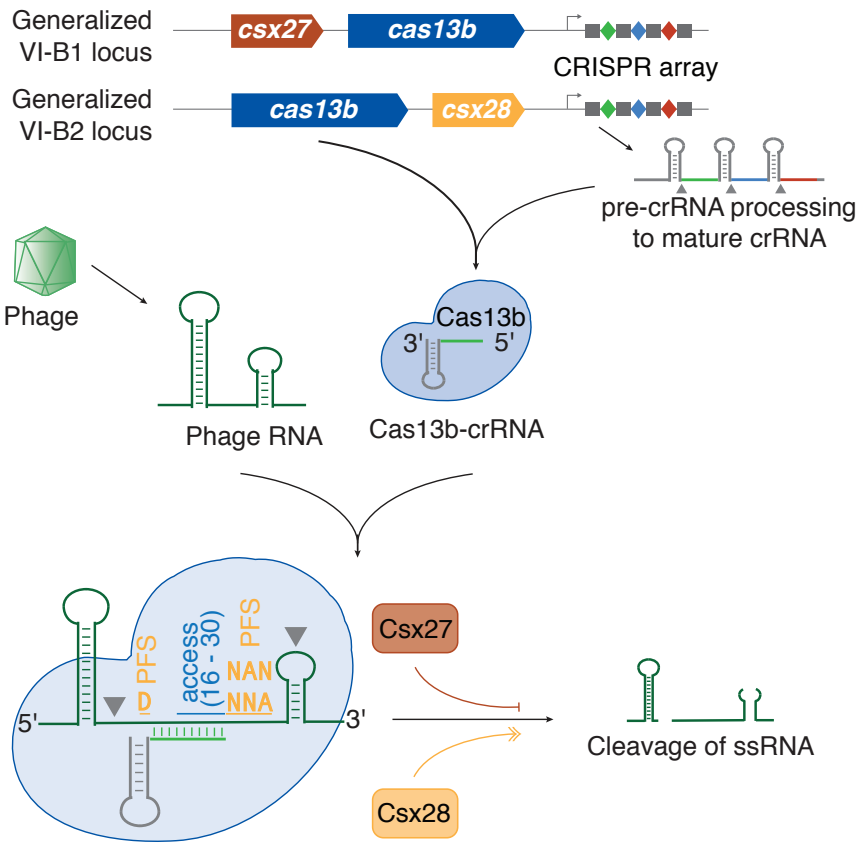


Figure 4-1: Summary of Cas13b and Type VI-B System. Figure is Graphical Abstract from Smargon et al. (2017).

possesses a 5′ direct repeat, Cas13b possesses a 3′ direct repeat, with a short and long variant in certain bacterial strains. As far as I understand, there is no precedent for this native dual repeat functionality in the CRISPR field. The longer direct repeats likely originated from an anomaly in adaptation, when a spacer integrated into the center of a direct repeat. The ability of the longer direct repeat variant to then incorporate new spacers may provide clues as to the unique adaptive mechanism of VI-B systems.

Finally, the 3′ protospacer-flanking sequence of Cas13a can be contrasted with the double-sided protospacer-flanking sequence of Cas13b, whose additional sequence requirements may make it more specific to RNA targeting. While certain class 1 CRISPR systems possess double-sided PFS variants, this finding is novel to class 2 CRISPR research. The additional finding that an adenosine residue must be located in one of two positions in the 3′ PFS suggests a certain versatility in how Cas13b interacts with target RNA on the basepair level. Solving the 3-dimensional structure of Cas13b should conclusively resolve this phenomenon.

The *E. coli* essential gene screen developed in the Cas13b study further revealed secondary structure requirements for RNA targeting. Unlike previously developed assays, the screen generalized to tens of thousands of unique spacer and protospacer sequences, and can be completed overnight after cloning. Surprisingly, the screen proved to be extremely robust and reproducible, with 60–80% of spacer depletions possessing a coefficient of variation of less than 0.2 and a minimum abundance of $1/3N$. Further implementations of the screen may prove useful in unbiased nucleic acid targeting assays of other CRISPR systems.

The greatest impact of the Cas13b study lies ahead. In the coming years, RNA targeting will be transformed by Cas13 just as DNA targeting was transformed by Cas9. As was discussed in Chapter 1, a number of proposed applications have yet to be realized–and undoubtedly many more remain to be hypothesized.

4.2 Future Research Directions

In the aftermath of the Cas13b study, several future research directions could be pursued. With the exponential growth in the number of deposited microbial genomes, a natural direction would be to continue searching for novel functional class 2 CRISPR systems through established computational and sequencing approaches. Aside from biocomputational work, many such proposed systems are waiting to be characterized experimentally.

In addition to searching for new systems, there are a few unanswered scientific questions about the CRISPR-Cas13b (VI-B) system. First, how do the VI-B CRISPR arrays adapt, if at all, in the absence of Cas1 and Cas2? One possibility is that they do not adapt, and formed around existing orphan CRISPR arrays. A second possibility is that VI-B systems lost their ability to adapt at some point in evolution. The seemingly prolific adaptation (abundance of spacers) of their CRISPR arrays argues against each of these possibilities, unless there is a strong selective pressure to preserve existing spacers. A third possibility is that VI-B systems adapt *in trans*, i.e., Cas1 and Cas2 are co-opted from other CRISPR systems in the genome. It is widely believed, however, that Cas1 and Cas2 co-evolve within a system to recognize a specific direct repeat. The diversity of other CRISPR systems present in VI-B genomes, as well as the robust conservation of the VI-B direct repeat, argue against this possibility. A fourth and final possibility is that VI-B systems adapt autonomously, without Cas1 or Cas2. If accurate, this would shatter the prevailing doctrine in the CRISPR field. Of the four possibilities, the second and third seem most probable, but the fourth tantalizes the highest experimental reward.

As a second follow-up research question, what are the mechanisms of the regulatory co-factors Csx27 and Csx28? Curiously, the downregulating Csx27 is often upstream of Cas13b, whereas the upregulating Csx28 is invariably downstream of Cas13b. Could these relative positions speak to some transcriptional feedback at

play? As far as the mechanism of interaction, do Csx27 and Csx28 interact with the Cas13b protein, mRNA encoding for Cas13b, or the CRISPR array itself? The predicted α-helices on both accessory proteins argue for binding with Cas13b. In the case of Csx27, there is precedent in the CRISPR field for α-helix-containing 'anti-CRISPRs' of Cas9 to inhibit DNA binding competitively (Pawluk et al., 2016). Similarly, Csx27 may competitively inhibit RNA binding to Cas13b. But what of the 'pro-CRISPR' Csx28? Does this agonist enhance RNA binding or RNA degradation? The presence of a putative HEPN domain in Csx28 hints at the latter.

As a third follow-up research question, what is the function of Cas13-mediated RNA degradation in bacteria? Does it truly protect against foreign invaders, or is it a form of programmed cell death via the collateral effect? While these questions were first posed when Cas13a was characterized, the differential regulation of Cas13b affords more nuanced possibilities in exploring their answers experimentally.

Beyond scientific follow-up research, the Cas13b study paves the road for future RNA-targeting tool development. Cas13b has already been engineered to edit RNA when fused to the double-stranded RNA-specific adenosine deaminase ADAR (Cox et al., 2017). In that study, several Cas13b orthologs were found to exhibit greater RNA interference in human cells than all other Cas13a/c orthologs tested (Figure 4-2A,B). One such ortholog, PspCas13b, interfered with RNA consistently more highly than the human cell-optimal Cas13a, LwaCas13a, when tiled along two transcripts (Figure 4-2C,D). In addition, PspCas13b had fewer detectable off-target sites, which may be attributable to its more constrained protospacer-flanking sequence (Figure 4-2E-G). The fact that there are numerous copies of a particular RNA molecule in a cell may explain the superior performance of PspCas13b in both RNA interference and targeting specificity.

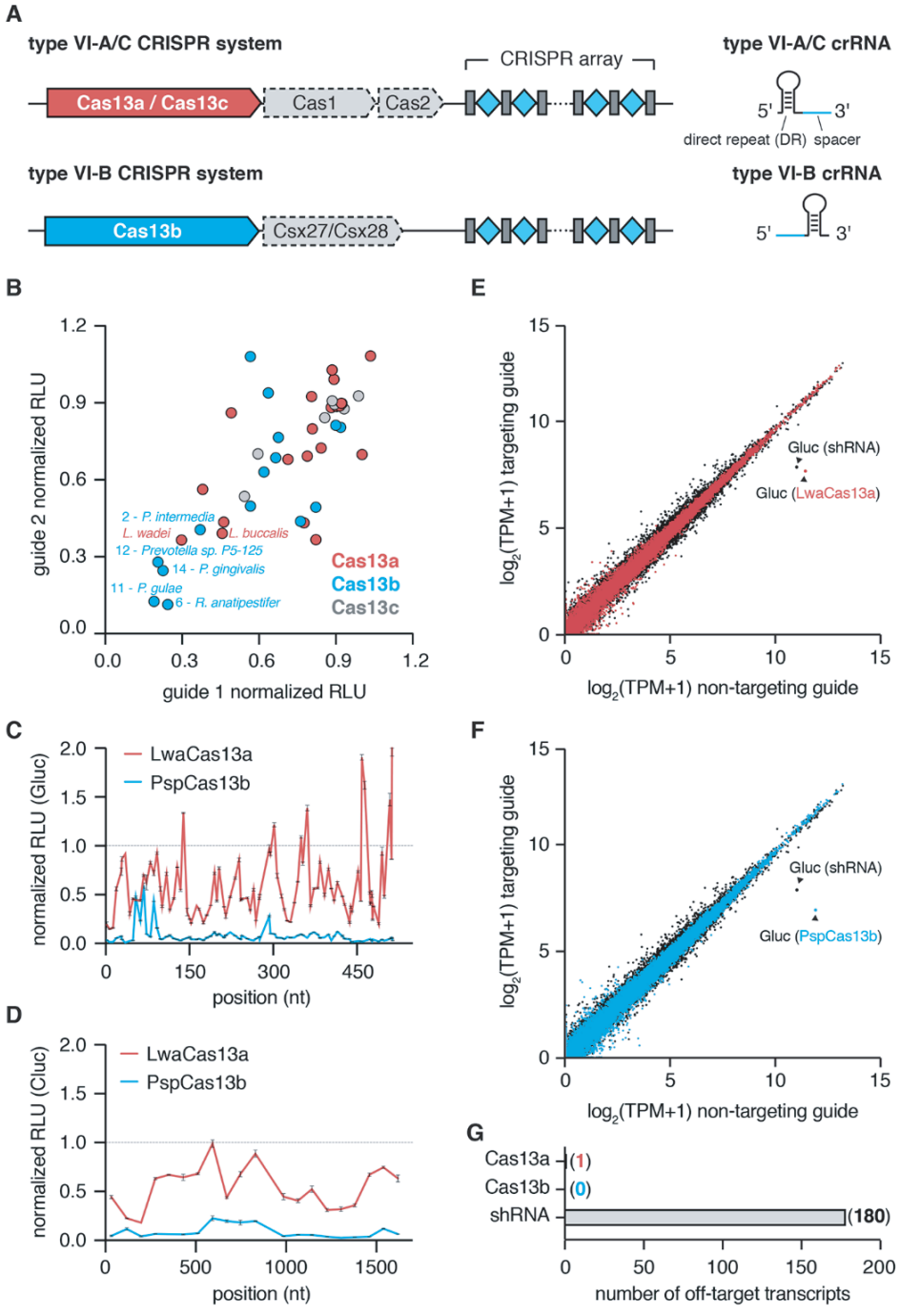

A
type VI-A/C CRISPR system
CRISPR array
type VI-A/C crRNA
Cas13a / Cas13c
Cas1
Cas2
5'
3'
direct repeat (DR)
spacer
type VI-B CRISPR system
type VI-B crRNA
Cas13b
Csx27/Csx28
5'
3'
B
guide 2 normalized RLU
guide 1 normalized RLU
2 - P. intermedia
L. wadei
L. buccalis
12 - Prevotella sp. P5-125
14 - P. gingivalis
11 - P. gulae
6 - R. anatipestifer
Cas13a
Cas13b
Cas13c
C
normalized RLU (Gluc)
LwaCas13a
PspCas13b
position (nt)
D
normalized RLU (Cluc)
LwaCas13a
PspCas13b
position (nt)
E
log2(TPM+1) targeting guide
log2(TPM+1) non-targeting guide
Gluc (shRNA)
Gluc (LwaCas13a)
F
log2(TPM+1) targeting guide
log2(TPM+1) non-targeting guide
Gluc (shRNA)
Gluc (PspCas13b)
G
Cas13a
Cas13b
shRNA
(1)
(0)
(180)
number of off-target transcripts

