Discovery of novel CRISPR enzymes for transcriptome engineering and human health

by

Omar O. Abudayyeh

S.B. Mechanical Engineering and Biological Engineering
Massachusetts Institute of Technology, 2012

SUBMITTED TO THE HARVARD-MIT PROGRAM IN HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

~~AUGUST 2018~~ [September 2018]

© Massachusetts Institute of Technology 2018. All rights reserved.

Signature of Author: Signature redacted
Department of Health Sciences and Technology
August 21st, 2018

Certified by: Signature redacted
Feng Zhang, PhD
Associate Professor of Biological Engineering and Brain and Cognitive Sciences
Thesis Supervisor

Accepted by: Signature redacted
Emery N. Brown, MD, PhD
Professor of Computational Neuroscience and Health Sciences and Technology
Director, Harvard-MIT Program in Health Sciences and Technology

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
OCT 09 2018
LIBRARIES
ARCHIVES

Discovery of novel CRISPR enzymes for transcriptome engineering and human health

by
Omar O. Abudayyeh

Submitted to the Harvard-MIT Program in Health Sciences and Technology in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Medical Engineering and Medical Physics

Abstract

RNA plays important and diverse roles in biology, yet molecular tools to measure and manipulate RNA are limited. Recently, the bacterial adaptive immune system, CRISPR, has revolutionized our ability to manipulate DNA, but no known RNA-targeting versions exist. To discover parallel bacterial RNA-targeting systems that could be used for transcriptome engineering, we developed a computational pipeline to mine for novel Class 2 CRISPR systems across more than 25,000 bacterial genomes. Among the many novel CRISPR systems, we found a programmable RNA-targeting CRISPR system, CRISPR-Cas13, that could provide immunity to *E. coli* against the ssRNA MS2 phage and biochemically characterized the enzyme.

We adapted CRISPR-Cas13 for modulating the transcriptome in mammalian and plant cells by heterologously expressing Cas13 and engineering the enzyme to precisely knockdown, bind, and edit RNA. Cas13 knockdown was as efficient as RNA interference, but much more specific, across many transcripts tested. RNA editing with Cas13 was also highly efficient, with up to 90% base editing rates, and as low as 20 off-targets with engineered specificity versions.

Lastly, we combined Cas13 with isothermal amplification to develop a CRISPR-based diagnostic (CRISPR-Dx), providing rapid DNA or RNA detection with single-molecule sensitivity and single-base mismatch specificity. We used this Cas13a-based molecular detection platform, termed SHERLOCK (Specific High Sensitivity Enzymatic Reporter UnLOCKing), to specifically detect pathogenic bacteria, genotype human DNA, and identify cell-free tumor DNA mutations. Our results establish CRISPR-Cas13 as a flexible platform for RNA targeting with wide applications in RNA biology, diagnostics, and therapeutics.

Thesis Supervisor: Feng Zhang
Title: James and Patricia Poitras Professor of Neuroscience at MIT
Associate Professor, Brain and Cognitive Sciences and Biological Engineering

Preface

The work presented here represents the culmination of four and a half years of work during my graduate studies in Feng Zhang's lab. Although I have been involved in many projects during this time, this thesis highlights six studies focused on the discovery of novel CRISPR effectors, the molecular characterization of these systems, and applications for cellular RNA tools and molecular diagnostics. The work has appeared in a number of published manuscripts and was made possible through the teamwork of numerous world-class collaborators, as is described in the cover page preceding each chapter. Below, I will share some thoughts about each study that will frame the work within a larger scientific context.

CHAPTER 2 – This study was my first foray into the world of enzymes beyond Cas9. In collaboration with Eugene Koonin's group, we sifted through many new candidate CRISPR systems and characterized them in bacteria and biochemically. This project was rather exciting, because at the time, most attention was focused on Cas9, and so the existence of enzymes beyond Cas9, like Cas12a, Cas12b, and Cas13, with vastly different properties, was very surprising.

CHAPTER 3 – The most amazing enzyme from Chapter 2 was Cas13, as it was a putative programmable RNase, but its mechanism was entirely a mystery. In this project, I delved into unraveling how a programmable RNase would function, discovering properties of the enzyme that greatly differed from Cas9, and finding a promiscuous cleavage effect that suggested a role in programmed cell death. This project was perhaps the most fun, as it involved cracking the Cas13 puzzle and required the most basic science approaches of any of the other studies.

CHAPTER 4 – Off the heels of understanding Cas13's molecular activity, we applied the unique collateral activity of Cas13 towards highly sensitive and specific nucleic acid testing. The assays were easy to design and always worked, yielding a robust molecular diagnostic platform that could have great impact in clinics and for global health. I most appreciate this study because it shows that the basic exploration of nature and enzymes can yield completely novel insights and technologies. I never intended to work on diagnostics, but in a way, nature guided us there.

CHAPTER 5 – In this study, we continued to develop the Cas13 diagnostic platform with new features, including portable and visual lateral flow readouts and signal amplification via crosstalk from other CRISPR enzymes, such as Csm6. We continued to explore enzyme orthologs and new classes of enzymes to produce an even better diagnostic.

CHAPTER 6 – While nucleic acid testing is certainly the killer application of Cas13, we were also interested in making better RNA tools in cells. In this work, we developed a programmable Cas13 tool for knocking down transcripts with better specificity and for imaging transcripts in live cells. This paper was a proof-of-principle that Cas13 could work in mammalian cells and that Cas13 could enable an RNA toolbox.

CHAPTER 7 – We continued to develop the RNA toolbox in this work. We first explored more Cas13 orthologs, settling on Cas13b, which demonstrated more robust and active knockdown activity in cells. We then used the Cas13 enzyme to recruit natural deaminase enzymes to allow for targeted RNA editing activity. RNA editing is an exciting area because it allows for the temporal modulation of genetic variants, which can be useful for many acute states of disease or reversible gene therapy. This study serves as a foundation for exploring RNA therapies further.

I hope these thoughts help to frame the work within the larger context of CRISPR enzyme discovery and molecular tool development, and to also demonstrate the unexpected twists and turns our research took. At many times, the paths were frustrating, but overall, we ended up at very interesting destinations. Overall, I hope to convey that science is tough, grueling, and filled with uncertainty, but those moments where you're left in absolute awe at the beauty of nature, make it all worth it in the end.

Omar Abudayyeh, writing in Cambridge, MA.

Acknowledgements

Graduate school has been a thrilling journey full of scientific growth and personal development. I would like to thank some of those who have helped and supported me during this journey. Although I have found research challenging at times, having incredibly supportive people around me has made the great times amazing and the bad times not so bad.

I would first like to thank my advisor and mentor Feng Zhang. From the early days of my rotation in the lab to now, Feng has inspired and driven me towards science. Feng has consistently given me his time for exciting science conversations at the bench, career advice whenever I need it, and frequent coffee trips around Kendall. His mentorship, inside the lab and out, has directly contributed to my success as a scientist and tremendous personal growth throughout graduate school. Feng is a thoughtful and compassionate leader with incredible vision, insight, and drive. His energy is truly infectious and his creativity and execution unparalleled. I want to thank Feng for investing time and energy in me and teaching me everything he knows in a patient and constructive manner. There is truly no better mentor one could ask for in graduate school and joining his lab was the best decision I have ever made.

I would also like to thank Jonathan Gootenberg for his partnership throughout graduate school. Although we were friends before graduate school, we have now become even closer colleagues. He is a very creative person to bounce ideas off of and to work with directly on scientific projects. All of the work in this thesis was done closely with Jonathan and I look forward to working with him more in the future.

I would also like to thank my many mentors and collaborators in the Zhang lab. Ophir Shalem and Silvana Konnerman were my early mentors in the lab and taught me so much about molecular biology and careful experimental design. I also worked with Patrick Hsu in my first year and very much enjoyed interacting with him. I would like to thank Julia Joung for many fruitful collaborations on a number of projects in the lab and her support throughout graduate school. I've worked with Brian Franklin, Patrick Essletzbichler, and Max Kellner very closely over the years and would like to thank them for their dedication and hard work on our projects. They have been the best students and research associates I have ever interacted with and are inspirational in their passion towards science. I would also like to thank David Cox who I worked closely with over the past year on our RNA editing work. I learned a lot from him and very much enjoyed the experience. I would also like to thank Ian Slaymaker for teaching me protein purification and all the tricks associated with it. I would additionally like to thank Rhiannon Macrae, Erin Blackwell, and Rich Belliveau who are the best support staff and have made the experience in the lab as smooth, enriching, and fun as possible.

I would like to also thank additional collaborators who I have worked with and learned from in graduate school, particularly Jim Collins, Pardis Sabeti, Deborah Hung, Eugene Koonin, and Eric Lander. Jim's excitement, energy, and insight make him a great person to work with and mentor to learn from. I would like to thank him for believing in our SHERLOCK work early on and supporting those efforts. I would also like to thank Eric Lander for creating a great scientific environment and his efforts in our collaborations. I have learned a great deal from his sharp insights and valuable advice. I would also like to thank Jacquin Niles, Martha Grey, Amy Cohen, and Matt Frosch who have consistently given me personal and career advice throughout my graduate career.

I would also like to thank my committee members David Bartel, Steven Elledge, Aviv Regev, and David Liu. I would especially like to thank them for taking time out of their busy schedules to be on my committee, providing rigorous feedback and insight, engaging me in scintillating scientific conversations, and giving helpful advice on my personal and career trajectory growth. I'd like to additionally thank Aviv for making herself available for many scientific discussions about our projects and beyond as well as valuable career advice.

I have had a number of funding sources throughout graduate school that I am truly thankful for because of the freedom and flexibility they have offered me. I would like to thank the Paul and Daisy Soros Fellowship for New Americans for giving me support early on and believing in me when I was still in medical school. This has been an incredible group of inspirational and supportive people. I would also like to thank the NSF Graduate Research Fellowship, National Defense Science and Engineering Graduate Fellowship, and NIH F30 National Research Service Award for additional funding and support over the years.

I would like to thank my friends at MIT, Harvard Medical School, and elsewhere for always being there and providing much needed relief from lab work. Particularly, I would like to thank Adam Akkad for being a fantastic roommate over a majority of my graduate school years and a friend since high school. He has been very supportive and caring, and is a great person to talk to about anything.

I want to also thank my parents, Osama and Ikhlas, for their unwavering support since I was born. They've poured unconditional love, lessons, and care into me and have supported me pursuing my dreams all the way to MIT and Harvard Medical School no matter the cost or distance. They have always been proud of everything I have done and none of my success would be possible without them.

Thank you again to everyone who has made my graduate school experience unforgettable and a long road well-traveled.

Table of Contents

Chapter 1

Introduction

All known life is connected by DNA. Four letters of DNA– A, T, C, and G - provide the code of life and when strung together form a unique blue print for each and every organism. In the case of the mammalian genome, billions of DNA letters come together to encode more than 20,000 genes, innumerable noncoding elements, and complex regulatory pathways we are only beginning to grasp. The first human genome was sequenced 17 years ago (Lander et al., 2001), and since then, millions more have followed, but how that information governs cellular, tissue, and even organismal function is still elusive. Much like engineers must debug a circuit to reverse engineer a device and understand its function, biologists need similar tools to reverse engineer cells and organisms. These tools would allow us to modify genes and study their function and could offer a promising path for eventually understanding the genetic circuits that drive life.

The central dogma of biology states that DNA is transcribed to RNA which is then translated into protein. Molecular biology and genetics attempt to study these processes, their regulation, and how these different molecules come together in a cell to drive complex function. Forward and reverse genetic approaches allow for perturbations to be made to DNA, RNA, or protein in order to understand how cellular function is affected. Therefore, precise tools to manipulate DNA are required to enable these functional biology experiments, and indeed many types of site-directed mutagenesis strategies have been developed, including designer meganucleases (Smith et al., 2006), Zinc finger nucleases (Miller et al., 2007; Urnov et al., 2005; Urnov et al., 2010), transcription activator like effector nucleases (TALENs) (Boch et al., 2009; Christian et al., 2010; Miller et al., 2011; Moscou and Bogdanove, 2009; Zhang et al., 2011a), and most recently CRISPR-Cas9 (Doudna and Charpentier, 2014; Hsu et al., 2014). Wielding these tools in creative ways will unlock the potential of molecular genetics to probe and understand cellular circuits.

The first tool to manipulate nucleic acid in a cell was recombinant DNA technology developed in the 1970s to express any protein of interest. Tools to directly probe the genome would eventually be developed, but technologies, such as Zinc finger nucleases and meganucleases, were not easy to use. Whereas zinc finger nucleases, meganucleases, and TALENs required extensive protein engineering work to reprogram the mutagenesis site, CRISPR-Cas9 allowed for easy reprogramming by a dual guide RNA that encodes targeting via complementarity in a 20-bp region. This has enabled research into studying how genetic variants and gene regulation affect phenotypes and diseases in cells. Additionally, numerous DNA-targeting technologies have been built with the Cas9 platform, including high-throughput genome wide screens (Konermann et al., 2015; Shalem et al., 2014; Wang et al., 2014), transcriptional activators (Bikard et al., 2013a; Chavez et al., 2015; Cheng et al., 2013; Farzadfard et al., 2013; Gilbert et al., 2014; Gilbert et al., 2013; Konermann et al., 2015; Maeder et al., 2013; Mali et al., 2013a; Perez-Pinera et al., 2013) and repressors (Gilbert et al., 2013), epigenetic modifiers (Gilbert et al., 2013; Hilton et al., 2015; Kearns et al., 2015; Vojta et al., 2016; Xu et al., 2016), genomic imagers (Chen et al., 2013), cellular lineage tracers (Frieda et al., 2017; Junker et al., 2017; Kalhor et al., 2017; McKenna et al., 2016), and direct base editors (Gaudelli et al., 2017; Komor et al., 2016). In addition to the vast amount of science being enabled by these tools, there is also great therapeutic potential for a platform that can efficiently knockout genes, modify DNA sequence, and tune gene regulation.

In this introduction, I will explore the development of CRISPR, efforts for expanding enzyme diversity, and delve into RNA targeting tools. This discussion will include CRISPR-based technologies for manipulating DNA and RNA regulation; traditional RNA tools for RNA knockdown, imaging, and editing; and nucleic acid sensing and how CRISPR could help improve molecular diagnostics. This background will help frame my thesis, which largely revolves around discovery of new CRISPR enzymes and applications for RNA manipulation and nucleic acid diagnostics.

1.1 CRISPR biology and development of genome editing tools

Although the CRISPR field has experienced a remarkable acceleration in discoveries and technology development the past few years, its beginnings date back to 1987. Nakata and colleagues in Japan were studying the iap enzyme responsible for izosyme conversion of alkaline phosphatase (Ishino et al., 1987). While studying the gene, they observed a series of 29-nt repeats downstream of the locus, but were unable to explain them. They remarked that the repeats were curious, as they appeared in noncoding regions and had interspaced regions of nonrepetitive sequences, but could not solve the problem of what their purpose was. As the sequencing revolution took off in the 1990s, many more bacteria and archaeal strains were sequenced, revealing the prevalence of similar repeat elements. In 2002, Jansen and Mojica named these interspaced repeat arrays with the acronym CRISPR, which stands for Clustered Regularly Interspaced Short Palindromic Repeats (Barrangou and Van der Oost, 2013; Jansen et al., 2002). Around this time, as more systems were sequenced, it became apparent that clusters of CRISPR-associated (Cas) genes were tightly associated with the arrays (Jansen et al., 2002). In 2005, by which time a large number of microbes and phages were sequenced, a systematic analysis of the inter-repeat spacer sequences showed similarities to extrachromosomal DNA or phage genomes (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). Because these repeat arrays were transcribed (Tang et al., 2002) and later studies found that archaea containing these systems were immune to phages targeted by these spacers (Mojica et al., 2005), it was suggested that CRISPR arrays could be an immune defense system with memory of past infections.

Although the exact mechanism of CRISPR defense was not clear, a series of experiments showed that dairy production bacteria could be protected by phage infection using CRISPR (Barrangou et al., 2007), CRISPR arrays are transcribed and processed into small crRNAs with spacers that guide Cas nuclease activity (Brouns et al., 2008), and that the target of CRISPR immunity is DNA (Marraffini and Sontheimer, 2008). By 2010, studies began to illuminate the biochemical mechanism of these systems including showing Cas9 alone is sufficient for target DNA cleavage (Garneau et al., 2010), Cas9 crRNA must hybridize to an additional trans-activating crRNA (tracrRNA) that is required for Cas9 binding and activation (Deltcheva et al., 2011), and the CRISPR system can be transplanted from endogenous hosts to *E. coli* and still be functional (Sapranauskas et al., 2011). In 2012, a study from

Charpentier and Doudna (Jinek et al., 2012) and a study from Siksnys (Gasiunas et al., 2012) showed that purified Cas9 can be guided by a hybridized crRNA:tracrRNA molecule to create a double stranded break in DNA *in vitro*. It was further shown that a chimeric guide RNA, or single-guide RNA (sgRNA), created by the fusion of the crRNA and tracrRNA was sufficient for programmable DNA cleavage by Cas9 (Jinek et al., 2012). In 2013, two studies showed that the Cas9 protein could be used for genome editing in mammalian cells by expression of the protein and guide off of DNA vectors (Cong et al., 2013; Mali et al., 2013c). Both crRNA:tracrRNA hybrids and sgRNAs could be used for genomic DNA cleavage in cells, allowing for either NHEJ- or HDR-mediated genome editing outcomes. Since then, thousands of labs have applied CRISPR for a variety of applications across diverse organisms (Doudna and Charpentier, 2014; Mohanraju et al., 2016).

As studies have shown, CRISPR proteins in bacteria and archaea constitute adaptive immune systems that capture fragments of genetic material from invading phages or mobile genetic elements and use these fragments to generate CRISPR RNAs (crRNA) which guide the cleavage of matching viral sequences upon future infections (Figure 1.1) (Marraffini, 2015). CRISPR-Cas systems display diverse sequence characteristics and architectural organization, providing a range of features that have utility in various genome editing applications. The highest-level organization of CRISPR loci separates systems into two groups: Class 1, which have multi-subunit effector complexes, and Class 2 systems, which only have single-protein effector complexes (Makarova et al., 2015a) (Figure 1.2). Class 2 systems are easier to engineer as tools, as only one protein must be reconstituted and expressed (Figure 1.3). While the Class 1 systems encompass a range of subtypes and Cas enzymes, Class 2 systems are less common, and initially thought to only contain CRISPR-Cas9 systems. This was expanded in 2015 to reflect the discovery of a second single-effector CRISPR enzyme, Cas12a (formerly called Cpf1) (Zetsche et al., 2015b) (Figure 1.4). These two Class 2 enzymes both target DNA in a programmable RNA guided fashion, but have different characteristics, such as sequence preferences, tracrRNA requirement, and cleavage overhangs.

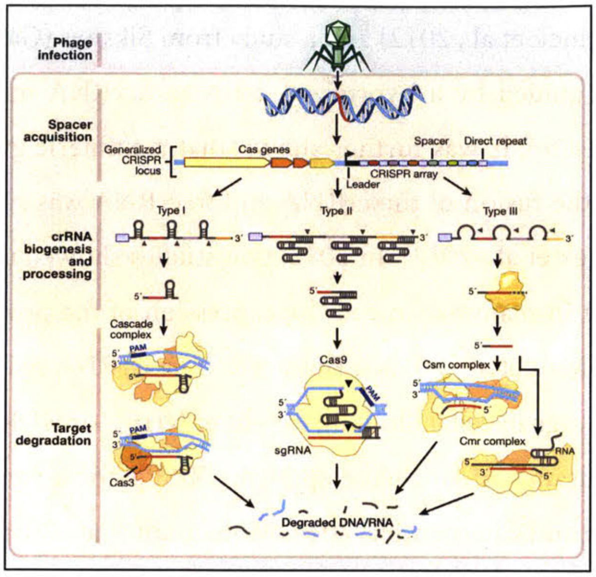


Figure 1.1: Class 1 and 2 CRISPR adaptive immunity.

CRISPR systems are adaptive immune systems in bacteria capable of cleaving invading phage genomes and protecting against infection. Adapted from (Hsu et al., 2014).

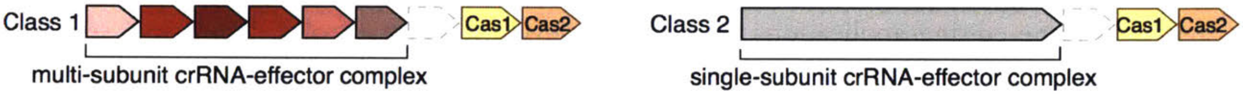

Class 1
Cas1
Cas2
multi-subunit crRNA-effector complex
Class 2
Cas1
Cas2
single-subunit crRNA-effector complex

Figure 1.2: Class 1 and Class 2 architecture.

CRISPR systems can be split into Class 1 or Class 2 systems depending on the number of effector protein subunits.

Components from CRISPR-Cas systems can be engineered and optimized to enable programmable targeting of specific DNA sequences. Typically, a Cas protein effector (*e.g.,* Cas9 or Cas12a) is combined with a customized guide RNA to establish a ribonucleoprotein complex capable of targeting DNA in a cell type of interest (Cong et al., 2013; Gasiunas et al., 2012; Jinek et al., 2012; Mali et al., 2013c; Zetsche et al., 2015b). In contrast to earlier DNA-editing enzymes (Kim and Kim, 2014) — such as meganucleases, zinc finger nucleases, and TALENs — CRISPR-Cas systems are targeted by Watson-Crick base-pairing between the guide RNA and target DNA or RNA, allowing rapid and flexible reprogramming by adjusting the sequence of the guide RNA. By modifying the Cas protein or RNA guide, the system can be used to recruit other molecules to act on the target sequence in different ways. In combination with appropriate cellular or molecular assays, this flexibility enables systematic forward and reverse genetic studies in mammalian cells.

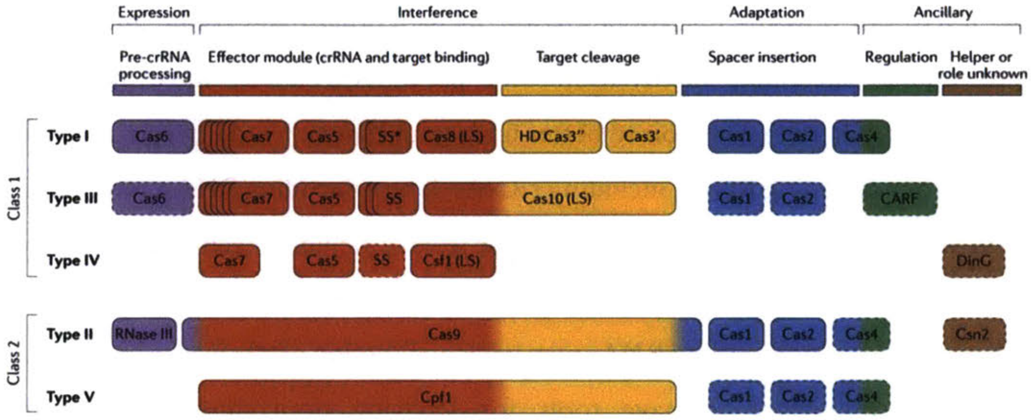


Figure 1.3: CRISPR classification.

Functional classification of Class 1 and 2 CRISPR systems. Reproduced from (Makarova et al., 2015a).

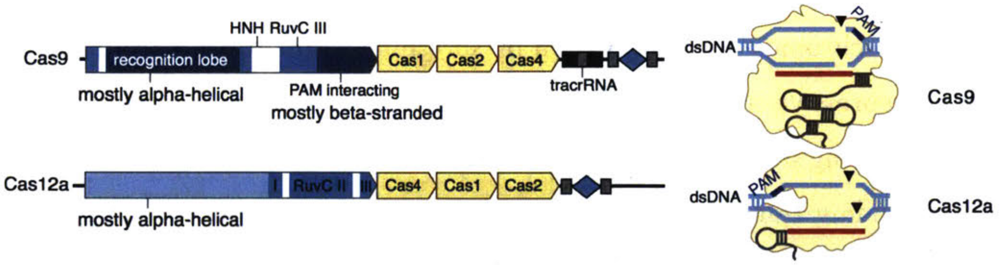


Figure 1.4: Cas9 and Cas12a CRISPR systems.

The Class 2 CRISPR enzymes Cas9 and Cas12a enact RNA-guided DNA cleavage in a programmable fashion. Shown are the typical locus organization of these systems and the structure of the protein:RNA complexes.

1.2 Importance of an RNA-targeting toolbox

RNAs are a diverse set of molecules in the cell that can broadly be categorized as either coding or noncoding. Coding RNAs are represented by messenger RNAs (mRNA) that are translated into proteins and are heavily regulated via splicing and interactions with RNA binding proteins (RBPs) and other RNAs (Hentze et al., 2018; Li et al., 2016). These regulatory interactions can dramatically reshape the transcriptomic profile of cells and are indispensable to proper cellular function. There is also emerging evidence that mRNA bases can be heavily modified, changing the interactions of these molecules with RBPs or how the nucleotides base-pair with other bases (Helm and Motorin, 2017). These modifications, including methylation and pseudouridylation, have been mapped via observational studies, but will need tools to better understand them. In addition to mRNAs, there are numerous classes of noncoding RNAs, including rRNAs, long noncoding RNAs, microRNAs, circular RNAs, and vault RNAs. These RNAs play diverse roles, including genome organization, gene transcription, splicing, transcript regulation, transcript stability and transport, rna editing, and translation. Because many non-coding RNAs may have no function at all and are products of spurious transcription, it is important to have tools for studying noncoding RNAs with different perturbational tools.

Long noncoding RNAs (lncRNAs) are a heterogeneous class of molecules, operationally defined as polyadenylated RNAs >200 nt that do not encode peptides. Prior to the genomic era, biochemical and genetic studies identified a handful of lncRNAs with important molecular functions, including *Xist,* which orchestrates X-chromosome inactivation by spreading across the X chromosome in *cis* and recruiting multiple repressive chromatin regulatory complexes to silence gene transcription (Plath et al., 2002). RNA sequencing studies found that nearly the whole genome is transcribed to some extent, and that these noncoding transcripts included thousands of polyadenylated and spliced lncRNAs (Consortium, 2012; Djebali et al., 2012; Kapranov et al., 2007). Although there are some notable exceptions, such as *Xist,* the function (if any) of most lncRNAs has remained elusive and will require new tools to help elucidate the mystery.

Initial tools for investigation of general RNA biology, such as antisense oligos (Wagner, 1994) or RNA interference (RNAi) (Elbashir et al., 2001; Fire et al., 1998; Root et al., 2006), have great utility, but are limited due to off-target effects (Jackson et al., 2003) and their reliance on endogenous machinery (Grimm et al., 2010; Valdmanis et al., 2016). The advent of protein-based systems, such as the MS2-MCP system(Bertrand et al., 1998) or IRP1, allowed for additional applications, such as imaging (Tyagi, 2009) or translational upregulation (De Gregorio et al., 1999). Some RNA binding proteins, including members of the pumilio (Pum/Puf), Pentatricopeptide, and Tristetraprolin protein families (Mackay et al., 2011; Yagi et al., 2013), show promise of reprogrammability and have been utilized for imaging (Adamala et al., 2016; Ozawa et al., 2007), translational modulation (Campbell et al., 2014), RNA knockdown (Choudhury et al., 2012), and splicing modulation in limited capacities (Wang et al., 2009). Analogous to zinc finger (Choo et al., 1994) or TALE (Boch et al., 2009; Moscou and Bogdanove, 2009) DNA-binding domains, the RNA recognition domains of these proteins are determined by their amino acid sequence, and Pum/Puf proteins cannot easily be retargeted. PUF domains recognize ssRNA by repeat domains that consist of a ~35-residue three-helix bundle that can bind to a single base (Edwards et al., 2001). Typically, the entire PUF domain will consist of eight of these repeats allowing recognition of an eight nucleotide RNA (Mackay et al., 2011). Certain residues in the repeat can be modified to change recognition to each of the four nucleotides. For example, glutamate and serine will encode for guanine recognition whereas glycine and asparagine will favor uracil (Cheong and Hall, 2006; Wang et al., 2002). The recognition code of PUF domains has been mapped over multiple studies allowing them to be reprogrammed theoretically to any sequence (Mackay et al., 2011). Despite this code, however, multiple designs must be tested and binding is not always specific.

Generally, four useful tools to probe both coding and noncoding transcripts would be valuable. Tools for modulating RNA levels, such as transcription activation or transcript knockdown, would be valuable for studying gene or transcript function. Tools to image RNAs would allow for a deeper understanding of RNA localization in the cell and dynamics during cell processes. RNA editing would allow for the temporal modulation of genetic variants for studying variant function and also therapeutics. RNA sensing would enable nucleic acid detection and quantification for both cellular and clinical applications. I'll explore these areas, including previous approaches, in the following sections.

1.3 Type III CRISPR systems target RNA

Developing CRISPR tools for targeting RNA would be ideal. However, only type III RNA-targeting CRISPR enzymes existed when starting the work of this thesis, and they are far too complex to engineer as cellular tools. Type III CRISPR systems are defined as Type III-A and Type III-B systems based on their effector complexes, Csm complex and Cmr complex, respectively (Makarova et al., 2011b). Type III-C and III-D systems also exist, but have not been characterized well (Makarova et al., 2015a). Several proteins make up these complexes with Csm3 (in III-A) or Cmr4 (in III-B) polymerizing as a protein backbone along the crRNA along with Csm2 (in III-A) or Cmr5 (in III-B) as the small subunit protein (Jackson and Wiedenheft, 2015). These proteins bend the crRNA with kinks, such that it binds every 6 nucleotides along a target RNA. Cmr3 and Csm4 also bind the 5' end of the crNRA handle, while Cas10 binds the protein complex as the large subunit (Osawa et al., 2015; Staals et al., 2014; Taylor et al., 2015). On the 3' end of the crRNA, Csm5, Cmr6, and Cmr1 bind the protein complex and cap the helical protein backbone formed by Csm3 or Cmr4. While there was confusion early on whether the Type III interference complex targeted DNA or RNA, later *in vitro* studies showed targeting and degradation of RNA targets (Hale et al., 2009; Marraffini and Sontheimer, 2008). The confusion was later reconciled, as it was shown *in vivo* that the Csm complex can cleave both DNA and RNA in a transcription-dependent manner in *Staphylococcus epidermidis* (Samai et al., 2015).

The multi-protein interference complex separates various functions amongst the constituent subunits (Wright et al., 2016). RNA interference is performed by the Csm3 backbone units that are situated along the target with cleavage every six nucleotides mediated by metal-independent RNase activity (Osawa et al., 2015; Staals et al., 2014; Tamulaitis et al., 2014; Taylor et al., 2015). In concert, Cas10 cleaves DNA exposed via a transcription bubble using a single catalytic site in the palm polymerase domain (Samai et al., 2015). This distinct co-transcriptional DNA and RNA cleavage activity allows bacterial hosts to tolerate temperate phages to persist until transcription and replication are activated. This can be advantageous for hosts as it can allow acquisition of beneficial phage genes, like antibiotic resistance genes, and when phages become lytic, can protect against replication (Wright et al., 2016).

Type III systems also lack a protospacer adjacent motif (PAM) requirement present in other CRISPR systems, which normally serves to restrict auto-immunity by permitting interference only for targets adjacent to the PAM. Instead, Type III systems check for complementarity between the direct repeat portion of the crRNA and the target, and do not cleave if there is a match (Marraffini and Sontheimer, 2010b; Wright et al., 2016). The programmable RNase activity of type III complexes make them a potential platform for developing RNA targeting tools. However, because of the large number of proteins that must be expressed and assembled, developing a robust set of tools has been difficult, especially in mammalian cells. The lack of a PAM requirement, which vastly expands the potential targeting space, could be a universal characteristic of CRISPR RNA-targeting systems perhaps yet to be found, and would be a useful feature for developing RNA targeting tools. Finding such a CRISPR system that only involves a single-protein effector for targeting RNA would be incredibly useful for building an RNA targeting platform.

1.4 RNA-targeting tools for modulating RNA levels

Several approaches using CRISPR have been developed for modulating RNA levels up or down. Beyond cleaving genomic sequences, CRISPR effectors can be used to direct protein or RNA cargo to specific locations in the genome. In such applications, functional proteins or RNAs are adjoined either to catalytically dead Cas9 (dCas9) (Gilbert et al., 2013; Mali et al., 2013a; Qi et al., 2013), in which the active sites of the nuclease domains have been mutated, or to the guide RNA via a number of different recruitment strategies. Emerging tools in this category include a variety of "epigenome editors", which recruit one or more proteins to alter gene regulatory processes.

CRISPR interference (CRISPRi) for gene repression. By itself, dCas9 delivery to gene promoters can physically block RNA polymerase or transcription factors from accessing DNA, leading to modest (20-40%) repression of endogenous gene expression (Gilbert et al., 2013). Repressive activity can be dramatically improved by fusing dCas9 to repressive chromatin regulators (Gilbert et al., 2013). For example, fusion to the Krüppel associated box (KRAB) domain leads to deposition of repressive histone modifications (including H3K9me3) and loss of activating modifications (including H3K27ac) at gene promoters, leading to 50 to 90% repression of gene expression when delivered in a critical window 0-200 bp downstream of a transcription start site (Gilbert et al., 2014; Gilbert et al., 2013) (Figure 1.5). Other repressive domains —including SID, LSD1, and DNMT3A proteins or protein domains (Bintu et al., 2016; Kearns et al., 2015; Konermann et al., 2013; Liu et al., 2016; Vojta et al., 2016) — have been explored and may have different kinetic properties for fine-tuning gene silencing.

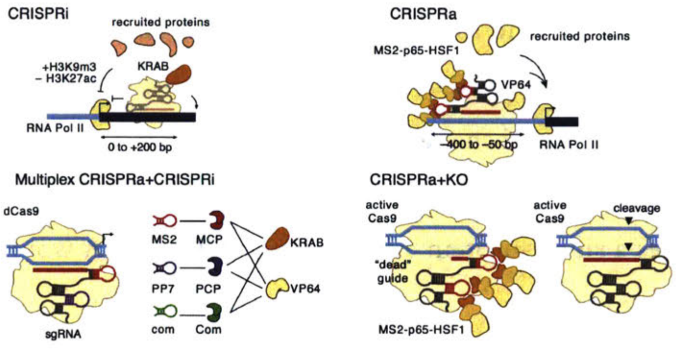


Figure 1.5: dCas9 CRISPR applications.

Various CRISPR applications using catalytically dead Cas9 to recruit effector enzymes.

CRISPR activation (CRISPRa). CRISPR-Cas9 tools for activation use Cas9 to recruit multiple transcriptional activators to achieve robust induction of gene expression at promoters (Bikard et al., 2013b; Chavez et al., 2015; Cheng et al., 2013; Farzadfard et al., 2013; Gilbert et al., 2014; Gilbert et al., 2013; Konermann et al., 2015; Maeder et al., 2013; Mali et al., 2013a; Perez-Pinera et al., 2013). Konermann *et al.* developed an approach called Synergistic Activation Mediator (SAM), in which MS2 stem loops are fused to modified Cas9 guide RNAs to recruit multiple activating proteins to the same site (Konermann et al., 2015) (Figure 1.5). Recruiting multiple activating proteins (VP64, HSF1, and p65) led to synergistic effects on gene expression compared to recruiting each protein individually (Konermann et al., 2015). In an alternative approach ("SunTag"), Tanenbaum et al. fused dCas9 to a repetitive GCN4 peptide-domain scaffold that is recognized by a GCN4-targeting nanobody fused to VP64, enabling simultaneous recruitment of up to 10 copies of VP64 (Tanenbaum et al., 2014). Each of these approaches appears to vary in efficiency across different genes in a way that is not yet entirely predictable, and the precise location of targeting (ideally, 50-400 bp upstream of the transcription start site) has an important effect on the efficiency of activation (Gilbert et al., 2014; Konermann et al., 2015).

RNA-targeting Cas9. Several other CRISPR-based approaches have been developed to enable RNA-targeted degradation or programmable protein recruitment. Although Cas9 is typically regarded as a DNA-targeting enzyme, certain natural Cas9 variants or engineered modifications have been shown to have the potential to also interact with and cleave RNA. Initial *in vitro* characterization of the Cas9 from *Streptococcus pyogenes* (SpCas9) demonstrated the possibility of RNA as a binding substrate

(Gasiunas et al., 2012), and later methods refined the activity *in vitro* via co-delivery of PAM duplex-forming small oligos, termed PAMmers (O'Connell et al., 2014). This engineered adaptation of SpCas9 for RNA binding was demonstrated to function *in vivo* for RNA imaging or cleavage of repeat-containing RNA (Batra et al., 2017; Nelles et al., 2016). Certain Cas9 orthologs also possess endogenous RNase activity and can cleave RNA without a PAMmer in *E. coli* (Strutt et al., 2018) or *in vitro* (Rousseau et al., 2018). RNA targeting with Cas9 in mammalian cells appears to require the introduction of the exogenous synthetic modified oligo (PAMmer) (Nelles et al., 2016), rendering Cas9 challenging to use for pooled screening approaches that require genetically-encodable constructs.

Beyond CRISPR, RNA interference is one of the most developed tools for knocking down RNAs. Since its discovery (Fire et al., 1998), there has been many studies exploring RNAi biology and applications, especially for clinical applications (Pecot et al., 2011). RNAi was exciting because it circumvented many of the specificity problems of small molecules and target accessibility of antibodies, promising programmable gene-specific targeting. RNAi works by cleaving target mRNA, which results in deadenylation, transcript destability, and transcript knockdown Double-stranded RNAs (dsRNAs) are recognized by the RNase Dicer, which cleaves them into small 21-23 bp fragments (Hannon and Rossi, 2004; Meister and Tuschl, 2004). These small fragments can guide the RNA-induced silencing complex (RISC) to bind to any complementary region in mRNA, allowing for cleavage (Martinez et al., 2002). Endogenous microRNAs (miRNAs), derived from non-coding RNAs in the cell, most typically target the 3' UTR where they are most effective at cleavage. When these dsRNA fragments are provided exogenously as siRNA, the RISC complex can be co-opted for programmable cleavage at a complementary target site and prevents translation via cleavage of the target RNA (Pecot et al., 2011). Although RNAi has offered major advances for studying genes, a big hurdle was it was not as specific as first believed. The siRNAs in a cell can easily bind to similar sequences, sometimes with only as much as 6-8 nt of similarity, especially inside the seed region (Jackson et al., 2006b). Although this is an evolutionary mechanism to allow for regulation of similar transcripts (Farh et al., 2005), as a tool this can have important phenotypic effects and can account for phenotypic changes that are independent of the expected target (Jackson et al., 2006a). Off-targets are most concerning when using RNAi as a tool for gene knockdown or high-throughput screens, as they can cause spurious phenotypes and lead to incorrect gene to phenotype connections. These off-targets are also concerning for clinical therapeutics. In some cases, where Dicer is properly expressed, there can be excessive overuse of the endogenous RNAi machinery by exogenous siRNA, causing toxicity due to inhibition of normal cellular processes. In one study, this caused hepatic failure and death in mice (Grimm et al., 2006). In other cases, oversaturation of RISC by exogenous siRNAs leads to derepression of regulated transcripts, causing unforeseen consequences in affected tissues (Khan et al., 2009). Additionally, many tissue types have low expression of Dicer, such as ovarian, lung, and breast cancers, preventing efficient use of RNAi (Merritt et al., 2008). As a result, co-opting an endogenous pathway for cellular perturbations or therapeutics can have negative consequences.

It would be ideal to find an exogenous RNA perturbation technology that does not affect endogenous cellular processes, has little effect on the transcriptome, and is highly specific. Particularly, finding an RNA-targeting CRISPR enzyme that is analogous to Cas9, without the needed addition of oligos or further engineering, would enable a range of applications. There are RNA-targeting enzymes within the CRISPR universe, but the RNA-targeting type III systems use large multi-protein complexes to perform programmable RNA cleavage (Hale et al., 2012; Hale et al., 2009). Because these complexes involve more than 10 protein subunits coming together, they are difficult to engineer in mammalian cells. These type III complexes can also activate RNA cleavage by an associated enzyme known as Csx1 or Csm6 (Kazlauskiene et al., 2017; Niewoehner et al., 2017). These enzymes contain the HEPN RNase domain, which is typically characterized by the RxxxxH amino acid motif (Anantharaman et al., 2013) and could be a promising alternative if they could be rendered easily programmable.

The HEPN superfamily is prevalent across prokaryotes and eukaryotes and include RNase LS and LsoA, KEN domains, and animal Sacsin proteins (Anantharaman et al., 2013). HEPN RNase activity is performed by a metal-independent endoRNase active site and typically functions in toxins. In prokaryotes, HEPN-containing proteins are in many toxin-antitoxin and abortive infection systems and play a role in both restriction modification and CRISPR-Cas immunity. Because of their prevalence in these systems, their existence is likely coupled to pathogen-targeting and strategies to protect against viruses. The Csx1 and Csm6 enzymes in type III CRISPR systems have been shown to be activated in response to foreign RNA and can non-specifically cleave and degrade RNA, likely serving an abortive infection function in cells (Kazlauskiene et al., 2017; Niewoehner et al., 2017). Given the prevalent nature of RNA-targeting in CRISPR systems, it is reasonable to speculate that there may be RNases that associate with CRISPR arrays and could function in a programmable fashion. Perhaps by searching for HEPN domains that co-localize with CRISPR arrays, it could be possible to identify programmable RNA-targeting CRISPR effectors that would serve as a defense against invading nucleic acid and could be ported to mammalian cells for RNA targeting applications.

1.5 RNA-targeting tools for RNA imaging

RNA imaging and tracking has been of significant interest ever since protein translation was linked to localized mRNA translation (Lehmann and Nusslein-Volhard, 1986). RNA localization has since become important across many biological settings, such as neurons where local mRNA translation at synapses can cause specific activity changes (Mannack et al., 2016). The oldest method for visualizing RNA was *in situ* hybridization (ISH), which used short fluorescently-labeled oligonucleotides to recognize transcripts, and has been further developed into various technologies, such as molecular beacons, which reduce background via a quenching mechanism. While these early technologies work great in fixed cells, they are difficult to deliver to live cells and are not optimized for the cellular setting, where they must overcome target secondary structure. As an alternative, bacteriophage-derived RNA binding proteins have been used for engineering RNA interactions in the cell. The MS2 coat protein (MCP) can recognize MS2 loops with high binding strength and specificity. By fusing GFP to the MCP and tagging target RNAs with multiple MS2 loops, target RNAs can be imaged in live cells (Bertrand et al., 1998). This MS2 system has even allowed for the imaging of single mRNA molecules in living mouse cells (Park et al., 2014). While these MS2 approaches have been successful, tagging RNAs with MS2 loops can affect their normal function and creating transgenic cell lines or organisms is very time consuming. Alternatively, Pumilio proteins from the PUF RNA binding protein family could be used for programmed RNA sensing without adding an RNA tag on the target. By using RNA binding proteins, multiplexed imaging would enable the study of different transcripts at once or for the localization of two transcripts together. Additionally, they would circumvent the high background fluorescence typical of RNA imaging tools by using a split-GFP system that only reconstitutes by targeting adjacent sites on a transcript (Ozawa et al., 2007; Tilsner et al., 2009). Pumilio proteins are still difficult to reprogram, however, limiting their broad applicability. A tool that is easily reprogrammable, can detect low transcript levels, and yield quantitation would be ideal.

1.6 RNA-targeting tools for RNA editing

Direct base editing of nucleic acid that circumvents endogenous repair pathways would allow for highly efficient cellular modification across a broad range of cell types and tissues. DNA base editing was first developed using CRISPR-Cas9 for site-directed recruitment of adenine deaminases.

CRISPR-mediated base editing is performed by fusing certain enzymes — such as APOBEC, AID, or ADAT engineered to target DNA — to a Cas9 protein in which one or both of its catalytic domains have been genetically inactivated (Gaudelli et al., 2017; Komor et al., 2016; Nishida et al., 2016) (Figure 1.6). These fusions currently enable specific and programmable conversions of C:G to T:A base-pairs or vice versa, enabling relatively efficient modification of just over half of all known pathogenic single-nucleotide variants (Gaudelli et al., 2017), although the requirement for a PAM sequence nearby reduces this number. An advantage of these approaches is that they can lead to DNA edits without introducing double-stranded DNA breaks, thereby avoiding certain error-prone repair mechanisms. For example, the APOBEC C-to-T base editor first enzymatically converts a target cytosine to uracil, and then creates a single-strand DNA break to engage the mismatch repair (MMR) machinery to repair the opposing guanine to adenine (Komor et al., 2016).

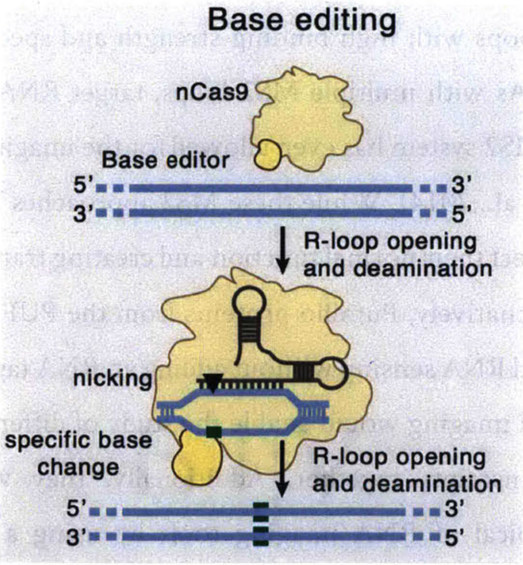


Figure 1.6: CRISPR base editing.

Base editing with Cas9 allows for direct adenosine to guanine or cytosine to thymine conversion in DNA.

While DNA base editing is efficient, it is restricted to PAM sites and still relies on some repair pathways, potentially limiting its use to dividing cells only. RNA editing offers a promising alternative, as it directly modifies the RNA bases without any repair and can be performed with high efficiency. RNA editing by adenosine deamination is a naturally occurring process in cells for the precise conversion of adenosines to inosines (Bass and Weintraub, 1987; Kim et al., 1994). The Adenosine Deaminase that Acts on RNA (ADAR) enzyme catalyzes the simple hydrolytic deamination reaction through a single active catalytic site (Melcher et al., 1996). During translation and other cellular processes, the inosines are interpreted as guanosines allowing for precise changes to the genetic code in a temporal fashion. Some of the best studied examples of ADAR editing are in neurons where RNA editing changes ion selectivity of ionotropic glutamate receptors, alters serotonin receptor function, inhibits voltage-dependent potassium channels, and modulates the transport rate of Na^+/K^+ ATPases (Montiel-Gonzalez et al., 2016).

ADAR enzymes contain two double-stranded RNA binding domains (dsRBDs) and a catalytic domain referred to as the deaminase domain ($ADAR_{DD}$). Normally, ADARs bind to specific secondary structures in a transcript via the dsRBDs, positioning the $ADAR_{DD}$ in the right orientation to deaminate a target adenosine. Beyond secondary structure, catalytic activity is also directed to specific adenosines because of enzymatic bias towards specific neighboring bases, namely a 5' uridine or adenine and a 3' guanine (Lehmann and Bass, 2000).

Two RNA editing strategies have been developed to redirect the $ADAR_{DD}$ towards a desired target adenosine for programmable A to I conversions. The first technology replaces the dsRBDs with a λN RNA binding peptide, which specifically recognizes the boxB hairpin fused to an RNA guide (Montiel-Gonzalez et al., 2013; Montiel-Gonzalez et al., 2016). The λN-boxB interaction is of nanomolar affinity, allowing for strong association of the $ADAR_{DD}$ with the guide RNA in cells. Genetically encoded plasmids carrying these components can be delivered to cells, allowing for around 20% editing in both epithelial cells and *Xenopus* oocytes (Montiel-Gonzalez et al., 2013). The guide design of this approach is complicated, however, limiting its use. An alternative approach uses a SNAP-tag to covalently link the $ADAR_{DD}$ to the guide RNA, which also has shown efficient editing, but has the limitation that a synthetic guide must be generated and linked to the protein (Stafforst and Schneider, 2012). The same group has more recently developed a genetically encoded version that uses the full ADAR2 enzyme and a unique hairpin on the guide RNA that the dsRBDs recognize (Wettengel et al., 2017). This approach achieves editing rates of up to 65% in cell culture, but can suffer from off-target editing since the dsRBDs are present and capable of recognizing their endogenous targets in cells. Neither approach has assayed transcriptome-wide off-targets, and so the true specificity of these technologies are hard to evaluate.

1.7 Nucleic acid detection technologies

While most CRISPR applications have focused on therapeutic genome editing or tools for studying cells, nucleic acid detection would benefit from the programmable and specific recognition of nucleic acid templates enabled by CRISPR proteins. Molecular nucleic acid testing (NAT) for infectious diseases and cancer are typically performed in central laboratories by skilled personnel due to the complexity of current systems (Niemz et al., 2011). The vision for point of care (POC) NAT would enable patient testing in hospital emergency rooms, primary care offices, at local clinics, and even at home for a variety of applications, including infectious disease, cancer, and genotyping (Figure 1.7). POC testing would also bring NAT to the developing world where central laboratories do not exist and skilled personnel are not available. Most current systems for clinical diagnostics are polymerase chain reaction (PCR)-based, can only be performed in labs by trained technicians, and require many hours for a result. A key barrier to PCR is the thermocycling steps that require complex instrumentation. Many iso-thermal nucleic acid amplification schemes have been developed in response to enable cheaper, quicker, and simpler tests, including recombinase polymerase amplification (RPA) and nicking enzyme amplification reaction (NEAR)(Niemz et al., 2011). These approaches replace the thermocycling steps with enzymatic components that allow for primer annealing and extension at a constant temperature of 37°C for RPA and 55°C for NEAR. Although these approaches are rapid, typically amplifying nucleic acids in 10-20 minutes, they suffer from poor specificity and primer bias, making multiplexing difficult. To overcome the specificity problem, a detection method must be coupled to these amplifications to allow for sequence-specific confirmation of the input target, either post-amplification or in real-time. Examples of detection methodologies include fluorescent oligonucleotide probes that are cleaved during the reaction, intercalating dyes, and riboswitch sensors that translate fluorescent proteins (Pardee et al., 2014; Pardee et al., 2016).

CRISPR enzymes could also serve as a sequence-specific detection method due to the specificity of the crRNA:target duplex recognition by the effector enzyme. To enable such a method however, the target recognition by the crRNA:enzyme complex must be coupled to some fluorescent readout.

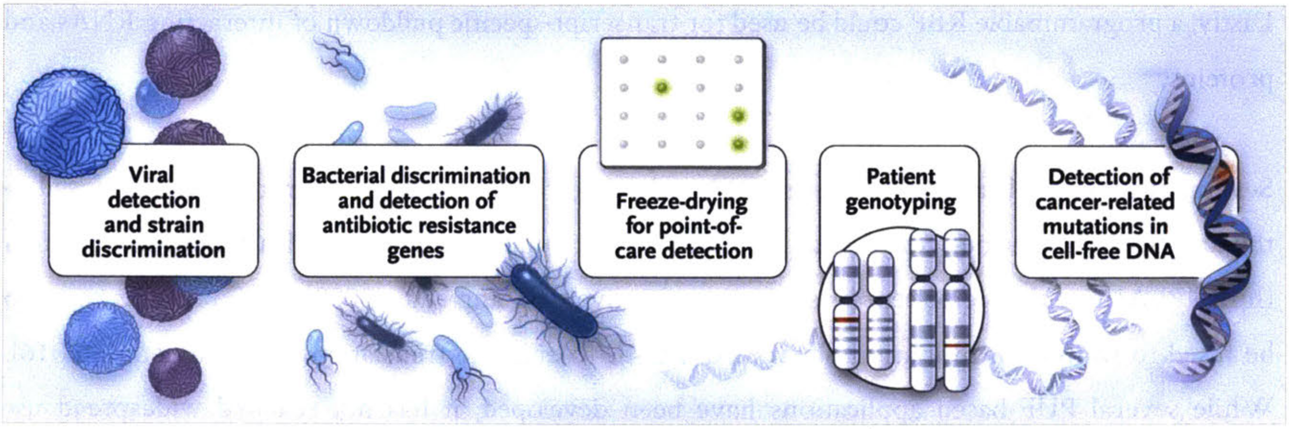

Viral detection and strain discrimination
Bacterial discrimination and detection of antibiotic resistance genes
Freeze-drying for point-of-care detection
Patient genotyping
Detection of cancer-related mutations in cell-free DNA

Figure 1.7: Molecular nucleic acid testing applications.

Examples of nucleic acid testing applications enabled by rapid isothermal amplification methods. Adapted from (Caliendo and Hodinka, 2017).

1.8 Other RNA-targeting applications

If a programmable RNA targeting enzyme could be identified and developed, a range of RNA tools could be realized (Figure 1.8). RNA cleavage would be the direct application of a natural programmable RNase, allowing for degradation of coding or noncoding transcripts. By catalytically inactivating the programmable RNase, an RNA binding platform would be possible with specific applications enabled by protein recruitment (Mackay et al., 2011). Translation could be stimulated by recruiting specific initiation factors, such as EIF4G, EIF4E, or EIF4A, to specific sites around the ribosome binding site (RBS) and could be repressed by blocking access to the RBS. Splicing of transcripts could also be modulated by recruiting certain splicing factors like A1 or RS domains allowing for exon inclusion or blocking a splice site for exon exclusion. By fusing to GFP, transcripts could be imaged in real-time and localized similar to applications with MS2, but without needing to modify the transcripts. Transcript localization could also be altered by changing the localization sequence on the programmable RBP. By fusing to RNA modifying enzymes, specific epitranscriptomic marks could be introduced in a site-specific manner, allowing elucidation of the role of many marks that have unknown function. For example, pseoduridine sites could be created by fusion to pseudouridine synthase 1-4 and methylation sites can be added to adenosines by methyltransferase-like 3. Additionally, RNA base editing enzymes, such as ADAR1/2 and APOPECs, could be recruited for site-specific editing of adenosine to inosine or cytosine to uridine, respectively. Lastly, a programmable RBP could be used for transcript-specific pulldown of interacting RNAs and proteins.

Several of these applications have been achieved using PUF proteins. PUF proteins can be fused to the arginine- and serine- rich domain of ASF or human heterogeneous nuclear ribonucleoprotein (hnRNP) A1 to enhance or repress splicing, respectively (Wang et al., 2009). PUF proteins can also be fused to translation initiation factors like EIF4E to enhance translation (Adamala et al., 2016). While several PUF-based applications have been developed, it has not reached widespread use because of the difficulty for most people to engineer and develop these constructs. Because of the limitations of existing tools, there remains a need for a genetically encodable, modular, and easily programmable RNA targeting platform. As more enzymes are identified to interact with RNA, more possibilities for RNA-targeting tools will emerge, allowing a robust toolbox for understanding RNA function in the cell. CRISPR based tools offer a potential platform for RNA tools if a natural RNA-targeting system could be discovered and tamed for mammalian application.

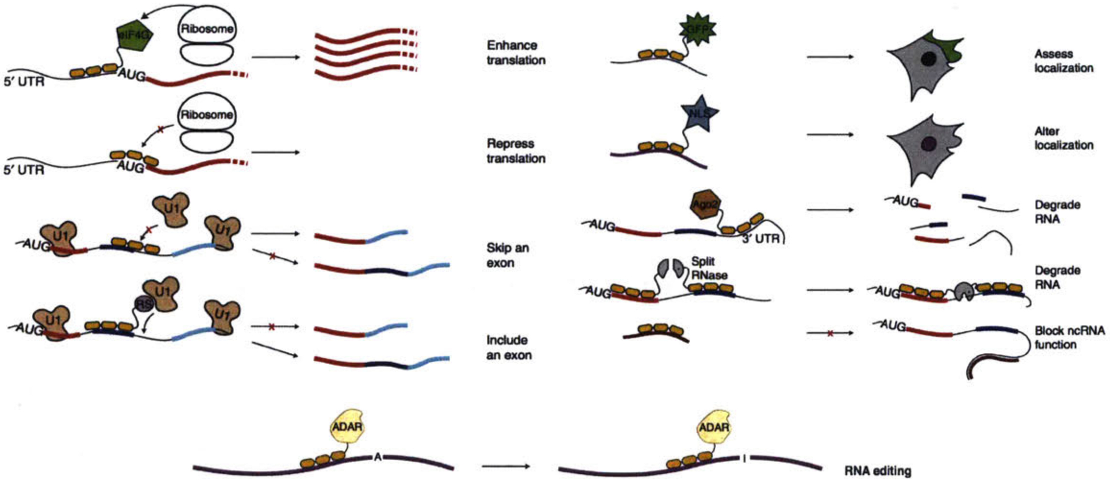


Figure 1.8: RNA targeting applications.

A variety of RNA perturbation tools can be developed with a programmable RNA binding protein. Partially reproduced from (Mackay et al., 2011).

In this thesis, I explore the evolutionary basis of CRISPR systems in the hope of expanding the framework and diversity of known CRISPR enzymes. We employ computational techniques to mine known bacterial genomes using signatures of CRISPR systems to expand the known set of Class 2 CRISPR systems (Chapter 2). In bacteria and biochemically, we characterize the Class 2 RNA-targeted RNA-guided CRISPR-Cas13 and uncover its unique mechanisms for programmable RNA cleavage (Chapter 3). Using its unique collateral RNase activity, we develop next generation molecular diagnostics for human health and agricultural applications (Chapters 4 and 5). Lastly, we transplant the Cas13 enzyme to mammalian cells and build RNA tools to allow for RNA knockdown and imaging (Chapter 6) as well as precise and efficient RNA base editing (Chapter 7). The series of work described here demonstrates the power of pursuing basic discovery of novel bacterial systems and great potential these new enzymes can have for building a comprehensive RNA toolbox with applications in basic science, biotechnology, therapeutics, and diagnostics.

Chapter 2

Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems

This chapter is adapted from the following article:

Shmakov, S.*, **Abudayyeh, O.O.***, Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., *et al.* (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell *60*, 385-397.

Contributions: Sergey Shmakov and Omar Abudayyeh are co-first authors (*). Eugene Koonin and Feng Zhang conceived the study; Kira Makarova, Yuri Wolf and Eugene Koonin designed the computational analyses; Sergey Shmakov, Kira Makarova, Yuri Wolf and Feng Zhang performed the computational analyses; Omar Abudayyeh, Jonathan Gootenberg, Feng Zhang and Konstantin Severinov designed the experiments; Omar Abudayyeh, Jonathan Gootenberg, Julia Joung, Silvana Konermann, Ekaterina Semenova and Leonid Minakhin performed the experiments; Omar Abudayyeh, Jonathan Gootenberg and Feng Zhang analyzed the results; Omar Abudayyeh, Kira Makarova, Feng Zhang and Eugene Koonin wrote the manuscript that was read, edited and approved by all authors.

2.1 Abstract

Microbial CRISPR-Cas systems are divided into Class 1, with multisubunit effector complexes, and Class 2, with single protein effectors. Currently, only two Class 2 effectors, Cas9 and Cpf1, are known. We describe here three distinct Class 2 CRISPR-Cas systems. The effectors of two of the identified systems, C2c1 and C2c3, contain RuvC-like endonuclease domains distantly related to Cpf1. The third system, C2c2, contains an effector with two predicted HEPN RNase domains. Whereas production of mature CRISPR RNA (crRNA) by C2c1 depends on tracrRNA, C2c2 crRNA maturation is tracrRNA independent. We found that C2c1 systems can mediate DNA interference in a 5'-PAM-dependent fashion analogous to Cpf1. However, unlike Cpf1, which is a single-RNA-guided nuclease, C2c1 depends on both crRNA and tracrRNA for DNA cleavage. Finally, comparative analysis indicates that Class 2 CRISPR-Cas systems evolved on multiple occasions through recombination of Class 1 adaptation modules with effector proteins acquired from distinct mobile elements.

2.2 Introduction

CRISPR (clustered regularly interspaced short palindromic repeat)-Cas (CRISPR-Associated proteins) are adaptive immune systems of archaea and bacteria (Barrangou and Marraffini, 2014; Koonin and Makarova, 2013; Marraffini and Sontheimer, 2010a). These systems have recently attracted much attention due to their unique, "Lamarckian" mode of action that retains "memories" from past infections and provides specific resistance to these infections via an RNA-guided process that has been harnessed to create powerful genome editing tools (Cho et al., 2013; Cong et al., 2013; Hwang et al., 2013; Jiang et al., 2013; Jinek et al., 2013; Mali et al., 2013b). The CRISPR-Cas systems show extreme diversity of Cas protein composition as well as genomic loci architecture (Makarova et al., 2011b; Makarova et al., 2015b).

Despite this diversity, CRISPR-Cas systems share a core set of features, indicative of a common origin. Most Cas proteins can be grouped into two main functional modules: the adaptation module, which delivers genetic material into CRISPR arrays to generate CRISPR RNAs (crRNAs), and the effector module, which, guided by crRNA, targets and cleaves invading nucleic acids (Makarova et al., 2011b;

Makarova et al., 2013). The adaptation modules are largely uniform across CRISPR-Cas systems and consist of two essential proteins, Cas1 and Cas2. By contrast, the effector modules show extreme variability. The latest classification of the CRISPR-Cas systems divides them into two classes based on the architecture of the effector modules (Figure 2.1A) (Makarova et al., 2015b). Class 1 systems, which encompass types I and III as well as the putative type IV, possess multi-subunit effector complexes comprised of multiple Cas proteins. Class 2 systems, which encompass type II and the putative type V, are characterized by effector complexes that consist of a single, large Cas protein.

The effector protein of type II CRISPR-Cas systems is Cas9, a large multidomain nuclease that ranges in size depending on the species from ~950 to over 1,600 amino acids and contains two nuclease domains, a RuvC-like (RNase H fold) domain and an HNH (McrA-like fold) domain (Makarova et al., 2006), for target DNA cleavage (Barrangou et al., 2007; Deltcheva et al., 2011; Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012; Sapranauskas et al., 2011). This multifunctional protein has been engineered into a key tool for genome editing. Recently, a second Class 2 effector protein, Cpf1, which contains a RuvC domain, but not an HNH domain (Makarova et al., 2015b; Schunder et al., 2013), has been shown to be an RNA-guided endonuclease that cleaves the target DNA via a staggered cut (Zetsche et al., 2015a). Based on their unique domain architecture, the Cpf1-containing systems have been categorized as type V CRISPR-Cas (Makarova et al., 2015b).

Although Class 2 systems are less common than Class 1 systems (Chylinski et al., 2014; Makarova et al., 2015b), it is likely that additional Class 2 systems, beyond those containing Cas9 and Cpf1 effector proteins, exist in the yet unexplored microbial diversity. Using a computational strategy, we identified three groups of candidate genomic loci encoding previously uncharacterized Class 2 variants. We experimentally demonstrate the functionality of two of the discovered systems, which have unique properties compared to Cas9. The characterization of these new systems provides evidence to suggest Class 2 systems originated by combination of Class 1 adaptation modules with effector proteins derived from different mobile elements.

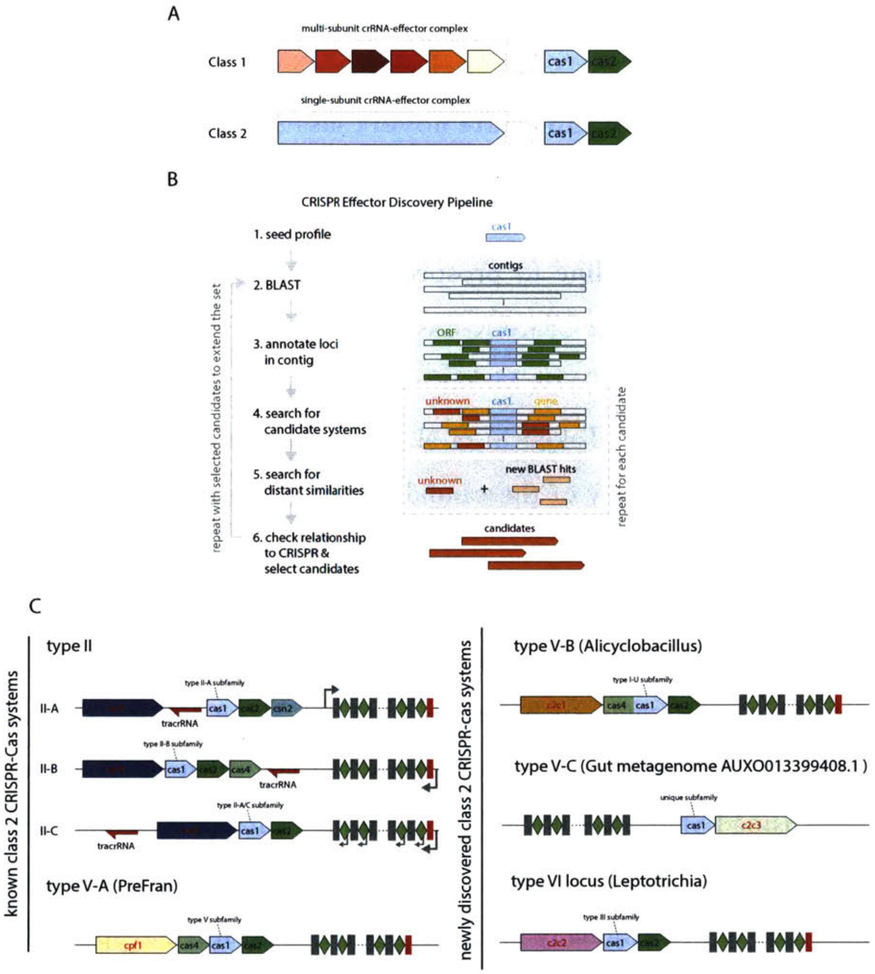

A
multi-subunit crRNA-effector complex
Class 1
cas1
cas2
single-subunit crRNA-effector complex
Class 2
cas1
cas2
B
CRISPR Effector Discovery Pipeline
1. seed profile
cas1
2. BLAST
contigs
3. annotate loci in contig
ORF
cas1
4. search for candidate systems
unknown
cas1
gene
5. search for distant similarities
unknown
new BLAST hits
6. check relationship to CRISPR & select candidates
candidates
repeat with selected candidates to extend the set
repeat for each candidate
C
known class 2 CRISPR-Cas systems
type II
II-A
type II-A subfamily
cas1
cas2
csn2
tracrRNA
II-B
type II-B subfamily
cas1
cas2
cas4
tracrRNA
II-C
type II-A/C subfamily
cas1
cas2
tracrRNA
type V-A (PreFran)
type V subfamily
cpf1
cas4
cas1
cas2
newly discovered class 2 CRISPR-cas systems
type V-B (Alicyclobacillus)
type I-U subfamily
c2c1
cas4
cas1
cas2
type V-C (Gut metagenome AUXO013399408.1)
unique subfamily
cas1
c2c3
type VI locus (Leptotrichia)
type III subfamily
c2c2
cas1
cas2

Figure 2.1: Prediction of candidate novel Class 2 CRISPR-Cas systems.

(A) Architectural principles of Class 1 (multi-protein effector complexes) and Class 2 (single-protein effector complexes) CRISPR-Cas systems.

(B) Schematic of the computational pipeline for identification of putative new Class 2 loci.

(C) Genomic architectures of the known and newly identified Class 2 CRISPR-Cas systems. The left panel shows the previously described three subtypes of type II and subtype V-A, and the right panel shows subtypes V-B and V-C, and type VI identified in this work. Subfamilies based on Cas1 are also indicated. The schematics include only the common genes represented in each subtype; the additional genes present in some variants are omitted. The red rectangle shows the degenerate repeat. The gray arrows show the direction of CRISPR array transcription. PreFran, *Prevotella-Francisella*.

See also Figure 2.S1.

2.3 Results

2.3.1 Computational prediction of candidate novel Class 2 CRISPR-Cas loci

We designed a computational pipeline to prospect the microbial genome sequence diversity to identify previously undetected Class 2 CRISPR-Cas loci (Figure 1B). Because most CRISPR-Cas loci include a *cas1* gene (Makarova et al., 2011b; Makarova et al., 2015b) and the Cas1 sequence is the most conserved among all Cas proteins (Takeuchi et al., 2012), we used *cas1* as the anchor to identify candidate loci. A substantial majority of the candidate CRISPR-Cas loci identified by the pipeline could be assigned to known subtypes (Chylinski et al., 2013; Chylinski et al., 2014; Fonfara et al., 2014; Makarova et al., 2011b; Makarova et al., 2015b). To identify novel Class 2 systems, we focused on unclassified candidate CRISPR-Cas loci containing long proteins (>500 aa) given that the presence of large single-subunit effector proteins, such as Cas9 and Cpf1, is the diagnostic feature of type II and type V systems, respectively. Based on this criterion, we identified 63 candidate loci that were analyzed individually using PSI-BLAST and HHpred. The protein sequences encoded in the candidate loci were used as queries to search metagenomic databases for additional homologs. In total, we discovered 53 loci (some of the originally identified 63 were discarded as spurious whereas several incomplete loci that lacked *cas1* were added) with characteristic features of Class 2 CRISPR-Cas systems that could be classified into three distinct groups based on the nature of the putative effector proteins (Figure 2.1C and Figure 2.S1).

The first group (Figure 2.1C and Figure 2.S1A), provisionally denoted C2c1 (Class 2 candidate 1), is represented in 18 bacterial genomes from four major taxa: *Bacilli*, *Verrucomicrobia*, α-proteobacteria, and δ-proteobacteria (Figure 2.S1A). The C2c1 loci encode a Cas1-Cas4 fusion, Cas2, and a large putative effector protein, and typically are adjacent to a CRISPR array (Figure 2.S1A). The loci in the second group include solely metagenomic sequences and thus could not be assigned to specific taxa. These loci encode only Cas1 and a large putative effector protein denoted C2c3 (Class 2 candidate 3; although the candidates were designated in the order of discovery, throughout the text, we juxtapose C2c1 and C2c3 as they contain distantly related effector proteins, discussed below) (Figure 2.1C and

Figure 2.S1B). The third group, denoted C2c2 (Class 2 candidate 2), was identified in 21 genomes from five major bacterial taxa: □-proteobacteria, *Bacilli, Clostridia, Fusobacteria,* and *Bacteroidetes* (Figure 2.1C and Figure 2.S1C). These loci encompass a large protein with no sequence similarity to C2c1, Cpf1, or Cas9. Although under our computational strategy, the originally identified C2c2 loci encompassed *cas1* and *cas2,* subsequent searches showed that the majority consists only of the *c2c2* gene and a CRISPR array (Figure 2.S1C). Such apparently incomplete loci could either encode defective CRISPR-Cas systems or might function with the adaptation module encoded elsewhere in the genome, as observed for some type III systems (Majumdar et al., 2015).

Typically, the sequence and structure of repeats in CRISPR arrays strongly correlate with the sequence of the respective Cas1 protein, which interacts with the repeats during spacer acquisition. However, despite the high similarity of the C2c1 system Cas1 proteins to each other, the CRISPR in the respective arrays are highly heterogeneous. All the repeats are 36-37 bp long and can be classified as unstructured. Among the C2c3 loci, only one contains a CRISPR array with unusually short, 17-18 nt spacers. The repeats in this array are 25 bp long and appear to be unstructured. The CRISPR arrays of the C2c2 loci are also highly heterogeneous (repeat length ranging from 35 to 39 bp) and unstructured.

Although bacteriophages infecting bacteria that harbor these newly discovered Class 2 CRISPR-Cas systems are virtually unknown, for each of these systems, we detected spacers that matched phages or predicted prophages. Although the majority of the spacers were not significantly similar to any available sequences, the existence of spacers matching phage genomes implies that at least some of these loci encode active, functional adaptive immunity systems. The low fraction of phage-specific spacers is typical of CRISPR-Cas systems and most likely reflects their dynamic evolution and the small fraction of virus diversity that is currently known. This interpretation is compatible with the observation that closely related bacterial strains encoding homologous CRISPR-Cas loci, e.g. the C2c2 loci from *Listeria weihenstephanensis* and *Listeria newyorkensis,* typically contain unrelated collections of spacers (Figure 2.S2)

2.3.2 C2c1 and C2c3 proteins contain RuvC-like nuclease domains and have a domain architecture resembling Cpf1

The lengths of C2c1 and C2c3 proteins range from ~1100 to ~1500 amino acids, similar to the typical lengths of Cas9 and Cpf1. Analogous to the previous findings for Cas9 and Cpf1 (Chylinski et al., 2014; Makarova and Koonin, 2015; Makarova et al., 2015b), the C-terminal regions of the C2c1 and C2c3 proteins are significantly similar to a subset of TnpB proteins encoded by transposons of the *IS605* family (Figure 2.2A and Figure 2.S3). However, in database searches, only C2c3 showed limited but significant similarity to Cpf1 within the TnpB homology regions, whereas C2c1 was not significantly similar to any of the other known or putative Class2 effector proteins. Moreover, the subsets of the TnpB proteins with significant similarity to the known (Cas9 and Cpf1) and putative (C2c1 and C2c3) Class 2 effectors did not overlap (Figure 2.2A and Figure 2.S3), suggesting that Cas9, Cpf1, C2c1, and C2c3 evolved independently from distinct transposable elements.

The TnpB homology regions of C2c1 and C2c3 contain the three catalytic motifs of the RuvC-like nuclease (Aravind et al., 2000), the region corresponding to the arginine-rich bridge helix, which is involved in crRNA-binding by Cas9, and a counterpart to the Zn finger of TnpB (the Zn-binding cysteine residues are conserved in C2c3 but are missing in the majority of Cpf1 and C2c1 proteins; Cpf1 and C2c1 contain multiple insertions and deletions in this region suggestive of functional divergence) (Figure 2.2A; Figures 2.S4 and 2.S5). The conservation of the catalytic residues implies that the RuvC homology domains of all these proteins are active nucleases. The N-terminal regions of C2c1 and C2c3 show no significant similarity to each other or any known proteins. Secondary structure predictions indicate that both these regions adopt a mixed □□□□conformation (Figures 2.S4 and 2.S5). Thus, the overall domain architectures of C2c1 and C2c3, and in particular the organization of the RuvC domain, resemble Cpf1 but are distinct from Cas9 (Figure 2.2A). Accordingly, we propose that the C2c1 and C2c3 loci are best classified as subtypes V-B and V-C, respectively, with Cpf1-encoding loci now designated subtype V-A.

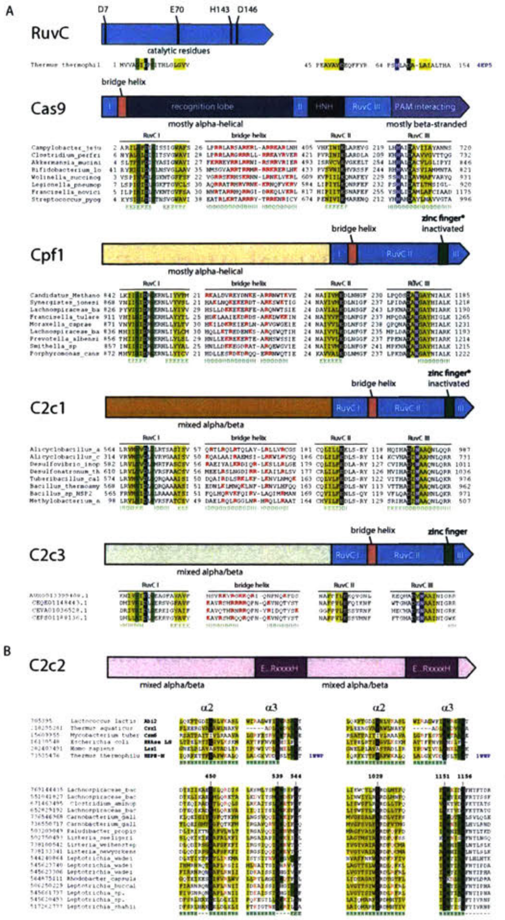


Figure 2.2: Domain architectures and conserved motifs of the Class 2 proteins.

(A) Types II and V: TnpB-derived nucleases. The top panel shows the RuvC nuclease from *Thermus thermophilus* (PDB ID: 4EP5) with the catalytic amino acid residues denoted. Underneath each domain architecture, an alignment of the conserved motifs in selected representatives of the respective protein family (a single sequence for RuvC) is shown. The catalytic residues are shown by white letters on a black background; conserved hydrophobic residues are highlighted in yellow; conserved small residues are highlighted in green; in the bridge helix alignment, positively charged residues are in red. Secondary structure prediction is shown underneath the aligned sequences: H denotes □-helix and E denotes extended conformation (□-strand). See also Figures 2.S4 and 2.S5.

(B) Type VI: predicted RNases containing two HEPN domains. The top alignment blocks include selected HEPN domains described previously and the bottom blocks include the catalytic motifs from the putative type VI effector proteins. The designations are as in (A).

2.3.3 C2c2 contains two HEPN domains and is predicted to possess RNase activity

Database searches detected no significant sequence similarity between C2c2 and any known proteins. However, inspection of multiple alignments of C2c2 protein sequences revealed two conserved R(N)xxxH motifs that are characteristic of HEPN (Higher Eukaryotes and Prokaryotes Nucleotide-binding) domains (Anantharaman et al., 2013; Grynberg et al., 2003). Additionally, a conserved glutamate embedded in a strongly predicted long □-helix and corresponding to the similar motif of HEPN domains was identified (Figure 2.2B; Figure 2.S6). The HEPN superfamily includes small (~150 aa) □-helical domains with extremely diverse sequences but highly conserved catalytic motifs shown or predicted to possess RNase activity (Anantharaman et al., 2013). Searching the Pfam database using the HHpred program and the C2c2 sequences as queries detected similarity to HEPN domains for both putative nuclease domains of C2c2 albeit not at a highly significant level. Importantly, however, these were the only HHpred-generated alignments in which the R(N)xxxH motifs were conserved. The identification of HEPN domains in C2c2 proteins is further supported by secondary structure predictions, which indicate that each motif is located within compatible structural contexts, and the predicted □-helical secondary structure of each putative domain is consistent with the HEPN fold (Figure 2.2B; Figure 2.S6). Outside of the two HEPN domains, the C2c2 sequence is predicted to adopt a mixed □□□ structure without discernible similarity to any known protein folds (Figure 2.S6). Given the unique predicted effector of C2c2, these systems qualify as a putative type VI CRISPR-Cas.

2.3.4 The candidate Class 2 CRISPR-Cas loci are expressed to produce mature crRNAs and encode putative tracrRNAs

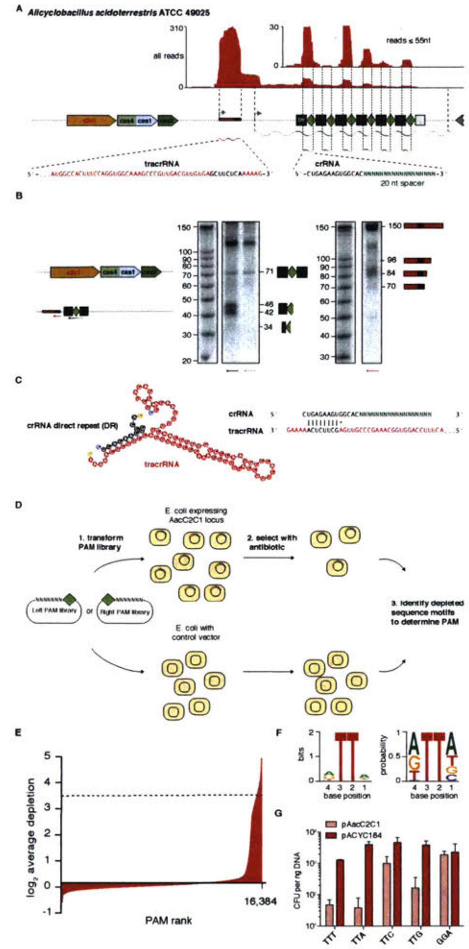

A
Alicyclobacillus acidoterrestris ATCC 49025
310
all reads
0
30
0
reads ≤ 55nt
c2c1
cas4
cas1
cas2
tracrRNA
5'-...AUGGCCACUUUCCAGGUGGCAAAGCCCGUUGACGUUGUGAGCUUCUCAAAAAG-3'
crRNA
5'-CUGAGAAGUGGCACNNNNNNNNNNNNNNNNNNNN-3'
20 nt spacer
B
150
100
90
80
70
60
50
40
30
20
71
46
42
34
150
100
90
80
70
60
50
40
30
150
96
84
70
C
crRNA direct repeat (DR)
tracrRNA
crRNA 5' CUGAGAAGUGGCACNNNNNNNNNNNNNNNNNNNN 3'
tracrRNA 3' GAAAAACUCUUCGAGUUGCCCGAAACGGUGGACCUUUCA...5'
D
E. coli expressing AacC2C1 locus
1. transform PAM library
2. select with antibiotic
Left PAM library
or
Right PAM library
3. identify depleted sequence motifs to determine PAM
E. coli with control vector
E
5
4
3
2
1
0
-1
log₂ average depletion
16,384
PAM rank
F
bits
2
1
0
4 3 2 1
base position
probability
1
0.5
0
4 3 2 1
base position
G
pAacC2C1
pACYC184
CFU per ng DNA
TTT
TTA
TTC
TTG
GGA

Figure 2.3: Functional validation of the *Alicyclobacillus acideoterrestris* C2c1 locus.

(A) RNA-sequencing shows the *A. acideoterrestris* C2c1 locus is highly expressed in the endogenous system, with processed crRNAs incorporating a 5' 14-nt DR and 20-nt spacer. A putative 79-nt tracrRNA is expressed robustly in the same orientation as the *cas* gene cluster (see also Figures 2.S7B and 2.S7C).

(B) Northern blot of RNAs expressed from endogenous locus (M) and a minimal first-spacer array (S) show processed crRNAs with a 5' DR and the presence of a small putative tracrRNA. Arrows indicate the probe positions and their directionality.

(C) *In silico* co-folding of the crRNA direct repeat and putative tracrRNA shows stable secondary structure and complementarity between the two RNAs. 5' bases are colored blue and 3' bases are colored orange (see also Figure 2.S7D).

(D) Schematic of the PAM determination screen.

(E) Depletion from the 5' left PAM library reveals a 5' TTN PAM. Depletion is measured as the negative $\log_2$ fold ratio and PAMs above a threshold of 3.5 are used to calculate the entropy score at each position.

(F) Sequence logo for the AacC2c1 PAM as determined by the plasmid depletion assay. *Left:* Letter height at each position is measured by entropy scores and error bars show the 95% Bayesian confidence interval. *Right:* Letter height at each position is measured by the relative frequency of the nucleotide (see also Figure 2.S7E).

(G) Validation of the AacC2c1 PAM by measuring interference with 8 different PAMs. PAMs matching the TTN motif show depletion as measured by cfus.

In addition to the adaptation and interference protein modules, type II, Cas9-based systems also use a small non-coding *trans*-activating CRISPR RNA (tracrRNA), which is typically encoded adjacent to the *cas* operon. The tracrRNA is partially complementary to repeat portions of the respective CRISPR array transcript (pre-crRNA) and is essential for its processing into crRNA which is catalyzed by RNase III recognizing the repeat-anti-repeat duplex (Chylinski et al., 2013; Chylinski et al., 2014; Deltcheva et al., 2011). We investigated whether the loci encoding Class 2 systems identified here also contain small RNAs with complementarity to cognate CRISPR repeats. We chose a representative C2c1 system from *Alicyclobacillus acidoterrestris* ATCC 49025 (Aac) for initial characterization and conducted whole-transcriptome RNA sequencing (RNA-seq) and Northern blotting to map transcription of small RNAs associated with the C2c1 locus. The CRISPR array was found to be actively transcribed in the same orientation as the *cas* gene cluster and shows robust processing of crRNAs that are 34 nt in length, with a 5' 14-nt direct repeat (DR) and a 20-nt spacer (Figure 2.3A). We also identified an abundant 79-nt small RNA encoded between the *cas2* gene and the CRISPR array and transcribed in the same orientation as the CRISPR array (Figure 2.3A, B). The internal region of this RNA contains a sequence complementary to the processed CRISPR repeat sequence (anti-repeat), suggesting that this transcript is the tracrRNA. *In silico* co-folding of the processed 14-nt CRISPR repeat with this putative tracrRNA predicts a stable secondary structure (Figure 2.3C).

Given that the putative tracrRNA in *A. acidoterrestris* contains a characteristic anti-repeat sequence, we sought to predict potential tracrRNAs for the rest of the identified C2c1, C2c2, and C2c3 loci by searching for anti-repeat sequences within each locus. In many CRISPR-Cas loci, the repeat located at the promoter-distal end of the CRISPR array is degenerate and has a sequence that is distinct from the rest of the repeats (Biswas et al., 2014). Such degenerate repeats were detected in several C2c1 and C2c2 systems (Figure 2.S1), allowing us to predict the direction of the array transcription. By integrating this information, we identified putative tracrRNAs in 4 of the 13 C2c1 and 4 of the 17 C2c2 loci (Figures 2.S1 and 2.S7A). However, in some subtype II-B and II-C loci, the CRISPR array is transcribed in the opposite direction, starting from the degenerate repeat (Sampson et al., 2013; Zhang et al., 2013). Accordingly, we attempted to predict the tracrRNA in different positions with respect to the CRISPR array but were unable to identify additional candidate tracrRNA sequences. However, not all Class 2 CRISPR systems require tracrRNA for crRNA maturation or effector function, as demonstrated by the Cpf1 systems (Zetsche et al., 2015a). Effectively identical patterns of RNA expression and processing were observed when the Aac C2c1 locus was expressed in the heterologous *E. coli* system (Figure 2.S7B).

Given the robust expression of the Aac locus and the identification of processed tracrRNA and crRNAs, we designed an interference screen to determine if the Aac C2c1 loci are active and to identify the protospacer adjacent motif (PAM), which in type II systems dictates where the effector protein will cleave (Figure 2.3D). Whereas the 3' PAM screen showed no significant depletion of PAMs, the 5' PAM library screening resulted in the identification of 364 significantly depleted PAMs (> 3.5 $\log_2$ fold depletion) (Figure 2.3E) which all had the sequence NNNNTTN (Figure 2.3F). Although there was a slight preference for bases other than C in the 5' position immediately adjacent to the protospacer, these results indicate that the 5' TTN motif is likely recognized by the AacC2c1 complex. We validated the proposed PAM using the first spacer of the AacC2c1 locus and all four TTN PAMs. The results of this experiment confirm that a 5' TTN PAM is necessary for interference and that interference is slightly reduced with the 5'TTC PAM (Figure 2.3G).

2.3.5 C2c1 is a dual-RNA-guided DNA endonuclease

We then sought to investigate whether C2c1 is an RNA-guided endonuclease, and to determine its RNA substrate requirements. We assayed *in vitro* DNA cleavage by incubating target DNA with protein lysate from human 293FT cells expressing C2c1 and *in vitro* transcribed crRNA and putative tracrRNA (Figure 2.4A). We designed crRNAs corresponding to the mature processed form that consisted of a 22-nt DR followed by a 20-nt spacer targeting a sequence from the human *EMX1* locus. To test cleavage of the *EMX1* target DNA, we used PCR to amplify a ~600 bp fragment containing the same DNA target site as the *EMX1*-targeting crRNA. *A. acidoterrestris* optimally grows at 50°C (Chang and Kang, 2004), and we observed most efficient AacC2c1-mediated RNA-guided, crRNA-specific and tracrRNA-dependent cleavage of the target DNA at 50°C (Figure 2.4B).

Because RNA-seq experiments identified putative tracrRNA transcripts of variable size (Figure 2.3A), we tested a series of 3'-truncated tracrRNAs and found that the shortest tracrRNA capable of supporting RNA-guided cleavage using C2c1 cell lysate was 78-nt in length (Figure 2.4C). Using this minimal tracrRNA, we showed that 50°C is indeed the optimal cleavage temperature and that there is no observable cleavage below 40°C (Figure 2.4D). To further validate the PAM requirements of C2c1, we designed a second crRNA targeting the protospacer-1 of the endogenous AacC2c1 CRISPR array (Figure 2.3F) and found that linear DNA molecules containing protospacer-1 preceded by TTT, TTA, and TTC PAMs but not GGA were efficiently cleaved (Figure 2.4E).

Given the demonstration that AacC2c1 is a dual-RNA-guided endonuclease, we hypothesized that, similar to Cas9 (Jinek et al., 2012), the C2c1 crRNA:tracrRNA duplex could be simplified into a single-guide RNA (sgRNA) by fusing the 3' end of the 78-nt tracrRNA with the 5' end of the crRNA (Figure 2.4F). Target cleavage activity similar to that obtained with the crRNA:tracrRNA duplex was observed for the sgRNA with both the *EMX1* and protospacer-1 plasmid targets (Figure 2.4G). Thus, these experiments demonstrate that the lysate of human cells expressing C2c1 can cleave target DNA, identify the temperature optimum of the enzyme and demonstrate the requirement for a crRNA:tracrRNA duplex and 5' PAM for AacC2c1 nuclease activity, in contrast to Cas9 which requires a 3' PAM (Jinek et al., 2012; Mojica et al., 2009) .

To validate the results obtained with heterologous expression and expand the findings to other type V-B systems, we screened the C2c1 locus from *Bacillus thermoamylovorans* (Bth). Whole-transcriptome sequencing of a synthesized BthC2c1 locus cloned into pET-28 in *E. coli* revealed strong processing of both spacers present in the array, as well as expression of a 91-nt RNA (Figure 2.S7C) that displayed secondary structure and repeat-anti-repeat base-paring similar to the putative Aac tracrRNA (Figure 2.S7D). To test for interference, we transformed the PAM library with the corresponding spacer into *E. coli* harboring the BthC2c1 locus and compared depletion to pET-28. In agreement with the results obtained for AacC2c1, this screen showed that BthC2c1 employs a 5' PAM with the consensus sequence ATTN (Figure 2.S7E).

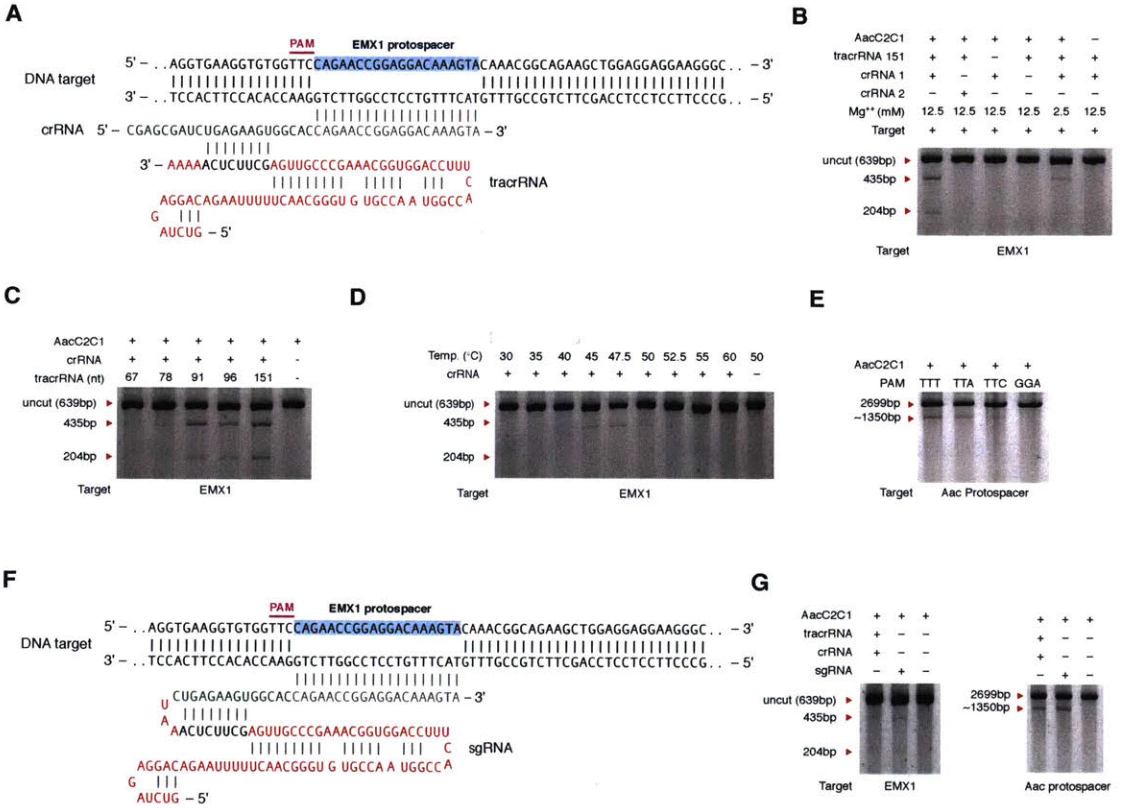

A
PAM
EMX1 protospacer
5' – . . AGGTGAAGGTGTGGTTCCAGAACCGGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGC . . – 3'
DNA target
3' – . . TCCACTTCCACACCAAGGTCTTGGCCTCCTGTTTCATGTTTGCCGTCTTCGACCTCCTCCTTCCCG . . – 5'
crRNA 5' – CGAGCGAUCUGAGAAGUGGCACCAGAACCGGAGGACAAAGTA – 3'
3' – AAAAACUCUUCGAGUUGCCCGAAACGGUGGACCUUU
C
tracrRNA
AGGACAGAAUUUUUCAACGGGU G UGCCA A UGGCCA
G
AUCUG – 5'
B
AacC2C1 + + + + + –
tracrRNA 151 + + – + + +
crRNA 1 + – + – + +
crRNA 2 – + – – – –
Mg++ (mM) 12.5 12.5 12.5 12.5 2.5 12.5
Target + + + + + +
uncut (639bp)
435bp
204bp
Target EMX1
C
AacC2C1 + + + + + +
crRNA + + + + + -
tracrRNA (nt) 67 78 91 96 151 -
uncut (639bp)
435bp
204bp
Target EMX1
D
Temp. (°C) 30 35 40 45 47.5 50 52.5 55 60 50
crRNA + + + + + + + + + –
uncut (639bp)
435bp
204bp
Target EMX1
E
AacC2C1 + + + +
PAM TTT TTA TTC GGA
2699bp
~1350bp
Target Aac Protospacer
F
PAM
EMX1 protospacer
5' – . . AGGTGAAGGTGTGGTTCCAGAACCGGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGC . . – 3'
DNA target
3' – . . TCCACTTCCACACCAAGGTCTTGGCCTCCTGTTTCATGTTTGCCGTCTTCGACCTCCTCCTTCCCG . . – 5'
U CUGAGAAGUGGCACCAGAACCGGAGGACAAAGTA – 3'
A AACUCUUCGAGUUGCCCGAAACGGUGGACCUUU
C
sgRNA
AGGACAGAAUUUUUCAACGGGU G UGCCA A UGGCCA
G
AUCUG – 5'
G
AacC2C1 + + +
tracrRNA + – –
crRNA + – –
sgRNA – + –
uncut (639bp)
435bp
204bp
Target EMX1
+ + +
+ – –
+ – –
– + –
2699bp
~1350bp
Aac protospacer

Figure 2.4: Characterization of the cleavage requirements of *A. acideoterrestris* C2c1.

(A) Schematic of the AacC2c1 crRNA and tracrRNA design hybridizing to the EMX1 target site.

(B) *In vitro* cleavage of the EMX1 target with the human cell lysate expressing AacC2c1 shows that *in vitro* targeting of AacC2c1 is robust and depends on tracrRNA. Non-targeting crRNA (crRNA 2) fails to cleave the EMX1 target, whereas crRNA 1 targeting EMX1 enabled strong cleavage in the presence of Mg++ and weak cleavage in the absence of Mg++.

(C) *In vitro* cleavage of the EMX1 target in the presence of a range of tracrRNA lengths identifies the 78 nt species as the minimal tracrRNA form, with increased cleavage efficiency for the 91nt form.

(D) Analysis of the temperature dependency of the *in vitro* cleavage of the EMX1 target shows that the optimal temperature range of robust AacC2c1 cleavage is between 40°C and 55°C

(E) *In vitro* validation of the AacC2c1 PAM requirements with four different PAMs. The PAMs matching the TTN motif are efficiently cleaved.