Figure 4-2: Characterization of a highly active Cas13b ortholog for RNA knockdown. (A) Schematic of stereotypical Cas13 loci and corresponding crRNA structure. **(B)** Evaluation of 19 Cas13a, 15 Cas13b, and 7 Cas13c orthologs for luciferase knockdown using two different guides. Orthologs with efficient knockdown using both guides are labeled with their host organism name. Values are normalized to a non-targeting guide with designed against the *E. coli LacZ* transcript, with no homology to the human transcriptome. **(C)** PspCas13b and LwaCas13 a knockdown activity (as measured by luciferase activity) using tiling guides against *Gluc*. Values represent mean +/− SEM Non-targeting guide is the same as in **(B)**. **(D)** PspCas13b and LwaCas13a knockdown activity (as measured by luciferase activity) using tiling guides against *Cluc*. Values represent mean +/− SEM Non-targeting guide is the same as in **(B)**. **(E)** Expression levels in $\log_2$ (transcripts per million (TPM+1)) values of all genes detected in RNA-seq libraries of non-targeting control (x-axis) compared to *Gluc*-targeting condition (y-axis) for LwaCas13a (red) and shRNA (black). Shown is the mean of three biological replicates. The *Gluc* transcript data point is labeled. Non-targeting guide is the same as in **(B)**. **(F)** Expression levels in $\log_2$ (transcripts per million (TPM+1)) values of all genes detected in RNA-seq libraries of non-targeting control (x-axis) compared to *Gluc*-targeting condition (y-axis) for PspCas13b (blue) and shRNA (black). Shown is the mean of three biological replicates. The *Gluc* transcript data point is labeled. Non-targeting guide is the same as in **(B)**. **(G)** Number of significant off-targets from *Gluc* knockdown for LwaCas13a, PspCas13b, and shRNA from the transcriptome wide analysis in **(E)** and **(F)**. Figure and figure legend are from Cox et al. (2017).

The structure of Cas13a has been solved (Liu et al., 2017; Knott et al., 2017), and it is only a matter of time before the same is true for Cas13b. With this structure in hand, rationally based design of Cas13b RNA-targeting tools will be achievable. Engineered Cas13b may be used to optimize nucleic acid detection (Gootenberg et al., 2017) or design more versatile RNA mammalian synthetic biology circuits (Isaacs et al., 2006; Chappell et al., 2015; Wroblewska et al., 2015; McKeague et al., 2016). Whichever direction scientists and engineers decide to take, Cas13b will likely play an important role in the future of RNA research.

Appendix A

Molecular Cell Paper Tables

A.1 Supplementary Tables

To download supplementary tables, visit `http://www.cell.com/molecular-cell/fulltext/S1097-2765(16)30866-8`.

Table S1. All crRNAs, nucleic acid targets, and primers used in biochemical experiments. Related to Figures 2-8, 3-4 and 3-8. See separate Excel file.

Table S2. All Cas13b plasmids used in this study. Related to Figures 3-1, 3-8, and 3-11. See separate Excel file.

Table S3. *E. coli* essential genes represented in *E. coli* essential gene screen library of spacers. Related to Figures 3-1, 3-9, and 3-11. See separate Excel file.

Table S4. Spacers from *E. coli* essential gene screen. Related to Figures 3-1, 3-9, and 3-11. See separate Excel file.

Table S5. Spacers from kanamycin validation screen. Related to Figure 3-1. See separate Excel file.

Table S6. Spacers targeting MS2 and pBLA plasmids. Related to Figures 3-8 and 3-11. See separate Excel file.

Table S7. EMSA raw data. Related to Figure 3-8. See separate Excel file.

A.2 Key Resources Table

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Chemicals, Peptides, and Recombinant Proteins		
BzCas13b	This study	Table S2
BzCas13b mutants (D1, D2, Q)	This study	Table S2
PbCas13b	This study	Table S2
SUPERase RNase Inhibitor	Thermo Fisher Scientific	AM2696
TURBO DNase	Life Technologies	AM2238
SUMO protease	Thermo Fisher Scientific	12588018
Critical Commercial Assays		
MiSeq Reagent Kit v3 (150 cycles)	Illumina	MS-102-3001
NextSeq 500/550 High Output v2 kit (150 cycles)	Illumina	FC-404-2002
HiScribe T7 High Yield RNA Synthesis kit	New England Biolabs	E2040S
HiScribe T7 Quick High Yield RNA Synthesis kit	New England Biolabs	E2050S
5′ oligonucleotide kit	VectorLabs	MB-9001

Deposited Data

EMSA raw data	This study	Table S7
Next-generation sequencing for bacterial RNA sequencing, E. coli essential gene screen, kanamycin validation screen	This study	`https://www.ncbi.nlm.nih.gov/bioproject/PRJNA358111`

Experimental Models: Cell Lines

Experimental Models: Organisms/Strains

B. zoohelcum	ATCC	43767
E. coli	ATCC	15597
E. coli bacteriophage MS2	ATCC	15597-B1
One Shot Stbl3 *E. coli*	Thermo Fisher Scientific	C737303
NEB 10-beta Competent *E. coli* (High Efficiency)	New England BioLabs	C3019H
MegaX DH10B T1R electrocompetent cells	Thermo Fisher Scientific	C640003
One Shot BL21(DE3)pLysE chemically competent *E. coli*	Invitrogen	C656503

Recombinant DNA

pMAX-GFP	Lonza	Not commercially available, except as part of a nucleofection kit: `http://bio.lonza.com/fileadmin/groups/FAQs/public/Technology_Flyer.pdf`
6x His/Twin Strep SUMO, a pET-based vector	Gift from Ilya Finkelstein	N/A
Plasmids generated in this study	This study	Table S2

Sequenced Based Reagents

List of spacers for *E. coli* essential gene screen	This study	Table S4
List of spacers for kanamycin validation screen	This study	Table S5
List of spacers for MS2 interference and pBLA assays	This study	Table S6
ssRNA targets	This study	Table S1
DNA primers	This study	Table S1
crRNAs	This study	Table S1

Software and Algorithms

PILER-CR	Edgar, 2007	http://drive5.com/pilercr/
BLASTP	Camacho et al., 2009	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
HHpred	Remmert et al., 2011	https://toolkit.tuebingen.mpg.de/hhpredl
BLOSUM62	Henikoff and Henikoff, 1992	
Vienna RNAfold	Anantharaman et al., 2013; Lorenz et al., 2011	http://rna.tbi.univie.ac.at/
CRISPRTarget	Biswas et al., 2013	http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis.html

Software and Algorithms (cont.)

Burrows-Wheeler Aligner	Li and Durbin, 2009	http://bio-bwa.sourceforge.net
The Galaxy Project	Center for Comparative Genomics and Bioinformatics at Penn State, and Department of Biology and at Johns Hopkins University	https://usegalaxy.org
Vienna RNAplfold	Bernhart et al., 2006	http://rna.tbi.univie.ac.at/
ImageJ	Wayne Rasband, NIH	https://imagej.nih.gov/ij/
GraphPad Prism version 7	GraphPad Software, La Jolla, California USA	https://www.graphpad.com/scientific-software/prism/
MATLAB	MathWorks, Natick, Massachusetts, United States	hhttps://www.mathworks.com/products/matlab.html

Appendix B

Diversity and Evolution of Class 2 CRISPR-Cas Systems

This appendix is derived in part from the *Nature Reviews Microbiology* class 2 CRISPR diversity and evolution analysis article (Shamkov et al., 2017). Full citation is as follows:

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., Severinov, K., Zhang, F. and Koonin, E.V. (2017). *Diversity and evolution of class 2 CRISPR-Cas systems.* Nat. Rev. Microbiol. 15, 169−182.

Contributions: I contributed to the discovery pipeline and parts of the text in this analysis article.

B.1 Abstract

Class 2 CRISPR-Cas systems are characterized by effector modules that consist of a single multidomain protein such as Cas9 or Cpf1. We designed a computational pipeline for the discovery of novel Class 2 variants and used it to identify six new CRISPR-Cas subtypes. The diverse properties of these new systems provide potential for the development of versatile tools for genome editing and regulation. We present a comprehensive census of Class 2 types and Class 2 subtypes in complete and draft bacterial and archaeal genomes, outline evolutionary scenarios for the independent

origin of different Class 2 CRISPR-Cas systems from mobile genetic elements, and propose an amended classification and nomenclature of CRISPR-Cas.

B.2 Introduction

CRISPR-Cas (clustered regularly interspersed palindromic repeats and CRISPR-associated genes) systems provide adaptive immunity in archaea and bacteria (Makarova et al., 2006; Barrangou et al., 2007; Barrangou, 2013; Marraffini, 2015; Mohanraju et al., 2016). The structural features and mechanisms of CRISPR-Cas are described in detail in many recent reviews (Barrangou, 2013; Marraffini, 2015; Mohanraju et al., 2016; van der Oost et al., 2014). Briefly, the CRISPR-Cas response consists of three stages. During the first stage, known as adaptation, the Cas1-Cas2 protein complex (which in some cases contains additional subunits) excises a segment of the target DNA (known as the protospacer) and inserts it between the repeats at the 5′ end of a CRISPR array, yielding a new spacer. In the expression and processing stage—a CRISPR array, together with the spacers, is transcribed into a long transcript known as the pre-CRISPR (cr) RNA and processed by a distinct complex of Cas proteins (which in some cases involves additional proteins and RNA molecules) into mature, small crRNAs. Finally, during the interference stage, a complex of Cas proteins (typically, a modified processing complex) employs the crRNA as a guide to cleave the target DNA or RNA. Similarly to other defense mechanisms, CRISPR-Cas systems have evolved in the context of an incessant arms race with mobile genetic elements, resulting in extreme diversification of the Cas protein sequences and in the architecture of the CRISPR-*cas* loci (Makarova et al., 2011a; Makarova et al., 2013; Takeuchi et al., 2012; Bondy-Denomy and Davidson, 2014; Bondy-Denomy et al., 2015; van Houte et al., 2016). Owing to this diversity and the lack of universal *cas* genes, a comprehensive classification of the CRISPR-Cas systems cannot be generated

as a single phylogenetic tree but requires a multipronged approach that combines the identification of signature genes with phylogenetic trees and the analysis of sequence similarity between partially conserved *cas* genes, as well as the comparison of the loci organization (Makarova et al., 2011b; Makarova and Koonin, 2015). The latest published CRISPR-Cas classification includes two classes that are subdivided into five types and 16 subtypes (Makarova et al., 2015). Shortly after this classification a sixth type and three additional subtypes were identified (Shmakov et al., 2015).

Class 1 CRISPR-Cas systems, with multisubunit effector complexes, are most common in bacteria and in archaea (including in all hyperthermophiles), comprising ∼90% of all identified CRISPR-*cas* loci (Makarova et al., 2015). The remaining ∼10% of CRISPR-Cas systems belong to Class 2 (which use a type II, V or VI effector protein); these systems are found almost exclusively in bacteria and have not been identified in hyperthermophiles (Makarova et al., 2015; Chylinski et al., 2014).

The CRISPR-Cas systems are characterized by pronounced functional and evolutionary modularity (Makarova et al., 2013). The adaptation module responsible for spacer acquisition shows limited variation among the diverse CRISPR-Cas systems (Makarova et al., 2015). By contrast, the CRISPR-Cas effector module that mediates the maturation of crRNAs, as well as target recognition and cleavage, is more versatile in gene composition and locus architecture; this led to the two classes of CRISPR-Cas systems being defined based on the different organization of their effector modules (Makarova et al., 2013). The effector complexes of Class 1 systems consist of 4 to 7 Cas protein subunits in an uneven stoichiometry, as exemplified by the CRISPR-associated complex for antiviral defense (Cascade) of the Type I systems (Brouns et al., 2008; Jore et al., 2011; Beloglazova et al., 2015; Jackson et al., 2014) and the Csm-Cmr complexes of the Type III systems (Rouillon et al., 2013; Staals et al., 2014; Osawa et al., 2015; Taylor et al., 2015). In contrast, the signature feature of Class 2 systems is an effector module that consists of a single, multi-domain protein. The

relatively simple architecture of its effector complex has made Class 2 CRISPR-Cas systems an attractive choice for use in the new generation of genome-editing tools (Jinek et al., 2012; Cong et al., 2013; Mali et al., 2013; Gasiunas et al., 2012).

Prior to the analysis reported here, five (predicted) Class 2 effectors had been described — Cas9, Cpf1, C2c1, C2c2 and C2c3 — the most common and best studied of which is the type II effector, Cas9. Cas9 is a crRNA-dependent endonuclease that contains two unrelated nuclease domains, RuvC and HNH, which are responsible for the cleavage of the displaced (non-target) and target DNA strands, respectively, in the crRNA-target DNA complex (Jinek et al., 2012; Gasiunas et al., 2012; Nishimasu et al., 2014; Nishimasu et al., 2015; Sternberg et al., 2015; Sapranauskas et al., 2011). The Type II CRISPR-Cas loci also encode a *trans*-acting crRNA (tracrRNA) that might have evolved from the corresponding CRISPR and that is essential for pre-crRNA processing and target recognition in Type II systems (Jinek et al., 2012; Deltcheva et al., 2011; Chylinski et al., 2013; Briner et al., 2014). The protein sequence of Cpf1, the prototype type V effector, contains only one readily detectable nuclease domain, RuvC (Makarova et al., 2015; Schunder et al., 2013; Zetsche et al., 2015). However, structures of Cpf1 complexed with the crRNA, or with both crRNA and target DNA, reveal a second nuclease domain with a unique fold that is functionally analogous to the HNH domain of Cas9 (Dong et al., 2016; Yamano et al., 2016). An important difference between Cpf1 and Cas9 is that Cpf1 is a single RNA-guided nuclease that does not require a tracrRNA. Furthermore, the Cpf1 protein itself is responsible for pre-crRNA processing, although the nature of its RNAse activity is not characterized (Fonfara et al.; 2016). Cpf1 also differs from Cas9 in its cleavage pattern and in its protospacer-adjacent motif (PAM), which determines which targets are cleaved (Zetsche et al., 2015). These differences suggest that the discovery of novel Class 2 effectors could enhance the application of CRISPR systems to genome engineering. Furthermore, the discovery of two, distantly related

Class 2 effector proteins, Cas9 and Cpf1, suggests that other, distinct variants of such systems could exist. Prompted by these findings, we developed a computational pipeline to systematically identify novel Class 2 CRISPR-Cas loci in genomic and metagenomic sequences. Using Cas1, the most conserved Cas protein, as a seed, we identified three previously unknown Class 2 subtypes, two of which contained effectors distantly related to Cpf1 and were included as additional subtypes in type V; the third novel Class 2 subtype became the new type VI subtype (Shmakov et al., 2015). The expression and ability to cause interference of two of these proteins, denoted C2c1 and C2c2, has been experimentally demonstrated (Shmakov et al., 2015; Abudayyeh et al., 2016).

In this Analysis article, we expand on our previous findings (Shmakov et al., 2015; Abudayyeh et al., 2016) and describe further analysis that we believe provides a comprehensive census of Class 2 effectors in sequenced bacterial and archaeal genomes. This new analysis stems from the observation that many known CRISPR-Cas systems are non-autonomous; that is, they depend on Cas1 and Cas2 proteins that are supplied by other CRISPR-Cas loci in the same genome and, as such, their loci lack *cas1* (Makarova et al., 2015) and will not have been detected in our previous analyses (Makarova et al., 2015; Shmakov et al., 2015). We extended the search for novel Class 2 systems by using the CRISPR array itself as the seed. As a result, we identify novel, putative Class 2 effectors that were missed in the previous analyses (Makarova et al., 2015; Shmakov et al., 2015) and which belong to at least three new CRISPR-Cas subtypes. We further discuss the evolutionary implications of our findings, including evidence of a crucial role for mobile genetic elements in the independent origin of different types and subtypes of Class 2 systems.

B.3 Comparative Genomics and Evolution

B.3.1 Subtypes V-A, V-B and V-C Identified with Cas1 Seed: Large Multidomain Effectors

The distinctive feature of type II and type V CRISPR-Cas sequences is the presence, in the multidomain effector proteins, of a RuvC-like nuclease domain. In the type II effector Cas9, the RuvC-like domain contains an inserted HNH nuclease domain (Figures B-1 and B-2). Other than the RuvC-like domain, the effector proteins of the three type V subtypes do not share any detectable sequence similarity to each other or to Cas9. However, the only available crystal structures of Class 2 effectors, specifically those of Cas9 and Cpf1, reveals that they have a common structural framework (see above) (Dong et al., 2016; Yamano et al., 2016). The structures of the putative large type V effectors that were discovered using the *cas1* seed, namely those of the subtypes V-B and V-C, are unsolved, but the subtype V-B effector C2c1 was shown to have robust interference activity (Shmakov et al., 2015). All the class V effectors identified at this stage share a similar, large size (typically, 1000 to 1,300 amino acid residues) and a single common domain, the RuvC-like endonuclease, although the sequence similarity between the effector proteins of different subtypes is extremely low. It is likely that all type V effectors adopt similar bilobed structures that hold together the crRNA and the target DNA, although the effector proteins of different subtypes do not appear to be directly related.

The search for homologs of the type II and type V effectors showed that the RuvC-like nuclease domains are related to TnpB proteins, an extremely abundant but poorly characterized family of nucleases that is encoded by many autonomous (that is, those that encode an active transposase, denoted TnpA, and mediate their own transposition) and even more numerous non-autonomous (that is, those that consist solely of the *tnpB* gene and rely on transposases from other elements for their

transposition) bacterial and archaeal transposons (Figure B-3a) (Pasternak et al., 2013; Bao et al., 2013; Kapitonov et al., 2015). In addition to the RuvC-like nuclease domain, TnpB proteins contain a predicted long, positively charged, long α-helix that appears to be the counterpart to the bridge helix, a common feature of Cas9 and Cpf1 (Figure B-2). Thus, similar to the Class 2 effectors, the TnpB proteins can be predicted to bind RNA. Moreover, it has been reported that a TnpB protein from the haloarchaeon *Halobacterium salinarium* binds short overlapping sense transcripts of its own gene (Gomes-Filho et al., 2015). Biochemical and biological characterization of TnpB should shed light on the evolution of the functions of Class 2 CRISPR-Cas effectors.

The closest relatives and possible ancestors of Cas9 have been identified on the basis of readily detectable sequence similarity and on the presence of the HNH insert in the RuvC-like nuclease domain of a distinct family of TnpB proteins that is denoted IscB (Insertion Sequences Cas9-like protein B) (Chylinski et al., 2014; Kapitonov et al., 2015). It is difficult to confidently trace a direct connection between type V effector proteins and a particular group of TnpB proteins because type V effector proteins show less similarity with TnpB proteins than Cas9 shows to IscB proteins Nevertheless, the effectors of the three subtypes of type V are similar to different TnpB families, suggesting independent origins of the effectors of different type V subtypes from the pool of *tnpB* genes (Shmakov et al., 2015).

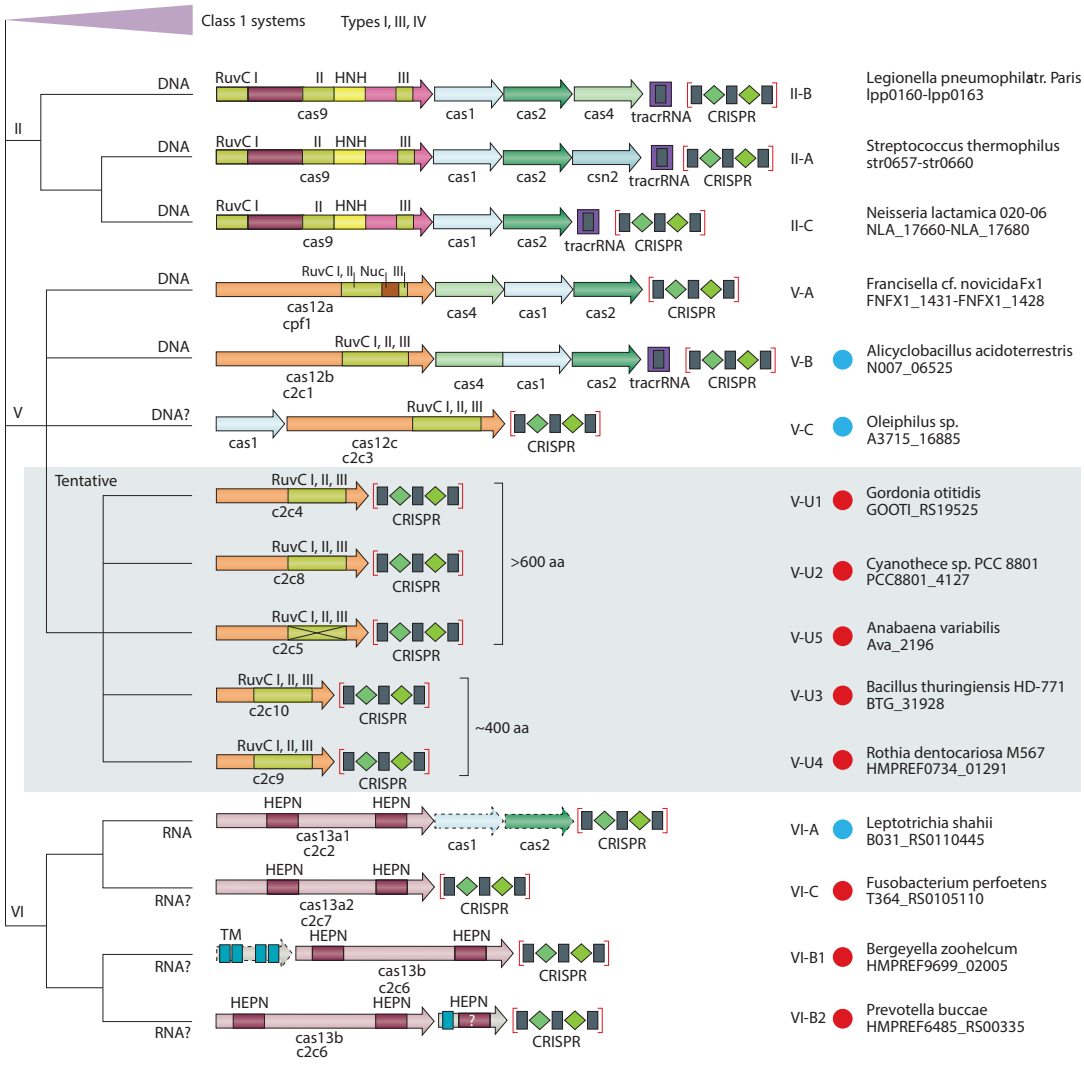

Class 1 systems
Types I, III, IV
II
V
VI
DNA
DNA
DNA
DNA
DNA
DNA?
Tentative
RNA
RNA?
RNA?
RNA?
RuvC I
II
HNH
III
cas9
cas1
cas2
cas4
csn2
tracrRNA
CRISPR
II-B
Legionella pneumophila str. Paris
lpp0160-lpp0163
II-A
Streptococcus thermophilus
str0657-str0660
II-C
Neisseria lactamica 020-06
NLA_17660-NLA_17680
RuvC I, II, Nuc, III
cas12a
cpf1
V-A
Francisella cf. novicida Fx1
FNFX1_1431-FNFX1_1428
RuvC I, II, III
cas12b
c2c1
V-B
Alicyclobacillus acidoterrestris
N007_06525
cas1
cas12c
c2c3
V-C
Oleiphilus sp.
A3715_16885
c2c4
V-U1
Gordonia otitidis
GOOTI_RS19525
c2c8
V-U2
Cyanothece sp. PCC 8801
PCC8801_4127
c2c5
V-U5
Anabaena variabilis
Ava_2196
>600 aa
c2c10
V-U3
Bacillus thuringiensis HD-771
BTG_31928
c2c9
V-U4
Rothia dentocariosa M567
HMPREF0734_01291
~400 aa
HEPN
cas13a1
c2c2
VI-A
Leptotrichia shahii
B031_RS0110445
cas13a2
c2c7
VI-C
Fusobacterium perfoetens
T364_RS0105110
TM
cas13b
c2c6
VI-B1
Bergeyella zoohelcum
HMPREF9699_02005
VI-B2
Prevotella buccae
HMPREF6485_RS00335