(F) Schematic of the chimeric AacC2c1 sgRNA shown hybridized to the EMX1 DNA target with repeat:anti-direct pairing between segments derived from the tracrRNA (red) and the crRNA (green)

(G) Comparison of the *in vitro* target cleavage in the presence of crRNA-tracrRNA AacC2c1 and sgRNA identifies comparable cleavage efficiencies.

2.3.6 Type VI C2c2 systems produce mature crRNA without tracrRNA

Using a similar approach, we investigated the functionality of the C2c2 loci. We synthesized the C2c2 locus of *Listeria seeligeri serovar* 1/2b str. SLCC3954 (Lse) and expressed it in *E. coli*. We observed a high level of expression and the formation of crRNAs with a 5' 29-nt DR and 15-18-nt spacers (Figure 2.5A). In contrast to the C2c1 loci, although this C2c2 locus contains a predicted tracrRNA (Figure 2.S1C), we did not observe its expression (Figure 2.5A). Thus, the secondary structure present in the pre-crRNA of this C2c2 locus could be sufficient for processing yielding the mature crRNA as well as crRNA loading onto the C2c2 protein. The RNA-folding of the processed crRNA shows a strongly predicted stem-loop within the direct repeat that might serve as a handle for the C2c2 protein (Figure 2.5A). In addition, we expressed the *Leptotrichia shahii* str. SLCC3954 C2c2 locus in *E. coli* and found that the CRISPR array is expressed and processed into 44-nt crRNAs (Figure 2.5B). We then used RNAseq to compare the expression of the *L. shahii* C2c1 locus in the endogenous and heterologous systems and in both cases, detected abundant, mature crRNA species but no tracrRNA (Figure 2.S7F,G). An additional, uncharacterized small RNA was expressed in the vicinity of the CRISPR array in *L. shahii* (Figure 2.S7F) but not in *E. coli* cells (Figure 2.S7G). *In silico* folding of the crRNA predicted secondary structure that was highly similar to that in *L. seeligeri* (Figure 2.S7F). However, co-folding with the highly expressed small RNA showed no stable structure or significant complementarity (not shown). The functional relevance of this RNA species in the C2c2 system remains to be determined.

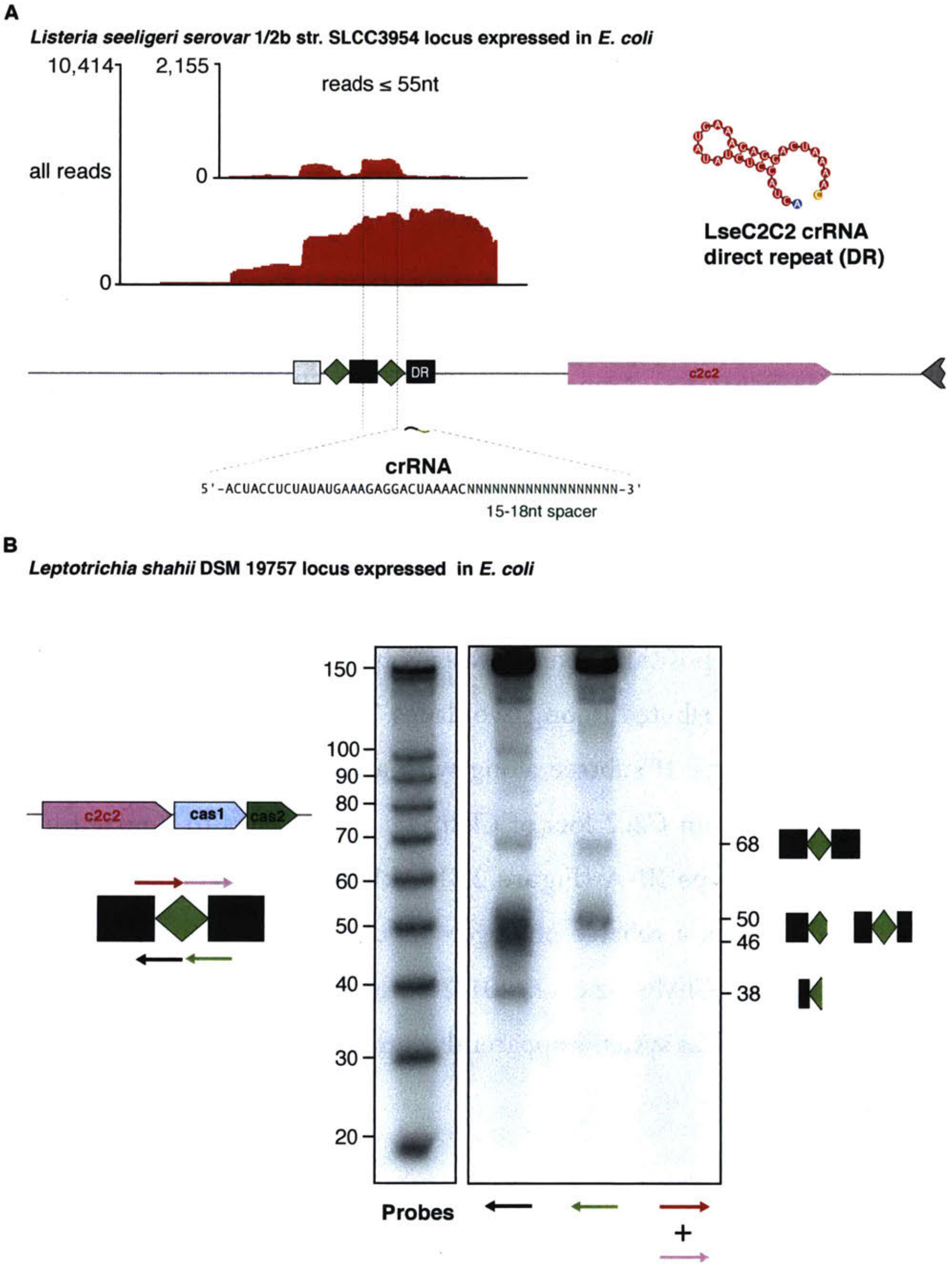


Figure 2.5: Expression and processing of C2c2 loci.

(A) RNA-sequencing of the *Listeria seeligeria serovar* 1/2b str. SLCC3954 C2c2 locus (see also Figures 2.S7F and 2.S7G).

(B) Northern blot analysis of the *Leptotrichia shahii* DSM 19757 shows processed crRNAs with a 5' DR. Arrows indicate the probe positions and their directionality.

2.3.7 The adaptation modules of distinct Class 2 systems evolved independently from different divisions of Class 1 systems

Cas1 is the most conserved Cas protein (Takeuchi et al., 2012) and the only one for which comprehensive phylogenetic analysis is feasible (Makarova et al., 2011b; Makarova et al., 2015b). In the phylogenetic tree of Cas1, putative subtype V-B (C2c1) is largely monophyletic and confidently clusters with type I-U (Figure 2.6). Among all the (putative) CRISPR-Cas loci, only type I-U and C2c1 encode a Cas1-Cas4 fusion. This derived shared character, together with the phylogenetic affinity of Cas1, indicates that the adaptation module of subtype V-B derives from that of type I-U. The type V-C Cas1 is the most diverged variant of Cas1 sequences discovered to date as indicated by the long branch in the phylogenetic tree (Figure 2.6). In the Cas1 tree, the type V-C branch is inside subtype I-B (Figure 2.6) although the position of such a fast evolving group should be taken with caution. The type VI Cas1 proteins are distributed among two clades. The first clade includes Cas1 from *Leptotrichia* and is located within the type II subtree along with a small type III-A branch. The second clade consists of Cas1 proteins from C2c2 loci of *Clostridia* and belongs to a mixed branch that mostly contains Cas1 proteins of type III-A (Figure 2.6). Although Cas2 is a small and relatively poorly conserved protein, for which a reliable phylogeny is difficult to obtain, all available data point to coevolution of *cas1* and *cas2* (Chylinski et al., 2014; Norais et al., 2013). Thus, the adaptation modules of the new Class 2 CRISPR-Cas systems apparently come from different variants of Class 1.

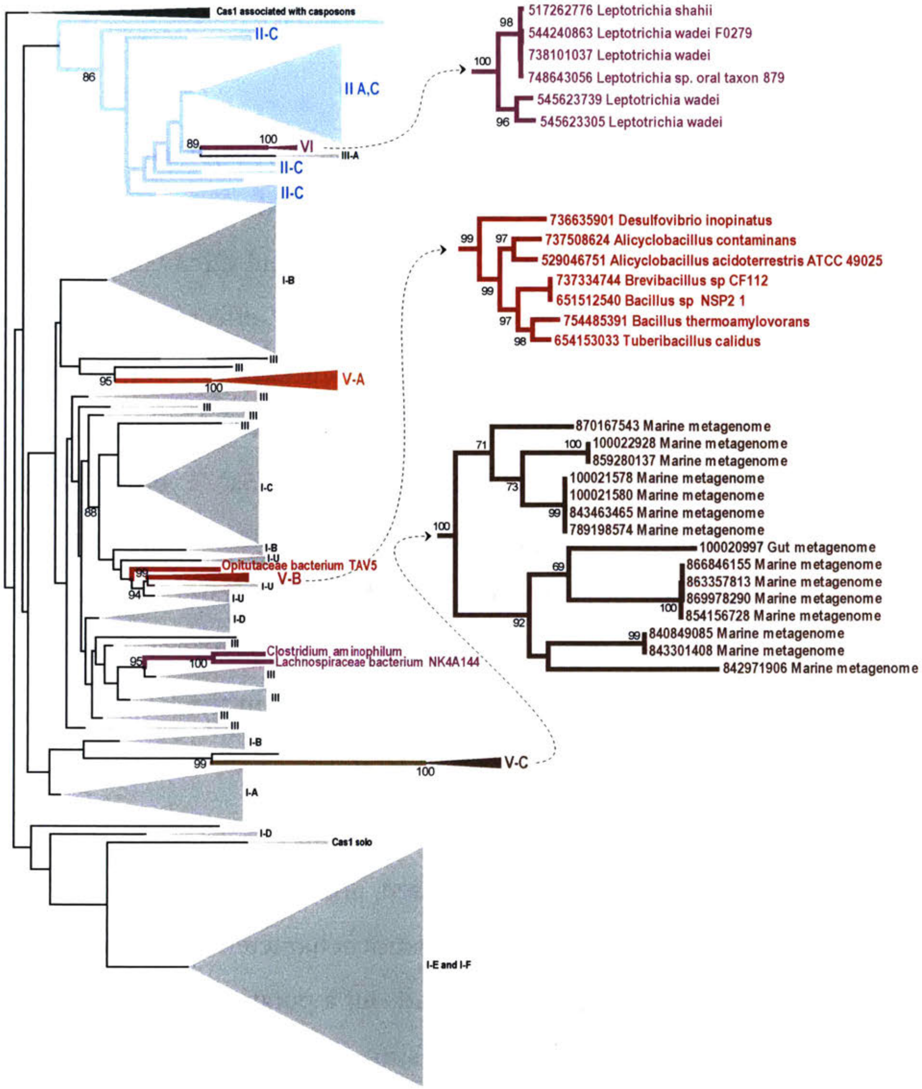

Cas1 associated with casposons
II-C
86
II A,C
89
100
VI
III-A
II-C
II-C
I-B
III
III
95
V-A
100
III
III
III
III
I-C
88
I-B
I-U
Opitutaceae bacterium TAV5
99
V-B
I-U
94
I-U
I-D
III
Clostridium aminophilum
Lachnospiraceae bacterium NK4A144
95
100
III
III
III
III
I-B
99
100
V-C
I-A
I-D
Cas1 solo
I-E and I-F
98
100
96
517262776 Leptotrichia shahii
544240863 Leptotrichia wadei F0279
738101037 Leptotrichia wadei
748643056 Leptotrichia sp. oral taxon 879
545623739 Leptotrichia wadei
545623305 Leptotrichia wadei
99
97
99
97
98
736635901 Desulfovibrio inopinatus
737508624 Alicyclobacillus contaminans
529046751 Alicyclobacillus acidoterrestris ATCC 49025
737334744 Brevibacillus sp CF112
651512540 Bacillus sp NSP2 1
754485391 Bacillus thermoamylovorans
654153033 Tuberibacillus calidus
100
71
100
73
99
69
100
92
99
870167543 Marine metagenome
100022928 Marine metagenome
859280137 Marine metagenome
100021578 Marine metagenome
100021580 Marine metagenome
843463465 Marine metagenome
789198574 Marine metagenome
100020997 Gut metagenome
866846155 Marine metagenome
863357813 Marine metagenome
869978290 Marine metagenome
854156728 Marine metagenome
840849085 Marine metagenome
843301408 Marine metagenome
842971906 Marine metagenome

Figure 2.6: Phylogenetic tree of Cas1.

The tree was constructed from a multiple alignment of 1498 Cas1 sequences which contained 304 phylogenetically informative positions. Branches, corresponding to Class 2 systems are highlighted: cyan, type II; orange, subtype V-A; red, subtype V-B; brown, subtype V-C; purple, type VI. Insets show the expanded branches of the novel (sub)types. The bootstrap support values are given as percentage points and shown only for few relevant branches. The complete tree with species names and bootstrap support values is available in Figure S7; the underlying alignment is available at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Class2/ (FASTA format).

See also Supplemental Experimental Procedures.

2.4 Discussion

Despite intense efforts to characterize the CRISPR-Cas systems, major aspects of the basic biology, diversity, and evolution of this remarkable defense strategy remain unknown. We describe here the discovery of three distinct Class 2 CRISPR-Cas systems, C2c1 and C2c3 (subtypes of the previously described putative type V), and C2c2 (putative type VI). Type V effector proteins resemble Cas9 in their overall domain architecture but contain only a single nuclease domain, the RuvC-like domain. The type V effector Cpf1 was recently shown to cleave double-stranded DNA, indicating that these enzymes use a different mechanism than Cas9 (Zetsche et al., 2015a). Type VI CRISPR-Cas systems contain a unique effector protein with two predicted HEPN domains, which typically possess RNAse activity (Anantharaman et al., 2013), suggesting that they might target and cleave mRNA. RNA cleavage has been reported for certain type III CRISPR-Cas systems (Hale et al., 2014; Hale et al., 2009; Peng et al., 2015). Alternatively, C2c2 could be the first DNase in the HEPN superfamily, perhaps with the two HEPN domains each cleaving one DNA strand.

We showed that two C2c1 CRISPR arrays are expressed, processed into mature crRNAs, and capable of interference in *E. coli.* These experiments reveal distinct characteristics of the C2c1 loci including: (i) a 5' processed DR in the crRNA, (ii) a 5' PAM, and (iii) a putative tracrRNA. The AT-rich PAM of C2c1 contrasts with the GC-rich PAMs of Cas9. Using expression of C2c1 in a human cell culture, we show that a tracrRNA is essential for *in vitro* cleavage of target DNA. This feature is in sharp contrast to the recently characterized Cpf1 nuclease (Zetsche et al., 2015a), which does not require a tracrRNA for DNA cleavage. These findings show that, despite their common overall layout, Class 2 CRISPR-Cas systems substantially differ in their requirements for PAM and tracrRNA.

We also showed that when the C2c2 locus from *L. seeligeri* is expressed in *E. coli,* it is processed into crRNAs containing a 29-nt 5' DR; similar results were obtained for the C2c2 locus of *L. shahii.* In this case, the degenerate repeat is at the beginning of the array, rather than at the end, as in most other CRISPR arrays, and the array and *cas* genes are transcribed co-directionally. We did not detect a putative tracrRNA in the C2c2 RNA-seq data. The predicted secondary structure of the 29-nt DR shows a stable hairpin handle which could be important for complex formation with the C2c2 effector protein. Together, these results strongly suggest that C2c2 loci are functionally active.

The discovery of three distinct Class 2 CRISPR-Cas systems combined with the results of previous analyses (Chylinski et al., 2014; Makarova et al., 2011b) reveals a dominant theme in their evolution. The effector proteins of two of the three types within this class appear to have evolved from a pool of transposable elements that encode TnpB proteins containing the RuvC-like nuclease domain. Cas9, the effector protein of type II systems, seems to be derived from a family of TnpB-like proteins with an HNH nuclease insert that is particularly abundant in Cyanobacteria (Chylinski et al., 2014). By contrast, it is hardly possible to trace Cpf1, C2c1, and C2c3 to a specific TnpB group; however, given that they contain distinct insertions between the RuvC-motifs and apparently unrelated N-terminal regions, the effector proteins of each subtype of type V likely evolved independently from different TnpB proteins.

The TnpB proteins seem to be "predesigned" for utilization in Class 2 CRISPR-Cas effector complexes, perhaps stemming from their predicted ability to cut single-stranded DNA while bound to an RNA molecule via the R-rich bridge helix, which in Cas9 has been shown to bind crRNA (Anders et al., 2014; Nishimasu et al., 2014).

With regard to the origin of the putative type VI systems, although HEPN domains so far have not been detected in bona fide transposons, they are characterized by high horizontal mobility and are integral to certain mobile elements such as toxin-antitoxin units (Anantharaman et al., 2013). Thus, type VI systems seem to fit the paradigm of the modular evolution of Class 2 CRISPR-Cas from mobile components. Given that the C2c2 protein is unrelated to the other Class 2 effectors, the discovery of type VI seems to clinch the case for the independent origins of different Class 2 variants.

In view of the emerging scenario of the evolution of Class 2 systems from mobile elements, it is instructive to examine the overall evolution of CRISPR-Cas loci and the contributions of mobile elements (Figure 2.7). The ancestral adaptive immunity system most likely originated via the insertion of a casposon (a Cas1-encoding transposon) next to a locus that encoded a primitive innate immunity system (Koonin and Krupovic, 2015; Krupovic et al., 2014). An additional important contribution was the incorporation of a toxin-antitoxin system that delivered the *cas2* gene, either in the ancestral casposon or in the evolving adaptive immunity locus. Given the wide spread of Class 1 systems in archaea and bacteria and the proliferation of the ancient RRM (RNA Recognition Motif) domains in them, there is little doubt that the ancestral system was of Class 1. The different types and subtypes of Class 2 then evolved via multiple substitutions of the gene block encoding the Class 1 effector complexes via insertion of transposable elements encoding various nucleases. This direction of evolution follows from the observation that the adaptation modules of different Class 2 variants derive from different Class 1 types (Figure 2.6).

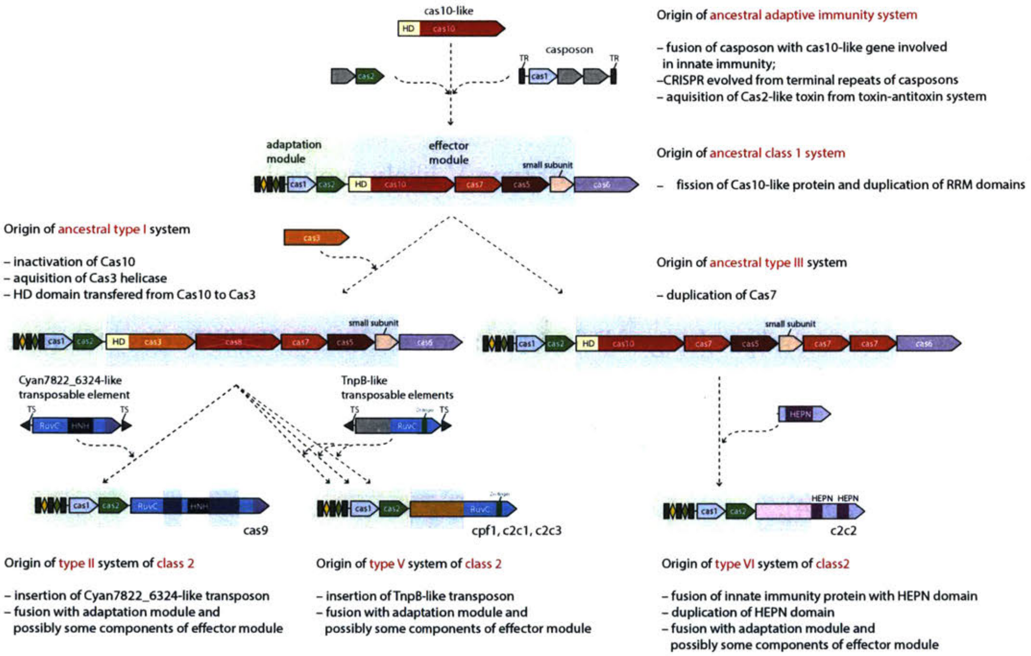

cas10-like
HD
cas10
casposon
TR
TR
cas2
cas1
Origin of ancestral adaptive immunity system
- fusion of casposon with cas10-like gene involved in innate immunity;
-CRISPR evolved from terminal repeats of casposons
- aquisition of Cas2-like toxin from toxin-antitoxin system
adaptation module
effector module
small subunit
cas1
cas2
HD
cas10
cas7
cas5
cas6
Origin of ancestral class 1 system
- fission of Cas10-like protein and duplication of RRM domains
Origin of ancestral type I system
- inactivation of Cas10
- aquisition of Cas3 helicase
- HD domain transfered from Cas10 to Cas3
cas3
Origin of ancestral type III system
- duplication of Cas7
small subunit
cas1
cas2
HD
cas3
cas8
cas7
cas5
cas6
small subunit
cas1
cas2
HD
cas10
cas7
cas5
cas7
cas7
cas6
Cyan7822_6324-like transposable element
TS
TS
RuvC
HNH
TnpB-like transposable elements
TS
Zn finger
TS
RuvC
HEPN
cas1
cas2
RuvC
HNH
cas9
cas1
cas2
RuvC
Zn finger
cpf1, c2c1, c2c3
cas1
cas2
HEPN
HEPN
c2c2
Origin of type II system of class 2
- insertion of Cyan7822_6324-like transposon
- fusion with adaptation module and possibly some components of effector module
Origin of type V system of class 2
- insertion of TnpB-like transposon
- fusion with adaptation module and possibly some components of effector module
Origin of type VI system of class2
- fusion of innate immunity protein with HEPN domain
- duplication of HEPN domain
- fusion with adaptation module and possibly some components of effector module

Figure 2.7: Evolutionary scenario for the CRISPR-Cas systems.

The scenario is a synthesis of the present and previous analyses (Chylinski et al., 2014; Makarova et al., 2011a; Makarova et al., 2015b; Makarova et al., 2013). The Cas8 protein is hypothesized to have evolved by inactivation of Cas10 (shown by white X) which was accompanied by a major acceleration of evolution. Abbreviations: TR, terminal repeats; TS, terminal sequences; HD, HD family endonuclease; HNH, HNH family endonuclease; RuvC, RuvC family endonuclease; HEPN, putative endoribonuclease of HEPN superfamily. Genes and portions of genes shown in gray denote sequences that are thought to have been encoded in the respective mobile elements but were eliminated in the course of evolution of CRISPR-Cas systems.

Strikingly, Class 2 CRISPR-Cas systems appear to have been completely derived from different mobile elements. There seem to have been at least two (subtype V-C) but typically, three or, for type II, even four mobile element contributors: (i) the ancestral casposon, (ii) the toxin-antitoxin module that gave rise to Cas2, (iii) a transposable element, in many cases a TnpB-encoding one, that was the ancestor of the Class 2 effector complex, and (iv) for type II, the HNH nuclease that could have been donated to the ancestral transposon by a group I or group II self-splicing intron (Stoddard, 2005) (Figure 2.7). The type V-C loci described here encode the ultimate minimalist CRISPR-Cas system, the only identified one that lacks Cas2; conceivably, the highly diverged subtype V-C Cas1 proteins are able to form the adaptation complex on their own, without the accessory Cas2 subunit.

Our report here of new varieties of Class 2 CRISPR-Cas systems could be only a sample of the additional variants that exist in nature, and although most if not all of the new CRISPR-Cas systems are expected to be rare, they could employ novel strategies and molecular mechanisms, providing a major resource for versatile applications in genome engineering and biotechnology. That the development of such new tools is realistic, is demonstrated by the activity of a C2c1nuclease in human cell lysate described here, and Cpf1-mediated genome editing in human cells (Zetsche et al., 2015a). In addition, the discovery of new variants will provide direct tests of the modular scenario of the evolution of CRISPR-Cas systems (Figure 2.7) and shed further light on the function of these diverse systems.

2.5 Experimental Procedures

2.5.1 Computational sequence analysis

The TBLASTN program with the E-value cut-off of 0.01 and low complexity filtering turned off parameters was used to search the NCBI WGS database using the Cas1 profile (Makarova et al., 2015b) as the query. Sequences of contigs or complete genome partitions where a Cas1 hit was identified were retrieved from the database, and regions 20 kb from the start of the *cas1* gene and 20 kb from the end of it were extracted and translated using GeneMarkS (Besemer et al., 2001). Predicted proteins from each Cas1-encoding region were searched against the collection of profiles from the

CDD database (Marchler-Bauer et al., 2013) and the specific Cas protein profiles (Makarova et al., 2015b) using the RPS-BLAST program (Marchler-Bauer et al., 2002). The previously developed procedure to assess the completeness and to classify CRISPR-Cas systems into the existing types and subtypes (Makarova et al., 2015b) was applied to each locus. Partial and/or unclassified loci that encompassed proteins larger than 500 amino acids were analyzed on a case-by-case basis. Specifically, each predicted protein encoded in these loci was searched against the NCBI non-redundant (NR) protein sequence database using PSI-BLAST (Altschul et al., 1997), with a cut-off e-value of 0.01 and composition based-statistics and low complexity filtering turned off. Each non-redundant protein identified in this search was searched against the WGS database using the TBLASTN program (Altschul et al., 1997). The HHpred program was used with default parameters to identify remote sequence similarity using as the queries all proteins identified in the BLAST searches (Soding et al., 2006). Multiple sequence alignments were constructed using MUSCLE (Edgar, 2004) and MAFFT (Katoh and Standley, 2013). Phylogenetic analysis was performed using the FastTree program with the WAG evolutionary model and the discrete gamma model with 20 rate categories (Price et al., 2010). Protein secondary structure was predicted using Jpred 4 (Drozdetskiy et al., 2015).

CRISPR repeats were identified using PILER-CR (Edgar, 2007) or, for degenerate repeats, CRISPRfinder (Grissa et al., 2007). The Mfold program (Zuker, 2003) was used to identify the most stable structure for the repeat sequences. The CRISPRmap method (Lange et al., 2013) was used for repeat classification.

The spacer sequences were searched against the NCBI nucleotide NR and WGS databases using MEGABLAST (Morgulis et al., 2008) with default parameters except that the word size was set at 20.

2.5.2 Bacterial RNA-sequencing

RNA was isolated from stationary phase bacteria by first resuspending the bacteria in TRIzol and then homogenizing the bacteria with zirconia/silica beads (BioSpec Products) in a BeadBeater (BioSpec Products) for 7 one-minute cycles. Total RNA was purified from homogenized samples with the Direct-Zol RNA miniprep protocol (Zymo), DNase treated with TURBO DNase (Life Technologies) and 3' dephosphorylated with T4 Polynucleotide Kinase (New England Biolabs). rRNA was removed with the bacterial Ribo-Zero rRNA removal kit (Illumina). RNA sequencing libraries were prepared from rRNA-depleted RNA using a derivative of the previously described CRISPR RNA sequencing method (Heidrich et al., 2015). Briefly, transcripts were poly-A tailed with *E. coli* Poly(A) Polymerase (New England Biolabs), ligated with 5' RNA adapters using T4 RNA Ligase 1 (ssRNA Ligase), High Concentration (New England Biolabs), and reverse transcribed with AffinityScript Multiple Temperature Reverse Transcriptase (Agilent Technologies). cDNA was PCR amplified with barcoded primers using Herculase II polymerase (Agilent Technologies) .

2.5.3 RNA-sequencing analysis

The prepared cDNA libraries were sequenced on an MiSeq (Illumina). 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 Picard tools (http://broadinstitute.github.io/picard) and these sequences were analyzed using Geneious 8.1.5. All the sequences obtained in this work were deposited in the Single Read Archive (SRA) database under the accession number PRJNA296743.

2.5.4 PAM Screen

Randomized PAM plasmid libraries were constructed using synthesized oligonucleotides (IDT) consisting of 7 randomized nucleotides either upstream or downstream of the spacer 1 target. The randomized ssDNA oligos were made double stranded by annealing to a short primer and using the large Klenow fragment for second strand synthesis. The dsDNA product was assembled into a linearized PUC19 using Gibson cloning. Stabl3 E. coli cells were transformed with the cloned products and more than 10^7 cells were collected and pooled. Plasmid DNA was harvested using a Qiagen maxi-prep kit. We transformed 360ng of the pooled library into *E. coli* cells transformed with the AacC2c1 locus, BthC2c1 locus, pACYC-184 and pET-28a. After transformation, cells were plated on ampicillin/chloramphenicol (Aac/pACYC-184) and ampicillin/kanamycin (Bth/pET-28a). After

16 hours of growth, >4*10^6 cells were harvested and plasmid DNA was extracted using a Qiagen maxi-prep kit. The target PAM region was amplified and sequenced using an Illumina MiSeq with single-end 150 cycles.

2.5.5 PAM validation

Sequences corresponding to both PAMs non-PAMs were cloned into digested pUC19 and ligated with T4 ligase (Enzymatics). Competent *E. coli* with either the AacC2c1 locus plasmid or pACYC184 control plasmid were transformed with 20ng of PAM plasmid and plated on LB agar plates supplemented with ampicillin and chloramphenicol. After 18 hours, colonies were counted with OpenCFU (Geissmann 2013).

2.5.6 *In vitro* lysate cleavage assay

Cleavage was performed using the lysate of HEK293 cells expressing C2c1 protein at 50°C, unless otherwise noted, in cleavage buffer (NEBuffer 3, 5mM DTT) for 1 hour. Each cleavage reaction used 200ng of target DNA and an equimolar ratio of crRNA:tracrRNA (500ng of crRNA). The RNA was pre-annealed by heating to 95°C and slowly cooling to 4°C. Target DNA consisted of either genomic PCR amplicons from the *EMX1* gene or the first protospacer of the AacC2c1 locus cloned into pUC19. The pUC19 protospacer construct was linearized by BsaI digestion prior to the cleavage reaction. Reactions were cleaned up using PCR purification columns (Qiagen) and run on 2% agarose E-gels (Life Technologies).

2.6 Acknowledgements

This work was published in (Shmakov et al., 2015). We thank Rhiannon Macrae and Fei Ann Ran for critical reading of the manuscript.

Chapter 3

C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector.

This chapter is adapted from the following article:

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., *et al.* (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573.

Contributions: Omar Abudayyeh, Jonathan Gootenberg, and Silvana Konermann are co-first authors (*). Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang conceived the study; Omar Abudayyeh and Jonathan Gootenberg designed and participated in all experiments; Silvana Konermann helped with biochemical Cas13 and transcription-dependent assays; Ian Slaymaker helped with Cas13 protein purification; Julia Joung helped with RNA sequencing cleavage analysis; David Cox helped with MS2 phage assays; Sergey Shmakov, Kira Makarova, Yuri Wolf, Ekaterina Semenova, Konstantin Severinov, Leonid Minakhin, and Aviv Regev helped with editing the manuscript; Eric Lander helped with screening analysis; Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang wrote the manuscript that was read, edited and approved by all authors.

3.1 Abstract

The CRISPR-Cas adaptive immune system defends microbes against foreign genetic elements via DNA or RNA-DNA interference. We characterize the Class 2 type VI-A CRISPR-Cas effector C2c2 and demonstrate its RNA-guided RNase function. C2c2 from the bacterium *Leptotrichia shahii* provides interference against RNA phage. *In vitro* biochemical analysis show that C2c2 is guided by a single crRNA and can be programmed to cleave ssRNA targets carrying complementary protospacers. In bacteria, C2c2 can be programmed to knock down specific mRNAs. Cleavage is mediated by catalytic residues in the two conserved HEPN domains, mutations in which generate catalytically inactive RNA-binding proteins. These results broaden our understanding of CRISPR-Cas systems and suggest that C2c2 can be used to develop new RNA-targeting tools.

3.2 Introduction

Almost all archaea and about half of bacteria possess Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated genes (CRISPR-Cas) adaptive immune systems (Makarova et al., 2011b; Makarova et al., 2015a), which protect microbes from viruses and other invading DNA through three steps: (i) adaptation, i.e., insertion of foreign nucleic acid segments (spacers) into the CRISPR array in between pairs of direct repeats (DRs), (ii) transcription and processing of the CRISPR array to produce mature CRISPR RNAs (crRNAs), and (iii) interference, whereby Cas enzymes are guided by the crRNAs to target and cleave cognate sequences in the respective invader genomes (Marraffini, 2015; van der Oost et al., 2009; Wright et al., 2016). All CRISPR-Cas systems characterized to date follow these three steps, although the mechanistic implementation and proteins involved in these processes display extensive diversity.

The CRISPR-Cas systems are broadly divided into two classes on the basis of the architecture of the interference module: Class 1 systems rely on multi-subunit protein complexes whereas Class 2 systems utilize single effector proteins (Makarova et al., 2015a). Within these two classes, types and subtypes are delineated according to the presence of distinct signature genes, protein sequence conservation, and organization of the respective genomic loci. Class 1 systems include type I, where interference is achieved through assembly of multiple Cas proteins into the Cascade complex, and type III systems, which rely on either the Csm (type III-A/D) or Cmr (Type III-B/C) effector complexes which are distantly related to the Cascade (Brouns et al., 2008; Hale et al., 2009; Jackson et al., 2014; Makarova et al., 2015a; Marraffini and Sontheimer, 2008; Mulepati et al., 2014; Sinkunas et al., 2013).

Class 2 CRISPR systems comprise type II, characterized by the single-component effector protein Cas9 (Barrangou et al., 2007; Deltcheva et al., 2011; Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012; Sapranauskas et al., 2011), which contains RuvC and HNH nuclease domains, and type V systems, which utilize single RuvC domain-containing effectors such as Cpf1 (Zetsche et al., 2015b), C2c1, and C2c3 (Shmakov et al., 2015). All functionally characterized systems, to date, have been reported to target DNA, and only the multi-component type III-A and III-B systems additionally target RNA (Hale et al., 2012; Hale et al., 2009; Jiang et al., 2016; Samai et al., 2015; Staals et al., 2013; Staals et al., 2014; Tamulaitis et al., 2014). However, the putative Class 2 type VI system is characterized by the presence of the single effector protein C2c2, which lacks homology to any known DNA nuclease domain but contains two Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domains (Shmakov et al., 2015). Given that all functionally characterized HEPN domains are RNases (Anantharaman et al., 2013), there is a possibility that C2c2 functions solely as an RNA-guided RNA-targeting CRISPR effector.

HEPN domains are also found in other Cas proteins. Csm6, a component of type III-A systems, and the homologous protein Csx1, in type III-B systems, each contain a single HEPN domain and have been biochemically characterized as ssRNA-specific endoribonucleases (Jiang et al., 2016; Niewoehner and Jinek, 2016; Sheppard et al., 2016). In addition, type III systems contain complexes of other Cas enzymes that bind and cleave ssRNA through acidic residues associated with RNA-recognition motif (RRM) domains. These complexes (Cas10-Csm in type III-A and Cmr in type III-B) carry out RNA-guided co-transcriptional cleavage of mRNA in concert with DNA target cleavage (Deng et al., 2013; Goldberg et al., 2014; Samai et al., 2015). In contrast, the roles of Csm6 and Csx1, which cleave their targets with little specificity, are less clear, although in some cases, RNA cleavage by Csm6 apparently serves as a second line of defense when DNA targeting fails (Jiang et al., 2016). Additionally, Csm6 and Csx1 have to dimerize to form a composite active site (Kim et al., 2013;

Niewoehner and Jinek, 2016; Sheppard et al., 2016), but C2c2 contains two HEPN domains, suggesting that it functions as a monomeric endoribonuclease.

As is common with Class 2 systems, type VI systems are simply organized. In particular, the type VI locus in *Leptotrichia shahii* contains Cas1, Cas2, C2c2 and a CRISPR array, which is expressed and processed into mature crRNAs (Shmakov et al., 2015). In all CRISPR-Cas systems characterized to date, Cas1 and Cas2 are exclusively involved in spacer acquisition (Datsenko et al., 2012; Diez-Villasenor et al., 2013; Heler et al., 2015; Nunez et al., 2014; Nunez et al., 2015; Yosef et al., 2012), suggesting that C2c2 is the sole effector protein which utilizes a crRNA guide to achieve interference, likely targeting RNA.

3.3 Results

3.3.1 Reconstitution of *L. shahii* C2c2 locus in *Escherichia coli* confers RNA-guided immunity

We explored whether LshC2c2 could confer immunity to MS2 (Tamulaitis et al., 2014), a lytic single-stranded (ss) RNA phage, without DNA intermediates in its life cycle, that infects *E. coli.* We constructed a low-copy plasmid carrying the entire LshC2c2 locus (pLshC2c2) to allow for heterologous reconstitution in *E. coli* (fig. 3.S1A). Because expressed mature crRNAs from the LshC2c2 locus have a maximum spacer length of 28nt (fig. 3.S1A) (Shmakov et al., 2015), we tiled all possible 28-nt target sites in the MS2 phage genome (Fig. 3.1A). This resulted in a library of 3,473 spacer sequences (along with 490 non-targeting guides designed to have a Levenshtein distance of ≥8 with respect to the MS2 and *E. coli* genomes) which we inserted between pLshC2c2 direct repeats (DRs). After transformation in of this construct into *E. coli,* we infected cells with varying dilutions of MS2 (10^{-1}, 10^{-3}, and 10^{-5}) and analyzed surviving cells to determine the spacer sequences carried by cells that survived the infection. Cells carrying spacers that confer robust interference against MS2 are expected to proliferate faster than those that lack such sequences. Following growth for 16 hours, we identified a number of spacers that were consistently enriched across three independent infection replicas in both the 10^{-1} and 10^{-3} dilution conditions, suggesting that they enabled interference against MS2. Specifically, 147 and 150 spacers showed >1.25 $\log_2$-fold enrichment in all three replicates for the 10^{-1} and 10^{-3} phage dilutions, respectively; of these two groups of top enriched spacers, 84 are shared (Figs. 3.1B, 3.S2A-G). Additionally, no non-targeting guides were found to be consistently enriched among the three 10^{-1}, 10^{-3}, or 10^{-5} phage replicates (fig. 3.S2D, G). We also analyzed the flanking regions of protospacers on the MS2 genome corresponding to the enriched spacers and found that spacers with a G immediately flanking the 3' end of the protospacer were less fit relative to all other nucleotides at this position (i.e. A, U, or C), suggesting that the 3' protospacer flanking site (PFS) affects the efficacy of C2c2-mediated targeting (Figs. 3.1C, 3.S2E-F, 3.S3). Although the PFS is adjacent to the protospacer target, we chose not to use the commonly used protospacer adjacent motif (PAM) nomenclature as it has come to connote a sequence used in self vs. non-self differentiation (Marraffini and Sontheimer, 2010b), which is irrelevant in a RNA-targeting system. It is worth noting that the avoidance of G by C2c2 echo the absence of PAMs applicable to other RNA-targeting CRISPR systems and effector proteins (Hale et al., 2014; Hale et al., 2012; Samai et al., 2015; Staals et al., 2014; Tamulaitis et al., 2014; Zhang et al., 2012).

The fact that only ~5% of crRNAs are enriched may reflect other factors influencing interference activity, such as accessibility of the target site that might be affected by RNA binding proteins or secondary structure. In agreement with this hypothesis, the enriched spacers tend to cluster into regions of strong interference where they are closer to each other than one would expect by random chance (fig. 3.S3F-G).

To validate the interference activity of the enriched spacers, we individually cloned four top-enriched spacers into pLshC2c2 CRISPR arrays and observed a 3- to 4-$\log_{10}$ reduction in plaque formation, consistent with the level of enrichment observed in the screen (Figs 3.1B, 3.S4). We cloned sixteen guides targeting distinct regions of the MS2 *mat* gene (4 guides per possible single-nucleotide PFS). All 16 crRNAs mediated MS2 interference, although higher levels of resistance were observed for the C, A, and U PFS-targeting guides (Figs. 3.1D, 3.1E, 3.S5), indicating that C2c2 can be effectively retargeted in a crRNA-dependent fashion to sites within the MS2 genome.

To further validate the observed PFS preference with an alternate approach, we designed a protospacer site in the pUC19 plasmid at the 5' end of the ß-lactamase mRNA, which encodes ampicillin resistance in *E. coli*, flanked by five randomized nucleotides at the 3' end. Significant depletion and enrichment was observed for the LshC2c2 locus (****, p<0.0001) compared to the pACYC184 controls (Fig. 3.S6A). Analysis of the depleted PFS sequences confirmed the presence of a PFS preference of H (Fig. 3.S6B).

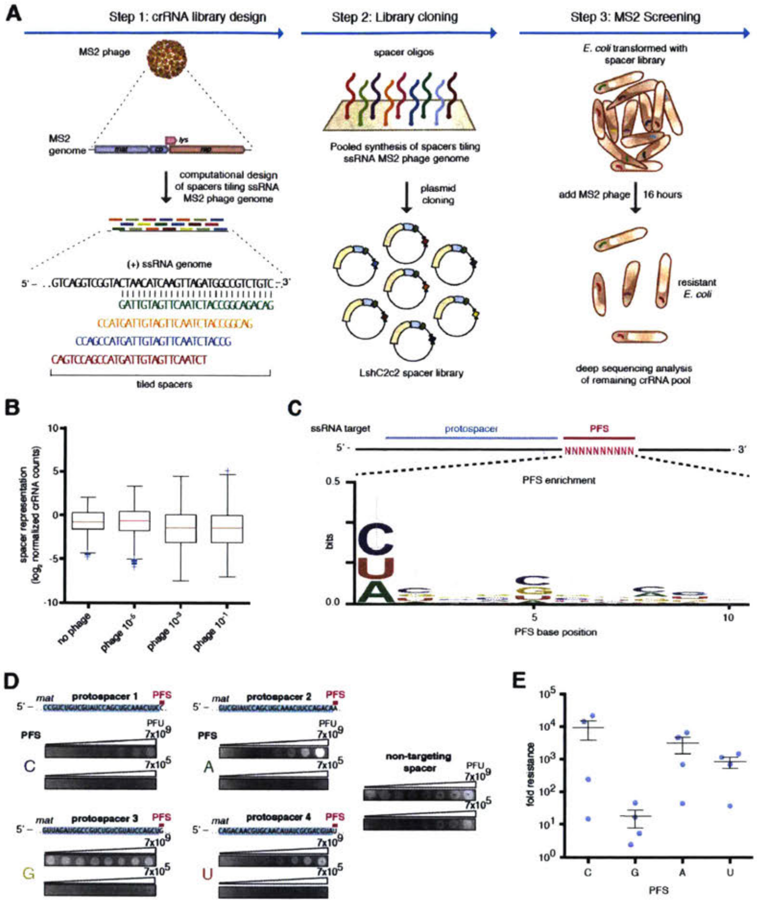


Figure 3.1: Heterologous expression of the *Leptotrichia shahii* C2c2 locus mediates robust interference of RNA phage in *Escherichia coli.*

A) Schematic for the MS2 bacteriophage interference screen. A library consisting of spacers targeting all possible sequences in the MS2 RNA genome was cloned into the LshC2c2 CRISPR array. Cells transformed with the MS2-targeting spacer library were then treated with phage and plated, and surviving cells were harvested. The frequency of spacers was compared to an untreated control (no phage), and enriched spacers from the phage-treated condition were used for the generation of PFS preference logos.

B) Box plot showing the distribution of normalized crRNA frequencies for the phage-treated conditions and control screen (no phage) biological replicates (n = 3). The box extends from the first to third quartile with whiskers denoting 1.5 times the interquartile range. The mean is indicated by the red horizontal bar. The 10^{-1} and 10^{-3} phage dilution distributions are significantly different than each of the control replicates (****, $p < 0.0001$ by ANOVA with multiple hypothesis correction).

C) Sequence logo generated from sequences flanking the 3' end of protospacers corresponding to enriched spacers in the 10^{-1} phage dilution condition, revealing the presence of a 3' H PFS (not G).

D) Plaque assay used to validate the functional significance of the H PFS in MS2 interference. All protospacers flanked by non-G PFSs exhibited robust phage interference. Spacer were designed to target the MS2 *mat* gene and their sequences are shown above the plaque images; the spacer used in the non-targeting control is not complementary to any sequence in either the *E. coli* or MS2 genome. Phage spots were applied as series of half-log dilutions.

E) Quantitation of MS2 plaque assay validating the H (non-G) PFS preference. 4 MS2-targeting spacers were designed for each PFS. Each point on the scatter plot represents the average of three biological replicates and corresponds to a single spacer. Bars indicate the mean of 4 spacers for each PFS and standard error (s.e.m).

3.3.2 C2c2 is a single-effector endoRNase mediating ssRNA cleavage with a crRNA guide

We purified the LshC2c2 protein (fig. 3.S7) and assayed its ability to cleave an *in vitro* transcribed 173-nt ssRNA target (Figs. 3.2A, 3.S8) containing a C PFS (ssRNA target 1 with protospacer 14). Mature LshC2c2 crRNAs contain a 28-nt direct repeat (DR) and a 28 nt spacer (fig. 3.S1A) (Shmakov et al., 2015). We therefore generated an *in-vitro*-transcribed crRNA with a 28-nt spacer complementary to protospacer 14 on ssRNA target 1. LshC2c2 efficiently cleaved ssRNA in a Mg^{2+}- and crRNA-dependent manner (Figs. 3.2B, 3.S9). We then annealed complementary RNA oligos to regions flanking the crRNA target site. This partially double-stranded RNA substrate was not cleaved by LshC2c2, suggesting it is specific for ssRNA (figs. 3.S10A-B).

We tested the sequence constraints of RNA cleavage by LshC2c2 with additional crRNAs complementary to ssRNA target 1 where protospacer 14 is preceded by each PFS variant. The results of this experiment confirmed the preference for C, A, and U PFSs, with little cleavage activity detected for the G PFS target (Fig. 3.2C). Additionally, we designed 5 crRNAs for each possible PFS (20 total) across the ssRNA target 1 and evaluated cleavage activity for LshC2c2 paired with each of these crRNAs. As expected, we observed less cleavage activity for G PFS-targeting crRNAs compared to other crRNAs tested (Fig. 3.2D).

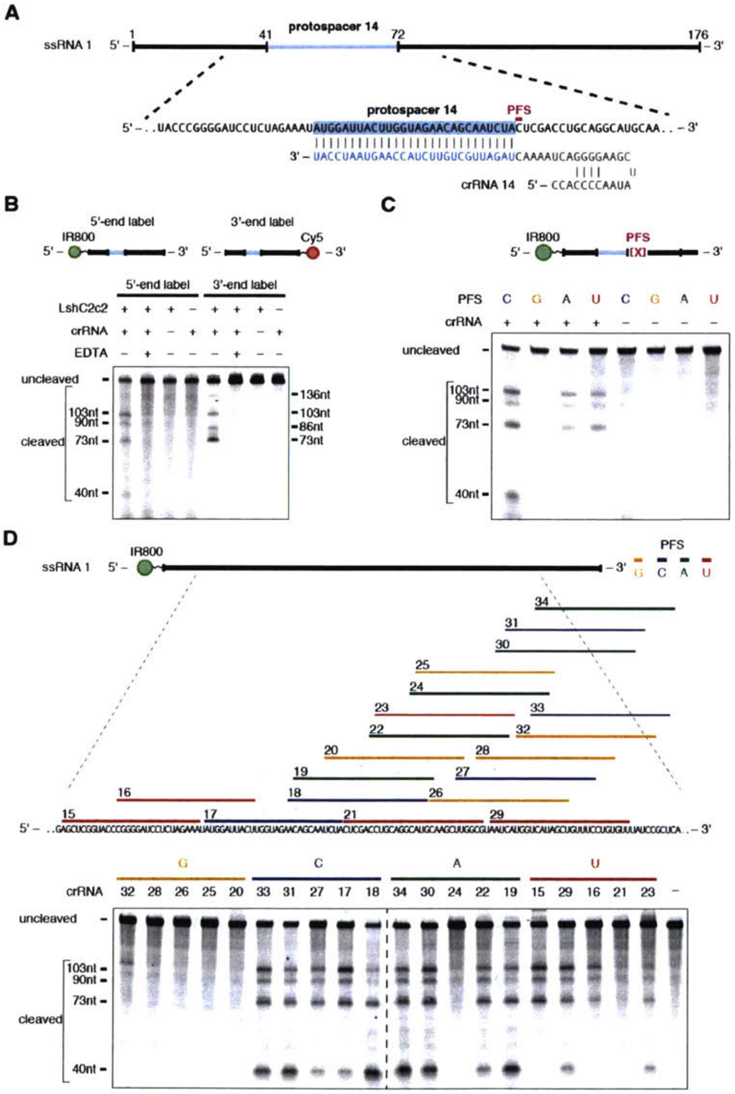


Figure 3.2: LshC2c2 and crRNA mediate RNA-guided ssRNA cleavage

A) Schematic of the ssRNA substrate being targeted by the crRNA. The protospacer region is highlighted in blue and the PFS is indicated by the magenta bar.

B) A denaturing gel demonstrating crRNA-mediated ssRNA cleavage by LshC2c2 after 1 hour of incubation. The ssRNA target is either 5' labeled with IRDye 800 or 3' labeled with Cy5.

Cleavage requires the presence of the crRNA and is abolished by addition of EDTA. Four cleavage sites are observed. Reported band lengths are matched from RNA sequencing.

C) A denaturing gel demonstrating the requirement for an H PFS (not G) after 3 hours of incubation. Four ssRNA substrates that are identical except for the PFS (indicated by the magenta X in the schematic) were used for the *in vitro* cleavage reactions. ssRNA cleavage activity is dependent on the nucleotide immediately 3' of the target site. Reported band lengths are matched from RNA sequencing.

D) Schematic showing five protospacers for each PFS on the ssRNA target (top). Denaturing gel showing crRNA-guided ssRNA cleavage activity after 1 hour of incubation. crRNAs correspond to protospacer numbering. Reported band lengths are matched from RNA sequencing.

We then generated a dsDNA plasmid library with protospacer 14 flanked by 7 random nucleotides to account for any PFS preference. When incubated with LshC2c2 protein and a crRNA complementary to protospacer 14, no cleavage of the dsDNA plasmid library was observed (fig. 3.S10C). We also did not observe cleavage when targeting a ssDNA version of ssRNA target 1 (fig. 3.S10D). To rule out co-transcriptional DNA cleavage, which has been observed in type III CRISPR-Cas systems (Samai et al., 2015), we recapitulated the *E. coli* RNA polymerase co-transcriptional cleavage assay (Samai et al., 2015) (fig. 3.S11A) expressing ssRNA target 1 from a DNA substrate. This assay of purified LshC2c2 and crRNA targeting ssRNA target 1 did not show any DNA cleavage (fig. 3.S11B). Together, these results indicate that C2c2 cleaves specific ssRNA sites directed by the target complementarity encoded in the crRNA, with a H PFS preference.

3.3.3 C2c2 cleavage depends on local target sequence and secondary structure

Given that C2c2 did not efficiently cleave dsRNA substrates and that ssRNA can form complex secondary structures, we reasoned that cleavage by C2c2 might be affected by secondary structure of the ssRNA target. Indeed, after tiling ssRNA target 1 with different crRNAs (Fig. 3.2D), we observed the same cleavage pattern regardless of the crRNA position along the target RNA. This observation suggests that the crRNA-dependent cleavage pattern was determined by features of the target sequence rather than the distance from the binding site. We hypothesized that the LshC2c2-crRNA complex binds the target and cleaves exposed regions of ssRNA within the secondary structure elements, with potential preference for certain nucleotides.

In agreement with this hypothesis, cleavage of three ssRNA targets with different sequences flanking identical 28-nt protospacers resulted in three distinct patterns of cleavage (Fig. 3.3A). RNA-sequencing of the cleavage products for the three targets revealed that cleavage sites mainly localized to uracil-rich regions of ssRNA or ssRNA-dsRNA junctions within the *in silico* predicted co-folds of the target sequence with the crRNA (Figs. 3.3B-C, 3.S12A-D). To test whether the LshC2c2-crRNA complex prefers cleavage at uracils, we analyzed the cleavage efficiencies of homopolymeric RNA targets (a 28-nt protospacer extended with 120 As or Us regularly interspaced by single bases of G or C to enable oligo synthesis) and found that LshC2c2 preferentially cleaved the uracil target compared to adenine (figs. 3.S12E, 3.S12F). We then tested cleavage of a modified version of ssRNA 4 which had its main site of cleavage, a loop, replaced with each of the four possible homopolymers and found that cleavage only occurred at the uracil homopolymer loop (fig. 3.S12G). To further test whether cleavage was occurring at uracil residues, we mutated single uracil residues in ssRNA 1 that showed cleavage in the RNA-sequencing (Fig. 3.3B) to adenines. This experiment showed that, by mutating each uracil residue, we could modulate the presence of a single cleavage band, consistent with LshC2c2 cleaving at uracil residues in ssRNA regions (Fig. 3.3D).

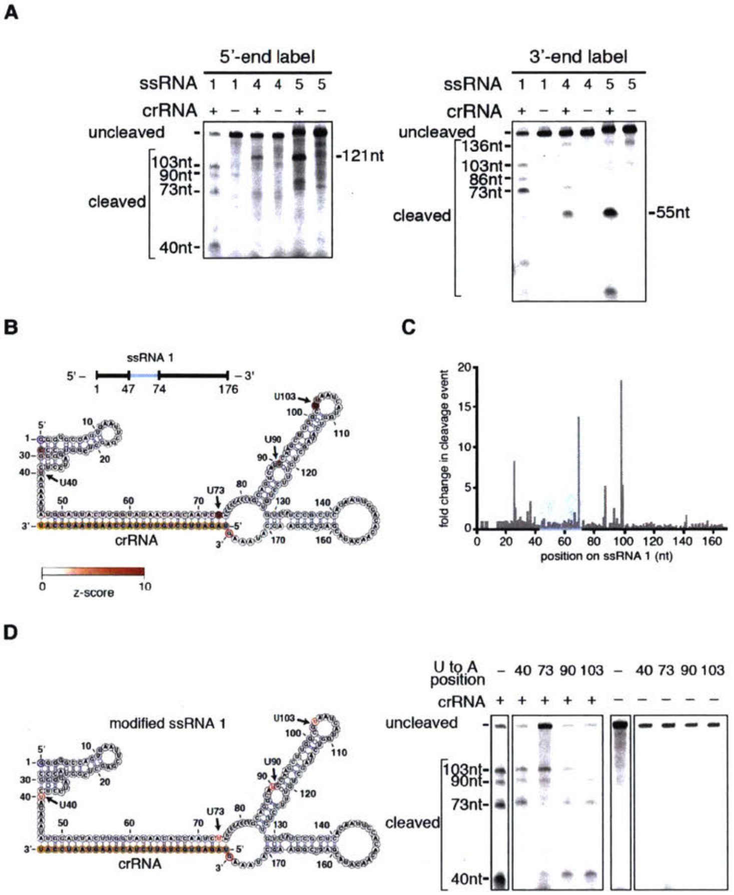


Figure 3.3: C2c2 cleavage sites are determined by secondary structure and target RNA sequence.

A) Denaturing gel showing C2c2-crRNA-mediated cleavage after 3 hours of incubation of three non-homopolymeric ssRNA targets (1, 4, 5; black, blue and green on figs 3B-C and S12A-D respectively) that share the same protospacer but are flanked by different sequences. Despite identical protospacers, different flanking sequences resulted in different cleavage patterns. Reported band lengths are matched from RNA sequencing.

B) The cleavage sites of non-homopolymer ssRNA target 1 were mapped with RNA-sequencing of the cleavage products. The frequency of cleavage at each base is colored according to the z-score and shown on the predicted crRNA-ssRNA co-fold secondary structure. Fragments used to generate the frequency analysis contained the complete 5' end. The 5' and 3' end of the ssRNA target are indicated by blue and red outlines, on the ssRNA and secondary structure, respectively. The 5' and 3' end of the spacer (outlined in yellow) is indicated by the blue and orange residues highlighted respectively. The crRNA nucleotides are highlighted in orange.

C) Plot of the frequencies of cleavage sites for each position of ssRNA target 1 for all reads that begin at the 5' end. The protospacer is indicated by the blue highlighted region.

D) Schematic of a modified ssRNA 1 target showing sites *(red)* of single U to A flips *(left)*. Denaturing gel showing C2c2-crRNA mediated cleavage of each of these single nucleotide variants after 3 hours of incubation *(right)*. Reported band lengths are matched from RNA sequencing.

3.3.4 The HEPN domains of C2c2 mediate RNA-guided ssRNA-cleavage

Bioinformatic analysis of C2c2 has suggested that the HEPN domains are likely to be responsible for the observed catalytic activity (Shmakov et al., 2015). Each of the two HEPN domains of C2c2 contains a dyad of conserved arginine and histidine residues (Fig. 3.4A), in agreement with the catalytic mechanism of the HEPN endoRNAse (Anantharaman et al., 2013; Niewoehner and Jinek, 2016; Sheppard et al., 2016). We mutated each of these putative catalytic residues separately to alanine (R597A, H602A, R1278A, H1283A) in the LshC2c2 locus plasmids and assayed for MS2 interference. None of the four mutant plasmids were able to protect *E. coli* from phage infection (Figs. 3.4B, 3.S13).

We purified the four single-point mutant proteins and assayed their ability to cleave 5'-end-labeled ssRNA target 1 (Fig. 3.4C). In agreement with our *in vivo* results, all four mutations abolished cleavage activity. The inability of either of the two wild-type HEPN domains to compensate for inactivation of the other implies cooperation between the two domains. These results agree with observations that several bacterial and eukaryotic single-HEPN proteins function as dimers (Kozlov et al., 2011; Niewoehner and Jinek, 2016; Sheppard et al., 2016).

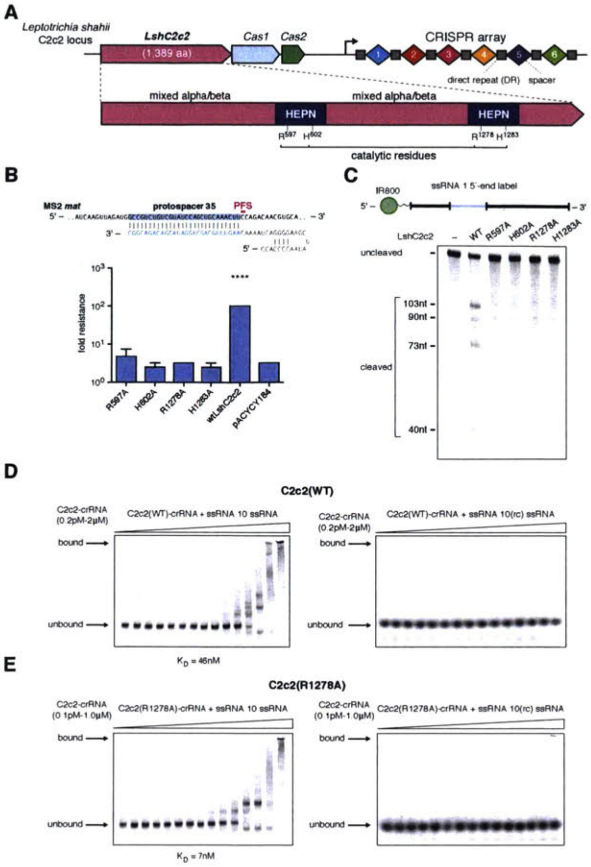


Figure 3.4: The two HEPN domains of C2c2 are necessary for crRNA-guided ssRNA cleavage but not for binding

A) Schematic of the *LshC2c2* locus and the domain organization of the LshC2c2 protein, showing conserved residues in HEPN domains (dark blue).

B) Quantification of MS2 plaque assay with HEPN catalytic residue mutants. For each mutant, the same crRNA targeting protospacer 35 was used. (n=3 biological replicates, ****, $p < 0.0001$ compared to pACYC184 by t-test. Bars represent mean ± s.e.m.)

C) Denaturing gel showing conserved residues of the HEPN motif, indicated as catalytic residues in panel A, are necessary for crRNA-guided ssRNA target 1 cleavage after 3 hours of incubation. Reported band lengths are matched from RNA sequencing.

D) Electrophoretic mobility shift assay (EMSA) evaluating affinity of the wild type LshC2c2-crRNA complex against a targeted (left) and a non-targeted (right) ssRNA substrate. The non-targeted ssRNA substrate is the reverse-complement of the targeted ssRNA 10. EDTA is supplemented to reaction condition to reduce any cleavage activity.

E) Electrophoretic mobility shift assay with LshC2c2(R1278A)-crRNA complex against on-target ssRNA 10 and non-targeting ssRNA (same substrate sequences as in D)

Catalytically inactive variants of Cas9 retain target DNA binding, allowing for the creation of programmable DNA-binding proteins (Gasiunas et al., 2012; Jinek et al., 2012). Electrophoretic mobility shift assays (EMSA) on both the wild-type (Fig. 3.4D) and R1278A mutant LshC2c2 (Fig. 3.4E) in complex with crRNA showed the wild-type LshC2c2 complex binding strongly (K_D ~ 46 nM, fig. 3.S14A) and specifically to 5'-end-labeled ssRNA target 10 but not to the 5'-end-labeled non-target ssRNA (the reverse complement of ssRNA target 10). The R1278A mutant C2c2 complex showed even stronger (K_D ~ 7 nM, fig. 3.S14B) specific binding, indicating that this HEPN mutation results in a catalytically inactive, RNA-programmable RNA-binding protein. The LshC2c2 protein or crRNA alone showed reduced levels of target affinity, as expected (fig. 3.S14C-E). Additionally, no specific binding of LshC2c2-crRNA complex to ssDNA was observed (fig. 3.S15).

These results demonstrate that C2c2 cleaves RNA via a catalytic mechanism distinct from other known CRISPR-associated RNases. In particular, the type III Csm and Cmr multiprotein complexes rely on acidic residues of RRM domains for catalysis, whereas C2c2 achieves RNA cleavage through the conserved basic residues of its two HEPN domains.

3.3.5 Sequence and structural requirements of C2c2 crRNA

Similar to the type V-B (Cpf1) systems (Zetsche et al., 2015b), the LshC2c2 crRNA contains a single stem loop in the direct repeat (DR), suggesting that the secondary structure of the crRNA could facilitate interaction with LshC2c2. We thus investigated the length requirements of the spacer sequence for ssRNA cleavage and found that LshC2c2 requires spacers of at least 22 nt length to efficiently cleave ssRNA target 1 (fig. 3.S16A). The stem-loop structure of the crRNA is also critical for ssRNA cleavage, because DR truncations that disturbed the stem loop abrogated target cleavage (fig. 3.S16B). Thus, a DR longer than 24 nt is required to maintain the stem loop necessary for LshC2c2 to mediate ssRNA cleavage.

Single base pair inversions in the stem that preserved the stem structure did not affect the activity of the LshC2c2 complex. In contrast, inverting all four G-C pairs in the stem eliminated the cleavage despite maintaining the duplex structure (fig. 3.S17A). Other perturbations, such as those that introduced kinks and reduced or increased base-pairing in the stem, also eliminated or drastically suppressed cleavage. This suggests that the crRNA stem length is important for complex formation and activity (fig. 3.S17A). We also found that loop deletions eliminated cleavage, whereas insertions and substitutions mostly maintained some level of cleavage activity (fig. 3.S17B). In contrast, nearly all substitutions or deletions in the region 3' to the DR prevented cleavage by LshC2c2 (fig 3.S18). Together, these results demonstrate that LshC2c2 recognizes structural characteristics of its cognate crRNA but is amenable to loop insertions and most tested base substitutions outside of the 3' DR region. These results have implications for the future application of C2c2-based tools that require guide engineering for recruitment of effectors or modulation of activity (Dahlman et al., 2015; Kiani et al., 2015; Konermann et al., 2015).

3.3.6 C2c2 cleavage is sensitive to double mismatches in the crRNA-target duplex

We tested the sensitivity of the LshC2c2 system to single mismatches between the crRNA guide and target RNA by mutating single bases across the spacer to the respective complementary bases (e.g., A to U). We then quantified plaque formation with these mismatched spacers in the MS2 infection assay and found that C2c2 was fully tolerant to single mismatches across the spacer as such mismatched spacers interfered with phage propagation with similar efficiency as fully matched spacers (figs. 3.S19A, 3.S20). However, when we introduced consecutive double substitutions in the spacer, we found a ~3 $\log_{10}$-fold reduction in the protection for mismatches in the center, but not at the 5'- or 3'-end, of the crRNA (figs. 3.19B, 3.S20). This observation suggests the presence of a mismatch-sensitive "seed region" in the center of the crRNA-target duplex.

We generated a set of *in vitro* transcribed crRNAs with mismatches similarly positioned across the spacer region. When incubated with LshC2c2 protein, all single mismatched crRNA supported cleavage (Fig. 3.S19C), in agreement with our *in vivo* findings. When tested with a set of consecutive and non-consecutive double mutant crRNAs, LshC2c2 was unable to cleave the target RNA if the mismatches were positioned in the center, but not at the 5'- or 3'-end of the crRNA (Fig. 3.S19D, 3.S21A), further supporting the existence of a central seed region. Additionally, no cleavage activity was observed with crRNAs containing consecutive triple mismatches in the seed region (fig. 3.S21B).

3.3.7 C2c2 can be reprogrammed to mediate specific mRNA knockdown *in vivo*

Given the ability of C2c2 to cleave target ssRNA in a crRNA sequence-specific manner, we tested whether LshC2c2 could be reprogrammed to degrade selected non-phage ssRNA targets, and particularly mRNAs, *in vivo*. We co-transformed *E. coli* with a plasmid encoding LshC2c2 and a crRNA targeting the mRNA of red fluorescent protein (RFP) as well as a compatible plasmid expressing RFP (Fig. 3.5A). For OD-matched samples, we observed an approximately 20% to 92% decrease in RFP positive cells for crRNAs targeting protospacers flanked by C, A, or U PFSs (Fig. 3.5B, C). As a control, we tested crRNAs containing reverse complements (targeting the dsDNA plasmid) of the top performing RFP mRNA-targeting spacers. As expected, we observed no decrease in RFP fluorescence by these crRNAs (Fig. 3.5B). We also confirmed that mutation of the catalytic arginine residues in either HEPN domain to alanine precluded RFP knockdown (fig. 3.S22). Thus, C2c2 is capable of general retargeting to arbitrary ssRNA substrates, governed exclusively by predictable nucleic-acid interactions.

When we examined the growth of cells carrying the RFP-targeting spacer with the greatest level of RFP knockdown, we noted that the growth rate of these bacteria was substantially reduced (Fig. 3.5A, spacer 7). We investigated whether the effect on growth was mediated by the RFP mRNA-targeting activity of LshC2c2 by introducing an inducible-RFP plasmid and an RFP-targeting LshC2c2 locus into *E. coli*. Upon induction of RFP transcription, cells with RFP knockdown showed substantial growth suppression, not observed in non-targeting controls (Fig. 3.5D, E). This growth restriction was dependent on the level of the RFP mRNA, as controlled by the concentration of the inducer anhydrotetracycline. In contrast, in the absence of RFP transcription, we did not observe any growth restriction nor did we observe any transcription-dependent DNA targeting in our biochemical experiment (fig. 3.S11). These results indicate that RNA-targeting is likely the primary driver of this growth restriction phenotype. We therefore surmised that, in addition to the cleavage of the target

RNA, C2c2 CRISPR systems might prevent virus reproduction also via non-specific cleavage of cellular mRNAs, causing programmed cell death (PCD) or dormancy (Hayes and Van Melderen, 2011; Makarova et al., 2009).

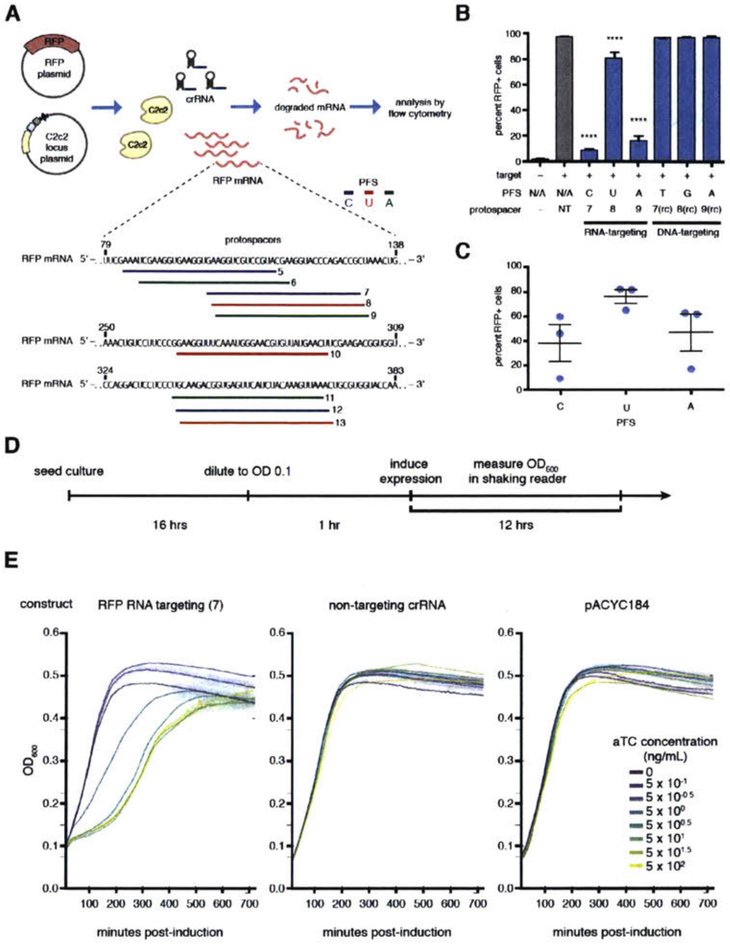


Figure 3.5: RFP mRNA knockdown by retargeting LshC2c2

A) Schematic showing crRNA-guided knockdown of RFP in *E. coli* heterologously expressing the LshC2c2 locus. Three RFP-targeting spacers were selected for each non-G PFS and each protospacer on the RFP mRNA is numbered.

B) RFP mRNA-targeting spacers effected RFP knockdown whereas DNA-targeting spacers (targeting the non-coding strand of the RFP gene on the expression plasmid, indicated as "rc" spacers) did not affect RFP expression. (n=3 biological replicates, ****, $p < 0.0001$ compared to non-targeting guide by ANOVA with multiple hypothesis correction. Bars represent mean ± s.e.m)

C) Quantification of RFP knockdown in *E. coli.* Three spacers each targeting C, U, or A PFS-flanking protospacers (9 spacers, numbered 5-13 as indicated in panel (A)) in the RFP mRNA were introduced and RFP expression was measured by flow cytometry. Each point on the scatter plot represents the average of three biological replicates and corresponds to a single spacer. Bars indicate the mean of 3 spacers for each PFS and errors bars are shown as the s.e.m.

D) Timeline of *E. coli* growth assay.

E) Effect of RFP mRNA targeting on the growth rate of *E. coli* transformed with an inducible RFP expression plasmid as well as the LshC2c2 locus with non-targeting, RNA targeting (spacer complementary to the RFP mRNA or RFP gene coding strand), and pACYC control plasmid at different anhydrotetracycline (aTc) concentrations.

3.3.8 C2c2 cleaves collateral RNA in addition to crRNA-targeted ssRNA

Cas9 and Cpf1 cleave DNA within the crRNA-target heteroduplex at defined positions, reverting to an inactive state after cleavage. In contrast, C2c2 cleaves the target RNA outside of the crRNA binding site at varying distances depending on the flanking sequence, presumably within exposed ssRNA loop regions (Figs. 3.3B, 3.3C, 3.S12A-D). This observed flexibility with respect to the cleavage distance led us to test whether cleavage of other, non-target ssRNAs also occurs upon C2c2 target binding and activation. Under this model, the C2c2-crRNA complex, once activated by binding to its target RNA, cleaves the target RNA as well as other RNAs non-specifically. We carried out *in vitro* cleavage reactions that included, in addition to LshC2c2 protein, crRNA and its target RNA, one of four unrelated RNA molecules without any complementarity to the crRNA guide (Fig. 3.6A). These experiments showed that, whereas the LshC2c2-crRNA complex did not mediate cleavage of any of the four collateral RNAs in the absence of the target RNA, all four were efficiently degraded in the presence of the target RNA (Figs. 3.6B, 3.S23A). Furthermore, R597A and R1278A HEPN mutants were unable to cleave collateral RNA (Fig. 3.S23B).

To further investigate the collateral cleavage and growth restriction *in vivo,* we hypothesized that if a PFS preference screen for LshC2c2 was performed in a transcribed region on the transformed plasmid, then we should be able to detect the PFS preference due to growth restriction induced by RNA targeting. We designed a protospacer site flanked by five randomized nucleotides at the 3' end in either a non-transcribed region or in a region transcribed from the *lac* promoter (fig. 3.S24A). The analysis of the depleted and enriched PFS sequences identified a H PFS only in the assay with the transcribed sequence but no discernable motif in the non-transcribed sequence (fig. 3.S24B-C).

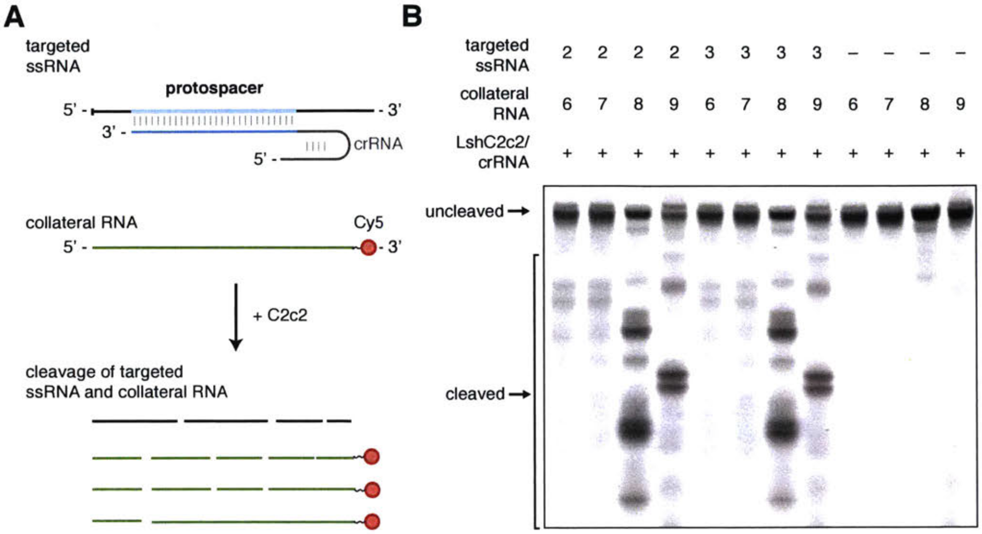


Figure 3.6: crRNA-guided ssRNA cleavage activates non-specific RNase activity.

A) Schematic of the biochemical assay used to detect crRNA-binding-activated non-specific RNase activity on non-crRNA-targeted collateral RNA molecules. The reaction consists of C2c2 protein, unlabeled crRNA, unlabeled target ssRNA, and a second ssRNA with 3' fluorescent labeling and is incubated for 3 hours. C2c2-crRNA mediates cleavage of the unlabeled target ssRNA as well as the 3'-end-labeled collateral RNA which has no complementarity to the crRNA.

B) Denaturing gel showing non-specific RNase activity against non-targeted ssRNA substrates in the presence of target RNA after 3 hours of incubation. The non-targeted ssRNA substrate is not cleaved in the absence of the crRNA-targeted ssRNA substrate.

These results suggest a HEPN-dependent mechanism whereby C2c2 in a complex with crRNA is activated upon binding to target RNA and subsequently cleaves non-specifically other available ssRNA targets. Such promiscuous RNA cleavage could cause cellular toxicity, resulting in the observed growth rate inhibition. These findings imply that, in addition to their likely role in direct suppression of RNA viruses, type VI CRISPR-Cas systems could function as mediators of a distinct variety of PCD or dormancy induction that is specifically triggered by cognate invader genomes (Fig. 3.7). Under this scenario, dormancy would slow the infection and supply additional time for adaptive immunity. Such a mechanism agrees with the previously proposed coupling of adaptive immunity and PCD during the CRISPR-Cas defensive response (Makarova et al., 2012).

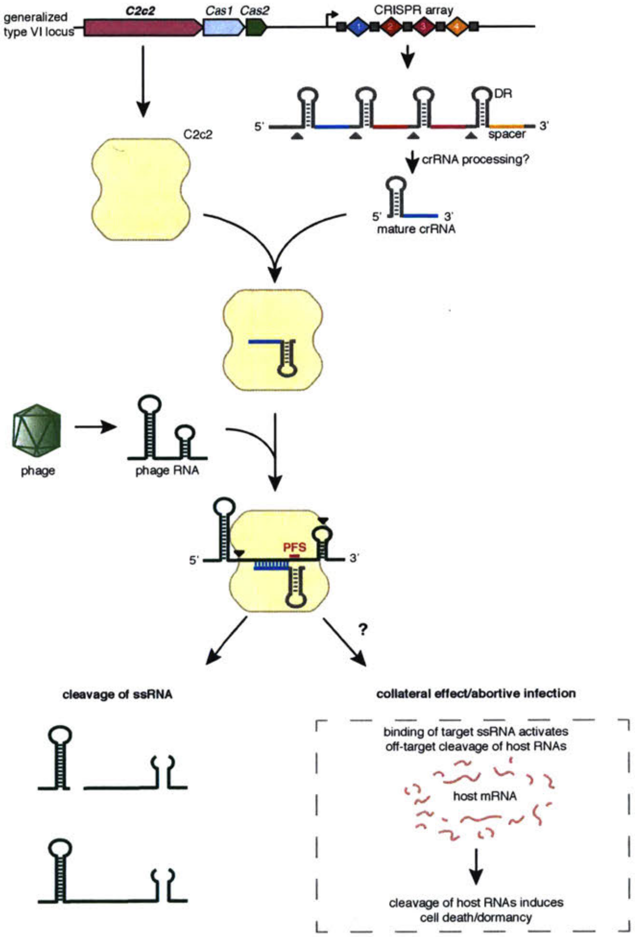

generalized type VI locus
C2c2
Cas1
Cas2
CRISPR array
1
2
3
4
DR
5'
3'
spacer
crRNA processing?
5'
3'
mature crRNA
C2c2
phage
phage RNA
5'
PFS
3'
?
cleavage of ssRNA
collateral effect/abortive infection
binding of target ssRNA activates off-target cleavage of host RNAs
host mRNA
cleavage of host RNAs induces cell death/dormancy

Figure 3.7: C2c2 as a putative RNA-targeting prokaryotic immune system

The C2c2-crRNA complex recognizes target RNA via base pairing with the cognate protospacer and cleaves the target RNA. In addition, binding of the target RNA by C2c2-crRNA activates a non-specific RNase activity which may lead to promiscuous cleavage of RNAs without complementarity to the crRNA guide sequence. Through this non-specific RNase activity, C2c2 may also cause abortive infection via programmed cell death or dormancy induction.

3.4 Discussion

In summary, the Class 2 type VI effector protein C2c2 is an RNA-guided RNase that can be efficiently programmed to degrade any ssRNA by specifying a 28-nt sequence on the crRNA (Fig. 3.7). C2c2 cleaves RNA through conserved basic residues within its two HEPN domains, in contrast to the catalytic mechanisms of other known RNases found in CRISPR-Cas systems (Benda et al., 2014; Tamulaitis et al., 2014). Alanine substitution of any of the four predicted HEPN domain catalytic residues converted C2c2 into an inactive programmable RNA-binding protein (dC2c2, analogous to dCas9). Many different spacer sequences work well in our assays although further screening will likely define properties and rules governing optimal function.

These results suggest a broad range of biotechnology applications and research questions (Abil and Zhao, 2015; Filipovska and Rackham, 2011; Mackay et al., 2011). For example, the ability of dC2c2 to bind to specified sequences could be used to (i) bring effector modules to specific transcripts to modulate their function or translation, which could be used for large-scale screening, construction of synthetic regulatory circuits and other purposes; (ii) fluorescently tag specific RNAs to visualize their trafficking and/or localization; (iii) alter RNA localization through domains with affinity for specific subcellular compartments; and (iv) capture specific transcripts (through direct pull down of dC2c2) to enrich for proximal molecular partners, including RNAs and proteins.

Active C2c2 also has many potential applications such as targeting a specific transcript for destruction, as performed here with RFP. In addition, C2c2, once primed by the cognate target, can cleave other (non-complementary) RNA molecules *in vitro* and inhibit cell growth *in vivo*. Biologically, this promiscuous RNase activity might reflect a PCD/dormancy-based protection mechanism of the type VI CRISPR-Cas systems (Fig. 3.7). Technologically, it might be used to trigger PCD or dormancy in specific cells such as cancer cells expressing a particular transcript, neurons of a given class, or cells infected by a specific pathogen.

Further experimental study is required to elucidate the mechanisms by which the C2c2 system acquires spacers and the classes of pathogens against which it protects bacteria. The presence of the conserved CRISPR adaptation module consisting of typical Cas1 and Cas2 proteins in the LshC2c2 locus suggests that it is capable of spacer acquisition. Although C2c2 systems lack reverse transcriptases, which mediate acquisition of RNA spacers in some type III systems (Silas et al., 2016), it is possible that additional host or viral factors could support RNA spacer acquisition. Additionally, or alternatively, type VI systems could acquire DNA spacers similar to other CRISPR-Cas variants but then target transcripts of the respective DNA genomes, eliciting PCD and abortive infection (Fig. 3.7).

The CRISPR-C2c2 system represent a distinct evolutionary path among Class 2 CRISPR-Cas systems. It is likely that other, broadly analogous Class 2 RNA-targeting immune systems exist, and further characterization of the diverse members of Class 2 systems will provide a deeper understanding of bacterial immunity and provide a rich starting point for the development of programmable molecular tools for *in vivo* RNA manipulation.

3.5 Experimental Procedures

3.5.1 Cloning of C2c2 locus and screening libraries for MS2 activity Screen

Genomic DNA from *Leptotrichia shahii DSM 19757* (ATCC, Manassas, VA) was extracted using the Blood & Cell Culture DNA Mini Kit (Qiagen, Hilden, Germany) and the C2c2 CRISPR locus was PCR amplified and cloned into a pACYC184 backbone with chloramphenicol resistance. For retargeting of the locus to MS2 phage or endogenous targets, the wild type spacers in the array were removed and replaced with a Eco31I landing site an additional spacer and a degenerate repeat, compatible with Golden Gate cloning.

A custom library consisting of all possible spacers targeting the genome of the bacteriophage MS2, excluding spacers containing the Eco31I restriction site, was synthesized by Twist Biosciences (San Francisco, CA), cloned into the retargeting backbone with Golden Gate cloning, transformed into Endura Duo electrocompetent cells (Lucigen, Middleton, WI) and subsequently purified using a NucleoBond Xtra MaxiPrep EF (Machery-Nagel, Düren, Germany).

3.5.2 Cloning of libraries and screening for β–lactamase and transcribed/non-transcribed PFS screens

Plasmid libraries for PFS screens were cloned from synthesized oligonucleotides (IDT, Coralville, IA) consisting of either 6 or 7 randomized nucleotides downstream of the spacer 1 target. To generate dsDNA fragments for cloning, these ssDNA oligonucleotides were annealed to a short primer for second strand synthesis by large Klenow fragment (New England Biolabs, Ipswich, MA). dsDNA fragments were Gibson cloned (New England Biolabs) into digested pUC19 at the 5'-region of β–lactamase, downstream of the lac promoter, or in a non-transcribed region of pUC19 and electroporated into Endura Duo electrocompetent cells (Lucigen). More than $5*10^6$ cells were collected, pooled, and harvested for plasmid DNA using a NucleoBond Xtra MaxiPrep EF (Machery-Nagel, Düren, Germany). To screen libraries, we co-transformed 50 ng of the pooled library and an equimolar amount of the LshC2c2 locus plasmid or pACYC184 plasmid control into *E. coli* cells (NovaBlue GigaSingles, EMD Millipore, Darmstadt, Germany). After transformation, cells were plated on ampicillin and chloramphenicol to select for both plasmids. After 16 hours of growth, $>1*10^6$ cells were harvested for plasmid DNA using a NucleoBond Xtra MaxiPrep EF (Machery-Nagel) The target PFS region was PCR amplified and sequenced using a MiSeq (Illumina, San Diego, CA) with a single-end 150 cycle kit.

3.5.3 Bacterial phage interference PFS screen assay

For the phage screen, 50ng of the plasmid library were transformed into NovaBlue(DE3) Competent Cells (EMD Millipore) followed by an outgrowth at 37°C for 30 minutes. Three different replicates of cells were then grown in Luria broth (LB, Miller's modification, 10g/L tryptone, 5g/L yeast extract, 5g/L NaCl, Sigma, St. Louis, MO) supplemented with 25 μg/mL chloramphenicol (Sigma) in a volume of 3.0mL for 30 minutes. Phage conditions were treated with $7*10^9$ (10^{-1} dilution), $7*10^7$ (10^{-3} dilution), or $7*10^5$ (10^{-5} dilution) PFU of Bacteriophage MS2 (ATCC). After 3 hours of shaking incubation at 37°C, samples were plated on LB-agar plates supplemented with chloramphenicol and harvested after 16 hours. DNA was extracted using NucleoBond Xtra MaxiPrep EF (Machery-Nagel), PCR amplified around the guide region, and sequenced using a MiSeq (Illumina) with a paired-end 150 cycle kit.

3.5.4 Bacterial phage interference assay for individual spacers

To test individual spacers for MS2 interference, the oligonucleotides encoding the spacer sequences flanked by Eco31I sites were ordered from IDT as complementary strands. The oligonucleotides (final concentration 10μM) were annealed in 10X T4 ligase buffer (New England Biolabs; final concentration 1X) supplemented with 5 units of T4 PNK (New England Biolabs). The oligonucleotides were phosphorylated by setting the thermocycler to 37°C for 30 minutes and then subsequently annealed by heating to 95°C for 5 minutes followed by a -5°C/minute ramp down to 25°C. Annealed oligos were then cloned into the locus backbone with Golden Gate cloning. Plasmids were transformed into C3000 strain *E. coli,* and the cultures were made competent with the Mix and

Go kit (Zymo Research, Irvine, CA). C3000 cells were seeded from an overnight culture grown to OD_{600} of 2, at which point they were diluted 1:13 in Top Agar (10g/L tryptone, 5g/L yeast extract, 10g/L NaCl, 7g/L agar) and poured on LB-chloramphenicol plates. Dilutions of MS2 phage in phosphate buffered saline were then spotted on the plates using a multichannel pipette, and plaque formation was recorded after overnight incubation.

3.5.5 RFP targeting assay

An ampicillin resistant RFP-expressing plasmid (pRFP) was transformed into DH5-alpha cells (New England Biolabs). Cells containing pRFP were then made chemically competent (Mix and Go, Zymo Research) to be used for downstream targeting experiments with pLshC2c2. Spacers targeting RFP mRNA were cloned into pLshC2c2 (as described above) and these plasmids were transformed into the chemically competent DH5-alpha pRFP cells. Cells were then grown overnight under double selection in LB and subjected to analysis by flow cytometry when they reached an OD600 of 4.0. Knockdown efficiency was quantified as the percent of RFP positive cells compared to a non-targeting spacer control (the endogenous LshC2c2 locus in pACYC184).

To interrogate the effect of LshC2c2 activity on the growth of the host cells, we created a TetR-inducible version of the RFP plasmid in pBR322 (pBR322_RFP). We transformed *E. coli* cells with this vector and then made them chemically competent (Mix and Go, Zymo Research) to prepare them for downstream experiments. We cloned pLshC2c2 plasmids with various spacers targeting RFP mRNA as well as their reverse complement controls and transformed them into *E. coli* cells carrying pBR322_RFP and streaked them on double-selection plates to maintain both plasmids. Colonies were then picked and grown overnight in LB with double selection. Bacteria were diluted to an OD_{600} of 0.1 and grown at 37C for 1 hour with chloramphenicol selection only. RFP expression was then induced using dilutions of anhydrotetracycline (Sigma) and OD_{600} measurements were taken every 6 minutes under continuous shaking in a Synergy 2 microplate reader (BioTek, Winooski, VT).

3.5.6 C2c2 protein purification

The mammalian codon-optimized gene for C2c2 (*Leptotrichia shahii)* was synthesized (GenScript, Jiangsu, China) and inserted into a bacterial expression vector (6-His-MBP-TEV, a pET based vector generously provided by Doug Daniels) using Golden Gate cloning. The LshC2c2 expression construct was transformed into One Shot® BL21(DE3)pLysE (Invitrogen, Carlsbad, CA) cells. 10mL of overnight culture were inoculated into 12 liters of Terrific Broth growth media (12g/L tryptone, 24g/L yeast extract, 9.4g/L K_2HPO, 2.2g/L KH_2PO_4, Sigma) supplemented with 100 µg/mL. Cells were then grown at 37 °C to a cell density of 0.2 OD_{600}, at which point the temperature was lowered to 21°C. At a cell density of 0.6 OD_{600}, MBP-LshC2c2 expression was induced by supplementing with IPTG to a final concentration of 500 µM. Induced culture was grown for 14-18 hours 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 1L of Lysis Buffer (50 mM Hepes pH 7, 2M NaCl, 5 mM $MgCl_2$, 20 mM imidazole) supplemented with protease inhibitors (cOmplete, EDTA-free, Roche Diagnostics Corporation, Indianapolis, IN). The resuspended cell paste was lysed by lysozyme (Sigma) addition and sonication (Sonifier 450, Branson, Danbury, CT). Lysate was cleared by centrifugation at 10,000g for 1 hour, and the supernatant was filtered through Stericup 0.45 micron filters (EMD Millipore). Filtered lysate was incubated with Ni-NTA superflow nickel resin (Qiagen) at 4°C for 1 hour with gentle agitation, and then applied to an Econo-column chromatography column (Bio-Rad Laboratories, Hercules, CA). Resin was washed with lysis Buffer and eluted with a gradient of imidazole. Fractions containing protein of the expected size for MBP-LshC2c2 were pooled and buffer exchanged into TEV Buffer (500 mM NaCl, 50 mM Hepes pH 7, 5 mM MgCl, 2 mM DTT) with Ultra-15 Centrifugal Filter Unit with 50 KDa cutoff (Amicon, EMD Millipore). TEV protease (Sigma) was added and incubated at 4°C overnight. After incubation, TEV cleavage was confirmed by SDS-PAGE and Coomassie staining, and the sample was 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, Chalfont Saint Giles, United Kingdom) via FPLC (AKTA Pure, GE Healthcare Life Sciences). The resulting fractions from gel filtration were tested for presence of LshC2c2 protein by SDS-PAGE; fractions containing LshC2c2 were pooled, buffer exchanged into Storage Buffer (1 M NaCl, 50 mM Tris-HCl pH 7.5, 5% glycerol, 2 mM DTT), concentrated, and either used directly for biochemical assays or frozen at -80°C for storage. To calculate the approximate size of recombinant LshC2c2, gel filtration standards were run on the same gel filtration column equilibrated in 2M NaCl, Hepes pH 7.0.

3.5.7 C2c2 HEPN mutant protein purification

Alanine mutants at each of the four HEPN catalytic residues were generated by Gibson cloning and transformed into One Shot® BL21(DE3)pLysE cells (Invitrogen). For each mutant, 6 L of Terrific Broth were used to generate cell paste. Protein purification was performed similarly to wild type C2c2 with exception of buffer composition being altered to increase stability of recombinant protein in solution. Detergent or glycerol was added to Lysis Buffer (50 mM Hepes pH 7, 1M NaCl, 5 mM $MgCl_2$, 20 mM imidazole, 1% Triton X-100), Imidazole Elution Buffer Buffer (50 mM Hepes pH 7, 1M NaCl, 5 mM $MgCl_2$, 200 mM imidazole, 0.01% Triton X-100, 10% glycerol) and TEV Buffer (500 mM NaCl, 50 mM Hepes pH 7, 5 mM MgCl, 1 mM DTT, 0.01% Triton X-100, 10% glycerol). In all situations where HEPN mutants were used biochemical analysis, wild type protein used for comparison was purified in the same manner.

3.5.8 Nucleic acid target preparation

DNA oligo templates for T7 transcription were ordered from IDT. Templates for crRNAs were annealed to a short T7 primer (final concentrations 10μM) and incubated with T7 polymerase overnight at 37°C using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). Target templates were PCR amplified to yield dsDNA and then incubated with T7 polymerase at 30°C overnight using the same kit.

5' end labeling was accomplished using the 5' oligonucleotide kit (VectorLabs, Burlingame, CA) and with a maleimide-IR800 probe (LI-COR Biosciences, Lincoln, NE). 3' end labeling was performed using a 3' oligonucleotide labeling kit (Sigma) using ddUTP-Cy5. Labeled probes were purified using Clean and Concentrator columns (Zymo Research).

dsRNA substrates were prepared by mixing 5'-end-labeled ssRNA targets with two-fold excess of non-labeled complementary ssRNA oligos in annealing buffer (30mM HEPES pH 7.4, 100 mM potassium acetate, and 2mM magnesium acetate). Annealing was performed by incubating the mixture for 1 minute at 95°C followed by a -1°C/minute ramp down to 23°C.

3.5.9 Nuclease Assay

Nuclease assays were performed with 160nM of end-labeled ssRNA target, 200nM purified LshC2c2, and 100nM crRNA, unless otherwise indicated, in nuclease assay buffer (40mM Tris-HCl, 60mM NaCl, 6mM MgCl2, pH 7.3). Reactions were allowed to proceed for 1 hour at 37°C (unless otherwise indicated) and were then quenched with proteinase buffer (proteinase K, 60mM EDTA, and 4M Urea) for 15 minutes at 37°C. The reactions were then denatured with 4.5M urea denaturing buffer at 95°C for 5 minutes. 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).

3.5.10 Electrophoretic mobility shift assay

Target ssRNA binding experiments were performed with a series of half-log complex dilutions (crRNA and LshC2c2) from 2μM to 0.2pM (or 1μM to 0.1pM in the case of R1278A LshC2c2). Binding assays were performed in nuclease assay buffer supplemented with 10mM EDTA to prevent cutting, 5% glycerol, and 10μg/mL heparin in order to avoid non-specific interactions of the complex with target RNA. Reactions were incubated at 37°C for 20 minutes 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).

3.5.11 Next-generation sequencing of in vitro cleaved RNA

In vitro nuclease assays were performed as described above using unlabeled ssRNA targets. After one hour, samples were quenched with proteinase K + EDTA and then column purified (Clean and

Concentrator, Zymo Research). The RNA samples were PNK treated in absence and presence of ATP to allow for the enrichment of 3'-P and 5'-OH ends, respectively. The samples were then polyphosphatase treated (Epicentre, Madison, WI) before being prepared for next-generation sequencing using the NEBNext Small RNA Library Prep Set for Illumina sequencing (New England Biolabs) with the PCR extension step increased to allow for longer templates to be included in the library. Libraries were sequenced on an MiSeq (Illumina) to sufficient depth and analyzed using the alignment tool BWA (Li and Durbin, 2009). Paired-end alignments were used to extract entire transcript sequences using Galaxy tools (https://usegalaxy.org/), and these sequences were analyzed using Geneious 8.1.5 (Biomatters, Auckland, New Zealand) and custom scripts (github.com/fengzhanglab).

3.5.12 In vitro co-transcriptional DNA cleavage assay

The *E. coli* RNAP co-transcriptional DNA cleavage assay was performed essentially as described previously (Samai et al., 2015). Briefly, 0.8pmol of ssDNA template strand were annealed with 1.6pmol of RNA in transcription buffer (from *E.coli* RNAP core enzyme, New England Biolabs) without magnesium to prevent RNA hydrolysis. 0.75ul of *E.coli* RNAP core enzyme and Magnesium were added and the reaction incubated at 25°C for 30min and then transferred to 37°C. 1pmol of freshly denatured nontemplate strand (NTS) were added and incubated at 37°C for 15min to obtain elongation complexes (ECs). 4pmol of LshC2C2-crRNA complexes along with 1.25 mM of RNTPs were added to the ECs and transcription was allowed to proceed for 1h at 37°C. DNA was resolved by denaturing gel electrophoresis on a 10% PAGE TBE-Urea (Invitrogen) gels following RNase and proteinase K treatment.