Figure B-1: The updated classification scheme for class 2 CRISPR-Cas systems. The class 1 systems are collapsed; all other systems shown are class 2 systems. New class 2 systems that were discovered using the computational pipeline in this study are indicated with blue circles for those that were described previously (Shmakov et al., 2015) and with red circles for those that are presented here for the first time. For each class 2 system subtype, as well as for the five distinct variants of the provisional V-uncharacterized (V-U) subtype, the locus organization and the domain architecture of the effector and accessory proteins are schematically shown. RuvC-I, RuvC-II and RuvC-III are the three distinct motifs that contribute to the nuclease catalytic center; numerals in the figure correspond to the respective RuvC motif. The portions of Cas9 proteins that roughly correspond to the recognition lobe and the protospacer-adjacent motif (PAM)-interacting domain are shown by maroon and pink shapes, respectively. The proposed new systematic gene names are shown in bold type in red boxes. Provisional gene name for effector protein candidates are shown below the respective shapes as follows: C2c1−10, class 2 candidate proteins 1-10; for subtype V-A, the previously introduced vernacular *cpf1* is indicated. For subtype VI-A, *cas1* and *cas2* are shown with dashed contours to indicate that only some of these loci include the adaptation module. For the V-U5 variant, the inactivation of the RuvC-like nuclease domain is indicated by a cross. The specific strains of bacteria in which these systems were identified and locus tags for the respective protein-coding genes are also indicated. The abbreviation TM indicates a predicted transmembrane helix. The predicted type of target, namely DNA or RNA, is indicated for each subtype. A question mark next to the target indicates that the activity is only predicted and has not been demonstrated experimentally. The target is not indicated for the type V-U systems because their RNA-guided interference capacity is questionable, which is additionally emphasized by shading. tracrRNA, *trans*-acting CRISPR RNA.

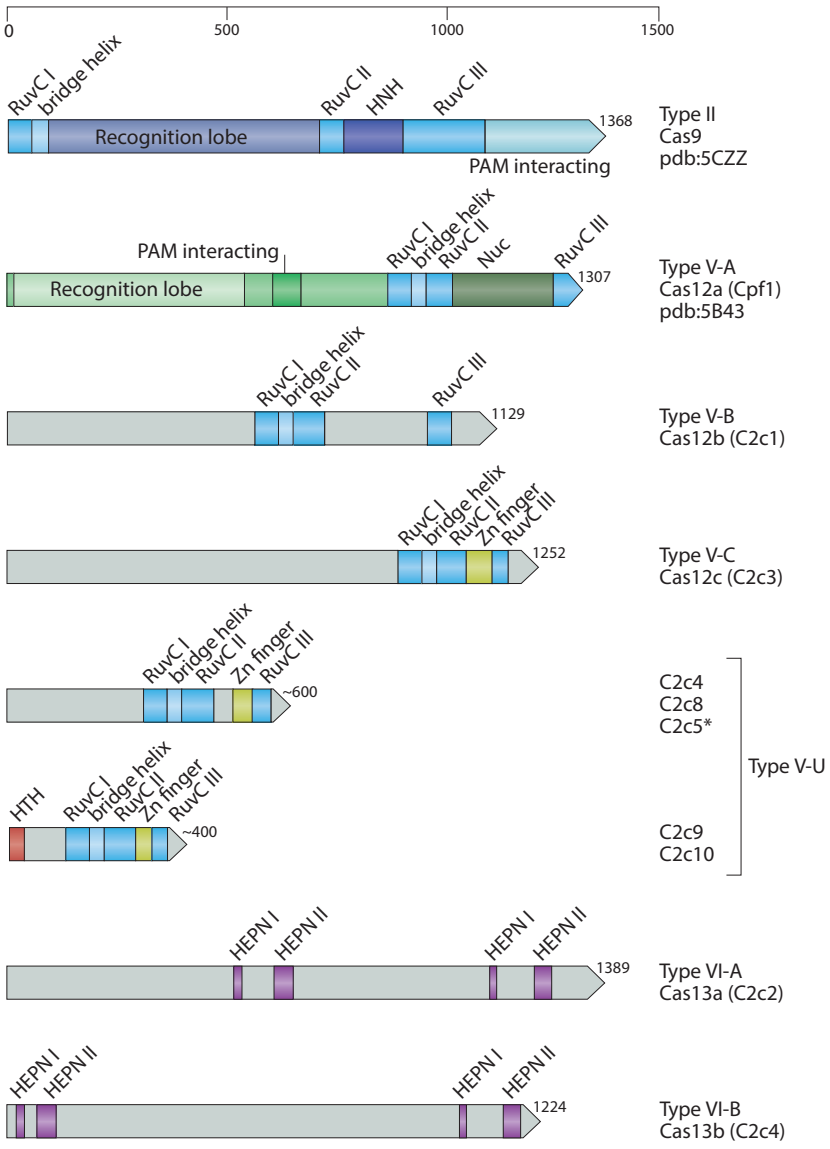

0
500
1000
1500
RuvC I
bridge helix
RuvC II
HNH
RuvC III
Recognition lobe
1368
PAM interacting
Type II
Cas9
pdb:5CZZ
PAM interacting
RuvC I
bridge helix
RuvC II
Nuc
RuvC III
Recognition lobe
1307
Type V-A
Cas12a (Cpf1)
pdb:5B43
RuvC I
bridge helix
RuvC II
RuvC III
1129
Type V-B
Cas12b (C2c1)
RuvC I
bridge helix
RuvC II
Zn finger
RuvC III
1252
Type V-C
Cas12c (C2c3)
RuvC I
bridge helix
RuvC II
Zn finger
RuvC III
~600
C2c4
C2c8
C2c5*
HTH
RuvC I
bridge helix
RuvC II
Zn finger
RuvC III
~400
C2c9
C2c10
Type V-U
HEPN I
HEPN II
HEPN I
HEPN II
1389
Type VI-A
Cas13a (C2c2)
HEPN I
HEPN II
HEPN I
HEPN II
1224
Type VI-B
Cas13b (C2c4)

Figure B-2: The domain architecture of class 2 CRISPR effector proteins. For the type II and subtype V-A effectors, the crystal structures (indicated here by their RCSB Protein Data Bank (PDB) accession numbers (5CZZ and 5B43, respectively)) are available and the corresponding domain architectures are shown in detail. For the remainder of the proteins, the grey areas indicate structurally and functionally uncharacterized portions. RuvC-I, RuvC-II and RuvC-III, as well as higher eukaryotes and prokaryotes nucleotide-binding I (HEPN I) and HEPN II, denote the catalytic motifs of the respective nuclease domains of the CRISPR effectors. The bridge helix corresponds to an arginine-rich region that follows the RuvC-I motif. Other domains shown in the figure are denoted as follows: PAM interacting, protospacer-adjacent motif (PAM)-interacting domain; HNH, HNH family endonuclease domain, zinc finger domain with a CXXC..CXXC motif (dots represent the variable distance between the two pairs of cysteines); HTH, putative DNA-binding helix-turn-helix domain; NUC, nuclease domain. The proteins and domains are shown approximately to scale. For each protein, the corresponding number of amino acids is indicated, and a ruler is shown on top of the figure to guide the eye. For the functionally characterized full-length effectors, the proposed new nomenclature (Cas12 and Cas13) is indicated, whereas for the uncharacterized putative effectors of type V-uncharacterized (V-U), only the provisional names are indicated. When, and if, functional evidence of a bona fide CRISPR response is reported for these effectors, they should be referred to as Cas12 proteins with the corresponding specifying letters. The putative V-U1, V-U2 and V-U5 effectors are larger than the typical TnpB proteins, whereas the V-U3 and V-U4 effectors are in the characteristic size range of TnpB. The asterisk at C2c5 indicates that this putative effector protein contains replacements of the catalytic residues of the RuvC-like nuclease domain and lacks the zinc finger.

B.3.2 Subtype V-U Identified with CRISPR Seed: Small Putative Effectors

The search for CRISPR-*cas* loci lacking the adaptation module (that is, loci that were identified with a CRISPR seed but not with a *cas1* seed) yielded several additional variants of putative type V systems (Figures B-1 and B-2) that might help explain how CRISPR-Cas effectors evolved from TnpB. The putative effector proteins of these loci that we have provisionally assigned to subtype V-U (where the 'U' stands for 'uncharacterized'; see below) share two features that distinguish them from the type II and type V effectors found at CRISPR-*cas* loci that contain Cas1 (Figure B-2). First, these proteins are much smaller than Class 2 effectors that contain Cas1, comprising

between ∼500 amino acids (only slightly larger than the typical size of TnpB) and ∼700 amino acids (between the size of TnpB and the typical size of the bona fide Class 2 effectors). Second, these putative effectors show a higher level of similarity to TnpB proteins than the larger type I and type V effectors. In particular, three groups of TnpB homologues included here in subtype V-U (denoted V-U1,2,5) showed evolutionary stability in terms of sequence conservation, consistent association with CRISPR arrays, and presence in distinct groups of bacteria (Figures B-1 and B-2; see below). A more detailed examination showed that, within each of these groups, in closely related bacterial genomes the respective loci were genuinely orthologous, as indicated by the gene synteny conservation.

In view of the identification of these smaller, CRISPR-associated TnpB homologs, we ran the pipeline with the requirement for the minimal length of the protein adjacent to the CRISPR-array lifted and examined the results for the presence of additional TnpB homologs. Numerous CRISPR-associated TnpB homologs were detected in the size range typical of the transposon-encoded TnpB, that is, ∼400 amino acids. Most of these loci were not evolutionarily conserved and were thus of questionable functional relevance. However, we additionally detected two distinct groups of such smaller, CRISPR-associated TnpBs (V-U3 and V-U4) with characteristics similar to those of the three subtype V-U groups with intermediately sized CRISPR-associated TnpBs (Figures B-1 and B-2).

Notably, the genes for the putative effectors of subtype V-U showed signs of purifying selection on protein sequences (as indicated by the low values of the non-synonymous to synonymous nucleotide substitutions, dN/dS), which was found to be particularly strong for the V-U3 group. Taken together, these observations imply that the respective TnpB homologs have CRISPR-dependent functions and, in our view, justify designating the respective loci subtype V-U.

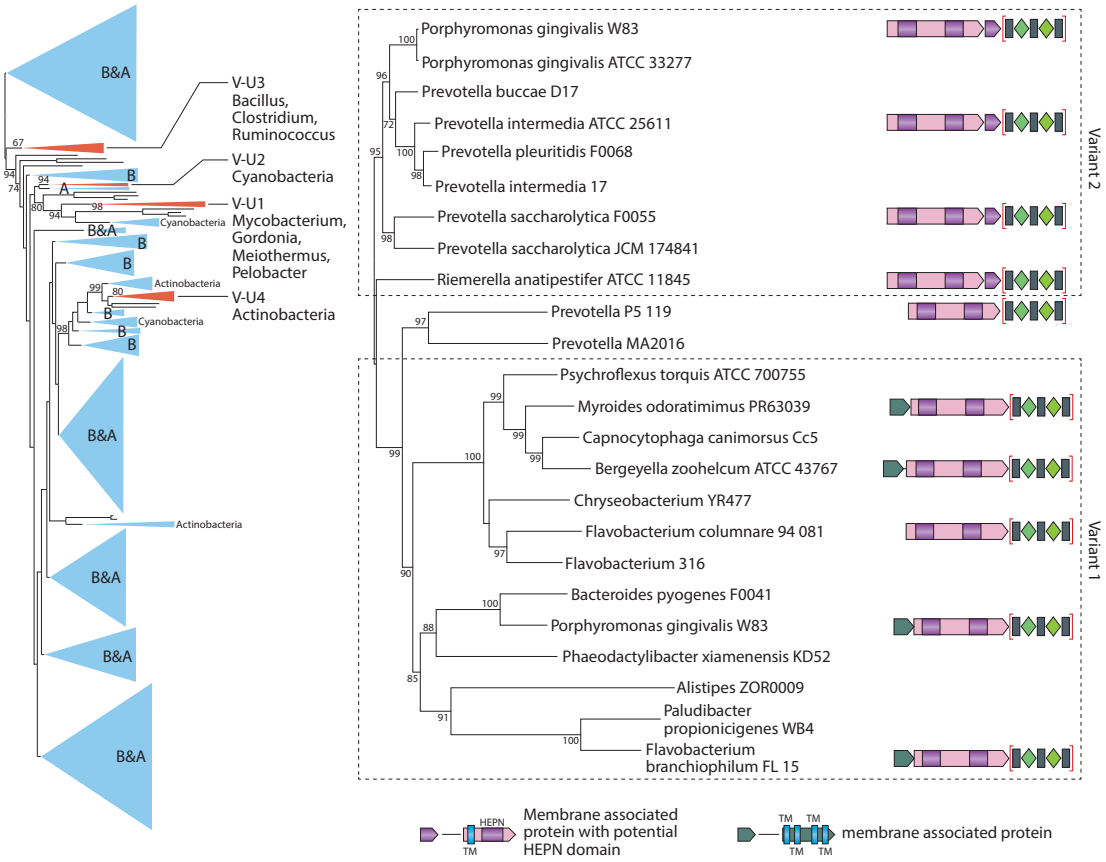
Whereas for the larger, bona fide type V effectors, low sequence conservation

precluded reliable phylogenetic analysis, a robust tree could be constructed for the smaller CRISPR-associated homologs together with the typical, transposon-encoded TnpB. The topology of this tree indicated that 4 of the 5 distinct variants of subtype V-U (hereinafter referred to as V-U1−V-U5) originated from different TnpB families (Figure B-3a), in agreement with the hypothesis on the independent evolution of different Class 2 subtype effectors from transposon-encoded nucleases. The fifth variant (V-U5), which is found in a variety of Cyanobacteria, consists of diverged TnpB homologs with multiple mutations in the catalytic motifs of their RuvC-like domain and was accordingly not included in the phylogeny here. Of the 5 stable variants, V-U1 is found in diverse bacteria whereas the remaining ones are largely limited in their spread to particular bacterial taxa (Figure B-3a). We further extended this evolutionary analysis to all putative type V effectors by building a cluster dendrogram based on the distances derived from profile to profile comparisons of the respective protein sequences. The results suggest that the effectors of each of the identified subtypes, as well as the 5 distinct variants within subtype V-U, originated independently from different TnpB families.

The subtype V-U TnpB-like proteins are too small to adopt a bilobed structure of sufficient size to accommodate the crRNA-target DNA complex, as the typical Class 2 effectors do, and therefore are unlikely to function in that capacity without additional partners. Furthermore, the subtype V-U loci lack any additional *cas* genes (Figure B-1) which, together with the above structural considerations, calls for caution in predicting that they harbour full-fledged CRISPR activity. Nevertheless, the evolutionarily stable association of at least 5 distinct V-U variants with CRISPR arrays implies that at least some of these proteins do perform CRISPR-dependent biological functions. Such functions might involve a typical CRISPR response that is aided by Cas proteins from other loci and/or by additional, non-Cas proteins. Remarkably, the CRISPR arrays associated with group V-U3, which is mostly found in

Bacilli and Clostridia, contain multiple spacers matching genomic sequences of bacteriophages that infect these bacteria. Furthermore, the sets of spacers within each subtype V-U group were completely different, even between closely related bacterial genomes, which implies active spacer turnover. The diversity of the spacers and the presence of the phage-specific spacers in V-U3 imply that at least some subtype V-U variants are functional CRISPR-Cas systems that are engaged in anti-phage adaptive immunity. Many of the complete genomes containing V-U3 and V-U4 loci lack any additional CRISPR-Cas systems, which makes it puzzling how these systems acquire their spacers. Alternatively, some of the V-U systems might have distinct regulatory roles that do not require the formation of a ternary complex with the crRNA and the DNA target; indeed, several non-defense functions of CRISPR-Cas have been described (Westra et al., 2014). This possibility is particularly plausible for the V-U5 variant, which appears to encompass a catalytically inactive TnpB homologue (Figure B-2, denoted C2c5). Furthermore, in genomes that contain the V-U2 and V-U5 loci, along with other CRISPR-Cas systems, the CRISPR sequences associated with the former loci are unique, suggesting that these type V-U systems have distinct functions.

The signature of type VI systems is the presence of an effector protein containing two HEPN domains (Figures B-1 and B-2). The HEPN domains are common in various defense systems, the experimentally characterized of which, such as the toxins of numerous prokaryotic toxin-antitoxin systems or eukaryotic RNase L, all possess RNase activity (Anantharaman et al., 2013; Makarova et al., 2014). Therefore, the first putative type VI effector, denoted C2c2, was predicted to function as an RNA-guided RNase (Shmakov et. al, 2015). Subsequently, this prediction was experimentally validated, and the type VI effectors were shown to protect against the RNA bacteriophage MS2 (Abudayyeh et al., 2016). In addition, a novel feature of C2c2 is that, once primed with the cognate target RNA, the effector turns into



a promiscuous RNase that has a toxic, growth-inhibitory effect on bacteria. These findings demonstrate a coupling between adaptive immunity and programmed cell death (or dormancy induction) that was previously predicted via comparative genomic analysis (Makarova et al., 2012) and mathematical modeling (Iranzo et al., 2015). More recently, the C2c2 protein was shown to mediate not only interference but also pre-crRNA processing (East-Seletsky et al., 2016). Since then, the tertiary structure of C2c2 (Cas13a) has been determined, confirming that it possesses two HEPN domains that interact to cleave target RNA and an additional catalytic site that controls pre-crRNA cleavage (Liu et al., 2017).

Figure B-3: Phylogenies of the type V and type VI-B effectors. A maximum-likelihood phylogenetic tree of TnpB nucleases, including the putative type V-uncharacterized (V-U) effectors that have a predicted active RuvC domain. The major subtrees of transposon-encoded TnpB proteins are collapsed and indicated by triangles; some of these large groups include *tnpB* genes that are adjacent to CRISPR arrays, but these do not show evolutionary stability and thus cannot be identified as effectors. The four distinct evolutionarily stable groups of CRISPR-associated TnpB assigned to subtype V-U are shown by red triangles. Altogether, the tree includes 1,770 unique TnpB sequences, 403 of which are TnpB proteins that are encoded next to TnpA (autonomous transposons); 168 of these *tnpB* genes are adjacent to CRISPR arrays, and of these, 49 are assigned to four variants of subtype V-U (none of these belongs to autonomous transposons). In the subtrees that include the subtype V-U variants, bootstrap values (percentages) are shown for those subtrees that include the distinct V-U variants. For each type V-U variant, the bacterial taxa that harbour the majority of the respective loci are indicated. Dominant bacterial or archaeal lineages, if there are any, are indicated in the triangles. **b** | Phylogenetic tree of the subtype VI-B Cas13b effector proteins. The tree was constructed as in part **a**, and the bootstrap values that are larger than 70% are indicated. The organization of typical *cas13b* loci for selected representatives (specifically those that are shown in bold) is schematically shown on the right. Variant 1 and variant 2 correspond to the two major branches of the tree and differ with respect to the domain architectures of the second smaller protein encoded in the locus; the domain architectures of these putative accessory proteins are shown above (for variant 1) and below (for variant 2) the respective loci schematics. The CRISPR arrays are shown schematically in brackets. TM indicates a predicted transmembrane domain, shown by blue boxes. Higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domains are shown as maroon boxes. A, diverse archaea; B, diverse bacteria.

B.3.3 Subtypes VI-B and VI-C Identified with CRISPR Seed: RNA-targeting CRISPR-Cas

The search for CRISPR-Cas loci using the CRISPR seed identified two additional large putative effectors that contain two HEPN domains and which we assigned to subtypes VI-B and VI-C, respectively (the C2c2-encoding loci accordingly became subtype VI-A). This classification of the type VI systems into separate subtypes is justified by the extremely low sequence similarity between the three groups of effectors, which is practically limited to the catalytic motif of the HEPN domain, the

different positions of the HEPN domains with the large protein sequences, and the additional features of the locus architecture in the case of subtype VI-B (Figures B-1 and B-2). Specifically, the two distinct variants of subtype VI-B, VI-B1 and VI-B2, both encode additional proteins that contain predicted transmembrane domains; VI-B1 encodes one of these and VI-B2 contains four (Figure B-3b). Phylogenetic analysis of the effector proteins suggests that the VI-B1 and VI-B2 variants diverged during evolution in accordance with the distinct architectures of the associated predicted membrane proteins (Figure B-3b). Furthermore, the single-transmembrane protein of VI-B1 encompasses an additional HEPN domain, the third one in the Type VI system (Figure B-3b). It was shown recently that the VI-B effector Cas13b has collateral RNA catalytic activity, which is differentially regulated in the VI-B1 and VI-B2 systems (Smargon et al., 2017).

Given that all of the putative type VI effectors discovered so far are similar in size to the active Class 2 effectors of subtype VI-A (Anantharaman et al., 2013), even those loci that lack *cas1* are likely to be functional CRISPR-Cas systems that rely on adaptation modules from other loci in the same genome. Moreover, given that RNA viruses only represent a minor part of the prokaryotic virome (Koonin et al., 2015), type VI systems might primarily elicit toxin activity in response to the active transcription of foreign DNA. This mechanism might not be limited to type VI given the presence of HEPN domains in poorly characterized Cas proteins in many CRISPR-Cas systems; indeed, the RNase activity of the HEPN-containing Csm6 and Csx1 proteins in type III systems has been demonstrated (Sheppard et al., 2016; Niewoehner et al., 2016), whereas their functions in the CRISPR response remain to be studied.

B.4 Census of Class 2 CRISPR-Cas Loci

The design of our CRISPR-Cas discovery pipeline implies that the analysis described in this article has identified nearly all variants of Class 2 systems that are present in the currently available bacterial and archaeal genomes. Given that the current databases include only a small fraction of the entire inferred microbial diversity of the biosphere (Curtis et al., 2002; Curtis et al., 2006; Fraser et al., 2009; Quince et al, 2008), the discovery of new CRISPR-Cas subtypes, or even of novel CRISPR-Cas types, is likely. However, such novel variants are expected to be either extremely rare or limited in their spread to specific groups of microbes that are at present poorly sampled.

B.4.1 Comprehensive Census of Class 2 CRISPR-Cas Loci in Bacteria and Archaea

Therefore we were interested in a comprehensive census of Class 2 types and subtypes in the current set of complete bacterial and archaeal genomes. To this end, we constructed sequence profiles for the effectors of all identified Class 2 subtypes (two separate profiles were used for the variants V-U1, V-U2 and V-U5 and the V-U3 and V-U4 variants were not included in the census because, in database searches, they cannot be readily distinguished from transposon-encoded TnpB) and compared these to the proteins encoded in the 4,961 completely sequenced prokaryotic genomes and 43,599 partial prokaryotic genomes available from the NCBI. This procedure should detect virtually all instances of each effector, including highly diverged variants. The neighborhoods of the respective genes were then examined for the presence of CRISPR arrays and additional *cas* genes as previously described (Makarova et al., 2015).