3.5.13 Computational analysis of in vivo PFS screens

To determine enriched spacers in the bacterial phage interference screens, spacer regions were extracted, counted, and normalized to total reads for each sample. For a given spacer, enrichment was measured as the $\log_2$ ratio compared to the no phage conditions, with a 1.0 psuedocount adjustment. 5' and 3' PFS regions from spacers above a 1.25 $\log_2$ enrichment threshold that occurred in all three biological replicates were used to generate sequence logos for the phage dilution samples (Crooks et al., 2004). Correlations between replicate conditions were measured using a Kendall's Tau rank correlation and the information coefficient, a mutual information based metric for ascertaining similarity (Konermann et al., 2015; Liberzon et al., 2015).

For the β-lactamase PFS screen and transcribed/non-transcribed pUC19 PFS screens, PFS regions were extracted, computationally collapsed to 5nt to increase coverage, counted, and normalized to total reads for each sample. For a given PFS, enrichment was measured as the log ratio compared to pACYC184 control, with a 0.01 psuedocount adjustment. PFSs above a 6 depletion threshold (β-lactamase screen), a 0.35 depletion threshold (transcribed pUC19 screen), or a 0.5 depletion threshold (non-transcribed pUC19 screen) that occurred in both biological replicates were collected and used to generate sequence logos (Crooks et al., 2004).

3.6 Acknowledgements

This work was published in (Abudayyeh et al., 2016). We would like to thank Paul Boutz, John Doench, Phillip Sharp, and Bernd Zetsche for helpful discussions and insights; Rich Belliveau for overall research support; Josh Francis and Daniel O'Connell for generous MiSeq instrument access; Doug Daniels and Colin Garvie for providing bacterial incubation space for protein purification; Rhiannon Macrae for critical reading of the manuscript; and the entire Zhang laboratory for support and advice. We would like to thank Nishant Ranu for generously providing pRFP and Doug Daniels for providing 6-His-MBP-TEV. \

Chapter 4

Nucleic acid detection with CRISPR-Cas13a/C2c2

This chapter is adapted from the following article:

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., et al. (2017b). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science

Contributions: Omar Abudayyeh and Jonathan Gootenberg are co-first authors (*). Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang conceived the study; Omar Abudayyeh and Jonathan Gootenberg designed and participated in all experiments; Jeong Wook Lee, Aaron Dy, and Nichole Daringer helped with pre-amplification strategies (e.g. NASBA and RPA) and experiments; Patrick Essletzbichler, Vanessa Verdine, and Julia Joung performed various SHERLOCK assays; Catherine Freije, Cameron Myhrvold, and Pardis Sabeti provided urine and serum Zika samples; Roby P. Bhattacharyya, Jonathan Livny, and Deborah Hung provided clinical bacterial isolates; Aviv Regev, Eugene Koonin, and Jim Collins helped with editing the manuscript; Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang wrote the manuscript that was read, edited and approved by all authors.

4.1 Abstract

Rapid, inexpensive, and sensitive nucleic acid detection may aid point-of-care pathogen detection, genotyping, and disease monitoring. The RNA-guided, RNA-targeting CRISPR effector Cas13a (previously known as C2c2) exhibits a "collateral effect" of promiscuous RNAse activity upon target recognition. We combine the collateral effect of Cas13a with isothermal amplification to establish a CRISPR-based diagnostic (CRISPR-Dx), providing rapid DNA or RNA detection with attomolar sensitivity and single-base mismatch specificity. We use this Cas13a-based molecular detection platform, termed SHERLOCK (Specific High Sensitivity Enzymatic Reporter UnLOCKing), to detect specific strains of Zika and Dengue virus, distinguish pathogenic bacteria, genotype human DNA, and identify cell-free tumor DNA mutations. Furthermore, SHERLOCK reaction reagents can be lyophilized for cold-chain independence and long-term storage, and readily reconstituted on paper for field applications.

4.2 Introduction

The ability to rapidly detect nucleic acids with high sensitivity and single-base specificity on a portable platform may aid in disease diagnosis and monitoring, epidemiology, and general laboratory tasks. Although methods exist for detecting nucleic acids (Du et al., 2017; Green et al., 2014; Kumar et al., 2014; Pardee et al., 2014; Pardee et al., 2016; Urdea et al., 2006), they have trade-offs among sensitivity, specificity, simplicity, cost, and speed. Microbial Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (CRISPR-Cas) adaptive immune systems contain programmable endonucleases that can be leveraged for CRISPR-based diagnostics (CRISPR-Dx). While some Cas enzymes target DNA (Shmakov et al., 2017b; Zetsche et al., 2015b), single effector RNA-guided RNases, such as Cas13a (previously known as C2c2) (Shmakov et al., 2017b), can be reprogrammed with CRISPR RNAs (crRNAs) to provide a platform for specific RNA sensing (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Shmakov et al., 2015; Smargon et al., 2017a). Upon recognition of its RNA target, activated Cas13a engages in "collateral" cleavage of nearby non-targeted RNAs (Abudayyeh et al., 2016). This crRNA-programmed collateral cleavage activity allows

Cas13a to detect the presence of a specific RNA *in vivo* by triggering programmed cell death (Abudayyeh et al., 2016) or *in vitro* by nonspecific degradation of labeled RNA (Abudayyeh et al., 2016; East-Seletsky et al., 2016). Here we describe SHERLOCK (Specific High Sensitivity Enzymatic Reporter UnLOCKing), an *in vitro* nucleic acid detection platform with attomolar sensitivity based on nucleic acid amplification and Cas13a-mediated collateral cleavage of a reporter RNA (East-Seletsky et al., 2016), allowing for real-time detection of the target (Fig. 4.1A).

4.3 Results

To achieve robust signal detection, we identified an ortholog of Cas13a from *Leptotrichia wadei* (LwCas13a), which displays greater RNA-guided RNase activity relative to *Leptotrichia shahii* Cas13a (LshCas13a) (Abudayyeh et al., 2016) (fig. 4.S1). LwCas13a incubated with ssRNA target 1 (ssRNA 1), crRNA, and reporter (quenched fluorescent RNA) (Fig. 1B) yielded a detection sensitivity of ~50 fM (Fig. 4.1C, 4.S2). Although this sensitivity is an improvement on previous studies with LbCas13a (East-Seletsky et al., 2016), attomolar sensitivity is required for many diagnostic applications (Barletta et al., 2004; Emmadi et al., 2011; Song et al., 2013). We therefore explored combining Cas13a-based detection with different isothermal amplification steps (fig. 4.S3, 4.S4A) (Compton, 1991; Piepenburg et al., 2006). Of the methods explored, recombinase polymerase amplification (RPA) (Piepenburg et al., 2006) afforded the greatest sensitivity and can be coupled with T7 transcription to convert amplified DNA to RNA for subsequent detection by LwCas13a. We refer to this combination of amplification by RPA, T7 RNA polymerase transcription of amplified DNA to RNA, and detection of target RNA by Cas13a collateral RNA cleavage-mediated release of reporter signal as SHERLOCK.

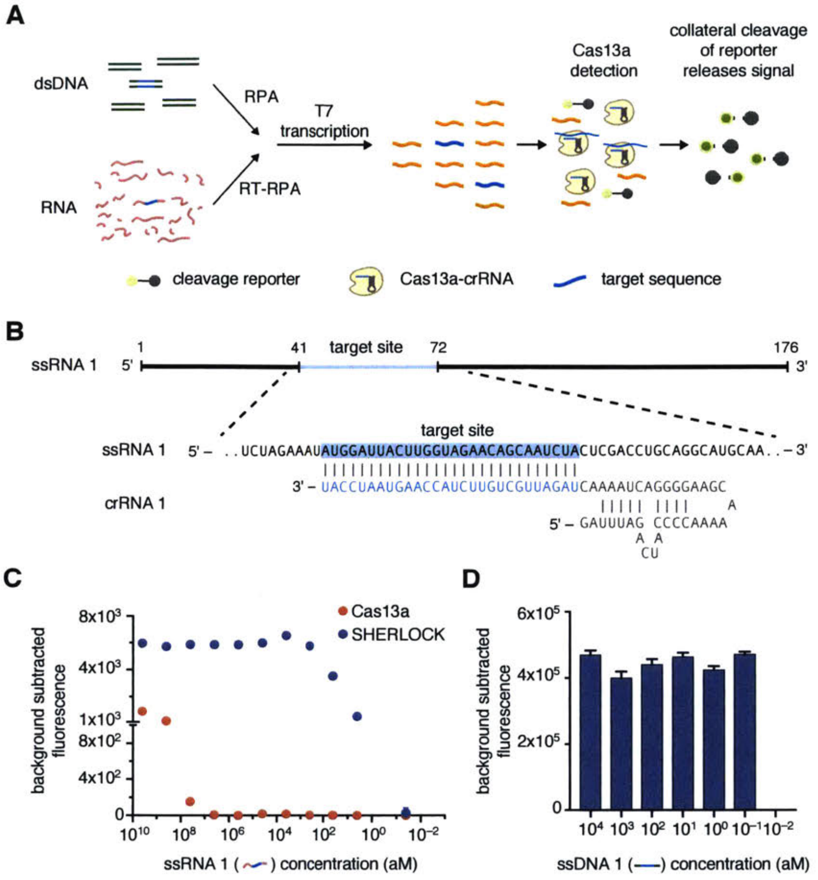


Figure 4.1: SHERLOCK is capable of single-molecule nucleic acid detection.

(A) Schematic of SHERLOCK.

(B) Schematic of ssRNA target detected via the Cas13a collateral detection. The target site is highlighted in blue.

(C) Cas13a detection of RNA with RPA amplification (SHERLOCK) can detect ssRNA target at concentrations down to ~2 aM, more sensitive than Cas13a alone. (n=4 technical replicates; bars represent mean ± s.e.m.)

(D) SHERLOCK is also capable of single-molecule DNA detection. (n=4 technical replicates; bars represent mean ± s.e.m.)

We first determined the sensitivity of SHERLOCK for detection of RNA (when coupled with reverse transcription) or DNA targets. We achieved single molecule sensitivity for both RNA and DNA, as verified by digital-droplet PCR (ddPCR) (Fig. 4.1C,D, 4.S4B,C). Attomolar sensitivity was maintained when we combined all SHERLOCK components in a single reaction, demonstrating the viability of this platform as a point-of-care (POC) diagnostic (fig. 4.S4D). SHERLOCK has similar levels of sensitivity as ddPCR and quantitative PCR (qPCR), two established sensitive nucleic acid detection approaches, whereas RPA alone was not sensitive enough to detect low levels of target (fig. 4.S5A-D). Moreover, SHERLOCK shows less variation than ddPCR, qPCR, and RPA, as measured by the coefficient of variation across replicates (fig. 4.S5E-F).

We next examined whether SHERLOCK would be effective in infectious disease applications that require high sensitivity. We produced lentiviruses harboring genome fragments of either Zika virus (ZIKV) or the related flavivirus Dengue (DENV) (Dejnirattisai et al., 2016) (Fig. 4.2A). SHERLOCK detected viral particles down to 2 aM and could discriminate between ZIKV and DENV (Fig. 4.2B). To explore the potential use of SHERLOCK in the field with paper-spotting and lyophilization (Pardee et al., 2016), we first demonstrated that Cas13a-crRNA complexes lyophilized and subsequently rehydrated could detect 20 fM of non-amplified ssRNA 1 (fig. 4.S6A) and that target detection was also possible on glass fiber paper (fig. 4.S6B). The other components of SHERLOCK are also amenable to freeze-drying: RPA is provided as a lyophilized reagent at ambient temperature, and we previously demonstrated that T7 polymerase tolerates freeze-drying (Pardee et al., 2014). In combination, freeze-drying and paper-spotting the Cas13a detection reaction resulted in comparable levels of sensitive detection of ssRNA 1 as aqueous reactions (fig. 4.S6C-E). Although paper-spotting and lyophilization slightly reduced the absolute signal of the readout, SHERLOCK (Fig. 4.2C) could readily detect mock ZIKV virus at concentrations as low as 20 aM (Fig. 4.2D).

SHERLOCK is also able to detect ZIKV in clinical isolates (serum, urine, or saliva) where titers can be as low as 2 x 10^3 copies/mL (3.2 aM) (Paz-Bailey et al., 2017). ZIKV RNA extracted from patient serum or urine samples and reverse transcribed into cDNA (Fig. 4.2E) could be detected at concentrations down to 1.25 x 10^3 copies/mL (2.1 aM), as verified by qPCR (Fig. 4.2F). Furthermore, the signal from patient samples was predictive of ZIKV RNA copy number and could be used to predict viral load (Fig. 4.S6F). To simulate sample detection without nucleic acid purification, we measured detection of ssRNA 1 spiked into human serum, and found that Cas13a could detect RNA in reactions containing as much as 2% serum (fig. 4.S6G).

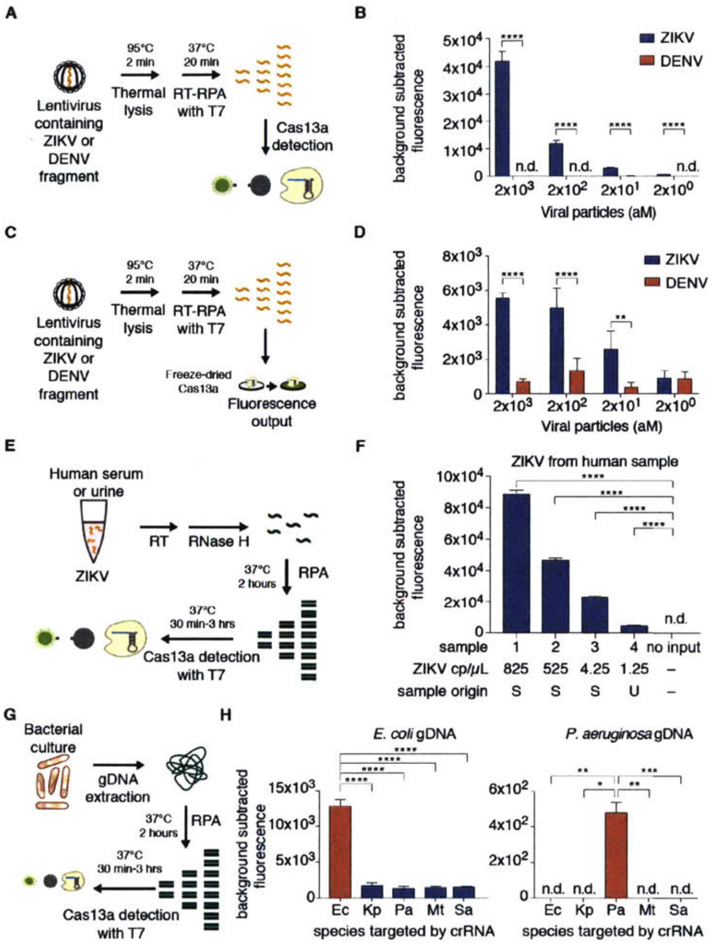


Figure 4.2: Cas13a detection can be used to sense viral and bacterial pathogens.

(A) Schematic of ZIKV RNA detection by SHERLOCK.

(B) SHERLOCK is capable of highly sensitive detection of the ZIKV lentiviral particles. (n=4 technical replicates, two-tailed Student t-test; ****, $p < 0.0001$; bars represent mean ± s.e.m.; n.d., not detected)

(C) Schematic of ZIKV RNA detection with freeze-dried Cas13a on paper

(D) Paper-based SHERLOCK is capable of highly sensitive detection of ZIKV lentiviral particles. (n=4 technical replicates, two-tailed Student t-test; **, $p < 0.01$; ****, $p < 0.0001$; bars represent mean ± s.e.m.)

(E) Schematic of SHERLOCK detection of ZIKV RNA isolated from human clinical samples.

(F) SHERLOCK is capable of highly sensitive detection of human ZIKV-positive serum (S) or urine (U) samples. Approximate concentrations of ZIKV RNA shown were determined by qPCR. (n=4 technical replicates, two-tailed Student t-test; ****, $p < 0.0001$; bars represent mean ± s.e.m.; n.d., not detected)

(G)Schematic of using SHERLOCK to distinguish bacterial strains using a universal 16S rRNA gene V3 RPA primer set.

(H)SHERLOCK achieves sensitive and specific detection of *E. coli* or *P. aeruginosa* gDNA. (n=4 technical replicates, two-tailed Student t-test; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$; bars represent mean ± s.e.m.). Ec, *Escherichia coli*; Kp, *Klebsiella pneumoniae*; Pa, *Pseudomonas aeruginosa*; Mt, *Mycobacterium tuberculosis*; Sa, *Staphylococcus aureus.*

Another important epidemiological application for CRISPR-dx is the identification of bacterial pathogens and detection of specific bacterial genes. We targeted the 16S rRNA gene V3 region, where conserved flanking regions allow universal RPA primers to be used across bacterial species and the variable internal region allows for differentiation of species. In a panel of five possible targeting crRNAs for different pathogenic strains and gDNA isolated from *E. coli* and *Pseudomonas aeruginosa* (Fig. 4.2G), SHERLOCK correctly genotyped strains and showed low cross-reactivity (Fig. 4.2H). Additionally, we were able to use SHERLOCK to distinguish between clinical isolates of *Klebsiella pneumoniae* with two different resistance genes: *Klebsiella pneumoniae* carbapenemase (KPC) and New Delhi metallo-beta-lactamase 1 (NDM-1) (Gupta et al., 2011) (fig. 4.S7).

To increase the specificity of SHERLOCK, we introduced synthetic mismatches in the crRNA:target duplex that enable LwCas13a to discriminate between targets that differ by a single-base mismatch (fig. 4.S8A,B). We designed multiple crRNAs with synthetic mismatches in the spacer sequences to detect either the African or American strains of ZIKV (Fig. 4.3A,B) and strain 1 or 3 of DENV (Fig. 4.3C,D). Synthetic mismatch crRNAs detected their corresponding strains with significantly higher signal (two-tailed Student t-test; $p < 0.01$) than the off-target strain, allowing for robust strain discrimination based off single mismatches (Fig. 4.3B,D, 4.S8C). Further characterization revealed that Cas13a detection achieves maximal specificity while maintaining on-target sensitivity when a mutation is in position 3 of the spacer and the synthetic mismatch is in position 5 (fig. 4.S9 and 4.S10).

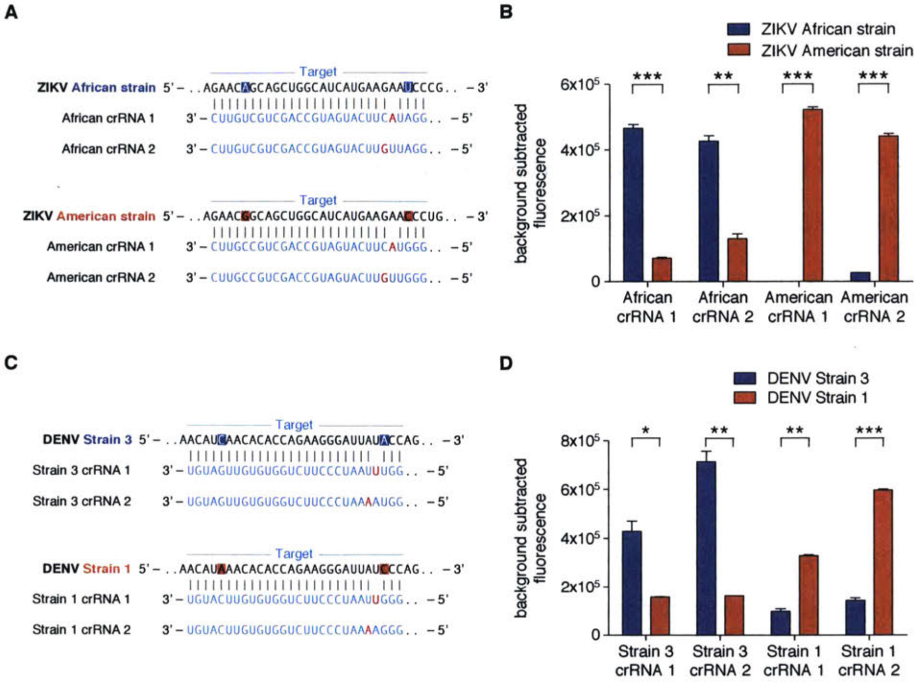


Figure 4.3: Cas13a detection can discriminate between similar viral strains.

(A)Schematic of ZIKV strain target regions and the crRNA sequences used for detection. SNPs in the target are highlighted red or blue and synthetic mismatches in the guide sequence are colored red.

(B)Highly specific detection of strain SNPs allows for the differentiation of ZIKV African versus American RNA targets using Cas13a. (n=2 technical replicates, two-tailed Student t-test; **, p < 0.01; ***, p < 0.001; bars represent mean ± s.e.m.)

(C)Schematic of DENV strain target regions and the crRNA sequences used for detection. SNPs in the target are highlighted red or blue and synthetic mismatches in the guide sequence are colored red.

(D)Highly specific detection of strain SNPs allows for the differentiation of DENV strain 1 versus strain 3 RNA targets using Cas13a. (n=2 technical replicates, two-tailed Student t-test; *, p < 0.05; **, p < 0.01; ***, p < 0.001; bars represent mean ± s.e.m.)

The ability to detect single-base differences opens the opportunity of using SHERLOCK for rapid human genotyping. We chose five loci spanning a range of health-related single-nucleotide polymorphisms (SNPs) and benchmarked SHERLOCK detection using 23andMe genotyping data as the gold standard at these SNPs (Eriksson et al., 2010) (Fig. 4.4A). We collected saliva from four human subjects with diverse genotypes across the loci of interest, and extracted genomic DNA either through column purification or direct heating for five minutes . SHERLOCK distinguished alleles with high significance and with enough specificity to infer both homozygous and heterozygous genotypes (Fig. 4.4B, 4.S11, 4.S12).

Finally, we sought to determine if SHERLOCK could detect low frequency cancer mutations in cell free (cf) DNA fragments, which is challenging because of the high levels of wild-type DNA in patient blood (Bettegowda et al., 2014; Newman et al., 2014; Qin et al., 2016). We first found that SHERLOCK could detect ssDNA 1 at attomolar concentrations diluted in a background of genomic DNA (fig. 4.S13A). Next, we found that SHERLOCK was also able to detect single nucleotide polymorphism (SNP)-containing alleles (fig. 4.S13B,C) at levels as low as 0.1% of background DNA, which is in the clinically relevant range. We then demonstrated that SHERLOCK could detect two different cancer mutations, EGFR L858R and BRAF V600E, in mock cfDNA samples with allelic fractions as low as 0.1% (Fig. 4.4C-F) .

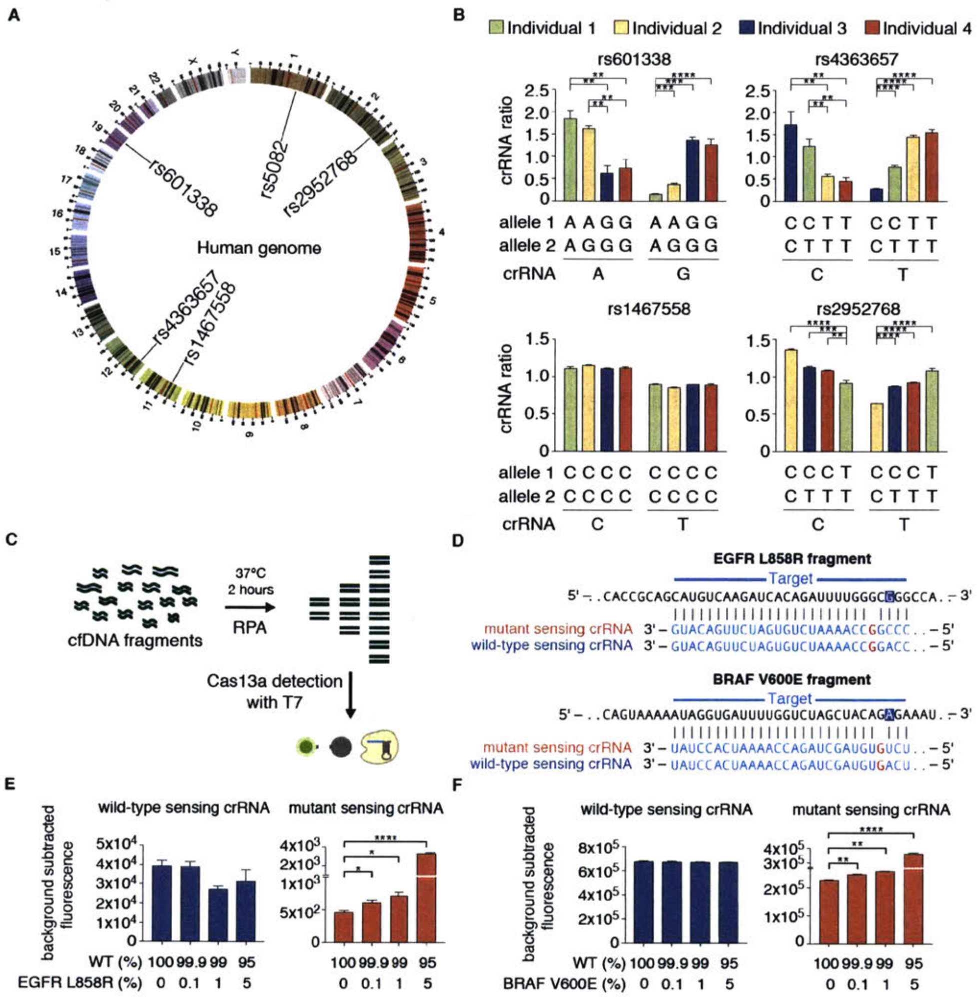


Figure 4.4: SHERLOCK can discriminate SNPs for human genotyping and cell-free allele DNA detection.

(A)Circos plot showing location of human SNPs detected with SHERLOCK.

(B)SHERLOCK can correctly genotype four different individuals at four different SNP sites in the human genome. The genotypes for each individual and identities of allele-sensing crRNAs are annotated below each plot. (n=4 technical replicates, two-tailed Student t-test; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$; bars represent mean ± s.e.m.)

(C)Schematic of cell-free DNA detection of cancer mutations using SHERLOCK.

(D)Sequences of two genomic loci assayed for cancer mutations in cell-free DNA. Shown are the target genomic sequence with the SNP highlighted in blue and the mutant/wild-type sensing crRNA sequences with synthetic mismatches colored in red.

(E,F) Cas13a can detect the mutant minor allele in mock cell-free DNA samples for the EGFR L858R (E) or the BRAF V600E (F) minor allele. (n=4 technical replicates, two-tailed Student t-test; *, $p < 0.05$; **, $p < 0.01$; ****, $p < 0.0001$; bars represent mean ± s.e.m.)

4.4 Conclusion

The SHERLOCK platform lends itself to further applications including (i) general RNA/DNA quantitation in lieu of specific qPCR assays, such as TaqMan, (ii) rapid, multiplexed RNA expression detection, and (iii) other sensitive detection applications, such as detection of nucleic acid contamination. Additionally, Cas13a could potentially detect transcripts within biological settings and track allele-specific expression of transcripts or disease-associated mutations in live cells. We have shown that SHERLOCK is a versatile, robust method to detect RNA and DNA, suitable for rapid diagnoses including infectious disease applications and sensitive genotyping. A SHERLOCK paper test can be redesigned and synthesized in a matter of days for as low as $0.61/test with confidence, as almost every crRNA tested resulted in high sensitivity and specificity. These qualities highlight the power of CRISPR-Dx and open new avenues for rapid, robust and sensitive detection of biological molecules.

4.5 Experimental Procedures

4.5.1 Cloning of Cas13a loci and proteins for expression

For the bacterial *in vivo* efficiency assay, Cas13a proteins from *Leptotrichia wadei* and *Leptotrichia shahii* were ordered as codon-optimized genes for mammalian expression (Genscript, Jiangsu, China) and cloned into pACYC184 backbones along with the corresponding direct repeats flanking either a beta-lactamase targeting or non-targeting spacer. Spacer expression was driven by a J23119 promoter.

For protein purification, mammalian codon-optimized Cas13a proteins were cloned into bacterial expression vector for protein purification (6x His/Twin Strep SUMO, a pET-based expression vector received as a gift from Ilya Finkelstein, University of Texas-Austin).

4.5.2 Bacterial *in vivo* Cas13a efficiency assay

LwCas13a and LshCas13a *in vivo* efficiency plasmids and a previously described beta-lactamase plasmid (Abudayyeh et al., 2016) were co-transformed into NovaBlue Singles competent cells (Millipore) at 90 ng and 25 ng, respectively. After transformation, dilutions of cells were plated on ampicillin and choramphicol LB-agar plate and incubated overnight at 37°C. Colonies were counted the next day.

4.5.3 LwCas13a protein purification

Cas13a bacterial expression vectors were transformed into Rosetta™ 2(DE3)pLysS Singles Competent Cells (Millipore). A 16 mL starter culture was grown overnight in Terrific Broth 4 growth media (12 g/L tryptone, 24 g/L yeast extract, 9.4 g/L K_2HPO, 2.2 g/L KH_2PO_4, Sigma) (TB), which was used to inoculate 4 L of TB for growth at 37°C and 300 RPM until an OD600 of 0.6. At this time, protein expression was induced by supplementation with IPTG (Sigma) to a final concentration of 500 uM, and cells were cooled to 18°C for 16 h for protein expression. Cells were then centrifuged at 5200 g for 15 min at 4°C. Cell pellet was harvested and stored at -80°C for later purification.

All subsequent steps of the protein purification were performed at 4°C. Cell pellet was crushed and resuspended in lysis buffer (20 mM Tris-HCl, 500 mM NaCl, 1 mM DTT, pH 8.0) supplemented with protease inhibitors (Complete Ultra EDTA-free tablets), lysozyme, and benzonase followed by sonication (Sonifier 450, Branson, Danbury, CT) with the following conditions: amplitude of 100 for 1 second on and 2 seconds off with a total sonication time of 10 min. Lysate was cleared by centrifugation for 1 hr at 4°C at 10,000 g and the supernatant was filtered through a Stericup 0.22 μm filter (EMD Millipore). Filtered supernatant was applied to StrepTactin Sepharose (GE) and incubated with rotation for 1 hr followed by washing of the protein-bound StrepTactin resin three times in lysis buffer. The resin was resuspended in SUMO digest buffer (30 mM Tris-HCl, 500 mM NaCl 1 mM DTT, 0.15% Igepal (NP-40), pH 8.0) along with 250 Units of SUMO protease (ThermoFisher) and incubated overnight at 4°C with rotation. Digestion was confirmed by SDS-PAGE and Coomassie Blue staining and the protein eluate was isolated by spinning the resin down.

For further cation exchange and gel filtration purification, protein was loaded onto a 5 mL HiTrap SP HP cation exchange column (GE Healthcare Life Sciences) via FPLC (AKTA PURE, GE Healthcare Life Sciences) and eluted over a salt gradient from 130 mM to 2M NaCl in elution buffer (20 mM Tris-HCl, 1 mM DTT, 5% glycerol, pH 8.0). The resulting fractions were tested for presence of LwCas13a by SDS-PAGE, and fractions containing the protein were pooled and concentrated via a Centrifugal Filter Unit to 1 mL in S200 buffer (10 mM HEPES, 1 M NaCl, 5 mM $MgCl_2$, 2 mM DTT, pH 7.0). The concentrated protein was loaded onto a gel filtration column (Superdex® 200 Increase 10/300 GL, GE Healthcare Life Sciences) via FPLC. The resulting fractions from gel filtration were analyzed by SDS-PAGE and fractions containing LwCas13a were pooled and buffer exchanged into Storage Buffer (600 mM NaCl, 50 mM Tris-HCl pH 7.5, 5% glycerol, 2mM DTT) and frozen at -80°C for storage.

4.5.4 Nucleic acid target and crRNA preparation

Nucleic acid targets were PCR amplified with KAPA Hifi Hot Start (Kapa Biosystems), gel extracted, and purified using MinElute gel extraction kit (Qiagen). Purified dsDNA was incubated with T7 polymerase overnight at 30°C using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs) and RNA was purified with the MEGAclear Transcription Clean-up kit (Thermo Fisher)

For preparation of crRNAs, constructs were ordered as DNA (Integrated DNA Technologies) with an appended T7 promoter sequence. crRNA DNA was annealed to a short T7 primer (final concentrations 10 uM) and incubated with T7 polymerase overnight at 37°C using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). crRNAs were purified using RNAXP clean beads (Beckman Coulter) at 2x ratio of beads to reaction volume, with an additional 1.8x supplementation of isopropanol (Sigma)

4.5.5 NASBA isothermal amplification

NASBA was performed as previously described (Pardee et al., 2016). For a 20 µL total reaction volume, 6.7 µL of reaction buffer (Life Sciences, NECB-24), 3.3 µL of Nucleotide Mix (Life Sciences, NECN-24), 0.5 µL of nuclease-free water, 0.4 µL of 12.5 µM NASBA primers, 0.1 µL of RNase inhibitor (Roche, 03335402001) and 4 µL of RNA input (or water for the negative control) were assembled at 4°C and incubated 65°C for 2 min and then 41°C for 10 min. 5 µL of enzyme mix (Life Sciences, NEC-1-24) was added to each reaction, and the reaction mixture was incubated at 41°C for 2 hr.

4.5.6 Recombinase Polymerase Amplification

Primers for RPA were designed using NCBI Primer-BLAST (Ye et al., 2012) using default parameters, with the exception of amplicon size (between 100 and 140 nt), primer melting temperatures (between 54°C and 67°C), and primer size (between 30 and 35 nt). Primers were then ordered as DNA (Integrated DNA Technologies).

RPA and RT-RPA reactions run were as instructed with TwistAmp® Basic or TwistAmp® Basic RT (TwistDx), respectively, with the exception that 280 mM MgAc was added prior to the input template. Reactions were run with 1 µL of input for 2 hr at 37°C, unless otherwise described.

4.5.7 LwCas13a collateral detection

Detection assays were performed with 45 nM purified LwCas13a, 22.5 nM crRNA, 125 nM quenched fluorescent RNA reporter (RNAse Alert v2, Thermo Scientific), 2 µL murine RNase inhibitor (New England Biolabs), 100 ng of background total human RNA (purified from HEK293FT culture), and varying amounts of input nucleic acid target, unless otherwise indicated, in nuclease assay buffer (40 mM Tris-HCl, 60 mM NaCl, 6 mM $MgCl_2$, pH 7.3). If the input was amplified DNA including a T7 promoter from a RPA reaction, the above Cas13a reaction was modified to include 1 mM ATP, 1 mM GTP, 1 mM UTP, 1 mM CTP, and 0.6 µL T7 polymerase mix (New England Biolabs). Reactions were allowed to proceed for 1-3 hr at 37°C (unless otherwise indicated) on a fluorescent plate reader (BioTek) with fluorescent kinetics measured every 5 min.

The single reaction combining RPA-DNA amplification, T7 polymerase conversion of DNA to RNA and Cas13a detection was performed by integrating the reaction conditions above with the RPA mix. Briefly, a 50 µL single reaction assay consisted of 0.48 µM forward primer, 0.48 µM reverse primer, 1x RPA rehydration buffer, varying amounts of DNA input, 45 nM LwCas13a recombinant protein, 22.5 nM crRNA, 250 ng background total human RNA, 200 nM substrate reporter (RNase alert v2), 4 µL murine RNase inhibitor (New England Biolabs), 2 mM ATP, 2 mM GTP, 2 mM UTP, 2 mM CTP, 1 µL T7 polymerase mix (New England Biolabs), 5 mM $MgCl_2$, and 14 mM MgAc.

4.5.8 Digital droplet PCR quantification

To confirm the concentration of ssDNA 1 and ssRNA 1 standard dilutions used in Figure 1C-D and for comparison of SHERLOCK sensitivity, we performed digital-droplet PCR (ddPCR). For DNA quantification, droplets were made using the ddPCR Supermix for Probes (no dUTP) (BioRad) with PrimeTime qPCR probes/primer assays (IDT) designed to target the ssDNA 1 sequence. For RNA quantification, droplets were made using the one-step RT-ddPCR kit for probes with PrimeTime qPCR probes/primer assays designed to target the ssRNA 1 sequence. Droplets were generated in either case using the QX200 droplet generator (BioRad) and transferred to a PCR plate. Droplet-based amplification was performed on a thermocycler as described in the kit protocol and nucleic acid concentrations were subsequently determined via measurement on a QX200 droplet reader.

4.5.9 Quantitative PCR (qPCR) analysis with TaqMan probes

To compare SHERLOCK quantification with other established methods, we performed qPCR on a dilution series of ssDNA 1. A TaqMan probe and primer set were designed against ssDNA 1 and synthesized with IDT. Assays were performed using the TaqMan Fast Advanced Master Mix (Thermo Fisher) and measured on a Roche LightCycler 480.

4.5.10 Real-time RPA with SYBR Green II

To compare SHERLOCK quantification with other established methods, we performed RPA on a dilution series of ssDNA 1. To quantitate accumulation of DNA in real-time, we added 1x SYBR Green II (Thermo Fisher) to the typical RPA reaction mixture described above, which provides a fluorescent signal that correlates with the amount of nucleic acid. Reactions were allowed to proceed for 1 hr at 37°C on a fluorescent plate reader (BioTek) with fluorescent kinetics measured every 5 min.

4.5.11 SHERLOCK freeze-drying and paper deposition

Glass fiber filter paper (Whatman, 1827-021) was autoclaved for 90 min (Consolidated Stills and Sterilizers, MKII) and blocked in 5% nuclease-free BSA (EMD Millipore, 126609-10GM) overnight. After rinsing the paper once with nuclease-free water (Life technologies, AM9932), RNases were removed via incubation with 4% RNAsecure^{TM} (Life technologies, AM7006) at 60°C for 20 min, and the paper was rinsed three more times with nuclease-free water to remove traces of RNAsecure. Treated papers were dried for 20 min at 80°C on a hot plate (Cole-Parmer, IKA C-Mag HS7) prior to use. 1.8 µL of Cas13a reaction mixture as indicated earlier was put onto the disc (2 mm) that was placed in black, clear bottom 384-well plate (Corning, 3544). For the freeze-dried test of SHERLOCK, the plate containing reaction mixture discs was flash frozen in liquid nitrogen and was freeze-dried overnight as previously described (Pardee et al., 2016). RPA samples were diluted 1:10 in nuclease-free water, and 1.8 µL of the mixture was loaded onto the paper discs and incubated at 37°C using a plate reader (BioTek Neo).

4.5.12 Lentivirus Preparation and Processing

Lentivirus preparation and processing was performed as previously described (1). Briefly, 10 μg pSB700 derivatives that include a ZIKV or DENV RNA fragment, 7.5 μg psPAX2, and 2.5 μg pMD2.G were transfected into HEK293FT cells (Life Technologies, R7007) using the HeBS-$CaCl_2$ method. 28 hr after changing media to fresh DMEM supplemented with 10% FBS, 1% penicillin-streptomycin and 4 mM GlutaMAX (ThermoFisher Scientific), the supernatant was filtered using a 0.45 μm syringe filter. ViralBind Lentivirus Purification Kit (Cell Biolabs, VPK-104) and Lenti-X Concentrator (Clontech, 631231) were used to purify and prepare lentiviruses from the supernatant. Viral concentration was quantified using QuickTiter Lentivirus Kit (Cell Biolabs, VPK-112). Viral samples were spiked into 7% human serum (Sigma, H4522), were heated to 95°C for 2 min and were used as input to RPA.

4.5.13 Isolation and cDNA purification of ZIKV human serum samples

Suspected ZIKV positive human serum or urine samples were inactivated with AVL buffer (Qiagen) and isolation of RNA was achieved with QIAamp Viral RNA minikit (Qiagen). Isolated RNA was converted into cDNA by mixing random primers, dNTPs, and sample RNA followed by heat denaturation for 7 min at 70 ℃. Denatured RNA was then reverse transcribed with Superscript III (Invitrogen) incubated at 22-25 ℃ for 10 min, 50 ℃ for 45 min, 55 ℃ for 15 min, and 80 ℃ for 10 min. cDNA was then incubated for 20 min at 37 ℃ with RNAse H (New England Biolabs) to destroy RNA in the RNA:cDNA hybrids.

4.5.14 Bacterial genomic DNA extraction

For experiments involving CRE detection, bacterial cultures were grown in lysogeny broth (LB) to mid-log phase, then pelleted and subjected to gDNA extraction and purification using the Qiagen DNeasy Blood and Tissue Kit, using the manufacturer's protocol for either Gram negative or Gram positive bacteria, as appropriate. gDNA was quantified by the Quant-It dsDNA (Thermo Scientific) assay on a Qubit fluorometer (Thermo Scientific) and its quality assessed via 200-300 nm absorbance spectrum on a Nanodrop spectrophotometer.

For experiments discriminating between *E. coli* and *P. aeruginosa,* bacterial cultures were grown to early stationary phase in Luria-Bertani (LB) broth. 1.0 mL of both *E. coli* and *P. aeruginosa* were processed using the portable PureLyse bacteria gDNA extraction kit (Claremont BioSolutions). 1X binding buffer was added to the bacterial culture before passing through the battery-powered lysis cartridge for three minutes. 0.5X binding buffer in water was used as a wash solution before eluting with 150 μL of water.

4.5.15 Genomic DNA extraction from human saliva

2 mL of saliva was collected from volunteers, who were restricted from consuming food or drink 30 min prior to collection. Samples were then processed using QIAamp® DNA Blood Mini Kit (Qiagen) as recommended by the kit protocol. For boiled saliva samples, 400 μL of phosphate buffered saline (Sigma) was added to 100 μL of volunteer saliva and centrifuged for 5 min at 1800 g. The supernatant was decanted and the pellet was resuspended in phosphate buffered saline with 0.2% Triton X-100 (Sigma) before incubation at 95°C for 5 min. 1 μL of sample was used as direct input into RPA reactions.

4.5.16 Synthetic standards for human genotyping

To create standards for accurate calling of human sample genotypes, we designed primers around the SNP target to amplify ~200 bp regions from human genomic DNA representing each of the two homozygous genotypes. The heterozygous standard was then made by mixing the homozygous standards in a 1:1 ratio. These standards were then diluted to equivalent genome concentrations (~0.56 fg/μL) and used as input for SHERLOCK alongside real human samples.

4.5.17 Detection of tumor mutant cell free-DNA (cfDNA)

Mock cfDNA standards simulating actual patient cfDNA samples were purchased from a commercial vendor (Horizon Discovery Group). These standards were provided as four allelic fractions (100%

WT and 0.1%, 1%, and 5% mutant) for both the BRAF V600E and EGFR L858R mutants. 3 μL of these standards were provided as input to SHERLOCK.

4.5.18 Analysis of SHERLOCK fluorescence data

To calculate background subtracted fluorescence data, the initial fluorescence of samples was subtracted to allow for comparisons between different conditions. Fluorescence for background conditions (either no input or no crRNA conditions) were subtracted from samples to generate background subtracted fluorescence.

crRNA ratios for SNP or strain discrimination were calculated to adjust for sample-to-sample overall variation as follows:

$$crRNA\ A_i\ ratio = \frac{(m+n)A_i}{\sum_{i=1}^{m} A_i + \sum_{i=1}^{n} B_i}$$

where A_i and B_i refer to the SHERLOCK intensity values for technical replicate i of the crRNAs sensing allele A or allele B, respectively, for a given individual. Since we typically have four technical replicates per crRNA, m and n are equal to 4 and the denominator is equivalent to the sum of all eight of the crRNA SHERLOCK intensity values for a given SNP locus and individual. Because there are two crRNAs, the crRNA ratio average across each of the crRNAs for an individual will always sum to two. Therefore, in the ideal case of homozygosity, the mean crRNA ratio for the positive allele crRNA will be two and the mean crRNA ratio for the negative allele crRNA will be zero. In the ideal case of heterozygosity, the mean crRNA ratio for each of the two crRNAs will be one.

4.6 Acknowledgements

This work was published in (Gootenberg et al., 2017c). We thank Fei Chen, Victor Rusu, Raj Gupta, Doug Daniels, Colin Garvie, Ilya Finkelstein, Viktor Adalsteinsson, A. Das, Eric Lander, Rhiannon Macrae, and Rich Belliveau for discussions and support. Human genotyping data was collected with

informed consent by the subjects and in consent with the guidelines of the approved MIT IRB protocol IRB-4062.

Chapter 5

Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6

This chapter is adapted from the following article:

Gootenberg, J.S.*, **Abudayyeh, O.O.***, Kellner, M.J., Joung, J., Collins, J.J., and Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science.

Contributions: Omar Abudayyeh and Jonathan Gootenberg are co-first authors (*). Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang conceived and designed the study. Omar Abudayyeh, Jonathan Gootenberg, and Max Kellner participated in the design and execution of all experiments. Max performed most SHERLOCK assays and protein purifications; Omar and Jonathan performed all lateral flow experiments; Julia Joung designed and performed the RNA motif screens and analysis. Omar Abudayyeh, Jonathan Gootenberg, Max Kellner, James Collins, and Feng Zhang wrote the paper with contributions from all authors.

5.1 Abstract

Rapid detection of nucleic acids is integral for clinical diagnostics and biotechnological applications. We recently developed a platform termed SHERLOCK (Specific High Sensitivity Enzymatic Reporter UnLOCKing) that combines isothermal pre-amplification with Cas13 to detect single molecules of RNA or DNA. Through characterization of CRISPR enzymology and application development, we report here four advances integrated into SHERLOCKv2: 1) 4-channel single reaction multiplexing using orthogonal CRISPR enzymes; 2) quantitative measurement of input down to 2 aM; 3) 3.5-fold increase in signal sensitivity by combining Cas13 with Csm6, an auxilary CRISPR-associated enzyme; and 4) lateral flow read-out. SHERLOCKv2 can detect Dengue or Zika virus ssRNA as well as mutations in patient liquid biopsy samples via lateral flow, highlighting its potential as a multiplexable, portable, rapid, and quantitative detection platform of nucleic acids.

5.2 Introduction

Versatile, rapid, and portable sensing of nucleic acids is vital for applications in human health. The RNA-targeting CRISPR-associated enzyme Cas13(Abudayyeh et al., 2016; Shmakov et al., 2015) has recently been adapted for such purpose. This detection platform, termed SHERLOCK (Specific High Sensitivity Enzymatic Reporter UnLOCKing) (Gootenberg et al., 2017c), can discriminate between inputs that differ by a single nucleotide at very low concentrations and can be lyophilized for portable deployment. However, this technology has several limitations, including the lack of quantitation and reliance on fluorescent detection equipment for readout. Here, we extend the SHERLOCK technology to address these limitations and further develop the utility of this platform.

5.3 Results

Many applications require detection of more than one target molecule in a single reaction, and we therefore sought to create a multiplex platform that relies on unique cleavage preferences of Cas enzymes(Abudayyeh et al., 2016; East-Seletsky et al., 2017; East-Seletsky et al., 2016; Gootenberg et al., 2017c). To identify possible candidate enzymes compatible with multiplexing, we biochemically characterized three members of the CRISPR-Cas13a family and fourteen members of the CRISPR-Cas13b family(Shmakov et al., 2017a; Smargon et al., 2017b) (fig. 5.S1, 5.S2). We profiled cleavage preferences on homopolymer reporters, and found that most orthologs preferred either uridine, a combination of bases, or adenine (fig. 5.S3) and cleavage could be improved with buffer and crRNA design optimization (fig. 5.S4-7). Among the adenine cleaving enzymes, PsmCas13b was more sensitive than LbaCas13a (fig. 5.S8). We refined the cleavage sequence preferences by evaluating collateral activity across di-nucleotide motifs (Fig. 5.1A), finding a large diversity of di-nucleotide cleavage motif preferences (figs. 5.S9-10). From these di-nucleotide cleavage screens, we found that the activities of LwaCas13a, CcaCas13b, LbaCas13a and PsmCas13b could all be independently measured with the four di-nucleotide reporters AU, UC, AC, and GA, respectively (Fig. 5.1B and fig. 5.S11). Additionally, using a random *in vitro* RNA library motif cleavage screen, we identified numerous RNA 6-mers that allowed for further orthogonality between Cas13 enzymes (fig. 5.S12-15).

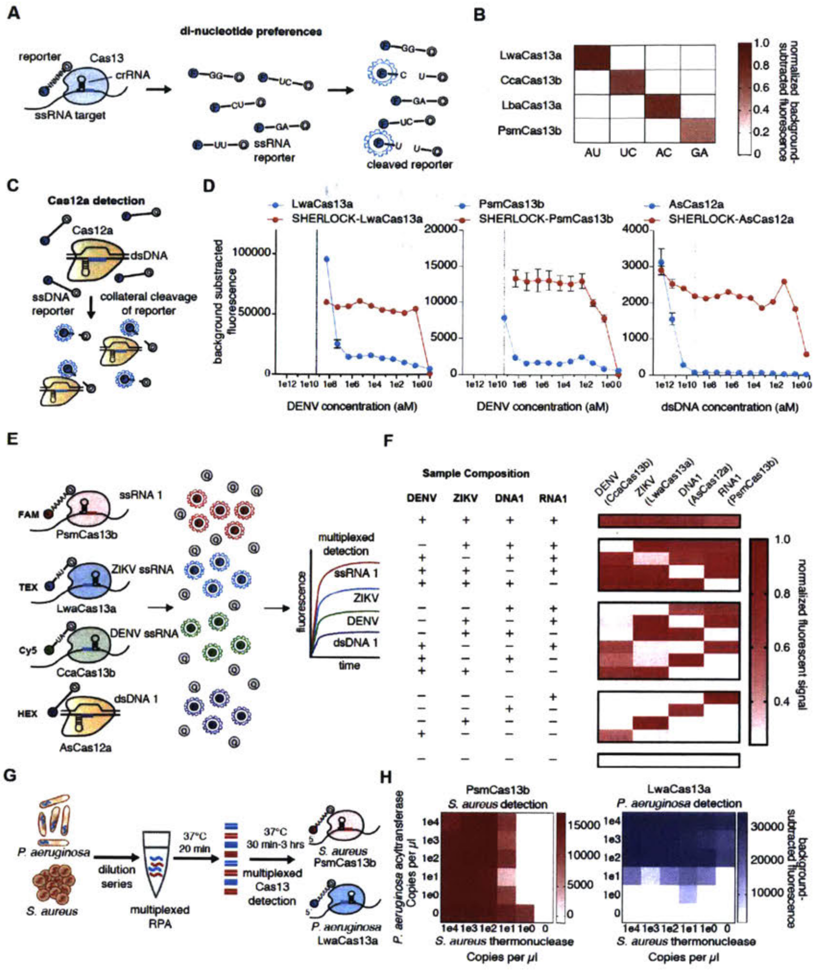


Figure 5.1: Multiplexed SHERLOCK detection with orthogonal collateral activity.

A)Schematic of assay for determining di-nucleotide preferences of Cas13a/b enzymes.

B)Collateral activity of LwaCas13a, CcaCas13b, LbaCas13a, and PsmCas13b on orthogonal di-nucleotide reporters.

C)Schematic of collateral activity of Cas12a activated by dsDNA.

D)Comparison of collateral activity and pre-amplification enhanced collateral activity (SHERLOCK) of LwaCas13a, PsmCas13b, and AsCas12a. The dotted line denotes 2e9 (aM), the limit of AsCas12a sensitivity without preamplification. Values represent mean +/– S.E.M.

E)Schematic of in-sample 4 channel multiplexing using orthogonal Cas13 and Cas12a enzymes.

F)In-sample multiplexed detection of ZIKV ssRNA, ssRNA 1, DENV ssRNA, and dsDNA 1 with LwaCas13a, PsmCas13b, CcaCas13b, and AsCas12a.

G)Schematic of in-sample multiplexed detection of *S. aureus* thermonuclease and *P. aeruoginosa* acyltransferase synthetic targets with LwaCas13a and PsmCas13b.

H)In-sample multiplexed RPA and collateral detection at decreasing concentrations of *S. aureus* thermonuclease and *P. aeruoginosa* acyltransferase synthetic targets with LwaCas13a and PsmCas13b.

Using these unique cleavage preferences, we were able to detect synthetic Zika virus (ZIKV) ssRNA in the HEX channel and synthetic Dengue virus (DENV) ssRNA in the FAM channel in the same reaction (fig. 5.S16). To expand the in-sample multiplexing capabilities of SHERLOCK, we engineered a detection system based on Cas12a, which also exhibits collateral activity(Chen et al., 2017) (Fig. 5.1C). Although AsCas12a collateral activity did not produce a detectable signal at input concentrations below 100nM, preamplification with recombinase polymerase amplification (RPA) enabled single-molecule detection at 2aM (Fig. 5.1D, 5.S17) (unless otherwise noted, all SHERLOCK reactions that involve a pre-amplification are performed in two steps with the RPA reaction being directly added into the Cas13 assay without any purification step). For triplex detection, we designed a LwaCas13a uridine reporter in the Cy5 channel, a PsmCas13b adenine reporter in the FAM channel, and an AsCas12a ssDNA reporter in the HEX channel (fig. 5.S18A). We were able to detect three targets (a synthetic ssDNA target, ZIKV ssRNA, and DENV ssRNA) in a single reaction (fig. 5.S18B). We further extended detection to four targets by leveraging orthogonal di-nucleotide motifs, with reporters for LwaCas13a, PsmCas13b, CcaCas13b, and AsCas12a in FAM, TEX, Cy5, and HEX channels, respectively (Fig. 5.1E), and were able to distinguish all combinations of targets (Fig. 5.1F). When combined with RPA, we detected two DNA targets (the *P. aeruginosa* acyltransferase gene and the *S. aureus* thermonuclease gene) (Fig. 5.1G) down to the attomolar range (Fig. 5.1H). Similarly, multiplexed SHERLOCK with PsmCas13b and LwaCas13a achieved attomolar multiplexed detection of ZIKV and DENV RNA dilutions as well as allele-specific genotyping of human saliva samples (fig. 5.S19). These advances in in-sample multiplexing via orthogonal base preferences allow for many targets to be detected at scale and for cheaper cost.

We next focused on tuning the output of the SHERLOCK signal to make it more quantitative, sensitive, and robust to broaden the utility of the technology. SHERLOCK relies on an exponential pre-amplification, which saturates quickly and hinders accurate quantitation, but we observed that more dilute primer concentrations increased both raw signal and quantitative accuracy, indicating that at lower primer concentrations, the reaction does not saturate (Fig. 5.2A,B and fig. 5.S20A-E). We tested a range of primer concentrations and found that 240nM exhibited the greatest correlation between signal and input (fig. 5.S20F), and quantification was sustainable across a large range of sample concentrations down to the attomolar range (Fig. 5.2C and fig. 5.S20G). Many applications of nucleic acid detection, such as HIV detection(Barletta et al., 2004; 2009), require single molecule/mL sensitivity, and we therefore tested if the detection limit could be pushed beyond 2aM, allowing for more dilute sample inputs into SHERLOCK. By scaling up the pre-amplification RPA step, we found that LwaCas13a could give detection signal for 200, 80, and 8zM input samples and allow for single-molecule volume inputs of 250µL and 540µL (fig. 5.S21A-B), and PsmCas13b could detect 200zM input samples in 250µL reactions (fig. 5.21C).

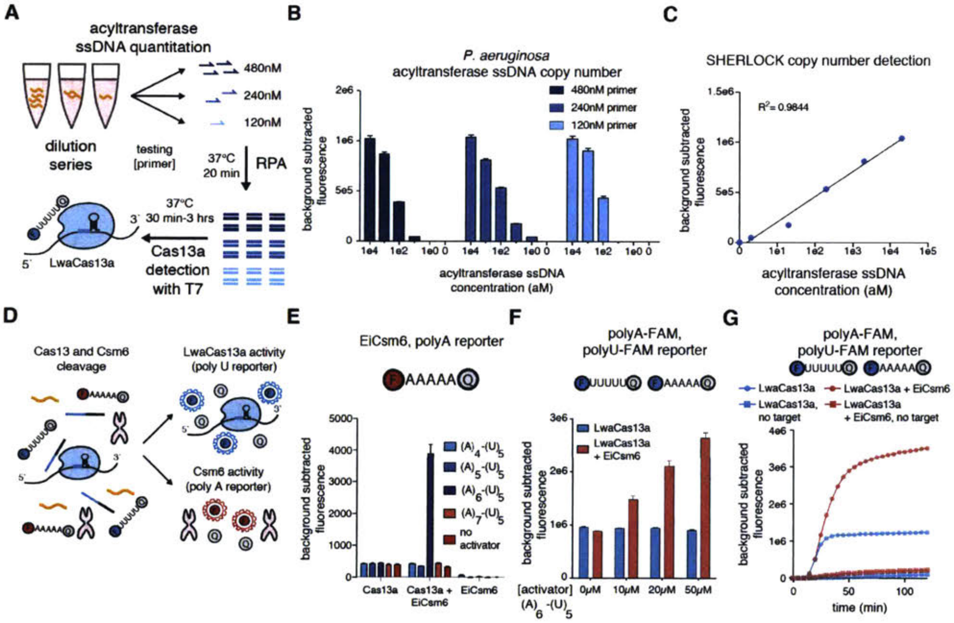


Figure 5.2: Single molecule quantitation and enhanced signal with SHERLOCK and Csm6

A)Schematic of DNA reaction scheme for quantitation of *P. aeroginosa* synthetic DNA

B)Quantitation of *P. aeroginosa* synthetic DNA at various RPA primer concentrations. Values represent mean +/– S.E.M.

C)Correlation of *P. aeroginosa* synthetic DNA concentration with detected fluorescence. Values represent mean +/– S.E.M.

D)Schematic of independent readout of LwaCas13a and Csm6 cleavage activity with orthogonal reporters.

E)Activation of EiCsm6 by LwaCas13a cleavage of adenine-uridine activators with different length adenine tracts. LwaCas13a is targeting synthetic DENV ssRNA. Values represent mean +/- S.E.M.

F)Combined LwaCas13a and EiCsm6 signal for increasing concentrations of $(A)_6$-$(U)_5$ activator detecting 20nM of DENV ssRNA. Values represent mean +/- S.E.M.

G)Kinetics of EiCsm6-enhanced LwaCas13a SHERLOCK detection of *P. aeruoginosa* acyltransferase synthetic target.

In order to amplify the detection signal, we leveraged the CRISPR type-III effector nuclease Csm6(Deng et al., 2013; Goldberg et al., 2014; Jiang et al., 2016; Niewoehner and Jinek, 2016; Samai et al., 2015; Staals et al., 2014; Tamulaitis et al., 2014), which is activated by cyclic adenylate molecules or linear adenine homopolymers terminated with a 2',3'-cyclic phosphate(Kazlauskiene et al., 2017; Niewoehner et al., 2017). LwaCas13a and PsmCas13b collateral activity generates cleavage products with hydroxylated 5' ends and 2',3'-cyclic phosphate ends (fig. 5.S22), suggesting that Cas13 collateral activity could generate Csm6 activating species, which would allow for amplified signal detection in the SHERLOCK assay. By testing RNA adenylate molecules of different lengths and 3' end modifications (Fig. 5.S23 and 5.S24A), we found that Csm6 from *Enterococcus italicus* (EiCsm6) and Csm6 from *Lactobacillus salivarius* (LsCsm6) were efficiently activated by hexadenylates containing 2',3'-cyclic phosphate ends (Fig. 5.S24B,C). Moreover, EiCsm6, LsCsm6, and Csm6 from *Thermus thermophilus* (TtCsm6) demonstrated a strong cleavage preference for A- and C-rich sensors based on sensor screening, enabling independent measurements of LwaCas13a and Csm6 cleavage activity in separate channels (Fig. 5.2D and fig. 5.S24B-D, 5.S25, 5.S26A-E).

To couple the activity of Cas13 with Csm6 activation, we designed protected RNA activators that contained a poly-A stretch followed by a protecting poly-U stretch that could be cleaved by a uracil preferring Cas13 enzyme, with the rationale that LwaCas13a could degrade all the uridines down to the homopolymeric A stretch since it had robust activity on UU and AU two-base motifs (fig. 5.S9). We found that, upon addition of target and LwaCas13a-crRNA complex, EiCsm6 and LsCsm6 were activated by the $(A)_6$-$(U)_5$ activator, consistent with the finding that the A_6 activator is optimal for

Csm6 activation and confirmed by mass spectrometry (Fig. 5.2E and fig. 5.S26F, 5.S27-5.S28). We combined the reporters for both Csm6 and Cas13 in the same reaction within the same fluorescence channel, and found that increasing the activator concentration increased the synergistic activation of Csm6 by Cas13 for DENV ssRNA detection (Fig. 5.2F), and that increasing the Csm6-specific polyA reporter also increased the Csm6 signal, leading to a larger increase in signal upon activator addition (fig. 5.S29A,B). After optimization (fig 5.S30), we found that Csm6-enhanced LwaCas13a increased the overall signal and kinetics of synthetic acyltransferase gene detection by SHERLOCK (Fig. 5.2G).

Another goal of SHERLOCKv2 was engineering a visual readout of activity requiring no additional instrumentation. We first tested a colorimetric RNase reporter based upon gold nanoparticle cluster disaggregation(Zhao et al., 2008a; Zhao et al., 2008b), but this readout required a level of RNase activity beyond what Cas13 collateral activity could achieve (fig. 5.S31). We then designed a lateral-flow readout that was based on the destruction of a FAM-biotin reporter, allowing for detection on commercial lateral flow strips. Abundant reporter accumulates anti-FAM antibody-gold nanoparticle conjugates at the first line on the strip, preventing binding of the antibody-gold conjugates to protein A on the second line; cleavage of reporter would reduce accumulation at the first line and result in signal on the second line (Fig. 5.3A). We tested this design for instrument-free detection of ZIKV or DENV ssRNA, and found that detection was possible in under 90 minutes with sensitivities down to the 2 aM condition (Fig. 5.3B,C and fig. 5.S32). Moreover, we found that we could do rapid genomic DNA extraction from human saliva (<10min) and input this directly into SHERLOCK without purification for rapid genotyping in under 23 minutes by fluorescence and 2 hours by lateral flow (fig. 5.S33). This exemplifies a closed-tube assay format with the entire SHERLOCK reaction being performed in a one-pot assay without any sample purification.

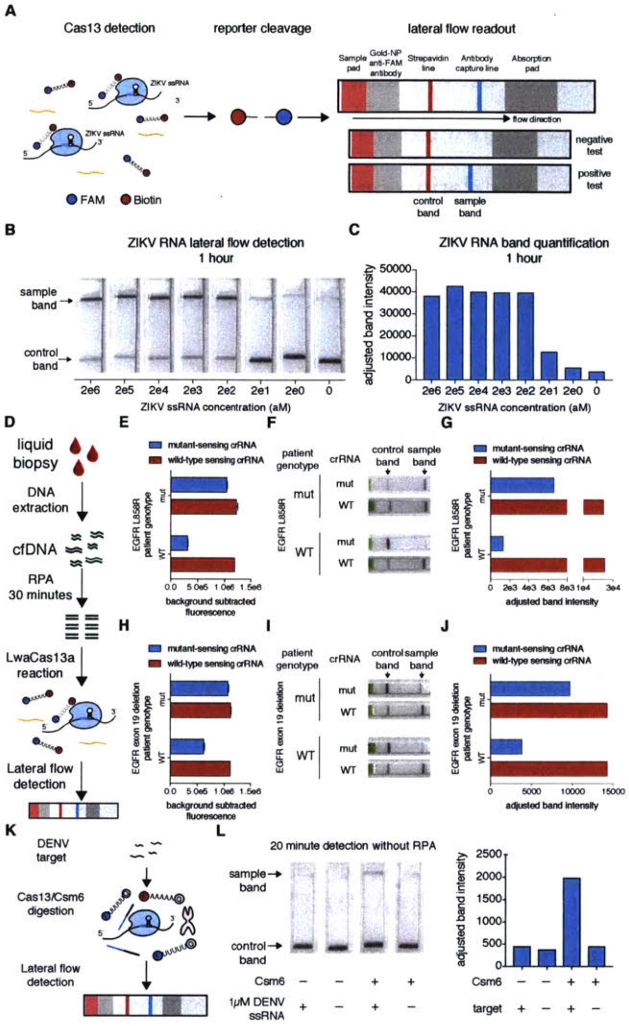


Figure 5.3: Adapting SHERLOCK for lateral flow detection

A)Schematic of lateral flow detection with SHERLOCK

B)Detection of synthetic ZIKV ssRNA using lateral flow SHERLOCK with 1 hour of LwaCas13a reaction

C)Quantitation of band intensity from detection in (B)

D)Schematic of lateral flow detection of therapeutically relevant EGFR mutations from patient liquid biopsy samples.

E)Detection of EGFR L858R mutation in patient-derived cell-free DNA samples with either L858R or WT cancer mutations. Values represent mean +/– S.E.M.

F)Lateral-flow detection of EGFR L858R mutation in patient-derived cell-free DNA samples with either L858R or WT alleles.

G)Quantitation of band intensity from detection in (E).

H)Detection of EGFR exon 19 deletion mutation in patient-derived cell-free DNA samples with either exon 19 deletion or WT alleles. Values represent mean +/- S.E.M.

I)Lateral-flow detection of EGFR exon 19 deletion mutation in patient-derived cell-free DNA samples with either exon 19 deletion or WT alleles.

J)Quantitation of band intensity from detection in (H).

K)Schematic of lateral flow readout of EiCsm6-enhanced LwaCas13a detection of DENV ssRNA

L)EiCsm6-enhanced lateral flow detection of synthetic DENV RNA in combination with LwaCas13a without preamplification by RPA. Band intensity quantitation is shown to the right.

We also applied the system to create a rapid and portable paper test for mutation detection in liquid biopsies of non-small cell lung cancer (NSCLC) patients. We designed SHERLOCK assays to detect either the EGFR L858R mutation or the exon 19 deletion (5 amino acids) and isolated cfDNA from patients with or without these mutations (Fig. 5.3D), as verified by targeted sequencing. SHERLOCK successfully detected these mutations, both with fluorescence based readout (Fig. 5.3E,H) and lateral flow-based readout (Fig. 5.3F,G,I,J fig. 5.S34A-D). Fluorescence-based SHERLOCK was also able to detect a different common EGFR mutation, T790M, in synthetic and patient cfDNA liquid biopsy samples (fig. 5.S34E,F).

To improve the robustness of the detection and reduce the likelihood of false positive readout, we combined Csm6 with Cas13 detection on lateral flow (Fig. 5.3K). We tested lateral flow reporters of various sequence and length in the presence of Csm6 and activator, and found that a long A-C reporter demonstrated strong cleavage signal (fig. 5.S35A,B). We used this reporter in combination with the Cas13 lateral flow reporter for rapid detection of DENV ssRNA relying solely on Csm6 for amplification (i.e., in the absence of RPA) (Fig. 5.3L). We subsequently combined RPA, Cas13/Csm6, and lateral flow readout to detect an acyltransferase target, and found that the increase in signal conferred by Csm6 allowed for more rapid detection by lateral flow (fig. 5.S35C-D) with reduced background.

Finally, we applied SHERLOCKv2 in a simulated approach that involves Cas13 serving as both a companion diagnostic and the therapy itself, as Cas13 has been developed for a variety of applications in mammalian cells including RNA knockdown, imaging, and editing (Abudayyeh et al., 2017; Cox et al., 2017)(Fig. 5.4A). We recently harnessed Cas13b from *Prevotella sp. P5-125* (PspCas13b) to correct mutations in genetic diseases using a system called RNA Editing for Programmable A-to-I Replacement (REPAIR)(Cox et al., 2017). To direct and monitor the outcome of a treatment, we tested if SHERLOCK could be used both for genotyping to guide the REPAIR treatment and as a readout of the edited RNA to track the efficiency of the therapy. We used a mutation in *APC* (APC:c.1262G>A) implicated in Familial adenomatous polyposis 1 (Fig. 5.4B,C) (Cottrell et al., 1992), and transfected synthetic healthy and mutant cDNAs of the fragment surrounding the mutation into HEK293FT cells. We harvested DNA from these cells and successfully genotyped the correct samples using single-sample multiplexed SHERLOCK with LwaCas13a and PsmCas13b (Fig. 5.4D). Concurrently, we designed and cloned guide RNAs for the REPAIR system and transfected cells that had the diseased genotype with the guide RNA and dPspCas13b-ADAR2$_{dd}$(E488Q) REPAIR system. We confirmed editing by next-generation sequencing (NGS) analysis, finding that 43% editing was achieved with the REPAIR system (Fig. 5.4E), and we were able to detect this editing with SHERLOCK (Fig. 5.4F and fig. 5.S36).

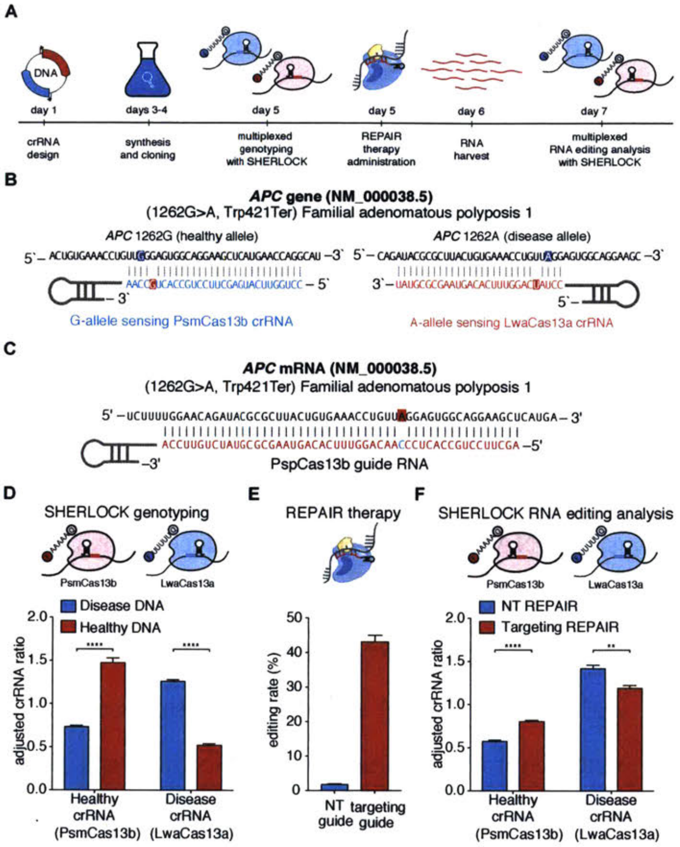

A
DNA
day 1
days 3-4
day 5
day 5
day 6
day 7
crRNA design
synthesis and cloning
multiplexed genotyping with SHERLOCK
REPAIR therapy administration
RNA harvest
multiplexed RNA editing analysis with SHERLOCK
B
APC gene (NM_000038.5)
(1262G>A, Trp421Ter) Familial adenomatous polyposis 1
APC 1262G (healthy allele)
APC 1262A (disease allele)
5`– ACUGUGAAACCUGUUGGGAGUGGCAGGAAGCUCAUGAACCAGGCAU –3`
AACCGUCACCGUCCUUCGAGUACUUGGUCC – 5`
– 3`
5`– CAGAUACGCGCUUACUGUGAAACCUGUUAGGAGUGGCAGGAAGC –3`
3`– UAUGCGCGAAUGACACUUUGGACUAUCC
5`–
G-allele sensing PsmCas13b crRNA
A-allele sensing LwaCas13a crRNA
C
APC mRNA (NM_000038.5)
(1262G>A, Trp421Ter) Familial adenomatous polyposis 1
5' – UCUUUUGGAACAGAUACGCGCUUACUGUGAAACCUGUUAGGAGUGGCAGGAAGCUCAUGA – 3'
ACCUUGUCUAUGCGCGAAUGACACUUUGGACAACCCUCACCGUCCUUCGA –5'
–3'
PspCas13b guide RNA
D
SHERLOCK genotyping
PsmCas13b
LwaCas13a
Disease DNA
Healthy DNA
adjusted crRNA ratio
2.0
1.5
1.0
0.5
0.0
****
****
Healthy crRNA (PsmCas13b)
Disease crRNA (LwaCas13a)
E
REPAIR therapy
editing rate (%)
50
40
30
20
10
0
NT guide
targeting guide
F
SHERLOCK RNA editing analysis
PsmCas13b
LwaCas13a
NT REPAIR
Targeting REPAIR
adjusted crRNA ratio
2.0
1.5
1.0
0.5
0.0
****
**
Healthy crRNA (PsmCas13b)
Disease crRNA (LwaCas13a)

Figure 5.4: Combined therapeutics and diagnostics with Cas13 enzymes

A)Schematic of timeline for detection of disease alleles, correction with REPAIR, and assessment of REPAIR correction.

B)Sequences of targets and crRNA designs used for detection of *APC* alleles.

C)Sequences of target and REPAIR guide design used for correction of *APC* alleles.

D)In-sample multiplexed detection of *APC* alleles from healthy- and disease-simulating samples with LwaCas13a and PsmCas13b. Adjusted crRNA ratio allows for comparisons between different crRNAs that will have different overall signal levels (see supplementary methods for more details). Values represent mean +/– S.E.M.

E)Quantitation of REPAIR editing efficiency at the targeted *APC* mutation. Values represent mean +/– S.E.M.

F)In-sample multiplexed detection of *APC* alleles from REPAIR targeting and non-targeting samples with LwaCas13a and PsmCas13b. Values represent mean +/– S.E.M.

5.4 Conclusion

The additional refinements presented here for Cas13-based detection allow for quantitative, visual, more sensitive, and multiplexed readouts, enabling additional applications for nucleic acid detection, especially in settings where portable and instrument-free analysis are necessary. SHERLOCKv2 can be used for multiplexed genotyping to inform pharmacogenomic therapeutic development and application, detecting genetically modified organisms in the field, or determining the presence of co-occurring pathogens. Moreover, the rapid, isothermal readout of SHERLOCKv2, enabled by lateral flow and Csm6, provides an opportunity for detection in settings where power or portable readers are unavailable, even for rare species like circulating DNA. In the future, it might be possible to make solution-based colorimetric readouts and multiplex lateral flow assays containing multiple test strips for different targets. Improved CRISPR-dx nucleic acid tests make it easier to detect the presence of nucleic acids in a range of applications across biotechnology and health and are now field-ready for rapid and portable deployment.

5.5 Experimental Procedures

5.5.1 Protein expression and purification of Cas13 and Csm6 orthologs

LwaCas13a expression and purification was carried out as described before(Gootenberg et al., 2017c) with minor modifications and is detailed below. LbuCas13a, LbaCas13a, Cas13b and Csm6 orthologs were expressed and purified with a modified protocol. In brief, bacterial expression vectors were transformed into Rosetta™ 2(DE3)pLysS Singles Competent Cells (Millipore). A 12.5 mL starter culture was grown overnight in Terrific Broth 4 growth media (Sigma) (TB), which was used to inoculate 4 L of TB for growth at 37°C and 300 RPM until an OD600 of 0.5. At this time, protein expression was induced by supplementation with IPTG (Sigma) to a final concentration of 500 µM, and cells were cooled to 18°C for 16 h for protein expression. Cells were then centrifuged at 5000 g for 15 min at 4°C. Cell pellet was harvested and stored at -80°C for later purification.

All subsequent steps of the protein purification were performed at 4°C. Cell pellet was crushed and resuspended in lysis buffer (20 mM Tris-HCl, 500 mM NaCl, 1 mM DTT, pH 8.0) supplemented with protease inhibitors (Complete Ultra EDTA-free tablets), lysozyme (500μg/1ml), and benzonase followed by high-pressure cell disruption using the LM20 Microfluidizer system at 27,000 PSI. Lysate was cleared by centrifugation for 1 hr at 4°C at 10,000 g. The supernatant was applied to 5mL of StrepTactin Sepharose (GE) and incubated with rotation for 1 hr followed by washing of the protein-bound StrepTactin resin three times in lysis buffer. The resin was resuspended in SUMO digest buffer (30 mM Tris-HCl, 500 mM NaCl 1 mM DTT, 0.15% Igepal (NP-40), pH 8.0) along with 250 Units of SUMO protease (250mg/ml) and incubated overnight at 4°C with rotation. The suspension was applied to a column for elution and separation from resin by gravity flow. The resin was washed two times with 1 column volume of Lysis buffer to maximize protein elution. The elute was diluted in cation exchange buffer (20 mM HEPES, 1 mM DTT, 5% glycerol, pH 7.0; pH 7.5 for LbuCas13a, LbaCas13a, EiCsm6, LsCsm6, TtCsm6) to lower the salt concentration in preparation for cation exchange chromatography to 250mM.

For cation exchange and gel filtration purification, protein was loaded onto a 5 mL HiTrap SP HP cation exchange column (GE Healthcare Life Sciences) via FPLC (AKTA PURE, GE Healthcare Life Sciences) and eluted over a salt gradient from 250 mM to 2M NaCl in elution buffer (20 mM HEPES, 1 mM DTT, 5% glycerol, pH 7.0; pH 7.5 for LbuCas13a, LbaCas13a). The resulting fractions were tested for presence of recombinant protein by SDS-PAGE, and fractions containing the protein were pooled and concentrated via a Centrifugal Filter Unit (Millipore 50MWCO) to 1 mL in S200 buffer (10 mM HEPES, 1 M NaCl, 5 mM MgCl2, 2 mM DTT, pH 7.0). The concentrated protein was loaded onto a gel filtration column (Superdex® 200 Increase 10/300 GL, GE Healthcare Life Sciences) via FPLC. The resulting fractions from gel filtration were analyzed by SDS-PAGE and fractions containing protein were pooled and buffer exchanged into Storage Buffer (600 mM NaCl, 50 mM Tris-HCl pH 7.5, 5% glycerol, 2mM DTT) and frozen at -80°C for storage.

5.5.2 Nucleic acid target and crRNA preparation

Nucleic acid targets for Cas12a and genomic DNA detection were PCR amplified with NEBNext PCR master mix, gel extracted, and purified using MinElute gel extraction kit (Qiagen). For RNA based detection, purified dsDNA was incubated with T7 polymerase overnight at 30°C using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs) and RNA was purified with the MEGAclear Transcription Clean-up kit (Thermo Fisher)

crRNA preparation was carried out as described before(Gootenberg et al., 2017c) with minor modifications and is detailed below. For preparation of crRNAs, constructs were ordered as ultramer DNA (Integrated DNA Technologies) with an appended T7 promoter sequence. crRNA DNA was annealed to a short T7 primer (final concentrations 10 uM) and incubated with T7 polymerase overnight at 37°C using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). crRNAs were purified using RNAXP clean beads (Beckman Coulter) at 2x ratio of beads to reaction volume, with an additional 1.8x supplementation of isopropanol (Sigma).

5.5.3 Recombinase Polymerase Amplification (RPA)

Primers for RPA were designed using NCBI Primer-BLAST(Ye et al., 2012) using default parameters, with the exception of amplicon size (between 100 and 140 nt), primer melting temperatures (between 54°C and 67°C), and primer size (between 30 and 35 nt). Primers were then ordered as DNA (Integrated DNA Technologies).

RPA and RT-RPA reactions run were as instructed with TwistAmp® Basic or TwistAmp® Basic RT (TwistDx), respectively, with the exception that 280 mM MgAc was added prior to the input template. Reactions were run with 1 µL of input for 1 hr at 37°C, unless otherwise described.

For SHERLOCK quantification of nucleic acid, RPA primer concentration tested at standard concentration (480nM final) and lower (240nM, 120nM,60nM, 24nM) to find the optimum concentration. RPA reactions were further run for 20 minutes.

When multiple targets were amplified with RPA, primer concentration was adjusted to a final concentration of 480nM. That is, 120nM of each primer for two primer pairs were added for duplex detection.

5.5.4 Fluorescent cleavage assay

Detection assays were carried out as described before(Gootenberg et al., 2017c) with minor modifications and the procedure is detailed below. Detection assays were performed with 45 nM purified Cas13, 22.5 nM crRNA, quenched fluorescent RNA reporter (125nM RNAse Alert v2, Thermo Scientific, homopolymer and di-nucleotide reporters (IDT); 250nM for polyA Trilink reporter), 0.5 µL murine RNase inhibitor (New England Biolabs), 25 ng of background total human RNA (purified from HEK293FT culture), and varying amounts of input nucleic acid target, unless otherwise indicated, in nuclease assay buffer (20 mM HEPES, 60 mM NaCl, 6 mM $MgCl_2$, pH 6.8). For Csm6 fluorescent cleavage reactions, protein was used at 10nM final concentration along with 500nM of 2', 3' cyclic phosphate oligoadenylate, 250nM of fluorescent reporter, and 0.5 µL murine RNase inhibitor in nuclease assay buffer (20 mM HEPES, 60 mM NaCl, 6 mM $MgCl_2$, pH 6.8). Reactions were allowed to proceed for 1-3 hr at 37°C (unless otherwise indicated) on a fluorescent plate reader (BioTek) with fluorescent kinetics measured every 5 min. In reactions involving AsCas12a, 45nM AsCas12a was included using recombinant protein from IDT. In the case of multiplexed reactions, 45nM of each protein and 22.5nM of each crRNA was used in the reaction.

5.5.5 SHERLOCK nucleic acid detection

Detection assays were performed with 45 nM purified Cas13, 22.5 nM crRNA, quenched fluorescent RNA reporter (125nM RNAse Alert v2, Thermo Scientific, homopolymer and di-nucleotide reporters (IDT), 250nM for polyA Trilink reporter), 0.5 µL murine RNase inhibitor (New England Biolabs), 25 ng of background total human RNA (purified from HEK293FT culture), and 1uL of RPA reaction in nuclease assay buffer (20 mM HEPES, 60 mM NaCl, 6 mM $MgCl_2$, pH 6.8), rNTP mix (1mM final, NEB), 0.6 µL T7 polymerase (Lucigen) and 3mM $MgCl_2$. Reactions were allowed to proceed for 1-3 hr at 37°C (unless otherwise indicated) on a fluorescent plate reader (BioTek) with fluorescent kinetics measured every 5 min.

For one-pot nucleic acid detection, the detection assay was carried out as described before (Gootenberg et al., 2017c) with minor modifications. A single 100 μL combined reaction assay consisted of 0.48 μM forward primer, 0.48 μM reverse primer, 1x RPA rehydration buffer, varying amounts of DNA input, 45 nM LwCas13a recombinant protein, 22.5 nM crRNA, 125 ng background total human RNA, 125 nM substrate reporter (RNase alert v2), 2.5 μL murine RNase inhibitor (New England Biolabs), 2 mM ATP, 2 mM GTP, 2 mM UTP, 2 mM CTP, 1 μL T7 polymerase mix (Lucigen), 5 mM MgCl2, and 14 mM MgAc. Reactions were allowed to proceed for 1-3 hr at 37°C (unless otherwise indicated) on a fluorescent plate reader (BioTek) with fluorescent kinetics measured every 5 min. For lateral flow readout, 20 uL of the combined reaction was added to 100uL of HybriDetect 1 assay buffer (Milenia) and run on HybriDetect 1 lateral flow strips (Milenia).

5.5.6 Nucleic acid labeling for cleavage fragment analysis

Target RNA was *in vitro* transcribed from a dsDNA template and purified as described above. The *in vitro* cleavage reaction was performed as described above for fluorescence cleavage reaction with the following modifications. Fluorescence reporter was substituted for 1μg RNA target and no background RNA was used. Cleavage reaction was carried out for 5 minutes (LwaCas13a) or 1 hour (PsmCas13b) at 37°C. The cleavage reaction was purified using the RNA clean & concentrator-5 kit (Zymo Research) and eluted in 10 uL UltraPure water (Gibco). Cleavage reaction was further labeled with a 10μg of maleimide IRDye 800CW (Licor) following the 5'EndTag labeling Reaction (Vector Laboratories) kit protocol. To determine the 5' end produced by Cas13 cleavage, the protocol was modified to either perform an Alkaline Phosphatase (AP) treatment or substitute with UltraPure water to only label 5'-OH containing RNA species, while undigested triphosphorylated (PPP) RNA species are only labeled when AP treatment is performed.

5.5.7 Mass Spectrometry for high resolution cleavage fragment analysis

For determining the cleavage ends produced by Cas13 collateral RNase activity by Mass Spectrometry, an *in vitro* cleavage reaction was performed as described above with the following modifications. Cas13 RNA target was used at 1 nM final concentration, Csm6 activator at 3μM final concentration and no background RNA was used. For control reactions, either Cas13 target was substituted by UltraPure water, or standard *in vitro* cleavage reaction was incubated with hexaadenylate containing a 2',3'cyclic phosphate activator in the absence of Cas13 target, Cas13 protein and Cas13 crRNA. The cleavage reactions were carried out for 1h at 37°C and purified using an New England Biolabs siRNA purification protocol. In brief, one-tenth volume of 3 M NaOAc, 2 μL of RNase-free Glycoblue (Thermofisher) and three volumes of cold 95% ethanol was added, placed at -20°C for 2 hours, and centrifuged for 15 minutes at 14,000g. The supernatant was removed and two volumes of 80% EtOH was added and incubated for 10 minutes at room temperature. The supernatant was decanted and samples centrifuged for 5 minutes at 14,000g. After air-drying the pellet, 50 μL of UltraGrade water added and sent on dry ice for Mass spectrometry analysis.

For mass spectrometry analysis, samples were diluted 1:1 with UltraGrade water and analyzed on Bruker Impact II q-TOF mass spectrometer in negative ion mode coupled to an Agilent 1290 HPLC. 10 μL were injected onto a PLRP-S column (50 mm, 5 um particle size, 1000 angstrom pore size PLRP-S column, 2.1 mm ID) using 0.1% ammonium hydroxide v/v in water as mobile phase A and acetonitrile as mobile phase B. The flow rate was kept constant throughout at 0.3 ml/minute. The mobile phase composition started at 0%B and was maintained for the first 2 minutes. After this point, the composition was changed to 100% B over the next 8 minutes and maintained for one minute. The composition was then returned to 0% B over 0.1 minute and then maintained for the following 4.9 minutes to allow the column to re-equilibrate to starting conditions. The mass spectrometer was tuned for large MW ions, and data was acquired between m/z 400-5000. The entire dataset from the mass spectrometer was calibrated by m/z using an injection of sodium formate. Data was analyzed using Bruker Compass Data Analysis 4.3 with a license for MaxEnt deconvolution algorithm to generate a calculated neutral mass spectrum from the negatively charged ion data.

5.5.8 Genomic DNA extraction from human saliva

Saliva DNA extraction was carried out as described before(Gootenberg et al., 2017c) with minor modifications and is detailed below. 2 mL of saliva was collected from volunteers, who were restricted from consuming food or drink 30 min prior to collection. Samples were then processed using QIAamp® DNA Blood Mini Kit (Qiagen) as recommended by the kit protocol. For boiled saliva samples, 400 µL of phosphate buffered saline (Sigma) was added to 100 µL of volunteer saliva and centrifuged for 5 min at 1800 g. The supernatant was decanted and the pellet was resuspended in phosphate buffered saline with 0.2% Triton X-100 (Sigma) before incubation at 95°C for 5 min. 1 µL of sample was used as direct input into RPA reactions.

5.5.9 Digital droplet PCR quantification

ddPCR quantification was carried out as described before(Gootenberg et al., 2017c) with minor modifications and is detailed below. To confirm the concentration of target dilutions, we performed digital-droplet PCR (ddPCR). For DNA quantification, droplets were made using the ddPCR Supermix for Probes (no dUTP) (BioRad) with PrimeTime qPCR probes/primer assays (IDT) designed for the target sequence. For RNA quantification, droplets were made using the one-step RT-ddPCR kit for probes with PrimeTime qPCR probes/primer assays designed for the target sequence. Droplets were generated in either case using the QX200 droplet generator (BioRad) and transferred to a PCR plate. Droplet-based amplification was performed on a thermocycler as described in the kit protocol and nucleic acid concentrations were subsequently determined via measurement on a QX200 droplet reader.

5.5.10 Cas13-Csm6 fluorescent cleavage assay

Cas13-Csm6 combined fluorescent cleavage assays were performed as described for standard Cas13 fluorescent cleavage reactions with the following modifications. Csm6 protein was added to 10 nM final concentration, 400 nM of Csm6 fluorescent reporter and 500 nM Csm6 activator unless otherwise indicated. For distinguishing Cas13 from Csm6 collateral RNase activity, two distinct fluorophores were used for fluorescence detection (FAM and HEX). Because of the interference of rNTPs with Csm6 activity, the IVT was performed in the RPA pre-amplification step and then 1µL of this reaction was added as input to the Cas13-Csm6 cleavage assay.

In the case where we tested a three-step Cas13-Csm6 cleavage assay, the RPA was performed normally as discussed above for varying times and then used as input to a normal IVT reaction for varying times. Then 1µL of the IVT was used as input to the Cas13-Csm6 reaction described in the previous paragraph.

5.5.11 Motif discovery screen with library

To screen for Cas13 cleavage preference, an *in vitro* RNA cleavage reaction was set up as described above with the following modifications. Cas13 target was used at 20nM, fluorescent reporter was substituted for 1 µM of DNA-RNA oligonucleotide (IDT) that contains a 6-mer stretch of randomized ribonucleotides flanked by DNA handles for NGS library preparation. Reactions were carried out for 60 minutes (unless otherwise indicated) at 37°C. The reactions were purified using the Zymo oligo-clean and concentrator-5 kit (Zymo research) and 15µL of UltraPure water was used for elution. 10µL of purified reaction was used for reverse transcription using a gene-specific primer that binds to the DNA handle.

Reverse transcription (RT) was carried out for 45 minutes at 42°C according to the qScript Flex cDNA-kit (quantabio) protocol. To assess cleavage efficiency and product purity, RT-reactions were diluted 1:10 in water and loaded on a Small RNA kit and run on a Bioanalyzer 2100 (Agilent). Four microliters of RT-reaction was used for the first-round of NGS library preparation. NEBNext (NEB) was used to amplify first strand cDNA with a mix of forward primers at 625 nM final and a reverse primer at 625 nM for 15 cycles with 3 minute initial denaturation at 98°C, 10s cycle denaturation at 98°C, 10s annealing at 63°C, 20s 72°C extension and 2 minute final extension extension at 72°C.

Two microliters of first round PCR reaction was used for second round PCR amplification to attach Illumina-compatible indices (NEB) for NGS sequencing. The same NEBNext PCR protocol was used for amplification. PCR product were analysed by agarose gel-electrophoresis (2% Sybr Gold E-Gel

Invitrogen system) and 5μL of each reaction was pooled. The pooled samples was gel extracted, quantified with Qubit DNA 2.0 DNA High sensitivity kit and normalized to 4 nM final concentration. The final library was diluted to 2 pM and sequenced on a NextSeq 500 Illumina system using a 75-cycle kit.

5.5.12 Motif Screen Analysis

To analyze depletion of preferred motifs from the random motif library screen, 6-mer regions were extracted from sequence data and normalized to overall read count for each sample. Normalized read counts were then used to generated log ratios, with psuedocount adjustment, between experimental conditions and matched controls. For Cas13 experiments, matched controls did not have target RNA added; for Csm6 and RNase A experiments, matched controls did not have enzyme. Log ratio distribution shape was used to determine cut-offs for enriched motifs. Enriched motifs were then used to determine occurrence of 1-, 2-, or 3- nucleotide combinations. Motif logos were generated using Weblogo3(Crooks et al., 2004).

5.5.13 Phylogenetic analysis of Cas13 protein and crRNA direct repeats

To study ortholog clustering, multiple sequence alignments were generated with Cas13a and Cas13b protein sequences in Geneious with MUSCLE and then clustered using Euclidean distance in R with the heatmap.2 function. To study direct repeat clustering, multiple sequence alignments were generated with Cas13a and Cas13b direct repeat sequences in Geneious using the Geneious algorithm and then clustered using Euclidean distance in R with the heatmap.2 function. To study clustering of orthologs based on di-nucleotide motif preference, the cleavage activity matrix was clustered using Euclidean distance in R using the heatmap.2 function.

5.5.14 Gold nanoparticle colorimetric

An RNA oligo was synthesized from IDT with thiols at the 5' and 3' ends. In order to deprotect the thiol groups, the oligo at a final concentration of 20mM was reduced in 150mM sodium phosphate buffer containing 100mM DTT for 2 hours at room temperature. The oligo were then purified using sephadex NAP-5 columns (GE Healthcare) into a final volume of 700µL water. As previously described(Zhao et al., 2008a), the reduced oligo at 10µM was added at a volume of 280µL to 600µL of 2.32nM 15nm-gold nanoparticles (Ted Pella), which is a 2000:1 ratio of oligo to nanoparticles. Subsequently, 10µL of 1M Tris-HCl at pH8.3 and 90µL of 1M NaCl were added to the oligo-nanoparticle mixture and incubated for 18 hours at room temperature with rotation. After 18 hours, additional 1M Tris-HCl (5µL at pH8.3) was added with 5M NaCl (50µL) and this was incubated for an additional 15 hours at room temperature with rotation. Following incubation, the final solution was centrifuged for 25 min at 22,000g. The supernatant was discarded and the conjugated nanoparticles were resuspended in 50µL of 200mM NaCl.

The nanoparticles were tested for RNase sensitivity using an RNase A assay. Varying amounts of RNase A (Thermo Fischer) were added to 1x RNase A buffer and 6µL of conjugated nanoparticles in a total reaction volume of 20µL. Absorbance at 520nm was monitored every 5 minutes for 3 hours using a plate spectrophotometer.

5.5.15 Lateral flow readout of Cas13 activity using FAM-biotin reporters

For lateral flow based on cleavage of a FAM-RNA-biotin reporter, non-RPA LwaCas13a reactions or SHERLOCK-LwaCas13a reactions were run for 1 hour, unless otherwise indicated, with 1uM final concentration of FAM-RNA-biotin reporter. After incubation, 20uL LwaCas13a reactions supernatant was added to 100uL of HybriDetect 1 assay buffer (Milenia) and run on HybriDetect 1 lateral flow strips (Milenia).

5.5.16 Cloning of REPAIR constructs, Mammalian cell transfection, RNA isolation and NGS library preparation for REPAIR

Constructs for simulating reversion of *APC* mutations and guide constructs for REPAIR were cloned as previously described(Cox et al., 2017). Briefly, 96 nt sequences centered on the APC:c.1262G>A mutation were designed and golden gate cloned under an expression vector, and corresponding guide sequences were golden gate cloned into U6 expression vectors for PspCas13b guides. To simulate patient samples, 300ng of either mutant or wildtype *APC* expression vector was transfected into HEK293FT cells with Lipofectamine 2000 (Invitrogen), and two days post-transfection DNA was harvested with Qiamp DNA Blood Midi Kit (Qiagen) following manufacturer's instructions. 20ng of DNA were used as input into SHERLOCK-LwaCas13a reactions.

RNA correction using the REPAIR system was performed as previously described(Cox et al., 2017): 150ng of dPspCas13b-ADAR(DD)E488Q, 200 ng of guide vector, and 30ng of *APC* expression vector were co-transfected, and two-days post transfection RNA was harvested using the RNeasy Plus Mini Kit (Qiagen) following manufacturer's instructions. 30ng of RNA was used as input into SHERLOCK-LwaCas13a reactions.

RNA editing fractions were independently determined by NGS as previously described. RNA was reverse transcribed with the qScript Flex kit (Quanta Biosciences) with a sequence specific primer. First strand cDNA was amplified with NEBNext High Fidelity 2X PCR Mastermix (New England Biosciences) with a mix of forward primers at 625nM final and a reverse primer at 625nM for 15 cycles with 3 minute initial denaturation at 98°C, 10 second cycle denaturation at 98°C, 30 second annealing at 65°C, 30 second 72°C extension and 2 minute final extension extension at 72°C. Two microliters of first round PCR reaction was used for second round PCR amplification to attach Illumina-compatible indices for NGS sequencing, with NEBNext, using the same protocol with 18 cycles. PCR products were analysed by agarose gel-electrophoresis (2% Sybr Gold E-Gel Invitrogen) and 5µL of each reaction was pooled. The pooled samples was gel extracted, quantified with Qubit DNA 2.0 DNA High sensitivity kit and normalized to 4nM final concentration, and read out with a 300 cycle v2 MiSeq kit (Illumina).