The most striking observation is the dramatic dominance of type II, which is represented in about 8% of bacterial genomes, among the Class 2 systems. Both type

V and type VI are more than an order of magnitude less abundant, in agreement with the expectation that the CRISPR-Cas types and subtypes remaining to be discovered are rare variants (Makarova et al., 2015). An intriguing question is whether the type II CRISPR-Cas system provides a substantial fitness advantage, perhaps being more efficient in defense and/or incurring a lower cost, as compared to other Class 2 variants. Most of the Class 2 subtypes are represented in taxonomically diverse bacteria and, furthermore, for type II and subtype V-A, the effector tree topologies differ from the topology of the species tree (Chylinski et al., 2014; Dong et al., 2016). These observations indicate that horizontal gene transfer might be the key process in CRISPR-Cas evolution. It is notable, however, that the relatively abundant subtype VI-B appears to be restricted to the phylum *Bacteroidetes*, perhaps reflecting some unique aspect of the biology of these bacteria. Similarly, the V-U5 variant, which contains an inactivated TnpB homolog, is limited to Cyanobacteria (see above), and could be involved in a distinct cyanobacterial regulatory pathway. As has been previously noted (Makarova et al., 2011b; Makarova et al., 2015), and as is emphasized by this present expansion of the diversity of Class 2 systems, apart from the identification of subtype V-A in mesophilic archaea in two instances, Class 2 systems are unique to bacteria. The exclusion of Class 2 systems from archaea, particularly from hyperthermophiles in which Class 1 systems are ubiquitous, implies that there is a major functional distinction between the two classes of CRISPR-Cas systems, the nature of which remains enigmatic.

B.4.2 Origins of Class 2 CRISPR-Cas Systems

Extending the previous hypothesis on the independent origins of the effectors in different types and subtypes of Class 2, we harness the findings on incomplete type V loci to propose a more specific evolutionary scenario (Figure B-4). As discussed above, at least 5 distinct variants within subtype V-U show a substantial degree of

evolutionary stability and consistent association with CRISPR arrays, and typically contain TnpB homologs that are intermediate in size between the compact transposon-encoded TnpB proteins and the large Class 2 effectors (Figures B-2 and B-3b). These groups of TnpB homologs might represent intermediate stages in independent paths to the emergence of new CRISPR-Cas variants. The other CRISPR-*tnpB* associations are not evolutionarily conserved and are likely to result from more or less random insertions of *tnpB* genes next to CRISPR arrays; some of these loci could represent the earliest stages of evolution of CRISPR-Cas systems.

All subtype V-U loci lack adaptation modules, suggesting that the first stage of evolution of new Class 2 CRISPR-Cas involves the random insertion of a TnpB-encoding element next to an orphan CRISPR array (Figure B-4). At the next step of evolution, the association between CRISPR and a TnpB derivative would become fixed in the microbial population, conceivably due to the emergence of a novel function, the exact nature of which remains to be understood. This would be accompanied by an increase in the size of the protein via internal duplications and/or the insertion of additional domains (Figure B-5). The final steps include further growth of the effector protein, resulting in the typical bilobed structure and, in some cases, its association with an adaptation module through recombination with a different CRISPR-Cas locus (Figure B-4). Compatible with this scenario, the Cas1 proteins of different subtypes of type II and of type V are homologous to different subtypes of type I (Shmakov et al., 2015). That the adaptation modules came last is strongly suggested by the fact that no subtype V-U loci contain the *cas1* and *cas2* genes, whereas many of the loci containing typical, large effector proteins do.

The above scenario might be challenged with respect to the directionality of evolution: the possibility could be considered that the transposon-encoded TnpB actually evolved from Class 2 effectors. However, the scenario in which transposon-encoded TnpB are the ancestral forms (Figure B-4) appears much more likely. First,

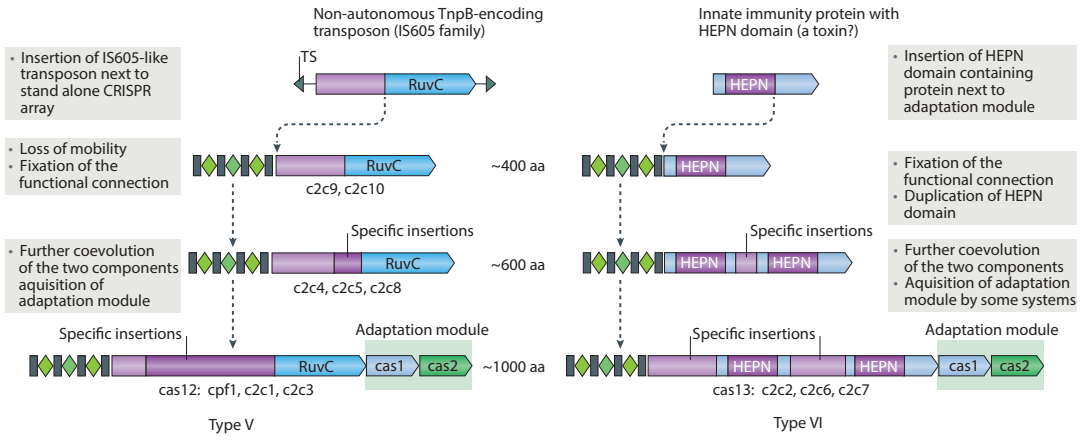


Figure B-4: Possible routes of evolution for Class 2 CRISPR-Cas systems. The figure depicts the three-step pathway of the evolutionary 'maturation' of type II, type V and type VI CRISPR-Cas systems. The systematic and/or provisional gene names are indicated below the respective 'mature' effector protein schematics and the proposed intermediate forms of type V systems. The first step involves the random insertion of a TnpB-encoding or insertion sequences Cas9-like protein B (IscB)-encoding transposon or a higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domain RNase-encoding gene next to a CRISPR cassette for type II, type V and type VI systems, respectively. During the second step, the functional connection between this protein and the CRISPR array is established and co-evolution begins, in particular, in the form of the accumulation of specific insertions that facilitate CRISPR RNA (crRNA) binding. For type V systems, the intermediate forms that correspond to the first and second step are identified as different type V-uncharacterized (V-U) variants. Additional components of the system could have originated during the second step, such as *trans*-acting CRISPR RNA (tracrRNA) in the case of type II systems. During the third step, further insertions lead to increased specificity of crRNA and target binding, and enable interactions with accessory proteins, such as Csn2 for type II-A and a protein with predicted transmembrane (TM) domains for type VI-B. The adaptation module is only inserted into some of the class 2 CRISPR-*cas* loci during the third step. TS, target site.

TnpB-encoding transposons (autonomous and non-autonomous, including some that have lost mobility) are far more abundant across a broad range of bacteria and archaea than Class 2 CRISPR-Cas systems, which are relatively rare and limited in their spread to a subset of bacterial phyla. Second, and perhaps more important, the Class 2 effectors are much larger and more complex than TnpB proteins, which makes them unlikely ancestral forms. Third, the TnpB proteins are encoded in transposons which, through their mobility, are well-suited to move into the vicinity of CRISPR arrays; by contrast CRISPR-Cas systems lack active mobility mechanisms. Finally, the observations reported here on the phylogeny of TnpB, in which the CRISPR-associated variants are lodged among the transposon-encoded TnpB (Figure B-3a), imply the ancestral status of TnpB.

Hypothetically, a similar scenario could apply to the type VI systems (Figure B-4). A comprehensive database search for HEPN domain-containing proteins encoded in the vicinity of CRISPR arrays failed to identify any evolutionarily stable configurations that might have been analogous to subtype V-U, while detecting numerous members of the HEPN-containing Cas protein families, Csm6 and Csx1. Thus, it seems possible that, during evolution, type VI systems recruited one of the HEPN-containing Cas proteins, followed by duplication of the HEPN domain and further expansion of the protein to the typical size of a Class 2 effector (Figure B-4). However, that type VI effectors directly originate from HEPN-containing toxins cannot be ruled out; further screening of new genomes and metagenomes for likely ancestors of the two-HEPN domain proteins should establish the origin of type VI effectors.

B.4.3 Amended Classification and Proposed Nomenclature

The systematic search for novel Class 2 CRISPR-Cas loci described here led to a major expansion of the known diversity of these systems. Instead of the two types and four subtypes included in the latest classification (Makarova et al., 2015), there

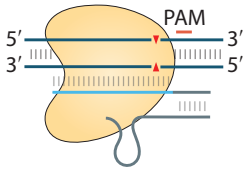
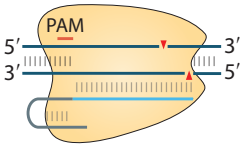
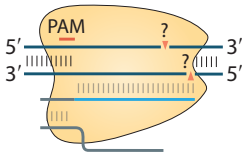
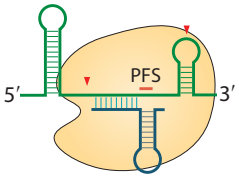
		Nuclease domains	tracrRNA	PAM	Substrate	Cleavage pattern
Type II Cas9		TnpB/RuvC+HNH	yes	3′, GC-rich	dsDNA	Blunt ends
Type V-A Cas12a (Cpf1)		TnpB/RuvC+Nuc	no	5′, AT-rich	dsDNA	Staggered ends, 5′overhangs
Type V-B Cas12b (C2c1)		TnpB/RuvC+?	yes	5′, AT-rich	dsDNA	?
Type VI-A Cas13a (C2c2)		2xHEPN	no	5′, non-G PFS_a	ssRNA	Cleaves ssRNA near uracil + collateral activity

Figure B-5: Functional diversity of the experimentally characterized Class 2 CRISPR-Cas systems. For each type of the class 2 CRISPR-Cas systems (and two subtypes in the case of type V), a schematic of the complex between the effector protein, the target, crRNA and, in the case of type II and type V-B systems, *trans*-acting CRISPR RNA (tracrRNA), is shown. The position of the protospacer adjacent motif (PAM) or the protospacer flanking site (PFS) is indicated by a red bar. The small red triangles show the position of the cut, or cuts, in the target DNA or RNA molecule. dsDNA, double-stranded DNA; ssRNA, single-stranded RNA.

are now three types and at least 10 subtypes (Figure B-1). Some uncertainty remains due to the lack of functional data on subtype V-U, but it appears likely that evolutionary stable and apparently functional variants that are currently grouped into this provisional subtype, particularly V-U3, will eventually be 'upgraded' to subtypes in their own right. The functional characterization of V-U variants will provide a more precise classification, although it is likely that many V-U loci do not encode typical, active CRISPR-Cas systems. Given the comprehensive character of the search described here, we expect that the only new variants yet to be discovered will be extremely rare or restricted in their spread to particular groups of prokaryotes that are not adequately represented in current sequence databases. In fact, two new type V CRISPR systems have been discovered in the genomes of uncultivated bacteria (Burstein et al., 2017).

We believe that the expansion of the CRISPR-Cas classification calls for the corresponding change to the nomenclature in which at least the experimentally characterized effectors and their homologues are given new names that correspond to numbered Cas proteins (Figure B-2). Thus, the type V effectors would become Cas12a, Cas12b and Cas12c, and those of type VI would become Cas13a, Cas13b and Cas13c (numerical continuity with Cas9 is not possible because Cas10 and Cas11 are already used for other proteins) (Makarova et al., 2015). The structure of Cas12b has been solved in complex with crRNA, tracrRNA and the DNA template (Liu et al., 2016; Yang et al., 2016), and it has been found to cleave target and non-target strands with its RuvC-like nuclease domain. We currently refrain from renaming the putative subtype V-U effectors until functional evidence of a bona fide CRISPR response for these effectors is reported, at which time we propose that they are referred to as Cas12 proteins.