5.5.17 Analysis of SHERLOCK fluorescence data

SHERLOCK fluorescence analysis was carried out as described before(Gootenberg et al., 2017c) with minor modifications and is detailed below. To calculate background subtracted fluorescence data, the initial fluorescence of samples was subtracted to allow for comparisons between different conditions. Fluorescence for background conditions (either no input or no crRNA conditions) were subtracted from samples to generate background subtracted fluorescence.

crRNA ratios for SNP discrimination were calculated to adjust for sample-to-sample overall variation as follows:

$$crRNA\ A_i\ ratio = \frac{(m+n)A_i}{\sum_{i=1}^{m} A_i + \sum_{i=1}^{n} B_i}$$

where A_i and B_i refer to the SHERLOCK intensity values for technical replicate i of the crRNAs sensing allele A or allele B, respectively, for a given individual. Since we typically have four technical replicates per crRNA, m and n are equal to 4 and the denominator is equivalent to the sum of all eight of the crRNA SHERLOCK intensity values for a given SNP locus and individual. Because there are two crRNAs, the crRNA ratio average across each of the crRNAs for an individual will always sum to two. Therefore, in the ideal case of homozygosity, the mean crRNA ratio for the positive allele crRNA will be two and the mean crRNA ratio for the negative allele crRNA will be zero. In the ideal case of heterozygosity, the mean crRNA ratio for each of the two crRNAs will be one. Because in SHERLOCKv2, we accomplish genotyping by measuring A_i and B_i in different color channels, we scaled the 530-color channel by 6 to match the intensity values in the 480-color channel.

5.6 Acknowledgements

This work was published in (Gootenberg et al., 2017a). We would like to thank Sonia Trauger and the Harvard Small Molecule Mass Spectrometry facility for mass spectrometry assistance, Jonathan Strecker and Ian Slaymaker for protein purification assistance, David Cox for assistance with Cas13b gene synthesis, Brian Franklin, Vanessa Verdine, and Anna Le for additional experimental assistance, Lynette Sholl and Jeff Golden for providing cell-free DNA samples, Liang Hao and Sangeeta Bhatia for assistance with gold nanoparticle experiments, Catherine Freije, Cameron Myrhvold, and Pardis

Sabeti for assistance with manuscript preparation, and Rhiannon Macrae, Rich Belliveau, Erin Blackwell, and the entire Zhang lab for discussions and support.

Chapter 6

RNA targeting with CRISPR–Cas13

This chapter is adapted from the following article:

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., et al. (2017). RNA targeting with CRISPR-Cas13. Nature 550, 280-284.

Contributions: Omar Abudayyeh and Jonathan Gootenberg are co-first authors (*). Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang conceived and designed the study. Omar Abudayyeh and Jonathan Gootenberg participated in the design and execution of all experiments. Patrick Essletzbichler performed biochemical characterization studies on the LwaCas13a protein. Julia Joung, and Vanessa Verdine prepared the vectors for arrayed knockdown screening experiments. Julia Joung also performed RNA immunoprecipitation experiments. Max Kellner performed the RNA integrity analysis. Omar Abudayyeh and Jonathan Gootenberg analyzed all data. Shuo Han performed select microscopy experiments. Joseph Belanto performed the plant protoplasts knockdown experiments. Omar Abudayyeh, Jonathan Gootenberg, Eric Lander, and Feng Zhang wrote the paper with input from Dan Voytas, Alice Ting, and Aviv Regev and help from all authors.

6.1 Abstract and introduction

RNA plays important and diverse roles in biology, but molecular tools to manipulate and measure RNA are limited. For example, RNA interference (RNAi)(Elbashir et al., 2001; Fire et al., 1998; Root et al., 2006) can efficiently knockdown RNAs, but it is prone to off-target effects(Jackson et al., 2003), and visualizing RNAs typically relies on the introduction of exogenous tags(Tyagi, 2009). Here, we demonstrate that the class 2 type VI(Shmakov et al., 2015; Shmakov et al., 2017a) RNA-guided RNA-targeting CRISPR-Cas effector Cas13a(Abudayyeh et al., 2016) (previously known as C2c2) can be engineered for mammalian cell RNA knockdown and binding. After initial screening of fifteen orthologs, we identified Cas13a from *Leptotrichia wadei (*LwaCas13a) as the most effective in an interference assay in *E. coli.* LwaCas13a can be heterologously expressed in mammalian and plant cells for targeted knockdown of either reporter or endogenous transcripts with comparable levels of knockdown as RNAi and improved specificity. Catalytically inactive LwaCas13a maintains targeted RNA binding activity, which we leveraged for programmable tracking of transcripts in live cells. Our results establish CRISPR-Cas13a as a flexible platform for studying RNA in mammalian cells.

6.2 Results

To achieve robust Cas13a-mediated RNA knockdown, we first evaluated fifteen Cas13a orthologs for protospacer flanking site (PFS) preference and activity using a previously described ampicillin resistance assay(Abudayyeh et al., 2016) (Fig. 6.1a and Extended Data Fig. 6.1a). This assay monitors Cas13a-mediated cleavage of the ß-*lactamase* (ampicillin resistance) transcript, resulting in bacterial death under ampicillin selection, which can be measured by quantifying surviving colonies. Using this approach, we found that the Cas13a ortholog from *L. wadei* (LwaCas13a) was most active, followed by the previously characterized LshCas13a (from *Leptotrichia shahii)* (Fig. 6.1b and Extended Data Fig. 6.1b)(Abudayyeh et al., 2016). Sequencing analysis of the PFS distributions from the LwaCas13a and LshCas13a screens revealed that most LwaCas13a PFS sequences were depleted (Extended Data Fig. 6.1c-e). Motif analysis of the depleted PFS sequences at varying thresholds revealed the expected 3' H motif for LshCas13a, but no significant PFS motif for LwaCas13a (Fig. 6.1c and Extended Data Fig. 6.1f,g). Consistent with these results, LwaCas13a was also found to be more active than LshCas13a as a nucleic acid sensor(Gootenberg et al., 2017b). Because of its high activity and lack of PFS in bacteria, we focused on LwaCas13a for further development.

In vitro cleavage reactions with LwaCas13a demonstrated programmable RNA cleavage with a crRNA encoding a 28-nt spacer (shorter than the 29-30 nt length found in the native *L. wadei* CRISPR array (Extended Data Fig. 6.2a)). These reactions confirmed the higher cleavage efficiency of LwaCas13a over LshCas13a (Extended Data Fig. 6.2b,c), and revealed similar biochemical characteristics for the two enzymes (Extended Data Fig. 6.2d-g). We found that LwaCas13a could cleave the corresponding pre-crRNA transcript from *L. wadei* (Extended Data Fig. 6.2h). We also explored the crRNA constraints on LwaCas13a cleavage by truncating the spacer, finding that LwaCas13a retained *in vitro* cleavage activity with spacer lengths as short as 20 nt (Extended Data Fig. 6.2i). Although guide lengths less than 20 nt no longer support catalytic activity, the LwaCas13-crRNA complex may still retain binding activity, providing an opportunity for orthogonal applications with a single enzyme(Dahlman et al., 2015).

We next evaluated the ability of LwaCas13a to cleave transcripts in mammalian cells. We cloned mammalian codon-optimized LwaCas13a into mammalian expression vectors with msfGFP fusions on the C- or N-terminus and either a dual-flanking nuclear export sequence (NES) or nuclear localization sequence (NLS) and evaluated expression and localization (Fig. 6.1d). We found that msfGFP-fused LwaCas13a constructs expressed well and localized effectively to the cytoplasm or nucleus according to the localization sequence. To evaluate the *in vivo* cleavage activity of LwaCas13a we developed a dual luciferase reporter system that expresses both Gaussia luciferase (Gluc) and Cypridinia luciferase (Cluc) under different promoters on the same vector, allowing one transcript to serve as the LwaCas13a target and the other to serve as a dosing control (Fig. 6.1e). We then designed guides against Gluc and cloned them into a tRNA^{Val} promoter-driven guide expression vector. We transfected the LwaCas13a expression vector, guide vector, and dual-luciferase construct into HEK293FT cells and measured luciferase activity at 48 hrs post-transfection. We found that LwaCas13a-msfGFP-NLS resulted in the highest levels of knockdown (75.7% for guide 1, 72.9% for guide 2), comparable to position-matched shRNA controls (78.3% for guide 1, 51.5% for guide 2) (Fig.

6.1f), which control for accessibility and sequence in the target region; we therefore used this design for all further knockdown experiments. We also found that knockdown is most efficient with a spacer length of 28 nt (73.8%), is dose-responsive to both the protein and guide transfected vector amounts, and is not sensitive to RNA polymerase III promoter choice. (Extended Data Fig. 6.3a-d).

We next tested knockdown in HEK293FT cells of three endogenous genes: *KRAS*, *CXCR4*, and *PPIB*. We observed varying levels of knockdown, and for *KRAS* and *CXCR4*, LwaCas13a knockdown (40.4% for *PPIB*, 83.9% for *CXCR4*, 57.5% for *KRAS*) was similar to RNAi with position-matched shRNAs (63.0% for *PPIB*, 73.9% for *CXCR4*, 44.3% for *KRAS*) (Fig. 6.1g). We also found that knockdown of *KRAS* was possible with either U6 or tRNA^{Val} promoters (Extended Data Fig. 6.3e). Similar results were obtained in the A375 melanoma cell line (Extended Data Fig. 3f). In all cases tested, knockdown was abolished by mutating the catalytic domain of LwaCas13a (Extended Data Fig. 6.3g). To test if LwaCas13a knockdown is efficient in plants, we targeted three rice (*Oryza sativa*) genes with three guides per transcript and co-transfected LwaCas13a and guide vectors into *O. sativa* protoplasts (Fig. 6.1h). After transfection, we observed >50% knockdown with seven out of the nine guides and maximal knockdown of 78.0% (Fig. 6.1i).

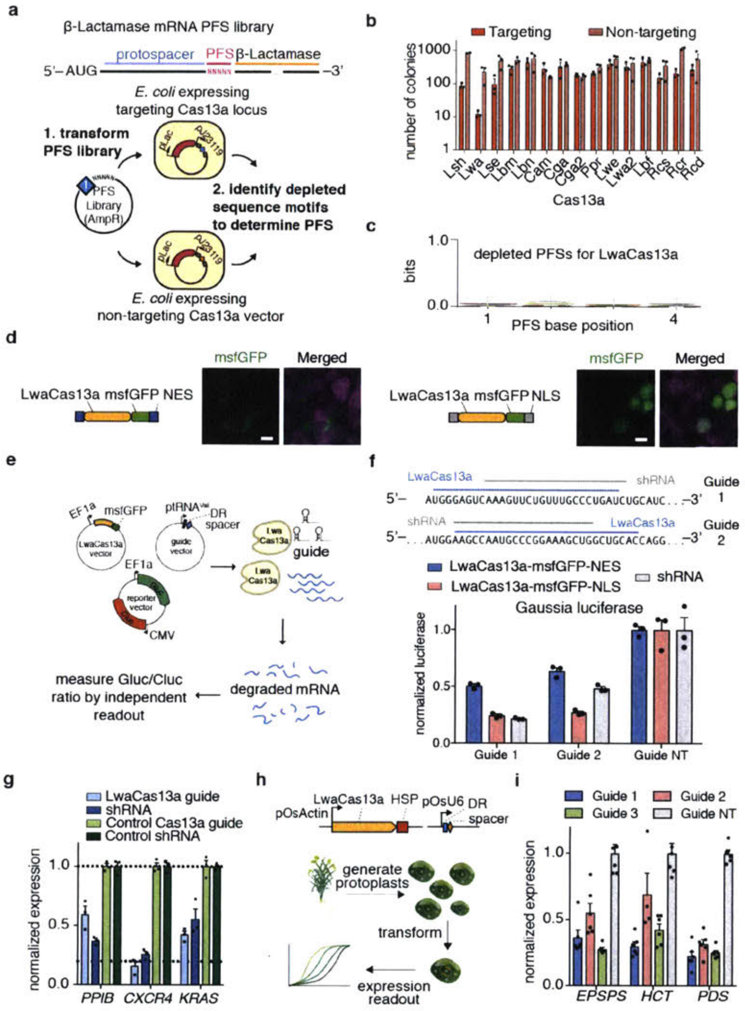


Figure 6.1: Cas13a from Leptotrichia wadei (LwaCas13a) is capable of eukaryotic transcript knockdown.

A)Schematic of protospacer flanking site (PFS) characterization screen of Cas13a orthologs.

B) Quantitation of Cas13a activity in *E. coli* measured by colony survival from PFS screen (n = 2 or 3).

C)*In vivo* PFS screening shows LwaCas13a has a minimal PFS preference. Error bars indicate an approximate Bayesian 95% confidence interval.

D)Imaging showing localization and expression of each of the mammalian constructs. Scale bars, 10μm.

E)Schematic of the mammalian luciferase reporter system used to evaluate knockdown.

F)Knockdown of Gaussia luciferase (Gluc) using engineered variants of LwaCas13a. Sequences for guides and shRNAs are shown above.

G)Knockdown of three different endogenous transcripts with LwaCas13a compared to corresponding RNAi constructs.

H)Schematic for LwaCas13a knockdown of transcripts in rice (*Oryza sativa)* protoplasts.

I)LwaCas13a knockdown of three transcripts in *O. sativa* protoplasts using three targeting guides per transcript (n = 4 or 6). All values are mean ± SEM with n = 3, unless otherwise noted.

To evaluate the range of efficiency of LwaCas13a knockdown, we tiled guides along the length of four transcripts: Gluc, Cluc, *KRAS*, and *PPIB* (Fig. 6.2a). The Gluc and Cluc tiling screens revealed guides with greater than 60% knockdown (Fig. 6.2b,c), with the majority of Gluc targeting guides exhibiting >50% knockdown and up to 83% knockdown. To compare LwaCas13a knockdown with RNAi, we selected the top three performing guides against Gluc and Cluc and compared them to position-matched shRNAs. We found that five out of six top performing guides achieved significantly higher levels of knockdown ($p < 0.05$) than their matched shRNAs (Extended Data Fig. 6.3h). For endogenous genes, we found that, while knockdown efficiency was transcript dependent, there was maximal knockdown of 85% and 75% for *KRAS* and *PPIB,* respectively (Fig. 6.2d,e). We selected the top three guides from the *KRAS* and *PPIB* tiling screens and observed robust knockdown with LwaCas13a (53.7%-88.8%) equivalent to levels attained by shRNA knockdown (61.8%-95.2%), with shRNA significantly better for 2 out of 6 guides ($p < 0.01$) and LwaCas13a significantly better for 2 out of 6 guides ($p < 0.01$) (Fig. 6.2f). LwaCas13a can also mediate significant knockdown of the nuclear transcripts *MALAT1* and *XIST*(Hutchinson et al., 2007), whereas position-matched shRNAs showed no detectable knockdown ($p > 0.05$) (Fig. 6.2g,h, Extended Data Fig. 6.3i)

LshCas13a activity is governed by target accessibility in *E. coli* (Abudayyeh et al., 2016), and we therefore used our data from the four tiling screens to investigate whether LwaCas13a activity is higher for guides located in regions of accessibility. We found that the most effective guides were closer together than expected by chance (Extended Data Fig. 6.4a), and predicted target accessibility could explain some of the variation in targeting efficacy (4.4%-16% of the variation in knockdown) (Extended Data Fig. 6.4b-d).

Because LwaCas13a can process its own pre-crRNA(East-Seletsky et al., 2016), it offers the possibility of streamlined multiplexed delivery of LwaCas13a guides(Zetsche et al., 2017). We designed five different guides against the endogenous *PPIB, CXCR4, KRAS, TINCR,* and *PCAT* transcripts and delivered the targeting system as a CRISPR array with 28-nt guides flanked by 36-nt DRs (representing an unprocessed DR and a truncated spacer), under expression of the U6 promoter. We found levels of knockdown for each gene that were comparable to single or pooled guide controls (Fig. 6.2i). To evaluate specificity in this context, we tested multiplexed delivery of three guides against *PPIB, CXCR4,* and *KRAS* or three variants where each one of the three guides was replaced with a non-targeting guide. We found that in each case where a guide was absent from the array, only the targeted transcripts were reduced (Fig. 6.2j).

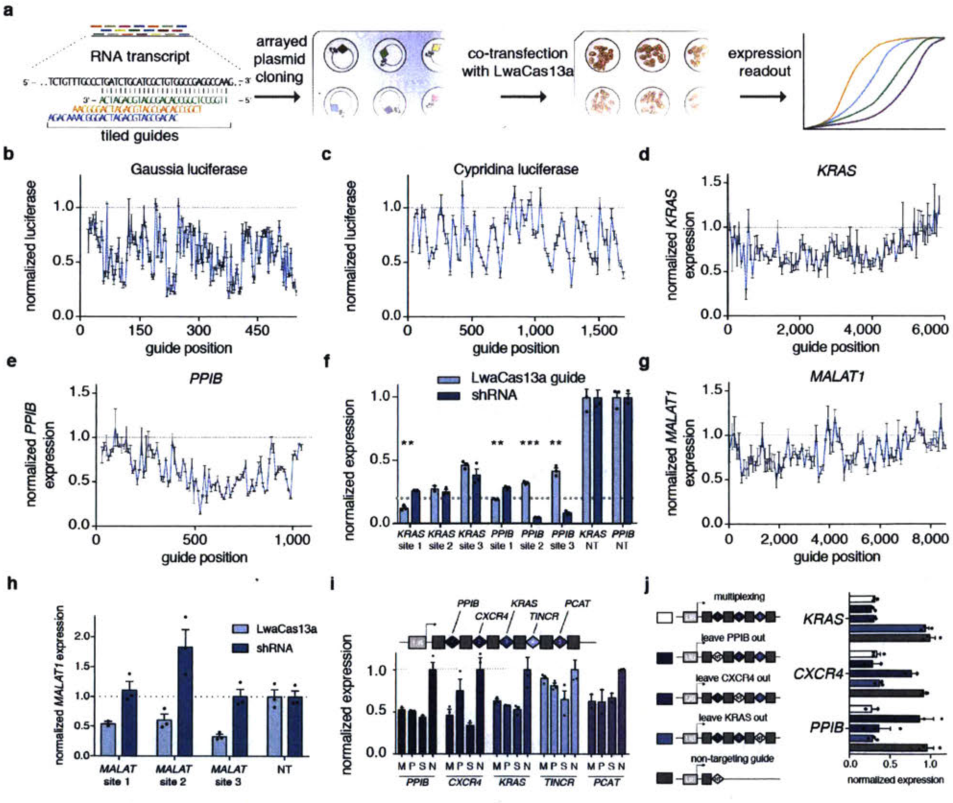


Figure 6.2: LwaCas13a arrayed screening of mammalian coding and non-coding RNA targets and multiplexed guide delivery.

A)Schematic of LwaCas13a arrayed screening.

B)Arrayed knockdown screen of 186 guides evenly tiled across the Gluc transcript.

C)Arrayed knockdown screen of 93 guides evenly tiled across Cluc.

D) Arrayed knockdown screen of 93 guides evenly tiled across *KRAS* transcript.

E)Arrayed knockdown screen of 93 guides evenly tiled across *PPIB* transcript.

F)Validation of the top three guides from the endogenous arrayed knockdown screens with shRNA comparisons (n = 2 or 3). ***p < 0.001; **p < 0.01; two-tailed student's T-test).

G)Arrayed knockdown screen of 93 guides evenly tiled across the *MALAT1* transcript.

H)Validation of top three guides from the endogenous arrayed *MALAT1* knockdown screen with shRNA comparisons (n= 2 or 3).

I)Multiplexed knockdown of five endogenous genes through delivery of five guides in a CRISPR array under the expression of a single promoter (n = 2 or 3).

J)Three-guide arrays containing combinations of targeting and non-targeting spacers showing sequence-specific multiplexed knockdown (n = 2 or 3). All values are mean ± SEM with n = 3 (n represents the number of transfection replicates), unless otherwise noted.

To further investigate the specificity of LwaCas13a *in vivo,* we introduced single mismatches into guides targeting either Gluc (Fig. 6.3a) or endogenous genes (Fig. 6.3b, Extended Data Fig. 6.5a,b), as well as double mismatches (Fig. 6.3c and Extended Data Fig. 6.5c), and found that knockdown was sensitive to mismatches in the central seed region of the guide:target duplex, which we additionally confirmed by biochemical profiling (Extended Data Fig. 6.5d-k). To comprehensively search for off-target effects of LwaCas13a knockdown, we performed transcriptome-wide mRNA sequencing. We targeted the Gluc transcript with LwaCas13a or a position matched-shRNA construct, and found significant knockdown of the target transcript ($p < .01$) (Fig. 6.3d,e). Similar results were found for the same comparison when targeting *KRAS* and *PPIB* ($p < .05$) (Extended Data Figure 6.6a,b). Differential expression analysis indicated hundreds of significant off-targets in each of the shRNA conditions but none in LwaCas13a conditions (Fig. 6.3f), despite comparable levels of knockdown of the target transcripts (30.5%, 43.5%, and 64.7% for shRNA, 62.6%, 27.1%, and 29.2% for LwaCas13a, for Gluc, *KRAS,* and *PPIB,* respectively) (Fig. 6.3g). Additional analysis of the Gluc targeting RNA-seq comparisons suggested the shRNA libraries show higher variability between targeting and non-targeting conditions compared to LwaCas13a because of these off-target effects (Extended Data Fig. 6.6c-f, 7).

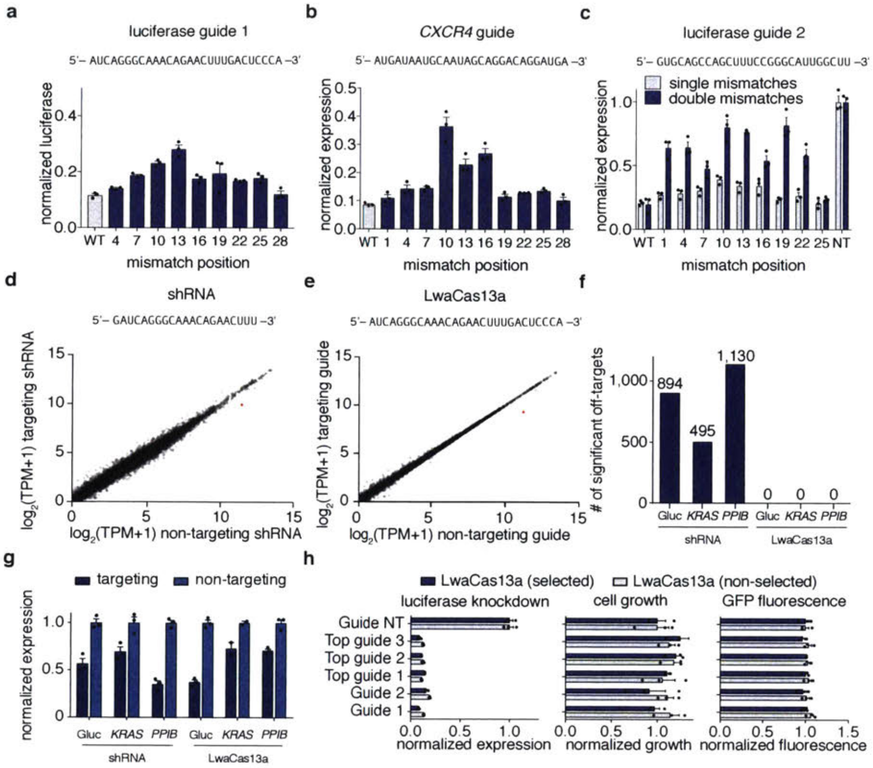


Figure 6.3: Evaluation of LwaCas13a knockdown specificity and comparisons to RNA interference.

A) Knockdown of Gluc evaluated with guides containing single mismatches at varying positions across the spacer sequence (shown above).

B) Knockdown of *CXCR4* evaluated with guides containing single mismatches at varying positions across the spacer sequence (shown above).

C) Knockdown of Gluc evaluated with guide 3 containing single or double mismatches at varying positions across the spacer sequence (shown above).

D) Expression levels in $\log_2$(transcripts per million (TPM)) values of all genes detected in RNA-seq libraries of non-targeting control (x-axis) compared to Gluc-targeting condition (y-axis) for shRNA. Shown is the mean of three biological replicates. The Gluc transcript data point is colored in red. The guide sequence used is shown above.

E) Expression levels in $\log_2$(transcripts per million (TPM)) values of all genes detected in RNA-seq libraries of non-targeting control (x-axis) compared to Gluc-targeting condition (y-axis) for LwaCas13a.

F) Differential gene expression analysis of six RNA-seq libraries (each with three biological replicates) comparing LwaCas13a knockdown to shRNA knockdown at three different genes (n =2 or 3).

G) Quantified mean knockdown levels for the targeted genes from the RNA-seq libraries.

H) Luciferase knockdown *(left)*, cell viability *(middle)*, and LwaCas13a-GFP expression *(right)* for cells transfected with LwaCas13a for 72 hours with and without selection. All values are mean ± SEM with n = 3 (n represents the number of transfection replicates), unless otherwise noted.

The collateral activity of LshCas13a has been directly observed biochemically *in vitro* and indirectly through growth suppression in bacteria(Abudayyeh et al., 2016), but the extent of this activity in mammalian cells is unclear. The multiplexed leave-one-out and RNA-seq analyses suggested a lack of collateral RNA degradation. We verified this by re-analyzing the knockdown tiling screens (Fig. 6.2b-e), finding that expression of the control gene did not correlate with the expression of the targeted gene (Gluc: R = -0.078, $p > 0.05$; *PPIB*: R = -0.058, $p > 0.05$; *KRAS*: R = -0.51, $p < 0.001$) (Extended Data Fig. 6.8a-h). Additionally, in the RNA-seq experiments there were no differentially expressed genes other than the target gene, indicating that LwaCas13a targeting does not lead to an observable cell stress response at the transcriptomic level(Subramanian et al., 2005) (Fig. 6.3d,e and Extended Data Fig. 6.6a,b), as would be reasonably expected if substantial collateral activity occurred. Furthermore, LwaCas13a-mediated knockdown of targeted transcripts did not affect the growth of mammalian cells expressing similar levels of LwaCas13a (Fig. 6.3h). Finally, because activation of non-specific RNA nucleases in mammalian cells results in detectable changes in RNA size distribution(Rath et al., 2015), we examined global RNA degradation in cells after LwaCas13a knockdown of Gluc transcripts and found no difference in the RNA integrity between targeting and non-targeting conditions ($p > 0.05$) (Extended Data Fig. 6.8i,j).

To expand the utility of LwaCas13a as a tool for studying RNA, we created a catalytically dead variant (dCas13a) by mutating catalytic arginine residues. We quantified RNA binding by dCas13a with RNA immunoprecipitation (RIP) (Fig. 6.4a) using guides containing the 36-nt DR and 28-nt spacers. We found that pulldown of dLwaCas13a targeted to either luciferase transcripts or *ACTB* mRNA (Fig.

6.4b) resulted in significant enrichment of the corresponding target over non-targeting controls (7.8-11.2x enrichment for luciferase and 2.1-3.3x enrichment for *ACTB*; p < 0.05), validating dLwaCas13a as a reprogrammable RNA binding protein.

One application for dLwaCas13a is as a transcript imaging platform. To reduce background noise due to unbound protein, we incorporated a negative-feedback (NF) system based upon zinc finger self-targeting and KRAB domain repression(Gross et al., 2013) (Fig. 6.4c). In comparison to dLwaCas13a, dLwaCas13a-NF effectively translocated from the nucleus to the cytoplasm when targeted to *ACTB* mRNA (Extended Data Fig. 6.9a). To further characterize translocation of dLwaCas13a-NF, we targeted *ACTB* transcripts with two guides and found that both guides increased translocation compared to a non-targeting guide (3.1-3.7x cellular/nuclear signal ratio; p < 0.001) (Fig. 6.4d,e and Extended Data Fig. 6.9b-d). To further validate dLwaCas13a-NF imaging, we analyzed the correlation of dLwaCas13a-NF signal to *ACTB* mRNA fluorescent *in situ* hybridization (FISH) signal (Extended Data Fig. 6.10a) and found that there was significant correlation and signal overlap for the targeting guides versus the non-targeting guide conditions (R = 0.27 and 0.30 for guide 1 and 2, respectively, and R = 0.00 for the non-targeting guide condition; p < 0.0001) (Extended Data Fig. 6.10b).

Using dLwaCas13a-NF, we investigated the accumulation of mRNA into stress granules(Nelles et al., 2016; Unsworth et al., 2010) by combining transcript imaging with visualization of stress granules marker *G3BP1*(Tourriere et al., 2003). In fixed samples, we found significant correlations between the dLwaCas13a-NF signal and the *G3BP1* fluorescence for *ACTB*-targeting guides compared to non-targeting controls (R = 0.49 and 0.50 for guide 1 and guide 2, respectively, and R = 0.08 for the non-targeting guide; p < 0.001) (Fig. 6.4f,g). We next performed stress granule tracking in live cells and found that dLwaCas13a-NF targeted to *ACTB* localized to significantly more stress granules per cell over time than the corresponding non-targeting control (p < 0.05) (Extended Data Fig. 6.10c,d).

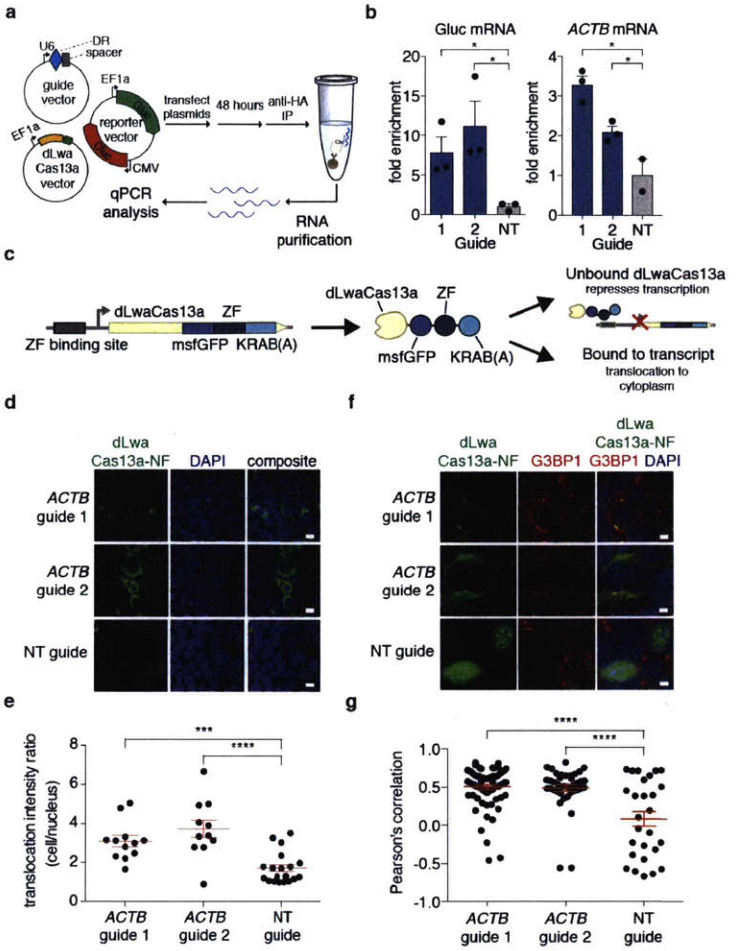


Figure 6.4: Catalytically-inactive LwaCas13a (dLwaCas13a) is capable of binding transcripts and tracking stress granule formation.

A) Schematic of RNA immunoprecipitation for quantitation of dLwaCas13a binding.

B) dLwaCas13a targeting Gluc and *ACTB* transcripts is significantly enriched compared to non-targeting controls. n = 2 or 3 (n represents the number of transfection replicates).

C) Schematic of dLwaCas13a-GFP-KRAB negative feedback (dLwaCas13a-NF) construct used for imaging.

D) Representative images for dLwaCas13a-NF imaging with multiple guides targeting *ACTB*. Scale bars, 10μm.

E) Quantitation of translocation of dLwaCas13a-NF. n = 12, 11, and 19 (Guides 1, 2, and NT) (n represents the number of individual cells analyzed).

F) Representative immunofluorescence images of HEK293FT cells treated with 400 uM sodium arsenite. Stress granules are indicated by *G3BP1* staining. Scale bars, 5μm.

G) *G3BP1* and dLwaCas13a-NF co-localization quantified per cell by Pearson's correlation. n = 75, 40, and 27 (Guides 1, 2, and NT) (n represents the number of individual cells analyzed). All values are mean ± SEM. ****p < 0.0001; ***p < 0.001; **p < 0.01; *p < 0.05. ns = not significant. A one-tailed student's t-test was used for comparisons in (**b**) and a two-tailed student's t-test was used for comparisons in (**e**) and (**g**).

6.3 Conclusion

These results show that LwaCas13a can be reprogrammed with guide RNAs to effectively knockdown or bind transcripts in mammalian cells. LwaCas13a knockdown is comparable to RNAi knockdown efficiency, but with substantially reduced off-targets. Furthermore, it can mediate nuclear RNA and multiplexed knockdown. Catalytically inactive dLwaCas13a can be used as a programmable RNA binding protein, which we adapted for live imaging transcript tracking. We anticipate that there will be a number of applications for LwaCas13a and dLwaCas13a, such as genome-wide pooled knockdown screening, interrogation of lncRNA and nascent transcript function, pulldown assays to study RNA-protein interactions, translational modulation, and RNA base editing. Importantly, we do not observe any evidence for collateral activity of LwaCas13a in mammalian cells. Our data show LwaCas13a functions in mammalian and plant cells with broad efficacy and high specificity, providing a platform for a range of transcriptome analysis tools.

6.4 Experimental Procedures

6.4.1 Cloning of orthologs for activity screen and recombinant expression

We synthesized human codon-optimized versions of fifteen Cas13a orthologs (Genscript, Jiangsu, China) and cloned them into pACYC184 under a pLac promoter. Adjacent to the Cas13a expression cassette, we cloned the ortholog's corresponding direct repeats flanking either a β-lactamase targeting or non-targeting spacer. Spacer array expression was driven by the J23119 promoter.

For purification of LwaCas13a, we cloned the mammalian codon-optimized LwaCas13a sequence into a bacterial expression vector for protein purification (6x His/Twin Strep SUMO, a pET-based expression vector received as a gift from Ilya Finkelstein, University of Texas-Austin).

6.4.2 Bacterial *in vivo* testing for Cas13a activity and PFS identity

Briefly, Cas13a is programmed to target a 5' stretch of sequence on the □-*lactamase* transcript flanked by randomized PFS nucleotides. Cas13a cleavage activity results in death of bacteria under ampicillin selection, and PFS depletion is subsequently analyzed by next generation sequencing. In order to allow for quantitative comparisons between orthologs, we cloned each Cas13a ortholog under a pLac promoter along with a single-spacer CRISPR array nearby under expression of the pJ23119 small RNA promoter.

To test for activity of Cas13a orthologs, 90 ng of ortholog expression plasmid with either targeting or non-targeting guide was co-transformed with 25 ng of a previously described β-lactamase target plasmid(Abudayyeh et al., 2016) into NovaBlue Singles competent cells (Millipore). Post-transformation, cells were diluted, plated on LB-agar supplemented with 100 μg/uL ampicillin and 25 μg/uL chloramphenicol, and incubated at 37°C overnight. Transformants were counted the next day.

For determination of LshCas13a and LwaCas13a PFS identity, 40 ng of ortholog expression plasmid with either targeting or non-targeting spacer was co-transformed with 25 ng of β-lactamase target plasmid into 2 aliquots of NovaBlue GigaSingles (Millipore) per biological replicate. Two biological replicates were performed. Post-transformation, cells were recovered at 37°C in 500 uL of SOC (ThermoFisher Scientific) per biological replicate for 1 hour, plated on bio-assay plates (Corning) with LB-agar (Affymetrix) supplemented with 100 μg/uL ampicillin and 25 μg/uL chloramphenicol, and incubated at 37°C for 16 hours. Colonies were then harvested by scraping, and plasmid DNA was purified with NuceloBond Xtra EF (Macherey-Nagel) for subsequent sequencing.

Harvested plasmid samples were prepared for next generation sequencing by PCR with barcoding primers and Illumina flow cell handles using NEBNext High Fidelity 2X Master Mix (New England Biosciences). PCR products were pooled and gel extracted using a Zymoclean gel extraction kit (Zymo Research) and sequenced using a MiSeq next generation sequencing machine (Illumina).

6.4.3 Computational analysis of PFS

From next generation sequencing of the LshCas13a and LwaCas13a PFS screening libraries, we aligned the sequences flanking the randomized PFS region and extracted the PFS identities. We collapsed PFS identities to 4 nucleotides to improve sequence coverage, counted the frequency of each unique PFS, and normalized to total read count for each library with a pseudocount of 1. Enrichment of each distribution as displayed in Extended Data Figure 1c. was calculated against the pACYC184 control (no protein/guide locus) as $-\log_2(f_{\text{condition}}/f_{\text{pACYC184}})$, where $f_{\text{condition}}$ is the frequency of PFS identities in the experimental condition and f_{pACYC184} is the frequency of PFS identities in the pACYC184 control. For analysis of a conserved PFS motif, top depleted PFS identities were calculated using each condition's non-targeting control as follows: $-\log_2(f_{\text{i,targeting}}/f_{\text{i,non-targeting}})$ where $f_{\text{i,targeting}}$ is the frequency of PFS identities in condition i with targeting spacer and $f_{\text{i,non-targeting}}$ is the frequency of PFS identities in condition i with non-targeting spacer.

6.4.4 Purification of LwaCas13a

Purification of LwaCas13a was performed as previously described(Gootenberg et al., 2017b). Briefly, LwaCas13a bacterial expression vectors were transformed into Rosetta 2(DE3)pLysS singles Competent Cells (Millipore) and 4 L of Terrific Broth 4 growth media (TB) was seeded with a starter culture. Cell protein expression was induced with IPTG and after overnight growth, the cell pellet was harvested and stored at -80°C. Following cell lysis, protein was bound using a StrepTactin Sepharose resin (GE) and protein was eluted by SUMO protease digestion (ThermoFisher). Protein was further purified by cation exchange using a HiTrap SP HP cation exchange column (GE Healthcare Life Sciences) and subsequently by gel filtration using a Superdex 200 Increase 10/300 GL column (GE Healthcare Life Sciences), both steps via FPLC (AKTA PURE, GE Healthcare Life Sciences). Final fractions containing LwaCas13a protein were pooled and concentrated into Storage Buffer (600 mM NaCl, 50 mM Tris-HCl pH 7.5, 5% Glycerol, 2 mM DTT) and aliquots were frozen at -80°C for long-term storage.

6.4.5 Cloning of mammalian expression constructs

The human codon optimized Cas13a gene was synthesized (Genscript) and cloned into a mammalian expression vector with either a nuclear export sequence (NES) or nuclear localization sequence (NLS) under expression of the EF1-α promoter. Because of the stability conferred by monomeric-super-folded GFP (msfGFP), we fused msfGFP to the C-terminus of LwaCas13a. The full-length direct-repeat of LwaCas13a was used for cloning the guide backbone plasmid with expression under a U6 promoter. The catalytically-inactive LwaCas13a-msfGFP construct (dead LwaCas13a or dLwaCas13a) was generated by introducing R474A and R1046A mutations in the two HEPN domains. A drug-selectable version of LwaCas13a-msfGFP was generated by cloning the protein into a backbone with the Blasticidin selection marker linked to the C-terminus via a 2A peptide sequence. The negative feedback version of the dLwaCas13a-msfGFP construct (dLwaCas13a-NF) was generated by cloning a zinc-finger binding site upstream of the promoter of dLwaCas13a-msfGFP and fusing a zinc finger and KRAB domain to the C-terminus.

The reporter luciferase construct was generated by cloning Cypridinia luciferase (Cluc) under expression of the CMV promoter and Gaussia luciferase (Gluc) under expression of the EF1-α promoter both on a single vector. Expression of both luciferases on a single vector allows one luciferase to serve as a dosing control for normalization of knockdown of the other luciferase, controlling for variation due to transfection conditions.

For the endogenous knockdown experiments in Fig. 6.1g, guides and shRNAs were designed using the RNAxs siRNA design algorithm(Tafer et al., 2008). The prediction tool was used to design shRNAs, and guides were designed in the same location to allow for comparison between shRNA and LwaCas13a knockdown.

For the plant knockdown experiments, the rice actin promoter (*pOsActin*) was PCR amplified from pANIC6A(Mann et al., 2012) and LwaCas13a was PCR amplified from human expression LwaCas13a constructs. These fragments were ligated into existing plant expression plasmids such that LwaCas13a was driven by the rice actin promoter and transcription was terminated by the HSP terminator while the LwaCas13a gRNAs were expressed from the rice U6 promoter (*pOsU6*).

6.4.6 Protoplast Preparation

Green rice protoplasts (*Oryza sativa* L. ssp. *japonica* var. Nipponbare) were prepared as previously described(Zhang et al., 2011b) with slight modifications. Seedlings were grown for 14 days and protoplasts were resuspended in MMG buffer containing 0.1 M $CaCl_2$. This modified MMG buffer was used to prepare fresh 40% PEG buffer as well as in place of WI buffer. Finally, protoplasts were kept in total darkness for 48 hours post-transformation. All other conditions were as previously described.

6.4.7 Nucleic acid target and crRNA preparation for *in vitro* reactions and collateral activity

For generation of nucleic acid targets, oligonucleotides were PCR amplified with KAPA Hifi Hot Start (Kapa Biosystems). dsDNA amplicons were gel extracted and purified using a MinElute gel extraction kit (Qiagen). The resulting purified dsDNA was transcribed via overnight incubation at 30°C with the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). Transcribed RNA was purified using the MEGAclear Transcription Clean-up kit (Thermo Fisher).

To generate crRNAs, oligonucleotides were ordered as DNA (Integrated DNA Technologies) with an additional 5' T7 promoter sequence. crRNA template DNA was annealed with a T7 primer (final concentrations 10 uM) and transcribed via overnight incubation at 37°C with the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). The resulting transcribed crRNAs were purified with RNAXP clean beads (Beckman Coulter), using a 2x ratio of beads to reaction volume, supplemented with additional 1.8x ratio of isopropanol (Sigma).

6.4.8 LwaCas13a cleavage and collateral activity detection

For biochemical characterization of LwaCas13a, assays were performed as previously described(Abudayyeh et al., 2016). Briefly, nuclease assays were performed with 160 nM of end-labeled ssRNA target, 200 nM purified LwaCas13a, and 100 nM crRNA, unless otherwise indicated. All assays were performed in nuclease assay buffer (40 mM Tris-HCl, 60 mM NaCl, 6 mM $MgCl_2$, pH 7.3). For array processing, 100 ng of *in vitro* transcribed array was used per nucelease assay. Reactions were allowed to proceed for 1 hour at 37°C (unless otherwise indicated) and were then quenched with proteinase buffer (proteinase K, 60 mM EDTA, and 4 M Urea) for 15 minutes at 37°C. The reactions were then denatured with 4.5 M urea denaturing buffer at 95°C for 5 minutes. 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).

Collateral activity detection assays were performed as previously described(Gootenberg et al., 2017b). Briefly, reactions consisted of 45 nM purified LwaCas13a, 22.5 nM crRNA, 125 nM quenched fluorescent RNA reporter (RNAse Alert v2, Thermo Scientific), 2 µL murine RNase inhibitor (New England Biolabs), 100 ng of background total human RNA (purified from HEK293FT culture), and varying amounts of input nucleic acid target, unless otherwise indicated, in nuclease assay buffer (40 mM Tris-HCl, 60 mM NaCl, 6 mM $MgCl_2$, pH 7.3). Reactions were allowed to proceed for 1-3 hr at 37°C (unless otherwise indicated) on a fluorescent plate reader (BioTek) with fluorescent kinetics measured every 5 min.

6.4.9 Cloning of tiling guide screens

For tiling guide screens, spacers were designed to target mRNA transcripts at even intervals to fully cover the entire length of the transcript. Spacers (ordered from IDT) were annealed and golden-gate cloned into LwaCas13a guide expression constructs with either a tRNA^{val} promoter (Gluc and Cluc screens) or U6 promoter (all endogenous screens).

6.4.10 Mammalian cell culture and transfection for knockdown with LwaCas13a

All mammalian cell experiments were performed in the HEK293FT line (ATCC) unless otherwise noted. HEK293FT cells were cultured in Dulbecco's Modified Eagle Medium with high glucose, sodium pyruvate, and GlutaMAX (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (VWR Seradigm) and 1X Penicillin-Streptomycin (Thermo Fisher Scientific). Cells were passaged to maintain confluency below 70%. For experiments involving A375 (ATCC), cells were cultured in RPMI Medium 1640 (Thermo Fisher Scientific) supplemented with 9% fetal bovine serum (VWR Seradigm) and 1X Penicillin-Streptomycin (Thermo Fisher Scientific).

To test knockdown of endogenous genes, Lipofectamine 2000 (Thermo Fisher Scientific) transfections were performed with 150 ng of LwaCas13a plasmid and 250 ng of guide plasmid per well, unless otherwise noted. Experiments testing knockdown of reporter plasmids were supplemented with 12.5 ng reporter construct per well. Sixteen hours prior to transfection, cells were plated in 96-well plates at approximately 20,000 cells/well and allowed to grow to 90% confluency overnight. For each well, plasmids were combined with Opti-MEM® I Reduced Serum Medium (Thermo Fisher) to a total of 25 uL, and separately 0.5 uL of Lipofectamine 2000 was combined with 24.5 uL of Opti-MEM. Plasmid and lipofectamine solutions were then combined, incubated for 5 minutes, and slowly pipetted onto cells to prevent disruption.

6.4.11 Transformation of green rice protoplasts

For the green rice experiments, plasmids expressing each LwaCas13a and the corresponding guide RNA were mixed in equimolar ratios such that a total of 30 μg of DNA was used to transform a total of 200,000 protoplasts per transformation.

6.4.12 Measurement of luciferase activity

Media containing secreted luciferase was harvested at 48 hours post transfection, unless otherwise noted. Media was diluted 1:5 in PBS and then luciferase activity was measured using the BioLux Cypridinia and Biolux Gaussia luciferase assay kits (New England Biolabs) on a Biotek Synergy 4 plate reader with an injection protocol. All replicates were performed as biological replicates.

6.4.13 Harvest of total RNA and quantitative PCR

For gene expression experiments in mammalian cells, cell harvesting and reverse transcription for cDNA generation was performed using a previously described modification(Joung et al., 2017) of the commercial Cells-to-Ct kit (Thermo Fisher Scientific) 48 hours post-transfection. Transcript expression was then quantified with qPCR using Fast Advanced Master Mix (Thermo Fisher Scientific) and TaqMan qPCR probes (Thermo Fisher Scientific) with *GAPDH* control probes (Thermo Fisher Scientific). All qPCR reactions were performed in 5 uL reactions with 4 technical replicates in 384-well format, and read out using a LightCycler 480 Instrument II (Roche). For multiplexed targeting reactions, readout of different targets was performed in separate wells. Expression levels were calculated by subtracting housekeeping control (*GAPDH*) Ct values from target Ct values to normalize for total input, resulting in ΔCt levels. Relative transcript abundance was computed as 2^(-ΔCt). All replicates were performed as biological replicates

For gene expression experiments in plant cells, total RNA was isolated after 48 hours of incubation using Trizol according to the manufacturer's protocol. One nanogram of total RNA was used in the SuperScript III Plantinum SYBR Green One-Step qRT-PCR Kit (Invitrogen) according to the manufacturer's protocol. All samples were run in technical triplicate of three biological replicates in a 384-well format on a LightCycler 480 Instrument (Roche). All PCR primers were verified as being specific based on melting curve analysis and are as follows: *OsEPSPS* (Os06g04280), 5' – TTG CCA TGA CCC TTG CCG TTG TTG – 3' and 5' – TGA TGA TGC AGT AGT CAG GAC CTT – 3'; *OsHCT* (Os11g07960), 5' – CAA GTT TGT GTA CCC GAG GAT TTG – 3' and 5' – AGC TAG TCC CAA TAA ATA TGC GCT – 3'; *OsEF1a* (Os03g08020), 5' – CTG TAG TCG TTG GCT GTG GT –

3' and 5' - CAG CGT TCC CCA AGA AGA GT - 3'. Primers for *OsEF1a* were previously described(Jain et al., 2006).

For analysis of RNA quality post-knockdown with LwaCas13a, total RNA was harvested by lysing cells using TRI Reagent® and purifying the RNA using the Direct-zol RNA MiniPrep Plus kit (Zymo). Four ng of total RNA was analyzed using a RNA 6000 Pico Bioanalyzer kit (Agilent).

6.4.14 Computational analysis of target accessibility

To first analyze target accessibility, top guides from the tiling screen were analyzed to determine whether they grouped closer together than expected under the assumption that if there were regions of accessibility, multiple guides in that region would be expected to be highly active. Top guides were defined as the top 20% of performing guides for the Gluc tiling screen and top 30% of performing guides for the Cluc, *KRAS*, and *PPIB* tiling screens. A null probability distribution was generated for pair-wise distances between guides by randomly simulated 10,000 guide positions and then compared to experimentally determined top guide pair-wise distances.

Accessibility was predicted using the RNApl fold algorithm in the Vienna RNA software suite(Bernhart et al., 2006). The default window size of 70 nt was used and probability of a target region being unpaired was calculated as the average of the 28 single-nt unpaired probabilities across the target region. These accessibility curves were smoothened and compared to smoothened knockdown curves across each of the four transcripts and correlations between the two factors and their significance were computed using Pearson's correlation coefficient using the SciPy Python package (pearsonr function). The probability space of these two factors was also visualized by performing 2D kernel density estimation across the two variables.

6.4.15 RNA sequencing and analysis

For specificity analysis of LwaCas13a knockdown, RNA sequencing was performed on mRNA from knockdown experiments involving both LwaCas13a and shRNA constructs. Total RNA was prepared from transfection experiments after 48 hours using the Qiagen RNeasy Plus Mini kit. mRNA was then extracted using the NEBNext Poly(A) mRNA Magnetic Isolation Module and RNA-seq libraries were prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina. RNA-sequencing libraries were sequenced on an Illumina NextSeq instrument with at least 10M reads per library.

An index was generated using the RefSeq GRCh38 assembly and reads were aligned and quantified using Bowtie and RSEM v1.2.31 using default parameters(Li and Dewey, 2011). Transcript per million (TPM) values were used for expression counts and were transformed to log-space by taking the $\log_2(TPM+1)$.

To find differentially expressed genes, Student's t-test was performed on the three targeting replicates versus the three non-targeting replicates. The statistical analysis was only performed on genes that had a $\log_2(TPM+1)$ value greater than 2.5 in at least two of the six replicates. Only genes that had a differential expression greater than 2 or less than 0.75 and a false discovery rate < 0.10 were reported to be significantly differentially expressed.

Cross-correlations between replicates and averages of replicates were performed using Kendall's tau coefficient. The variation of shRNA versus LwaCas13a libraries was analyzed by considering the distribution of standard deviations for gene expression across the 6 replicates (3 targeting and 3 non-targeting replicates) and plotted as violin plots.

6.4.16 Cell viability assay

Mammalian cells were transfected with luciferase reporter target, guide plasmid, and either LwaCas13a or drug-selectable LwaCas13a. Twenty-four hours post-transfection, cells were split 1:5 into fresh media and drug-selectable LwaCas13a samples were supplemented with 10 ug/mL Blasticidin S (Thermo Fisher Scientific). After 48 hours of additional growth, cells were assayed for luciferase knockdown, maintenance of LwaCas13a expression via GFP fluorescence measurement on a multimode plate reader (Biotek Neo2), and cell growth by CellTiter-Glo® Luminescent Cell Viability Assay (Promega).

6.4.17 Quantifying dLwaCas13a binding with RIP

For RNA immunoprecipitation (RIP) experiments, HEK293FT cells were plated in 6-well plates and transfected with 1.3 ug of dLwaCas13a expression plasmid and 1.7 ug of guide plasmid, with an additional 150 ng of reporter plasmid for conditions involving reporter targeting. 48 hours post transfection, cells were washed twice with ice-cold PBS (Sigma) and fixed with 0.2% paraformaldehyde (Electron Microscopy Sciences) in PBS for 15 minutes at room temperature. After fixation, the paraformaldehyde was removed, 125 mM glycine in PBS was added to quench crosslinking, and the cells were incubated for 10 minutes. Cells were washed twice with ice-cold PBS, harvested by scraping, and the cell suspension was centrifuged at 800 g for 4 minutes to pellet the cells. The supernatant was removed and the pellet was washed with PBS prior to lysis. Cells were lysed with 200 μL of 1X RIPA Buffer (Cell Signaling) supplemented with cOmplete™ ULTRA Tablets, EDTA-free (Sigma) and Ribonuclease inhibitor (Sigma R1158). Cells were allowed to lyse on ice for 10 minutes and then sonicated for 2 minutes with a 30 sec on/30 sec off cycle at low intensity on a Bioruptor sonicator (Diagenode). Insoluble material was pelleted by centrifugation at 16,000 g for 10 minutes at 4°C, and the supernatant containing cleared lysate was used for pulldown with magnetic beads.

To conjugate antibodies to magnetic beads, 100 μL/sample of Dynabeads® Protein A for Immunoprecipitation (Thermo Fisher Scientific) were pelleted by application of a magnet, and the supernatant was removed. Beads were resuspended in 200 μL of wash buffer (PBS supplemented with 0.02% Tween-20 (Sigma)) and 5 μg of rabbit anti-Mouse IgG (Sigma M7023) was added. The sample was incubated for 10 minutes at room temperature on a rotator to allow antibody to conjugate to the beads. After incubation, beads were pelleted via magnet, supernatant was removed, and beads were washed twice with wash buffer. The pellet was resuspended in 100 μL wash buffer and split into two 50 μL volumes for conjugation of anti-HA antibody (Thermo Fisher Scientific 26183) or IgG antibody control (Sigma I5381). For each antibody, 2.5 μg of antibody was added with 200 μL wash buffer and incubated for 10 minutes at room temperature on a rotator. Post-incubation, beads were pelleted via magnet and washed twice with wash buffer, and resuspended in 200 µL 1X RIPA with Ribonuclease inhibitor (Sigma R1158) and protease inhibitor cocktail (Sigma P8340). 100 µL of sample lysate was added to beads and rotated overnight at 4°C.

After incubation with sample lysate, beads were pelleted, washed three times with 1X RIPA, 0.02% Tween-20, and then washed with DNase buffer (350 mM Tris-HCl [pH 6.5]; 50 mM MgCl2; 5 mM DTT). Beads were resuspended in DNase buffer and TURBO DNase (Life Technologies) was added to final concentration of 0.08 units/µl. DNase was incubated 30 minutes at 37°C on a rotator. Proteins were then digested by addition of Proteinase K (New England Biosciences) to a final concentration of 0.1 units/µl and incubated at 37°C with rotation for an additional 30 minutes. For denaturation and purification, urea (Sigma) was added to a final concentration of 2.5 M, samples were incubated for 30 minutes, and RNA was purified using a Direct-Zol RNA miniprep (Zymo Research). Purified RNA was reverse transcribed to cDNA using the qScript Flex cDNA (Quantabio) and pulldown was quantified with qPCR using Fast Advanced Master Mix and TaqMan qPCR probes. All qPCR reactions were performed in 5 µL reactions with 4 technical replicates in 384-well format, and read out using a LightCycler 480 Instrument II. Enrichment was quantified for samples as compared to their matched IgG antibody controls.

6.4.18 Translocation measurement of LwaCas13a and LwaCas13a-NF

HEK293FT cells were plated in 24-well tissue culture plates on poly-D-lysine coverslips (Corning) and transfected with 150 ng dLwaCas13a-NF vector and 300 ng guides for imaging *ACTB*. For translocation experiments, cells were fixed with 4% PFA and permeabilized with 0.2% Triton X-100 after 48 hours and mounted using anti-fade mounting medium with DAPI (Vectashield). Confocal microscopy was performed using a Nikon Eclipse Ti1 with Andor Yokagawa Spinning disk Revolution WD system.

Nuclear export of dLwaCas13a-NF with guides targeting *ACTB* mRNA was analyzed by measuring the average cytoplasmic and nuclear msfGFP fluorescence and comparing the ratio across many cells between targeting and non-targeting conditions.

6.4.19 Fluorescent *in situ* hybridization (FISH) of *ACTB* transcript

HEK293FT cells were plated in 24-well tissue culture plates on poly-D-lysine coverslips (Corning) and transfected with 75 ng dLwaCas13a-NF vector and 250 ng guides for imaging *ACTB.* After 48 hours, cells were fixed with 4% PFA for 45 minutes. The QuantiGene viewRNA ISH Cell assay kit (Affymetrix) was used for performing FISH on the cell samples according to the manufacturer's protocol. After finishing the FISH procedure, coverslips were mounted using anti-fade mounting medium (Vectashield). Confocal microscopy was performed using a Nikon Eclipse Ti1 with Andor Yokagawa Spinning disk Revolution WD system.

6.4.20 Tracking of LwaCas13a to stress granules

HEK293FT cells were plated in 24-well tissue culture plates on poly-D-lysine coverslips (Corning) and transfected with 75 ng dLwaCas13a-NF vector and 250 ng guides for imaging *ACTB.* For stress granule experiments, 200 μM sodium arsenite was applied for 1 hour prior to fixing and permeabilizing the cells. For immunofluorescence of *G3BP1*, cells were blocked with 20% goat serum, and incubated over night at room temperature with anti-*G3BP1* primary antibody (Abnova H00010146-B01P). Cells were then incubated for 1 hour with secondary antibody labeled with Alexa Fluor 594 and mounted using anti-fade mounting medium with DAPI (Vectashield). Confocal microscopy was performed using a Nikon Eclipse Ti1 with Andor Yokagawa Spinning disk Revolution WD system.

Stress granule co-localization with dLwaCas13a-NF was calculated using the average msfGFP and *G3BP1* signal per cell using Pearson's correlation coefficient. The colocalization analyses were performed in the image analysis software FIJI (Schindelin et al., 2012) using the Coloc 2 plugin.

For live imaging experiments, HEK293FT cells were plated in 96-well tissue culture plates and transfected with 150 ng dLwaCas13a-NF vector, 300 ng guides for imaging *ACTB*, and 5 ng of *G3BP1*-RFP reporter. After 48 hours, the cells were subjected to 0 μM or 400 μM sodium arsenite and imaged every 15 minutes every 2 hours on an Opera Phenix High Content Screening System (PerkinElmer) using the spinning disk confocal setting with 20x water objective. Cells were maintained at 37 ℃ in a humidified chamber with 50% CO_2. Live cell dLwaCas13a-NF colocalization with *G3BP1*-RFP in stress granules was measured using the Opera Phenix Harmony software (PerkinElmer).

6.5 Acknowledgements

This work was published in (Abudayyeh et al., 2017). We thank Maria Alimova, David Feldman, Fei Chen, John Doench, Jesse Engreitz, Naomi Habib, Danielle Tenen, A. Allen, Rhiannon Macrae, and Rich Belliveau for discussions and support.

Chapter 7

RNA editing with CRISPR-Cas13

This chapter is adapted from the following article:

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. Science 358, 1019-1027.

Contributions: David Cox, Omar Abudayyeh, and Jonathan Gootenberg are co-first authors (*). David Cox, Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang conceived and designed the study. David Cox, Omar Abudayyeh, and Jonathan Gootenberg participated in the design and execution of all experiments. Brian Franklin cloned plasmid constructs, generated sequencing libraries, and performed luciferase assays. Max Kellner assigned with plasmid clonings. Julia Joung helped with sequencing library generation. David Cox, Omar Abudayyeh, Jonathan Gootenberg, and Feng Zhang wrote the paper with input from all authors.

7.1 Abstract

Nucleic acid editing holds promise for treating genetic disease, particularly at the RNA level, where disease-relevant sequences can be rescued to yield functional protein products. Type VI CRISPR-Cas systems contain the programmable single-effector RNA-guided RNases Cas13. Here, we profile Type VI systems to engineer a Cas13 ortholog capable of robust knockdown and demonstrate RNA editing by using catalytically-inactive Cas13 (dCas13) to direct adenosine to inosine deaminase activity by ADAR2 to transcripts in mammalian cells. This system, referred to as RNA Editing for Programmable A to I Replacement (REPAIR), has no strict sequence constraints, can be used to edit full-length transcripts containing pathogenic mutations. We further engineer this system to create a high specificity variant, REPAIRv2, that is 919 times more specific than REPAIRv1 as well as minimize the system to ease viral delivery. REPAIR presents a promising RNA editing platform with broad applicability for research, therapeutics, and biotechnology.

7.2 Introduction

Precise nucleic acid editing technologies are valuable for studying cellular function and as novel therapeutics. Current editing tools, based on programmable nucleases such as the prokaryotic clustered regularly interspaced short palindromic repeats (CRISPR)-associated nucleases Cas9 (Cong et al., 2013; Komor et al., 2017; Mali et al., 2013c; Wu et al., 2014) or Cpf1(Zetsche et al., 2015b), have been widely adopted for mediating targeted DNA cleavage which in turn drives targeted gene disruption through non-homologous end joining (NHEJ) or precise gene editing through template-dependent homology-directed repair (HDR) (Kim and Kim, 2014). NHEJ utilizes host machineries that are active in both dividing and post-mitotic cells and provides efficient gene disruption by generating a mixture of insertion or deletion (indel) mutations that can lead to frame shifts in protein coding genes. HDR, in contrast, is mediated by host machineries whose expression is largely limited to replicating cells. Accordingly, the development of gene-editing capabilities for post-mitotic cells remains a major challenge. DNA base editors, consisting of a fusion between Cas9 nickase and cytidine deaminase can mediate efficient cytidine to uridine conversions within a target window and significantly reduce the formation of double-strand break induced indels (Komor et al., 2016; Nishida et al., 2016). However the potential targeting sites of DNA base editors are limited by the requirement of Cas9 for a protospacer adjacent motif (PAM) at the editing site (Kim et al., 2017). Here, we describe the development of a precise and flexible RNA base editing technology using the type VI CRISPR-associated RNA-guided RNase Cas13 (Abudayyeh et al., 2016; Shmakov et al., 2015; Shmakov et al., 2017a; Smargon et al., 2017b).

Cas13 enzymes have two Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) endoRNase domains that mediate precise RNA cleavage with a preference for targets with protospacer flanking site (PFS) motif observed biochemically and in bacteria (Abudayyeh et al., 2016; Shmakov et al., 2015). Three Cas13 protein families have been identified to date: Cas13a (previously known as C2c2), Cas13b, and Cas13c (Shmakov et al., 2017a; Smargon et al., 2017b). We recently reported that Cas13a enzymes can be adapted as tools for nucleic acid detection (Gootenberg et al., 2017c) as well as mammalian and plant cell RNA knockdown and transcript tracking (Abudayyeh et al., 2017). Interestingly, the biochemcial PFS was not required for RNA interference with Cas13a (Abudayyeh et al., 2017). The programmable nature of Cas13 enzymes makes them an attractive starting point to develop tools for RNA binding and perturbation applications.

The adenosine deaminase acting on RNA (ADAR) family of enzymes mediates endogenous editing of transcripts via hydrolytic deamination of adenosine to inosine, a nucleobase that is functionally equivalent to guanosine in translation and splicing (Nishikura, 2010; Tan et al., 2017). There are two functional human ADAR orthologs, ADAR1 and ADAR2, which consist of N-terminal double stranded RNA-binding domains and a C-terminal catalytic deamination domain. Endogenous target sites of ADAR1 and ADAR2 contain substantial double stranded identity, and the catalytic domains require duplexed regions for efficient editing *in vitro* and *in vivo* (Bass and Weintraub, 1988; Matthews et al., 2016). Importantly, the ADAR catalytic domain is capable of deaminating target adenosines without any protein co-factors *in vitro* (Zheng et al., 2017). ADAR1 has been found to target mainly repetitive regions whereas ADAR2 mainly targets non-repetitive coding regions (Tan et al., 2017). Although ADAR proteins have preferred motifs for editing that could restrict the potential flexibility of targeting, hyperactive mutants, such as ADAR2(E488Q) (Kuttan and Bass, 2012), relax sequence constraints and increase adenosine to inosine editing rates. ADARs preferentially deaminate adenosines mispaired with cytidine bases in RNA duplexes (Wong et al., 2001), providing a promising opportunity for precise base editing. Although previous approaches have engineered targeted ADAR fusions via RNA guides (Fukuda et al., 2017; Montiel-Gonzalez et al., 2013; Montiel-Gonzalez et al., 2016; Wettengel et al., 2017), the specificity of these approaches has not been reported and their respective targeting mechanisms rely on RNA-RNA hybridization without the assistance of protein partners that may enhance target recognition and stringency.

Here we assay a subset of the family of Cas13 enzymes for RNA knockdown activity in mammalian cells and identify the Cas13b ortholog from *Prevotella sp. P5-125* (PspCas13b) as the most efficient and specific for mammalian cell applications. We then fuse the ADAR2 deaminase domain (ADAR2$_{DD}$) to catalytically inactive PspCas13b and demonstrate RNA editing for programmable A to I (G) replacement (REPAIR) of reporter and endogenous transcripts as well as disease-relevant mutations. Lastly, we employ a rational mutagenesis scheme to improve the specificity of dCas13b-ADAR2$_{DD}$ fusions to generate REPAIRv2 with more than 919-fold higher specificity.

7.3 Results

7.3.1 Comprehensive Characterization of Cas13 Family Members in Mammalian Cells

We previously developed LwaCas13a for mammalian knockdown applications, but it required an monomeric superfolder GFP (msfGFP) stabilization domain for efficient knockdown and, although the specificity was high, knockdown levels were not consistently below 50% (Abudayyeh et al., 2017). We sought to identify a more robust RNA-targeting CRISPR system by characterizing a genetically diverse set of Cas13 family members to assess their RNA knockdown activity in mammalian cells (Fig. 7.1A). We generated mammalian codon-optimized versions of multiple Cas13 proteins, including 21 orthologs of Cas13a, 15 of Cas13b and 7 of Cas13c, and cloned them into an expression vector with N- and C-terminal nuclear export signal (NES) sequences and a C-terminal msfGFP to enhance protein stability. To assay interference in mammalian cells, we designed a dual reporter construct expressing the independent *Gaussia* (*Gluc*) and *Cypridinia* (*Cluc*) luciferases under separate promoters, which allows one luciferase to function as a measure of Cas13 interference activity and the other to serve as an internal control. For each Cas13 ortholog, we designed protospacer flanking site (PFS)-compatible guide RNAs, using the Cas13b PFS motifs derived from an ampicillin interference assay (fig 7.S1) and the 3' H (not G) PFS from previous reports of Cas13a activity (Abudayyeh et al., 2016).

We transfected HEK293FT cells with Cas13-expression, guide RNA, and reporter plasmids and then quantified levels of Cas13 expression and the targeted *Gluc* 48 hours later (Fig. 7.1B, fig. 7.S2A). Testing two guide RNAs for each Cas13 ortholog revealed a range of activity levels, including five Cas13b orthologs with similar or increased interference across both guide RNAs relative to the recently characterized LwaCas13a (Figure 7.1B), and we observed only a weak correlation between Cas13 expression and interference activity (fig. 7.S2B-D). We selected the top five Cas13b orthologs, as well as the top two Cas13a orthologs for further engineering.

We next tested Cas13-mediated knockdown of Gluc without msfGFP, to select orthologs that do not require stabilization domains for robust activity. We hypothesized that Cas13 activity could be affected by subcellular localization, as we previously reported for optimization of LwaCas13a (Abudayyeh et al., 2017). Therefore, we tested the interference activity of the seven selected Cas13 orthologs C-terminally fused to one of six different localization tags without msfGFP. Using the luciferase reporter assay, we identified the top three Cas13b designs with the highest level of interference activity: Cas13b from *Prevotella sp. P5-125* (PspCas13b) and Cas13b from *Porphyromonas gulae* (PguCas13b) C-terminally fused to the HIV Rev gene NES and Cas13b from *Riemerella anatipestifer* (RanCas13b) C-terminally fused to the MAPK NES (fig. 7.S3A). To further distinguish activity levels of the top orthologs, we compared the three optimized Cas13b constructs to the optimal LwaCas13a-msfGFP fusion and to shRNA for their ability to knockdown the endogenous *KRAS* transcript using position-matched guides (fig. 7.S3B). We observed the highest levels interference for PspCas13b (average knockdown 62.9%) and thus selected this for further comparison to LwaCas13a.

To more rigorously define the activity of PspCas13b and LwaCas13a, we designed position-matched guides tiling along both *Gluc* and *Cluc* transcripts and assayed their activity using our luciferase reporter assay. We tested 93 and 20 position-matched guides targeting *Gluc* and *Cluc*, respectively, and found that PspCas13b had consistently increased levels of knockdown relative to LwaCas13a (average of 92.3% for PspCas13b vs. 40.1% knockdown for LwaCas13a) (Fig. 7.1C,D).

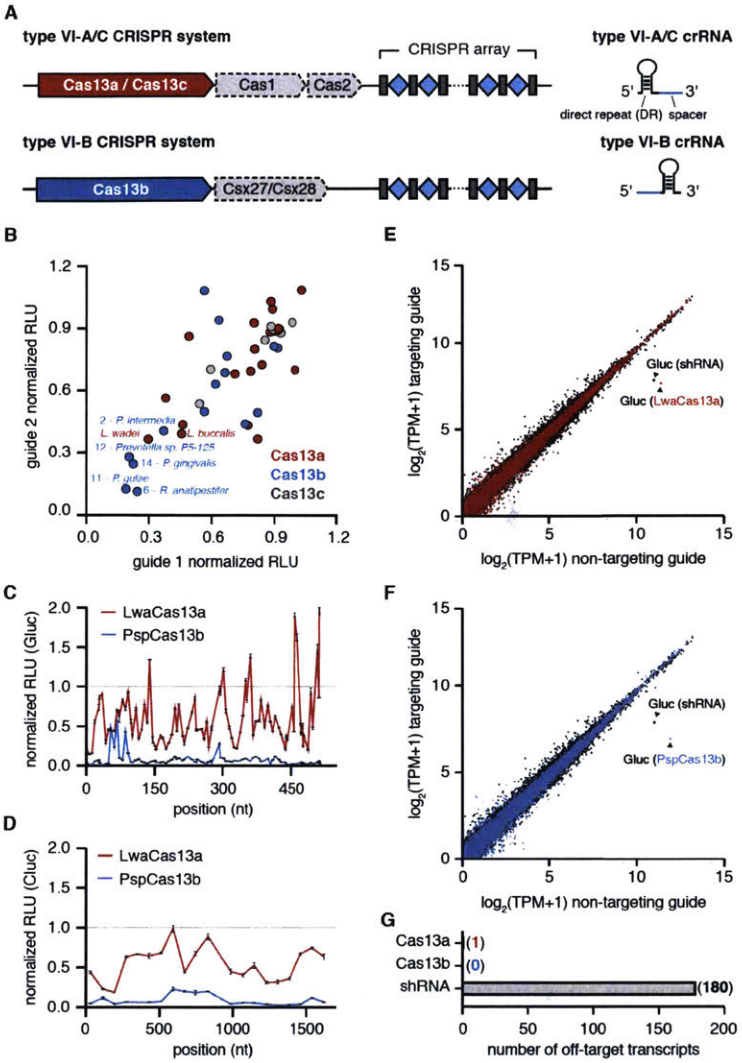


Figure 7.1: Characterization of a highly active Cas13b ortholog for RNA knockdown

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 LwaCas13a knockdown activity (as measured by luciferase activity) using tiling guides against *Gluc*. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig. 1B.

D)PspCas13b and LwaCas13a knockdown activity (as measured by luciferase activity) using tiling guides against *Cluc*. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig. 1B.

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 Fig1B.

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 Fig. 1B.

G)Number of significant off-targets from *Gluc* knockdown for LwaCas13a, PspCas13b, and shRNA from the transcriptome wide analysis in E and F.

7.3.2 Specificity of Cas13 mammalian interference activity

To characterize the interference specificities of PspCas13b and LwaCas13a we designed a plasmid library of luciferase targets containing single mismatches and double mismatches throughout the target sequence and the three flanking 5' and 3' base pairs (fig. 7.S3C). We transfected HEK293FT cells with either LwaCas13a or PspCas13b, a fixed guide RNA targeting the unmodified target sequence, and the mismatched target library corresponding to the appropriate system. We then performed targeted RNA sequencing of uncleaved transcripts to quantify depletion of mismatched target sequences. We found that LwaCas13a and PspCas13b had a central region that was relatively intolerant to single mismatches, extending from base pairs 12-26 for the PspCas13b target and 13-24 for the LwaCas13a target (fig. 7.S3D). Double mismatches were even less tolerated than single mutations, with little knockdown activity observed over a larger window, extending from base pairs 12-29 for PspCas13b and 8-27 for LwaCas13a in their respective targets (fig. 7.S3E). Additionally, because there are mismatches included in the three nucleotides flanking the 5' and 3' ends of the target sequence, we could assess PFS constraints on Cas13 knockdown activity. Sequencing showed that almost all PFS combinations allowed robust knockdown, indicating that a PFS constraint for interference in mammalian cells likely does not exist for either enzyme tested. These results indicate that Cas13a and Cas13b display similar sequence constraints and sensitivities against mismatches.

We next characterized the interference specificity of PspCas13b and LwaCas13a across the mRNA fraction of the transcriptome. We performed transcriptome-wide mRNA sequencing to detect significant differentially expressed genes. LwaCas13a and PspCas13b demonstrated robust knockdown of *Gluc* (Fig. 7.1E,F) and were highly specific compared to a position-matched shRNA, which showed hundreds of off-targets (Fig. 7.1G), consistent with our previous characterization of LwaCas13a specificity in mammalian cells (Abudayyeh et al., 2017).

7.3.3 Cas13-ADAR fusions enable targeted RNA editing

Given that PspCas13b achieved consistent, robust, and specific knockdown of mRNA in mammalian cells, we envisioned that it could be adapted as an RNA binding platform to recruit RNA modifying domains, such as the deaminase domain of ADARs (ADAR_{DD}) for programmable RNA editing. To engineer a PspCas13b lacking nuclease activity (dPspCas13b, referred to as dCas13b hereafter), we mutated conserved catalytic residues in the HEPN domains and observed loss of luciferase RNA knockdown (fig. 7.S4A). We hypothesized that a dCas13b-ADAR_{DD} fusion could be recruited by a guide RNA to target adenosines, with the hybridized RNA creating the required duplex substrate for ADAR activity (Fig. 7.2A). To enhance target adenosine deamination rates we introduced two additional modifications to our initial RNA editing design: we introduced a mismatched cytidine opposite the target adenosine, which has been previously reported to increase deamination frequency, and fused dCas13b with the deaminase domains of human ADAR1 or ADAR2 containing hyperactivating mutations to enhance catalytic activity (ADAR1_{DD}(E1008Q) (Wang et al., 2015) or ADAR2_{DD}(E488Q) (Kuttan and Bass, 2012)).

To test the activity of dCas13b-ADAR_{DD} we generated an RNA-editing reporter on *Cluc* by introducing a nonsense mutation (W85X (UGG->UAG)), which could functionally be repaired to the wildtype codon through A->I editing (Fig. 7.2B) and then be detected as restoration of Cluc luminescence. We evenly tiled guides with spacers of 30, 50, 70 or 84 nucleotides in length across the target adenosine to determine the optimal guide placement and design (Fig. 7.2C). We found that dCas13b-ADAR1$_{DD}$ required longer guides to repair the *Cluc* reporter, while dCas13b-ADAR2$_{DD}$ was functional with all guide lengths tested (Fig. 7.2C). We also found that the hyperactive E488Q mutation improved editing efficiency, as luciferase restoration with the wildtype ADAR2$_{DD}$ was reduced (fig. 7.S4B). From this demonstration of activity, we chose dCas13b-ADAR2$_{DD}$(E488Q) for further characterization and designated this approach as RNA Editing for Programmable A to I Replacement version 1 (REPAIRv1).

To validate that restoration of luciferase activity was due to *bona fide* editing events, we directly measured REPAIRv1-mediated editing of *Cluc* transcripts via reverse transcription and targeted next-generation sequencing. We tested 30- and 50-nt spacers around the target site and found that both guide lengths resulted in the expected A to I edit, with 50-nt spacers achieving higher editing percentages (Fig. 7.2D,E, fig. 7.S4C). We also observed that 50-nt spacers had an increased propensity for editing at non-targeted adenosines within the sequencing window, likely due to increased regions of duplex RNA (Fig. 7.2E, fig. 7.S4C).

We next targeted an endogenous gene, *PPIB*. We designed 50-nt spacers tiling *PPIB* and found that we could edit the *PPIB* transcript with up to 28% editing efficiency (Fig. 7.S4D). To test if REPAIR could be further optimized, we modified the linker between dCas13b and ADAR2$_{DD}$(E488Q) (fig. 7.S4E) and found that linker choice modestly affected luciferase activity restoration. Additionally, we tested the ability of dCas13b and guide alone to mediate editing events, finding that the ADAR deaminase domain is required for editing (fig. 7.S5A-D).

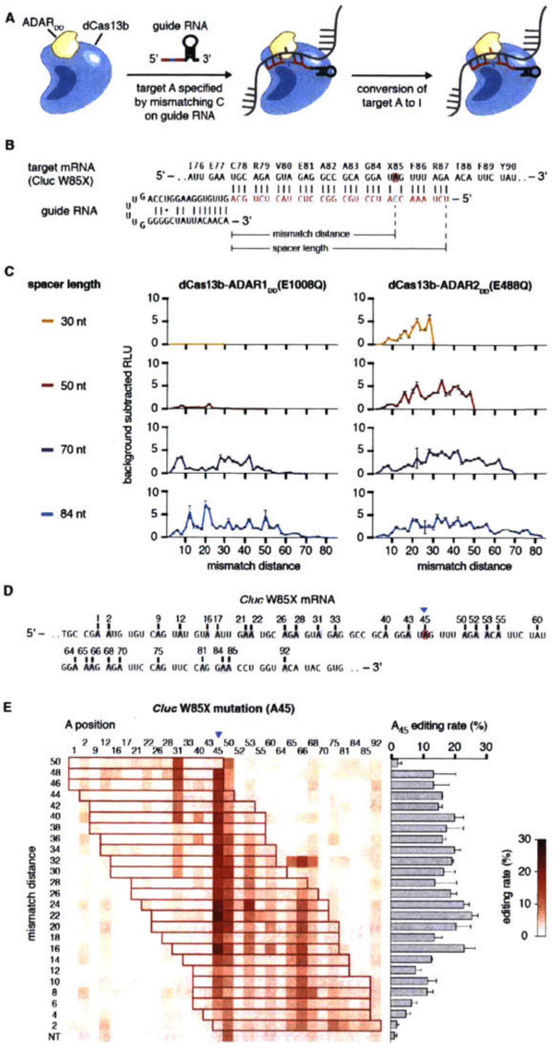


Figure 7.2: Engineering dCas13b-ADAR fusions for RNA editing

A)Schematic of RNA editing by dCas13b-ADAR$_{DD}$ fusion proteins. Catalytically dead Cas13b (dCas13b) is fused to the deaminase domain of human ADAR (ADAR$_{DD}$), which naturally deaminates adenosines to insosines in dsRNA. The crRNA specifies the target site by hybridizing to the bases surrounding the target adenosine, creating a dsRNA structure for editing, and recruiting the dCas13b-ADAR$_{DD}$ fusion. A mismatched cytidine in the crRNA opposite the target adenosine enhances the editing reaction, promoting target adenosine deamination to inosine, a base that functionally mimics guanosine in many cellular reactions. B)Schematic of *Cypridina* luciferase W85X target and targeting guide design. Deamination of the target adenosine restores the stop codon to the wildtype tryptophan. Spacer length is the region of the guide that contains homology to the target sequence. Mismatch distance is the number of bases between the 3' end of the spacer and the mismatched cytidine. The cytidine mismatched base is included as part of the mismatch distance calculation.

C)Quantification of luciferase activity restoration for dCas13b-ADAR1$_{DD}$(E1008Q) (left) and dCas13b-ADAR2$_{DD}$(E488Q) (right) with tiling guides of length 30, 50, 70, or 84 nt. All guides with even mismatch distances are tested for each guide length. Values are background subtracted relative to a 30nt non-targeting guide that is randomized with no sequence homology to the human transcriptome.

D)Schematic of the sequencing window in which A to I edits were assessed for *Cypridinia* luciferase W85X.

E)Sequencing quantification of A to I editing for 50-nt guides targeting *Cypridinia* luciferase W85X. Blue triangle indicates the targeted adenosine. For each guide, the region of duplex RNA is outlined in red. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig. 2C.

7.3.4 Defining the sequence parameters for RNA editing

Given that we could achieve precise RNA editing at a test site, we wanted to characterize the sequence constraints for programming the system against any RNA target in the transcriptome. Sequence constraints could arise from dCas13b targeting limitations, such as the PFS, or from ADAR sequence preferences (Lehmann and Bass, 2000). To investigate PFS constraints on REPAIRv1, we designed a plasmid library carrying a series of four randomized nucleotides at the 5' end of a target site on the *Cluc* transcript (Fig. 7.3A). We targeted the center adenosine within either a UAG or AAC motif and found that for both motifs, all PFSs demonstrated detectable levels of RNA editing, with a majority of the PFSs having greater than 50% editing at the target site (Fig. 7.3B). Next, we sought to determine if the ADAR2$_{DD}$ in REPAIRv1 had any sequence constraints immediately flanking the targeted base, as has been reported previously for ADAR2$_{DD}$ (Lehmann and Bass, 2000). We tested every possible combination of 5' and 3' flanking nucleotides directly surrounding the target adenosine (Fig. 7.3C), and found that REPAIRv1 was capable of editing all motifs (Fig. 7.3D). Lastly, we analyzed whether the identity of the base opposite the target A in the spacer sequence affected editing efficiency and found that an A-C mismatch had the highest luciferase restoration, in agreement with previous reports of ADAR2 activity, with A-G, A-U, and A-A having drastically reduced REPAIRv1 activity (fig. 7.S5E).

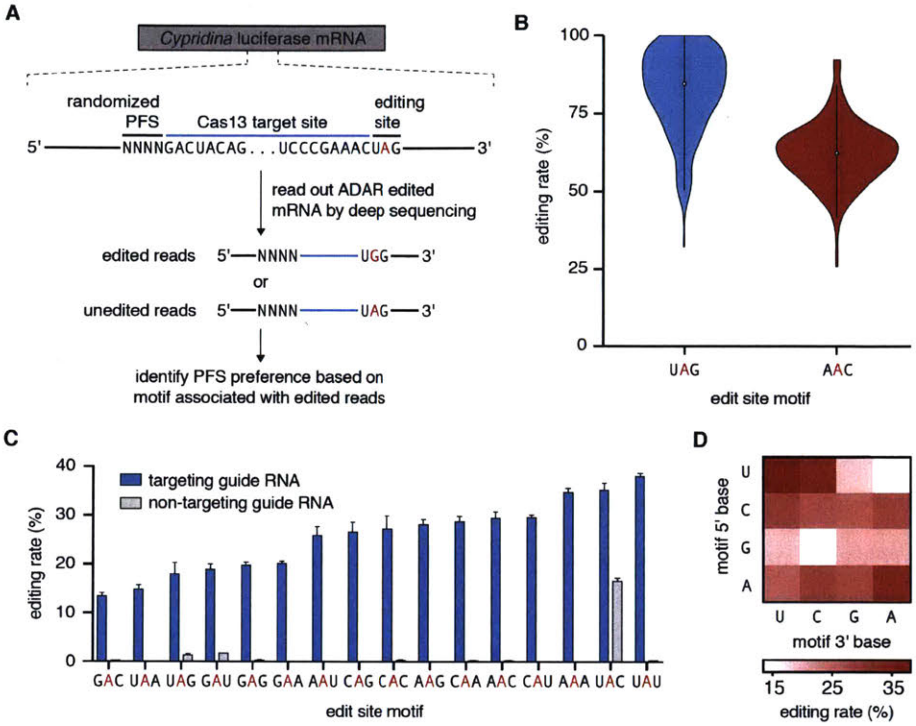


Figure 7.3: Measuring sequence flexibility for RNA editing by REPAIRv1

A)Schematic of screen for determining Protospacer Flanking Site (PFS) preferences of RNA editing by REPAIRv1. A randomized PFS sequence is cloned 5' to a target site for REPAIR editing. Following exposure to REPAIR, deep sequencing of reverse transcribed RNA from the target site and PFS is used to associate edited reads with PFS sequences.

B)Distributions of RNA editing efficiencies for all 4-N PFS combinations at two different editing sites

C)Quantification of the percent editing of REPAIRv1 at Cluc W85 across all possible 3 base motifs. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig. 2C.

D)Heatmap of 5' and 3' base preferences of RNA editing at Cluc W85 for all possible 3 base motifs.

7.3.5 Correction of disease-relevant human mutations using REPAIRv1

To demonstrate the broad applicability of the REPAIRv1 system for RNA editing in mammalian cells, we designed REPAIRv1 guides against two disease relevant mutations: 878G>A (*AVPR2* W293X) in X-linked Nephrogenic diabetes insipidus and 1517G>A (*FANCC* W506X) in Fanconi anemia. We transfected expression constructs for cDNA of genes carrying these mutations into HEK293FT cells and tested whether REPAIRv1 could correct the mutations. Using guide RNAs containing 50-nt spacers, we were able to achieve 35% correction of *AVPR2* and 23% correction of *FANCC* (Fig. 7.4A-D). We then tested the ability of REPAIRv1 to correct 34 different disease-relevant G>A mutations and found that we were able to achieve significant editing at 33 sites with up to 28% editing efficiency (Fig. 7.4E). The mutations we chose are only a fraction of the pathogenic G to A mutations (5,739) in the ClinVar database, which also includes an additional 11,943 G to A variants (Fig. 7.4F and fig. 7.S6). Because there are no sequence constraints (Fig. 7.3), REPAIRv1 is capable of potentially editing all these disease relevant mutations, especially given that we observed editing regardless of the target motif (Fig. 7.3C and Fig. 7.4G).

Delivering the REPAIRv1 system to diseased cells is a prerequisite for therapeutic use, and we therefore sought to design REPAIRv1 constructs that could be packaged into therapeutically relevant viral vectors, such as adeno-associated viral (AAV) vectors. AAV vectors have a packaging limit of 4.7kb, which cannot accommodate the large size of dCas13b-ADAR$_{DD}$ (4,473 bp) along with promoter and expression regulatory elements. To reduce the size, we tested a variety of N-terminal and C-terminal truncations of dCas13 fused to ADAR2$_{DD}$(E488Q) for RNA editing activity. We found that all C-terminal truncations tested were still functional and able to restore luciferase signal (fig. 7.S7), and the largest truncation, C-terminal Δ984-1090 (total size of the fusion protein 4,152bp) was small enough to fit within the packaging limit of AAV vectors.

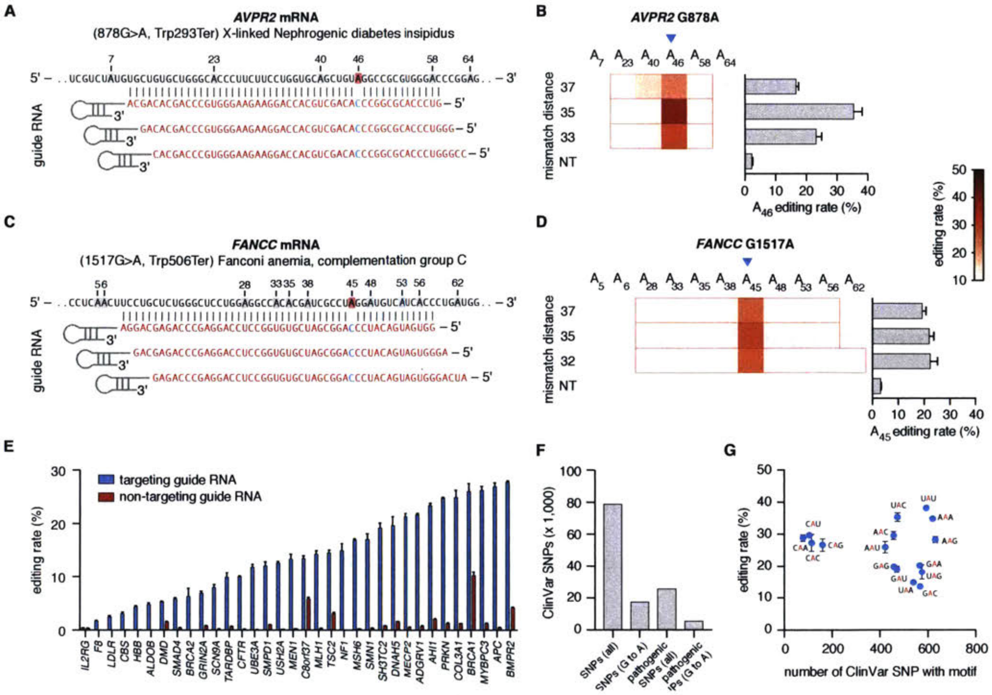


Figure 7.4: Correction of disease-relevant mutations with REPAIRv1

A)Schematic of target and guide design for targeting *AVPR2* 878G>A.

B) The 878G>A mutation (indicated by blue triangle) in *AVPR2* is corrected to varying levels using REPAIRv1 with three different guide designs. For each guide, the region of duplex RNA is outlined in red. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig. 2C.

C)Schematic of target and guide design for targeting *FANCC* 1517G>A.

D) The 1517G>A mutation (indicated by blue triangle) in *FANCC* is corrected to varying levels using REPAIRv1 with three different guide designs. For each guide, the region of duplex RNA is outlined in red. The heatmap scale bar is the same as in panel B. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig. 2C.

E)Quantification of the percent editing of 34 different disease-relevant G>A mutations selected from ClinVar using REPAIRv1. Non-targeting guide is the same as in Fig. 2C.

F)Analysis of all the possible G>A mutations that could be corrected using REPAIR as annotated in the ClinVar database.

G)The distribution of editing motifs for all G>A mutations in ClinVar is shown versus the editing efficiency by REPAIRv1 per motif as quantified on the *Gluc* transcript. Values represent mean +/– S.E.M.

7.3.6 Transcriptome-wide specificity of REPAIRv1

Although RNA knockdown with PspCas13b was highly specific in our luciferase tiling experiments, we observed off-target adenosine editing within the guide:target duplex (Fig. 7.2E). To see if this was a widespread phenomenon, we tiled an endogenous transcript, *KRAS*, and measured the degree of off-target editing near the target adenosine (Fig. 7.5A). We found that for *KRAS*, while the on-target editing rate was 23%, there were many sites around the target site that also had detectable A to I edits (Fig. 7.5B).

Because of the observed off-target editing within the guide:target duplex, we initially evaluated transcriptome-wide off-targets by performing RNA sequencing on all mRNAs with 12.5X coverage. Of all the editing sites across the transcriptome, the on-target editing site had the highest editing rate, with 89% A to I conversion. We also found that there was a substantial number of A to I off-target events, with 1,732 off-targets in the targeting guide condition and 925 off-targets in the non-targeting guide condition, with 828 off-targets shared between the targeting and non-targeting guide conditions (Fig. 7.5C,D). Given the high number of overlapping off-targets between the targeting and non-targeting guide conditions, we reasoned that the off-targets may arise from ADAR_{DD}. To test this hypothesis, we repeated the Cluc targeting experiment, this time comparing transcriptome changes for REPAIRv1 with a targeting guide, REPAIRv1 with a non-targeting guide, REPAIRv1 alone, or ADAR_{DD}(E488Q) alone (fig. 7.S8). We found differentially expressed genes and off-target editing events in each condition (fig. 7.S8B,C). Interestingly, there was a high degree of overlap in the off-target editing events between ADAR_{DD}(E488Q) and all REPAIRv1 off-target edits, supporting the hypothesis that REPAIR off-target edits are driven by dCas13b-independent ADAR_{DD}(E488Q) editing events (fig. 7.S8D).

Next, we sought to compare two RNA-guided ADAR systems that have been described previously (fig. 7.S9A). The first utilizes a fusion of ADAR2$_{DD}$ to the small viral protein lambda N (λN), which binds to the BoxB-λ RNA hairpin (Montiel-Gonzalez et al., 2013). A guide RNA with double BoxB-λ hairpins guides ADAR2$_{DD}$ to edit sites encoded in the guide RNA (Montiel-Gonzalez et al., 2016). The second design utilizes full-length ADAR2 (ADAR2) and a guide RNA with a hairpin that the double strand RNA binding domains (dsRBDs) of ADAR2 recognize (Fukuda et al., 2017; Wettengel et al., 2017). We analyzed the editing efficiency of these two systems compared to REPAIRv1 and found that the BoxB-ADAR2 and full-length ADAR2 systems demonstrated 50% and 34.5% editing rates, respectively, compared to the 89% editing rate achieved by REPAIRv1 (fig. 7.S9B-E). Additionally, the BoxB and full-length ADAR2 systems created 1,814 and 66 observed off targets, respectively, in the targeting guide conditions, compared to the 2,111 off targets in the REPAIRv1 targeting guide condition. Notably, all the conditions with the two ADAR2$_{DD}$-based systems (REPAIRv1 and BoxB) showed a high percentage of overlap in their off-targets whereas the full-length ADAR2 system had a largely distinct set of off-targets (fig. 7.S9F). The overlap in off-targets between the targeting and non-targeting conditions and between REPAIRv1 and BoxB conditions suggests ADAR2$_{DD}$ drives off-targets independent of dCas13 targeting (fig. 7.S9F).

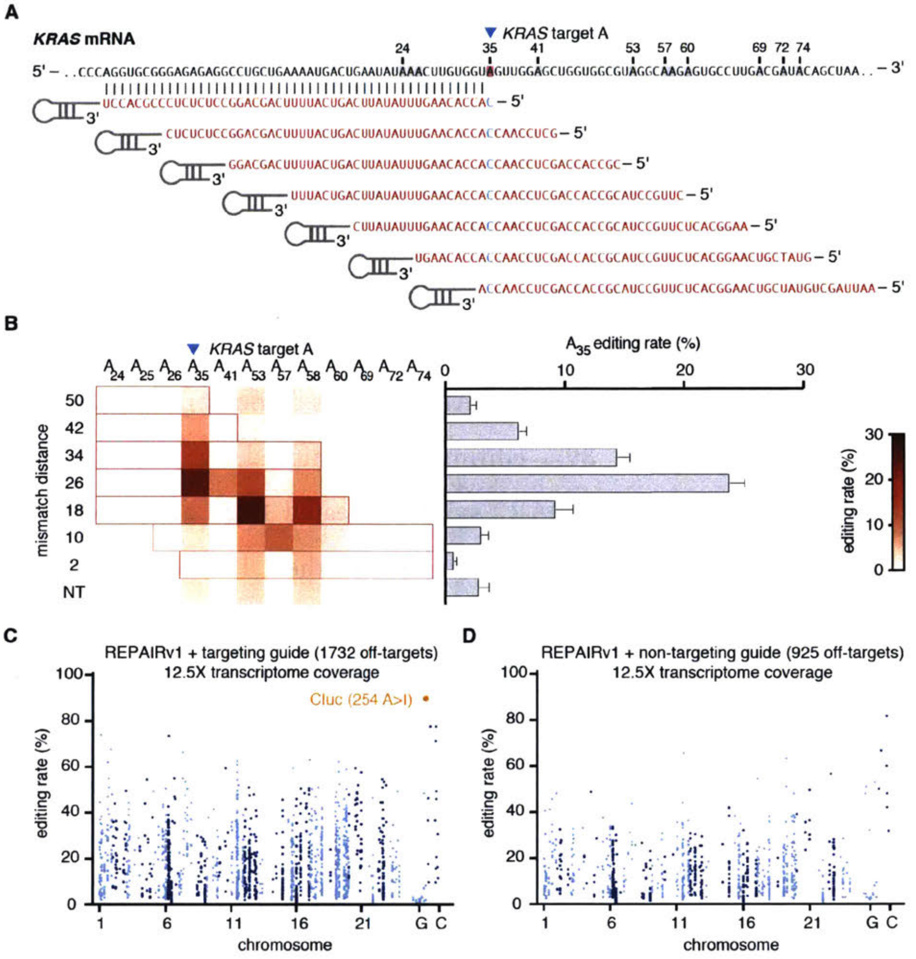


Figure 7.5: Characterizing specificity of REPAIRv1

A)Schematic of *KRAS* target site and guide design.