B.5 Concluding Remarks

The genomic analysis presented here expands the diversity of Class 2 CRISPR-Cas systems. In particular, the inclusion of non-autonomous CRISPR-Cas systems lacking the adaptation module, combined with the search of expanded genomic and metagenomics databases, led to the discovery of three new subtypes which, together with our previous analysis, increases the number of Class 2 subtypes from 4 to 10. Furthermore, one of the new subtypes, V-U, at present, is a collection of diverse variants, some of which are expected to become new subtypes once they have been functionally characterized. It seems especially notable that the newly discovered Class 2 systems all fall into the two previously defined subclasses, those that cleave the non-target strand of the target dsDNA using a RuvC-like nuclease and those that attack RNA targets with a two-HEPN RNase. The apparent repeated emergence of these CRISPR-Cas varieties might reflect strict demands for protein structure to accommodate the crRNA and the target molecule, to which only a few protein folds are conducive.

The new Class 2 variants show some unprecedented functional features, for example, subtype V-A do not require a tracrRNA whereas other variants, such as subtype VI-A (and likely all type VI systems) exclusively target RNA and appear to induce a toxic response in cells. The subtype V-U is expected to show even more unusual properties. This functional diversity provides the potential for the development of new, versatile genome editing and regulation tools. We provide indications that different Class 2 types and subtypes independently originate from mobile elements that encode diverse TnpB proteins (types II and types V) and from HEPN domain-containing proteins (type VI) that ultimately originate from mRNA-cleaving toxins. The remarkable diversity notwithstanding, we believe that the computational pipeline applied here provides for a nearly exhaustive identification of Class 2 systems. Whatever additional variants remain to be found, they will be either extremely rare or

confined to bacterial phyla that are currently unknown or poorly sampled. However, as shown by the example of type VI, despite the rarity and/or narrow spread of such variants, their biological features could be of major interest and potential value for new applications.

Bibliography

[1] Abil, Z. and Zhao, H. (2015). *Engineering reprogrammable RNA-binding proteins for study and manipulation of the transcriptome*. Mol. Biosyst. 11, 2658−2665.

[2] Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., Severinov, K., Regev, A., Lander, E.S., Koonin, E.V. and Zhang, F. (2016). *C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector*. Science 353, aaf5573.

[3] Abudayyeh, O.O., Gootenberg, J.S., Essletzbichler, P., Han, S., Joung, J., Belanto, J.J., Verdine, V., Cox, D.B.T., Kellner, M.J., Regev, A., Lander, E.S., Voytas, D.F., Ting, A.Y. and Zhang, F. (2017) *RNA targeting with CRISPR-Cas13*. Nature 550, 280–284.

[4] Adamala, K.P., Martin-Alarcon D.A. and Boyden, E.S. (2016). *Programmable RNA-binding protein composed of repeats of a single modular unit*. Proc. Natl. Acad. Sci. USA 113(19), E2579–2588.

[5] Anantharaman, V., Makarova, K.S., Burroughs, A.M., Koonin, E.V. and Aravind, L. (2013). *Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing*. Biol. Direct 8, 15.

[6] Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. and Mori, H. (2006). *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Mol. Syst. Biol. 2, 2006.0008.

[7] Bao, W. and Jurka, J. (2013). *Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements.* Mob. DNA 4, 12.

[8] Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007). *CRISPR provides acquired resistance against viruses in prokaryotes.* Science 315, 1709−1712.

[9] Barrangou, R. (2013). *CRISPR-Cas systems and RNA-guided interference.* Wiley Interdiscip. Rev. RNA 4, 267−278.

[10] Beloglazova, N., Kuznedelov, K., Flick, R., Datsenko, K.A., Brown, G., Popovic, A., Lemak, S., Semenova, E., Severinov, K. and Yakunin, A.F. (2015). *CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli.* Nucleic Acids Res. 43, 530−543.

[11] Bernhart, S.H., Hofacker, I.L. and Stadler, P.F. (2006). *Local RNA base pairing probabilities in large sequences.* Bioinformatics 22, 614−615.

[12] Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H. and Long, R.M. (1998) *Localization of ASH1 mRNA Particles in Living Yeast.* Mol. Cell 2, 437–445.

[13] Biswas, A., Gagnon, J.N., Brouns, S.J., Fineran, P.C. and Brown, C.M. (2013). *CRISPRTarget: bioinformatic prediction and analysis of crRNA targets.* RNA Biol. 10, 817−827.

[14] Bondy-Denomy, J. and Davidson, A.R. (2014). *To acquire or resist: the complex biological effects of CRISPR-Cas systems.* Trends Microbiol. 22, 218−225.

[15] Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M.F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K.L. and Davidson, A.R. (2015). *Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins.* Nature 526, 136−139.

[16] Brouns, S. J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V. and van der Oost, J. (2008). *Small CRISPR RNAs guide antiviral defense in prokaryotes.* Science 321, 960−964.

[17] Briner, A.E., Donohoue, P.D., Gomaa, A.A., Selle, K., Slorach, E.M., Nye, C.H., Haurwitz, R.E., Beisel, C.L., May, A.P. and Barrangou R. (2014). *Guide RNA functional modules direct Cas9 activity and orthogonality.* Mol. Cell 56, 333−339.

[18] Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A. and Banfield, J.F. (2017). *New CRISPR-Cas systems from uncultivated microbes.* Nature 542, 237−241.

[19] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009). *BLAST+: architecture and applications.* BMC Bioinformatics 10, 421.

[20] Chappell, J., Watters, K.E., Takahashi, M.K. and Lucks, J.B. (2015). *A renaissance in RNA synthetic biology: new mechanisms, applications and tools for the future.* Curr. Opin. Chem. Biol. 28, 47−56.

[21] Chylinski, K., Le Rhun, A. and Charpentier, E. (2013). *The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems.* RNA Biol. 10, 726−737.

[22] Chylinski, K., Makarova, K.S., Charpentier, E. and Koonin, E.V. (2014). *Classification and evolution of type II CRISPR-Cas systems.* Nucleic Acids Res. 42, 6091−6105.

[23] Coller, J.M., Gray, N.K. and Wickens, M.P. (1998). *mRNA stabilization by poly(A) binding protein is independent of poly(A) and requires translation.* Genes & Dev 12, 3226–3235.

[24] Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A. and Zhang, F. (2013). *Multiplex genome engineering using CRISPR/Cas systems.* Science 339, 819−823.

[25] Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J. and Zhang, F. (2017). *RNA-editing with CRISPR-Cas13.* doi:10.1126/science.aaq0180

[26] Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004). *WebLogo: a sequence logo generator.* Genome Res. 14, 1188−1190.

[27] Curtis, T.P., Sloan, W.T. and Scannell, J.W (2002). *Estimating prokaryotic diversity and its limits.* Proc. Natl. Acad. Sci. USA 99, 10494−10499.

[28] Curtis, T.P., Head, I.M., Lunn, M., Woodcock, S., Schloss, P.D. and Sloan, W.T. (2006). *What is the extent of prokaryotic diversity?* Philos. Trans. R. Soc. Lond. B Biol. Sci. 361, 2023−2037.

[29] Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., Charpentier, E. (2011). *CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.* Nature 471, 602−607.

[30] Dong, D., Ren, K., Qiu, X., Zheng, J., Guo, M., Guan, X., Liu, H., Li, N., Zhang, B., Yang, D., Ma, C., Wang, S., Wu, D., Ma, Y., Fan, S., Wang, J.,

Gao, N. and Huang, Z. (2016). *The crystal structure of Cpf1 in complex with CRISPR RNA.* Nature 532, 522−526.

[31] *CRISPR systems in prokaryotic immunity.* The Doudna Lab, UC Berkeley. Retrieved December 9, 2017 from `http://rna.berkeley.edu/crispr.html`

[32] East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H., Tjian, R. and Doudna, J.A. (2016). *Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection.* Nature 538, 270−273.

[33] Edgar, R.C. (2007). *PILER-CR: fast and accurate identification of CRISPR repeats.* BMC Bioinformatics 8, 18.

[34] Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. and Tuschl, T. (2001). *Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells.* Nature 411, 494–498.

[35] Filipovska, A. and Rackham, O. (2011). *Designer RNA-binding proteins: new tools for manipulating the transcriptome.* RNA Biol. 8, 978−983.

[36] Fonfara, I., Richter, H., Bratovic, M., Le Rhun, A. and Charpentier, E. (2016). *The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA.* Nature 532, 517−521.

[37] Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G. and Hanage, W.P. (2009). *The bacterial species challenge: making sense of genetic and ecological diversity.* Science 323, 741−746.

[38] Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. (2012). *Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria.* Proc. Natl. Acad. Sci. USA 109, E2579−E2586.

[39] Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balázsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C. anderson, I., Gelfand, M.S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M.V., Grechkin, Y., Mseeh, F., Fonstein, M.Y., Overbeek, R., Barabási, A.L., Oltvai, Z.N. and Osterman, A.L. (2003). *Experimental determination and system level analysis of essential genes in Escherichia coli MG1655.* J. Bacteriol. 185, 5673−5684.

[40] Gomes-Filho, J. V., Zaramela, L.S., Italiani, V.C., Baliga, N.S., Vêncio, R.Z. and Koide, T. (2015). *Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea.* RNA Biol. 12, 490−500.

[41] Gootenberg, J.S., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A., Myhrvold, C., Bhattacharyya, R.P., Livny, J., Regev, A., Koonin, E.V., Hung, D.T., Sabeti, P.C., Collins, J.J. and Zhang F. (2017). *Nucleic acid detection with CRISPR-Cas13a/C2c2.* Science 356, 438–442.

[42] Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M. and Terns, M.P. (2009). *RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex.* Cell 139, 945−956.

[43] Heidrich, N., Dugar, G., Vogel, J. and Sharma, C.M. (2015). *Investigating CRISPR RNA biogenesis and function using RNA-seq.* Methods Mol. Biol. 1311, 1−21.

[44] Henikoff, S. and Henikoff, J.G. (1992). *Amino acid substitution matrices from protein blocks.* Proc. Natl. Acad. Sci. USA 89, 10915−10919.

[45] Hildebrand, A., Remmert, M., Biegert, A. and Söding, J. (2009). *Fast and accurate automatic structure prediction with HHpred.* Proteins 77 (Suppl 9), 128−132.

[46] Iranzo, J., Lobkovsky, A.E., Wolf, Y.I. and Koonin, E.V. (2015). *Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems.* BMC Evol. Biol. 15, 43.

[47] Isaacs, F.J., Dwyer, D.J. and Collins, J.J. (2006). *RNA synthetic biology.* Nat Biotechnol. 24(5), 545−554.

[48] Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A. (1987). *Nucleotide Sequence of the iap Gene, Responsible for Alkaline Phosphatase Isozyme Conversion in Escherichia coli, and Identification of the Gene Product.* J. Bacteriol. 169, 5429–5433.

[49] Jackson, R.N., Golden, S.M., van Erp, P.B., Carter, J., Westra, E.R., Brouns, S.J., van der Oost, J., Terwilliger, T.C., Read, R.J. and Wiedenheft, B. (2014). *Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli.* Science 345, 1473−1479.

[50] Jiang, W., Samai, P. and Marraffini, L.A. (2016). *Degradation of phage transcripts by CRISPR-associated rnases enables type III CRISPR-Cas immunity.* Cell 164, 710−721.

[51] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. (2012). *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.* Science 337, 816−821.

[52] Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., Beijer, M.R., Barendregt, A., Zhou, K., Snijders, A.P., Dickman, M.J., Doudna, J.A., Boekema, E.J., Heck, A.J., van der Oost, J., Brouns, S.J. (2011). *Structural basis for CRISPR RNA-guided DNA recognition by Cascade.* Nat. Struct. Mol. Biol. 18, 529−536.

[53] Kapitonov, V.V., Makarova, K.S. and Koonin, E.V. (2015). *ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs.* J. Bacteriol. 198, 797−807.