B)Quantification of percent A to I editing for tiled *KRAS*-targeting guides. Editing percentages are shown for the on-target (blue triangle) and neighboring adenosine sites. For each guide, the region of duplex RNA is outlined in red. Values represent mean +/– S.E.M.

C)Transcriptome-wide sites of significant RNA editing by REPAIRv1 (150ng REPAIR vector transfected) with *Cluc* targeting guide. The on-target site *Cluc* site (254 A>I) is highlighted in orange.

D)Transcriptome-wide sites of significant RNA editing by REPAIRv1 (150ng REPAIR vector transfected) with non-targeting guide. Non-targeting guide is the same as in Fig. 2C.

7.3.7 Improving specificity of REPAIR through rational protein engineering

To improve the specificity of REPAIRv1, we employed structure-guided protein engineering of ADAR2$_{DD}$(E488Q). Because of the guide-independent nature of the off-targets, we hypothesized that destabilizing ADAR2$_{DD}$(E488Q)-RNA binding would selectively decrease off-target editing, but maintain on-target editing due to increased local concentration from dCas13b tethering of ADAR2$_{DD}$(E488Q) to the target site. We mutated residues in ADAR2$_{DD}$(E488Q) previously determined to contact the duplex region of the target RNA (Fig. 7.6A) (Matthews et al., 2016). To assess efficiency and specificity, we tested 17 single mutants with both targeting and non-targeting guides, under the assumption that background luciferase restoration in the non-targeting condition would be indicative of broader off-target activity. We found that mutations at the selected residues had significant effects on the luciferase activity for targeting and non-targeting guides (Fig. 7.6A,B, fig. 7.S10A). A majority of mutants either significantly improved the luciferase activity for the targeting guide or increased the ratio of targeting to non-targeting guide activity, which we termed the specificity score (Fig. 7.6A,B).

We selected a subset of these mutants (Fig. 7.6B) for transcriptome-wide specificity profiling by next generation sequencing. As expected, off-targets measured from transcriptome-wide sequencing correlated with our specificity score (fig. 7.S10B) for mutants. We found that with the exception of ADAR2$_{DD}$(E488Q/R455E), all sequenced REPAIRv1 mutants could effectively edit the reporter transcript (Fig. 7.6C), with many mutants showing reduction in the number of off-targets (Fig. 7.6C, fig 7.S10C, 7.S11). We further explored the surrounding motifs of off-targets for the various specificity mutants, and found that REPAIRv1 and most of the engineered variants exhibited a strong 3' G preference for their off-target edits, in agreement with the characterized ADAR2 motif (fig. 7.S12A) (Lehmann and Bass, 2000).

We focused on the mutant ADAR2$_{DD}$(E488Q/T375G), as it had the highest percent editing of the four mutants with the lowest numbers of transcriptome-wide off targets and termed it REPAIRv2. Compared to REPAIRv1, REPAIRv2 exhibited increased specificity, with a reduction from 18,385 to 20 transcriptome-wide off-targets by high-coverage sequencing (125X coverage, 10ng DNA transfection) (Fig. 7.6D). In the region surrounding the targeted adenosine in *Cluc*, REPAIRv2 also had reduced off-target editing, visible in sequencing traces (Fig. 7.6E). In motifs derived from the off-target sites, REPAIRv1 presented a strong preference towards 3' G, but showed off-targeting edits for all motifs (fig. 7.S12B); by contrast, REPAIRv2 only edited the strongest off-target motifs (fig. 7.S12C). The distribution of edits on transcripts was heavily skewed for REPAIRv1, with highly-edited genes having over 60 edits (fig. 7.S13A), whereas REPAIRv2 only edited one transcript (*EEF1A1*) multiple times (fig. 7.S13B). REPAIRv1 off-target edits were predicted to result in numerous variants, including 1000 missense base changes (fig. 7.S13C) with 93 events in genes related to cancer processes (fig. 7.S13D). In contrast, REPAIRv2 only had 6 predicted base changes (fig. 7.S10E), none of which were in cancer-related genes (fig. 7.S13F). Analysis of the sequence surrounding off-target edits for REPAIRv1 or v2 did not reveal homology to guide sequences, suggesting that off-targets are likely dCas13b-independent (fig. 7.S14), consistent with the high overlap of off-targets between REPAIRv1 and the ADAR deaminase domain (fig. 7.S8D). To directly compare REPAIRv2 against other programmable ADAR systems, we repeated our *Cluc* targeting experiments with all systems at two different dosages of ADAR vector, finding that REPAIRv2 had comparable on-target editing to BoxB and ADAR2 but with significantly fewer off-target editing events at both dosages (fig. 7.S15). REPAIRv2 had enhanced specificity compared to REPAIRv1 at both dosages (fig. 7.S15B), a finding that also extended to two guides targeting distinct sites on *PPIB* (fig. 7.S16A-D). It is also worth noting that, in general, the lower dosage condition (10 ng) had fewer off-targets than the higher dosage condition (150 ng) (fig. 7.S5).

To assess editing specificity with greater sensitivity, we sequenced the low dosage condition (10 ng of transfected DNA) of REPAIRv1 and v2 at significantly higher sequencing depth (125X coverage of the transcriptome). Increased numbers of off-targets were found at higher sequencing depths corresponding to detection of rarer off-target events (fig. 7.S17). Furthermore, we speculated that different transcriptome states could also potentially alter the number of off-targeting events. Therefore, we tested REPAIRv2 activity in the osteosarcoma U2OS cell line, observing 6 and 7 off-targets for the targeting and non-targeting guide, respectively (fig. 7.S18).

We targeted REPAIRv2 to endogenous genes to test if the specificity-enhancing mutations reduced nearby edits in target transcripts while maintaining high-efficiency on-target editing. For guides targeting either *KRAS* or *PPIB*, we found that REPAIRv2 had no detectable off-target edits, unlike

REPAIRv1, and could effectively edit the on-target adenosine at efficiencies of 27.1% (*KRAS*) or 13% (*PPIB*) (Fig. 7.6F). This specificity extended to additional target sites, including regions that demonstrate high-levels of background in non-targeting conditions for REPAIRv1, such as other *KRAS* or *PPIB* target sites (fig. 7.S19). Overall, REPAIRv2 eliminated off-targets in duplexed regions around the edited adenosine and showed dramatically enhanced transcriptome-wide specificity.

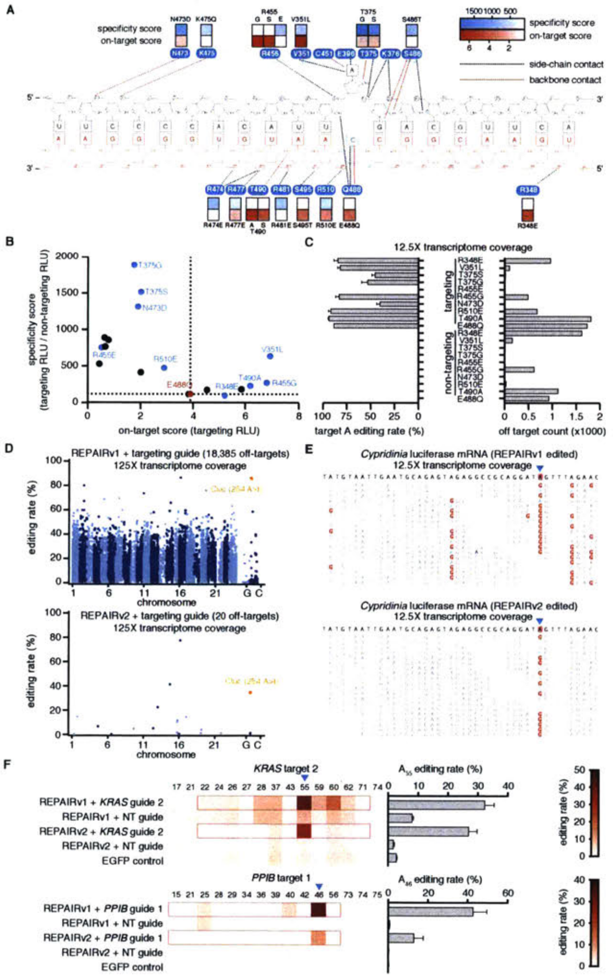


Figure 7.6: Rational mutagenesis of ADAR2 to improve the specificity of REPAIRv1

A)Quantification of luciferase signal restoration (on-target score, red boxes) by various dCas13-ADAR2$_{DD}$ mutants as well as their specificity score (blue boxes) plotted along a schematic of the contacts between key ADAR2 deaminase residues and the dsRNA target (target strand shown in gray; the non-target strand is shown in red). All deaminase mutations were made on the dCas13-ADAR2$_{DD}$(E488Q) background. The specificity score is defined as the ratio of the luciferase signal between targeting guide and non-targeting guide conditions. Schematic of ADAR2 deaminase domain contacts with dsRNA is adapted from ref (Zheng et al., 2017).

B)Quantification of luciferase signal restoration by various dCas13-ADAR2 mutants versus their specificity score. Non-targeting guide is the same as in Fig. 2C.

C)Quantification of on-target editing and the number of significant off-targets for each dCas13-ADAR2$_{DD}$(E488Q) mutant by transcriptome wide sequencing of mRNAs. Values represent mean +/- S.E.M. Non-targeting guide is the same as in Fig. 2C.

D)Transcriptome-wide sites of significant RNA editing by REPAIRv1 (top) and REPAIRv2 (bottom) with a guide targeting a pretermination site in *Cluc*. The on-target *Cluc* site (254 A>I) is highlighted in orange. 10 ng of REPAIR vector was transfected for each condition.

E)Representative RNA sequencing reads surrounding the on-target *Cluc* editing site (254 A>I; blue triangle) highlighting the differences in off-target editing between REPAIRv1 (top) and REPAIRv2 (bottom). A>I edits are highlighted in red; sequencing errors are highlighted in blue. Gaps reflect spaces between aligned reads. Non-targeting guide is the same as in Fig. 2C.

F)RNA editing by REPAIRv1 and REPAIRv2 with guides targeting an out-of-frame UAG site in the endogenous *KRAS* and *PPIB* transcripts. The on-target editing fraction is shown as a sideways bar chart on the right for each condition row. For each guide, the region of duplex RNA is outlined in red. Values represent mean +/- S.E.M. Non-targeting guide is the same as in Fig. 2C.

7.4 Discussion

We show here that the RNA-guided RNA-targeting type VI-B CRISPR effector Cas13b is capable of highly efficient and specific RNA knockdown, providing the basis for improved tools for interrogating essential genes and non-coding RNA as well as controlling cellular processes at the transcript level. Catalytically inactive Cas13b (dCas13b) retains programmable RNA binding capability, which we leveraged here by fusing dCas13b to the adenosine deaminase domain of ADAR2 to achieve precise A to I edits, a system we term REPAIRv1 (RNA Editing for Programmable A to I Replacement version 1). Further engineering of the system produced REPAIRv2, which has dramatically higher specificity than previously described RNA editing platforms (Montiel-Gonzalez et al., 2016; Stafforst and Schneider, 2012) while maintaining high levels of on-target efficacy.

Although Cas13b exhibits high fidelity, our initial results with dCas13b-ADAR2$_{DD}$(E488Q) fusions revealed a substantial number of off-target RNA editing events. To address this, we employed a rational mutagenesis strategy to vary the ADAR2$_{DD}$ residues that contact the RNA duplex, identifying a variant, ADAR2$_{DD}$(E488Q/T375G), capable of precise, efficient, and highly specific editing when fused to dCas13b. Editing efficiency with this variant was comparable to or better than that achieved with two currently available systems, BoxB-ADAR2$_{DD}$(E488Q) or ADAR2 editing. Moreover, the REPAIRv2 system created only 20 observable off-targets in the whole transcriptome, at least an order of magnitude better than both alternative editing technologies. While it is possible that ADAR could deaminate adenosine bases on the DNA strand in RNA-DNA heteroduplexes (Zheng et al., 2017), it is unlikely to do so in this case as Cas13b does not bind DNA efficiently and that REPAIR is cytoplasmically localized. Additionally, the lack of homology of off-target sites to the guide sequence and the strong overlap of off-targets with the ADAR$_{DD}$(E488Q)-only condition suggest that off-targets are not mediated by off-target guide binding. Deeper sequencing and novel inosine enrichment methods could further refine our understanding of REPAIR specificity in the future.

The REPAIR system offers many advantages compared to other nucleic acid editing tools. First, the exact target site can be encoded in the guide by placing a cytidine within the guide extension across from the desired adenosine to create a favorable A-C mismatch ideal for ADAR editing activity. Second, Cas13 has no targeting sequence constraints, such as a PFS or PAM, and no motif preference surrounding the target adenosine, allowing any adenosine in the transcriptome to be potentially targeted with the REPAIR system. The lack of motif for ADAR editing, in contrast with previous literature, is likely due to the increased local concentration of REPAIR at the target site due to dCas13b binding. We do note that DNA base editors can target either the sense or anti-sense strand, while the REPAIR system is limited to transcribed sequences, thereby constraining the total number of possible editing sites. However, due to the less constrained nature of targeting with REPAIR, this system can affect more edits within ClinVar (Fig. 7.4C) than Cas9-DNA base editors. Third, the REPAIR system directly deaminates target adenosines to inosines and does not rely on endogenous repair pathways, such as base-excision or mismatch repair, to generate desired editing outcomes.

Therefore, REPAIR should be able to mediate efficient RNA editing even in post-mitotic cells such as neurons. Fourth, in contrast to DNA editing, RNA editing is transient and can be more easily reversed, allowing the potential for temporal control over editing outcomes. The temporary nature of REPAIR-mediated edits will likely be useful for treating diseases caused by temporary changes in cell state, such as local inflammation and could also be used to treat disease by modifying the function of proteins involved in disease-related signal transduction. For instance, REPAIR editing would allow the re-coding of some serine, threonine and tyrosine residues that are the targets of kinases. Phosphorylation of these residues in disease-relevant proteins affects disease progression for many disorders including Alzheimer's disease and multiple neurodegenerative conditions (Ballatore et al., 2007). REPAIR might also be used to transiently or even chronically change the sequence of expressed, risk-modifying G to A variants to decrease the chance of entering a disease state for patients. For instance, REPAIR could be used to functionally mimic A to G alleles of *IFIH1* that protect against autoimmune disorders such as type I diabetes, immunoglobulin A deficiency, psoriasis, and systemic lupus erythematosus (Ferreira et al., 2010; Li et al., 2010).

The REPAIR system provides multiple opportunities for additional engineering. Cas13b possesses pre-crRNA processing activity (Smargon et al., 2017b), allowing for multiplex editing of multiple variants, any one of which alone may not affect disease, but together might have additive effects and disease-modifying potential. Extension of our rational design approach, such as combining promising mutations and directed evolution, could further increase the specificity and efficiency of the system, while unbiased screening approaches could identify additional residues for improving REPAIR activity and specificity.

Currently, the base conversions achievable by REPAIR are limited to generating inosine from adenosine; additional fusions of dCas13 with other catalytic RNA editing domains, such as APOBEC, could enable cytidine to uridine editing. Additionally, mutagenesis of ADAR could relax the substrate preference to target cytidine, allowing for the enhanced specificity conferred by the duplexed RNA substrate requirement to be exploited by C to U editors. Adenosine to inosine editing on DNA substrates may also be possible with catalytically inactive DNA-targeting CRISPR effectors, such as dCas9 or dCpf1, either through formation of DNA-RNA heteroduplex targets (Zheng et al., 2017) or mutagenesis of the ADAR domain.

We have demonstrated the use of the PspCas13b enzyme as both an RNA knockdown and RNA editing tool. The dCas13b platform for programmable RNA binding has many applications, including live transcript imaging, splicing modification, targeted localization of transcripts, pull down of RNA-binding proteins, and epitranscriptomic modifications. Here, we used dCas13 to create REPAIR, adding to the existing suite of nucleic acid editing technologies. REPAIR provides a new approach for treating genetic disease or mimicking protective alleles, and establishes RNA editing as a useful tool for modifying genetic function.

7.5 Experimental Procedures

7.5.1 Design and cloning of bacterial constructs

Mammalian codon optimized Cas13b constructs were cloned into the chloramphenicol resistant pACYC184 vector under control of the Lac promoter. Two corresponding direct-repeat (DR) sequences separated by BsaI restriction sites were then inserted downstream of Cas13b, under control of the pJ23119 promoter. Last, oligos for targeting spacers were phosphorylated using T4 PNK (New England Biolabs), annealed and ligated into BsaI digested vectors using T7 ligase (Enzymatics) to generate targeting Cas13b vectors.

7.5.2 Bacterial PFS screens

Ampicillin resistance plasmids for PFS screens were cloned by inserting PCR products containing Cas13b targets with two 5' randomized nucleotides and four 3' randomized nucleotides separated by a target site immediately downstream of the start codon of the ampicillin resistance gene *bla* using NEB Gibson Assembly (New England Biolabs). 100 ng of ampicillin-resistant target plasmids were then electroporated with 65-100 ng chloramphenicol-resistant Cas13b bacterial targeting plasmids into Endura Electrocompetent Cells (Lucigen). Plasmids were added to cells, incubated for 15 minutes on ice, electroporated using the manufacturer's recommended settings, and then 950 uL of recovery media was added to cells before a one-hour outgrowth at 37° C. The outgrowth was plated onto chloramphenicol and ampicillin double selection plates. Serial dilutions of the outgrowth were used to estimate the cfu/ng DNA. 16 hours post plating, cells were scraped off plates and surviving plasmid DNA was harvested using the Qiagen Plasmid Plus Maxi Kit (Qiagen). Surviving Cas13b target sequences and their flanking regions were amplified by PCR and sequenced using an Illumina NextSeq. To assess PFS preferences, the positions containing randomized nucleotides in the original library were extracted, and sequences depleted relative to the vector only condition and that were present in both bioreplicates were extracted using custom python scripts. The -$\log_2$ of the ratio of PFS abundance in the Cas13b condition compared to the vector only control was then used to calculate preferred motifs. Specifically, all sequences having -$\log_2$(sample/vector) depletion ratios above a specific threshold were used to generate weblogos of sequence motifs (weblogo.berkeley.edu).

7.5.3 Design and cloning of mammalian constructs for RNA interference

To generate vectors for testing Cas13 orthologs in mammalian cells, mammalian codon optimized Cas13a, Cas13b, and Cas13c genes were PCR amplified and golden-gate cloned into a mammalian expression vector containing dual NLS sequences and a C-terminal msfGFP, under control of the EF1alpha promoter. For further optimization Cas13 orthologs were golden-gate cloned into destination vectors containing different C-terminal localization tags under control of the EF1alpha promoter.

The dual luciferase reporter was cloned by PCR amplifying *Gaussia* and *Cypridinia* luciferase coding DNA, the EF1alpha and CMV promoters and assembled using the NEB Gibson Assembly (New England Biolabs).

For expression of mammalian guide RNAs for Cas13a, Cas13b, or Cas13c orthologs, the corresponding direct repeat sequences were synthesized with golden-gate acceptor sites and cloned under U6 expression via restriction digest cloning. Individual guides were then cloned into the corresponding expression backbones for each ortholog by golden-gate cloning.

7.5.4 Measurement of Cas13 expression in mammalian cells

Dual-NLS Cas13-msfGFP constructs were transfected into HEK293FT cells with targeting and non-targeting guides. GFP fluorescence was measured 48 hours post transfection in the non-targeting guide condition using a plate reader.

7.5.5 Cloning of pooled mismatch libraries for Cas13 interference specificity

Pooled mismatch library target sites were created by PCR using a forward primer containing the semi-degenerate target sequences and a constant reverse primer off of a *Gluc* template. The semi-degenerate forward oligo had at each position of the Cas13 target, plus the 5' and 3' three flanking bases, a nucleotide mixture containing 94% of the correct base and 2% of each incorrect base. The mismatch library amplicon was then cloned into the dual luciferase reporter in place of wild-type *Gluc* using NEB Gibson assembly (New England Biolabs).

7.5.6 Design and cloning of mammalian constructs for RNA editing

PspCas13b was made catalytically inactive (dPspCas13b) via two histidine to alanine mutations (H133A/H1058A) at the catalytic site of the HEPN domains. The deaminase domains of human ADAR1 and ADAR2 were synthesized and PCR amplified for Gibson cloning into pcDNA-CMV vector backbones and were fused to dPspCas13b at the C-terminus via GS or GSGGGGS linkers. For the experiment in which we tested different linkers we cloned the following additional linkers between dPspCas13b and ADAR2$_{\text{DD}}$: GGGGSGGGGSGGGGS, EAAAK, GGSGGSGGSGGSGGSGGS, and SGSETPGTSESATPES (XTEN). Specificity mutants were generated by Gibson cloning the appropriate mutants into the dPspCas13b-GSGGGGS backbone.

The luciferase reporter vector for measuring RNA editing activity was generated by creating a W85X mutation (TGG>TAG) in the luciferase reporter plasmid used for knockdown experiments. This reporter vector expresses functional *Gluc* as a normalization control, but a defective *Cluc* due to the addition of the W85X pretermination site. To test ADAR editing motif preferences, we cloned every possible motif around the adenosine at codon 85 (XAX) of *Cluc*.

7.5.7 Testing PFS preferences for dCas13b

For testing PFS preference of REPAIR, we cloned a pooled plasmid library containing a 6 basepair degenerate PFS sequence upstream of a target region and adenosine editing site. The library was synthesized as an ultramer from Integrated DNA Technologies (IDT) and was made double stranded via annealing a primer and using the Klenow fragment of DNA polymerase I (New England Biolabs) to fill in the sequence. This dsDNA fragment containing the degenerate sequence was then Gibson cloned into the digested reporter vector and this was then isopropanol precipitated and purified. The cloned library was then electroporated into Endura competent *E. coli* cells (Lucigen) and plated on 245mm x 245mm square bioassay plates (Nunc). After 16 hours, colonies were harvested and midiprepped using endotoxin-free MACHEREY-NAGEL midiprep kits. Cloned libraries were verified by next-generation sequencing.

7.5.8 Cloning pathogenic G>A mutations for assaying REPAIR activity

For cloning disease-relevant mutations for testing REPAIR activity, 34 G>A mutations related to disease pathogenesis as defined in ClinVar were selected and 200-bp regions surrounding these mutations were golden-gate cloned between mScarlett and EGFP under a CMV promoter. Two additional G>A patient mutations in *AVPR2* and *FANCC* and their cDNA sequences were synthesized and Gibson cloned under expression of EF1alpha.

7.5.9 Guide cloning for REPAIR

For expression of mammalian guide RNAs for REPAIR, the PspCas13b direct repeat sequences were synthesized with golden-gate acceptor sites and cloned under U6 expression via restriction digest cloning. Individual guides were then cloned into this expression backbone by golden-gate cloning.

7.5.10 Mammalian cell culture

Mammalian cell culture experiments were performed in the HEK293FT line (American Type Culture Collection (ATCC)), which was grown in Dulbecco's Modified Eagle Medium with high glucose, sodium pyruvate, and GlutaMAX (Thermo Fisher Scientific), additionally supplemented with 1× penicillin–streptomycin (Thermo Fisher Scientific) and 10% fetal bovine serum (VWR Seradigm). Cells were maintained at confluency below 80%. The U2OS specificity experiment was performed using the U2OS cell line from ATCC and cells were cultured in ATCC-formulated McCoy's 5a Medium Modified.

Unless otherwise noted, all transfections were performed with Lipofectamine 2000 (Thermo Fisher Scientific) in 96-well plates coated with poly-D-lysine (BD Biocoat). Cells were plated at approximately 20,000 cells/well 16 hours prior to transfection to ensure 90% confluency at the time of transfection. For each well on the plate, transfection plasmids were combined with Opti-MEM I Reduced Serum Medium (Thermo Fisher) to a total of 25 μl. Separately, 24.5 μl of Opti-MEM was combined with 0.5 μl of Lipofectamine 2000. Plasmid and Lipofectamine solutions were then combined and incubated for 5 minutes, after which they were pipetted onto cells. The U2OS transfections were performed using Lipofectamine 3000 according to the manufacturer's protocol.

7.5.11 Mammalian cell RNA knockdown assays

To assess RNA targeting in mammalian cells with reporter constructs, 150 ng of Cas13 construct was co-transfected with 300 ng of guide expression plasmid and 12.5 ng of the knockdown reporter construct. 48 hours post-transfection, media containing secreted luciferase was removed from cells, diluted 1:5 in PBS, and measured for activity with BioLux Cypridinia and Biolux Gaussia luciferase assay kits (New England Biolabs) on a plate reader (Biotek Synergy Neo2) with an injection protocol. All replicates performed are biological replicates.

For targeting of endogenous genes, 150 ng of Cas13 construct was co-transfected with 300 ng of guide expression plasmid. 48 hours post-transfection, cells were lysed and RNA was harvested and reverse transcribed using a previously described(Joung et al., 2017) modification of the Cells-to-Ct kit (Thermo Fisher Scientific). cDNA expression was measured via qPCR using TaqMan qPCR probes for the *KRAS* transcript (Thermo Fisher Scientific), *GAPDH* control probes (Thermo Fisher Scientific), and Fast Advanced Master Mix (Thermo Fisher Scientific). qPCR reactions were read out on a LightCycler 480 Instrument II (Roche), with four 5 μl technical replicates in 384-well format.

7.5.12 Evaluation of RNA specificity using pooled libraries of mismatched targets

The ability of Cas13 to interfere with the mismatched target library was tested using HEK293FT cells seeded in 6-well plates. ~70% confluent cells were transfected using 2400 ng Cas13 vector, 4800 ng of guide, and 240 ng of mismatched target library. 48 hours post-transfection, cells were harvested and RNA was extracted using the QIAshredder (Qiagen) and the Qiagen RNeasy Mini Kit. 1 μg of extracted RNA was reverse transcribed using the qScript Flex cDNA synthesis kit (Quantabio) following the manufacturer's gene-specific priming protocol with a *Gluc* specific RT primer. cDNA was then amplified and sequenced on an Illumina NextSeq.

Sequencing was analyzed by counting reads per sequence and depletion scores were calculated by determining the $\log_2$(-read count ratio) value, where read count ratio is the ratio of read counts in the targeting guide condition versus the non-targeting guide condition. This score represents the level of Cas13 activity on the sequence, with higher values representing stronger depletion and thus higher Cas13 cleavage activity. Separate distributions for the single mismatch and double mismatch sequences were determined and plotted as heatmaps with a depletion score for each mismatch identity. For double mismatch sequences the average of all possible double mismatches at a given position were plotted.

7.5.13 Transcriptome-wide profiling of Cas13 in mammalian cells by RNA sequencing

For measurement of transcriptome-wide specificity, 150 ng of Cas13 construct, 300 ng of guide expression plasmid, and 15 ng of the knockdown reporter construct were co-transfected; for shRNA conditions, 300 ng of shRNA targeting plasmid, 15 ng of the knockdown reporter construct, and 150 ng of EF1-alpha driven mCherry (to balance reporter load) were co-transfected. 48 hours post-transfection, RNA was purified with the RNeasy Plus Mini kit (Qiagen), mRNA was isolated using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs), and prepared for sequencing with the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs). RNA sequencing libraries were then sequenced on a NextSeq (Illumina).

To analyze transcriptome-wide sequencing data, reads were aligned to the RefSeq GRCh38 assembly using Bowtie and RSEM version 1.2.31 with default parameters(Li and Dewey, 2011). Transcript expression was quantified as $\log_2(\text{TPM} + 1)$, genes were filtered for $\log_2(\text{TPM} + 1) > 2.5$. For selection of differentially expressed genes, only genes with differential changes of >2 or <.75 were considered. Statistical significance of differential expression was evaluated using a Student's t-test on three targeting replicates versus non-targeting replicates, and filtered for a false discovery rate of <0.01% by the Benjamini-Hochberg procedure.

7.5.14 REPAIR editing in mammalian cells

To assess REPAIR activity in mammalian cells, we transfected 150 ng of REPAIR vector, 300 ng of guide expression plasmid, and 40 ng of the RNA editing reporter. After 48 hours, RNA from cells was harvested and reverse transcribed using a method previously described(Joung et al., 2017) with a gene specific reverse transcription primer. The extracted cDNA was then subjected to two rounds of PCR to add Illumina adaptors and sample barcodes using NEBNext High-Fidelity 2X PCR Master Mix (New England Biolabs). The library was then subjected to next generation sequencing on an Illumina NextSeq or MiSeq. RNA editing rates were then evaluated at all adenosines within the sequencing window.

In experiments where the luciferase reporter was targeted for RNA editing, we also harvested the media with secreted luciferase prior to RNA harvest. In this case, because corrected *Cluc* might be at low levels, we did not dilute the media. We measured luciferase activity with BioLux Cypridinia and Biolux Gaussia luciferase assay kits (New England Biolabs) on a plate reader (Biotek Synergy Neo2) with an injection protocol. All replicates performed are biological replicates.

7.5.15 PFS binding mammalian screen

To determine the contribution of the PFS to editing efficiency in mammalian cells, 625 ng of PFS target library, 4.7 µg of guide, and 2.35 µg of REPAIR were co-transfected in HEK293FT cells plated in 25 cm^2 flasks. Plasmids were mixed with 33 µl of PLUS reagent (Thermo Fisher Scientific), brought to 533 µl with Opti-MEM, incubated for 5 minutes, combined with 30 µl of Lipofectamine 2000 and 500 µl of Opti-MEM, incubated for an additional 5 minutes, and then pipetted onto cells. 48 hours post-transfection, RNA was harvested with the RNeasy Plus Mini kit (Qiagen), reverse transcribed with qScript Flex (Quantabio) using a gene specific primer, and amplified with two rounds of PCR using NEBNext High-Fidelity 2X PCR Master Mix (New England Biolabs) to add Illumina adaptors and sample barcodes. The library was sequenced on an Illumina NextSeq, and RNA editing rates at the target adenosine were mapped to PFS identity. To increase coverage, the PFS was computationally collapsed to 4 nucleotides adjacent to the 5' end of the target sequence. REPAIR editing rates were calculated for each PFS, averaged over biological replicates with non-targeting rates for the corresponding PFS subtracted.

7.5.16 Whole-transcriptome sequencing to evaluate ADAR editing specificity

For analyzing off-target RNA editing sites across the transcriptome, we harvested total RNA from cells 48 hours post-transfection using the RNeasy Plus Miniprep kit (Qiagen). The mRNA fraction was then enriched using a NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) and this RNA was then prepared for sequencing using an NEBNext Ultra RNA Library Prep Kit for Illumina (NEB). The libraries were then sequenced on an Illumina NextSeq and loaded such that there were at least 5 million reads per sample.

7.5.17 RNA editing analysis for targeted and transcriptome-wide experiments

Analysis of the transcriptome-wide editing RNA sequencing data was performed on the FireCloud computational framework (https://software.broadinstitute.org/firecloud/) using a custom workflow we developed: https://portal.firecloud.org/#methods/m/rna_editing_final_workflow/rna_editing_final_workflow/1. For analysis, unless otherwise denoted, sequence files were randomly downsampled to 5 million reads. For the high-coverage sequencing analysis, samples were randomly downsampled to 5 million, 15 million, or 50 million reads. An index was generated using the RefSeq GRCh38 assembly with *Gluc* and *Cluc* sequences added, and reads were aligned and quantified using Bowtie/RSEM version 1.3.0. Alignment BAMs were then sorted and analyzed for RNA editing sites using REDitools (Picardi et al., 2015; Picardi and Pesole, 2013) with the following parameters: -t 8 -e -d -l -U [AG or TC] -p -u -

m20 -T6-0 -W -v 1 -n 0.0. Any significant edits found in untransfected or EGFP-transfected conditions were considered to be SNPs or artifacts of the transfection and filtered out from the analysis of off-targets. Off-targets were considered significant if the Fisher's exact test yielded a p-value less than 0.05 after multiple hypothesis correction by Benjamini Hochberg correction and at least 2 of 3 biological replicates identified the edit site. Overlap of edits between samples was calculated relative to the maximum possible overlap, equivalent to the fewer number of edits between the two samples. The percentage of overlapping edit sites was calculated as the number of shared edit sites divided by minimum number of edits of the two samples, multiplied by 100. For the high-coverage sequencing analysis, an additional layer of filtering for known SNP positions was performed using the Kaviar (Glusman et al., 2011) method for identifying SNPs.

For analyzing the predicted variant effects of each off-target, the list of off-target edit sites was analyzed using the variant annotation integrator (https://genome.ucsc.edu/cgi-bin/hgVai) as part of the UCSC genome browser suite of tools using the SIFT and PolyPhen-2 annotations. To predict whether the off-target genes are oncogenic, a database of oncogenic annotations from the COSMIC catalogue of somatic mutations in cancer was used to characterize off-target genes (cancer.sanger.ac.uk).

For analyzing whether the REPAIR constructs perturbed RNA levels, the transcript per million (TPM) values output from the RSEM analysis were used for expression counts and transformed to log-space by taking the $\log_2$(TPM+1). To find differentially regulated genes, a Student's t-test was performed on three targeting guide replicates versus three non-targeting guide replicates. The statistical analysis was only performed on genes with $\log_2$(TPM+1) values greater than 2.5 and genes were only considered differentially regulated if they had a fold change greater than 2 or less than 0.8. Genes were reported if they had a false discovery rate (Benjamini Hochberg correction) of less than 0.01.

7.6 Acknowledgements

This work was published in (Cox et al., 2017). We would like to thank Ruchi Munshi, Alex Baumann, Anthony Philippakis, and Eric Banks for computational guidance, Christoph Muus for sequencing assistance, Aaron Smargon for gene synthesis of a Cas13b orthologs 1-8, Vanessa Verdine for assistance with RNA sequencing library preparation, Rhiannon Macrae, Rich Belliveau, and the entire Zhang lab for discussions and support.

Chapter 8

Conclusion

The generation of somatic or germline genome edits allows for the study and modeling of genetic variants and holds the promise for modulating human traits and treating all genetic diseases. The development of CRISPR-Cas9 for genome editing has enabled the unprecedented speed and scale at which scientists can study diseases, generate mouse models, and develop human gene therapy due to the high efficiency of mutagenesis and the ease of programmability. While the Cas9 technology has matured, we have demonstrated that there is more Class 2 enzyme diversity beyond Cas9 and that the enzymes, such as Cpf1 (i.e. Cas12a), Cas12b, and Cas13, can be characterized and used for novel genome editing applications. In particular, we focus on the RNA-guided RNA targeting Cas13 system, determining its biochemical mechanism in detail and biological function in bacteria as an adaptive immune system against RNA phage or transcripts of DNA phages. While characterizing the system, we discovered a peculiar phenomenon called the collateral effect in which activated Cas13 complex bound to target ssRNA would cleave any other RNAs in solution. We harnessed this effect to cleave fluorescent reporters in response to the presence of specific nucleic acid targets and developed it into a highly sensitive, specific, cheap, and rapid diagnostic called SHERLOCK. Because mammalian RNA perturbation tools are lacking, we further optimized Cas13 expression for mammalian cells and developed tools to knockdown RNA, image transcripts, and precisely and efficiently perform A to G base edits in RNA. These applications highlight the promise of new enzyme discovery, and for Cas13 to enable a comprehensive RNA targeting toolbox for biology and therapeutics.

8.1 Comparison of characterized Class 2 CRISPR enzymes

The diversity of enzymes found in nature, especially bacteria, has evolved over billions of years and serves as a rich resource for biology and biotechnology. In this thesis alone, we found numerous new CRISPR enzyme systems with unique and interesting biochemical properties and numerous applications for biotechnology with immense potential to affect human health and society. Related to the work of this thesis, we also discovered numerous DNA targeting systems, including Cas12a and Cas12b, in our original computational work that we have shown also function for genome editing in mammalian cells with very different properties than Cas9. For instance, Cas12a has no tracrRNA, a very short crRNA, and has a AT-rich PAM allowing for expanded targeting range in genomes. Cas12a is more specific owing to two seed regions in the guide:target duplex and has worked in numerous organisms beyond humans. As a direct result of this thesis, there are now four Class 2 enzymes known and characterized, including Cas9, Cas12a, Cas12b, and Cas13 (Figure 8.1 and 8.2). Their diverse properties are described below and summarized in Table 8.1.

Cas9. The most well-characterized single-protein Cas effector is Cas9 (Class 2, type II). Cas9 is a dual-guide RNA-dependent endonuclease that contains two nuclease domains, RuvC and HNH, that coordinate to cleave both strands of a double-stranded DNA (dsDNA) target complementary to the crRNA spacer sequence (Doudna and Charpentier, 2014; Hsu et al., 2014). Cas9 can be engineered to facilitate genome editing in eukaryotic cells (Cong et al., 2013; Mali et al., 2013c). The two guide RNAs, crRNA and its trans-activating crRNA (tracrRNA) (Deltcheva et al., 2011), can be covalently linked to form a single-guide RNA (sgRNA) to simplify the RNA components (Jinek et al., 2012). Targeting is PAM-dependent (*e.g.,* Cas9 from *Streptococcus pyogenes* requires a 5'-NGG-3' PAM at the 3' end of the target sequence), restricting editing to genomic sites that contain the appropriate PAM sequence.

Cas12. CRISPR-Cas12a systems (originally called Cpf1) also target DNA using a single-protein endonuclease and can be engineered to facilitate genome editing in eukaryotic cells (Zetsche et al., 2015b). Like Cas9, Cas12a cleaves dsDNA using two DNase domains, the RuvC and nuclease domains (Yamano et al., 2016; Zetsche et al., 2015b). By contrast, Cas12a does not require a tracrRNA (Zetsche et al., 2015b) and possesses its own RNase domain capable of cleaving a pre-crRNA array into individual mature crRNAs for convenient delivery of multiple guide RNAs (Fonfara et al., 2016; Zetsche et al., 2017). Cas12a-mediated dsDNA cleavage results in 5' staggered ends, (Zetsche et al., 2015b) and the PAM sequence and location differ from those used by Cas9, with most Cas12a orthologs requiring a 5'-TTTV-3' PAM at the 5' end of the target site and cleaving distal to the PAM (Zetsche et al., 2015b). There is also the Cas12b system (formerly called C2c1) that functions similar to Cas12a in having staggered cleavage overhangs and an AT-rich PAM, but requires an accessory tracrRNA for functional activity (Shmakov et al., 2015).

Cas13. The recent discovery of single-effector CRISPR systems that target RNA (Shmakov et al., 2015; Shmakov et al., 2017a) has further expanded our conception of the diversity of these bacterial enzymes. Members of the Cas13 family contain two domains with homology to higher eukaryote's and prokaryote's nucleotide-binding (HEPN) RNase domains (Anantharaman et al., 2013). Cas13a, like DNA-targeting CRISPR systems, is guided to target RNA sequences by a complementary crRNA, and hybridization between the target RNA and its complementary crRNA leads to cleavage of the target RNA at multiple sites within single-stranded regions (Abudayyeh et al., 2016). Continued protein discovery has since uncovered many Cas13 subfamilies (Cas13a, Cas13b, Cas13c, and Cas13d) (Abudayyeh et al., 2016; East-Seletsky et al., 2017; East-Seletsky et al., 2016; Konermann et al., 2018; Shmakov et al., 2015; Shmakov et al., 2017a; Smargon et al., 2017b; Yan et al., 2018). In bacteria, both Cas13a and Cas13b have a sequence constraint known as the protospacer flanking site (PFS) (Abudayyeh et al., 2016; Smargon et al., 2017b). In bacteria, Cas13a from *Leptotrichia shahii* has a 3' H (not G) requirement (Abudayyeh et al., 2016), and in mammalian cells it appears that the PFS does not affect targeting (Abudayyeh et al., 2017; Cox et al., 2017). An additional unique feature of these enzymes observed *in vitro* is that, upon target RNA binding, the activated Cas13 complex is capable of cleaving both the targeted transcript and other nearby non-complementary RNAs via a "collateral" activity (Abudayyeh et al., 2016), which has been harnessed for nucleic acid detection applications (East-Seletsky et al., 2017; East-Seletsky et al., 2016; Gootenberg et al., 2018; Gootenberg et al., 2017c).

These examples illustrate the value of continued exploration of the diversity of CRISPR systems for the development of molecular tools. Many CRISPR systems remain unexplored and may harbor novel combinations of sequence flexibility, endonuclease capabilities, or other unexpected functions to facilitate genome-engineering applications.

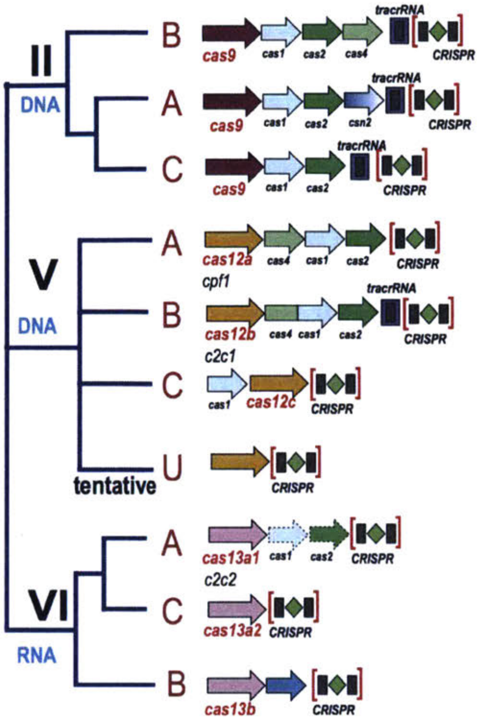


Figure 8.1: Phylogenetic tree and locus architecture of Class 2 CRISPR systems.

Two DNA-targeting and one RNA-targeting Class 2 systems have been characterized. Reproduced from (Makarova et al., 2017).

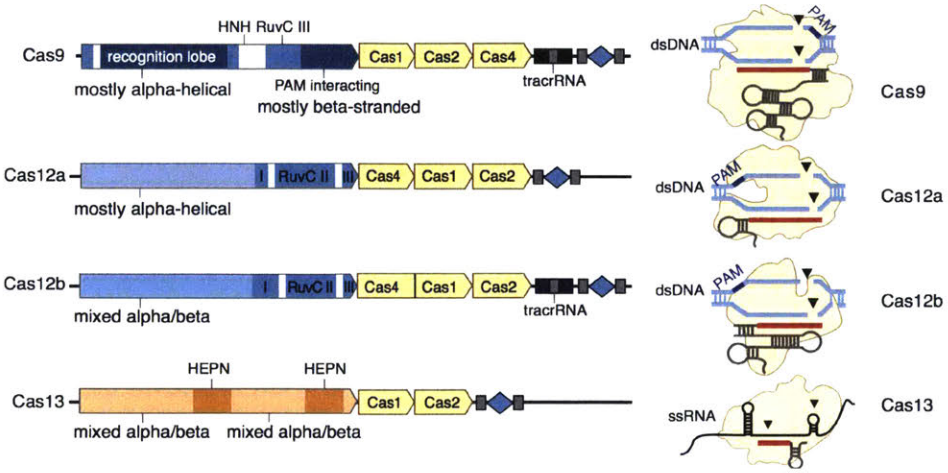


Figure 8.2: Genomic architecture and complex structure of the Class 2 enzymes.

Table 8.1: Comparison of the biochemical features of Class 2 CRISPR enzymes.

	Cas9	Cas12a/Cpf1	Cas12b/C2c1	Cas13a/C2c2
DR Position	3’	5’	5’	5’
PAM/PFS Position	3’	5’	5’	3’
PAM Identity	GC rich	AT rich	AT rich	not G
tracrRNA	✓	✗	✓	✗
cleavage	blunt	staggered	Staggered	ssRNA breaks
target	dsDNA	dsDNA	dsDNA	ssRNA
Collateral activity	No	Yes	No	Yes

8.2 Nucleic acid testing platform with CRISPR diagnostics

Nucleic acid testing with Cas13 is an exciting new direction that has resulted in a new field of CRISPR diagnostics. Although we did not set out to develop next generation molecular diagnostics, the work presented here is a classic example of how research can guide us in unexpected directions. CRISPR enzymes were previously only used for studying biology or generating therapeutics, but new discoveries have propelled the field into diagnostics. The Cas13 collateral effect was uniquely poised for highly sensitive diagnostics because of its signal amplification and high nucleotide specificity due to the inherent specificity of target recognition by the guide/protein complex. By combining Cas13 detection with isothermal pre-amplification strategies, we generated a diagnostic solution that is cheap, rapid, and highly portable and that does not require complex skill to run. Nucleic acid tests that are cheap, quick, and simple-to-use are urgently needed in fields such as infectious disease, cancer cell free DNA monitoring, and agriculture. For example, sepsis diagnostic evaluation can take days, something a sick person cannot afford. Having molecular diagnostics that can guide treatment decision making in a matter of minutes to an hour can save lives. This is especially true in the developing world where portable, point-of-care solutions are necessary, especially during epidemics. While the SHERLOCK technology is quite mature, there is still more work to distill the test into a small handheld device and have additional readouts, such as mobile electronic or colorimetric solution versions. Additionally, more clinical evaluation on large patient sample sets are necessary to evaluate the sensitivity and specificity of the test and robustness towards sample-to-sample variability. Nevertheless, the SHERLOCK platform is well poised to improve diagnostics in numerous fields, including human and animal healthcare and agriculture.

8.3 An expanding RNA toolbox with CRISPR-Cas13

Beyond nucleic acid diagnostics, Cas13 has immense potential for generating a diverse set of tools for studying and perturbing RNA. Although some RNA tools existed before, they were based on protein recognition of RNA requiring intense protein engineering to reprogram akin to Zinc finger nucleases or TALENs. We envision Cas13 as a platform for any RNA perturbations, including the ability to manipulate translation, splicing, localization, transcript levels, base identity, and base modifications. In this thesis, we demonstrate highly specific RNA knockdown that is as efficient as RNA interference, but orders of magnitude more specific. We also generate catalytically inactive dead Cas13 versions for serving as a scaffold for targeted recruitment of enzymes, such as GFP for transcript imaging or ADAR for base editing. An exciting area of future development is recruitment of enzymes for targeted base modification, such as methylation or pseudouridylation, as epitranscriptomics is emerging as an abundant phenomenon in cells, but tools for studying them are lacking.

These new RNA-targeted tools open new opportunities to study many additional aspects of RNA biology. Here, I will highlight key developments and potential applications with Cas13 (Figure 8.3).

RNA knockdown. In light of the challenges in distinguishing the effects of DNA- and RNA-mediated mechanisms (Section 2.2), technologies for RNA knockdown with CRISPR enzymes have the potential to enable more specific studies of RNA-mediated functions because they avoid manipulating DNA sequence. To this end, we recently showed that the Cas13 platform can knock down RNAs with efficiency similar to RNA interference (Figure 8.3) (Abudayyeh et al., 2017; Cox et al., 2017). At the same time, Cas13 knockdown of a reporter transcript (luciferase) had minimal effects on the rest of the transcriptome, whereas RNAi knockdown of the same transcript led to significant effects on ~900 other genes (Abudayyeh et al., 2017; Cox et al., 2017; Konermann et al., 2018). Thus, RNA knockdown screens with Cas13 have the potential to reduce the off-target effects that have confounded RNAi screens, and, in contrast to approaches that rely on endogenous RNAi proteins, the Cas13 enzymes can be further optimized and engineered for additional functionality or specificity. For example, adding a nuclear localization tag to Cas13a led to successful knockdown of nuclear-localized lncRNAs, while RNAi was ineffective (Abudayyeh et al., 2017). We also found that a second Cas13 enzyme, Cas13b from *Prevotella sp. P5-125*, was far more effective at RNA knockdown (>90% for each site targeted) than our original studies using LwaCas13a (Cox et al., 2017). Further optimizations of the system — including adapting other Cas13 proteins and guide RNAs for viral delivery, learning design rules for efficient guides, and exploring or optimizing Cas13 proteins for high stability and specificity — should soon enable RNA-targeted forward genetic screens to directly study RNA function.

Post-transcriptional RNA modifications. Post-transcriptional modifications of RNA nucleotides — including N6-methyladenosine (m6A), pseudouridine, and inosine — appear to be a critical layer of post-transcriptional regulation (Helm and Motorin, 2017; Hsu et al., 2017; Lewis et al., 2017; Schwartz, 2016). RNA mapping technologies have revealed that these modifications can affect tens of thousands of sites in both coding and noncoding RNAs (Li et al., 2016), but the functions of most of these modifications are poorly understood. We developed a Cas13 approach to perform adenine-to-inosine (A-to-I) editing, a natural modification that alters the sequence of RNA. In this approach, named RNA Editing for Programmable A to I Replacement (REPAIR), we fused catalytically dead Cas13 to the ADAR2 enzyme to enable targetable A-to-I editing activity (Cox et al., 2017) (Figure 8.3). By extending the guide RNA and including a cytosine mismatch in the guide across from the targeted adenosine, a more optimal double-stranded RNA substrate with a bubble was created, maximizing ADAR2 enzymatic activity. This approach achieved up to 90% A-to-I editing of targeted sites on reporter transcripts and up to 40% on endogenous transcripts. Off-target editing observed with the initial system (>18,000 edited off-target sites) was reduced through a rational mutagenesis strategy, which produced a more specific version of Cas13 that retained on-target editing activity but reduced off-target activity (20 edited off-target sites) (Cox et al., 2017). Because of the high specificity, high efficiency, lack of sequence restriction, and precise modification site, REPAIR provides a useful tool for studying the effects of sequence variants and inosine modifications, and for developing therapeutics.

Programmable and precise base modification with REPAIR is also a promising avenue for temporal modification of genetic variants in cells. The modulation of enzymatic function via sites of phosphorylation or other post-translational modifications will allow for the precise modulation of signaling processes in a cell and can also be useful for therapeutics. For instance, in the case of acute liver failure, ß-catenin can be activated via mutagenesis of phosphorylation sites, leading to regeneration of hepatocytes and prevention of fibrosis. Additionally, many acute states of disease such as infectious disease, cancer, auto-immune disease, pain, or migraines could be targeted with REPAIR. Targeted cellular differentiation via temporal genetic modification would also be possible. REPAIR delivered via AAV also offers the opportunity for long term expression and permanent correction of genetic disease that could be reversed. A major advantage of REPAIR is it directly deaminates the riboadenosine, not requiring endogenous repair pathways, and thus would be highly efficient in post-mitotic cells like neurons. For the full potential of RNA editing to be realized, more work needs to be done to demonstrate REPAIR in *in vivo* models of disease and characterize the immunogenic effects of REPAIR expression. Additionally, continued protein engineering and directed evolution of base editing enzymes will allow for additional base modifications beyond just adenosine to inosine editing. Similar approaches with alternative RNA editing or modification enzymes, such as RNA methyltransferases, cytidine deaminases, and pseudouridine synthases, may enable precise introduction of RNA modifications and studies to connect these modifications with their cellular and molecular functions (Figure 8.3).

RNA splicing, localization, and more. RNA-targeted CRISPR tools are still in their infancy, and — just as an explosion of dCas9 fusion tools have begun to reveal biological insights into transcriptional regulation — dCas13 fusions will likely have a major impact in illuminating post-transcriptional regulation. Indeed, studies using previous RNA-targeting technologies have highlighted additional possibilities for manipulating RNA regulation with new CRISPR tools. For example, fusion of PUF proteins to the arginine- and serine-rich domain of SRSF1 or the glycine-rich domain of HNRNPA1 enabled the creation of programmable activators and repressors of splicing, respectively (Cheong and Hall, 2006). RNA recruitment systems using MS2 and PP7 (or other systems that involve tagging an endogenous RNA with the recognition sequence for an exogenous RNA-binding protein) have been used to test the sufficiency of specific RNA-protein interactions for RNA function (Bos et al., 2016; Chen and Varani, 2013) and to track transcripts via fluorescence (Park et al., 2010). Early studies have shown the potential for Cas13 tools to modulate the expression of specific RNA isoforms by binding and blocking splice sites (Konermann et al., 2018) (Figure 8.3). The ability to combine each of these tools with transcriptome-wide screening will enable systematic perturbation studies of many aspects of RNA biology, including translation, splicing, RNA localization, and post-transcriptional modifications (Figure 8.3).

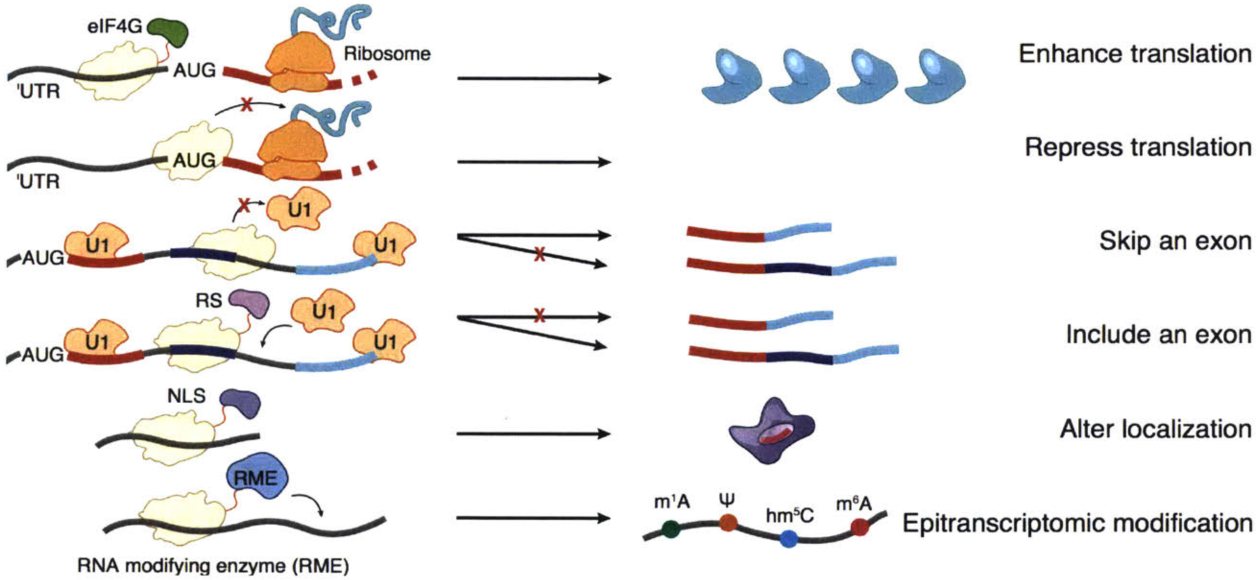


Figure 8.3: Applications that can be developed with Cas13 RNA tools.

CRISPR studies of RNA regulation will not only reveal the functions of key molecular processes, but will also help us to directly manipulate them in humans. Pre-clinical development for CRISPR-based therapeutics is already underway, including approaches for genome editing *ex vivo* (where editing is performed outside the body and edited cells are transplanted into the patient) and *in vivo* (where CRISPR tools are delivered to edit cells directly in the tissue of interest). Many challenges remain, including optimizing the efficiency and specificity of CRISPR editing, developing new methods for efficient and precise delivery to target tissues and cell types, and minimizing unintended consequences of these manipulations on immune responses or target cell functions. Nonetheless, overcoming these challenges promises to enable a flexible class of therapeutics that leverage a programmable suite of CRISPR tools to manipulate DNA or RNA.

While therapeutic development efforts have largely focused on modification of protein-coding genes, approaches for targeting DNA and RNA regulatory mechanisms may prove advantageous for certain indications. For example, therapeutic targeting of DNA enhancers, which can have cell-type specific activities, may enable more precise modulation of gene function in cell types relevant to disease (Canver et al., 2015). RNA-targeted therapeutic approaches may enable specific regulation of entities that are difficult to manipulate at the level of DNA, such as degradation of toxic repeat expansion RNAs (Batra et al., 2017) or selective activation of specific RNA splice isoforms (Konermann et al., 2018; Palacino et al., 2015).

8.4 Limitations of CRISPR-Cas13

Despite many of the advances Cas13 enables, there are still limitations that need to be addressed. Cas13 catalytic activity could be improved, as knockdown in mammalian cells varies widely and is not always robust. Although discovery of Cas13b yielded a version that could reach up to 95% protein knockdown, there are still many guides that do not work and often endogenous transcript knockdown does not exceed 50%. Newer Cas13 orthologs or programmable RNA targeting tools other than Cas13 could potentially be better. In addition, finding orthologs that bind the target tighter could improve certain applications as well. Although I was able to develop a transcript imaging tool with Cas13, many guides did not work, and imaging was limited to highly expressed transcripts. Based on biochemical data, the K_d for Cas13 is only in the low nM range, which is much weaker than the MS2 to MS2-binding protein interaction, which is in the lower pM range. In order to make transcript imaging more robust and capable of imaging lowly expressed transcripts, either engineering Cas13 to have a stronger K_d or finding tighter-binding Cas13 orthologs is necessary. In general, having a Cas13 capable of strong binding would improve many other applications, including splicing modulation, which relies on blocking splicing factors, and RNA-editing, which requires robust target binding. An additional consideration for RNA editing is that binding must not be too tight as to block ribosome readthrough and so there is likely an optimal binding strength. A deeper understanding of Cas13b activation upon target binding and the resultant conformational shift will also enable more applications. For instance, two domains of split-GFP could be placed in specific positions on the Cas13 protein and would reassociate upon the conformational change of Cas13 induced by target recognition. This principle could be applied to any split-effector application, such as target-dependent killing of cells via split Caspase proteins that are reconstituted on activated Cas13 molecules.

8.5 Looking beyond CRISPR for next generation technologies

As more bacteria and archaea are sequenced around the world (Figure 8.4), additional computational mining of genomes will yield even newer systems with unimaginable properties that could be useful biotechnologically. Even recently, additional mining of more bacterial databases, yielded a newer Cas13d that was smaller and more amenable for packaging in AAV delivery vectors and was shown to be more superior for RNA knockdown due to its stability in mammalian cells (Konermann et al., 2018). Given the rate at which microorganisms are being sequenced (Figure 8.4), there is tremendous opportunity to find systems beyond CRISPR that play roles in important bacterial processes, such as cell defense, metabolism, and genome regulation. Enzymes in these systems could be useful for human genome editing, especially as more efficient methods for gene insertion and deletion are needed that are independent of endogenous repair pathways.

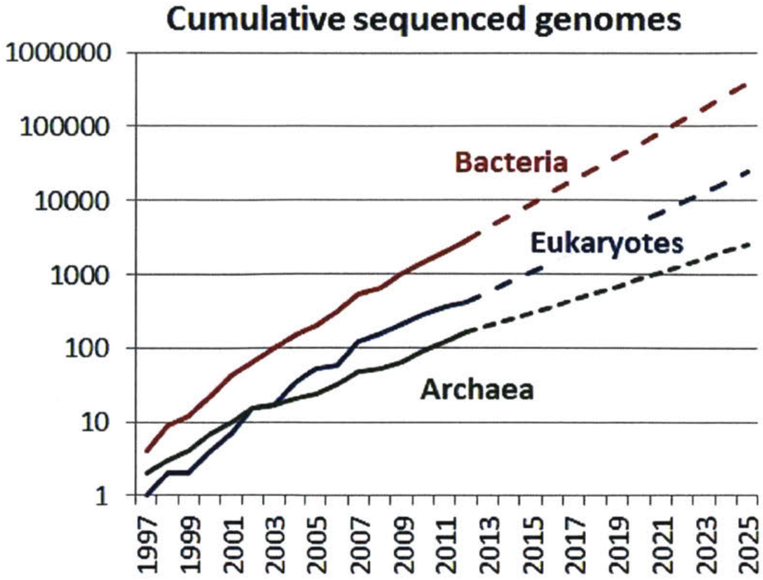


Figure 8.4: Cumulative sequenced genomes.

Projected number of sequenced prokaryotic genomes over time. Adapted from (Andrew, 2013).

The idea that many gene systems are waiting to be explored in microbial genomes was recently explored in a systematic study of phage defense systems. Doron et al. developed a computational approach to uncover genes that cluster together in defense islands nearby known phage defense genes (Doron et al., 2018). This approach uncovered known defense genes, such as restriction enzymes or CRISPR systems, as well as hundreds of more gene families, nine of which were experimentally verified to protect model bacteria from phage infection. These proteins adopted unique mechanisms, including the use of bacterial flagella and chromosome maintenance complexes, and many of the genes showed similarity to Toll-interleukin receptor (TIR) domains involved in innate immunity in animals and plants (Doron et al., 2018). While only a handful of these systems were experimentally explored, there were tens of thousands of new phage defense systems waiting to be explored. In addition to expanding our understanding of the rich molecular processes at play in bacteria, there is also the potential for many new biotechnological tools using these proteins once their mechanisms are deciphered. With each new enzyme discovered, whether within CRISPR or beyond, novel biology will be understood, new biotechnologies developed, and the closer we will come to better understanding disease and eventually curing all genetic disease.

Appendix A

References

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., 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.
3. 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.
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 U S A *113*, E2579-2588.
5. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.
6. 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.
7. Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569-573.
8. Andrew, S. (2013). Analysis of Sequenced Genomes.
9. Aravind, L., Makarova, K.S., and Koonin, E.V. (2000). SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. Nucleic Acids Res *28*, 3417-3432.
10. Ballatore, C., Lee, V.M., and Trojanowski, J.Q. (2007). Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. Nat Rev Neurosci *8*, 663-672.

11. Barletta, J.M., Edelman, D.C., and Constantine, N.T. (2004). Lowering the detection limits of HIV-1 viral load using real-time immuno-PCR for HIV-1 p24 antigen. Am J Clin Pathol *122*, 20-27.

12. 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.

13. Barrangou, R., and Marraffini, L.A. (2014). CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Mol Cell *54*, 234-244.

14. Barrangou, R., and Van der Oost, J., eds. (2013). CRISPR-Cas Systems. RNA-mediated Adaptive Immunity in Bacteria and Archaea (Heidelberg: Springer).

15. Bass, B.L., and Weintraub, H. (1987). A developmentally regulated activity that unwinds RNA duplexes. Cell *48*, 607-613.

16. Bass, B.L., and Weintraub, H. (1988). An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell *55*, 1089-1098.

17. Batra, R., Nelles, D.A., Pirie, E., Blue, S.M., Marina, R.J., Wang, H., Chaim, I.A., Thomas, J.D., Zhang, N., Nguyen, V., Aigner, S., Markmiller, S., Xia, G., Corbett, K.D., Swanson, M.S., and Yeo, G.W. (2017). Elimination of Toxic Microsatellite Repeat Expansion RNA by RNA-Targeting Cas9. Cell *170*, 899-912 e810.

18. Benda, C., Ebert, J., Scheltema, R.A., Schiller, H.B., Baumgartner, M., Bonneau, F., Mann, M., and Conti, E. (2014). Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. Mol Cell *56*, 43-54.

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

20. 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.

21. Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res *29*, 2607-2618.

22. Bettegowda, C., Sausen, M., Leary, R.J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B.R., Wang, H., Luber, B., Alani, R.M., *et al.* (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med *6*, 224ra224.

23. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L. (2013a). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic acids research *41*, 7429-7437.

24. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L.A. (2013b). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res *41*, 7429-7437.

25. Bintu, L., Yong, J., Antebi, Y.E., McCue, K., Kazuki, Y., Uno, N., Oshimura, M., and Elowitz, M.B. (2016). Dynamics of epigenetic regulation at the single-cell level. Science *351*, 720-724.

26. Biswas, A., Fineran, P.C., and Brown, C.M. (2014). Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. Bioinformatics *30*, 1805-1813.

27. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. Science *326*, 1509-1512.

28. Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology *151*, 2551-2561.
29. Bos, T.J., Nussbacher, J.K., Aigner, S., and Yeo, G.W. (2016). Tethered Function Assays as Tools to Elucidate the Molecular Roles of RNA-Binding Proteins. Adv Exp Med Biol *907*, 61-88.
30. 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.
31. Caliendo, A.M., and Hodinka, R.L. (2017). A CRISPR Way to Diagnose Infectious Diseases. N Engl J Med *377*, 1685-1687.
32. Campbell, Z.T., Valley, C.T., and Wickens, M. (2014). A protein-RNA specificity code enables targeted activation of an endogenous human transcript. Nat Struct Mol Biol *21*, 732-738.
33. Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.C., Zhang, F., Orkin, S.H., and Bauer, D.E. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192-197.
34. Chang, S.S., and Kang, D.H. (2004). Alicyclobacillus spp. in the fruit juice industry: history, characteristics, and current isolation/detection procedures. Crit Rev Microbiol *30*, 55-74.
35. Chavez, A., Scheiman, J., Vora, S., Pruitt, B.W., Tuttle, M., E, P.R.I., Lin, S., Kiani, S., Guzman, C.D., Wiegand, D.J., Ter-Ovanesyan, D., Braff, J.L., Davidsohn, N., Housden, B.E., Perrimon, N., Weiss, R., Aach, J., Collins, J.J., and Church, G.M. (2015). Highly efficient Cas9-mediated transcriptional programming. Nat Methods *12*, 326-328.
36. Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., and Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell *155*, 1479-1491.
37. Chen, C.Y. (2014). DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Front Microbiol *5*, 305.
38. Chen, J.S., Ma, E., Harrington, L.B., Tian, X., and Doudna, J.A. (2017). CRISPR-Cas12a target binding unleashes single-stranded DNase activity. bioRxiv.
39. Chen, Y., and Varani, G. (2013). Engineering RNA-binding proteins for biology. FEBS J *280*, 3734-3754.
40. Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B., and Jaenisch, R. (2013). Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res *23*, 1163-1171.
41. Cheong, C.G., and Hall, T.M. (2006). Engineering RNA sequence specificity of Pumilio repeats. Proc Natl Acad Sci U S A *103*, 13635-13639.
42. Cho, S.W., Kim, S., Kim, J.M., and Kim, J.S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol *31*, 230-232.
43. Choo, Y., Sanchez-Garcia, I., and Klug, A. (1994). In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. Nature *372*, 642-645.

44. Choudhury, R., Tsai, Y.S., Dominguez, D., Wang, Y., and Wang, Z. (2012). Engineering RNA endonucleases with customized sequence specificities. Nat Commun *3*, 1147.

45. Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. Genetics *186*, 757-761.

46. 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.

47. 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.

48. Compton, J. (1991). Nucleic acid sequence-based amplification. Nature *350*, 91-92.

49. 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.

50. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

51. Cottrell, S., Bicknell, D., Kaklamanis, L., and Bodmer, W.F. (1992). Molecular analysis of APC mutations in familial adenomatous polyposis and sporadic colon carcinomas. Lancet *340*, 626-630.

52. 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. Science *358*, 1019-1027.

53. Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome research *14*, 1188-1190.

54. Dahlman, J.E., Abudayyeh, O.O., Joung, J., Gootenberg, J.S., Zhang, F., and Konermann, S. (2015). Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. Nat Biotechnol *33*, 1159-1161.

55. Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun *3*, 945.

56. De Gregorio, E., Preiss, T., and Hentze, M.W. (1999). Translation driven by an eIF4G core domain in vivo. EMBO J *18*, 4865-4874.

57. Dejnirattisai, W., Supasa, P., Wongwiwat, W., Rouvinski, A., Barba-Spaeth, G., Duangchinda, T., Sakuntabhai, A., Cao-Lormeau, V.M., Malasit, P., Rey, F.A., Mongkolsapaya, J., and Screaton, G.R. (2016). Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. Nat Immunol *17*, 1102-1108.

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

59. Deng, L., Garrett, R.A., Shah, S.A., Peng, X., and She, Q. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Mol Microbiol *87*, 1088-1099.

60. Diez-Villasenor, C., Guzman, N.M., Almendros, C., Garcia-Martinez, J., and Mojica, F.J. (2013). CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biol *10*, 792-802.

61. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

62. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. Science *359*.

63. Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096.

64. Drozdetskiy, A., Cole, C., Procter, J., and Barton, G.J. (2015). JPred4: a protein secondary structure prediction server. Nucleic Acids Res *43*, W389-394.

65. Du, Y., Pothukuchy, A., Gollihar, J.D., Nourani, A., Li, B., and Ellington, A.D. (2017). Coupling Sensitive Nucleic Acid Amplification with Commercial Pregnancy Test Strips. Angew Chem Int Ed Engl *56*, 992-996.

66. East-Seletsky, A., O'Connell, M.R., Burstein, D., Knott, G.J., and Doudna, J.A. (2017). RNA Targeting by Functionally Orthogonal Type VI-A CRISPR-Cas Enzymes. Mol Cell *66*, 373-383 e373.

67. 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.

68. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

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

70. Edwards, T.A., Pyle, S.E., Wharton, R.P., and Aggarwal, A.K. (2001). Structure of Pumilio reveals similarity between RNA and peptide binding motifs. Cell *105*, 281-289.

71. 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.

72. Emmadi, R., Boonyaratanakornkit, J.B., Selvarangan, R., Shyamala, V., Zimmer, B.L., Williams, L., Bryant, B., Schutzbank, T., Schoonmaker, M.M., Amos Wilson, J.A., Hall, L., Pancholi, P., and Bernard, K. (2011). Molecular methods and platforms for infectious diseases testing a review of FDA-approved and cleared assays. J Mol Diagn *13*, 583-604.

73. Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., and Mountain, J. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet *6*, e1000993.

74. Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science *310*, 1817-1821.

75. Farzadfard, F., Perli, S.D., and Lu, T.K. (2013). Tunable and multifunctional eukaryotic transcription factors based on CRISPR/Cas. ACS Synth Biol *2*, 604-613.

76. Ferreira, R.C., Pan-Hammarstrom, Q., Graham, R.R., Gateva, V., Fontan, G., Lee, A.T., Ortmann, W., Urcelay, E., Fernandez-Arquero, M., Nunez, C., Jorgensen, G., Ludviksson, B.R., Koskinen, S., Haimila, K., Clark, H.F., Klareskog, L., Gregersen, P.K., Behrens, T.W., and Hammarstrom, L. (2010). Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. Nat Genet *42*, 777-780.

77. Filipovska, A., and Rackham, O. (2011). Designer RNA-binding proteins: New tools for manipulating the transcriptome. RNA Biol *8*, 978-983.
78. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.
79. Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lecrivain, A.L., Bzdrenga, J., Koonin, E.V., and Charpentier, E. (2014). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res *42*, 2577-2590.
80. 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.
81. Frieda, K.L., Linton, J.M., Hormoz, S., Choi, J., Chow, K.K., Singer, Z.S., Budde, M.W., Elowitz, M.B., and Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. Nature *541*, 107-111.
82. Fukuda, M., Umeno, H., Nose, K., Nishitarumizu, A., Noguchi, R., and Nakagawa, H. (2017). Construction of a guide-RNA for site-directed RNA mutagenesis utilising intracellular A-to-I RNA editing. Sci Rep *7*, 41478.
83. Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67-71.
84. 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 U S A *109*, E2579-2586.
85. Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. Nature *551*, 464-471.
86. Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., Qi, L.S., Kampmann, M., and Weissman, J.S. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647-661.
87. Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., Lim, W.A., Weissman, J.S., and Qi, L.S. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442-451.
88. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L., and Roach, J.C. (2011). Kaviar: an accessible system for testing SNV novelty. Bioinformatics *27*, 3216-3217.
89. Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature *514*, 633-637.
90. Gootenberg, J.S., Abudayyeh, O., Kellner, M., Joung, J., Collins, J.J., and Zhang, F. (2017a). SHERLOCKv2: Multiplexed and portable nucleic acid detection platform with Cas13, Cpf1, and Csm6. Submitted.
91. Gootenberg, J.S., Abudayyeh, O.O., Kellner, M.J., Joung, J., Collins, J.J., and Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science.

92. 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. (2017b). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science.

93. 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. (2017c). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science *356*, 438-442.

94. Green, A.A., Silver, P.A., Collins, J.J., and Yin, P. (2014). Toehold switches: de-novo-designed regulators of gene expression. Cell *159*, 925-939.

95. Grimm, D., Streetz, K.L., Jopling, C.L., Storm, T.A., Pandey, K., Davis, C.R., Marion, P., Salazar, F., and Kay, M.A. (2006). Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. Nature *441*, 537-541.

96. Grimm, D., Wang, L., Lee, J.S., Schurmann, N., Gu, S., Borner, K., Storm, T.A., and Kay, M.A. (2010). Argonaute proteins are key determinants of RNAi efficacy, toxicity, and persistence in the adult mouse liver. J Clin Invest *120*, 3106-3119.

97. Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res *35*, W52-57.

98. Gross, G.G., Junge, J.A., Mora, R.J., Kwon, H.B., Olson, C.A., Takahashi, T.T., Liman, E.R., Ellis-Davies, G.C., McGee, A.W., Sabatini, B.L., Roberts, R.W., and Arnold, D.B. (2013). Recombinant probes for visualizing endogenous synaptic proteins in living neurons. Neuron *78*, 971-985.

99. Grynberg, M., Erlandsen, H., and Godzik, A. (2003). HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins. Trends Biochem Sci *28*, 224-226.

100. Gupta, N., Limbago, B.M., Patel, J.B., and Kallen, A.J. (2011). Carbapenem-resistant Enterobacteriaceae: epidemiology and prevention. Clin Infect Dis *53*, 60-67.

101. Hale, C.R., Cocozaki, A., Li, H., Terns, R.M., and Terns, M.P. (2014). Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex. Genes Dev *28*, 2432-2443.

102. Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M., and Terns, M.P. (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Mol Cell *45*, 292-302.

103. 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.

104. Hannon, G.J., and Rossi, J.J. (2004). Unlocking the potential of the human genome with RNA interference. Nature *431*, 371-378.

105. Hayes, F., and Van Melderen, L. (2011). Toxins-antitoxins: diversity, evolution and function. Crit Rev Biochem Mol Biol *46*, 386-408.

106. 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.

107. Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D., and Marraffini, L.A. (2015). Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature.

108. Helm, M., and Motorin, Y. (2017). Detecting RNA modifications in the epitranscriptome: predict and validate. Nat Rev Genet *18*, 275-291.

109. Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol.

110. Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol *33*, 510-517.

111. Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262-1278.

112. Hsu, P.J., Shi, H., and He, C. (2017). Epitranscriptomic influences on development and disease. Genome Biol *18*, 197.

113. Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B., and Chess, A. (2007). A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics *8*, 39.

114. Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol *31*, 227-229.

115. 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.

116. Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol *21*, 635-637.

117. Jackson, A.L., Burchard, J., Leake, D., Reynolds, A., Schelter, J., Guo, J., Johnson, J.M., Lim, L., Karpilow, J., Nichols, K., Marshall, W., Khvorova, A., and Linsley, P.S. (2006a). Position-specific chemical modification of siRNAs reduces "off-target" transcript silencing. RNA *12*, 1197-1205.

118. Jackson, A.L., Burchard, J., Schelter, J., Chau, B.N., Cleary, M., Lim, L., and Linsley, P.S. (2006b). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. RNA *12*, 1179-1187.

119. 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). Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science *345*, 1473-1479.

120. Jackson, R.N., and Wiedenheft, B. (2015). A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. Mol Cell *58*, 722-728.

121. Jain, M., Nijhawan, A., Tyagi, A.K., and Khurana, J.P. (2006). Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. Biochem Biophys Res Commun *345*, 646-651.

122. Jansen, R., Embden, J.D., Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol *43*, 1565-1575.

123. Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol *31*, 233-239.

124. 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.

125. 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.

126. Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife *2*, e00471.

127. Joung, J., Konermann, S., Gootenberg, J.S., Abudayyeh, O.O., Platt, R.J., Brigham, M.D., Sanjana, N.E., and Zhang, F. (2017). Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc *12*, 828-863.

128. Junker, J.P., Spanjaard, B., Peterson-Maduro, J., Alemany, A., Hu, B., Florescu, M., and van Oudenaarden, A. (2017). Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars. bioRxiv.

129. Kalhor, R., Mali, P., and Church, G.M. (2017). Rapidly evolving homing CRISPR barcodes. Nat Methods *14*, 195-200.

130. Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., *et al.* (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science (New York, NY) *316*, 1484-1488.

131. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol *30*, 772-780.

132. Kazlauskiene, M., Kostiuk, G., Venclovas, C., Tamulaitis, G., and Siksnys, V. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science *357*, 605-609.

133. Kearns, N.A., Pham, H., Tabak, B., Genga, R.M., Silverstein, N.J., Garber, M., and Maehr, R. (2015). Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nat Methods *12*, 401-403.

134. Khan, A.A., Betel, D., Miller, M.L., Sander, C., Leslie, C.S., and Marks, D.S. (2009). Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. Nat Biotechnol *27*, 549-555.

135. Kiani, S., Chavez, A., Tuttle, M., Hall, R.N., Chari, R., Ter-Ovanesyan, D., Qian, J., Pruitt, B.W., Beal, J., Vora, S., Buchthal, J., Kowal, E.J., Ebrahimkhani, M.R., Collins, J.J., Weiss, R., and Church, G. (2015). Cas9 gRNA engineering for genome editing, activation and repression. Nat Methods *12*, 1051-1054.

136. Kim, H., and Kim, J.S. (2014). A guide to genome engineering with programmable nucleases. Nat Rev Genet *15*, 321-334.

137. Kim, U., Wang, Y., Sanford, T., Zeng, Y., and Nishikura, K. (1994). Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. Proc Natl Acad Sci U S A *91*, 11457-11461.

138. Kim, Y.B., Komor, A.C., Levy, J.M., Packer, M.S., Zhao, K.T., and Liu, D.R. (2017). Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nat Biotechnol *35*, 371-376.

139. 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.

140. Komor, A.C., Badran, A.H., and Liu, D.R. (2017). CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. Cell *168*, 20-36.

141. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420-424.

142. Konermann, S., Brigham, M.D., Trevino, A., Hsu, P.D., Heidenreich, M., Cong, L., Platt, R.J., Scott, D.A., Church, G.M., and Zhang, F. (2013). Optical control of mammalian endogenous transcription and epigenetic states. Nature *500*, 472-476.

143. 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.

144. Konermann, S., Lotfy, P., Brideau, N.J., Oki, J., Shokhirev, M.N., and Hsu, P.D. (2018). Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. Cell.

145. Koonin, E.V., and Krupovic, M. (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat Rev Genet *16*, 184-192.

146. Koonin, E.V., and Makarova, K.S. (2009). CRISPR-Cas: an adaptive immunity system in prokaryotes. F1000 Biol Rep *1*, 95.

147. Koonin, E.V., and Makarova, K.S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol *10*, 679-686.

148. Kozlov, G., Denisov, A.Y., Girard, M., Dicaire, M.J., Hamlin, J., McPherson, P.S., Brais, B., and Gehring, K. (2011). Structural Basis of Defects in the Sacsin HEPN Domain Responsible for Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS). J Biol Chem *286*, 20407-20412.

149. Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E.V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biology *12*, 36.

150. Kumar, R.M., Cahan, P., Shalek, A.K., Satija, R., DaleyKeyser, A.J., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J.J., Ferrante, T.C., Regev, A., Daley, G.Q., and Collins, J.J. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature *516*, 56-61.

151. Kuttan, A., and Bass, B.L. (2012). Mechanistic insights into editing-site specificity of ADARs. Proc Natl Acad Sci U S A *109*, E3295-3304.

152. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

153. Lange, S.J., Alkhnbashi, O.S., Rose, D., Will, S., and Backofen, R. (2013). CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. Nucleic Acids Res *41*, 8034-8044.

154. Lehmann, K.A., and Bass, B.L. (2000). Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry *39*, 12875-12884.

155. Lehmann, R., and Nusslein-Volhard, C. (1986). Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of oskar, a maternal gene in Drosophila. Cell *47*, 141-152.

156. Lewis, C.J., Pan, T., and Kalsotra, A. (2017). RNA modifications and structures cooperate to guide RNA-protein interactions. Nat Rev Mol Cell Biol *18*, 202-210.

157. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

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

159. Li, X., Xiong, X., and Yi, C. (2016). Epitranscriptome sequencing technologies: decoding RNA modifications. Nat Methods *14*, 23-31.

160. Li, Y., Liao, W., Cargill, M., Chang, M., Matsunami, N., Feng, B.J., Poon, A., Callis-Duffin, K.P., Catanese, J.J., Bowcock, A.M., Leppert, M.F., Kwok, P.Y., Krueger, G.G., and Begovich, A.B. (2010). Carriers of rare missense variants in IFIH1 are protected from psoriasis. J Invest Dermatol *130*, 2768-2772.

161. Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst *1*, 417-425.

162. Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. Cell *167*, 233-247 e217.

163. 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.

164. Maeder, M.L., Linder, S.J., Cascio, V.M., Fu, Y., Ho, Q.H., and Joung, J.K. (2013). CRISPR RNA-guided activation of endogenous human genes. Nat Methods *10*, 977-979.

165. Majumdar, S., Zhao, P., Pfister, N.T., Compton, M., Olson, S., Glover, C.V., 3rd, Wells, L., Graveley, B.R., Terns, R.M., and Terns, M.P. (2015). Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. RNA *21*, 1147-1158.

166. 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.

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

168. 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.

169. 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. (2011b). Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol *9*, 467-477.

170. Makarova, K.S., and Koonin, E.V. (2015). Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol *1311*, 47-75.

171. 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., *et al.* (2015a). An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol *13*, 722-736.
172. 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., *et al.* (2015b). An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol.
173. Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2009). Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. Biol Direct *4*, 19.
174. 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.
175. Makarova, K.S., Zhang, F., and Koonin, E.V. (2017). SnapShot: Class 2 CRISPR-Cas Systems. Cell *168*, 328-328 e321.
176. Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013a). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol *31*, 833-838.
177. Mali, P., Esvelt, K.M., and Church, G.M. (2013b). Cas9 as a versatile tool for engineering biology. Nat Methods *10*, 957-963.
178. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013c). RNA-guided human genome engineering via Cas9. Science *339*, 823-826.
179. Mann, D.G., Lafayette, P.R., Abercrombie, L.L., King, Z.R., Mazarei, M., Halter, M.C., Poovaiah, C.R., Baxter, H., Shen, H., Dixon, R.A., Parrott, W.A., and Neal Stewart, C., Jr. (2012). Gateway-compatible vectors for high-throughput gene functional analysis in switchgrass (Panicum virgatum L.) and other monocot species. Plant Biotechnol J *10*, 226-236.
180. Mannack, L.V., Eising, S., and Rentmeister, A. (2016). Current techniques for visualizing RNA in cells. F1000Res *5*.
181. Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res *30*, 281-283.
182. Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Lu, S., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., and Bryant, S.H. (2013). CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res *41*, D348-352.
183. Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. Nature *526*, 55-61.
184. Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science *322*, 1843-1845.
185. Marraffini, L.A., and Sontheimer, E.J. (2010a). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet *11*, 181-190.
186. Marraffini, L.A., and Sontheimer, E.J. (2010b). Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature *463*, 568-571.
187. Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell *110*, 563-574.

188. Matthews, M.M., Thomas, J.M., Zheng, Y., Tran, K., Phelps, K.J., Scott, A.I., Havel, J., Fisher, A.J., and Beal, P.A. (2016). Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. Nat Struct Mol Biol *23*, 426-433.

189. McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

190. Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature *431*, 343-349.

191. Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P.H., and Higuchi, M. (1996). A mammalian RNA editing enzyme. Nature *379*, 460-464.

192. Merritt, W.M., Lin, Y.G., Han, L.Y., Kamat, A.A., Spannuth, W.A., Schmandt, R., Urbauer, D., Pennacchio, L.A., Cheng, J.F., Nick, A.M., *et al.* (2008). Dicer, Drosha, and outcomes in patients with ovarian cancer. N Engl J Med *359*, 2641-2650.

193. Miller, J.C., Holmes, M.C., Wang, J., Guschin, D.Y., Lee, Y.L., Rupniewski, I., Beausejour, C.M., Waite, A.J., Wang, N.S., Kim, K.A., Gregory, P.D., Pabo, C.O., and Rebar, E.J. (2007). An improved zinc-finger nuclease architecture for highly specific genome editing. Nature biotechnology *25*, 778-785.

194. 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., *et al.* (2011). A TALE nuclease architecture for efficient genome editing. Nature biotechnology *29*, 143-148.

195. 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.

196. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733-740.

197. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol *60*, 174-182.

198. Montiel-Gonzalez, M.F., Vallecillo-Viejo, I., Yudowski, G.A., and Rosenthal, J.J. (2013). Correction of mutations within the cystic fibrosis transmembrane conductance regulator by site-directed RNA editing. Proc Natl Acad Sci U S A *110*, 18285-18290.

199. Montiel-Gonzalez, M.F., Vallecillo-Viejo, I.C., and Rosenthal, J.J. (2016). An efficient system for selectively altering genetic information within mRNAs. Nucleic Acids Res *44*, e157.

200. Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R., and Schaffer, A.A. (2008). Database indexing for production MegaBLAST searches. Bioinformatics *24*, 1757-1764.

201. Moscou, M.J., and Bogdanove, A.J. (2009). A simple cipher governs DNA recognition by TAL effectors. Science *326*, 1501.

202. Mulepati, S., Heroux, A., and Bailey, S. (2014). Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science *345*, 1479-1484.

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

204. Newman, A.M., Bratman, S.V., To, J., Wynne, J.F., Eclov, N.C., Modlin, L.A., Liu, C.L., Neal, J.W., Wakelee, H.A., Merritt, R.E., Shrager, J.B., Loo, B.W., Jr., Alizadeh, A.A., and Diehn, M. (2014). An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med *20*, 548-554.
205. Niemz, A., Ferguson, T.M., and Boyle, D.S. (2011). Point-of-care nucleic acid testing for infectious diseases. Trends Biotechnol *29*, 240-250.
206. Niewoehner, O., Garcia-Doval, C., Rostol, J.T., Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini, L.A., and Jinek, M. (2017). Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature *548*, 543-548.
207. 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.
208. Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M., Mochizuki, M., Miyabe, A., Araki, M., Hara, K.Y., Shimatani, Z., and Kondo, A. (2016). Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science *353*.
209. Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem *79*, 321-349.
210. 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.
211. Norais, C., Moisan, A., Gaspin, C., and Clouet-d'Orval, B. (2013). Diversity of CRISPR systems in the euryarchaeal Pyrococcales. RNA Biol *10*, 659-670.
212. Nunez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nature structural & molecular biology *21*, 528-534.
213. Nunez, J.K., Lee, A.S., Engelman, A., and Doudna, J.A. (2015). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature.
214. O'Connell, M.R., Oakes, B.L., Sternberg, S.H., East-Seletsky, A., Kaplan, M., and Doudna, J.A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature *516*, 263-266.
215. Organization, W.H. (2009). In Guidelines for Using HIV Testing Technologies in Surveillance: Selection, Evaluation and Implementation: 2009 Update (Geneva).
216. 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.
217. Ozawa, T., Natori, Y., Sato, M., and Umezawa, Y. (2007). Imaging dynamics of endogenous mitochondrial RNA in single living cells. Nat Methods *4*, 413-419.
218. Palacino, J., Swalley, S.E., Song, C., Cheung, A.K., Shu, L., Zhang, X., Van Hoosear, M., Shin, Y., Chin, D.N., Keller, C.G., *et al.* (2015). SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. Nature chemical biology *11*, 511-517.
219. Pardee, K., Green, A.A., Ferrante, T., Cameron, D.E., DaleyKeyser, A., Yin, P., and Collins, J.J. (2014). Paper-based synthetic gene networks. Cell *159*, 940-954.
220. Pardee, K., Green, A.A., Takahashi, M.K., Braff, D., Lambert, G., Lee, J.W., Ferrante, T., Ma, D., Donghia, N., Fan, M., Daringer, N.M., Bosch, I., Dudley, D.M., O'Connor, D.H., Gehrke, L., and Collins, J.J. (2016). Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. Cell *165*, 1255-1266.

221. Park, H.Y., Buxbaum, A.R., and Singer, R.H. (2010). Single mRNA tracking in live cells. Methods Enzymol *472*, 387-406.

222. Park, H.Y., Lim, H., Yoon, Y.J., Follenzi, A., Nwokafor, C., Lopez-Jones, M., Meng, X., and Singer, R.H. (2014). Visualization of dynamics of single endogenous mRNA labeled in live mouse. Science *343*, 422-424.

223. Paz-Bailey, G., Rosenberg, E.S., Doyle, K., Munoz-Jordan, J., Santiago, G.A., Klein, L., Perez-Padilla, J., Medina, F.A., Waterman, S.H., Gubern, C.G., Alvarado, L.I., and Sharp, T.M. (2017). Persistence of Zika Virus in Body Fluids - Preliminary Report. N Engl J Med.

224. Pecot, C.V., Calin, G.A., Coleman, R.L., Lopez-Berestein, G., and Sood, A.K. (2011). RNA interference in the clinic: challenges and future directions. Nat Rev Cancer *11*, 59-67.

225. Peng, W., Feng, M., Feng, X., Liang, Y.X., and She, Q. (2015). An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. Nucleic Acids Res *43*, 406-417.

226. Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., Guilak, F., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Methods *10*, 973-976.

227. Picardi, E., D'Erchia, A.M., Montalvo, A., and Pesole, G. (2015). Using REDItools to Detect RNA Editing Events in NGS Datasets. Curr Protoc Bioinformatics *49*, 12 12 11-15.

228. Picardi, E., and Pesole, G. (2013). REDItools: high-throughput RNA editing detection made easy. Bioinformatics *29*, 1813-1814.

229. Piepenburg, O., Williams, C.H., Stemple, D.L., and Armes, N.A. (2006). DNA detection using recombination proteins. PLoS Biol *4*, e204.

230. Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A., and Panning, B. (2002). Xist RNA and the mechanism of X chromosome inactivation. Annual review of genetics *36*, 233-278.

231. Pourcel, C., Salvignol, G., and Vergnaud, G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology *151*, 653-663.

232. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

233. Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173-1183.

234. Qin, Z., Ljubimov, V.A., Zhou, C., Tong, Y., and Liang, J. (2016). Cell-free circulating tumor DNA in cancer. Chin J Cancer *35*, 36.

235. Rath, S., Donovan, J., Whitney, G., Chitrakar, A., Wang, W., and Korennykh, A. (2015). Human RNase L tunes gene expression by selectively destabilizing the microRNA-regulated transcriptome. Proc Natl Acad Sci U S A *112*, 15916-15921.

236. Root, D.E., Hacohen, N., Hahn, W.C., Lander, E.S., and Sabatini, D.M. (2006). Genome-scale loss-of-function screening with a lentiviral RNAi library. Nat Methods *3*, 715-719.

237. Rousseau, B.A., Hou, Z., Gramelspacher, M.J., and Zhang, Y. (2018). Programmable RNA Cleavage and Recognition by a Natural CRISPR-Cas9 System from Neisseria meningitidis. Mol Cell *69*, 906-914 e904.

238. Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., and Marraffini, L.A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell *161*, 1164-1174.
239. Sampson, T.R., Saroj, S.D., Llewellyn, A.C., Tzeng, Y.L., and Weiss, D.S. (2013). A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. Nature *497*, 254-257.
240. 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.
241. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. Nat Methods *9*, 676-682.
242. 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.
243. Schwartz, S. (2016). Cracking the epitranscriptome. RNA *22*, 169-174.
244. Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84-87.
245. 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.
246. 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.
247. 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. (2017a). Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol *15*, 169-182.
248. 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. (2017b). Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol.
249. Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A.M., and Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science *351*, aad4234.
250. Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. EMBO J *32*, 385-394.
251. Smargon, A.A., Cox, D.B., 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. (2017a). Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. Mol Cell.
252. Smargon, A.A., Cox, D.B., 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. (2017b). Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. Mol Cell *65*, 618-630 e617.

253. Smith, J., Grizot, S., Arnould, S., Duclert, A., Epinat, J.C., Chames, P., Prieto, J., Redondo, P., Blanco, F.J., Bravo, J., Montoya, G., Paques, F., and Duchateau, P. (2006). A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. Nucleic acids research *34*, e149.

254. Soding, J., Remmert, M., Biegert, A., and Lupas, A.N. (2006). HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res *34*, W374-378.

255. Song, L., Shan, D., Zhao, M., Pink, B.A., Minnehan, K.A., York, L., Gardel, M., Sullivan, S., Phillips, A.F., Hayman, R.B., Walt, D.R., and Duffy, D.C. (2013). Direct detection of bacterial genomic DNA at sub-femtomolar concentrations using single molecule arrays. Anal Chem *85*, 1932-1939.

256. 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., 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.

257. Staals, R.H., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K., *et al.* (2014). RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Mol Cell *56*, 518-530.

258. Stafforst, T., and Schneider, M.F. (2012). An RNA-deaminase conjugate selectively repairs point mutations. Angew Chem Int Ed Engl *51*, 11166-11169.

259. Stoddard, B.L. (2005). Homing endonuclease structure and function. Quarterly reviews of biophysics *38*, 49-95.

260. Strutt, S.C., Torrez, R.M., Kaya, E., Negrete, O.A., and Doudna, J.A. (2018). RNA-dependent RNA targeting by CRISPR-Cas9. Elife *7*.

261. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

262. 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.

263. 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.

264. 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.

265. Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., *et al.* (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature *550*, 249-254.

266. Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell *159*, 635-646.

267. Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Huttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc Natl Acad Sci U S A *99*, 7536-7541.

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

269. Tilsner, J., Linnik, O., Christensen, N.M., Bell, K., Roberts, I.M., Lacomme, C., and Oparka, K.J. (2009). Live-cell imaging of viral RNA genomes using a Pumilio-based reporter. Plant J *57*, 758-770.

270. Tourriere, H., Chebli, K., Zekri, L., Courselaud, B., Blanchard, J.M., Bertrand, E., and Tazi, J. (2003). The RasGAP-associated endoribonuclease G3BP assembles stress granules. J Cell Biol *160*, 823-831.

271. Tyagi, S. (2009). Imaging intracellular RNA distribution and dynamics in living cells. Nat Methods *6*, 331-338.

272. Unsworth, H., Raguz, S., Edwards, H.J., Higgins, C.F., and Yague, E. (2010). mRNA escape from stress granule sequestration is dictated by localization to the endoplasmic reticulum. FASEB J *24*, 3370-3380.

273. Urdea, M., Penny, L.A., Olmsted, S.S., Giovanni, M.Y., Kaspar, P., Shepherd, A., Wilson, P., Dahl, C.A., Buchsbaum, S., Moeller, G., and Hay Burgess, D.C. (2006). Requirements for high impact diagnostics in the developing world. Nature *444 Suppl 1*, 73-79.

274. Urnov, F., Miller, J., Lee, Y.-L., Beausejour, C., Rock, J., Augustus, S., Jamieson, A., Porteus, M., Gregory, P., and Holmes, M. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature *435*, 646-651.

275. Urnov, F., Rebar, E., Holmes, M., Zhang, H., and Gregory, P. (2010). Genome editing with engineered zinc finger nucleases. Nature reviews Genetics *11*, 636-646.

276. Valdmanis, P.N., Gu, S., Chu, K., Jin, L., Zhang, F., Munding, E.M., Zhang, Y., Huang, Y., Kutay, H., Ghoshal, K., Lisowski, L., and Kay, M.A. (2016). RNA interference-induced hepatotoxicity results from loss of the first synthesized isoform of microRNA-122 in mice. Nat Med *22*, 557-562.

277. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci *34*, 401-407.

278. Vojta, A., Dobrinic, P., Tadic, V., Bockor, L., Korac, P., Julg, B., Klasic, M., and Zoldos, V. (2016). Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Res *44*, 5615-5628.

279. Wagner, R.W. (1994). Gene inhibition using antisense oligodeoxynucleotides. Nature *372*, 333-335.

280. Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80-84.

281. Wang, X., McLachlan, J., Zamore, P.D., and Hall, T.M. (2002). Modular recognition of RNA by a human pumilio-homology domain. Cell *110*, 501-512.

282. Wang, Y., Cheong, C.G., Hall, T.M., and Wang, Z. (2009). Engineering splicing factors with designed specificities. Nat Methods *6*, 825-830.

283. Wang, Y., Havel, J., and Beal, P.A. (2015). A Phenotypic Screen for Functional Mutants of Human Adenosine Deaminase Acting on RNA 1. ACS Chem Biol *10*, 2512-2519.

284. Watson, J.D. (2014). Molecular biology of the gene, Seventh edition edn (Boston: Pearson).

285. Wettengel, J., Reautschnig, P., Geisler, S., Kahle, P.J., and Stafforst, T. (2017). Harnessing human ADAR2 for RNA repair - Recoding a PINK1 mutation rescues mitophagy. Nucleic Acids Res *45*, 2797-2808.

286. Wong, S.K., Sato, S., and Lazinski, D.W. (2001). Substrate recognition by ADAR1 and ADAR2. RNA *7*, 846-858.

287. Wright, A.V., Nunez, J.K., and Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29-44.

288. Wright, A.V., Sternberg, S.H., Taylor, D.W., Staahl, B.T., Bardales, J.A., Kornfeld, J.E., and Doudna, J.A. (2015). Rational design of a split-Cas9 enzyme complex. Proc Natl Acad Sci U S A *112*, 2984-2989.

289. Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., Jaenisch, R., Zhang, F., and Sharp, P.A. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nature biotechnology *32*, 670-676.

290. Xu, X., Tao, Y., Gao, X., Zhang, L., Li, X., Zou, W., Ruan, K., Wang, F., Xu, G.L., and Hu, R. (2016). A CRISPR-based approach for targeted DNA demethylation. Cell Discov *2*, 16009.

291. Yagi, Y., Hayashi, S., Kobayashi, K., Hirayama, T., and Nakamura, T. (2013). Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. PLoS One *8*, e57286.

292. Yamano, T., 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.

293. Yan, W.X., Chong, S., Zhang, H., Makarova, K.S., Koonin, E.V., Cheng, D.R., and Scott, D.A. (2018). Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. Mol Cell.

294. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*, 134.

295. Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res *40*, 5569-5576.

296. Zetsche, B., Gootenberg, J., Abudayyeh, O., Slaymaker, I., Makarova, K., Volz, S., Joung, J., Essletzbichler, P., Van der Oost, J., Regev, A., Koonin, E., and Zhang, F. (2015a). Cpf1 is a single RNA-guided endonuclease of a novel Class 2 CRISPR-Cas system. Cell *in press*.

297. 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. (2015b). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759-771.

298. Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E.M., Winblad, N., Choudhury, S.R., Abudayyeh, O.O., Gootenberg, J.S., Wu, W.Y., Scott, D.A., Severinov, K., van der Oost, J., and Zhang, F. (2017). Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. Nat Biotechnol *35*, 31-34.

299. Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G.M., and Arlotta, P. (2011a). Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat Biotechnol *29*, 149-153.

300. Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., Albers, S.V., Naismith, J.H., Spagnolo, L., and White, M.F. (2012). Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell *45*, 303-313.

301. Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol Cell *50*, 488-503.

302. Zhang, Y., Su, J., Duan, S., Ao, Y., Dai, J., Liu, J., Wang, P., Li, Y., Liu, B., Feng, D., Wang, J., and Wang, H. (2011b). A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. Plant Methods *7*, 30.

303. Zhao, W., Ali, M.M., Aguirre, S.D., Brook, M.A., and Li, Y. (2008a). Paper-based bioassays using gold nanoparticle colorimetric probes. Anal Chem *80*, 8431-8437.

304. Zhao, W., Lam, J.C., Chiuman, W., Brook, M.A., and Li, Y. (2008b). Enzymatic cleavage of nucleic acids on gold nanoparticles: a generic platform for facile colorimetric biosensors. Small *4*, 810-816.

305. Zheng, Y., Lorenzo, C., and Beal, P.A. (2017). DNA editing in DNA/RNA hybrids by adenosine deaminases that act on RNA. Nucleic Acids Res *45*, 3369-3377.

306. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res *31*, 3406-3415.

Appendix B

Supplementary Figures

10.1 Chapter 2 Supplementary Figures

10.1.1 Figure 2.S1

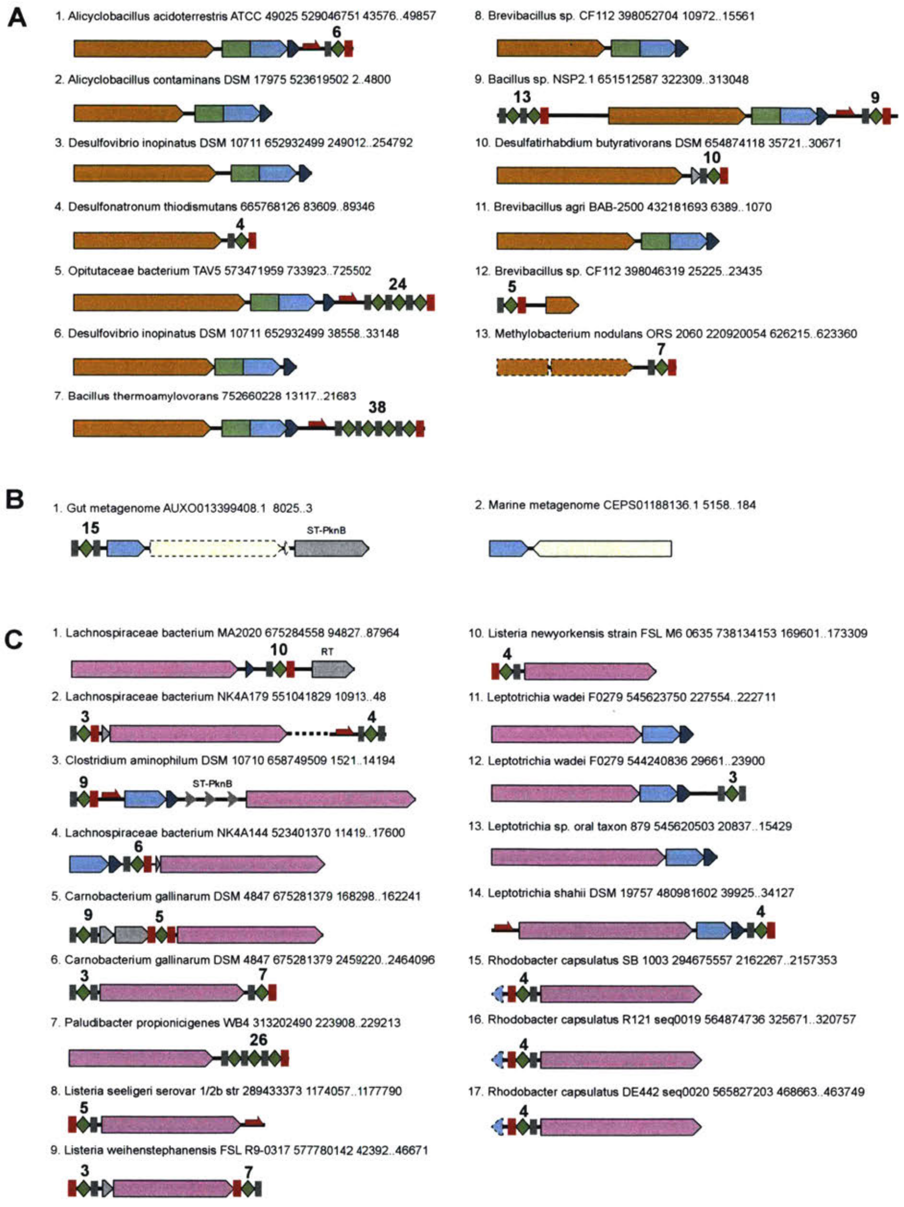

A
1. Alicyclobacillus acidoterrestris ATCC 49025 529046751 43576..49857
6
2. Alicyclobacillus contaminans DSM 17975 523619502 2..4800
3. Desulfovibrio inopinatus DSM 10711 652932499 249012..254792
4. Desulfonatronum thiodismutans 665768126 83609..89346
4
5. Opitutaceae bacterium TAV5 573471959 733923..725502
24
6. Desulfovibrio inopinatus DSM 10711 652932499 38558..33148
7. Bacillus thermoamylovorans 752660228 13117..21683
38
8. Brevibacillus sp. CF112 398052704 10972..15561
9. Bacillus sp. NSP2.1 651512587 322309..313048
13
9
10. Desulfatirhabdium butyrativorans DSM 654874118 35721..30671
10
11. Brevibacillus agri BAB-2500 432181693 6389..1070
12. Brevibacillus sp. CF112 398046319 25225..23435
5
13. Methylobacterium nodulans ORS 2060 220920054 626215..623360
7
B
1. Gut metagenome AUXO013399408.1 8025..3
15
ST-PknB
2. Marine metagenome CEPS01188136.1 5158..184
C
1. Lachnospiraceae bacterium MA2020 675284558 94827..87964
10
RT
2. Lachnospiraceae bacterium NK4A179 551041829 10913..48
3
4
3. Clostridium aminophilum DSM 10710 658749509 1521..14194
9
ST-PknB
4. Lachnospiraceae bacterium NK4A144 523401370 11419..17600
6
5. Carnobacterium gallinarum DSM 4847 675281379 168298..162241
9
5
6. Carnobacterium gallinarum DSM 4847 675281379 2459220..2464096
3
7
7. Paludibacter propionicigenes WB4 313202490 223908..229213
26
8. Listeria seeligeri serovar 1/2b str 289433373 1174057..1177790
5
9. Listeria weihenstephanensis FSL R9-0317 577780142 42392..46671
3
7
10. Listeria newyorkensis strain FSL M6 0635 738134153 169601..173309
4
11. Leptotrichia wadei F0279 545623750 227554..222711
12. Leptotrichia wadei F0279 544240836 29661..23900
3
13. Leptotrichia sp. oral taxon 879 545620503 20837..15429
14. Leptotrichia shahii DSM 19757 480981602 39925..34127
4
15. Rhodobacter capsulatus SB 1003 294675557 2162267..2157353
4
16. Rhodobacter capsulatus R121 seq0019 564874736 325671..320757
4
17. Rhodobacter capsulatus DE442 seq0020 565827203 468663..463749
4

Figure 2.S1: Genomic architectures of all identified candidate novel CRISPR-Cas loci.

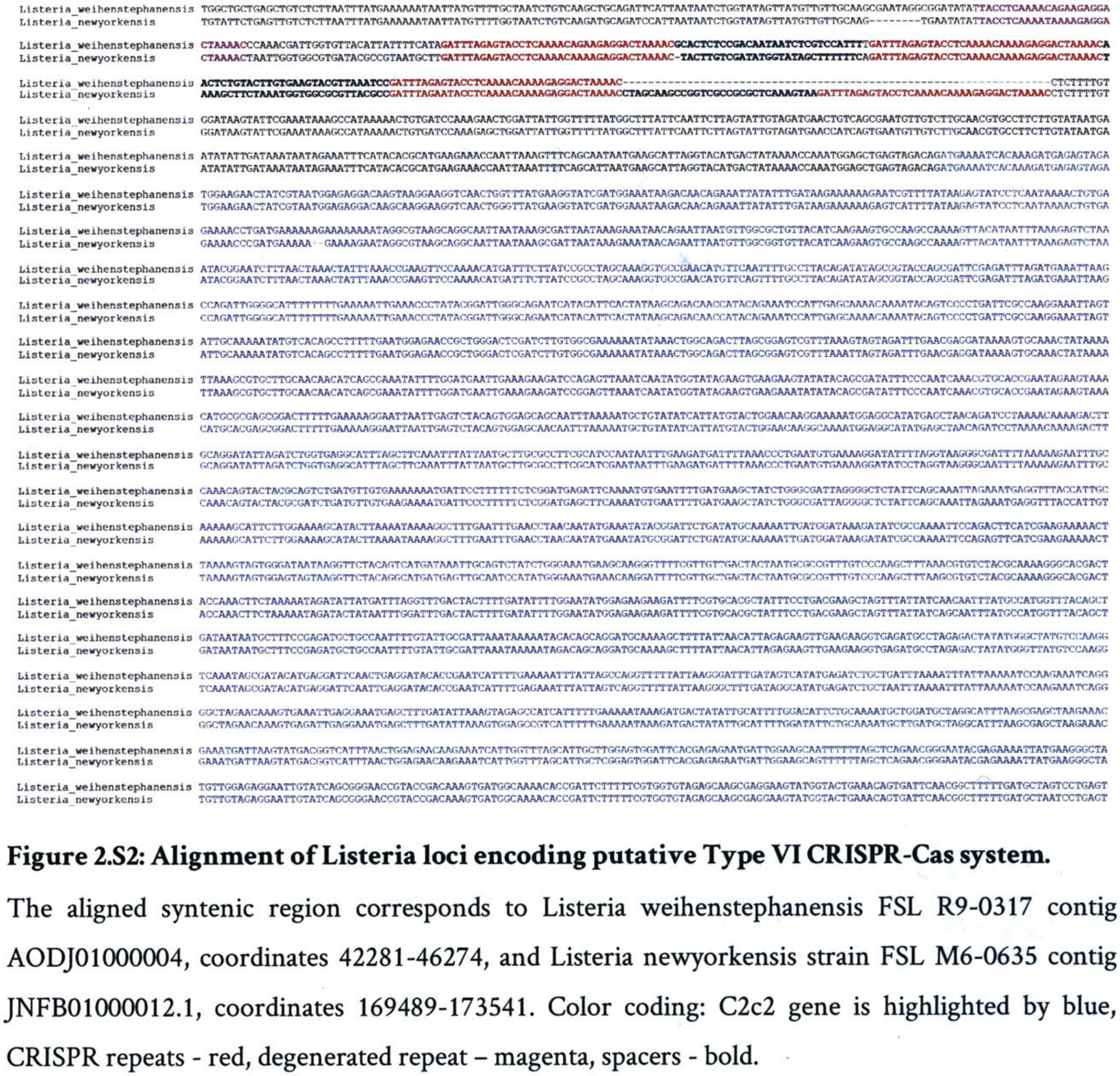
(A) The C2c1 loci (subtype V-B).

(B) The C2c3 loci (subtype V-C)

(C) The C2c2 loci (type VI).

Genes are colored according to the schematic in Figure 1. The number of repeats in CRISPR arrays is indicated. For each genomic contig, Genbank numeric ID and the coordinates of the locus are indicated. Additional designations: ST-PknB, serine/threonine protein kinase; RT, reverse transcriptase.

10.1.2 Figure 2.S2



```
Listeria_weihenstephanensis TGGCTGCTGAGCTGTCTCTTAATTTATGAAAAAATAATTATGTTTTGCTAATCTGTCAAGCTGCAGATTCATTAATAATCTGGTATAGTTATGTTGTTGCAAGCGAATAGGCGGATATATTACCTCAAAACAGAAGAGGA
Listeria_newyorkensis       TGTATTCTGAGTTGTCTCTTAATTTATGAAAAAATAATTATGTTTTGGTAATCTGTCAAGATGCAGATCCATTAATAATCTGGTATAGTTATGTTGTTGCAAG--------TGAATATATTACCTCAAAATAAAAGAGGA

Listeria_weihenstephanensis CTAAAACCCAAACGATTGGTGTTACATTATTTTCATAGATTTAGAGTACCTCAAAACAGAAGAGGACTAAAACGCACTCTCCGACAATAATCTCGTCCATTTTGATTTAGAGTACCTCAAAACAAAAGAGGACTAAAACA
Listeria_newyorkensis       CTAAAACTAATTGGTGGCGTGATACGCCGTAATGCTTGATTTAGAGTACCTCAAAACAAAAGAGGACTAAAAC-TACTTGTCGATATGGTATAGCTTTTTTCAGATTTAGAGTACCTCAAAACAAAAGAGGACTAAAACT

Listeria_weihenstephanensis ACTCTGTACTTGTGAAGTACGTTAAATCCGATTTAGAGTACCTCAAAACAAAAGAGGACTAAAAC------------------------------------------------------------------CTCTTTTGT
Listeria_newyorkensis       AAAGCTTCTAAATGGTGGCGCGTTACGCCGATTTAGAATACCTCAAAACAAAAGAGGACTAAAACCTAGCAAGCCGGTCGCCGCGCTCAAAGTAAGATTTAGAGTACCTCAAAACAAAAGAGGACTAAAACCTCTTTTGT

Listeria_weihenstephanensis GGATAAGTATTCGAAATAAAGCCATAAAAACTGTGATCCAAAGAACTGGATTATTGGTTTTTATGGCTTTATTCAATTCTTAGTATTGTAGATGAACTGTCAGCGAATGTTGTCTTGCAACGTGCCTTCTTGTATAATGA
Listeria_newyorkensis       GGATAAGTATTCGAAATAAAGCCATAAAAACTGTGATCCAAAGAGCTGGATTATTGGTTTTTATGGCTTTATTCAATTCTTAGTATTGTAGATGAACCATCAGTGAATGTTGTCTTGCAACGTGCCTTCTTGTATAATGA

Listeria_weihenstephanensis ATATATTGATAAATAATAGAAATTTCATACACGCATGAAGAAACCAATTAAAGTTTCAGCAATAATGAAGCATTAGGTACATGACTATAAAACCAAATGGAGCTGAGTAGACAGATGAAAATCACAAAGATGAGAGTAGA
Listeria_newyorkensis       ATATATTGATAAATAATAGAAATTTCATACACGCATGAAGAAACCAATTAAATTTTCAGCATTAATGAAGCATTAGGTACATGACTATAAAACCAAATGGAGCTGAGTAGACAGATGAAAATCACAAAGATGAGAGTAGA

Listeria_weihenstephanensis TGGAAGAACTATCGTAATGGAGAGGACAAGTAAGGAAGGTCAACTGGTTTATGAAGGTATCGATGGAAATAAGACAACAGAAATTATATTTGATAAGAAAAAAGAATCGTTTTATAAGAGTATCCTCAATAAAACTGTGA
Listeria_newyorkensis       TGGAAGAACTATCGTAATGGAGAGGACAAGCAAGGAAGGTCAACTGGGTTATGAAGGTATCGATGGAAATAAGACAACAGAAATTATATTTGATAAGAAAAAAGAGTCATTTTATAAGAGTATCCTCAATAAAACTGTGA

Listeria_weihenstephanensis GAAAACCTGATGAAAAAAGAAAAAAAATAGGCGTAAGCAGGCAATTAATAAAGCGATTAATAAAGAAATAACAGAATTAATGTTGGCGCTGTTACATCAAGAAGTGCCAAGCCAAAAGTTACATAATTTAAAGAGTCTAA
Listeria_newyorkensis       GAAAACCCGATGAAAAA--GAAAAGAATAGGCGTAAGCAGGCAATTAATAAAGCGATTAATAAAGAAATAACAGAATTAATGTTGGCGGTGTTACATCAAGAAGTGCCAAGCCAAAAGTTACATAATTTAAAGAGTCTAA

Listeria_weihenstephanensis ATACGGAATCTTTAACTAAACTATTTAAACCGAAGTTCCAAAACATGATTTCTTATCCGCCTAGCAAAGGTGCCGAACATGTTCAATTTTGCCTTACAGATATAGCGGTACCAGCGATTCGAGATTTAGATGAAATTAAG
Listeria_newyorkensis       ATACGGAATCTTTAACTAAACTATTTAAACCGAAGTTCCAAAACATGATTTCTTATCCGCCTAGCAAAGGTGCCGAACATGTTCAGTTTTGCCTTACAGATATAGCGGTACCAGCGATTCGAGATTTAGATGAAATTAAG

Listeria_weihenstephanensis CCAGATTGGGGCATTTTTTTTGAAAAATTGAAACCCTATACGGATTGGGCAGAATCATACATTCACTATAAGCAGACAACCATACAGAAATCCATTGAGCAAAACAAAATACAGTCCCCTGATTCGCCAAGGAAATTAGT
Listeria_newyorkensis       CCAGATTGGGGCATTTTTTTTGAAAAATTGAAACCCTATACGGATTGGGCAGAATCATACATTCACTATAAGCAGACAACCATACAGAAATCCATTGAGCAAAACAAAATACAGTCCCCTGATTCGCCAAGGAAATTAGT

Listeria_weihenstephanensis ATTGCAAAAATATGTCACAGCCTTTTTGAATGGAGAACCGCTGGGACTCGATCTTGTGGCGAAAAAATATAAACTGGCAGACTTAGCGGAGTCGTTTAAAGTAGTAGATTTGAACGAGGATAAAAGTGCAAACTATAAAA
Listeria_newyorkensis       ATTGCAAAAATATGTCACAGCCTTTTTGAATGGAGAACCGCTGGGACTCGATCTTGTGGCGAAAAAATATAAACTGGCAGACTTAGCGGAGTCGTTTAAATTAGTAGATTTGAACGAGGATAAAAGTGCAAACTATAAAA

Listeria_weihenstephanensis TTAAAGCGTGCTTGCAACAACATCAGCGAAATATTTTGGATGAATTGAAAGAAGATCCAGAGTTAAATCAATATGGTATAGAAGTGAAGAAGTATATACAGCGATATTTCCCAATCAAACGTGCACCGAATAGAAGTAAA
Listeria_newyorkensis       TTAAAGCGTGCTTGCAACAACATCAGCGAAATATTTTGGATGAATTGAAAGAAGATCCGGAGTTAAATCAATATGGTATAGAAGTGAAGAAATATATACAGCGATATTTCCCAATCAAACGTGCACCGAATAGAAGTAAA

Listeria_weihenstephanensis CATGCGCGAGCGGACTTTTTGAAAAAGGAATTAATTGAGTCTACAGTGGAGCAGCAATTTAAAAATGCTGTATATCATTATGTACTGGAACAAGGAAAAATGGAGGCATATGAGCTAACAGATCCTAAAACAAAAGACTT
Listeria_newyorkensis       CATGCACGAGCGGACTTTTTGAAAAAGGAATTAATTGAGTCTACAGTGGAGCAACAATTTAAAAATGCTGTATATCATTATGTACTGGAACAAGGCAAAATGGAGGCATATGAGCTAACAGATCCTAAAACAAAAGACTT

Listeria_weihenstephanensis GCAGGATATTAGATCTGGTGAGGCATTTAGCTTCAAATTTATTAATGCTTGCGCCTTCGCATCCAATAATTTGAAGATGATTTTAAACCCTGAATGTGAAAAGGATATTTTAGGTAAGGGCGATTTTAAAAAGAATTTGC
Listeria_newyorkensis       GCAGGATATTAGATCTGGTGAGGCATTTAGCTTCAAATTTATTAATGCTTGCGCCTTCGCATCGAATAATTTGAAGATGATTTTAAACCCTGAATGTGAAAAGGATATCCTAGGTAAGGGCAATTTTAAAAAGAATTTGC

Listeria_weihenstephanensis CAAACAGTACTACGCAGTCTGATGTTGTGAAAAAAATGATTCCTTTTTTCTCGGATGAGATTCAAAATGTGAATTTTGATGAAGCTATCTGGGCGATTAGGGGCTCTATTCAGCAAATTAGAAATGAGGTTTACCATTGC
Listeria_newyorkensis       CAAACAGTACTACGCGATCTGATGTTGTGAAGAAAATGATTCCCTTTTTCTCGGATGAGCTTCAAAATGTGAATTTTGATGAAGCTATCTGGGCGATTAGGGGCTCTATTCAGCAAATTAGAAATGAGGTTTACCATTGT

Listeria_weihenstephanensis AAAAAGCATTCTTGGAAAAGCATACTTAAAATAAAAGGCTTTGAATTTGAACCTAACAATATGAAATATACGGATTCTGATATGCAAAAATTGATGGATAAAGATATCGCCAAAATTCCAGACTTCATCGAAGAAAAACT
Listeria_newyorkensis       AAAAAGCATTCTTGGAAAAGCATACTTAAAATAAAAGGCTTTGAATTTGAACCTAACAATATGAAATATGCGGATTCTGATATGCAAAAATTGATGGATAAAGATATCGCCAAAATTCCAGAGTTCATCGAAGAAAAACT

Listeria_weihenstephanensis TAAAAGTAGTGGGATAATAAGGTTCTACAGTCATGATAAATTGCAGTCTATCTGGGAAATGAAGCAAGGGTTTTCGTTGTTGACTACTAATGCGCCGTTTGTCCCAAGCTTTAAACGTGTCTACGCAAAAGGGCACGACT
Listeria_newyorkensis       TAAAAGTAGTGGAGTAGTAAGGTTCTACAGGCATGATGAGTTGCAATCCATATGGGAAATGAAACAAGGATTTTCGTTGCTGACTACTAATGCGCCGTTTGTCCCAAGCTTTAAGCGTGTCTACGCAAAAGGGCACGACT

Listeria_weihenstephanensis ACCAAACTTCTAAAAATAGATATTATGATTTAGGTTTGACTACTTTTGATATTTTGGAATATGGAGAAGAAGATTTTCGTGCACGCTATTTCCTGACGAAGCTAGTTTATTATCAACAATTTATGCCATGGTTTACAGCT
Listeria_newyorkensis       ACCAAACTTCTAAAAATAGATACTATAATTTGGATTTGACTACTTTTGATATTTTGGAATATGGAGAAGAAGATTTTCGTGCACGCTATTTCCTGACGAAGCTAGTTTATTATCAGCAATTTATGCCATGGTTTACAGCT

Listeria_weihenstephanensis GATAATAATGCTTTCCGAGATGCTGCCAATTTTGTATTGCGATTAAATAAAAATACACAGCAGGATGCAAAAGCTTTTATTAACATTAGAGAAGTTGAAGAAGGTGAGATGCCTAGAGACTATATGGGCTATGTCCAAGG
Listeria_newyorkensis       GATAATAATGCTTTCCGAGATGCTGCCAATTTTGTATTGCGATTAAATAAAAATAGACAGCAGGATGCAAAAGCTTTTATTAACATTAGAGAAGTTGAAGAAGGTGAGATGCCTAGAGACTATATGGGTTATGTCCAAGG

Listeria_weihenstephanensis TCAAATAGCGATACATGAGGATTCAACTGAGGATACACCGAATCATTTTGAAAAATTTATTAGCCAGGTTTTTATTAAGGGATTTGATAGTCATATGAGATCTGCTGATTTAAAATTTATTAAAAATCCAAGAAATCAGG
Listeria_newyorkensis       TCAAATAGCGATACATGAGGATTCAATTGAGGATACACCGAATCATTTTGAGAAATTTATTAGTCAGGTTTTTATTAAGGGCTTTGATAGGCATATGAGATCTGCTAATTTAAAATTTATTAAAAATCCAAGAAATCAGG

Listeria_weihenstephanensis GGCTAGAACAAAGTGAAATTGAGGAAATGAGCTTTGATATTAAAGTAGAGCCATCATTTTTGAAAAATAAAGATGACTATATTGCATTTTGGACATTCTGCAAAATGCTGGATGCTAGGCATTTAAGCGAGCTAAGAAAC
Listeria_newyorkensis       GGCTAGAACAAAGTGAGATTGAGGAAATGAGCTTTGATATTAAAGTGGAGCCGTCATTTTTGAAAAATAAAGATGACTATATTGCATTTTGGATATTCTGCAAAATGCTTGATGCTAGGCATTTAAGCGAGCTAAGAAAC

Listeria_weihenstephanensis GAAATGATTAAGTATGACGGTCATTTAACTGGAGAACAAGAAATCATTGGTTTAGCATTGCTTGGAGTGGATTCACGAGAGAATGATTGGAAGCAATTTTTTAGCTCAGAACGGGAATACGAGAAAATTATGAAGGGCTA
Listeria_newyorkensis       GAAATGATTAAGTATGACGGTCATTTAACTGGAGAACAAGAAATCATTGGTTTAGCATTGCTCGGAGTGGATTCACGAGAGAATGATTGGAAGCAGTTTTTTAGCTCAGAACGGGAATACGAGAAAATTATGAAGGGCTA

Listeria_weihenstephanensis TGTTGGAGAGGAATTGTATCAGCGGGAACCGTACCGACAAAGTGATGGCAAAACACCGATTCTTTTTCGTGGTGTAGAGCAAGCGAGGAAGTATGGTACTGAAACAGTGATTCAACGGCTTTTTGATGCTAGTCCTGAGT
Listeria_newyorkensis       TGTTGTAGAGGAATTGTATCAGCGGGAACCGTACCGACAAAGTGATGGCAAAACACCGATTCTTTTTCGTGGTGTAGAGCAAGCGAGGAAGTATGGTACTGAAACAGTGATTCAACGGCTTTTTGATGCTAATCCTGAGT
```

Figure 2.S2: Alignment of Listeria loci encoding putative Type VI CRISPR-Cas system.

The aligned syntenic region corresponds to Listeria weihenstephanensis FSL R9-0317 contig AODJ01000004, coordinates 42281-46274, and Listeria newyorkensis strain FSL M6-0635 contig JNFB01000012.1, coordinates 169489-173541. Color coding: C2c2 gene is highlighted by blue, CRISPR repeats - red, degenerated repeat – magenta, spacers - bold.

10.1.3 Figure 2.S3

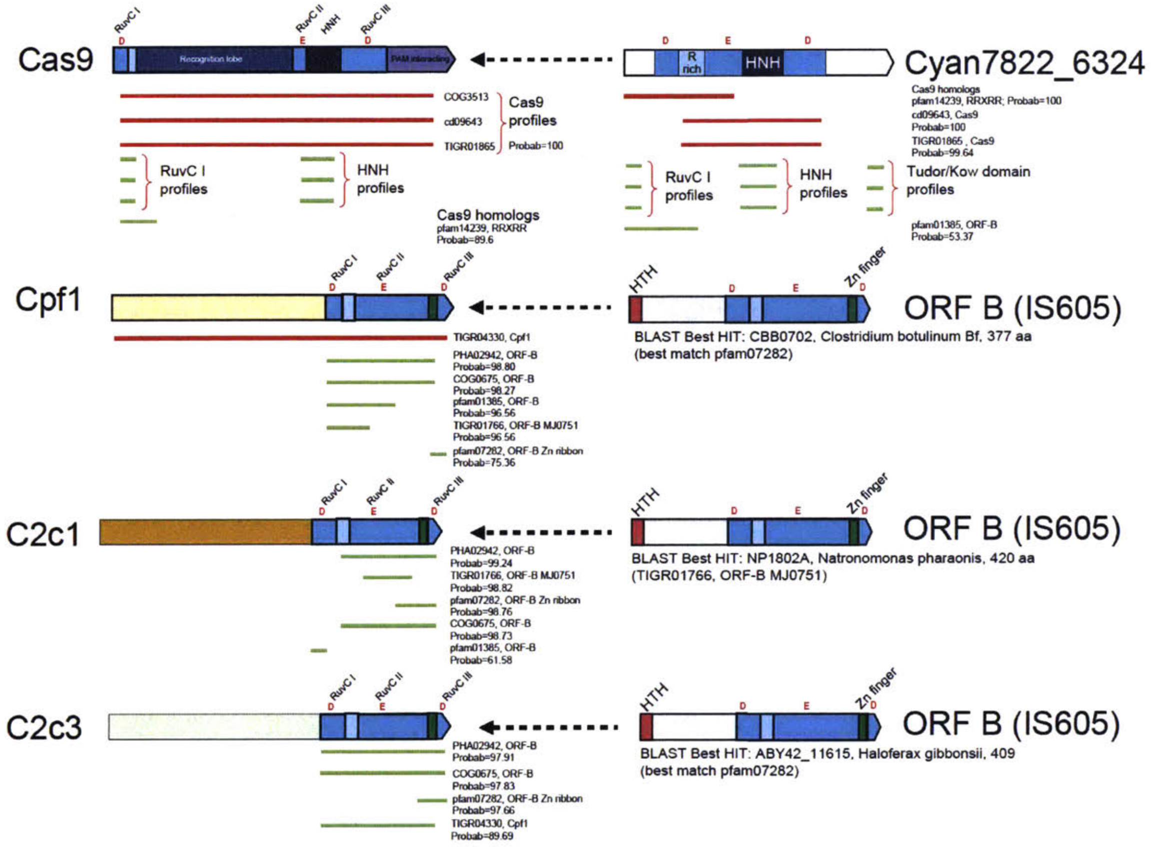


Figure 2.S3:The closest homologs of the new type V effector proteins among the transposon-encoded proteins: non-overlapping sets of homologs.

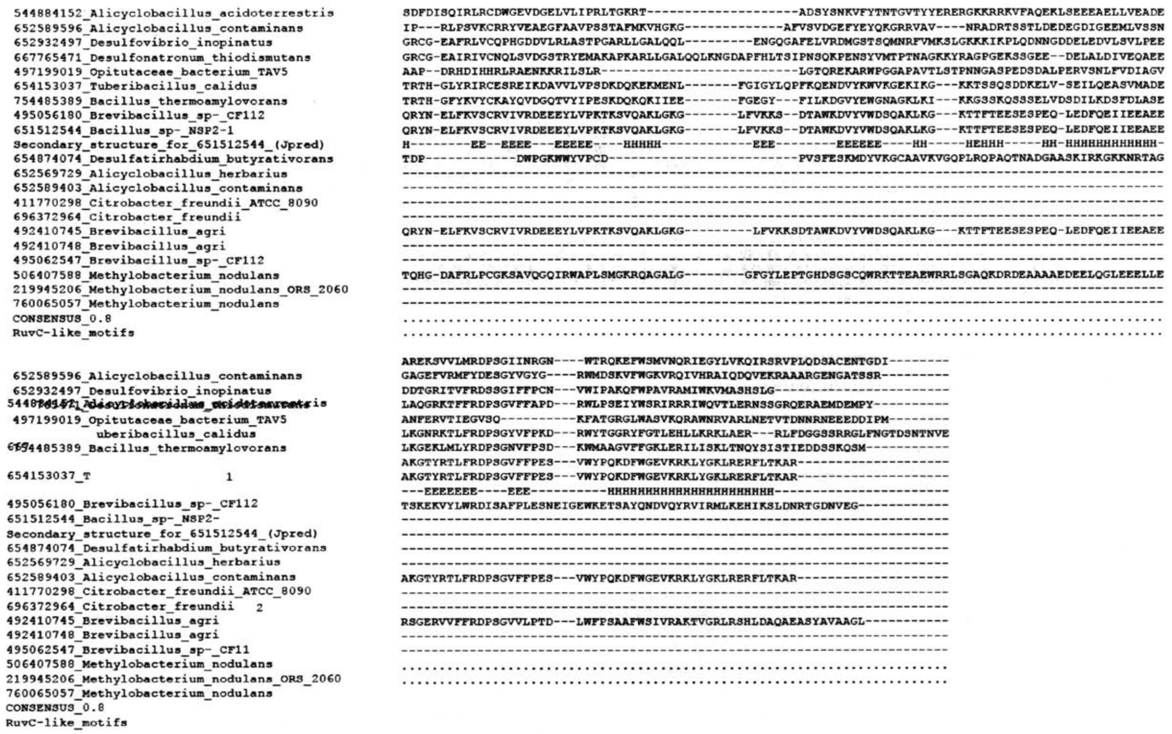
10.1.4 Figure 2.S4

```
544884152_Alicyclobacillus_acidoterrestris    ------------------------MAVKSIKVKLRLDDMPE---IRAG------LWKLHKEVNAGVRYYTEWLSLLRQEN---LYRRSPNGDGEQECDKT
652589596_Alicyclobacillus_contaminans        ----------------------------------------------------------------------------------------------------
652932497_Desulfovibrio_inopinatus            ------------------------MPTRTINLKLVLGKNPENATLRRA------LFSTHRLVNQATKRIEEFLLLCRGEA------YRTVDNEGKEAEIP
667765471_Desulfonatronum_thiodismutans       ---------------------------------MVLGRKDDTAELRRA------LWTTHEHVNLAVAEVERVLLRCRGRS---YWTLDR---RGDPVHVP
497199019_Opitutaceae_bacterium_TAV5          ------------------MSLNRIYQGRVAAVETGTALAKGNVEWMPAAGGDEVLWQHHELFQAAINYYLVALLALADKNNPVLGPLISQMDNPQSPYHV
654153037_Tuberibacillus_calidus              ------------------------MATKSFILKMKTKNNPQ---LRLS------LWKTHELFNFGVAYYMDLLSLFRQKD---LYMHNDE-DPDHPVVLK
754485389_Bacillus_thermoamylovorans          ------------------------MATRSFILKIEP--NEE---VKKG------LWKTHEVLNHGIAYYMNILKLIRQEA---IYEHHEQ-DPKNPKKVS
495056180_Brevibacillus_sp-_CF112             ----------------------------------------------------------------------------------------------------
651512544_Bacillus_sp-_NSP2-1                 ------------------------MAIRSIKLKLKTHTGPEAQNLRKG------IWRTHRLLNEGVAYYMKMLLLFRQES---TGERPKEELQEELICHI
Secondary_structure_for_651512544_(Jpred)     *---------------------------EEEEEEEE----HHHHHHHH------HHHHHHHHHHHHHHHHHHHHHHH-----------HHHHHHHHHHHH
654874074_Desulfatirhabdium_butyrativorans    ----------------MPLSNNPPVTQRAYTLRLR-GADPSDLSWREA------LWHTHEAVNKGAKVFGDWLLTLRGGLDHTLADTKVKGGKGKPDRDP
652569729_Alicyclobacillus_herbarius          MLKQAVLGNGPLINWEKNVKRGKGMATKSIKVKLRLGKHPD---IRAG------IWQLHKAANAGVRYYTEWLSLMRQKN---LYTRGP--KGEQQLYRS
652589403_Alicyclobacillus_contaminans        ------------------------MSVKSIKFKLMIG-GPQYTRIRRG------IYKTHEVFNEGVRYYQEWLLLMRQGD---VYRYQDD---KPEIVLS
411770298_Citrobacter_freundii_ATCC_8090      ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                  ------------------------MEKRDERFQL------------------------HQRVKFQIRVLAQIMRMANKQ---------------------
492410748_Brevibacillus_agri                  ------------------------MAIRSIKLKLKTHTGPEAQNLRKG------IWRTHRLLNEGVAYYMKMLLLFRQES---TGERPKEELQEELICHI
495062547_Brevibacillus_sp-_CF112             ------------------------MAIRSIKLKLKTHTGPEAQNLRKG------IWRTHRLLNEGVAYYMKMLLLFRQES---TGERPKEELQEELICHI
506407588_Methylobacterium_nodulans           -------------------------------------------------------------------------MLTKQD---------------------
219945206_Methylobacterium_nodulans_ORS_2060  ------------------------MPVRSLKLKIVVPRHPSELEKAQA------LWSTHRLVNEAVSFYEQKLLLLRGET-----------YSTSDGSVP
760065057_Methylobacterium_nodulans           ----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                 *.........................................................H.............L...........................
RuvC-like_motifs                              //..................................................................................................

544884152_Alicyclobacillus_acidoterrestris    AEECKAELLERLRAR----------------------------------------------------QVENGHRGPAGSDDELLQLARQL----------
652589596_Alicyclobacillus_contaminans        ------------------------------------------------------------------------------MGFNTAELLRKV----------
652932497_Desulfovibrio_inopinatus            RHAVQEEALAFAKAA-----------------------------------------------------QRHNGCISTYEDQEILDVLRQL----------
667765471_Desulfonatronum_thiodismutans       ESQVAEDALAMAREA----------------------------------------------------QRRNGWP-VVGEDEEILLALRYL----------
497199019_Opitutaceae_bacterium_TAV5          WGSFRRQGRQRTGLS----------------------------------------------------QAVAPYITPGNNAPTLDEVFRSILAGNPTDRAT
654153037_Tuberibacillus_calidus              KEEIQERLWMKVRET----------------------------------------------------QQKNGFHGEV-SKDEVLETLRAL----------
754485389_Bacillus_thermoamylovorans          KAEIQAELWDFVLKM----------------------------------------------------QKCNSFTHEV-DKDVVFNILREL----------
495056180_Brevibacillus_sp-_CF112             ----------------------------------------------------------------------------------------------------
651512544_Bacillus_sp-_NSP2-1                 REQQQRNQADKNTQA----------------------------------------------------LPL----------DKALEALRQL----------
Secondary_structure_for_651512544_(Jpred)     HHHHHH---------------------------------------------------------------H----------HHHHHHHHHH----------
654874074_Desulfatirhabdium_butyrativorans    TPEERKARRILLALSWLSVESKLGAPSSYIVASGDEPAKDRNDNVVSALEEILQSRKVAKSEIDDWKRDCSASLSAAIRDDAVWVNRSKV----------
652569729_Alicyclobacillus_herbarius          GEQCRRELLQRLRER----------------------------------------------------QRLNGRTDEPGTDEELLKVARQI----------
652589403_Alicyclobacillus_contaminans        AEHCKRELLRRLRQV----------------------------------------------------QKENVGR-TSHTDEELLQVMRAL----------
411770298_Citrobacter_freundii_ATCC_8090      ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                  ----------------------------------------------------------------------------------------------------
492410748_Brevibacillus_agri                  REQQQRNQADKNTQA----------------------------------------------------LPL----------DKALEALRQL----------
495062547_Brevibacillus_sp-_CF112             REQQQRNQADKNTQA--------------------------------------------------------------LPLDKALEALRQL----------
506407588_Methylobacterium_nodulans           ----------------------------------------------------------------------------------------------------
219945206_Methylobacterium_nodulans_ORS_2060  QDEVRRQLLEQAREA----------------------------------------------------QARNGG--SGGSDDEIVRLCRSL----------
760065057_Methylobacterium_nodulans           -----------------------------------------------------------------------------------------M----------
CONSENSUS_0.8                                 ....................................................................................................
RuvC-like_motifs                              ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris    YELLVPQ---AIGAK------GDAQQIARKFLSPLADKDA------------------------------------------------------------
652589596_Alicyclobacillus_contaminans        EEEMRKT---SVGFD-------------------------------------------------------------------------------------
652932497_Desulfovibrio_inopinatus            YERLVPSVNENNE--------AGDAQAANAWVSPLMSAESEGGLSVYDKVLDPPPVWMKLKEEKAPGWEAASQIWIQSD---------------------
667765471_Desulfonatronum_thiodismutans       YEQIVPS---CLLDDLGKPLKGDAQKIGTNYAGPLFDSD--TCRRDEGKDVACCGPFHEVAGKYLGALPEWATPISKQEFDGKDASHLRFKATGGDDAFF
497199019_Opitutaceae_bacterium_TAV5          LDAALMQLLKACDGA------GAIQQEGRSYWPKFCDPDSTANFAGDPAMLRREQHRLLLPQVLHDPAITHDSPALGSFDTYSIAT--------------
654153037_Tuberibacillus_calidus              YEELVPS---AVGKS------GEANQISNKYLYPLTDPA--S----------------------------------------------------------
754485389_Bacillus_thermoamylovorans          YEELVPS---SVEKK------GEANQLSNKFLYPLVDPN--S----------------------------------------------------------
495056180_Brevibacillus_sp-_CF112             ----------------------------------------------------------------------------------------------------
651512544_Bacillus_sp-_NSP2-1                 YELLVPS---SVGQS------GDAQIISRKFLSPLVDPN--S----------------------------------------------------------
Secondary_structure_for_651512544_(Jpred)     HHHHHHH---HH-----------HHHHHHHH---------------------------------------------------------------------
654874074_Desulfatirhabdium_butyrativorans    FDEAVKSVGSSLTREEAWDMLERFFGSRDAYLTPMKDPEDKSSETEQEDKAKDLVQKAGQWLSSRYGTSEGADFCRMSDIYGKIAAWADNASQGGSSTVD
652569729_Alicyclobacillus_herbarius          YEVLVPQ---SIGKS------GDAQQLASNFLSPLVDPN-------------------------------------------------------------
652589403_Alicyclobacillus_contaminans        YELIVPS---AVGKK------GDAASLSRKFLSPLAWKD--S----------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090      ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                  ----------------------------------------------------------------------------------------------------
492410748_Brevibacillus_agri                  YELLVPS---SVGQS------GDAQIISRKFLSPLVDPN--S----------------------------------------------------------
495062547_Brevibacillus_sp-_CF112             YELLVPS---SVGQS------GDAQIISRKFLSPLVDPNS------------------------------------------------------------
506407588_Methylobacterium_nodulans           ----------------------------------------------------------------------------------------------------
219945206_Methylobacterium_nodulans_ORS_2060  YEAIVLA-------------DDANAQLANAFLGPLTDPNSAGFLEAFNKVDRPAPSWLDQVPASDPIDPAVLAEANAWLDTD------------------
760065057_Methylobacterium_nodulans           YEAIVLA-------------DDANAQLANAFLGPLTDPNSAGFLEAFNKVDRPAPSWLDQVPASDPIDPAVLAEANAWLDTD------------------
CONSENSUS_0.8                                 ....................................................................................................
RuvC-like_motifs                              ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris    ---------------VGGLGIAKAGNKPR---------WVRMREAGEPGWEEEKEKAETRKSADRTA---DVLRALA--------------------DFG
652589596_Alicyclobacillus_contaminans        -------------------------------------------------------------TDNPFA---------------------------------
652932497_Desulfovibrio_inopinatus            ---------------EGQSLLNKPGSPPR---------WIRKLRSGQP-WQDDFVSDQKKKQDELTKGNAPLIKQLK--------------------EMG
667765471_Desulfonatronum_thiodismutans       RVSIEKANAWYEDPANQDALKNKAYNKDD---------WKKEKDKGISSWAVKYIQKQLQLGQDPRT---EVRRKLWL-------------------ELG
497199019_Opitutaceae_bacterium_TAV5          ---------------PDTRTPQLTGPKARARLEQAITLWRVRLPESAADFDRLASSLKKIPDDDSRLNLQGYVGSSAKGEVQARLFALLLFRHLERSSFT
654153037_Tuberibacillus_calidus              ---------------QSGKGTANSGRKPR---------WKKLKEAGDPSWKDAYEKWEKERQEDPKL---KILAALQ--------------------SFG
754485389_Bacillus_thermoamylovorans          ---------------QSGKGTASSGRKPR---------WYNLKIAGDPSWEEEKKKWEEDKKKDPLA---KILGKLA--------------------EYG
495056180_Brevibacillus_sp-_CF112             ----------------------------------------------------------------------------------------------------
651512544_Bacillus_sp-_NSP2-1                 ---------------EGGKGTSKAGAKPT---------WQKKKEANDPTWEQDYEKWKKRREEDPTA---SVITTLE--------------------EYG
Secondary_structure_for_651512544_(Jpred)     -------------------EEE----------------HHHHHH-------HHHHHHHHHH----HH---HHHHHHH--------------------H--
654874074_Desulfatirhabdium_butyrativorans    DLVSELRQHFDTKESKATNGLDWIIGLSS---------YTGHTPNPVHELLRQNTSLNKSHLDDLKKKANTRAESCKS-------------------KIG
652569729_Alicyclobacillus_herbarius          --------------SKGGQGQSNAGRKPA---------WQKMRDEGNPGWVAAKERYEQRKATDPTK---KMIEMLD--------------------GLG
652589403_Alicyclobacillus_contaminans        ---------------KGLTGESKAGNKPR---------WKRLQEQGLP-YEEEYNRWLREKESDPAK---HIPAQLA--------------------SMG
411770298_Citrobacter_freundii_ATCC_8090      ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                  --------------------------------------------------------------------------------------------------YG
492410748_Brevibacillus_agri                  ---------------EGGKGTSKAGAKPT---------WQKKKEANDPTWEQDYEKWKKRREEDPTA---SVITTLE--------------------EYG
495062547_Brevibacillus_sp-_CF112             ---------------EGGKGTSKAGAKPT---------WQKKKEANDPTWEQDYEKWKKRREEDPTA---SVITTLE--------------------EYG
506407588_Methylobacterium_nodulans           ----------------------------------------------------------------------------------------------------
219945206_Methylobacterium_nodulans_ORS_2060  ---------------AGRAWLVDTGAPPR---------WRSLAAKQDPIWPREFARKLGELRKEAASGTSAIIKALKR-------------------DFG
760065057_Methylobacterium_nodulans           ---------------AGRAWLVDTGAPPR---------WRSLAAKQDPIWPREFARKLGELRKEAASGTSAIIKALKR-------------------DFG
CONSENSUS_0.8                                 ....................................................................................................
RuvC-like_motifs                              ....................................................................................................
```

```
544884152_Alicyclobacillus_acidoterrestris    ---LKPLMRVYTDSE---MSSVEWKPLRKGQAVRTWD------------------RDMFQQAIERMMSWE-------SWNQRVGQEYAKLVEQKNRFEQK
652589596_Alicyclobacillus_contaminans        ----------------------------------------------------------------------------------------------------
652932497_Desulfovibrio_inopinatus            ---LLPLVNPF----------FRHLLDPEGKGVSPWD------------------RLAVRAAVAHFISWE-------SWNHRTRAEYNSLKLRRDEFEAA
667765471_Desulfonatronum_thiodismutans       ---LLPLFIPVFDK---------------TMVGNLWN------------------RLAVRLALAHLLSWE-------SWNHRAVQDQALARAKRDELAAL
497199019_Opitutaceae_bacterium_TAV5          LGLLRSATPPPKNAETPPPAGVPLPAASAADPVRIARGK----------------RSFVFRAFTSLPCWHGGDNIHPTWKSFDIAAFKYALTVINQIEEK
654153037_Tuberibacillus_calidus              ---LIPLFRPFTENDHKAVISVKWMPKSKNQSVRKFD------------------KDMFNQAIERFLSWE-------SWNEKVAEDYEKTVSIYESLQKE
754485389_Bacillus_thermoamylovorans          ---LIPLFIPFTDSNEPIVKEIKWMEKSRNQSVRRLD------------------KDMFIQALERFLSWE-------SWNLKVKEEYEKVEKEHKTLEER
495056180_Brevibacillus_sp-_CF112             -------------------------------------------------------------------------------------------------MPK
651512544_Bacillus_sp-_NSP2-1                 ---IRPIFPLYTNT----VTDIAWLPLQSNQFVRTWD------------------RDMLQQAIERLLSWE-------SWNKRVQEEYAKLKEKMAQLNEQ
Secondary_structure_for_651512544_(Jpred)     -----HHHHHH-----------EE----------HHH------------------HHHHHHHHHHHH-HH-------HHHHHHHHHHHHHHHHHHHHHHH
654874074_Desulfatirhabdium_butyrativorans    SKGQRPYSDAILNDVES-VCGFTYRVDKDGQPVSVADYSKYDVDYKWGTARHYIFAVMLDHAARRISLAH-------KWIKRAEAERHKFEEDAKRIANV
652569729_Alicyclobacillus_herbarius          ---LKPLFSVFTET---YTTGVKWKDLSKRQGVRTWD---------------------------------------------------------------
652589403_Alicyclobacillus_contaminans        ---LKPFLKVFTES----TEGIAWLPLAKDQGVRTWD------------------RDMFQQAIEGLLSWE-------SWNRRV-----------------
411770298_Citrobacter_freundii_ATCC_8090      ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                  ----------------------------------------------------------------------------------------------------
492410748_Brevibacillus_agri                  ---IRPIFPLYTNT----VTDIAWLPLQSNQFVRTWD------------------RDMLQQAIERLLSWE-------SWNKRVQEEYAKLKEKMAQLNEQ
495062547_Brevibacillus_sp-_CF112             ---IRPIFPLYTNT----VTDIAWLPLQSNQFVRTWD------------------RDMLQQAIERLLSWE-------SWNKRV-----------------
506407588_Methylobacterium_nodulans           ----------------------------------------------------------------------------------------------------
219945206_Methylobacterium_nodulans_ORS_2060  ---VLPLFQP----------SLAPRILGSRSSLTPWD------------------RLAFRLAVGHLLSWE-------SWCTRARDEHTARVQRLEQFSSA
760065057_Methylobacterium_nodulans           ---VLPLFQP----------SLAPRILGSRSSLTPWD------------------RLAFRLAVGHLLSWE-------SWCTRARDEHTARVQRLEQFSSA
CONSENSUS_0.8                                 ....................................................................................................
RuvC-like_motifs                              ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris    NF--VGQEHLVHLVNQLQQDMKEASPGLESKEQTA----------------------------------HYVTGRALRGSDKVFEKWGKLAPDAP-----
652589596_Alicyclobacillus_contaminans        ---------------------------------------------------------------------HRITRRAIRGWDRIAEAWRRLPPDAP-----
652932497_Desulfovibrio_inopinatus            S---DEFKDDFTLLRQYEAKRHSTLKSIALADDSNP---------------------------------YRIGVRSLRAWNRVREEWIDKGAT-------
667765471_Desulfonatronum_thiodismutans       FL--GMEDGFAGLREYELRRNESIKQHAFEPVDRP----------------------------------YVVSGRALRSWTRVREEWLRHGDT-------
497199019_Opitutaceae_bacterium_TAV5          TKERQKECAELETDFDYMHGRLAKIPVKYTTGEAEPPPILANDLRIPLLRELLQNIKVDTALTDGEAVSYGLQRRTIRGFRELRRIWRGHAPAGTVFSSE
654153037_Tuberibacillus_calidus              LK--GISTKAFEIMERVEKAYEAHLRE-ITFSNST----------------------------------YRIGNRAIRGWTEIVKKWMKLDPSAP-----
754485389_Bacillus_thermoamylovorans          IK--E-DIQAFKSLEQYEKERQEQLLR-DTLNTNE----------------------------------YRLSKRGLRGWREIIQKWLKMDENEP-----
495056180_Brevibacillus_sp-_CF112             IL--RGHK-WISLLEQYEENRERELRENMTAANDK----------------------------------YRITKRQMKGWNELYELWSTFPASAS-----
651512544_Bacillus_sp-_NSP2-1                 LE--GGQE-WISLLEQYEENRERELRENMTAANDK----------------------------------YRITKRQMKGWNELYELWSTFPASAS-----
Secondary_structure_for_651512544_(Jpred)     ------HH-HHHHHHHHHHHHHHHHH--------H----------------------------------HHHHHEEE---HHHHHHH-------------
654874074_Desulfatirhabdium_butyrativorans    PARA-----------REWLDSFCKERSVTSGAVEP----------------------------------YRIRRRAVDGWKEVVAAWSKSDCKST-----
652569729_Alicyclobacillus_herbarius          ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans        ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090      ----------------------------------------------------------------------------------MYELWSTFPASAS-----
696372964_Citrobacter_freundii                ----------------------------------------------------------------------------------LYELWSTFPASAS-----
492410745_Brevibacillus_agri                  ------------------------------------------------------------------------------------DRWDE-----------
492410748_Brevibacillus_agri                  LE--GGQE-WISLLEQYEENRERELRENMTAANDK----------------------------------YRITKRQMKGWNELYELWSTFPASAS-----
495062547_Brevibacillus_sp-_CF112             ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans           ----------------------------------------------------------------------------------------------------
219945206_Methylobacterium_nodulans_ORS_2060  HLKGDLATKVSTLREYERARKEQIAQLGLPMGERD----------------------------------FLITVRMTRGWDDLREKWRRSGDKG------
760065057_Methylobacterium_nodulans           HLKGDLATKVSTLREYERARKEQIAQLGLPMGERD----------------------------------FLITVRMTRGWDDLREKWRRSGDKG------
CONSENSUS_0.8                                 ..........................................................................R...........W.............
RuvC-like_motifs                              ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris    -FDLYDAEIKNVQRRNTRRFGSHDLF-AKLAEPEYQALW--------RE------DA-SFLTRYAVYNSILRKLNHAK-MFATFTLPDATAHPIWTRFDK
652589596_Alicyclobacillus_contaminans        -ESEYIEAFKDIQRKNPRKIGSEPLF-KNLAAPGVRSEL--------LN------NP-QVLITFAKYNELQRQLAKAK-QFAQKTLPHPVFHPVWVRYDK
652932497_Desulfovibrio_inopinatus            -EEQRVTILSKLQTQLRGKFGDPDLF-NWLAQDRHVHLW----------------SPRDSVTPLVRINAVDKVLRRRK-PYALMTFAHPRFHPRWILYEA
667765471_Desulfonatronum_thiodismutans       -QESRKNICNRLQDRLRGKFGDPDVF-HWLAEDGQEALW--------KE-------R-DCVTSFSLLNDADGLLEKRK-GYALMTFADARLHPRWAMYEA
497199019_Opitutaceae_bacterium_TAV5          LKEKLAGELRQFQTDNSTTIGSVQLFNELIQNPKYWPIWQAPDVETARQWADAGFAD-DPLAALVQEAELQEDIDALK-APVKLTPADPEYSRRQYDFNA
654153037_Tuberibacillus_calidus              -QGNYLDVVKDYQRRHPRESGDFKLF-ELLSRPENQAAW--------RE------YP-EFLPLYVKYRHAEQRMKTAK-KQATFTLCDPIRHPLWVRYEE
754485389_Bacillus_thermoamylovorans          -SEKYLEVFKDYQRKHPREAGDYSVY-EFLSKKENHFIW--------RN------HP-EYPYLYATFCEIDKKKKDAK-QQATFTLADPINHPLWVRFEE
495056180_Brevibacillus_sp-_CF112             -HEQYKEALKRVQQRLRGRFGDAHFF-QYLMEEKNRLIW--------KG------NP-QRIHYFVARNELTKRLEEAK-QSATMTLPNARKHPLWVRFDA
651512544_Bacillus_sp-_NSP2-1                 -HEQYKEALKRVQQRLRGRFGDAHFF-QYLMEEKNRLIW--------KG------NP-QRIHYFVARNELTKRLEEAK-QSATMTLPNARKHPLWVRFDA
Secondary_structure_for_651512544_(Jpred)     ---HHHHHHHHHHHH-------HHHH-HHHHHHH-HHH---------------------EEEEEHHHHHHHHHHHHHH-HH--E---------EEEEEE-
654874074_Desulfatirhabdium_butyrativorans    -EDRIAAARALQDDSEIDKFGDIQLF-EALAEDDALCVW--------HKDGEATNEP-DFQPLIDYSLAIEAEFKKRQFKVPAYRHPDELLHPVFCDFGK
652569729_Alicyclobacillus_herbarius          -RDMFQSL-----------------------SERSGVID--------VG-------S-HTVHHIDLATASDAQIQYEL----------------------
652589403_Alicyclobacillus_contaminans        -REEYDALSARVYAYHAKHFAD---------QPGWAVYW--------PQ------SQ-P-----------------------------------------
411770298_Citrobacter_freundii_ATCC_8090      -HEQYKEALKRVQQRLRGRFGDAHFF-QYLMEEKNRLIW--------KG------NP-QRIHYFVARNELTKRLEEAK-QSATMTLPNARKHPLWVRFDA
696372964_Citrobacter_freundii                -HEQYKEALKRVQQRLRGRFGDAHFF-QYLMEEKNRLIW--------KG------NP-QRIHYFVARNELTKRLEEAK-QSATMTLPNARKHPLWVRFDA
492410745_Brevibacillus_agri                  -LDSLKQAVEQKKSPL---------------DQTDRTFW--------EG----------------IVCDLTKVLPRNE----------------------
492410748_Brevibacillus_agri                  -HEQYKEALKRVQQRLRGRFGDAHFF-QYLMEEKNRLIW--------KG------NP-QRIHYFVARNELTKRLEEAK-QSATMTLPNARKHPLWVRFDA
495062547_Brevibacillus_sp-_CF112             -QEEYAKLKEKMA--------------QLNEQLEGGQEW----------------------------CTLSR----------------------------
506407588_Methylobacterium_nodulans           -KQQKITYCTNMNEVFEAKLGSADLL----------LNW--------DH----------------LRGRIRDRVDAGDIGSAFLKLALDVAHVLPDGVDD
219945206_Methylobacterium_nodulans_ORS_2060  -QEALHAIIATEQTRKRGRFGDPDLF-RWLARPENHHVW-------ADG------HA-DAVGVLARVNAMERLVERSR-DTALMTLPDPVAHPRSAQWEA
760065057_Methylobacterium_nodulans           -QEALHAIIATEQTRKRGRFGDPDLF-RWLARPENHHVW-------ADG------HA-DAVGVLARVNAMERLVERSR-DTALMTLPDPVAHPRSAQWEA
CONSENSUS_0.8                                 ....................G....F...L........W....................................................H........
RuvC-like_motifs                              ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris    LGG-NLHQYT-FLFNEFGERRHAIRFHKLLKV-------ENGVAREVDDVTVPISMSEQLDNL----------------LPRDPNEPIALYFRDYGAEQ-
652589596_Alicyclobacillus_contaminans        LGG-NLHHYQ-IEPAVHANDTHKVKFSSLL------LPQEDGSYAEVKDVTVSLAPSLQFPTGLVHPKVTTPPRTGLVTVMDEEAGKPVVCYRDRGHDA-
652932497_Desulfovibrio_inopinatus            PGGSNLRQYA-------LDCTENALHITLPLL----VDDAHGTWIEKK-IRVPLAPSGQIQDLT---------------LEKLEKKKNRLYY--RSGFQ-
667765471_Desulfonatronum_thiodismutans       PGGSNLRTYQ-------IRKTENGLWADVVLL----SPRNESAAVEEKTFNVRLAPSGQLSNVSFDQ------------IQKGSKMVGRCRY--QSANQ-
497199019_Opitutaceae_bacterium_TAV5          VSKFGAGSRS---ANRHEPGQTERGHNTFTTEIAARNAADGNRWRATH-VRIHYSAPRLLRDGLRRP------------DTDGNEALEAVPWLQPMMEAL
654153037_Tuberibacillus_calidus              RSGTNLNKYRLIMNE-------KEKVVQFDRLICLN---ADGHYEEQEDVTVPLAPSQQFDDQIKFS------------SEDTGKGKHNFSYYHKGINY-
754485389_Bacillus_thermoamylovorans          RSGSNLNKYRILTEQLHTEKLKKKLTVQLDRLIYPT---ESGGWEEKGKVDIVLLPSRQFYNQIFLD------------IEE--KGKHAFTYKDESIKF-
495056180_Brevibacillus_sp-_CF112             RGG-NLQDY---YLTAEADKPRSRRFVTFSQL----IWPSESGWMEKKDVEVELALSRQFYQQV--K------------LLKNDKGKQKIEFKDKGSGS-
651512544_Bacillus_sp-_NSP2-1                 RGG-NLQDY---YLTAEADKPRSRRFVTFSQL----IWPSESGWMEKKDVEVELALSRQFYQQV--K------------LLKNDKGKQKIEFKDKGSGS-
Secondary_structure_for_651512544_(Jpred)     --------E---EEEEE-------HEEEE---------------EE---EEEEEEHHHHHHHHH--H------------H--------EEEEEE------
654874074_Desulfatirhabdium_butyrativorans    SRWKI---NYDVHKNVQAPFYRGLCLTLWTGSEIK---PVPLCWQSKR-LTRDLALGNNHRNDAASA------------VTRADRLGRAASNVTKSDMV-
652569729_Alicyclobacillus_herbarius          ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans        ---------------------------------------RQKGWVKMK----------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090      RGG-NLQDY---YLTAEADKPRSRRFVTFSQL----IWPSESGWMEKKDVEVELALSRQFYQQV--K------------LLKNDKGKQKIEFKDKGSGS-
696372964_Citrobacter_freundii                RGG-NLQDY---YLTAEADKPRSRRFVTFSQL----IWPSESGWMEKKDVEVELALSRQFYQQV--K------------LLKNDKGKQKIEFKDKGSGS-
492410745_Brevibacillus_agri                  -----------------------------------------ADW--------------------------------------------------------
492410748_Brevibacillus_agri                  RGG-NLQDY---YLTAEADKPRSRRFVTFSQL----IWPSESGWMEKKDVEVELALSRQFYQQV--K------------LLKNDKGKQKIEFKDKGSGS-
495062547_Brevibacillus_sp-_CF112             ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans           Q-----------------------------------LARAAFHFQSAKGAKSKHADSVQA----------------------------------------
219945206_Methylobacterium_nodulans_ORS_2060  EGGSNLRNYQ------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans           EGGSNLRNYQ------------------------------------------------------------------------------------------
CONSENSUS_0.8                                 ....................................................................................................
RuvC-like_motifs                              ....................................................................................................
```

```
544884152_Alicyclobacillus_acidoterrestris      ----HFTGEFGGA----KIQCRRDQLA--------------HMHR--RRGARDVYLNVSVRVQSQSE--ARGERRPPYAAVFRLVG-----DNHRAFVHF
652589596_Alicyclobacillus_contaminans          ----LVPVAFGGA----KLQFNRAHLS-----------AGYRKGVLSAGGGGSIYFNVTLDVQVPNE---------------------RDVSKTFSFSRD
652932497_Desulfovibrio_inopinatus               ----QFAGLAGGA----EVLFHRPYME-----------HDERSEESLLERPGAVWFKLTLDVATQAP---------------------PNWLDGKGRVRT
667765471_Desulfonatronum_thiodismutans         ----QFEGLLGGA----EILFDRKRIA-----------NEQHGATDLASKPGHVWFKLTLDVRPQAP---------------------QGWLDGKGRPAL
497199019_Opitutaceae_bacterium_TAV5            APLPTLPQDLTG-----MPVFLMPDVT--------------------LSGERRILLNLPVTLEPAALVEQLGNAGRWQNQFFG--------SREDPFALR
654153037_Tuberibacillus_calidus                ----ELKGTLGGA----RIQFDREHLL-----------R-RQGVK--AGNVGRIFLNVTLNIEPMQPFSRSGNLQTSVGKALKVYV--DGYPKVVNFKPK
754485389_Bacillus_thermoamylovorans            ----PLKGTLGGA----RVQFDRDHLR-----------RYPHKVE--SGNVGRIYFNMTVNIEPT---------ESPVSKSLKIHR--DDFPKFVNFKPK
495056180_Brevibacillus_sp-_CF112               ----TFNGHLGGA----KLQLERGDLE-----------KEEKNFE--DGEIGSVYLNVVIDFEPL-QEVKNGRVQAPYGQVLQLIRRPNEFPKVTTYKSE
651512544_Bacillus_sp-_NSP2-1                   ----TFNGHLGGA----KLQLERGDLE-----------KEEKNFE--DGEIGSVYLNVVIDFEPL-QEVKNGRVQAPYGQVLQLIRRPNEFPKVTTYKSE
Secondary_structure_for_651512544_(Jpred)       ----EEEEEE--E----EEEE----HH-----------HHHH----------EEEEEEEEEEE----HH--------HHHHHHHHHH---------HHHH
654874074_Desulfatirhabdium_butyrativorans      ----NITGLFEQADWNGRLQAPRQQLEAIAVVRDNPRLSEQERNLRMCGMIEHIRWLVTFSVKLQPQ-------GPWCAYAEQ--------HGLNTNPQY
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ----TFNGHLGGA----KLQLERGDLE-----------KEEKTSR--TGKSAAFTLTL------------------------------------------
696372964_Citrobacter_freundii                  ----TFNGHLGGA----KLQLERGDLE-----------KEEKTSR--TGKSAAFTLTL------------------------------------------
492410745_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
492410748_Brevibacillus_agri                    ----TFNGHLGGA----KLQLERGDLE-----------KEEKNFE--DGEIGSVYLNVVIDFEPL-QEVKNGRVQAPYGQVLQLIRRPNEFPKVTTYKSE
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             ----------------------------------------------------------------------------------------------------
219945206_Methylobacterium_nodulans_ORS_2060    ------LEAVGG-----ELQITLPLLK--------------------AADDGRC----------------------------------------------
760065057_Methylobacterium_nodulans             ------LEAVGG-----ELQITLPLLK--------------------AADDGRC----------------------------------------------
CONSENSUS_0.8                                   ....................................................................................................
RuvC-like_motifs                                ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris      DKLSDYLAEHPDDGKLGSEGLLSGLRVMSVDLGLRTSASISVFRVARKDELK-------------------------------PNSKGRVPFFFPIKGND
652589596_Alicyclobacillus_contaminans          RDLVSLKAEELKRYMETKPLGMPGVRVMSVDLGVRYGAAISVFEVKPFAEVR-------------------------------KDK-----LHYPITGCE
652932497_Desulfovibrio_inopinatus               PPEVHHFKTALSNKSKHTRTLQPGLRVLSVDLGMRTFASCSVFELIEGKPETG--------------------------------------RAFPVADER
667765471_Desulfonatronum_thiodismutans         PPEAKHFKTALSNKSKFADQVRPGLRVLSVDLGVRSFAACSVFELVRGGPDQ-------------------------------G-------TYFPAADGR
497199019_Opitutaceae_bacterium_TAV5            WPADGAVKTAKGKTHIPWHQDRDHFTVLGVDLGTRDAGALALLNVTAQKPAK------------------------------------------PVHRII
654153037_Tuberibacillus_calidus                ELTEHIKESEKNTLTLGVESLPTGLRVMSVDLGQRQAAAISIFEVVSEKPD--------------------------------DNK-----LFYPVKDTD
754485389_Bacillus_thermoamylovorans            ELTEWIKDSKGKKLKSGIESLEIGLRVMSIDLGQRQAAAASIFEVVDQKPDI-------------------------------EGK-----LFFPIKGTE
495056180_Brevibacillus_sp-_CF112               QLVEWIKASPQHSA--GVESLASGFRVMSIDLGLRAAAATSIFSVEESSDKN-------------------------------AAD-----FSYWIEGTP
651512544_Bacillus_sp-_NSP2-1                   QLVEWIKASPQHSA--GVESLASGFRVMSIDLGLRAAAATSIFSVEESSDKN-------------------------------AAD-----FSYWIEGTP
Secondary_structure_for_651512544_(Jpred)       HHHHHHH-----------HHHHHHHHHEEE---HHHH--EEEEEE----------------------------------------------EEEE-----
654874074_Desulfatirhabdium_butyrativorans      WPHADTNRDRKVHARLILPRLP-GLRVLSVDLGHRYAAACAVWEAVNTETVKEACQNVGRDMPKEHDLYLHIKVKKQGIGKQTEVDKTTIYRRIGADTLP
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                  ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
492410748_Brevibacillus_agri                    QLVEWIKASPQHSA--GVESLASGFRVMSIDLGLRAAAATSIFL--------------------------------------------------------
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             -----------------------GLRVLSIDLGVRSFATCSVFELKDTAPTT-------------------------------G-------VAFPLAEFR
219945206_Methylobacterium_nodulans_ORS_2060    ----------------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans             ----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                   ..........................V..VDLG.R.................................................................
RuvC-like_motifs                                ..............................D.....................................................................

544884152_Alicyclobacillus_acidoterrestris      N------LVAVHERSQLLKLPGET------------------ESKDLRAIREERQRTLRQLRTQLAYLRLLVRCGS-EDVGRR-----ERSWAKLIEQPV
652589596_Alicyclobacillus_contaminans          G------FVAEHERSVILKLPGE--------------------GVRTAGKQSERKQALAAIRAEMSILRKWLRVSQVTEEDRA-----KAVRGLLEDERG
652932497_Desulfovibrio_inopinatus               SMDSPNKLWAKHERSFKLTLPGET------------------PSRKEEEERSIARAEIYALKRDIQRLKSLLRLGEEDNDNRR-----DALLEQFFKGWG
667765471_Desulfonatronum_thiodismutans         TVDDPEKLWAKHERSFKITLPGEN------------------PSRKEEIARRAAMEELRSLNGDIRRLKAILRLSVLQEDDPR-----TEHLRLFMEAIV
497199019_Opitutaceae_bacterium_TAV5            GEADGRTWYASLADARMIRLPGEDARLFVRGKLVQEPYGERGRNASLLEWEDARNIILRLGQNPDELLGADPRRHSYPEINDKLLVALRRAQARLARLQN
654153037_Tuberibacillus_calidus                -------LFAVHRTSFNIKLPGEK------------------RTERRMLEQQKRDQAIRDLSRKLKFLKNVLNMQKLEKTDER-----EKRVNRWIKDRE
754485389_Bacillus_thermoamylovorans            -------LYAVHRASFNIKLPGET------------------LVKSREVLRKAREDNLKLMNQKLNFLRNVLHFQQFEDITER-----EKRVTKWISRQE
495056180_Brevibacillus_sp-_CF112               -------LVAVHHRSYMLRLPGEQ------------------VEKQVMEKRDERFQLHQRVKFQIRVLAQIMRMAN-KQYGDR-----WDELDSLKQAVE
651512544_Bacillus_sp-_NSP2-1                   -------LVAVHQRSYMLRLPGEQ------------------VEKQVMEKRDERFQLHQRVKFQIRVLAQIMRMAN-KQYGDR-----WDELDSLKQAVE
Secondary_structure_for_651512544_(Jpred)       -------EEEEE--EEEEE-----------------------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH------HHH-----HHHHHHHHHHHH
654874074_Desulfatirhabdium_butyrativorans      DGRPHPAPWARLDRQFLIKLQGE--------------------EKDAREASNEEIWALHQMECKLDRTKPLIDRLIASGWGLL-----KRQMARLDALKE
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                  ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
492410748_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             -------LWAVHERSFTLELPGEN------------------VGAAGQQWRAQADAELRQLRGGLNRHRQLLRAATVQ----------KGERDAYLTDLR
219945206_Methylobacterium_nodulans_ORS_2060    ----------------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans             ----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                   .........A.........L.GE.............................................................................
RuvC-like_motifs                                ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris      DAANHMTPDWREAFENELQK-----------------------------------------------LKSLHGICSDKE--------W------------
652589596_Alicyclobacillus_contaminans          GGWTMDPGEDSDHQPLQQFL-----------------------------------------------HEARLAVGELVNLVHLSPAEW------------
652932497_Desulfovibrio_inopinatus               EEDVVPGQA----------------------------------------------------------FPRSLFQGLGAAPFRSTPELW------------
667765471_Desulfonatronum_thiodismutans         DDPAKSA------------------------------------------------------------LNAELFKGFGDDRFRSTPDLW------------
497199019_Opitutaceae_bacterium_TAV5            RSWRLRDLAESDKALDEIHAERAGE------------------------------------------KPSPLPPLARDDAIKSTDEA-------------
654153037_Tuberibacillus_calidus                REEENP------VYVQEFEM-----------------------------------------------ISKVLY-SPHSV--------W------------
754485389_Bacillus_thermoamylovorans            NSDVPL------VYQDELIQ-----------------------------------------------IRELMY-KPYKD--------W------------
495056180_Brevibacillus_sp-_CF112               QKKSPLDQTDRTFWEGIVCD-----------------------------------------------LTKVLP-RNEAD--------W------------
651512544_Bacillus_sp-_NSP2-1                   QKKSPLDQTDRTFWEGIVCD-----------------------------------------------LTKVLP-RNEAD--------W------------
Secondary_structure_for_651512544_(Jpred)       --------------------------------------------------------------------EEEE---------------H------------
654874074_Desulfatirhabdium_butyrativorans      LGWIPAPDSSENLSREDGEAKDYRESLAVDDLMFSAVRTLRLALQRHGNRARIAYYLISEVKIRPGGIQEKLDENGRIDLLQDALALWHELFSSPGWRDE
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ---------------SLISN-----------------------------------------------LCKK-----------------------------
696372964_Citrobacter_freundii                  ---------------SLISN-----------------------------------------------LCKK-----------------------------
492410745_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
492410748_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             EAWSAKE-----LWPFEASL-----------------------------------------------LSELERCSTVADPL------W------------
219945206_Methylobacterium_nodulans_ORS_2060    -------------------------------------------------------------------IDTPL----------------------------
760065057_Methylobacterium_nodulans             -------------------------------------------------------------------IDTPL----------------------------
CONSENSUS_0.8                                   ....................................................................................................
RuvC-like_motifs                                ....................................................................................................
```

```
544884152_Alicyclobacillus_acidoterrestris      ----------------------------------MDAVYESVRRVWRHMGKQVRDWRKDVRSGERPKIRGYAKDV------------------------V
652589596_Alicyclobacillus_contaminans          ----------------------------------ERAVIERHRRLERITASHIRVFQTMRKVWGKRRNEDAAH--------------------------T
652932497_Desulfovibrio_inopinatus              ----------------------------------RQHCQTYYDKAEACLAKHISDWRKRTRPRPTSREMWYKTRSY-----------------------H
667765471_Desulfonatronum_thiodismutans         ----------------------------------KQHCHFFHDKAEKVVAERFSRWRTETRPKSSSWQDWRERRGYA-----------------------
497199019_Opitutaceae_bacterium_TAV5            ----------------------------------LLSQRDIIRRSFVQIANLILPLRGRRWEWRPHVEVPDCHILA--------QSDPGTDDTKRLVAGQ
654153037_Tuberibacillus_calidus                ----------------------------------VDQLKSIHRKLEEQLGKEISKWRQSISQ-GRQG--------------------------------V
754485389_Bacillus_thermoamylovorans            ----------------------------------VAFLKQLHKRLEVEIGKEVKHWRKSLSD-GRKG--------------------------------L
495056180_Brevibacillus_sp-_CF112               ----------------------------------EQAVVQIHRKAEEYVGKAVQAWRKRFAA-DERKG-------------------------------I
651512544_Bacillus_sp-_NSP2-1                   ----------------------------------EQAVVQIHRKAEEYVGKAVQAWRKRFAA-DERKG-------------------------------I
Secondary_structure_for_651512544_(Jpred)       ----------------------------------HHHHHHHHHHHHHHHHHHHHHHHHHHHH--------------------------------------
654874074_Desulfatirhabdium_butyrativorans      AAKQLWDSRIATLAGYKAPEENGDNVSDVAYRKKQQVYREQLRNVAKTLSGDVITCKELSDAWKERWEDEDQRWKKLLRWFKDWVLPSGTQANNATIRNV
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                  ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                    ----------------------------------EQAVVQIHRKAEEYVGKAVQAWRKRFAADERKG--------------------------------I
492410748_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             ----------------------------------QDTCKRAARLYRTEFGAVVSEWRSRTRSREDRK--------------------------------Y
219945206_Methylobacterium_nodulans_ORS_2060    ----------------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans             ----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                   ....................................................................................................
RuvC-like_motifs                                ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris      GGNSIEQIEYLERQYKFLKSWSFFGKVSGQ-----VIRAEKGSR--FAITLREHIDHAKEDRLKKLADRIIMEALGYVYALDERGKGKW-----------
652589596_Alicyclobacillus_contaminans          GGISLAHIEHLIQQRKLFIRWSTHARTYGE-----VRRLPKHEG--FAKRLQKHTNHVKEDRIKKLADMIVMAARGYRF---LDKRARW----------V
652932497_Desulfovibrio_inopinatus              GGKSIWMLEYLDAVRKLLLSWSLRGRTYGA-----INRQDTARFGSLASRLLHHINSLKEDRIKTGADSIVQAARGYIP---LPHGKGW-----------
667765471_Desulfonatronum_thiodismutans         GGKSYWAVTYLEAVRGLILRWNMRGRTYGE-----VNRQDKKQFGTVASALLHHINQLKEDRIKTGADMIIQAARGFVP---RKNGAGW-----------
497199019_Opitutaceae_bacterium_TAV5            RGISHERIEQIEELRRRCQSLNRALRHKPGERPVLGRPAKGEEIADPCPALLEKINRLRDQRVDQTAHAILAAALGVRLRAPSKDRAER--RHRDIHGEY
654153037_Tuberibacillus_calidus                YGISLKNIEDIEKTRRLLFRWSMRPENPGE-----VKQLQPGER--FAIDQQNHLNHLKDDRIKKLANQIVMTALGYRY---DGKRKKW----------I
754485389_Bacillus_thermoamylovorans            YGISLKNIDEIDRTRKFLLRWSLRPTEPGE-----VRRLEPGQR--FAIDQLNHLNALKEDRLKKMANTIIMHALGYCY---DVRKKKW----------Q
495056180_Brevibacillus_sp-_CF112               AGLSMWNIEELEGLRKLLISWSRRSRNPQE-----VNRFERGHT--SHQRLLTHIQNVKEDRLKQLSHAIVMTALGYVY---DERKQEW----------C
651512544_Bacillus_sp-_NSP2-1                   AGLSMWNIEELEGLRKLLISWSRRTRNPQE-----VNRFERGHT--SHQRLLTHIQNVKEDRLKQLSHAIVMTALGYVY---DERKQEW----------C
Secondary_structure_for_651512544_(Jpred)       --HHHHHHHHHHHHHHHHHHHHHHHHHH--------------------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH---H----EE----------E
654874074_Desulfatirhabdium_butyrativorans      GGLSLSRLATITEFRRKVQVGFF-TRLRPD-----GTRHEIGEQ--FGQKTLDALELLREQRVKQLASRIAEAALGIG----SEGGKGWDGGKRPRQRIN
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                  ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                    AGLSMWNIEELEGLRKLLISWSRRTRNPQE-----VNRFERGHT--SHQRLLTHIQNVKEDRLKQLSHAIVMTALGYVY---DERKQEW-----------
492410748_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             AGKSMWSVQHLTDVRRFLQSWSLAGRASGD-----IRRLDRERGGVFAKDLLDHIDALKDDRLKTGADLIVQAARGF-----QRNEFGY----------W
219945206_Methylobacterium_nodulans_ORS_2060    ----------------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans             ----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                   .G.S.........................................................R....A..I...A.G........................
RuvC-like_motifs                                ....................................................................................................

544884152_Alicyclobacillus_acidoterrestris      VAKYPPCQLILLEELSEYQFNNDRPPSENNQLMQWSHRGVFQELINQAQVHDLL-------------------VGTMYAAFSSRFDARTGAPGIRCRRVP
652589596_Alicyclobacillus_contaminans          KTRHAPCDLILFEDLSRYRFTMDRPPTENSQLMNWSHRELLKTVKMQAALFGIG-------------------VGTVPAAFTSRFDAQTGAPGLRCKRVT
652932497_Desulfovibrio_inopinatus              EQRYEPCQLILFEDLARYRFRVDRPRRENSQLMQWNHRAIVAETTMQAELYGQI-------------------VENTAAGFSSRFHAATGAPGVRCRFLL
667765471_Desulfonatronum_thiodismutans         VQVHEPCRLILFEDLARYRFRTDRSRRENSRLMRWSHREIVNEVGMQGELYGLH-------------------VDTTEAGFSSRYLASSGAPGVRCRHLV
497199019_Opitutaceae_bacterium_TAV5            ERFRAPADFVVIENLSRYLSSQDRARSENTRLMQWCHRQIVQKLRQLCETYGIP-------------------VLAVPAAYSSRFSSRDGSAGFRAVHLT
654153037_Tuberibacillus_calidus                A-KHPACQLVLFEDLSRYAFYDERSRLENRNLMRWSRREIPKQVAQIGGLYGLL-------------------VGEVGAQYSSRFHAKSGAPGIRCRVVK
754485389_Bacillus_thermoamylovorans            A-KNPACQIILFEDLSNYNPYEERSRFENSKLMKWSRREIPRQVALQGEIYGLQ-------------------VGEVGAQFSSRFHAKTGSPGIRCSVVT
495056180_Brevibacillus_sp-_CF112               A-EYPACQVILFENLSQYRSNLDRSTKENSTLMKWAHRSIPKYVHMQAEPYGIQ-------------------IGDVRAEYSSRFYAKTGTPGIRCKKVR
651512544_Bacillus_sp-_NSP2-1                   A-EYPACQVILFENLSQYRSNLDRSTKENSTLMKWAHRSIPKYVHMQAEPYGIQ-------------------IGDVRAEYSSRFYAKTGTPGIRCKKVR
Secondary_structure_for_651512544_(Jpred)       --------EEEEHH---H-------HHHHHHHHHHHHHHHHHHHHHHHHH-----------------------EEEE----------------E------
654874074_Desulfatirhabdium_butyrativorans      DSRFAPCHAVVIENLANYRPDETRTRLENRRLMTWSASKVHKYLSEACQLNGLY-------------------LCTVSAWYTSRQDSRTGAPGIRCQDVS
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                  ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                    CAEYPACQVILFENLSQYRSNLDRSTKENSTLMKWAHRSIPKYVHMQAEPYGIQ-------------------IGDVRAEYSSRFYAKTGTPGIRCKKVR
492410748_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             VQKHAPCHVILFEDLSRYRMRTDRPRRENSQLMQWAHRGVPDMVGMQGEIYGIQDRRDPDSARKHARQPLAAFCLDTPAAFSSRYHASTMTPGIRCHPLR
219945206_Methylobacterium_nodulans_ORS_2060    ----------------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans             ----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                   .........I..E.L..Y.....R...EN..LM.W....I......................................A...SR.....G..G.R.....
RuvC-like_motifs                                ............E.......................................................................................

544884152_Alicyclobacillus_acidoterrestris      ARCTQEHNPEPFPWWLNKFV-----------------VEHTLDACPLRADD---------LIPTGEGEIFVSP-FSAEEGDFHQIHADLNAAQNLQQRLW
652589596_Alicyclobacillus_contaminans          KQDKEKTPFWLI----------------------QFAEITGVNVTNVEPGQ---------LIPVDGGEWFVSPKGPRAADGLKCVHADINAAHNLQRRFW
652932497_Desulfovibrio_inopinatus              ERDFDNDLPKPYLLRELSWMLGNTKV-------ESEEEKLRLLSEKIRPGS---------LVPWDGGEQFATL--HPKRQTLCVIHADMNAAQNLQRRFF
667765471_Desulfonatronum_thiodismutans         EEDFHDGLPGMHLVGELDWLLPKDKD-------RTANEARRLLGGMVRPGM---------LVPWDGGELFATL---NAASQLHVIHADINAAQNLQRRFW
497199019_Opitutaceae_bacterium_TAV5            PDHRHRMPWSRILARLKAHEEDGKRLEKTVLDEARAVRGLFDRLDRFNAGHVPGKPWRTLLAPLPGGPVFVPL------GDATPMQADLNAAINIALRGI
654153037_Tuberibacillus_calidus                EHELYITEGGQKVRNQKFLD----SL-------VENNIIEPDDARRLEPGD---------LIRDQGGDKFATL---DERGELVITHADINAAQNLQKRFW
754485389_Bacillus_thermoamylovorans            K---------EKLQDNRFFK----NL-------QREGRLTLDKIAVLKEGD---------LYPDKGGEKFISL---SKDRKLVTTHADINAAQNLQKRFW
495056180_Brevibacillus_sp-_CF112               GQDL------QGRRFENLQK----RL-------VNEQFLTEEQVKQLRPGD---------IVPDDSGELFMTLTDGSGSKEVVFLQADINAAHNLQKRFW
651512544_Bacillus_sp-_NSP2-1                   GQDL------QGRRFENLQK----RL-------VNEQFLTEEQVKQLRPGD---------IVPDDSGELFMTLTDGSGSKEVVFLQADINAAHNLQKRFW
Secondary_structure_for_651512544_(Jpred)       -----------------HHH----HH-------HH---------------------------------EEEE----------EEEHHHHHHHHHHHHHHH
654874074_Desulfatirhabdium_butyrativorans      VREFMQSPFWRKQVKQAEAKHDENKG------DARERFLCELNKTWKAKTPAEWKKAGFVRIPLRGGEIFVSA--DSKSPSAKGIHADLNAAANIGLRAL
652569729_Alicyclobacillus_herbarius            ----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans          ----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090        ----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                  ----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                    GQDL------QGRRFENLQK----RL-------VNEQFLTEEQVKQLRPGD---------IVPDDSGELFMTLTDGSGSKEVVFLQADINAAHNLQKRFW
492410748_Brevibacillus_agri                    ----------------------------------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF112               ----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans             KREFEDQGFLELLKREN----------------------EGLDLNGYKPGD---------LVPLPGGEVFVCL----NANGLSRIHADINAAQNLQRRFW
219945206_Methylobacterium_nodulans_ORS_2060    ----------------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans             ----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                   ..............................................................P...G..F................AD.NAA.N...R..
RuvC-like_motifs                                .....................................................<--...............................D............
```



```
544884152_Alicyclobacillus_acidoterrestris        SDFDISQIRLRCDWGEVDGELVLIPRLTGKRT---------------------ADSYSNKVFYTNTGVTYYERERGKKRRKVFAQEKLSEEEAELLVEADE
652589596_Alicyclobacillus_contaminans            IP---RLPSVKCRRYVEAEGFAAVPSSTAFMKVHGKG--------------AFVSVDGEFYEYQKGRRVAV----NRADRTSSTLDEDEGDIGEEMLVSSN
652932497_Desulfovibrio_inopinatus                GRCG-EAFRLVCQPHGDDVLRLASTPGARLLGALQQL---------ENGQGAFELVRDMGSTSQMNRFVMKSLGKKKIKPLQDNNGDDELEDVLSVLPEE
667765471_Desulfonatronum_thiodismutans           GRCG-EAIRIVCNQLSVDGSTRYEMAKAPKARLLGALQQLKNGDAPFHLTSIPNSQKPENSYVMTPTNAGKKYRAGPGEKSSGEE--DELALDIVEQAEE
497199019_Opitutaceae_bacterium_TAV5              AAP--DRHDIHHRLRAENKKRILSLR---------------------------LGTQREKARWPGGAPAVTLSTPNNGASPEDSDALPERVSNLFVDIAGV
654153037_Tuberibacillus_calidus                  TRTH-GLYRIRCESREIKDAVVLVPSDKDQKEKMENL-------FGIGYLQPFKQENDVYKWVKGEKIKG---KKTSSQSDDKELV-SEILQEASVMADE
754485389_Bacillus_thermoamylovorans              TRTH-GFYKVYCKAYQVDGQTVYIPESKDQKQKIIEE-------FGEGY---FILKDGVYEWGNAGKLKI---KKGSSKQSSSELVDSDILKDSFDLASE
495056180_Brevibacillus_sp-_CF112                 QRYN-ELFKVSCRVIVRDEEEYLVPKTKSVQAKLGKG-------LFVKKS--DTAWKDVYVWDSQAKLKG---KTTFTEESESPEQ-LEDFQEIIEEAEE
651512544_Bacillus_sp-_NSP2-1                     QRYN-ELFKVSCRVIVRDEEEYLVPKTKSVQAKLGKG-------LFVKKS--DTAWKDVYVWDSQAKLKG---KTTFTEESESPEQ-LEDFQEIIEEAEE
Secondary_structure_for_651512544_(Jpred)         H--------EE--EEEE---EEEEE----HHHHH-----------EEE--------EEEEEE----HH-----HEHHH-----HH-HHHHHHHHHHHH-
654874074_Desulfatirhabdium_butyrativorans        TDP-----------DWPGKWWYVPCD---------------------------PVSFESKMDYVKGCAAVKVGQPLRQPAQTNADGAASKIRKGKKNRTAG
652569729_Alicyclobacillus_herbarius              -----------------------------------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans            -----------------------------------------------------------------------------------------------------
411770298_Citrobacter_freundii_ATCC_8090          -----------------------------------------------------------------------------------------------------
696372964_Citrobacter_freundii                    -----------------------------------------------------------------------------------------------------
492410745_Brevibacillus_agri                      QRYN-ELFKVSCRVIVRDEEEYLVPKTKSVQAKLGKG---------LFVKKSDTAWKDVYVWDSQAKLKG---KTTFTEESESPEQ-LEDFQEIIEEAEE
492410748_Brevibacillus_agri                      -----------------------------------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF112                 -----------------------------------------------------------------------------------------------------
506407588_Methylobacterium_nodulans               TQHG-DAFRLPCGKSAVQGQIRWAPLSMGKRQAGALG--------GFGYLEPTGHDSGSCQWRKTTEAEWRRLSGAQKDRDEAAAAEDEELQGLEEELLE
219945206_Methylobacterium_nodulans_ORS_2060      -----------------------------------------------------------------------------------------------------
760065057_Methylobacterium_nodulans               -----------------------------------------------------------------------------------------------------
CONSENSUS_0.8                                     .....................................................................................................
RuvC-like_motifs                                  .....................................................................................................

                                                  AREKSVVLMRDPSGIINRGN----WTRQKEFWSMVNQRIEGYLVKQIRSRVPLQDSACENTGDI--------
 652589596_Alicyclobacillus_contaminans           GAGEFVRMFYDESGYVGYG----RWMDSKVFWGKVRQIVHRAIQDQVEKRAAARGENGATSSR---------
 652932497_Desulfovibrio_inopinatus               DDTGRITVFRDSSGIFFPCN---VWIPAKQFWPAVRAMIWKVMASHSLG-----------------------
5448?4152?Alicyclobacillus_acidoterrestris        LAQGRKTFFRDPSGVFFAPD---RWLPSEIYWSRIRRRIWQVTLERNSSGRQERAEMDEMPY----------
 497199019_Opitutaceae_bacterium_TAV5             ANFERVTIEGVSQ----------KFATGRGLWASVKQRAWNRVARLNETVTDNNRNEEEDDIPM--------
           uberibacillus_calidus                  LKGNRKTLFRDPSGYVFPKD---RWYTGGRYFGTLEHLLKRKLAER---RLFDGGSSRRGLFNGTDSNTNVE
6654485389_Bacillus_thermoamylovorans             LKGEKLMLYRDPSGNVFPSD---KWMAAGVFFGKLERILISKLTNQYSISTIEDDSSKQSM-----------
                                                  AKGTYRTLFRDPSGVFFPES---VWYPQKDFWGEVKRKLYGKLRERFLTKAR--------------------
654153037_T                      1                AKGTYRTLFRDPSGVFFPES---VWYPQKDFWGEVKRKLYGKLRERFLTKAR--------------------
                                                  ---EEEEEEE----EEE----------HHHHHHHHHHHHHHHHHHHHHH-----------------------
495056180_Brevibacillus_sp-_CF112                 TSKEKVYLWRDISAFPLESNEIGEWKETSAYQNDVQYRVIRMLKEHIKSLDNRTGDNVEG------------
651512544_Bacillus_sp-_NSP2-                      ------------------------------------------------------------------------
Secondary_structure_for_651512544_(Jpred)         ------------------------------------------------------------------------
654874074_Desulfatirhabdium_butyrativorans        ------------------------------------------------------------------------
652569729_Alicyclobacillus_herbarius              ------------------------------------------------------------------------
652589403_Alicyclobacillus_contaminans            AKGTYRTLFRDPSGVFFPES---VWYPQKDFWGEVKRKLYGKLRERFLTKAR--------------------
411770298_Citrobacter_freundii_ATCC_8090          ------------------------------------------------------------------------
696372964_Citrobacter_freundii   2                ------------------------------------------------------------------------
492410745_Brevibacillus_agri                      RSGERVVFFRDPSGVVLPTD---LWFPSAAFWSIVRAKTVGRLRSHLDAQAEASYAVAAGL-----------
492410748_Brevibacillus_agri                      ------------------------------------------------------------------------
495062547_Brevibacillus_sp-_CF11                  ------------------------------------------------------------------------
506407588_Methylobacterium_nodulans               ........................................................................
219945206_Methylobacterium_nodulans_ORS_2060      ........................................................................
760065057_Methylobacterium_nodulans
CONSENSUS_0.8
RuvC-like_motifs
```

Figure 2.S4: Multiple alignment of C2c1 protein family

The alignment was built using MUSCLE program and modified manually on the basis of local PSI-BLAST pairwise alignments. Each sequence is labelled with GenBank Identifier (GI) number and systematic name of an organism. Secondary structure was predicted by Jpred and shown underneath the sequence which was used as a query (designations: H- alpha helix , E–beta strand). CONSENSUS was calculated for each alignment column by scaling the sum-of pairs score within the column between those of a homogeneous column (the same residue in all aligned sequences) and a random column with homogeneity cutoff 0.8. Active site motifs of RuvC-like domain are shown below alignment.