[54] Kim, Y.K., Kim, Y.G. and Oh, B.H. (2013). *Crystal structure and nucleic acid-binding activity of the CRISPR-associated protein Csx1 of Pyrococcus furiosus.* Proteins 81, 261−270.

[55] Knott, G.J., East-Seletsky, A., Cofsky, J.C., Holton, J.M., Charles, E., O'Connell, M.R. and Doudna, J.A. (2017). *Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme.* Nat. Struct. Mol. Biol. 24, 825–833.

[56] Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., Nureki, O. and Zhang, F. (2015). *Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex.* Nature 517, 583−588.

[57] Koonin, E.V., Dolja, V.V. and Krupovic, M. (2015). *Origins and evolution of viruses of eukaryotes: the ultimate modularity.* Virology 479−480, 2−25.

[58] Koonin, E.V., Makarova, K.S. and Zhang, F. (2017). *Diversity, classification and evolution of CRISPR-Cas systems.* Curr. Opin. Microbiol. 37, 67–78.

[59] Li, H. and Durbin, R. (2009). *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics 25, 1754−1760.

[60] Liu, L., Chen, P., Wang, M., Li, X., Wang, J., Yin, M. and Wang, Y. (2016). *C2c1-sgRNA complex structure reveals RNA-guided DNA cleavage mechanism.* Mol. Cell 65, 310−322.

[61] Liu, L., Li, X., Ma, J., Li, Z., You, L., Wang, J., Wang, M., Zhang, X. and Wang, Y. (2017). *The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a.* Cell, 170, 714–726.

[62] Liu, L., Li, X., Wang, J., Wang, M., Chen, P., Yin, M., Li, J., Sheng, G. and Wang, Y.(2017). *Two distant catalytic sites are responsible for C2c2 RNase activities.* Cell 168, 121−134.e12.

[63] Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011). *ViennaRNA Package 2.0.* Algorithms Mol. Biol. 6, 26.

[64] Mackay, J.P., Font, J. and Segal, D.J. (2011). *The prospects for designer single-stranded RNA-binding proteins.* Nat. Struct. Mol. Biol. 18, 256−261.

[65] Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006). *A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.* Biol. Direct 1, 7.

[66] Makarova, K.S., Aravind, L., Wolf, Y.I. and Koonin, E.V. (2011). *Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems.* Biol. Direct 6, 38.

[67] Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., van der Oost, J. and Koonin, E.V. (2011). *Evolution and classification of the CRISPR-Cas systems.* Nat. Rev. Microbiol. 9, 467−477

[68] Makarova, K.S., Anantharaman, V., Aravind, L. and Koonin, E.V. (2012). *Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes.* Biol. Direct 7, 40.

[69] Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2013). *The basic building blocks and evolution of CRISPR-Cas systems.* Biochem. Soc. Trans. 41, 1392−1400.

[70] Makarova, K.S., Anantharaman, V., Grishin, N.V., Koonin, E.V. and Aravind, L. (2014). *CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems.* Front. Genet. 5, 102.

[71] Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J., Terns, R.M., Terns, M.P., White, M.F., Yakunin, A.F., Garrett, R.A., van der Oost, J., Backofen, R., Koonin, E.V. (2015). *An updated evolutionary classification of CRISPR-Cas systems.* Nat. Rev. Microbiol. 13, 722−736.

[72] Makarova, K.S. and Koonin, E.V. (2015). *Annotation and classification of CRISPR-Cas systems.* Methods Mol. Biol. 1311, 47−75.

[73] Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013). *RNA-guided human genome engineering via Cas9.* Science 339, 823−826.

[74] Mali, P., Esvelt, K.M. and Church, G.M. (2013). *Cas9 as a versatile tool for engineering biology.* Nat. Methods 10, 957−963.

[75] Marraffini, L.A. (2015). *CRISPR-Cas immunity in prokaryotes.* Nature 526, 55−61.

[76] McKeague, M., Wong, R.S. and Smolke, C.D. (2016). *Opportunities in the design and application of RNA for gene expression control.* Nucleic Acids Res. 44(7), 2987−2999.

[77] Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., Dulay, G.P., Hua, K.L., Ankoudinova, I., Cost, G.J., Urnov, F.D., Zhang, H.S., Holmes, M.C., Zhang, L., Gregory, P.D. and Rebar, E.J. (2011) *A TALE nuclease architecture for efficient genome editing.* Nat Biotechnol. 29(2),143–148.

[78] Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E.V. and van der Oost, J. (2016). *Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems.* Science 353, aad5147.

[79] Möller, S., Croning, M.D. and Apweiler, R. (2001). *Evaluation of methods for the prediction of membrane spanning regions.* Bioinformatics 17, 646−653.

[80] Nelles, D.A., Fang, M.Y., O'Connell, M.R., Xu, J.L., Markmiller, S.J., Doudna, J.A., Yeo, G.W. (2016). *Programmable RNA Tracking in Live Cells with CRISPR/Cas9.* Cell 165, 488–496.

[81] Niewoehner, O. and Jinek, M. (2016). *Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6.* RNA 22, 318−329.

[82] Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F. and Nureki O. (2014). *Crystal structure of Cas9 in complex with guide RNA and target DNA.* Cell 156, 935−949.

[83] Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F. and Nureki, O. (2015). *Crystal structure of Staphylococcus aureus Cas9.* Cell 162, 1113−1126.

[84] Osawa, T., Inanaga, H., Sato, C. and Numata, T. (2015). *Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog.* Mol. Cell 58, 418−430.

[85] Pasternak, C., Dulermo, R., Ton-Hoang, B., Debuchy, R., Siguier, P., Coste, G., Chandler, M. and Sommer, S. (2013). *ISDra2 transposition in Deinococcus radiodurans is downregulated by TnpB.* Mol. Microbiol. 88, 443−455.

[86] Pawluk, A., Amrani, N., Zhang, Y., Garcia, B., Hidalgo-Reyes, Y., Lee, J., Edraki, A., Shah, M., Sontheimer, E., Maxwell, K.L. and Davidson, A.R. (2016). *Naturally Occurring Off-Switches for CRISPR-Cas9.* Cell 167, 1829–1838.

[87] Quince, C., Curtis, T. P. and Sloan, W. T. (2008). *The rational exploration of microbial diversity.* ISME J. 2, 997−1006.

[88] Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., Koonin, E.V., Sharp, P.A. and Zhang, F. (2015). *In vivo genome editing using Staphylococcus aureus Cas9.* Nature 520, 186−191.

[89] Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C.V., Spagnolo, L. and White, M.F. (2013). *Structure of the CRISPR interference complex CSM reveals key similarities with cascade.* Mol. Cell 52, 124−134.

[90] Remmert, M., Biegert, A., Hauser, A. and Söding, J. (2011). *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat. Methods 9, 173−175.

[91] Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and Siksnys, V. (2011). *The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli.* Nucleic Acids Res. 39, 9275−9282.

[92] Schunder, E., Rydzewski, K., Grunow, R. and Heuner, K. (2013). *First indication for a functional CRISPR/Cas system in Francisella tularensis.* Int. J. Med. Microbiol. 303, 51−60.

[93] Sheppard, N.F., Glover, C.V., 3rd, Terns, R.M. and Terns, M.P. (2016). *The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease.* RNA 22, 216−224.

[94] Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., Zhang, F. and Koonin, E.V. (2015). *Discovery and functional characterization of diverse class 2 CRISPR-Cas systems.* Mol. Cell 60, 385−397.

[95] Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., Severinov, K., Zhang, F. and Koonin, E.V. (2017). *Diversity and evolution of class 2 CRISPR-Cas systems.* Nat. Rev. Microbiol. 15, 169−182.

[96] Smargon, A.A. (2016) *An Expanded Search for RNA-Programmable Genomic Engineering Effectors* (Master's Thesis). Retrieved from DSpace@MIT: `http://hdl.handle.net/1721.1/105959`.

[97] Smargon, A.A., Cox, D.B.T., Pyzocha, N.K., Zheng, K., Slaymaker, I.M., Gootenberg, J.S., Abudayyeh, O.A., Essletzbichler, P., Shmakov, S., Makarova, K.S., Koonin, E.V. and Zhang, F. (2017). *Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28.* Mol. Cell 65, 618−630.e7.

[98] Staals, R.H., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D.W., van Duijn, E., Barendregt, A., Vlot, M., Koehorst, J.J., Sakamoto, K., Masuda, A., Dohmae, N., Schaap, P.J., Doudna, J.A., Heck, A.J.R., Yonekura, K., van der Oost, J. and Shinkai, A. (2013). *Structure and activity of the RNA-targeting type III-B CRISPR-Cas complex of Thermus thermophilus.* Mol. Cell 52, 135−145.

[99] Staals, R.H., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K., Sakamoto, K., Suzuki, T., Dohmae, N., Yokoyama, S., Schaap, P.J., Urlaub, H., Heck, A.J., Nogales, E., Doudna, J.A., Shinkai, A. and van der Oost, J. (2014). *RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus.* Mol. Cell 56, 518−530.

[100] Sternberg, S.H., LaFrance, B., Kaplan, M. and Doudna, J.A. (2015). *Conformational control of DNA target cleavage by CRISPR-Cas9.* Nature 527, 110−113.

[101] Tafer, H., Ameres, S.L., Obernosterer, G., Gebeshuber, C.A., Schroeder, R., Martinez, J. and Hofacker, I.L. (2008). *The impact of target site accessibility on the design of effective siRNAs.* Nat. Biotechnol. 26, 578−583.

[102] Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, C., Nwokeoji, A.O., Dickman, M.J., Horvath, P. and Siksnys, V. (2014). *Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus.* Mol. Cell 56, 506−517.

[103] Taylor, D. W., Zhu, Y., Staals, R.H., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E. and Doudna, J.A. (2015). *Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning.* Science 348, 581−585.

[104] Takeuchi, N., Wolf, Y.I., Makarova, K.S. and Koonin, E.V. (2012). *Nature and intensity of selection pressure on CRISPR-associated genes.* J. Bacteriol. 194, 1216−1225.

[105] Van der Oost, J., Westra, E.R., Jackson, R.N. and Wiedenheft, B. (2014). *Unravelling the structural and mechanistic basis of CRISPR-Cas systems.* Nat. Rev. Microbiol. 12, 479−492 (2014).

[106] van Houte, S., Ekroth, A.K., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A. and Westra, E.R. (2016). *The diversity-generating benefits of a prokaryotic adaptive immune system.* Nature 532, 385−388.

[107] Wahlestedt, C. (2013). *Targeting long non-coding RNA to therapeutically up-regulate gene expression.* Nat. Rev. Med. 12, 433–446.

[108] Westra, E.R., Buckling, A. and Fineran, P.C. (2014). *CRISPR-Cas systems: beyond adaptive immunity.* Nat. Rev. Microbiol. 12, 317−326.

[109] Wright, A.V., Nuñez, J.K. and Doudna, J.A. (2016). *Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering.* Cell 164, 29−44.

[110] Wroblewska, L., Kitada, T., Endo, K., Siciliano, V., Stillo, B., Saito, H. and Weiss, R. (2015). *Mammalian synthetic circuits with RNA binding proteins for RNA-only delivery.* Nat. Biotechnol. 33, 839–841.

[111] Yamano, Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I.M., Li, Y., Fedorova, I., Nakane, T., Makarova, K.S., Koonin, E.V., Ishitani, R., Zhang, F. and Nureki, O. (2016). *Crystal structure of Cpf1 in complex with guide RNA and target DNA.* Cell 165, 949−962.

[112] Yang, H., Gao, P., Rajashankar, K.R. and Patel, D.J. (2016). *PAM-dependent target DNA recognition and cleavage by C2c1 CRISPR-Cas endonuclease.* Cell 167, 1814−1828.e12.

[113] Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S.J., Cunningham, F., Aken, B.L., Zerbino, D.R. and Flicek, P. (2016). *Ensembl 2016.* Nucleic Acids Res. 44 (D1), D710−D716.

[114] Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., Koonin, E.V. and Zhang, F. (2015). *Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system.* Cell 163, 759−771.