10.1.5 Figure 2.S5

```
100000002   CEPX01008730.1  -------------MRSNYHGGRNARQWRKQISGLA-RRTKETVFTYKFPLETDAAEI--DFDKAVQTYGIAEGVGHGSLIGLVCAFHLSGFRLFSKAGEAMAFRNRSRYPTDAFAEKLSAIMGI
100020996 AUXO013399408.1  -----MKKFELKQNFRNNYSG-KTLRNFRQTLAQIANKKSSDSILTIKFKLDCSKTGKLPKYENLISLYDTIEDIKKGTLSYYLFTLIVSGFKFFGSASQAKAFSTKDIFKDNDFYNQFKIQSHL
100022927   CEQE01148443.1  -------------MRSNYHGGRNARQWRKQISGLA-RRTKETVFTYKFPLETDAAEI--DFDKAVQTYGIAEGVGHGSLIGLVCAFHLSGFRLFSKAGEAMAFRNRSRYPTDAFAEKLSAIMGI
100000004   CEVA01036528.1  ---------------------MNARDWRKHVGVLA-QQHKETTRTYTFPLDTTGSAI--DFDAALQAYNAVEGVGYGSLLGLACAVHLSGFRLFSTGKEAATFRNRARYPNAAFQAALRKELGT
100021577   CEPS01188136.1  MVAGLKKIKRDGVTMKSNYHGGVKARAWRKRIGGLA-RRQKETVFTYKFPLETEEAGI--DFDKAVQTYGIAEGISQGSLIGLVCAFHLSGFRLFSKADETKAFCNQGRYPNQAFAEKLRNELSV
Jpred Secondary structure   ----HHHH----EEE--------HHHHHHHH----H-H----EEEEEE--------------HHHHHHHEE--------HHHHHHHHH------------HHHH-------HHHHHHHHHHHHH
CONSENSUS_0.8               ........................R..RK.I..LA.....ET..T..F.L.........FD..I..Y...E.V..G.L........LSGFR.F....EA..F.....Y....F.........

100000002   CEPX01008730.1  QLPTLSPEGLDLIFQSPPRSRDGIAPVWSENEVRNRLYTNWTGRGPANKPDEHLL-EIAGEIAKQVFPKFGGWDDLASDPDKALAAADKYFQSQ-GDFPSIASLPAAIMLSPANSTVDFEGD--Y
100020996 AUXO013399408.1  DLPDFVPSKIYQRLKKNVRSTNGKDNAFKASVIVAEYRKEIGKLKNKDESSEHQCEELFKKIGTALETRFSSWQDLINNCSTGCEIIDEILNDSFGTLPSIKKM--VLASTTQSSDGEQDGI--A
100022927   CEQE01148443.1  QLPTLSPEGLDLIFQSPPRSRDGIAPVWSENEVRNRLYTNWTGRGPANKPDEHLL-EIAGEIAKQVFPKFGGWDDLASDPDKALAAADKYFQSQ-GDFPSIASLPAAIMLSPANSTVDFEGD--Y
100000004   CEVA01036528.1  TITTLTPETLDRLFSSRPKRRNGVPLPWNQDSIRDRLYTNWVKPRPGDTPDAVLF-QIATGIAQEITEDVSSWTDLAKNSDRGLKAAHRYFARV-GGFPAFDNLTPPATVQPTDTTIDYDPNAPF
100021577   CEPS01188136.1  TLPKLSPQSLDVLFQSSPKSKNGVAPEWSKNAIRNRLYTNWTGKGAGTNPDEHLL-EIAEDIAAEIDSDLDGWKDLEEHPEKGLSAADRYFQAQ-GDFPSLTGLPPSVPLTPQNSTVAFEGD--P
Jpred Secondary structure   E--------HHHHH---------------HHHHHHHHHH------------HHHH-HHHHHHHHHH--------------------HHHHHHH----------------------EEEE------
CONSENSUS_0.8               .L....P..L............G.........I.......................EI...I...I......W.DL..................G..PS...L..........S.........

100000002   CEPX01008730.1  IAIDPAAETL----LHQAVSRCAARLGRERPDLDQNKGPFVSSLQDALVSSQNNGLSWLFGVGFQHWKEKSPKELIDEYKVPADQHGAVTQVKSFVDAIPLNPLFDTTH----YGEFRASVAGKV
100020996 AUXO013399408.1  IAYDPDSTFIKSDELLNPYFAVATILKSMPPEIQQDKKS--AYVKANLTTPTHNALSWIFGKGLTLFQTESTEKLCAMFNV--SDKRVIEQVQDAAKAVKLPAELDLNHCTLKFQDFRSSLGGHL
100022927   CEQE01148443.1  IAIDPAAETL----LHQAVSRCAARLGRERPDLDQNKGPFVSSLQDALVSSQNNGLSWLFGVGFQHWKEKSPKELIDEYKVPADQHGAVTQVKSFVDAIPLNPLFDTTH----YGEFRASVAGKV
100000004   CEVA01036528.1  HLVSHADQTL----IHQSISLCAHRIRQEDPALDPNKSGFIKQLQNNFLSQTFYGLSWLFGAGYVHFRECTANDLAIQYGIPNNCRDGIHQIKSFADAILPNTFFEKKH----YRKDSRSVGKKA
100021577   CEPS01188136.1  VCLNPSDNTL----LHQAVARCAGRILQEQPNLSPDKNRFINQLQDELVSSQNNGLSWLFGVGFKYWKEMSVDQLADDYKVKSTDLDALKQVKSFIDAIPLNPLFDTPH----YGEFRASVAGKM
Jpred Secondary structure   EEE------H----HHHHHHHHHHHHHHH--------HHHHHHHHHHHH-------EEEEE---HHHH---HHHHHHH--------HHHHHHHHHHHH--------------------HHHHHH
CONSENSUS_0.8               .........L....L.......A..L....P.L...K......L...L.......LSWLFG.G.......S...L...Y.V.......I.QVK....AI......D..H....Y.....SV....

100000002   CEPX01008730.1  RSWVANYWKRLLDLKSLLATTE-FTLPESISDPKAVSLFSGLLVDPQG-----LKKVADSL--PARLVSAEEAIDRLMGVGIP--TAADIAQVERVADEIGAFIGQVQQFNNQVKQKLENLQDAD
100020996 AUXO013399408.1  DSWTTNYLKRLDELNDLL-----LNLPKNLSLPDIFMIDGKDFIEYSGCNRDEIQQMIDFVVNEQNRIKLQESLNALLGKGNNQICSDDISTVKDFSEIVNSLHSFVQQIDNSLEQSSNEANSIF
100022927   CEQE01148443.1  RSWVANYWKRLLDLKSLLATTE-FTLPESISDPKAVSLFSGLLVDPQG-----LKKVADSL--PARLVSAEEAIDRLMGVGIP--TAADIAQVERVADEIGAFIGQVQQFNNQVKQKLENLQDAD
100000004   CEVA01036528.1  KSWISNYWQRLLQLQTWVDDHTWVTLPQELTEAQFKPLFRGLLVDAVE-----LMAIAERL--PQRLADCRDSLDCLMGKGPQAATKNDVEIVEKVREEIESFVGQIEQLGNQLRHQLENENN--
100021577   CEPS01188136.1  RSWVKNYWKRLLDLKSQLGTAN-INLPEGLDEQRAENLFSGLLIDSKG-----LRQVTDKL--PSRLKKAEDTIDRLMGDGNP--TSDDIEQVETVAAEISAFIGQVEQFNNQLEQRLENPLEGD
Jpred Secondary structure   HHHHHHHHHHHHHHHHH--------------HHHHHHHH-HEEE---------HHHHHHHH--HHHHHHHHHHHHHHH-------------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-------
CONSENSUS_0.8               .SW..NY.KRL..L...L.......LP..L.......L.....VD........L....D.L...........L..LMG.G.......DI..V......I......V.Q..N.L.Q.........

100000002   CEPX01008730.1  DE---EFLKGLKIEL-PSGDKEPPAINRISGGAPDAAAEISELEEKLQRLLDARSEHFQTISEWAEENAVTLDPIAAMVELER-LRLAERGATGDPEEYALRLLLQRIGRLANRVSPVSAGSIRE
100020996 AUXO013399408.1  SELKKKIEKNEKWDIWKNNLKKIPKLNKLSGGVPDAWKEIREIEQKFHEISENQKKHFTEVMEWIDAGNGTIDIFESRFKYDELLKKSKKNNLQSADELAFRSVLNKLGRFARQGNDLVCEKIKN
100022927   CEQE01148443.1  DE---EFLKGLKIEL-PSGDKEPPAINRISGGAPDAAAEISELEEKLQRLLDARSEHFQTISEWAEENAVTLDPIAAMVELER-LRLAERGATGDPEEYALRLLLQRIGRLANRVSPVSAGSIRE
100000004   CEVA01036528.1  DQVHRDNLHQLKNRL-PLDLRRPQALNKISGGVPDVAKSIRGLETQLDQVLKERRSHFGRLTKWAKECGITLDPLQPLIESEK-QRVAERGSAHDAKELAIRLLLQRIGRLGHRLSPTNATAIQE
100021577   CEPS01188136.1  DE---TFLKQLKIDL-PAEFKKPPAINRISGGSPDPTAEIAELEEKLDRLMSARKEHYETIAEWASANKVTLDPMEAMTTLEA-QRLTERGAEGDQEEFALRLLLQRIGRLANRLSPQGATAIRD
Jpred Secondary structure   HH---HHHHHHH-------------------------HHHHHHHHHHHHHHHH--HHHHHHHHHHH----------HHHHHH-HHHHHH-------HHHHHHHHHHHHHHHHHH------HHHHHH
CONSENSUS_0.8               .E.........K..L.....K....LN.ISGG.PD...EI..LE.K..........HF..I.EW......TLD........E...R...R.......E.A.R.LL.RIGR............I..

100000002   CEPX01008730.1  LLKP--VFMEEREFNLFFHNRLGSLYRSPYSTSRHQPFSIDVG-KAKAIDWIAGLDQISSDIEKALSGAGEALGDQLRDWINLAGFAISQRLRGLP--DTVPNALAQVRCPD---DVRIPPLLAM
100020996 AUXO013399408.1  WFKEQNIFDSSKDFNRYFINQKGFIFKHPSSKKDNSPYNLSANLLEKRYEVTNTVGALLEQCESDPAIVNDPFS--MRSLVEFRALWFSINISGISKEQHIPTKIAQPKLDDSTYQESVSPTLKY
100022927   CEQE01148443.1  LLKP--VFMEEREFNLFFHNRLGSLYRSPYSTSRHQPFSIDVG-KAKAIDWIAGLDQISSDIEKALSGAGEALGDQLRDWINLAGFAISQRLRGLP--DTVPNALAQVRCPD---DVRIPPLLAM
100000004   CEVA01036528.1  LLRP--VFAVKREFNLFFHNHMGALYRSPYSTSRHQPFQINVD-VAHGTDWIGTIETLIQNLFTQIQ--DDAL---LRDLVQLEGFVFSHKLRALP--GVIPSELARPNNLQ---QMGLPALLLV
100021577   CEPS01188136.1  LLRP--VFTEKREFNLFFHNRMGSLYRSPYSTSRHQPFTIDVA-VAKNTDWMDALDGIAETIMKGLSQAGDELSLRLRDWINISGFSLSQRLRGLP--DTVPGELALVRSAD---DVRIPPMLAL
Jpred Secondary structure   HHHH--HHH---HHHHHHHHH---------------EEEEEE-E------HHHHHHHHHHHHHHHHHH---HHHHHHHHHH------HHHHH------------EEEE------------HHHH
CONSENSUS_0.8               ......VF...REFN.FF.N..G.LYR.P.S.....PF.I.........D.......L...........D.....LR..V.......S..L.GL.....IP..LA.....D......I...L..

100000002   CEPX01008730.1  LLEEDDIARDVCLKAFNLYVSAINGCLFGALREGFIVRTRFQRIGTDQIHYVPKDKAWEYPDRLNTAKGPINAAVSSD---WIEK--DGAVIKPVETVRNL-SSTGFAGAG-VSEYLVQAPHDWY
100020996 AUXO013399408.1  RLEKEQITSSELNSIFTVYKSLLSGLSIRLSRNSFYLRTKFSWIGNNSLIYCPKETTWKIPAAYFKSDLWNEYKDKQILIVNEEY--DVDVVKTFESVYKIVKSKDNNEKNRILPLLKQLPHDWM
100022927   CEQE01148443.1  LLEEDDIARDVCLKAFNLYVSAINGCLFGALREGFIVRTRFQRIGTDQIHYVPKDKAWEYPDRLNTAKGPINAAVSSD---WIEK--DGAVIKPVETVRNL-SSTGFAGAG-VSEYLVQAPHDWY
100000004   CEVA01036528.1  LLQADQVHRETVLRVFNLYGSAINGYLFQALRPGFIVRAGFQRLETKKLRYVPKAQSWQYPDRLHHAKSAIKNSLSAG---WIKKNHQGAIL-PQKTLTALVKQKSLKDTG-VPEYLVQAPHDWY
100021577   CEPS01188136.1  QLEEDEVSREVCLKAFNLYVSAINGCLFRALREGFIVRTKFQRLERDVLSYVPKTKLWNYPQRLDTARGPIHSALAAA---WINK--EGSVIDPVETVTAL-SDTGFSDDG-IPEYLVQAPHDWY
Jpred Secondary structure   H--------HHHHHHHHHHHHHH--HHHHHHH---EEE-HHHHHHHHHHHH--------------------HHHHHHH---HH-------EE-HHHHHHH---------------HHH-------
CONSENSUS_0.8               .LE.D.I........F.LY.S.I.G......R..F.VR..F..I....L.Y.PK...W..P...........................VI...ETV..L...........V...L.Q.PHDW.

100000002   CEPX01008730.1  TPLDLRDV-AHLVTGLPVEK---NITKLKRL--TNRTAFRMVGASSFKTHLDSVLLSDKIKLGDFTIIIDQHYRQSVTY-GGKVKISYEPERLQVEAAVPVVDTRDRTVPEPDTLFDHIVAIDLG
100020996 AUXO013399408.1  FKLPFGASNAEKCKVLKLEK---NNKKFKPLSVSKDSLARLSGPSTYFNQIDEIMMNDESELSEMTLLADEPVRQQMS--NGKIEII--PDDYVMSLAIPIT-RSLKKGNTESFPFKNIVSIDQG
100022927   CEQE01148443.1  TPLDLRDV-AHLVTGLPVEK---NITKLKRL--TNRTAFRMVGASSFKTHLDSVLLSDKIKLGDFTIIIDQHYRQSVTY-GGKVKISYEPERLQVEAAVPVVDTRDRTVPEPDTLFDHIVAIDLG
100000004   CEVA01036528.1  VPIDLRGP-AIPIEGLTVGTEGPELTQLGPM--KDDCAFRAIGPSSFKSKIDAGLLPQDVKYGDMTLIFDQHYQQSISFANGTFSIQYQPTSLQVKAAIPVVDKRPRDTRNNSHLYDRIVAIDLG
100021577   CEPS01188136.1  TPIDLRDI-SKPVSGLPVKK---NITGLKRQ--KKQTAFRMVGPSSFKSHLDSTLLSEEVKLGDFTLIFDQYYKQRVSY-NGRVKITFEPDRLHVEAAVPVIDKRVRPSTEEDALFDHLLAIDLG
Jpred Secondary structure   ---------------------------HHH--HH---EEEE----------HHHH--------EEEEHHHHHHHHH------EEEEEE----E-EE----E----------HHHHHHHHHHH---
CONSENSUS_0.8               ..L......A.....L.V....................R..G.S.F...ID..LL....K..D.TLI.DQ...Q..S...G...I...P....V..AIPV.....R........F..IV.ID.G
RuvC-like_motifs            ...........................................................................................................................D..

100000002   CEPX01008730.1  ERSVGFAVFDIKSCLRTGEVKPIHDNNGNPVVGTVAVPSIRRLMKAVRSHRRRRQPNQKVNQTYSTALQNYRENVIGDVCNRIDTLMERYNAFPVLEFQIKNFQAGAKQLEIVYGS---------
100020996 AUXO013399408.1  EAGFAYAVFKLSDC-GNERAEPIAT-------GLIPIPSIRRLIHSVKKYRGKKQRIQNFNQKFDSTMFTLRENVTGDICGLIVALMKKYNAFPILEKQVGNLESGSKQLMLVYKAVNSKFLAAK
100022927   CEQE01148443.1  ERSVGFAVFDIKSCLRTGEVKPIHDNNGNPVVGTVAVPSIRRLMKAVRSHRRRRQPNQKVNQTYSTALQNYRENVIGDVCNRIDTLMERYNAFPVLEFQIKNFQAGAKQLEIVYGSVLHRYTFSG
100000004   CEVA01036528.1  ERKIGYAIFDLKQVLKSEQLEPMRE-DGKPLIGSISIRSIRGLMKAVQTHRNRRQPNYRIDQTYSKALMHYRESVIGDVCNAIDTLCARYGGFPVLESSVRNFEVGSAQLKTVYGSVSRRYTWSA
100021577   CEPS01188136.1  EKRVGYAVYDIKACLRTGDIKPLEDGDGKPIVGSVAVPSIRRLMKAVRSHRQQRQPNQKVNQTYSTALMNYRENVIGDVCNRIDTLMEKYNAFPVLESSVMNFEAGSRQLEMVYGSVLHRYTYSK
Jpred Secondary structure   ----EEEEEEEEHHHH-------------EEEEEE----HHHHHHHHHHHHHH------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-------------------EEEHH-HHHHHHH---
CONSENSUS_0.8               E....YAVF.L..........P..........G.I.I.SIR.L...V..HR..RQ.....NQ.Y...L...RENV.GDVC..I..L...Y..FPVLE..V.N.E.GS.QL..VY..V........
RuvC-like_motifs            ..............................................................................................E.............................

100000002   CEPX01008730.1  ----------------------------------------------------------------------------------------------------------------------------
100020996 AUXO013399408.1  VDMQNDQRRSWWYQGN-----SWNTPILRISNPNQSNNKNIVKNINGKKYEELKIYPGYSVSAYMTSCICHVCGRNALELLKNDDSTGKVKKYQINQDGEVTIGGEVIKLY-------------
100022927   CEQE01148443.1  VDAHKAKRREYWYNGE-----LWEHPYLMAKKWNEETNS------MSGAPKPVSLFPGVTVNAARTSQICHQCQRNPMSHLRG--LTGTI---EISSDGLLELDDGTIRLFETSDYDEDKFKQSR
100000004   CEVA01036528.1  VDAHKNQRQQYWLGGTKDKIPIWTHPYLMTREWDEKNSK------WSNRSKPLKMHPGVEVHPAGTSQICHQCKRNPIGALWN--VADTV---VLDDQGQLDLDDGTIRLNSGY-IDTTEIKRAR
100021577   CEPS01188136.1  IDAHTAKRKEYWYTGE-----YWDHPYLMAHKWNERTRS------YSGSLSALTLYPGVMVHPAGTSQRCHQCKRNPMVEIKQ--LTGQV---EINADGSLELDDGTICLYEGYDYSPEEYKKAK
Jpred Secondary structure   ----------EEE---------------EH--------------------EEEEE----EE----------------HHHEEE-----EE---EE-------------EEEEE-------HHHHHH
CONSENSUS_0.8               VD.....R...W..G.......W..P.L...........................PG..V....TS..CH.C.RN....L........V........G........I.L...............

100000002   CEPX01008730.1  --------------------------------------------------------------------------------------------
100020996 AUXO013399408.1  RKPDRLTPVKNLA--KKGNRERTYASINERAP-#--MSKDTTQSRYFCVFKNCPCHNKEQHADVNAAINIGRRFLKDCILDDNKEKD----
100022927   CEQE01148443.1  REKRRLDANVLLSGRHRAEYIYTVAKRNLRRPPKNVMTKDTTQSRYTCLYKNCS---WTGHADENAAINIGRRYLAERIDMPASKTKAAV-
100000004   CEVA01036528.1  RKKIRLPENKPLTGSHKTSHVRAVARRNLRQPPKSTRAKDTTQSRYTCLYVDCG---HECHADENAAINIGRKYLQERIHIEASRQALSTR
100021577   CEPS01188136.1  REKRRLDPNVPLSGRHQAKHVSAVAKRNLRRPTVSMMSGDTTQARYVCLYTDCD---FTGHADENAAINIGWKYLTERIALSESKDKAGV-
Jpred Secondary structure   HH----------------EE-HHHHH------EEEE-------EEEEEEE-------------------HHHHHHHHHH----------
CONSENSUS_0.8               R...RL.....L...........A..N.R.P.......DTTQ.RY.C....C.......HAD.NAAINIG...L...I............
RuvC-like_motifs            .................................................................D..........................
```

Figure 2.S5: Multiple alignment of C2c3 protein family

The alignment was built using MUSCLE program. Each sequence is labelled with local assigned number and the Genbank ID for metagenomics contig coding for respective C2c3 protein. Secondary structure was predicted by Jpred and shown underneath the alignment (designations: H- alpha helix, E – beta strand). CONSENSUS was calculated for each alignment column by scaling the sum-of-pairs score within the column between those of a homogeneous column (the same residue in all aligned sequences) and a random column with homogeneity cutoff 0.8. Active site motifs of RuvC-like domain are shown below alignment for the C-terminal domain.

10.1.6 Figure 2.S6

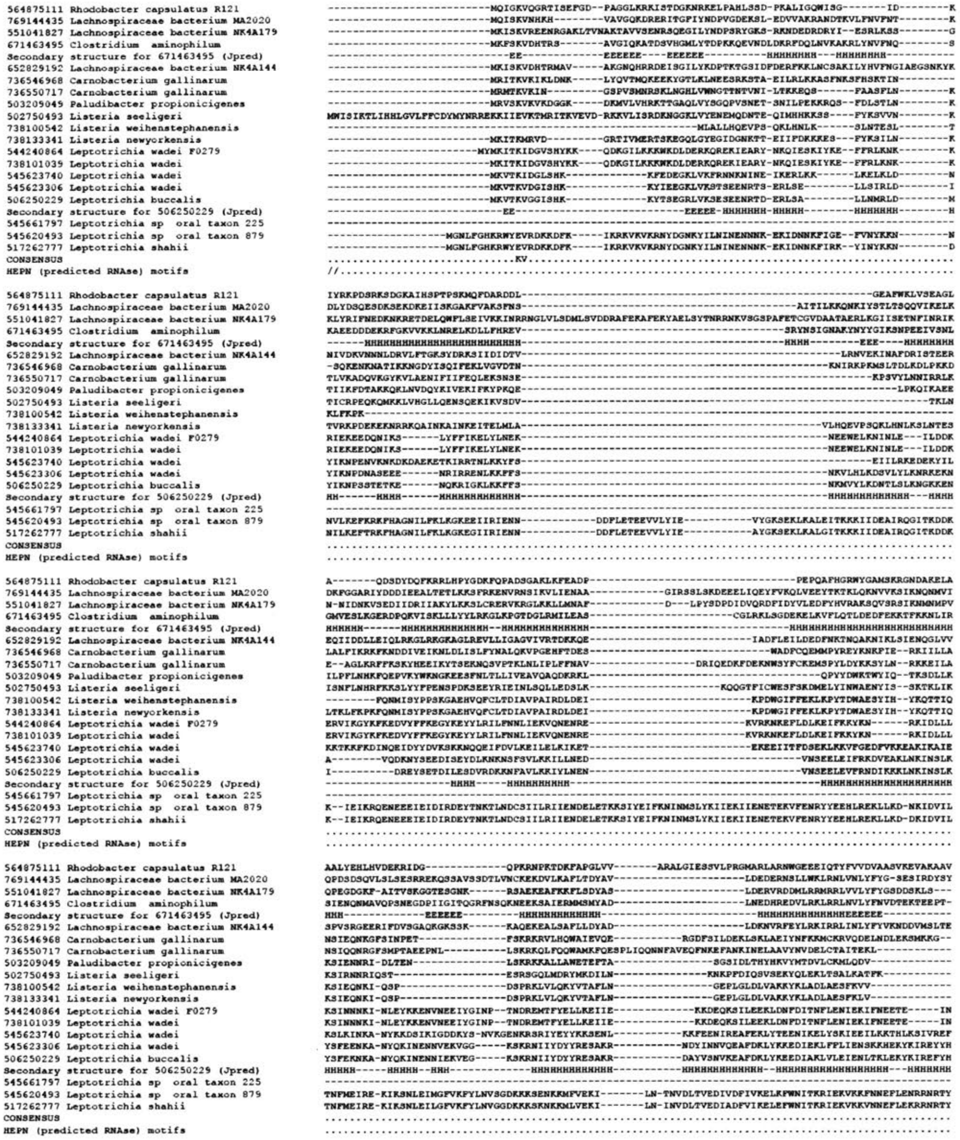


Figure 2.S6: Multiple alignment of C2c2 protein family.

The alignment was built using MUSCLE program and modified manually on t he basis of local PSIBLAST pairwise alignments. Each sequence is labelled with GenBank Identifier (GI) number and systematic name of an organism. Secondary structure was predicted by Jpred and shown underneath the sequence which was used as a query (designations: H- alpha helix, E – be ta strand). CONSENSUS was calculated for each alignment column by scaling the sum-of-pairs score within the column between those of a homogeneous column (the same residue in all aligned sequences) and a random column with homogeneity cutoff 0.8. A ctive site motifs of HEPN domain are shown below alignment.

10.1.7 Figure 2.S7

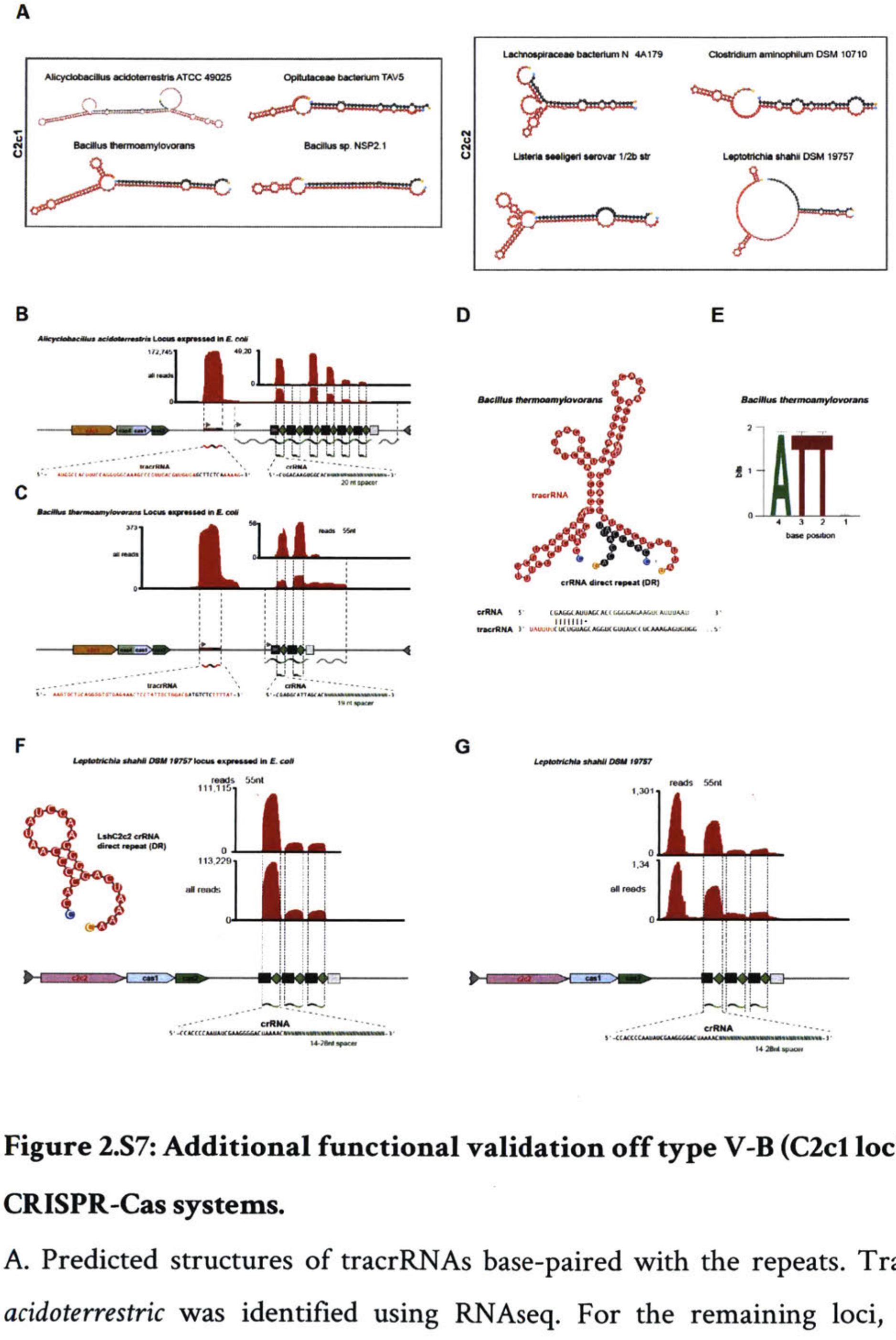

A
C2c1
Alicyclobacillus acidoterrestris ATCC 49025
Opitutaceae bacterium TAV5
Bacillus thermoamylovorans
Bacillus sp. NSP2.1
C2c2
Lachnospiraceae bacterium N 4A179
Clostridium aminophilum DSM 10710
Listeria seeligeri serovar 1/2b str
Leptotrichia shahii DSM 19757
B
Alicyclobacillus acidoterrestris Locus expressed in E. coli
172,745
all reads
0
49,20
0
tracrRNA
crRNA
20 nt spacer
C
Bacillus thermoamylovorans Locus expressed in E. coli
373
all reads
0
58
0
reads 55nt
tracrRNA
crRNA
19 nt spacer
D
Bacillus thermoamylovorans
tracrRNA
crRNA direct repeat (DR)
crRNA 5' CGAGGCAUUAGCACCGGGGAGAAGUCAUUUAAU 3'
tracrRNA 3' 5'
E
Bacillus thermoamylovorans
2
1
0
bits
4 3 2 1
base position
F
Leptotrichia shahii DSM 19757 locus expressed in E. coli
LshC2c2 crRNA direct repeat (DR)
reads 55nt
111,115
0
113,229
all reads
0
c2c2
cas1
crRNA
5'-CCACCCCAAUAUCGAAGGGGACUAAAACNNNNNNNNNNNNNNNNNNNNNNNN-3'
14-28nt spacer
G
Leptotrichia shahii DSM 19757
reads 55nt
1,301
0
1,34
all reads
0
c2c2
cas1
crRNA
5'-CCACCCCAAUAUCGAAGGGGACUAAAACNNNNNNNNNNNNNNNNNNNNNNNN-3'
14-28nt spacer

Figure 2.S7: Additional functional validation off type V-B (C2c1 loci) and type VI (C2c2 loci) CRISPR-Cas systems.

A. Predicted structures of tracrRNAs base-paired with the repeats. TracrRNA for *Alicyclobacillus acidoterrestric* was identified using RNAseq. For the remaining loci, putative tracrRNAs were identified based on presence of an anti-direct repeat (DR) sequence. Anti-DRs were identified using Geneious (www.geneious.com) by searching for sequences within each respective CRISPR locus that are highly similar to the DR. The 5' and 3' ends of each putative tracrRNA was determined by computational prediction of bacterial transcription start and termination sites using BPROM (www.softberry.com) and ARNOLD (rna.ig-mors.u-psud.fr/toolbox/arnold/) respectively. Co-folding predictions were generated using Geneious. 5' ends are colored blue and 3' ends are colored orange.

B. Heterologous expression of the Alicyclobacillus acideoterrestris C2c1 locus in pACYC-84 transformed into E. coli shows identical results to the expression observed in the endogenous strain (Fig. 4A). Processed crRNAs have a 5' 14-nt DR and 20-nt spacer and a putative 79-nt tracrRNA is expressed robustly.

C. The Bacillus thermoamylovorans locus was heterologously expressed in E. coli. The putative tracrRNA is robustly expressed and processed to 91 nt. Processed crRNAs are also present with a 5' 14 nt DR and 19 nt spacer.

D.In silico co-folding of the crRNA direct repeat and putative tracrRNA shows stable secondary structure and complementarity between the two RNAs. 5' bases are colored blue and 3' bases are colored orange.

E. Depletion from the 5' left PAM library reveals a 5' ATTN PAM. Depletion is measured as the negative log2 fold ratio and PAMs above a threshold of 3.5 are used to calculate the entropy score at each position.

F. In silico folding of the L. shahii crRNA DR predicts stable secondary structure and RNA-sequencing of the L. shahii DSM 19757 locus expressed in E. coli. Processing of the CRISPR array in the 3' to 5' direction (direction of the locus) is observed. crRNAs are processed to have a 5' DR that is 28nt in length and spacers with lengths 14-28 nt.

G. RNA-sequencing of the endogenous L. shahii DSM 19757 C2c2 locus shows results to those in E. coli (F).

10.2 Chapter 3 Supplementary Figures

10.2.1 Figure 3.S1

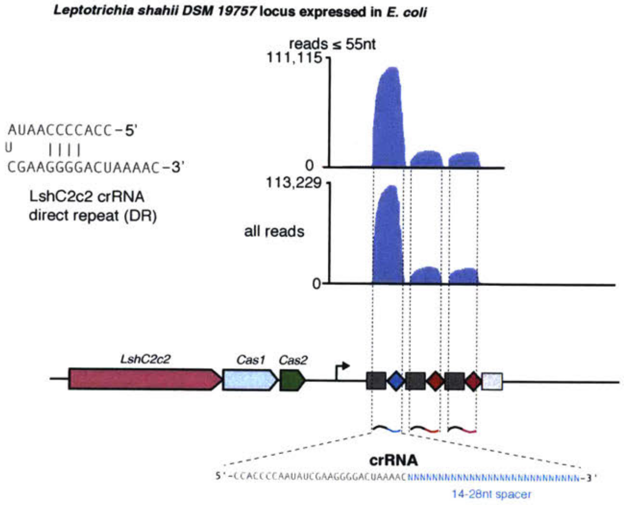


Figure 3.S1: RNA-sequencing of the *Leptotrichia shahii* locus heterologously expressed in *E. coli* and spacer analysis.

Heterologous expression of the LshC2c2 locus reveals processing of the array. Insert: *In silico* co-folding analysis of a mature direct repeat.

10.2.2 Figure 3.S2

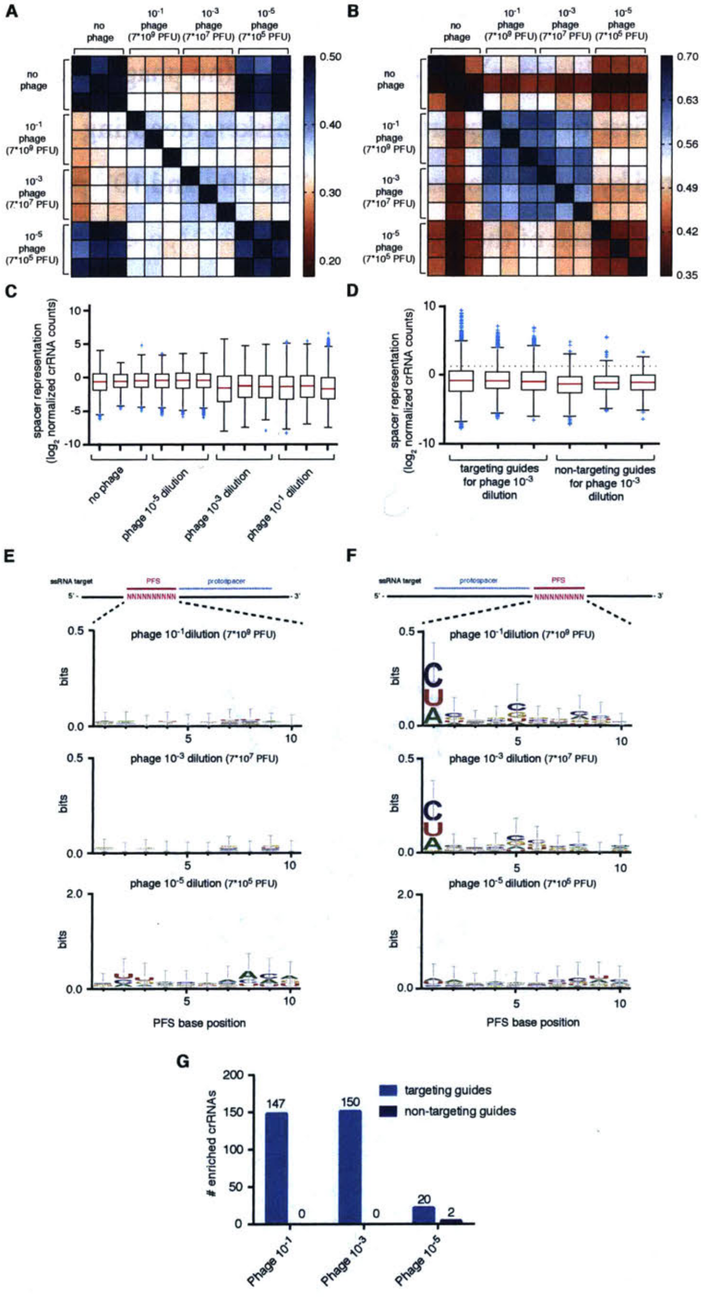


Figure 3.S2: MS2 phage screen replicates show agreement and do not have a 5' PFS

(A) Rank correlation (Kendall) of normalized crRNA count distributions between replicate conditions in the screen. (B) Information coefficient representing the mutual information between the normalized crRNA count distributions of replicate conditions in the screen. This metric highlights how similar the 10^{-1} and 10^{-3} dilution conditions are and how little information is shared between $10^{-1}/10^{-3}$ and 10^{-5}/no phage groups. A higher information coefficient represents strong correlation and is computed as previously described (Konermann et al., 2015; Liberzon et al., 2015). (C) Box plot showing the distribution of normalized crRNA frequencies for the phage-treated conditions (10^{-1}, 10^{-3}, and 10^{-5} dilutions) and control screen (no phage) biological replicates ($n = 3$). The box extends from the first to third quartile with whiskers denoting 1.5 times the interquartile range. The mean is indicated by the red horizontal bar. The 10^{-1} and 10^{-3} phage dilution distributions are significantly different than each of the control replicates (****, $p < 0.0001$). (C) Box plot showing the distribution of normalized crRNA frequencies (normalized to no phage condition) for targeting and non-targeting guides from the 10^{-3} diluted phage conditions (n=3). (D) (E) Sequence logo of 5' sequences from enriched spacers in each of the phage dilutions. (F) Sequence logos of 3' sequences from enriched spacers in the 10^{-1} and 10^{-3} phage dilutions show a PFS. (G) The number of targeting and non-targeting control spacers that are consistently enriched (the exact number is above each bar). A spacer is considered enriched only if it has a $\log_2$ normalized crRNA fold change > 1.25 in all three replicates. Within the consistently enriched crRNAs, there are 84 sequences that are shared between the 10^{-1} and 10^{-3} conditions.

10.2.3 Figure 3.S3

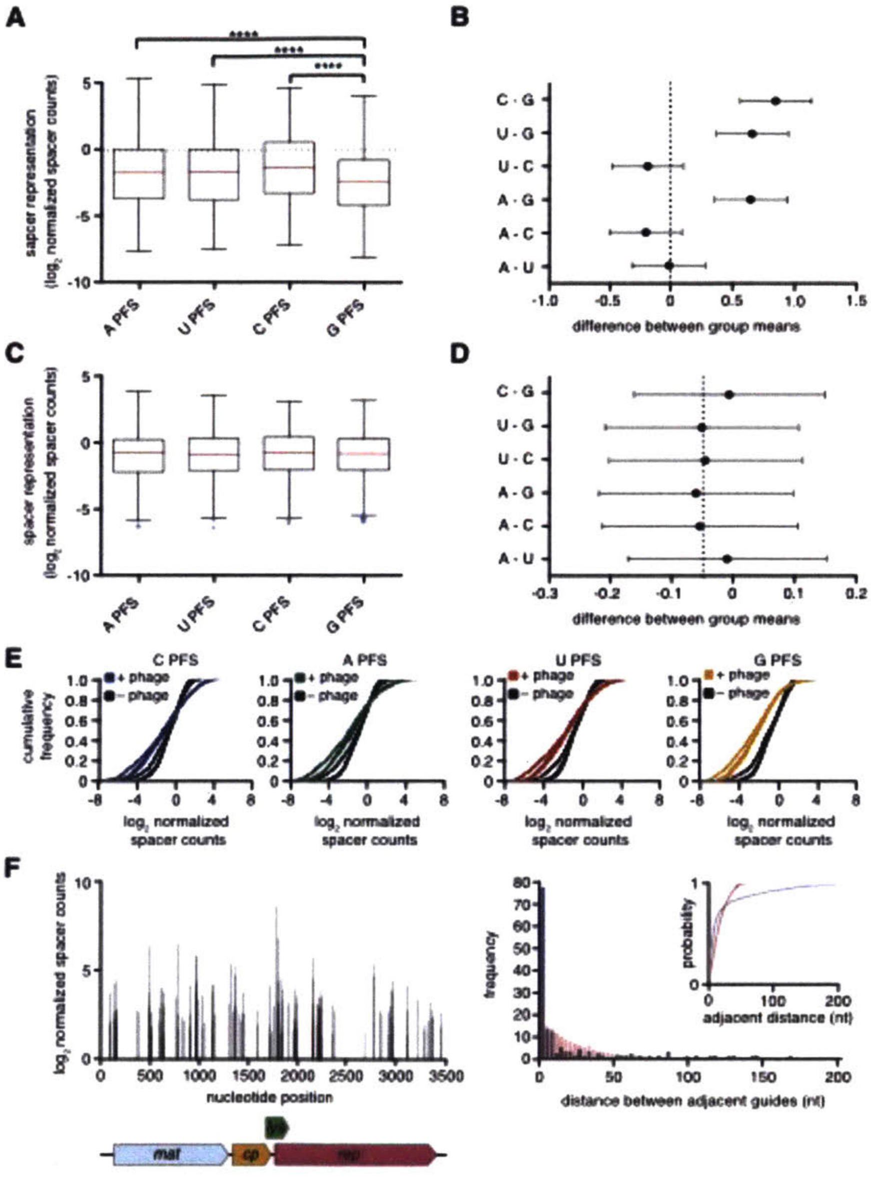


Figure 3.S3: MS2 phage screen spacer representation across each PFS.

(A) Box plot showing the distribution of spacer frequencies with spacers grouped by their 3' PFS for 10^{-3} phage treated conditions. Box extends from the first to third quartile with the whiskers denoting 1.5 times the interquartile range. ****, p < 0.0001. (B) Multiple comparison test (ANOVA with Tukey correction) between all possible PFS pairs for the 10^{-3} phage treated spacer distributions. Plotted are the 95% confidence intervals for difference in means between the compared PFS pairs. (C) Box plot showing the distribution of spacer frequencies with spacers grouped by their 3' PFS for non-phage treated conditions. Box extends from the first to third quartile with the whiskers denoting 1.5 times the interquartile range. (D) Multiple comparison test (ANOVA with Tukey correction) between all possible PFS pairs for the non-phage treated spacer distributions. Plotted are the 95% confidence intervals for difference in means between the compared PFS pairs. (E) Cumulative frequency plots for the $\log_2$ normalized spacer counts. Spacers are separated by respective PFS to show the enrichment differences between the 10^{-3} phage and control PFS distributions. (F) The enriched spacers from the 10^{-3} dilution condition are plotted according to their position along the MS2 genome. The corresponding gene positions are mapped below. (G) Frequency distributions of the non-redundant nearest-neighbor pairwise-distances between all 150 enriched guides in the 10^{-3} phage dilution condition (blue) and in a bootstrapped simulation of 150 randomly chosen guides (n=10,000) (red). A significant difference between both distributions is observed using a two-sample Kolmogorov-Smirnov statistic (****, $p < 0.0001$). Inset: the cumulative frequency distributions for the non-redundant nearest-neighbor pairwise-distance distributions for the 10^{-3} phage dilution condition (blue) and bootstrapped simulation (red).

10.2.4 Figure 3.S4

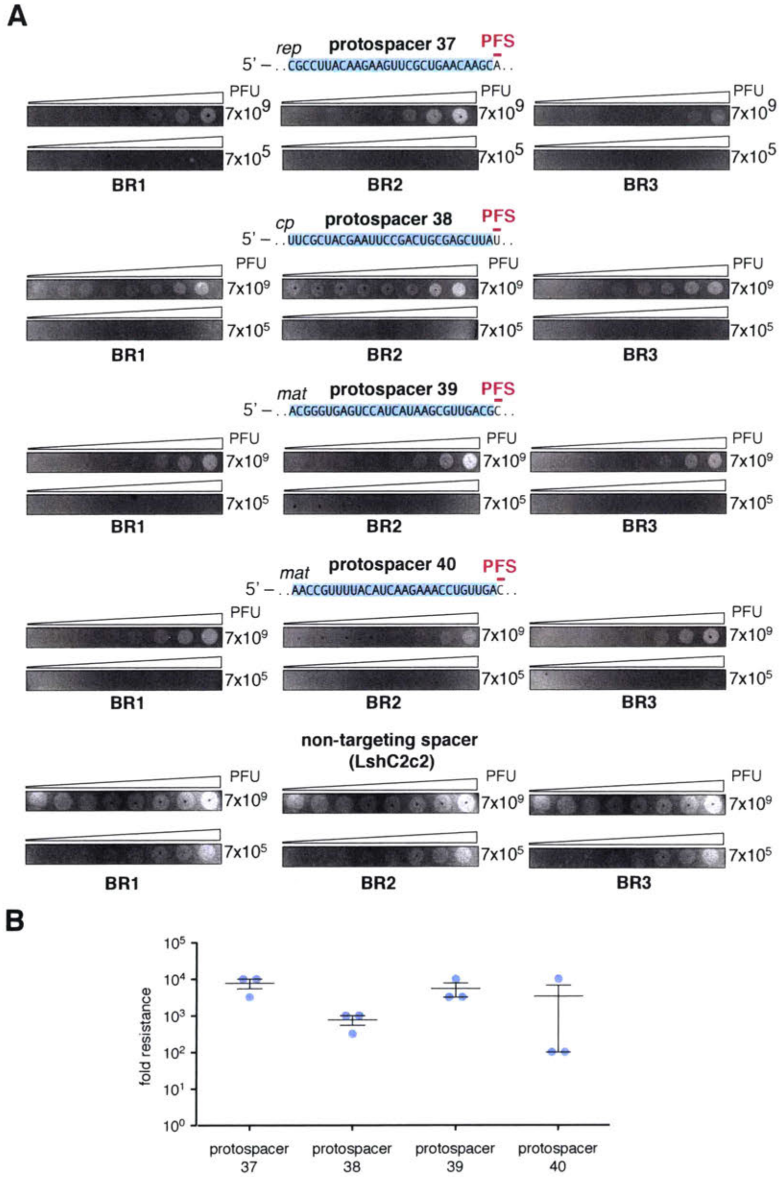


Figure 3.S4: Top hits from MS2 phage screen show interference in plaque assay

(A) Images from validation of MS2 screen by plaque assay showing reduced plaque formation in top hits. Phage dilutions were spotted on bacteria plates at decreasing numbers of plaque forming units (PFU). Spacer targets are shown above images; biological replicates are labeled BR1, BR2, or BR3. Non-targeting control is the native LshC2c2 locus. (B) Quantitation of MS2 plaque assay demonstrating interference by top hits. Interference was quantified by highest dilution without plaques. Bars plotted are the mean ± s.e.m.

10.2.5 Figure 3.S5

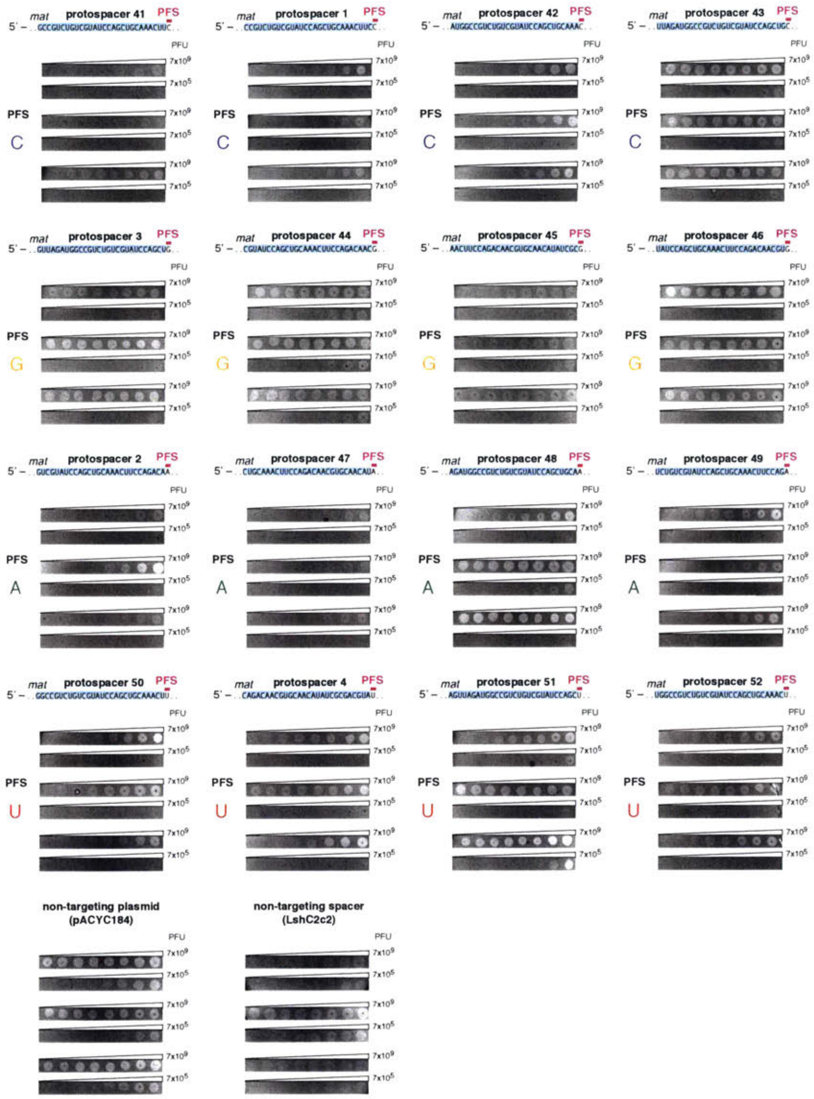

mat protospacer 41 PFS
5' – ..GCCGUCUGUCGUAUCCAGCUGCAAACUUC..
mat protospacer 1 PFS
5' – ..CCGUCUGUCGUAUCCAGCUGCAAACUUCC..
mat protospacer 42 PFS
5' – ..AUGGCCGUCUGUCGUAUCCAGCUGCAAAC..
mat protospacer 43 PFS
5' – ..UUAGAUGGCCGUCUGUCGUAUCCAGCUGC..
PFS
C
mat protospacer 3 PFS
5' – ..GUUAGAUGGCCGUCUGUCGUAUCCAGCUG..
mat protospacer 44 PFS
5' – ..CGUAUCCAGCUGCAAACUUCCAGACAACG..
mat protospacer 45 PFS
5' – ..AACUUCCAGACAACGUGCAACAUAUCGCG..
mat protospacer 46 PFS
5' – ..UAUCCAGCUGCAAACUUCCAGACAACGUG..
PFS
G
mat protospacer 2 PFS
5' – ..GUCGUAUCCAGCUGCAAACUUCCAGACAA..
mat protospacer 47 PFS
5' – ..CUGCAAACUUCCAGACAACGUGCAACAUA..
mat protospacer 48 PFS
5' – ..AGAUGGCCGUCUGUCGUAUCCAGCUGCAA..
mat protospacer 49 PFS
5' – ..UCUGUCGUAUCCAGCUGCAAACUUCCAGA..
PFS
A
mat protospacer 50 PFS
5' – ..GGCCGUCUGUCGUAUCCAGCUGCAAACUU..
mat protospacer 4 PFS
5' – ..CAGACAACGUGCAACAUAUCGCGACGUAU..
mat protospacer 51 PFS
5' – ..AGUUAGAUGGCCGUCUGUCGUAUCCAGCU..
mat protospacer 52 PFS
5' – ..UGGCCGUCUGUCGUAUCCAGCUGCAAACU..
PFS
U
non-targeting plasmid (pACYC184)
non-targeting spacer (LshC2c2)
PFU
7x10^9
7x10^5

Figure 3.S5: MS2 plaque assay validates the 3' H PFS.

Four spacers for each possible 3' PFS (A, G, C, and U) are cloned into the pLshC2c2 vector and tested for MS2 phage restriction in a plaque forming assay. The images show significantly reduced plaque formation for A, C, and U PFSs, and less restriction for the G PFS. Phage dilutions were spotted on bacteria plates at decreasing numbers of plaque forming units (PFU). Spacer targets are shown above images; three biological replicates are vertically stacked under each protospacer sequence. Non-targeting controls are the native LshC2c2 locus and the pACYC184 backbone.

10.2.6 Figure 3.S6

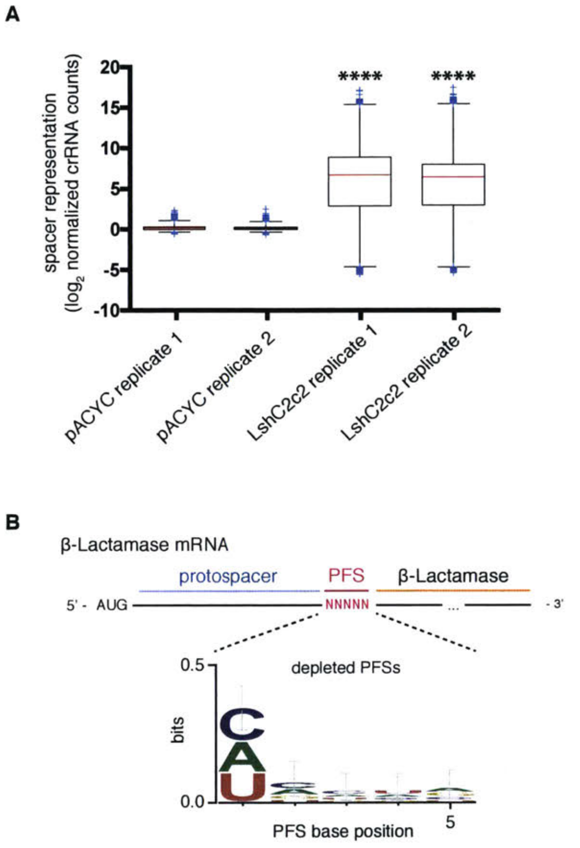


Figure 3.S6. A PFS screen in the β-lactamase mRNA reveals a 3' H PFS

(A) Comparison of the normalized crRNA count distributions for the pACYC control and LshC2c2 replicates (n=2). Box plots are shown with boxes extending from the first quartile to third quartile and whiskers denoting 1.5 times the interquartile range. Significant enrichment and depletion is observed in the LshC2c2 replicates (****, $p < 0.0001$). (B) A 3' PFS is observed from the depleted PFSs ($-\log_2$ normalized PFS count fold change >6.0).

10.2.7 Figure 3.S7

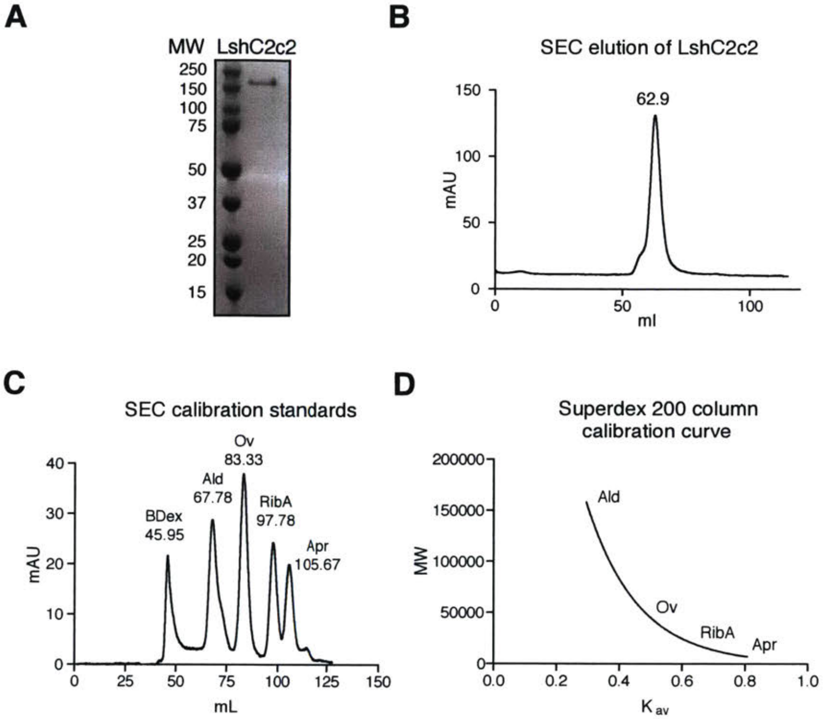


Figure 3.S7: Protein purification of LshC2c2.

(A) Coomassie blue stained acrylamide gel of purified LshC2c2 stepwise purification. A strong band just above 150 kD is consistent with the size of LshC2c2 (171 kD). (B) Size exclusion gel filtration of LshC2c2. LshC2c2 eluted at a size approximately >160 kD (62.9 mL). (C) Protein standards used to calibrate the Superdex 200 column. BDex = Blue Dextran (void volume), Ald = Aldolase (158 kD), Ov = Ovalbumin (44 kD), RibA = Ribonuclease A (13.7 kD), Apr = Aprotinin (6.5 kD). (D) Calibration curve of the Superdex 200 column. K_{av} is calculated as (elution volume – void volume)/(geometric column volume – void volume). Standards were plotted and fit to a logarithmic curve.

10.2.8 Figure 3.S8

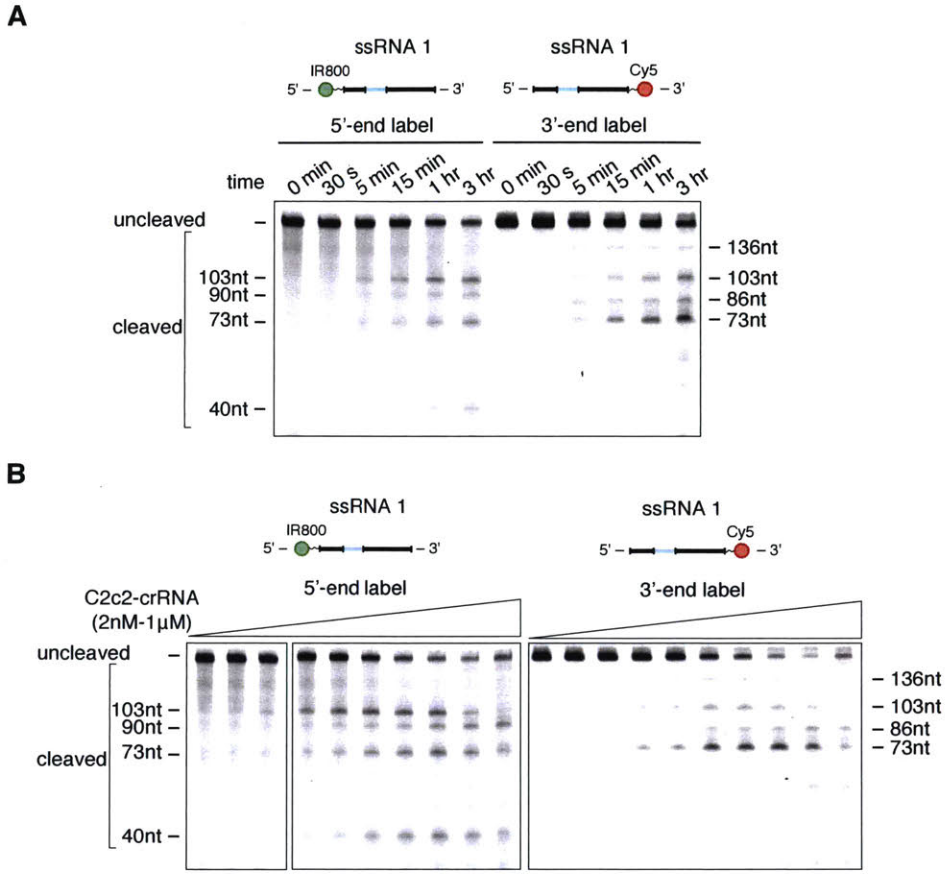


Figure 3.S8: Further *in vitro* characterization of the RNA cleavage kinetics of LshC2c2.

(A) Denutaring gel of a time series of LshC2c2 ssRNA cleavage using a 5'- and 3-end-labeled target 1.

(B) A denaturing gel after 1 hour of RNA-cleavage of 5'- and 3-end-labeled target 1 using LshC2c2-crRNA complex that is serially diluted in half-log steps. Reported band lengths are matched from RNA sequencing.

10.2.9 Figure 3.S9

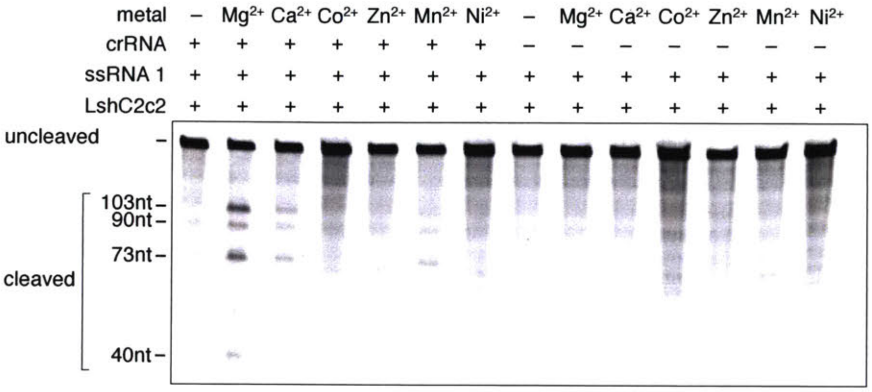

metal – Mg2+ Ca2+ Co2+ Zn2+ Mn2+ Ni2+ – Mg2+ Ca2+ Co2+ Zn2+ Mn2+ Ni2+
crRNA + + + + + + + – – – – – – –
ssRNA 1 + + + + + + + + + + + + + +
LshC2c2 + + + + + + + + + + + + + +
uncleaved
cleaved
103nt
90nt
73nt
40nt

Figure 3.S9: Characterization of the metal dependence of LshC2c2 RNA cleavage.

A variety of divalent metal cations are supplemented for the LshC2c2 cleavage reaction using 5'-end-labeled target 1 incubated for 1 hour. Significant cleavage is only observed for Mg^{+2}. Weak cleavage is observed for Ca^{+2} and Mn^{+2}. Reported band lengths are matched from RNA sequencing.

10.2.10 Figure 3.S10

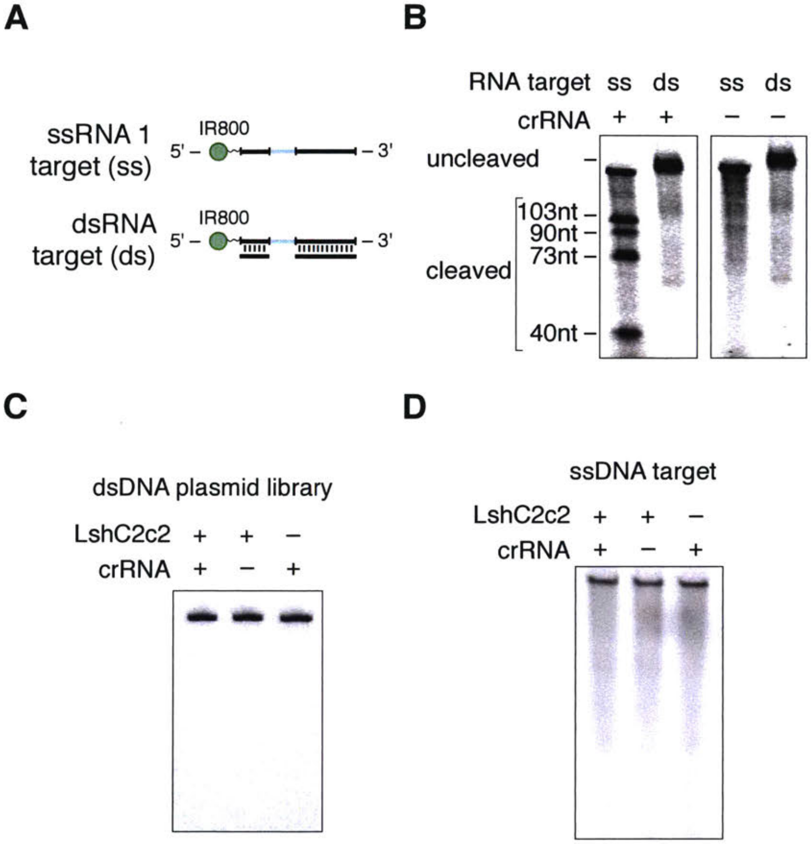


Figure 3.S10: LshC2c2 has no observable cleavage activity when using dsRNA, dsDNA, or ssDNA substrates.

(A) A schematic of the partial dsRNA target. 5'-end-labeled target 1 is annealed to two shorter RNAs that are complementary to the regions flanking the protospacer site. This partial dsRNA is a more stringent test for dsRNA cutting since it should still allow for LshC2c2 complex binding to ssRNA. (B) LshC2c2 cleavage activity after 1 hour of incubation with a dsRNA target shown in (A) compared to the ssRNA target 1. No cleavage is observed when using the dsRNA substrate. Reported band lengths are matched from RNA sequencing. (C) LshC2c2 cleavage of a dsDNA plasmid library incubated for 1 hour. A plasmid library was generated to have seven randomized nucleotides 5' of protospacer 14 to account for any sequence requirements for dsDNA cleavage. No cleavage is observed for this dsDNA library. (D) A ssDNA version of target 1 is tested for cleavage by LshC2c2 after 1 hour of incubation. No cleavage is observed.

10.2.11 Figure 3.S11

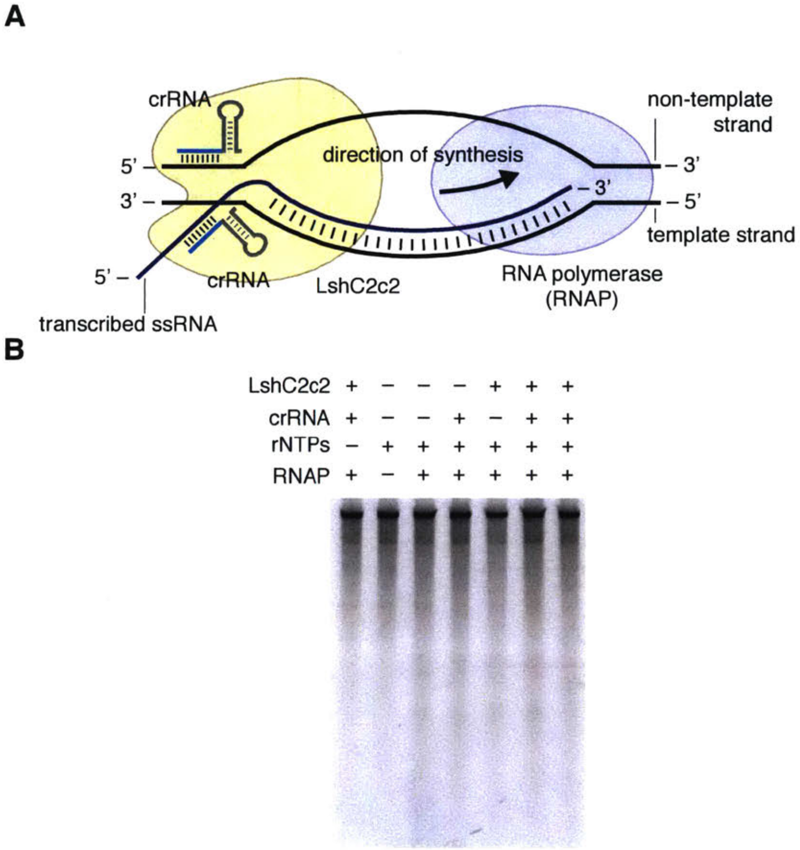


Figure 3.S11: LshC2c2 has no observable cleavage activity on dsDNA targets in a co-transcriptional cleavage assay

(A) Schematic of co-transcriptional cleavage assay. C2c2 was incubated with *E. coli* RNA polymerase (RNAP) elongation complexes and rNTP as previously described (Samai et al., 2015). (B) LshC2c2 cleavage of DNA target after co-transcriptional cleavage assay. No cleavage is observed.

10.2.12 Figure 3.S12

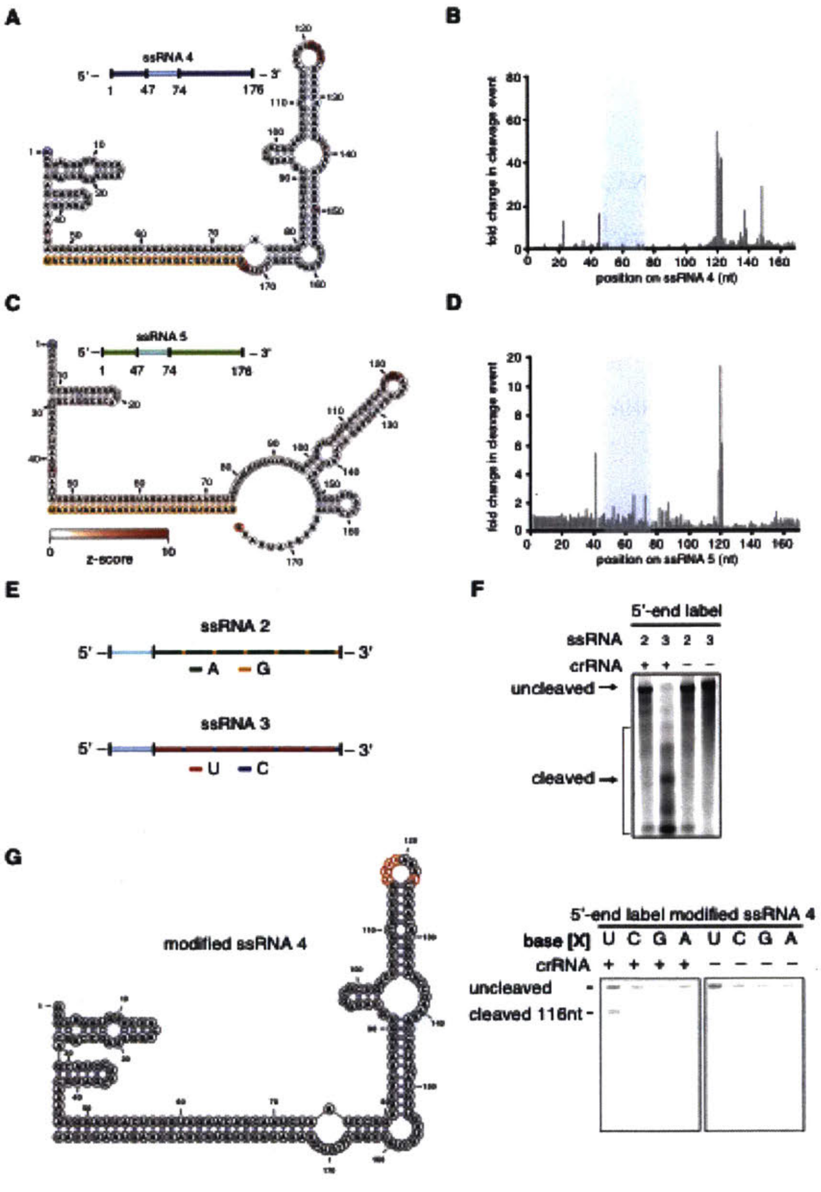


Figure 3.S12: Figure 4. LshC2c2 prefers cleavage at uracil residues.

(A,C) The cleavage sites of non-homopolymer ssRNA targets 4 (A), and 5 (C) were mapped with RNA-sequencing of the cleavage products. The frequency of cleavage at each base is colored according to the z-score and shown on the predicted crRNA-ssRNA co-fold secondary structure. Fragments used to generate the frequency analysis contained the complete 5' end. The 5' and 3' end of the ssRNA target are indicated by blue and red outlines, on the ssRNA and secondary structure, respectively. The 5' and 3' end of the spacer (outlined in yellow) is indicated by the blue and orange residues highlighted respectively. (B,D) Plot of the frequencies of cleavage sites for each position of ssRNA targets 4 and 5 for all reads that begin at the 5' end. The protospacer is indicated by the blue highlighted region. (E) Schematic of homopolymer ssRNA targets. The protospacer is indicated by the light blue bar. Homopolymer stretches of A (green) and U (red) bases are interspaced by individual bases of G (orange) and C (purple). (F) Denaturing gel showing C2c2-crRNA-mediated cleavage patterns of each homopolymer after 3 hours of incubation. (G) Schematic of ssRNA 4 modified with a hompolymer stretch in the highlighted loop (red) for each of the four possible nucleotides (*left*). Denaturing gel showing C2c2-crRNA-mediated cleavage for each of the four possible homopolymer targets after 3 hours of incubation. Reported band lengths are matched from RNA sequencing.

10.2.13 Figure 3.S13

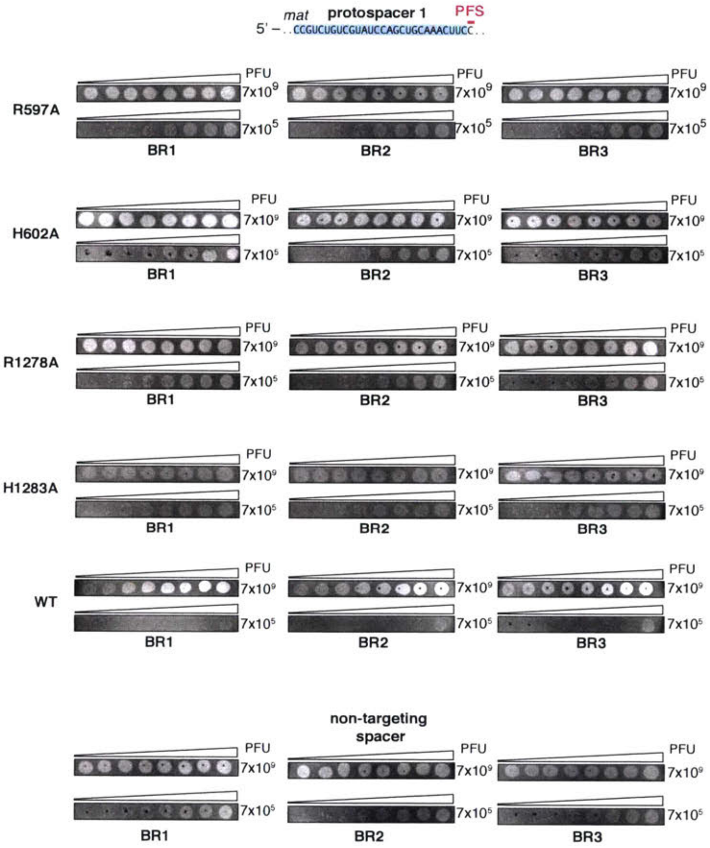


Figure 3.S13: MS2 restriction assay reveals that single HEPN mutants abrogate LshC2c2 activity.

All four possible single HEPN mutants were generated in the pLshC2c2 vector (R597A, H602A, R1278A, and H1283A) with protospacer 1. Images from plaque assay testing these HEPN mutant loci show similar plaque formation to the non-targeting locus and is significantly higher than the WtC2c2 locus. Phage dilutions were spotted on bacteria plates at decreasing numbers of plaque forming units (PFU). Spacer targets are shown above images; biological replicates are labeled BR1, BR2, or BR3. Non-targeting control is the native LshC2c2 locus.

10.2.14 Figure 3.S14

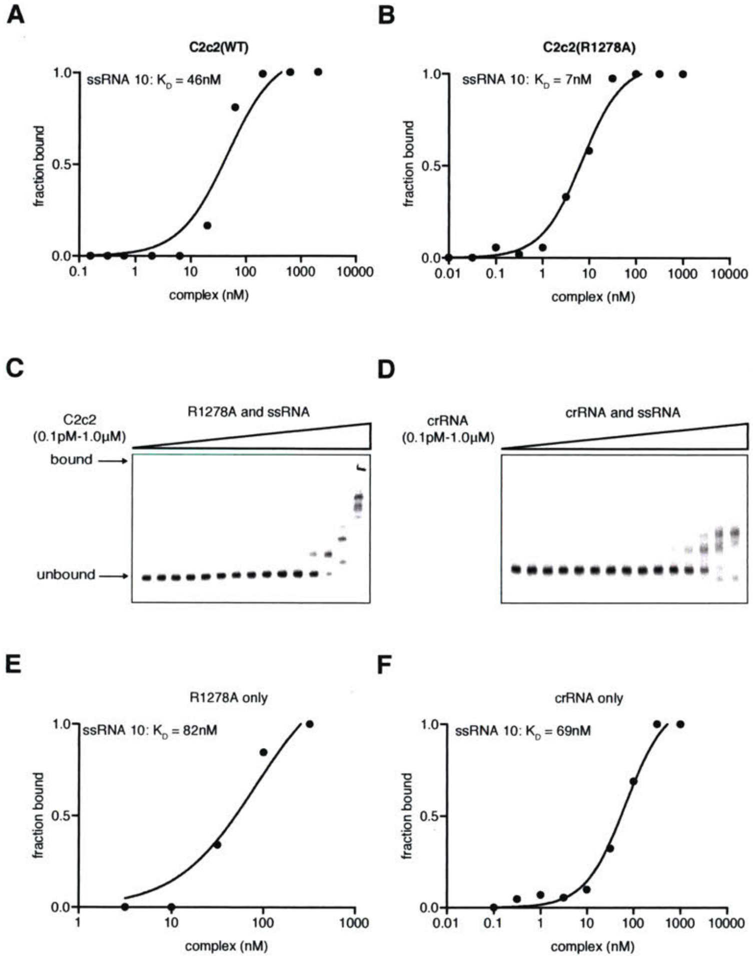


Figure 3.S14: Quantitation of LshC2c2 binding

(A) Calculation of binding affinity for wildtype LshC2c2-crRNA complex and on-target ssRNA. Fraction of protein bound was quantified by densitometry from Fig. 4D and K_D was calculated by fitting to binding isotherm. (B) Calculation of binding affinity for HEPN mutant R1278A LshC2c2-crRNA complex and on-target ssRNA. Fraction of protein bound was quantified by densitometry from Fig. 4E and K_D was calculated by fitting to binding isotherm. (C) Electrophoretic mobility shift assay with HEPN mutant R1278A LshC2c2 against on-target ssRNA in the absence of crRNA. EDTA is supplemented to reaction condition. (D) Electrophoretic mobility shift assay crRNA against on-target ssRNA. EDTA is supplemented to reaction condition. (E) Calculation of binding affinity for HEPN mutant R1278A LshC2c2 and on-target ssRNA in the absence of crRNA. Fraction of protein bound was quantified by densitometry from Fig. S12C and K_D was calculated by fitting to binding isotherm. (F) Calculation of binding affinity for crRNA and on-target ssRNA. Fraction of crRNA bound was quantified by densitometry from Fig. S12D and K_D was calculated by fitting to binding isotherm.

10.2.15 Figure 3.S15

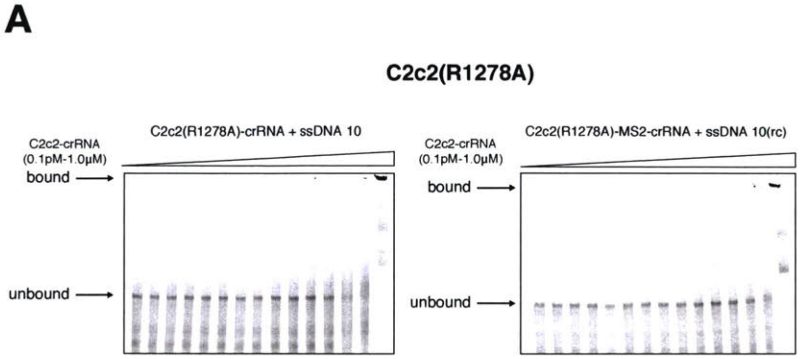


Figure 3.S15. LshC2c2-crRNA complex has little binding affinity for ssDNA targets.

Electrophoretic mobility shift assay with HEPN mutant R1278A LshC2c2 against on-target ssDNA and non-complementary ssDNA (reverse complement). EDTA is supplemented to reaction condition.

10.2.16 Figure 3.S16

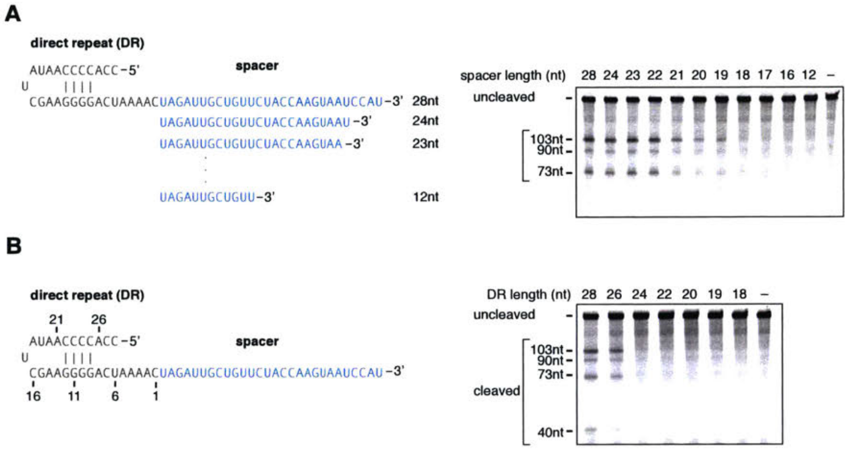

A
direct repeat (DR)
AUAACCCCACC-5'
spacer
U
CGAAGGGGACUAAAACUAGAUUGCUGUUCUACCAAGUAAUCCAU-3' 28nt
UAGAUUGCUGUUCUACCAAGUAAU-3' 24nt
UAGAUUGCUGUUCUACCAAGUAA-3' 23nt
UAGAUUGCUGUU-3' 12nt
spacer length (nt) 28 24 23 22 21 20 19 18 17 16 12 –
uncleaved -
103nt -
90nt -
73nt -
B
direct repeat (DR)
21 26
AUAACCCCACC-5'
spacer
U
CGAAGGGGACUAAAACUAGAUUGCUGUUCUACCAAGUAAUCCAU-3'
16 11 6 1
DR length (nt) 28 26 24 22 20 19 18 –
uncleaved -
103nt -
90nt -
73nt -
cleaved
40nt -

Figure 3.S16. Spacer and direct repeat lengths affect the RNA-guided RNase activity of LshC2c2.

(A) Denaturing gel showing crRNA-guided cleavage of ssRNA 1 as a function of spacer length after 3 hours of incubation. Reported band lengths are matched from RNA sequencing. (B) Denaturing gel showing crRNA-guided cleavage of ssRNA 1 as a function of the direct repeat length after 3 hours of incubation. Reported band lengths are matched from RNA sequencing.

10.2.17 Figure 3.S17

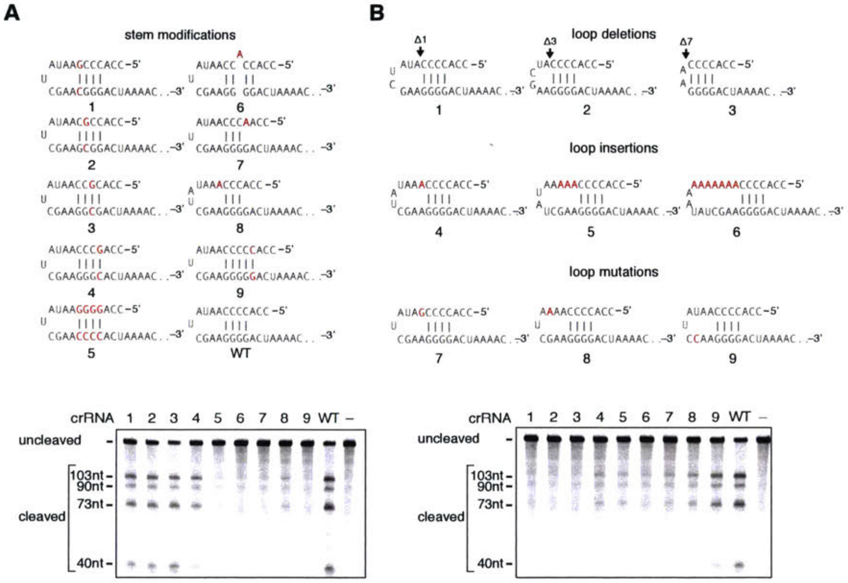


Figure 3.S17. RNA-guided RNase activity of LshC2c2 is dependent on direct repeat structure and sequence.

(A) Schematic showing modifications to the crRNA direct repeat stem (top). Altered bases are shown in red. Denaturing gel showing crRNA-guided cleavage of ssRNA 1 by each modified crRNA after 3 hours of incubation (bottom). Reported band lengths are matched from RNA sequencing. (B) Schematic showing modifications to the loop region of the crRNA direct repeat (top). Altered bases are shown in red and deletion lengths are indicated by arrows. Denaturing gel showing crRNA-guided cleavage of ssRNA 1 by each modified crRNA after 3 hours of incubation (bottom). Reported band lengths are matched from RNA sequencing.

10.2.3 Figure 3.S18

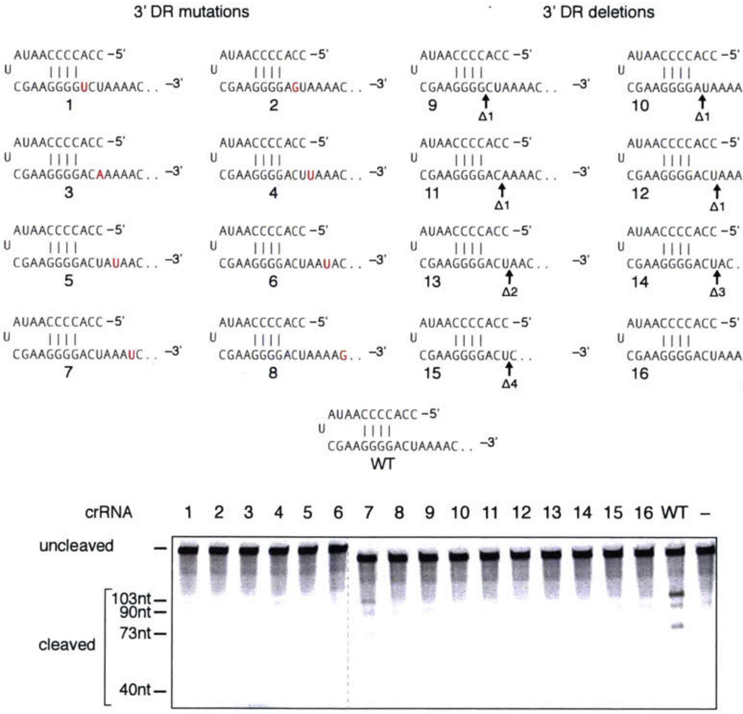


Figure 3.S18. The effect of 3' modifications to the crRNA DR.

Schematic shows the modifications made to the 3' end of the DR: single mutations *(top-left)* or deletions *(top-right)*. Altered bases are shown in red and deletion lengths are indicated by arrows. Denaturing gel depicting LshC2c2 cleavage activity by each modified crRNA after 3 hours of incubation *(bottom)*. Reported band lengths are matched from RNA sequencing.

10.2.19 Figure 3.S19

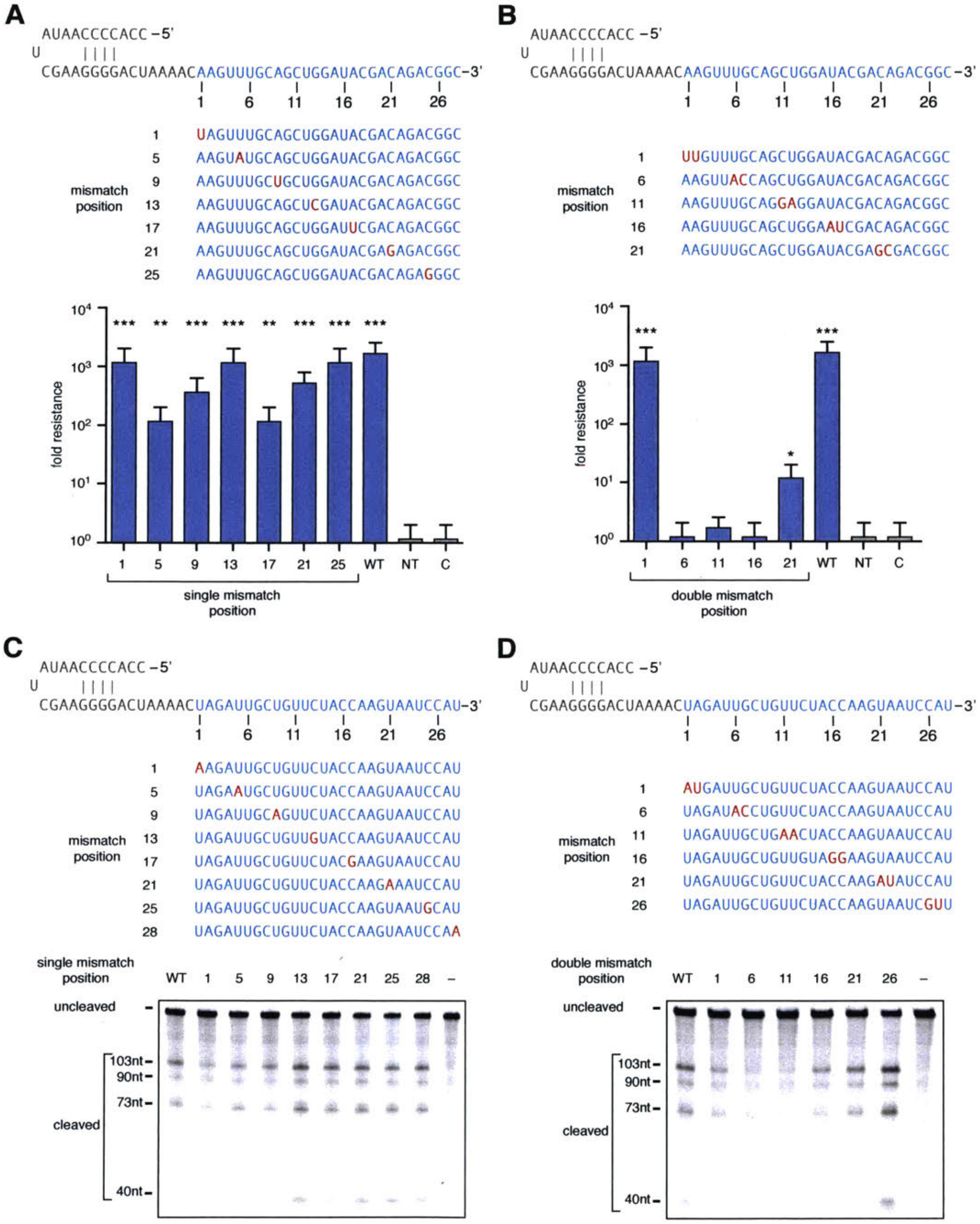


Figure 3.S19: Effect of RNA target-crRNA mismatches on LshC2c2 RNase activity.

(A) Quantification of MS2 plaque assays testing single mismatches at various positions in the spacer. Single mismatches have minimal effect on phage interference. Locations and identity of mismatches are shown in red. . (n=3 biological replicates. **, $p < 0.01$; ***, $p < 0.001$ compared to pACYC184 by t-test. Bars represent mean ± s.e.m). (B) Quantification of MS2 plaque assays testing double mismatches at various positions in the spacer. Consecutive double mismatches in the middle of the spacer eliminate phage interference. Locations and identity of mismatches are shown in red. (n=3 biological replicates. ***, $p < 0.001$ compared to pACYC184 by t-test. Bars represent mean ± s.e.m). (C) Schematic showing the position and identity of single mismatches (red) in the crRNA spacer (top). Denaturing gel showing cleavage of ssRNA 1 guided by crRNAs with single mismatches in the spacer after 3 hours of incubation (bottom). Reported band lengths are matched from RNA sequencing. (D) Schematic showing the position and identity of pairs of mismatches (red) in the crRNA spacer (top). Denaturing gel showing cleavage of ssRNA 1 guided by crRNAs with pairs of mismatches in the spacer after 3 hours of incubation (bottom). Reported band lengths are matched from RNA sequencing.

10.2.20 Figure 3.S20

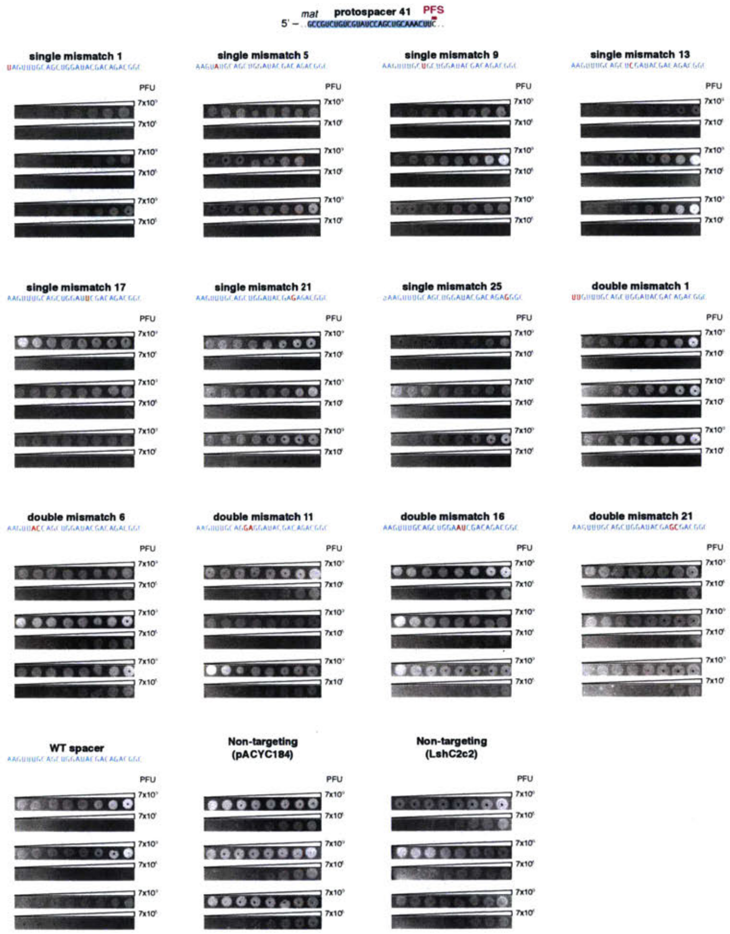

mat
protospacer 41
PFS
5' – ..GCCGUCUGUCGUAUCCAGCUGCAAACUUC..
single mismatch 1
UAGUUUGCAGCUGGAUACGACAGACGGC
single mismatch 5
AAGUAUGCAGCUGGAUACGACAGACGGC
single mismatch 9
AAGUUUGCUGCUGGAUACGACAGACGGC
single mismatch 13
AAGUUUGCAGCUCGAUACGACAGACGGC
single mismatch 17
AAGUUUGCAGCUGGAUUCGACAGACGGC
single mismatch 21
AAGUUUGCAGCUGGAUACGAGAGACGGC
single mismatch 25
AAGUUUGCAGCUGGAUACGACAGAGGGC
double mismatch 1
UUGUUUGCAGCUGGAUACGACAGACGGC
double mismatch 6
AAGUUACCAGCUGGAUACGACAGACGGC
double mismatch 11
AAGUUUGCAGGAGGAUACGACAGACGGC
double mismatch 16
AAGUUUGCAGCUGGAAUCGACAGACGGC
double mismatch 21
AAGUUUGCAGCUGGAUACGAGCGACGGC
WT spacer
AAGUUUGCAGCUGGAUACGACAGACGGC
Non-targeting (pACYC184)
Non-targeting (LshC2c2)
PFU
7x10

Figure 3.S20: MS2 restriction assay testing the effect of single and double mismatches on LshC2c2 activity.

pLshC2c2 with protospacer 41 was modified to have a series of single mismatches and consecutive double mismatches as shown. Images from plaque assay testing these mismatched spacers reveals reduced plaque formation for the single-mismatch spacers on-par with the fully complementary spacer. The double mismatch spacers show increased plaque formation for a seed region in the middle of spacer sequence. Phage dilutions were spotted on bacteria plates at decreasing numbers of plaque forming units (PFU). Spacer targets are shown above images; biological replicates are labeled BR1, BR2, or BR3. Non-targeting control is the native LshC2c2 locus.

10.2.21 Figure 3.S21

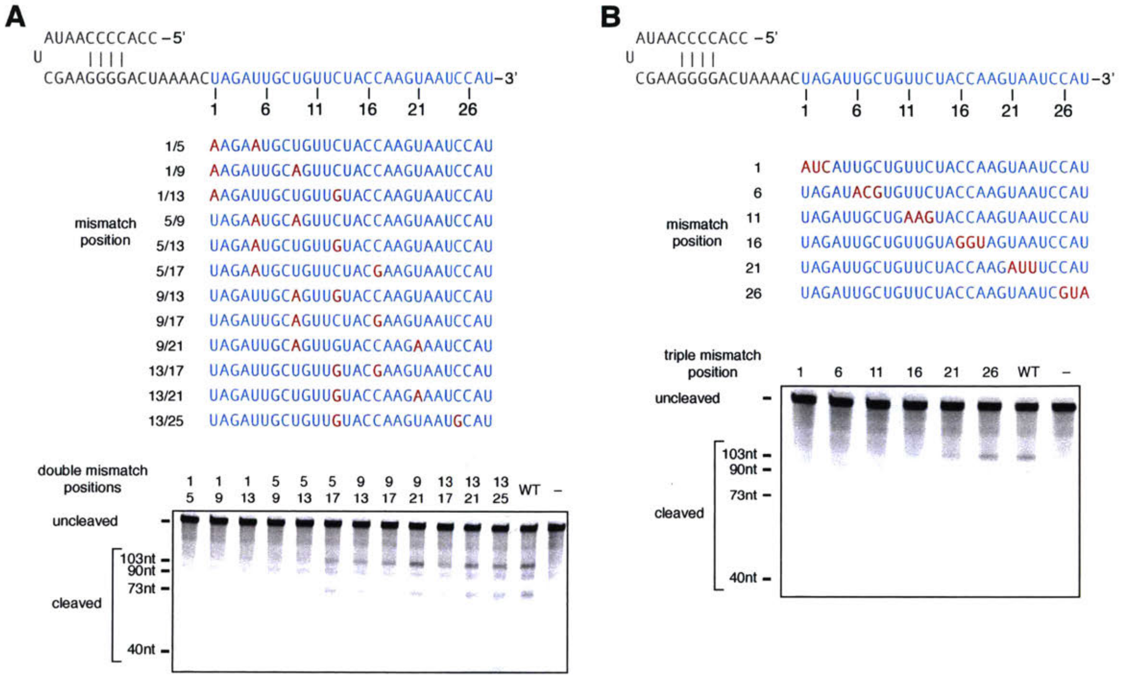

A
AUAACCCCACC-5'
U
CGAAGGGGACUAAAACUAGAUUGCUGUUCUACCAAGUAAUCCAU-3'
1 6 11 16 21 26
mismatch position
1/5 AAGAAUGCUGUUCUACCAAGUAAUCCAU
1/9 AAGAUUGCAGUUCUACCAAGUAAUCCAU
1/13 AAGAUUGCUGUUGUACCAAGUAAUCCAU
5/9 UAGAAUGCAGUUCUACCAAGUAAUCCAU
5/13 UAGAAUGCUGUUGUACCAAGUAAUCCAU
5/17 UAGAAUGCUGUUCUACGAAGUAAUCCAU
9/13 UAGAUUGCAGUUGUACCAAGUAAUCCAU
9/17 UAGAUUGCAGUUCUACGAAGUAAUCCAU
9/21 UAGAUUGCAGUUGUACCAAGAAAUCCAU
13/17 UAGAUUGCUGUUGUACGAAGUAAUCCAU
13/21 UAGAUUGCUGUUGUACCAAGAAAUCCAU
13/25 UAGAUUGCUGUUGUACCAAGUAAUGCAU
double mismatch positions
1 5, 1 9, 1 13, 5 9, 5 13, 5 17, 9 13, 9 17, 9 21, 13 17, 13 21, 13 25, WT, –
uncleaved
cleaved
103nt
90nt
73nt
40nt
B
AUAACCCCACC-5'
U
CGAAGGGGACUAAAACUAGAUUGCUGUUCUACCAAGUAAUCCAU-3'
1 6 11 16 21 26
mismatch position
1 AUCAUUGCUGUUCUACCAAGUAAUCCAU
6 UAGAUACGUGUUCUACCAAGUAAUCCAU
11 UAGAUUGCUGAAGUACCAAGUAAUCCAU
16 UAGAUUGCUGUUGUAGGUAGUAAUCCAU
21 UAGAUUGCUGUUCUACCAAGAUUUCCAU
26 UAGAUUGCUGUUCUACCAAGUAAUCGUA
triple mismatch position
1 6 11 16 21 26 WT –
uncleaved
cleaved
103nt
90nt
73nt
40nt

Figure 3.S21. The effect of triple mismatches on LshC2c2-crRNA cleavage activity.

(A) Schematic showing the position and identity of non-consecutive triple mismatches (red) in the crRNA spacer (top). Denaturing gel depicting LshC2c2 cleavage activity with crRNAs bearing triple non-consecutive mismatches between the spacer and ssRNA target region after 3 hours of incubation (bottom). Reported band lengths are matched from RNA sequencing. (B) Schematic showing the position and identity of consecutive triple mismatches (red) in the crRNA spacer (top). Denaturing gel depicting LshC2c2 cleavage activity with crRNAs bearing triple consecutive mismatches between the spacer and ssRNA target region after 3 hours of incubation (bottom). Reported band lengths are matched from RNA sequencing.

10.2.22 Figure 3.S22

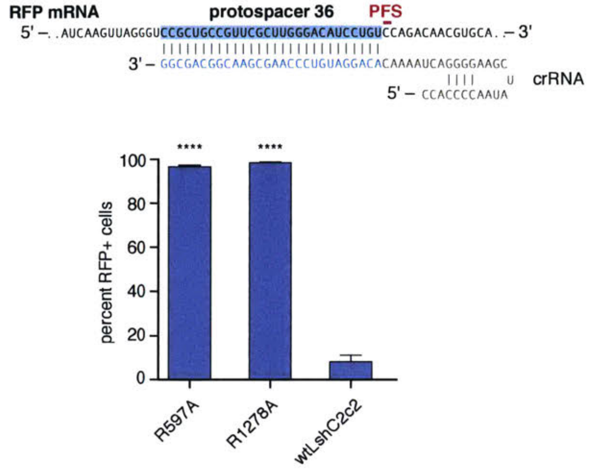


Figure 3.S22: HEPN mutant LshC2c2 are tested for RFP mRNA targeting activity.

The pLshC2c2 vector with protospacer 36 was modified to have the single HEPN mutants R597A and R1278A (one in each of the HEPN domains). These mutations resulted in little detectable RFP knockdown as measured by flow cytometry on the *E. coli*. (n=3 biological replicates. ***, p < 0.001 compared to wtLshC2c2 by t-test. Bars represent mean ± s.e.m).

10.2.23 Figure 3.S23

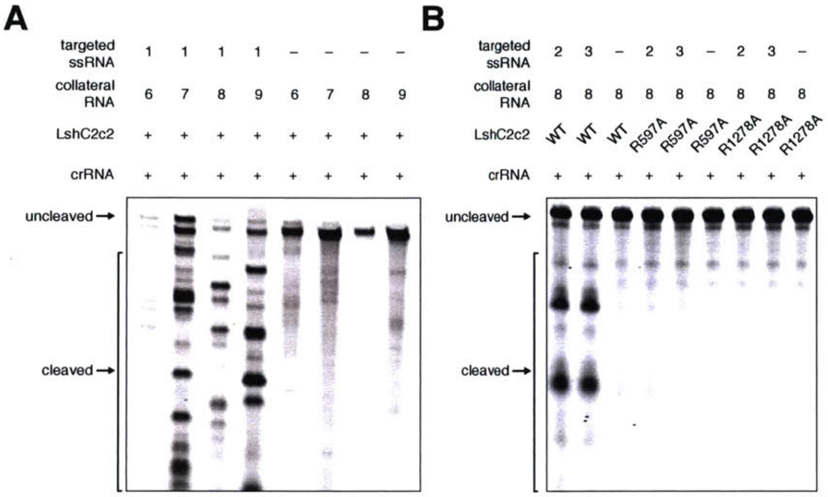


Figure 3.S23: Biochemical characterization of the collateral cleavage effect.

(A) LshCc2 is incubated for 3 hours with a crRNA targeting protospacer 14 with and without unlabeled ssRNA target 1 (contains protospacer 14). When LshC2c2 is in the presence of target 1, significant cleavage activity is observed for 5' fluorescently labeled non-complementary targets 6-9. (B) HEPN mutant collateral activity is compared to WT C2c2. The proteins are incubated for 3 hours with crRNA complementary to protospacer 14 and with and without unlabeled homopolymer targets 2 or 3 (both containing protospacer 14). The collateral effect is no longer observed with the HEPN mutant proteins on the 3' fluorescently labeled non-complementary target 8.

10.2.24 Figure 3.S24

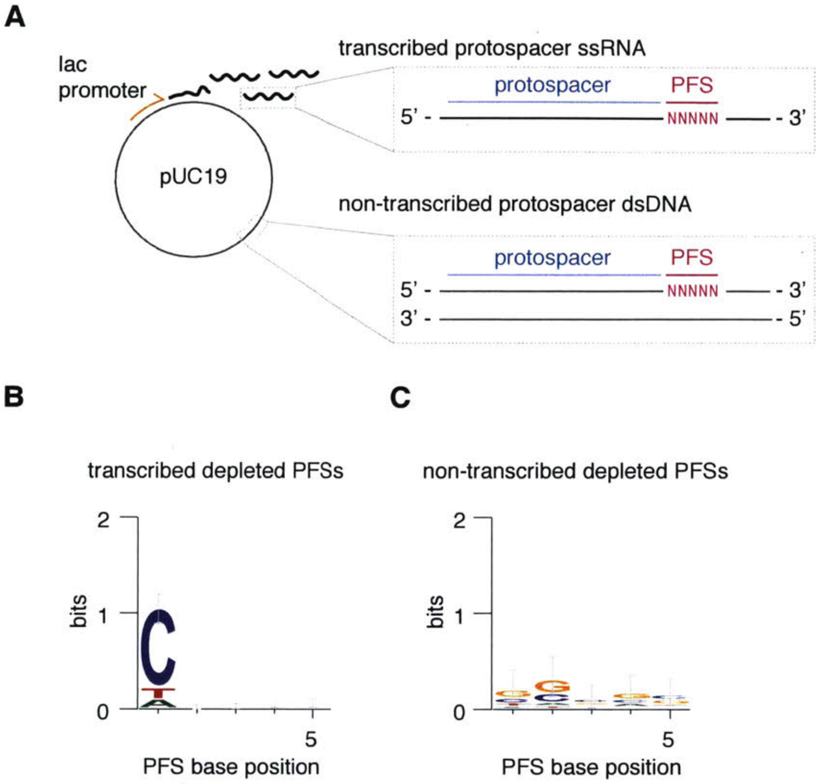


Figure 3.S24. *In vivo* collateral effect reveals a 3' H PFS with a PFS screen in a transcribed region.

(A) Schematic for a transcribed and non-transcribed PFS screen. (B) 3' PFS motif for a PFS screen designed in a transcribed plasmid region. (C) No PFS is observed for a PFS screen in a non-transcribed region.

10.3 Chapter 4 Supplementary Figures

10.3.1 Figure 4.S1

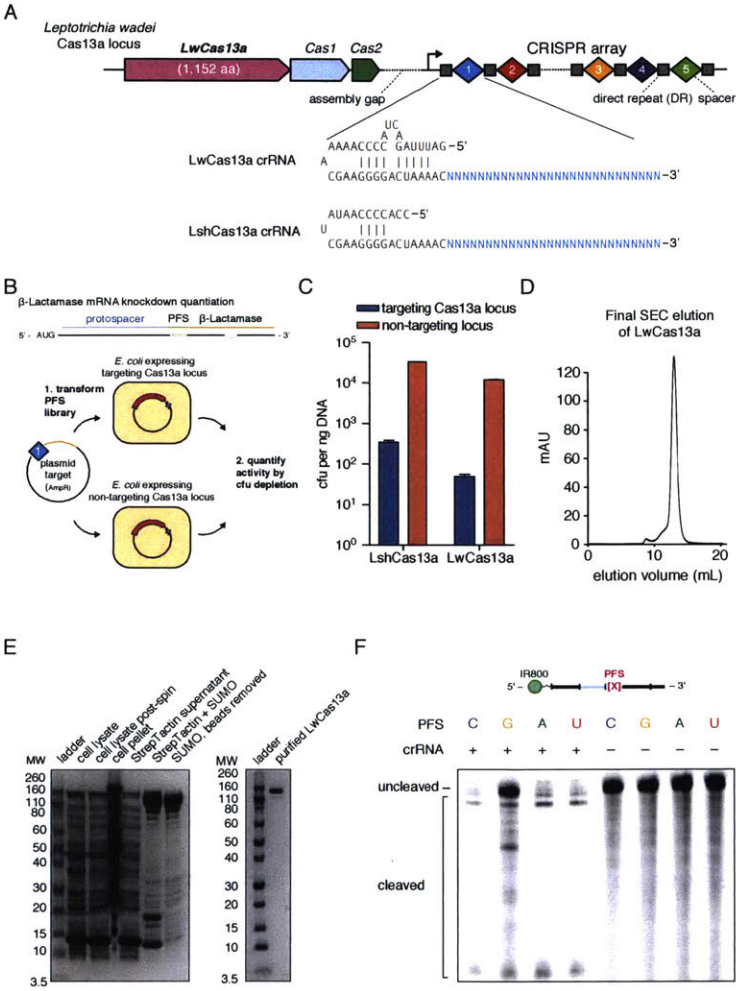


Figure 4.S1. LwCas13a is capable of RNA-guided RNA interference and cleavage.

(A) Schematic of the CRISPR/Cas13a locus from *Leptotrichia wadei*. Representative crRNA structures from LwCas13a and LshCas13a systems are shown.

(B) Schematic of *in vivo* bacterial assay for Cas13a activity. A protospacer is cloned upstream of the beta-lactamase gene in an ampicillin-resistance plasmid, and this construct is transformed into *E. coli* expressing Cas13a in conjunction with either a targeting or non-targeting spacer. Successful transformants are counted to quantify activity.

(C) Quantitation of LwCas13a and LshCas13a *in vivo* activity. (n=2 biological replicates; bars represent mean ± s.e.m.)

(D) Final size exclusion gel filtration of recombinant LwCas13a protein.

(E) Coomassie blue stained acrylamide gel of LwCas13a stepwise purification.

(F) Activity of LwCas13a against different PFS targets. LwCas13a was targeted against fluorescent RNA with variable 3' PFS flanking the spacer, and reaction products were visualized on denaturing gel. LwCas13a shows a slight preference against a G PFS.

10.3.2 Figure 4.S2

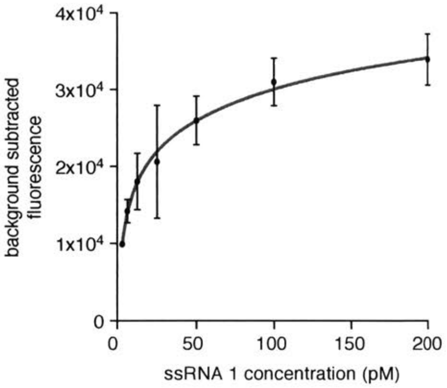


Figure 4.S2. Detection with LwCas13a is quantitative.

(A) Fluorescence measurements from Cas13a detection without amplification are correlated with input RNA concentration. (n=2 biological replicates; bars represent mean ± s.e.m.)

10.3.3 Figure 4.S3

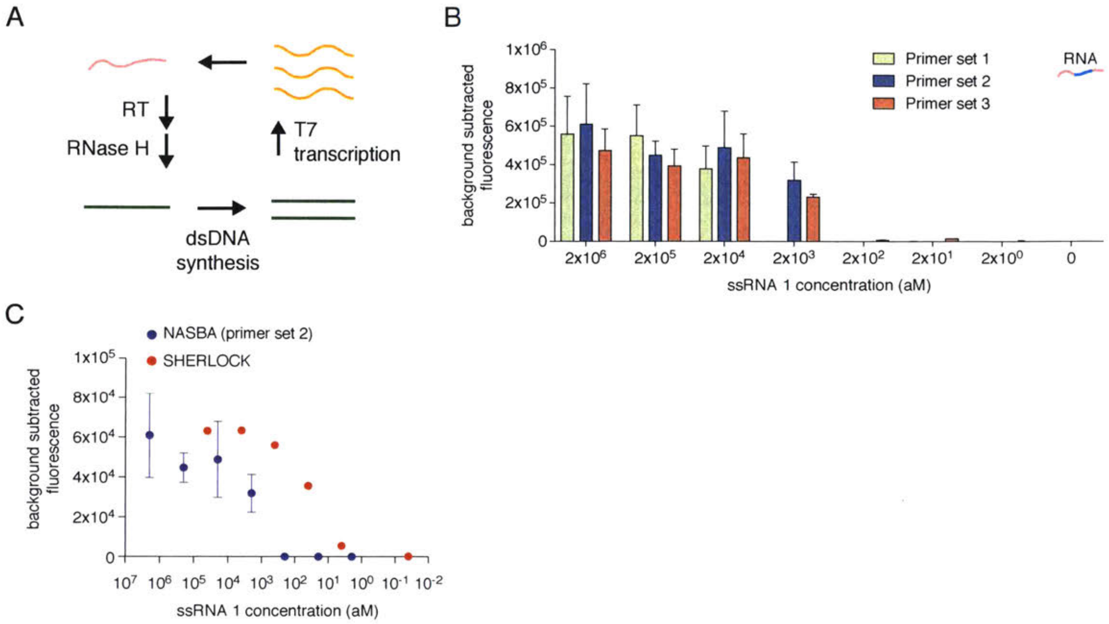


Figure 4.S3. Nucleic acid amplification with NASBA followed by Cas13a detection.

A. Schematic of the NASBA reaction.

B. Detection of nucleic acid target ssRNA 1 amplified by NASBA with three different primer sets and then subjected to Cas13a collateral detection using a quenched fluorescent probe.

C. Comparison of detection of ssRNA 1 by NASBA with primer set 2 and SHERLOCK. (n=2 technical replicates; bars represent mean ± s.e.m.)

10.3.4 Figure 4.S4

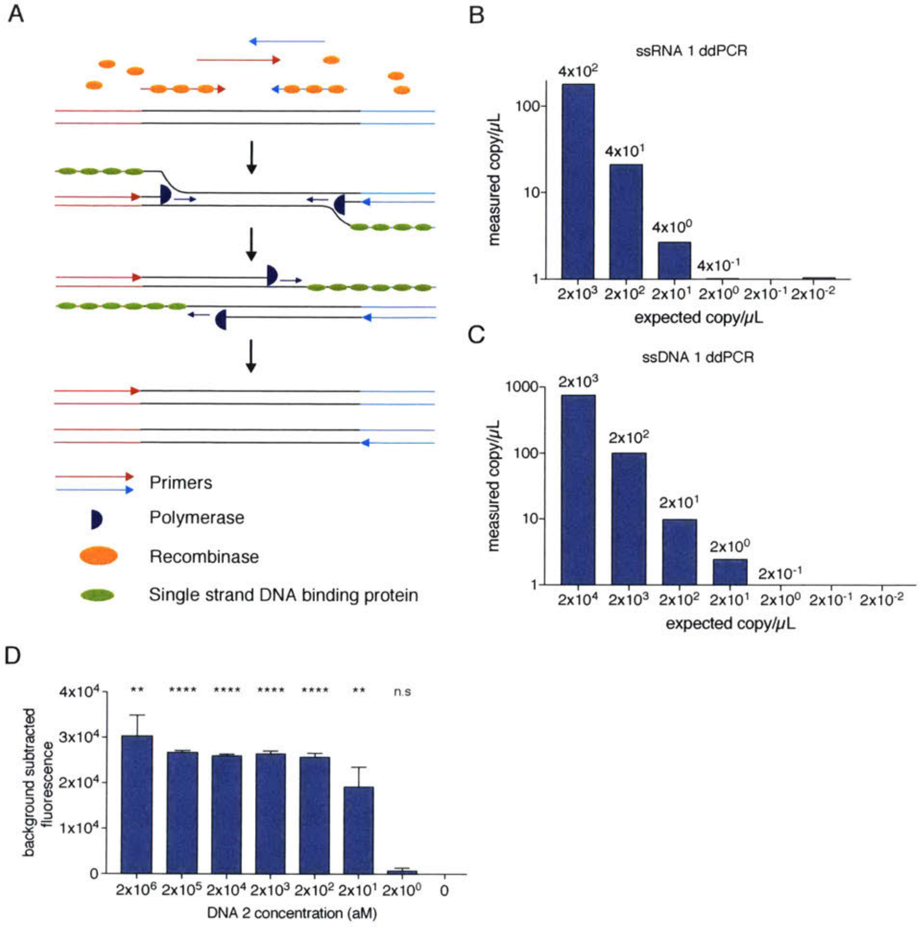


Figure 4.S4. Nucleic acid amplification with RPA and single-reaction SHERLOCK.

(A) Schematic of the RPA reaction, showing the participating components in the reaction.

(B) Digital-droplet PCR quantitation of ssRNA 1 for dilutions used in Fig. 1C. Adjusted concentrations for the dilutions based on the ddPCR results are shown above bar graphs.

(C) Digital-droplet PCR quantitation of ssDNA 1 for dilutions used in Fig. 1D. Adjusted concentrations for the dilutions based on the ddPCR results are shown above bar graphs.

(D) The RPA, T7 transcription, and Cas13a detection reactions are compatible and achieve single molecule detection of DNA 2 when incubated simultaneously. (n=3 technical replicates, two-tailed Student t-test; n.s., not significant; **, $p < 0.01$; ****, $p < 0.0001$; bars represent mean ± s.e.m.)

10.3.5 Figure 4.S5

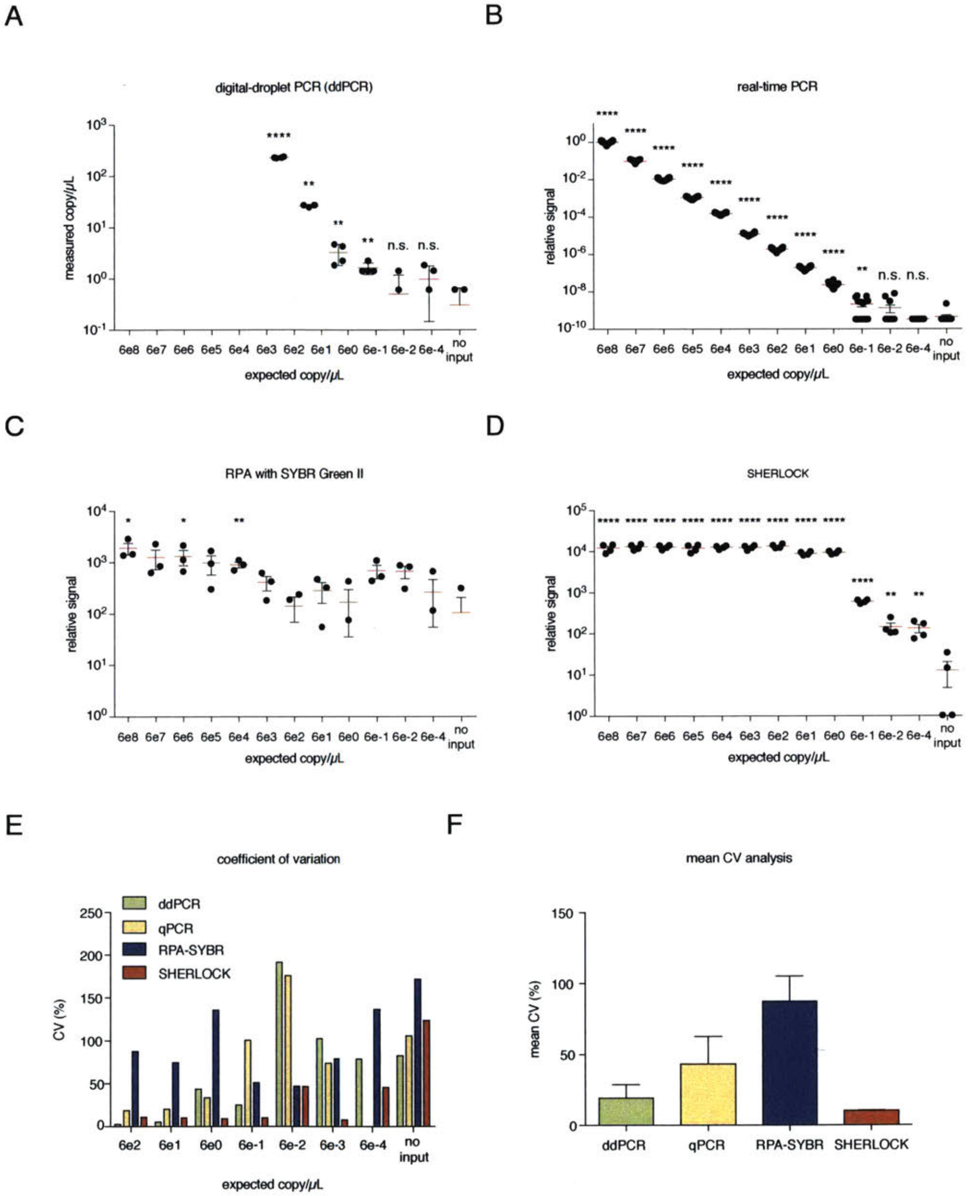


Figure 4.S5. Comparison of SHERLOCK to other sensitive nucleic acid detection tools.

(A) Detection analysis of ssDNA 1 dilution series with digital-droplet PCR. (n=4 technical replicates, two-tailed Student t-test; n.s., not significant; *, $p < 0.05$; **, $p < 0.01$; ****, $p < 0.0001$; red lines represent mean, bars represent mean ± s.e.m. Samples with measured copy/μL below 10^{-1} not shown.)

(B) Detection analysis of ssDNA 1 dilution series with quantitative PCR. (n=16 technical replicates, two-tailed Student t-test; n.s., not significant; **, $p < 0.01$; ****, $p < 0.0001$; red lines represent mean, bars represent mean ± s.e.m. Samples with relative signal below 10^{-10} not shown.)

(C) Detection analysis of ssDNA 1 dilution series with RPA with SYBR Green II. (n=4 technical replicates, two-tailed Student t-test; *, $p < 0.05$; **, $p < 0.01$; red lines represent mean, bars represent mean ± s.e.m. Samples with relative signal below 10^{0} not shown.)

(D) Detection analysis of ssDNA 1 dilution series with SHERLOCK. (n=4 technical replicates, two-tailed Student t-test; **, $p < 0.01$; ****, $p < 0.0001$; red lines represent mean, bars represent mean ± s.e.m. Samples with relative signal below 10^{0} not shown.)

(E) Percent coefficient of variation for a series of ssDNA 1 dilutions for four types of detection methods.

(F) Mean percent coefficient of variation for the 6e2, 6e1, 6e0, and 6e-1 ssDNA 1 dilutions for four types of detection methods. (bars represent mean ± s.e.m.)

10.3.6 Figure 4.S6

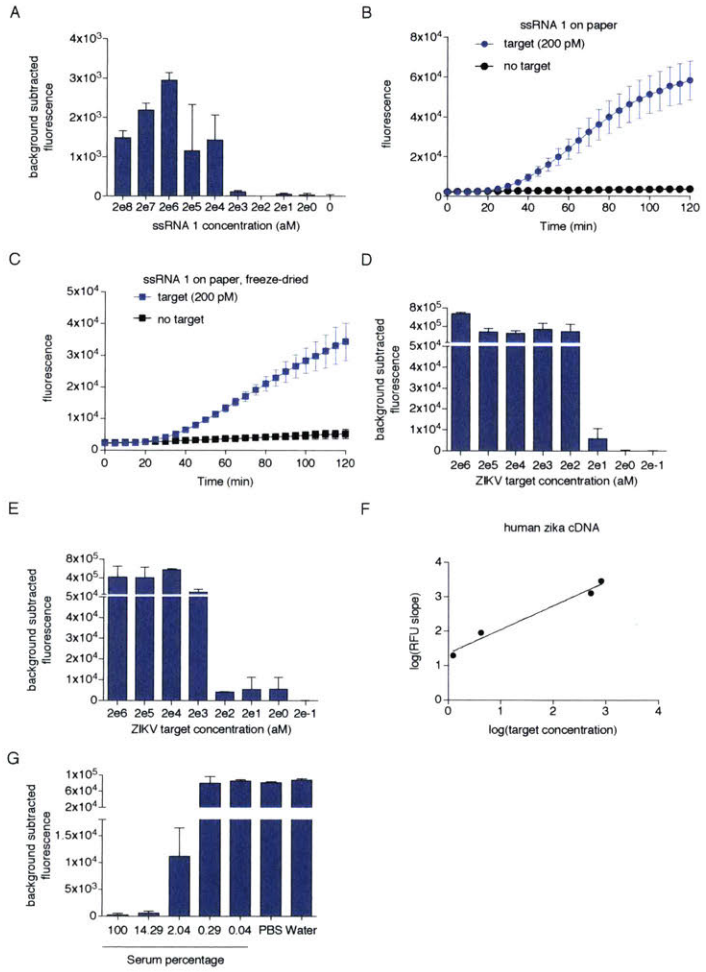


Figure 4.S6. Development of SHERLOCK as a point-of-care diagnostic.

(A) Freeze-dried Cas13a is capable of sensitive detection of ssRNA 1 in the low femtomolar range. (n=2 technical replicates; bars represent mean ± s.e.m.)

(B, C) Cas13a is capable of rapid detection of a 200 pM ssRNA 1 target on paper as spotted as liquid (B) or freeze-dried form (C). (n=3 technical replicates; bars represent mean ± s.d.)

(D, E) The SHERLOCK reaction is capable of sensitive detection of synthesized ZIKV RNA fragments in solution (D) and in freeze-dried form (E) (n=3 technical replicates; bars represent mean ± s.e.m.)

(F) Quantitative curve for human ZIKV cDNA detection with SHERLOCK showing significant correlation between input concentration and detected fluorescence.

(G) Cas13a detection of ssRNA 1 performed in the presence of varying amounts of human serum. (n=2 technical replicates; bars represent mean ± s.e.m.)

10.3.7 Figure 4.S7

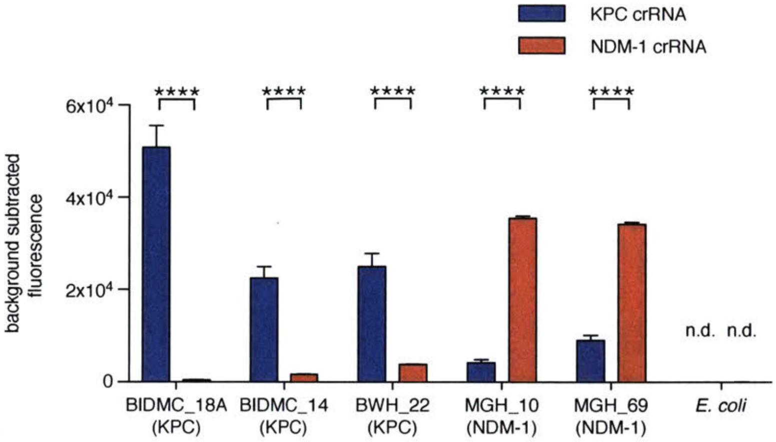


Figure 4.S7. Detection of carbapanem resistance in clinical bacterial isolates.

Detection of two different carbapenem-resistance genes (KPC and NDM-1) from five clinical isolates of *Klebsiella pneumoniae* and an *E. coli* control. (n=4 technical replicates, two-tailed Student t-test; ****, p < 0.0001; bars represent mean ± s.e.m.; n.d., not detected)

10.3.8 Figure 4.S8

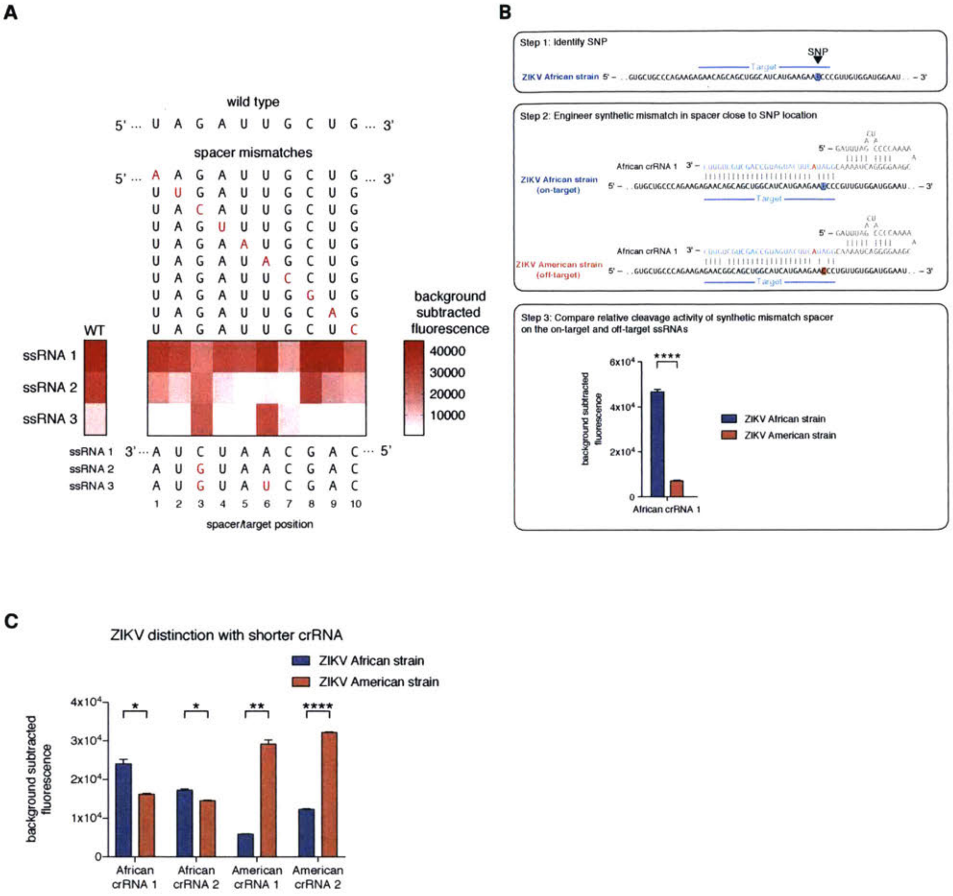


Figure 4.S8. Engineering Cas13a to have single-base specificity.

(A) Cas13a is not sensitive to single mismatches, but can distinguish between single nucleotide differences in target when loaded with crRNAs with additional mismatches. ssRNA 1-3 were detected with 11 crRNAs, with 10 spacers containing synthetic mismatches at various positions in the crRNA. Mismatched spacers did not show reduced collateral cleavage of ssRNA 1, but showed inhibited collateral cleavage of mismatched targets ssRNA 2 and ssRNA 3.

(B) Schematic of the process for rational design of single-base specific spacers with synthetic mismatches. Synthetic mismatches are placed in proximity to the SNP or base of interest.

(C) Highly specific detection of strain SNPs allows for the differentiation of ZIKV African versus American RNA targets differing by only one nucleotide using Cas13a detection with truncated (23 nt) crRNAs. ($n=2$ technical replicates, one-tailed Student t-test; *, $p < 0.05$; **, $p < 0.01$; ****, $p < 0.0001$; bars represent mean ± s.e.m.)

10.3.9 Figure 4.S9

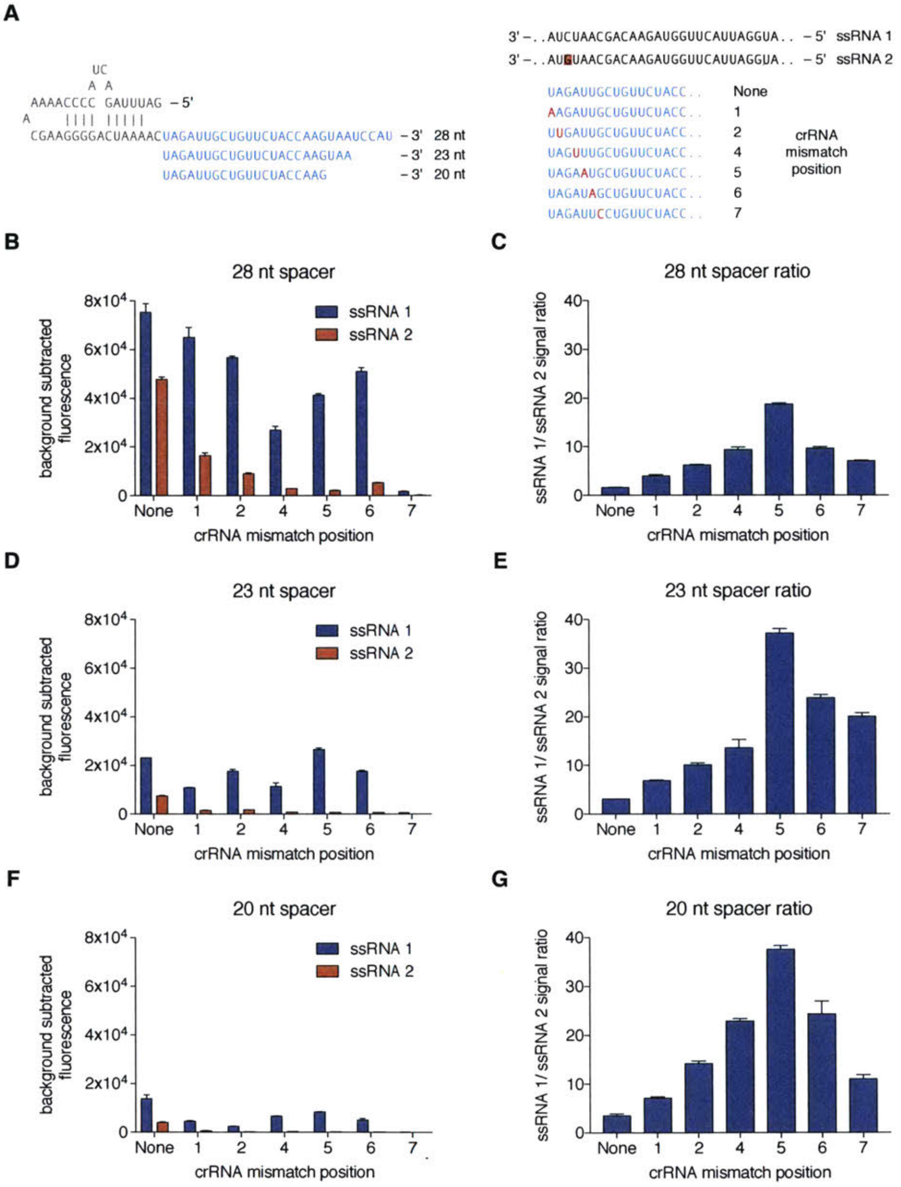


Figure 4.S9. Characterization of LwCas13a sensitivity to truncated spacers and single mismatches in the target sequence.

(A) Sequences of truncated spacer crRNAs used in (B)-(G). Also shown are sequences of ssRNA 1 and 2, which has a single base-pair difference highlighted in red. crRNAs containing synthetic mismatches are displayed with mismatch positions colored in red.

(B) Collateral cleavage activity on ssRNA 1 and 2 for 28 nt spacer crRNA with synthetic mismatches at positions 1-7. (n=4 technical replicates; bars represent mean ± s.e.m.)

(C) Specificity ratios of crRNA tested in (B). Specificity ratios are calculated as the ratio of the on-target RNA (ssRNA 1) collateral cleavage to the off-target RNA (ssRNA 2) collateral cleavage. (n=4 technical replicates; bars represent mean ± s.e.m.)

(D) Collateral cleavage activity on ssRNA 1 and 2 for 23 nt spacer crRNA with synthetic mismatches at positions 1-7. (n=4 technical replicates; bars represent mean ± s.e.m.)

(E) Specificity ratios of crRNA tested in (D). Specificity ratios are calculated as the ratio of the on-target RNA (ssRNA 1) collateral cleavage to the off-target RNA (ssRNA 2) collateral cleavage. (n=4 technical replicates; bars represent mean ± s.e.m.)

(F) Collateral cleavage activity on ssRNA 1 and 2 for 20 nt spacer crRNA with synthetic mismatches at positions 1-7. (n=4 technical replicates; bars represent mean ± s.e.m.)

(G) Specificity ratios of crRNA tested in (F). Specificity ratios are calculated as the ratio of the on-target RNA (ssRNA 1) collateral cleavage to the off-target RNA (ssRNA 2) collateral cleavage. (n=4 technical replicates; bars represent mean ± s.e.m.)

10.3.10 Figure 4.S10

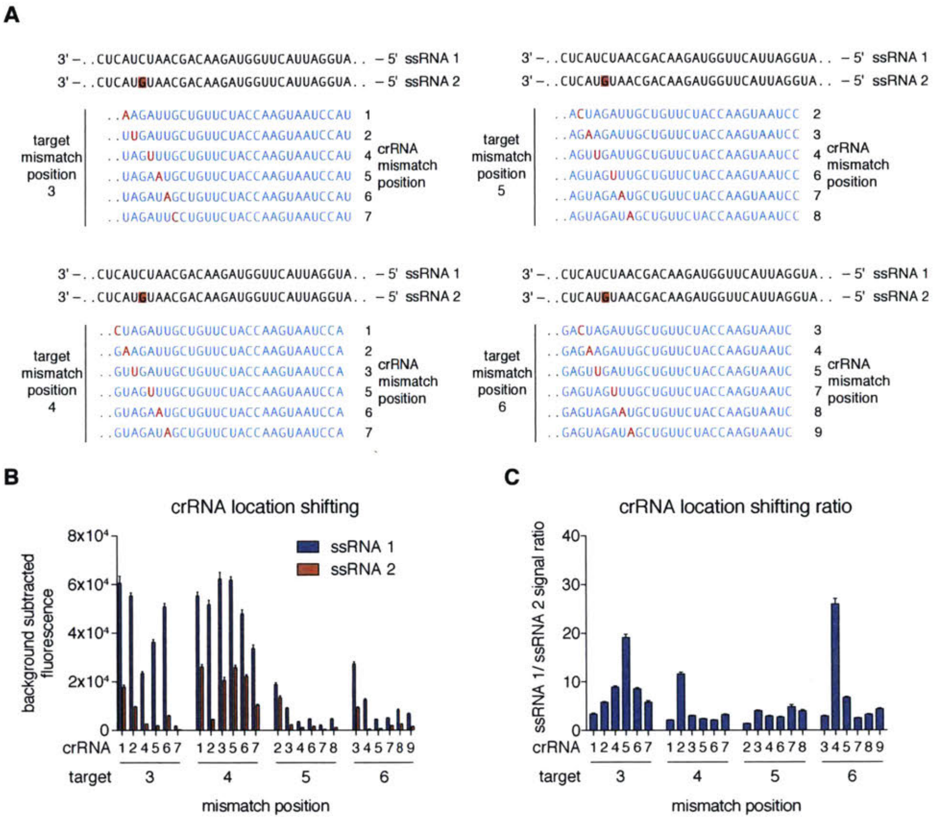


Figure 4.S10. Identification of ideal synthetic mismatch position relative to mutations in the target sequence.

(A) Sequences for evaluation of the ideal synthetic mismatch position to detect a mutation between ssRNA 1 and ssRNA 2. On each of the targets, crRNAs with synthetic mismatches at the colored (red) locations are tested. Each set of synthetic mismatch crRNAs is designed such that the mutation location is shifted in position relative to the sequence of the spacer. Spacers are designed such that the mutation is evaluated at positions 3, 4, 5, and 6 within the spacer.

(B) Collateral cleavage activity on ssRNA 1 and 2 for crRNAs with synthetic mismatches at varying positions. There are four sets of crRNAs with the mutation at either position 3, 4, 5, or 6 within the spacer:target duplex region. (n=4 technical replicates; bars represent mean ± s.e.m.)

(C) Specificity ratios of crRNA tested in (B). Specificity ratios are calculated as the ratio of the on-target RNA (ssRNA 1) collateral cleavage to the off-target RNA (ssRNA 2) collateral cleavage. (n=4 technical replicates; bars represent mean ± s.e.m.)

10.3.11 Figure 4.S11

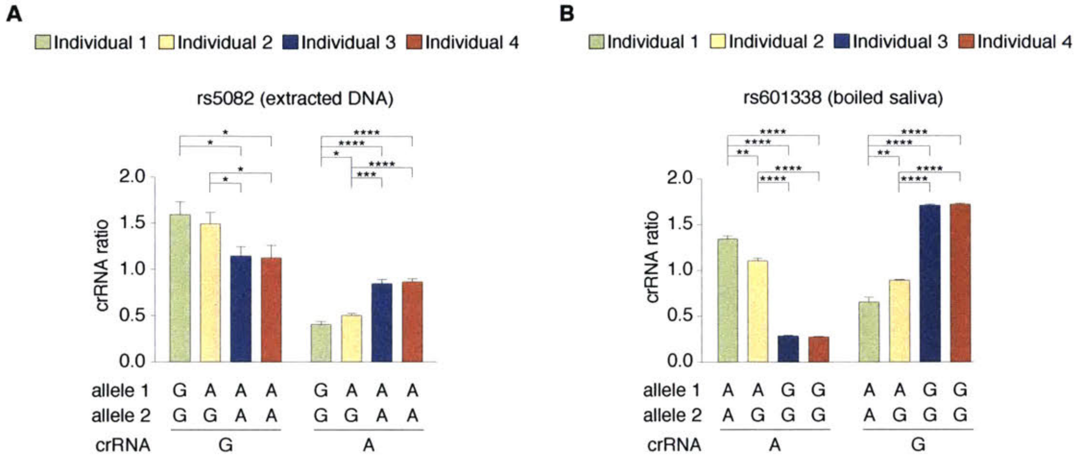
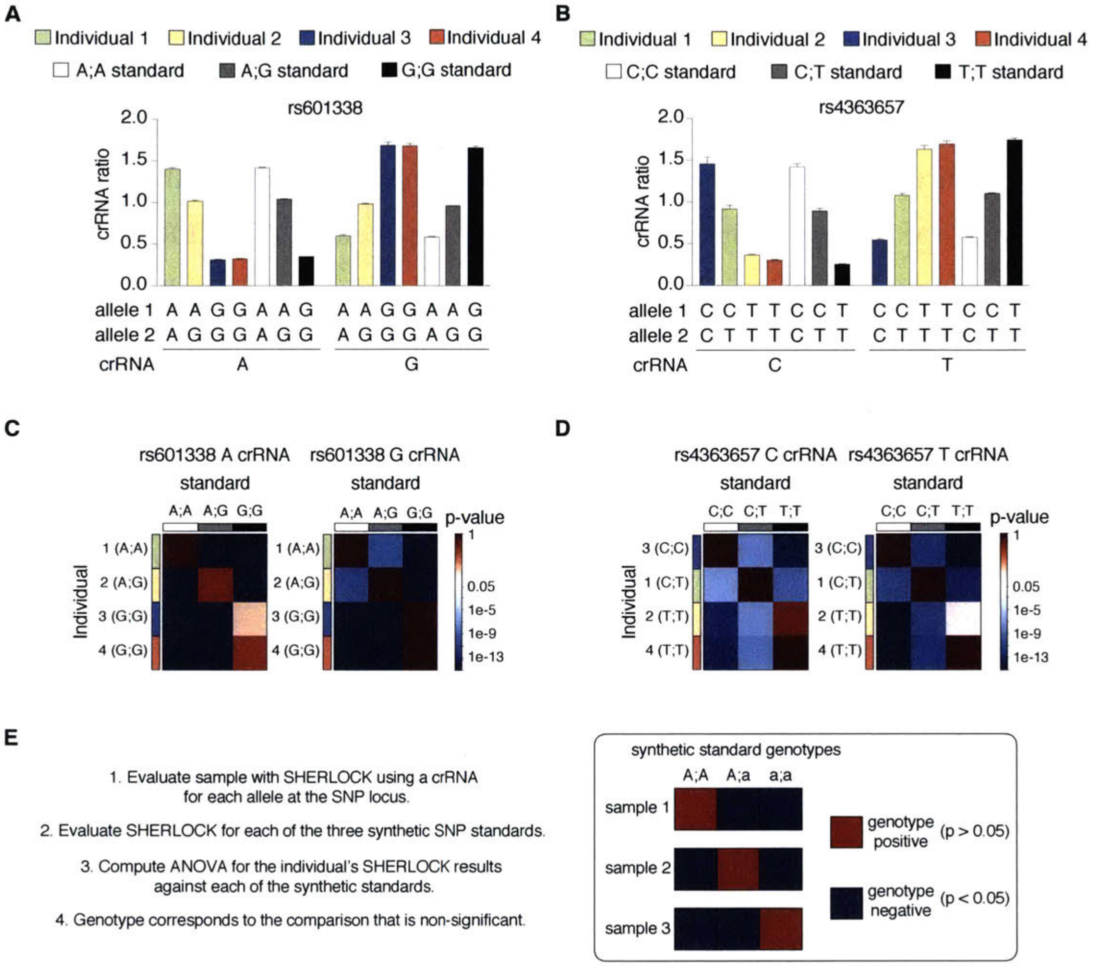


Figure 4.S11. Genotyping with SHERLOCK at an additional locus and direct genotyping from boiled saliva.

(A) SHERLOCK can distinguish between genotypes at the rs5082 SNP site. (n=4 technical replicates, one-tailed Student t-test; *, $p < 0.05$; ***, $p < 0.001$; ****, $p < 0.0001$; bars represent mean ± s.e.m.)

(B) SHERLOCK can distinguish between genotypes at the rs601338 SNP site in genomic DNA directly from centrifuged, denatured, and boiled saliva. (n=4 technical replicates, two-tailed Student t-test; **, $p < 0.01$; ****, $p < 0.001$; bars represent mean ± s.e.m.)

10.3.12 Figure 4.S12



1. Evaluate sample with SHERLOCK using a crRNA for each allele at the SNP locus.

2. Evaluate SHERLOCK for each of the three synthetic SNP standards.

3. Compute ANOVA for the individual's SHERLOCK results against each of the synthetic standards.

4. Genotype corresponds to the comparison that is non-significant.

Figure 4.S12. Development of synthetic genotyping standards to accurately genotype human SNPs.

(A) Genotyping with SHERLOCK at the rs601338 SNP site for each of the four individuals compared against PCR-amplified genotype standards. (n=4 technical replicates; bars represent mean ± s.e.m.)

(B) Genotyping with SHERLOCK at the rs4363657 SNP site for each of the four individuals compared against PCR-amplified genotype standards. (n=4 technical replicates; bars represent mean ± s.e.m.)

(C) Heatmaps of computed p-values between the SHERLOCK results for each individual and the synthetic standards at the rs601338 SNP site. A heatmap is shown for each of the allele-sensing crRNAs. The heatmap color map is scaled such that insignificance ($p > 0.05$) is red and significance ($p < 0.05$) is blue. (n=4 technical replicates, one-way ANOVA)

(D) Heatmaps of computed p-values between the SHERLOCK results for each individual and the synthetic standards at the rs4363657 SNP site. A heatmap is shown for each of the allele-sensing crRNAs. The heatmap color map is scaled such that insignificance ($p > 0.05$) is red and significance ($p < 0.05$) is blue. (n=4 technical replicates, one-way ANOVA)

(E) A guide for understanding the p-value heatmap results of SHERLOCK genotyping. Genotyping can easily be called by choosing the allele that corresponds to a p-value > 0.05 between the individual and allelic synthetic standards. Red blocks correspond to non-significant differences between the synthetic standard and individual's SHERLOCK result and thus a genotype-positive result. Blue blocks correspond to significant differences between the synthetic standard and individual's SHERLOCK result and thus a genotype-negative result.

10.3.13 Figure 4.S13

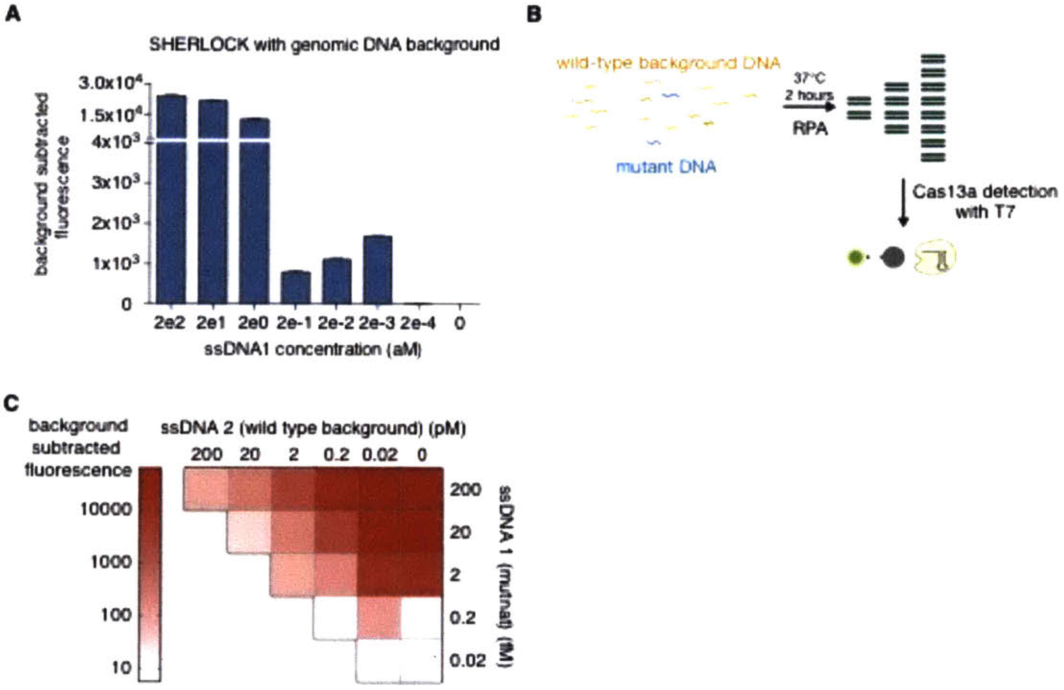


Figure 4.S13. Detection of ssDNA 1 as a small fraction of mismatched background target.

(A) SHERLOCK detection of a dilution series of ssDNA 1 on a background of human genomic DNA. Note that there should be no sequence similarity between the ssDNA 1 target being detected and the background genomic DNA. (n=2 technical replicates; bars represent mean ± s.e.m.)

(B) Schematic of SHERLOCK detection of ssDNA 1 (mutant DNA) on a background of ssDNA 2, which differs from ssDNA 1 by only a single mismatch (wild-type background DNA).

(C) SHERLOCK achieves single nucleotide specificity detection of ssDNA 1 in the presence of ssDNA 2, which differs by only a single mismatch. Various concentrations of ssDNA 1 were combined with a background excess of ssDNA 2 and detected by SHERLOCK.

10.4 Chapter 5 Supplementary Figures

10.4.1 Figure 5.S1

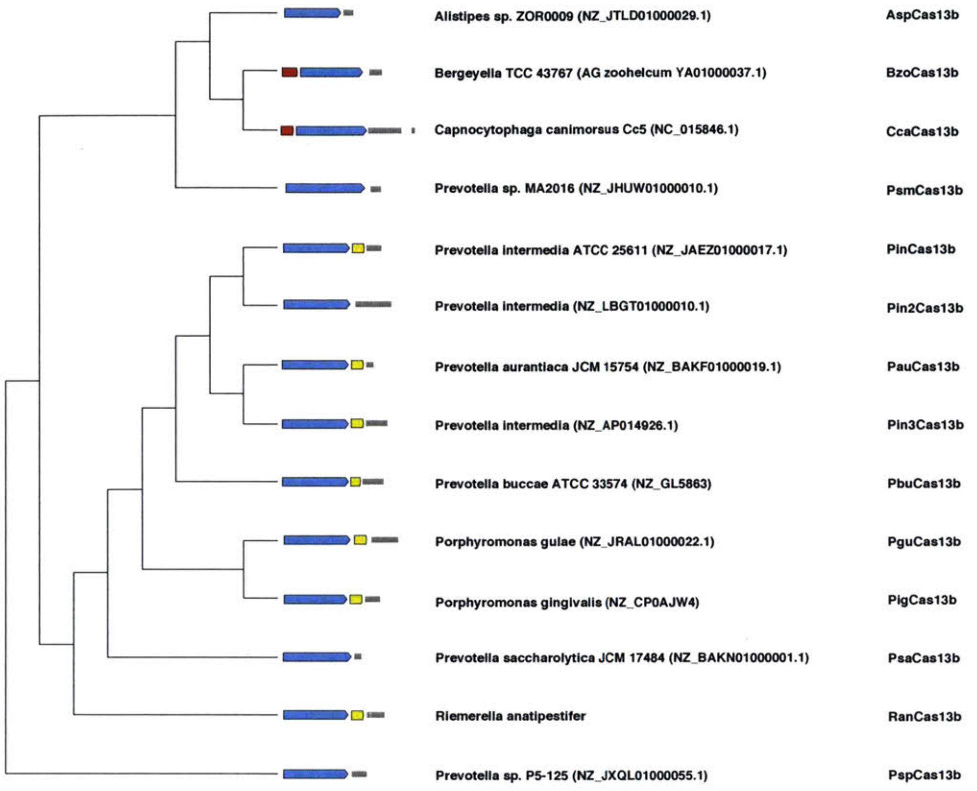

Alistipes sp. ZOR0009 (NZ_JTLD01000029.1)
AspCas13b
Bergeyella TCC 43767 (AG zoohelcum YA01000037.1)
BzoCas13b
Capnocytophaga canimorsus Cc5 (NC_015846.1)
CcaCas13b
Prevotella sp. MA2016 (NZ_JHUW01000010.1)
PsmCas13b
Prevotella intermedia ATCC 25611 (NZ_JAEZ01000017.1)
PinCas13b
Prevotella intermedia (NZ_LBGT01000010.1)
Pin2Cas13b
Prevotella aurantiaca JCM 15754 (NZ_BAKF01000019.1)
PauCas13b
Prevotella intermedia (NZ_AP014926.1)
Pin3Cas13b
Prevotella buccae ATCC 33574 (NZ_GL5863)
PbuCas13b
Porphyromonas gulae (NZ_JRAL01000022.1)
PguCas13b
Porphyromonas gingivalis (NZ_CP0AJW4)
PigCas13b
Prevotella saccharolytica JCM 17484 (NZ_BAKN01000001.1)
PsaCas13b
Riemerella anatipestifer
RanCas13b
Prevotella sp. P5-125 (NZ_JXQL01000055.1)
PspCas13b

Figure 5.S1: Cas13b orthologs evaluated for *in vitro* collateral activity.

Tree of 15 Cas13b orthologs purified and evaluated for in vitro collateral activity. Cas13b gene (blue), Csx27 gene (red), Csx28 gene (yellow), and CRISPR array (grey) are shown.

10.4.2 Figure 5.S2

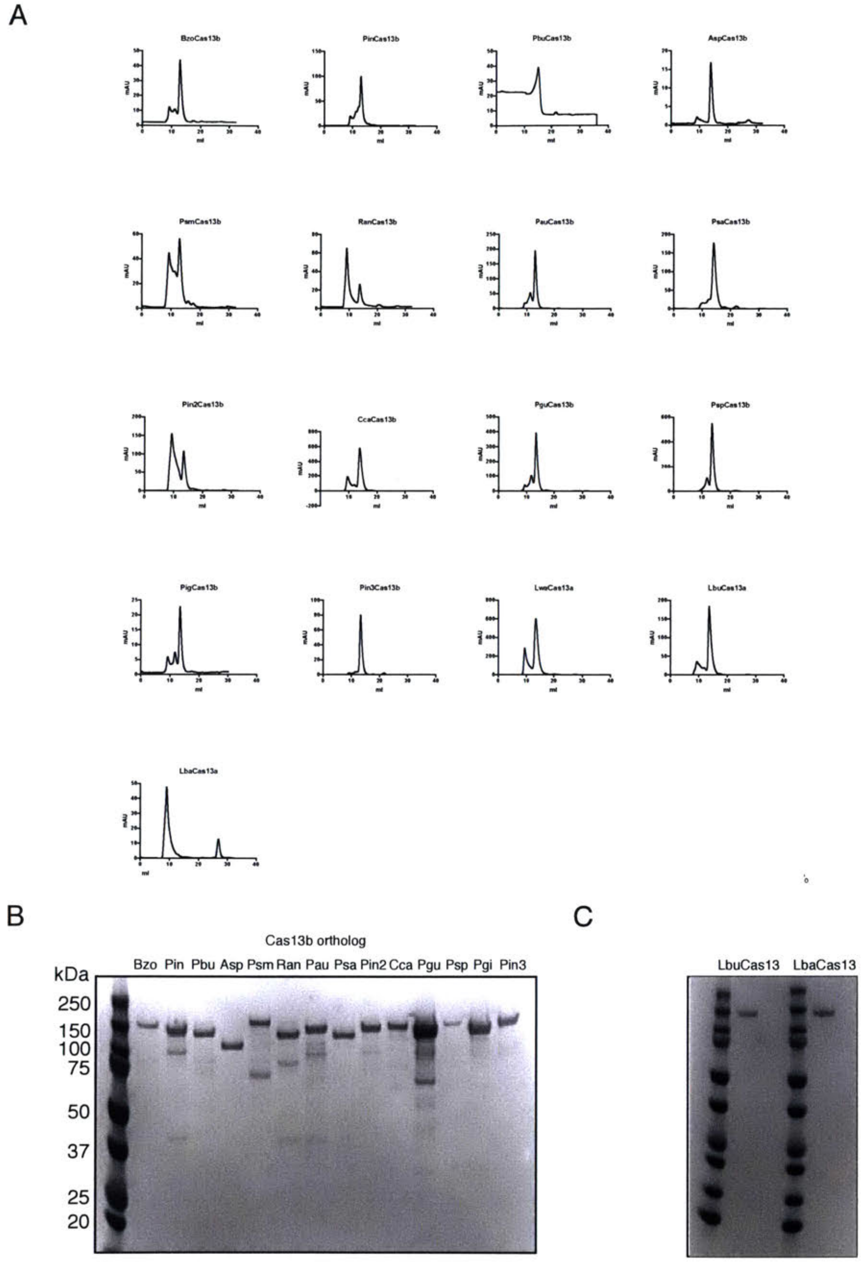


Figure 5.S2: Protein purification of Cas13 orthologs

A) Chromatograms of size exclusion chromatography for Cas13b, LwCas13a and LbaCas13a. Measured UV absorbance (mAU) is shown against the elution volume (ml)

B) SDS-PAGE gel of purified Cas13b orthologs. Fourteen Cas13b orthologs are loaded from left to right. A protein ladder is shown to the left.

Final SDS-PAGE gel of LbaCas13a and LbuCas13a. Two dilutions of LbaCas13a and LbuCas13a are shown.

10.4.3 Figure 5.S3

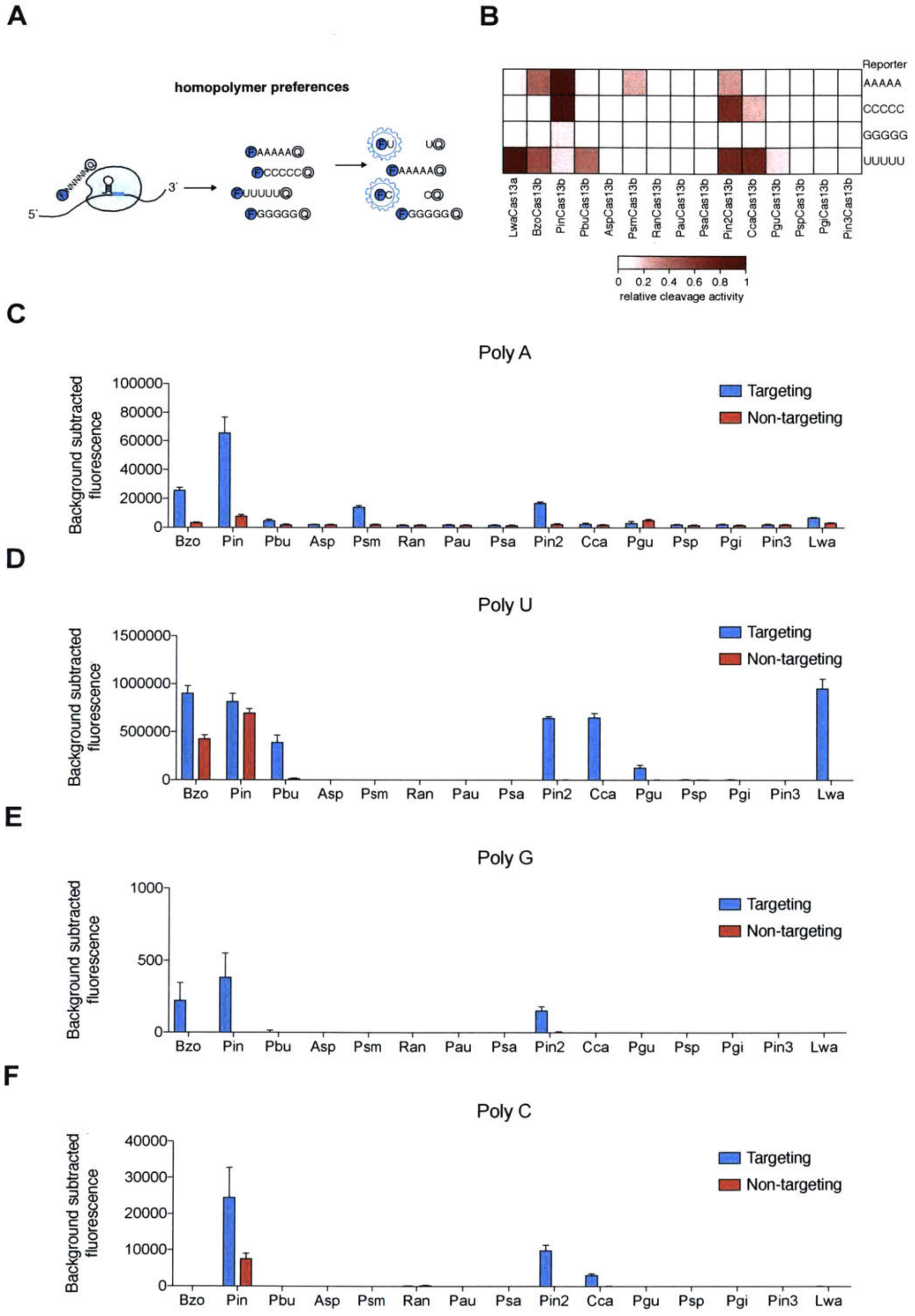


Figure S3: Base preference of Cas13b ortholog collateral cleavage.

A) Schematic of assay for determining hompolymer preferences of Cas13a/b enzymes.

B) Heatmap of the base preference of 15 Cas13b orthologs targeting ssRNA 1 with reporters consisting of a homopolymer of A, C, G, or U bases.

C) Cleavage activity of fourteen Cas13b orthologs targeting ssRNA 1 using a homopolymer adenine sensor five nucleotides long.

D) Cleavage activity of fourteen Cas13b orthologs targeting ssRNA 1 using a homopolymer uridine sensor five nucleotides long.

E) Cleavage activity of fourteen Cas13b orthologs targeting ssRNA 1 using a homopolymer guanine sensor five nucleotides long.

F) Cleavage activity of fourteen Cas13b orthologs targeting ssRNA 1 using a homopolymer cytidine sensor five nucleotides long.

10.4.4 Figure 5.S4

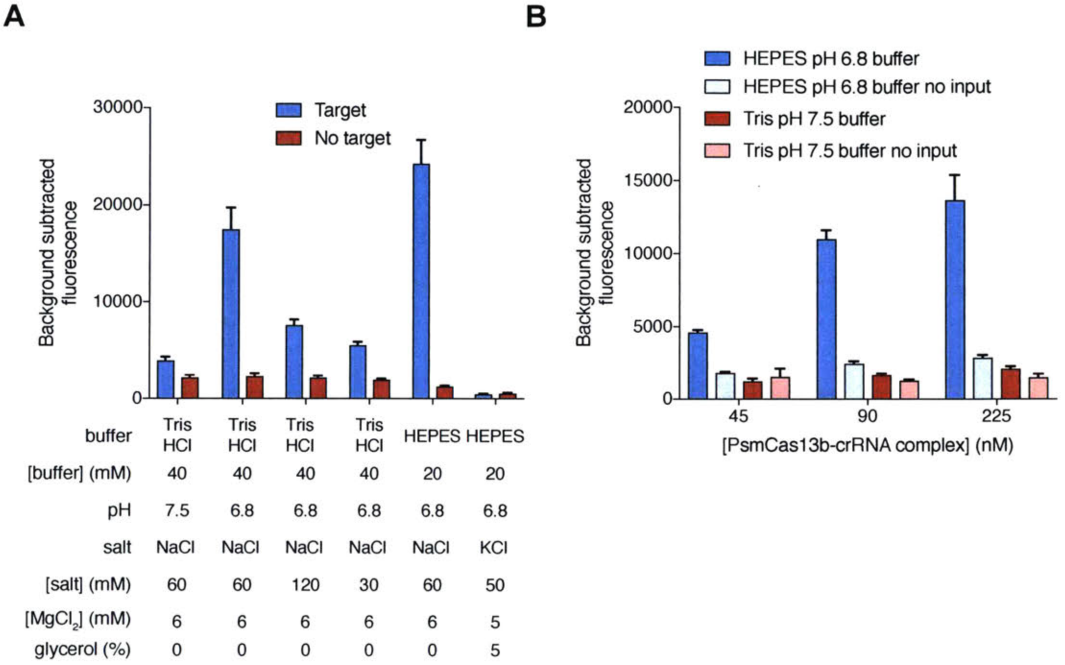


Figure 5.S4: Buffer optimization of PsmCas13b cleavage activity.

A) A variety of buffers are tested for their effect on PsmCas13b collateral activity after targeting ssRNA 1.

B) The optimized buffer is compared to the original buffer at different PsmCas13b-crRNA complex concentrations.

10.4.5 Figure 5.S5

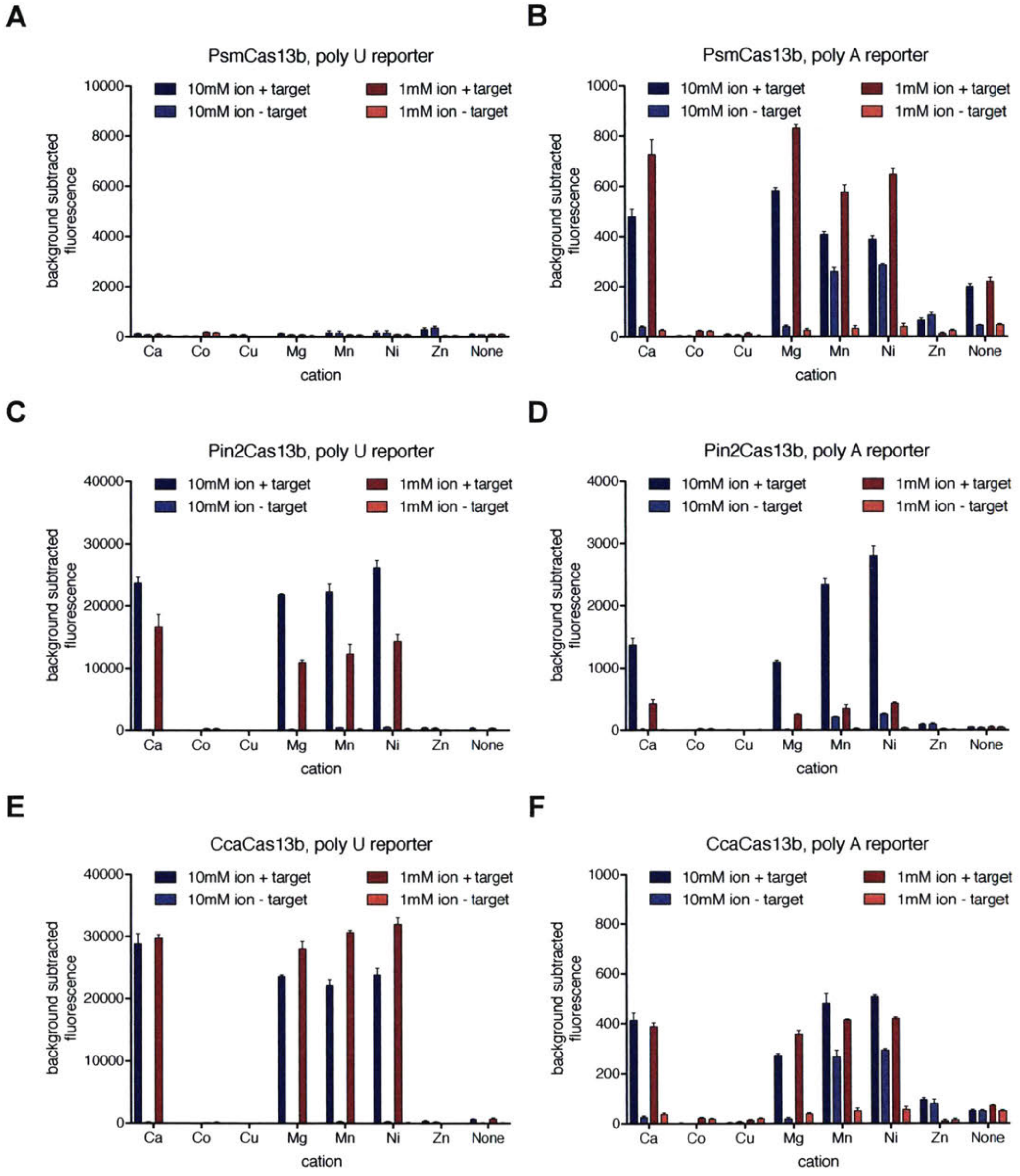


Figure 5.S5: Ion preference of Cas13 orthologs for collateral cleavage.

A) Cleavage activity of PsmCas13b with a fluorescent poly U sensor for divalent cations Ca, Co, Cu, Mg, Mn, Ni, and Zn. PsmCas13b is incubated with a crRNA targeting a synthetic ssRNA 1.

B) Cleavage activity of PsmCas13b with a fluorescent poly A sensor for divalent cations Ca, Co, Cu, Mg, Mn, Ni, and Zn. PsmCas13b is incubated with a crRNA targeting a synthetic ssRNA 1.

C) Cleavage activity of Pin2Cas13b with a fluorescent poly U sensor for divalent cations Ca, Co, Cu, Mg, Mn, Ni, and Zn. Pin2Cas13b is incubated with a crRNA targeting a synthetic ssRNA 1.

D) Cleavage activity of Pin2Cas13b with a fluorescent poly A sensor for divalent cations Ca, Co, Cu, Mg, Mn, Ni, and Zn. Pin2Cas13b is incubated with a crRNA targeting a synthetic ssRNA 1.

E) Cleavage activity of CcaCas13b with a fluorescent poly U sensor for divalent cations Ca, Co, Cu, Mg, Mn, Ni, and Zn. CcaCas13b is incubated with a crRNA targeting a synthetic ssRNA 1.

F) Cleavage activity of CcaCas13b with a fluorescent poly A sensor for divalent cations Ca, Co, Cu, Mg, Mn, Ni, and Zn. CcaCas13b is incubated with a crRNA targeting a synthetic ssRNA 1.

10.4.6 Figure 5.S6

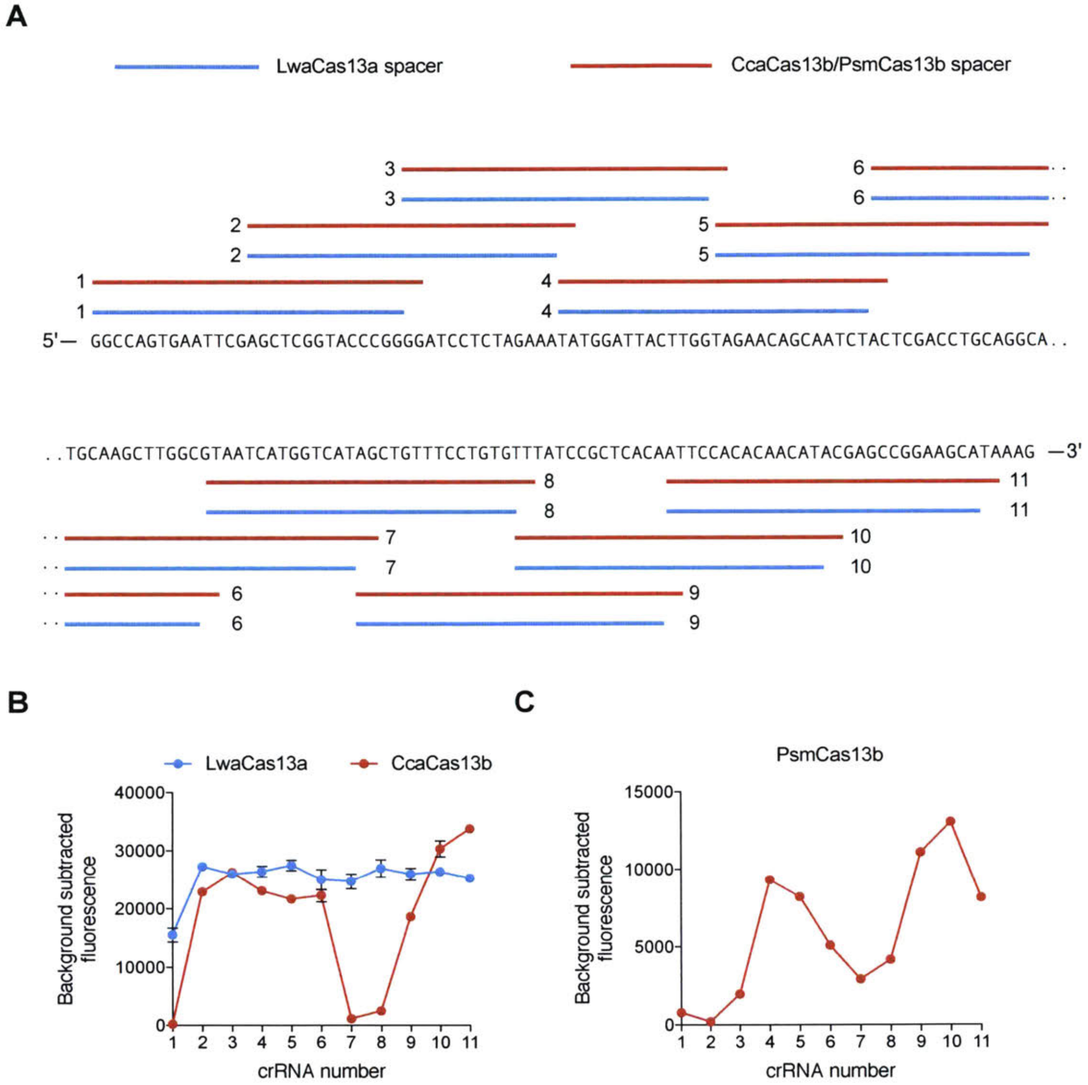


Figure 5.S6: Testing Cas13 ortholog reprogrammability with crRNAs tiling ssRNA 1.

A) Schematic of locations tiled crRNA targeting ssRNA 1.

B) Cleavage activity of LwaCas13a and CcaCas13b with crRNAs tiled across ssRNA1.

C) Cleavage activity of PsmCas13b with crRNAs tiled across ssRNA1.

10.4.7 Figure 5.S7

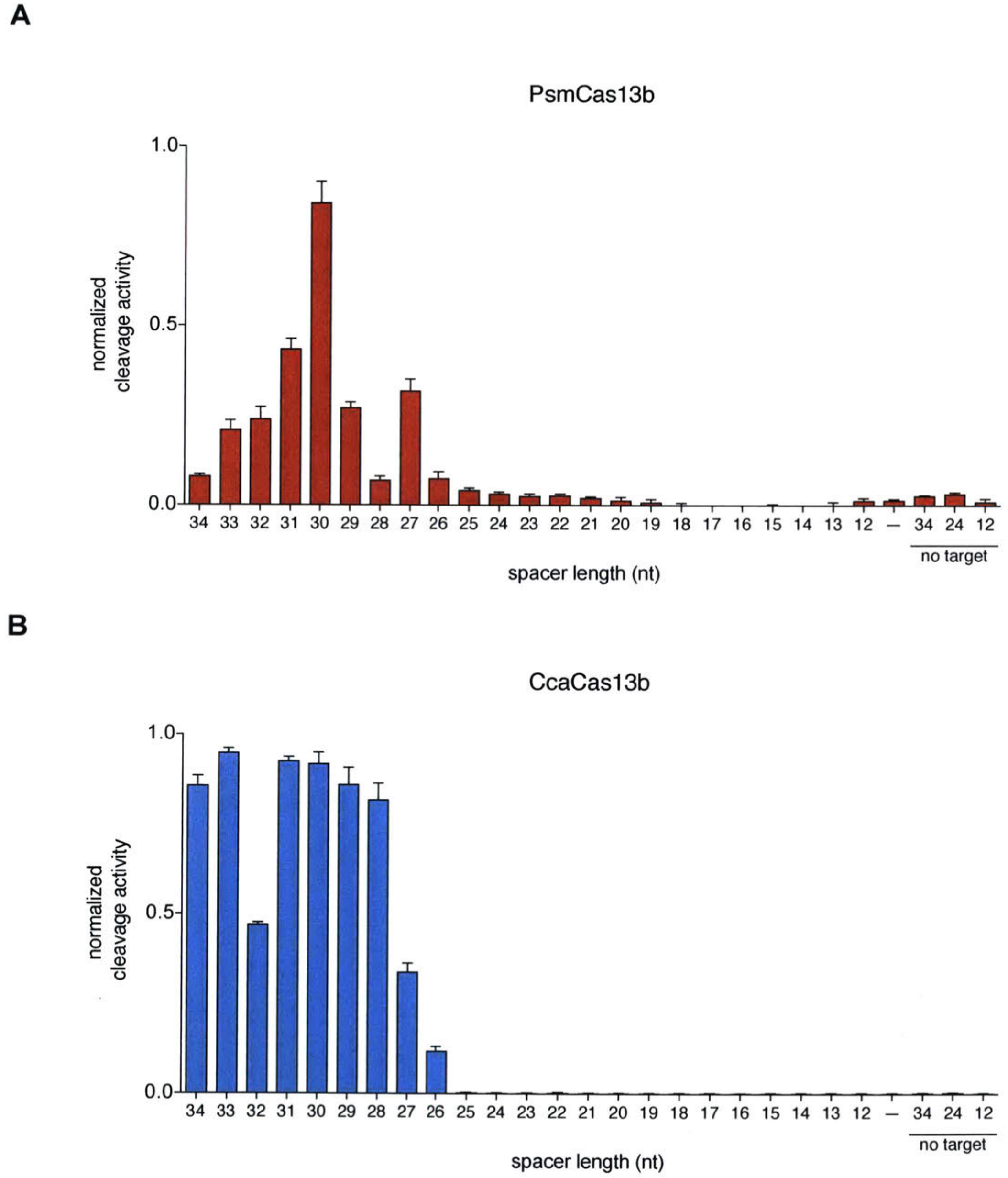


Figure 5.S7: Effect of crRNA spacer length on Cas13 ortholog cleavage

A) Cleavage activity of PsmCas13b with ssRNA1-targeting crRNAs of varying spacer lengths.

B) Cleavage activity of CcaCas13b with ssRNA1-targeting crRNAs of varying spacer lengths.

10.4.8 Figure 5.S8

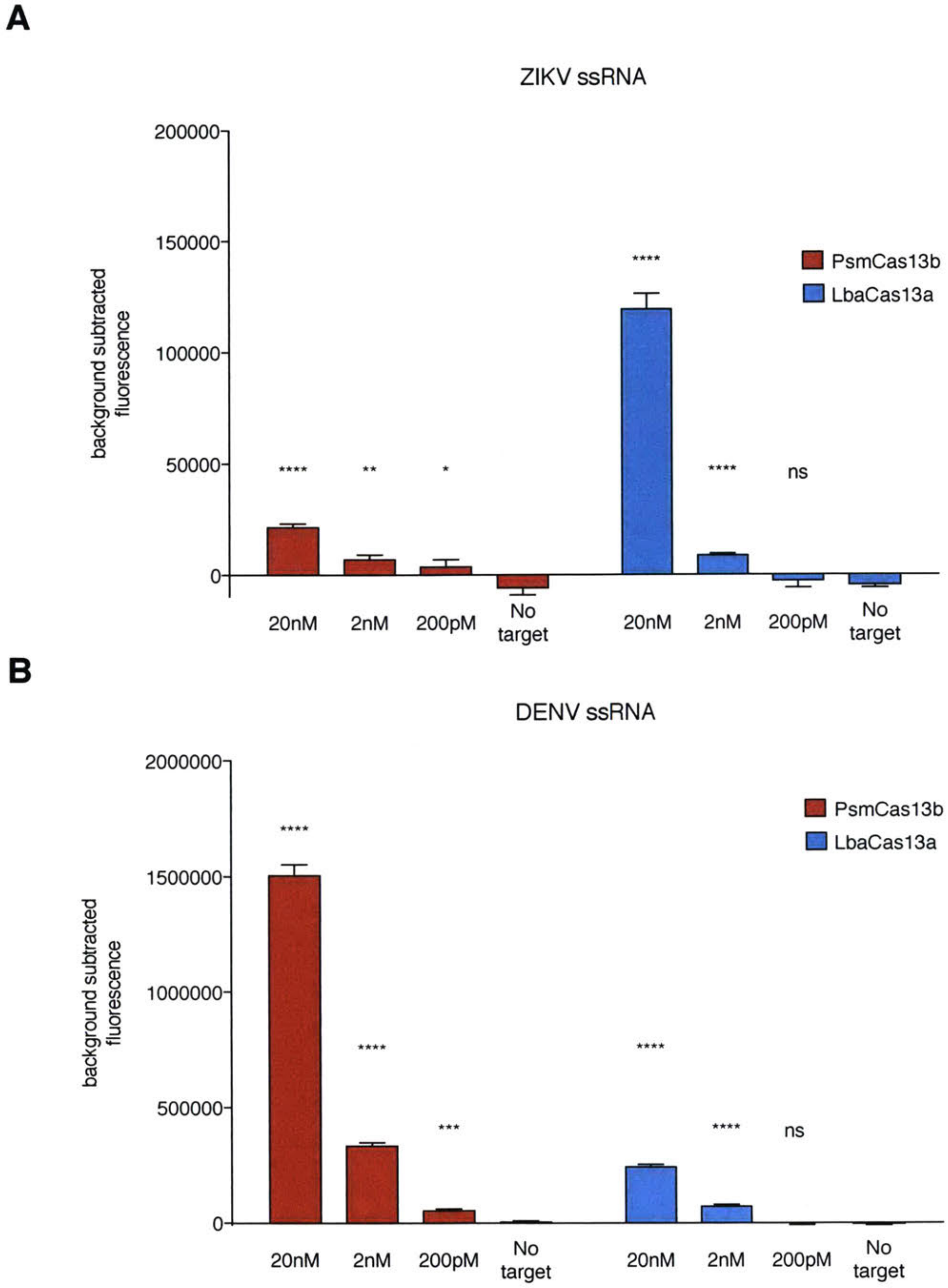


Figure 5.S8: Comparison of cleavage activity for Cas13 orthologs with adenine cleavage preference

A) Cleavage activity of PsmCas13b and LbaCas13a incubated with respective crRNAs targeting the ZIKV ssRNA target at different concentrations (n=4 technical replicates, two-tailed

Student t-test; n.s., not significant; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p<0.0001$; bars represent mean ± s.e.m.).

B) Cleavage activity of PsmCas13b and LbaCas13a incubated with respective crRNAs targeting a synthetic DENV ssRNA target at different concentrations (n=4 technical replicates, two-tailed Student t-test; n.s., not significant; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p<0.0001$; bars represent mean ± s.e.m.).

10.4.9 Figure 5.S9

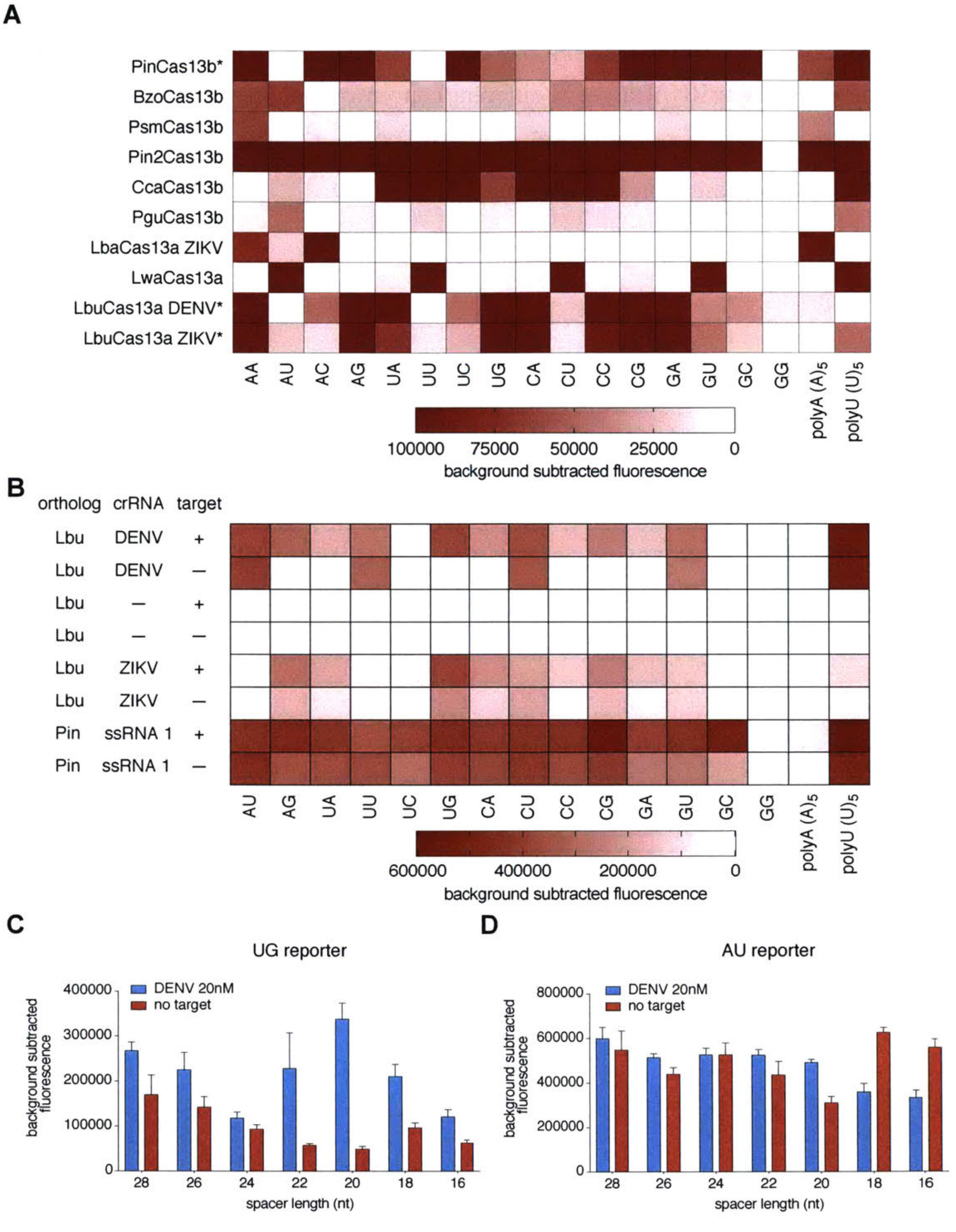


Figure 5.S9: Di-nucleotide preferences of Cas13a/b enzymes

A) Heatmap of the di-nucleotide base preference of 10 Cas13a/b orthologs targeting ssRNA 1, unless otherwise indicated, with reporters consisting of a di-nucleotide of A, C, G, or U RNA bases. (*) represent non-background subtracted orthologs with high background activity.

B) Heatmap of the di-nucleotide base preference of the high background activity orthologs LbuCas13a and PinCas13b tested on indicated targets.

C) Cleavage activity of LbuCas13a on AU di-nucleotide motif with and without 20nM DENV ssRNA target tested with varying spacer lengths.

D) Cleavage activity of LbuCas13a on UG di-nucleotide motif with and without 20nM DENV ssRNA target tested with varying spacer lengths.

10.4.10 Figure 5.S10

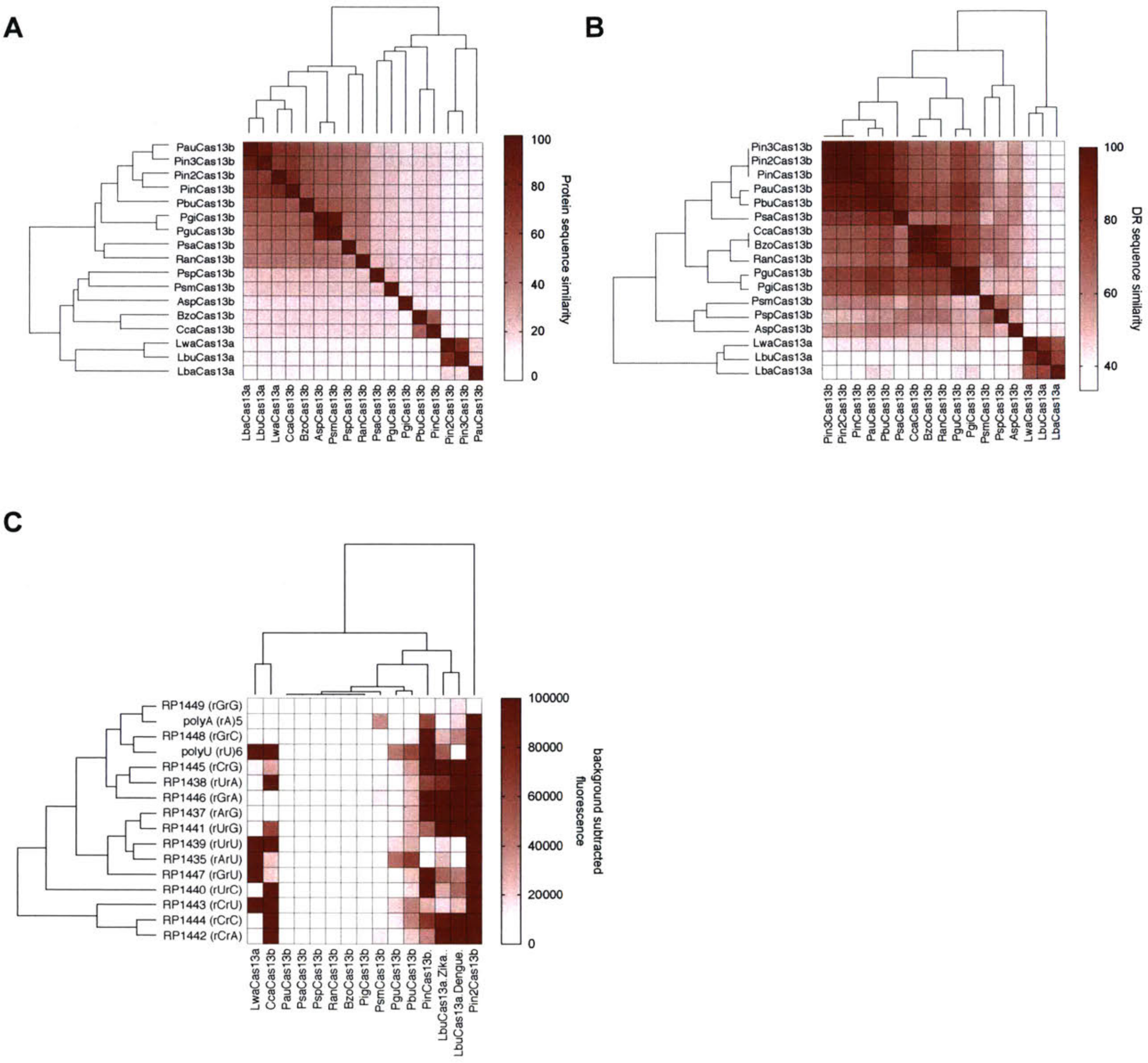


Figure 5.S10: Relationship of Cas13 families with di-nucleotide cleavage preferences

A) Protein sequence similarity matrix based on multiple protein sequence alignment of several Cas13a and Cas13b ortholog members. Clustering is shown based on Euclidean distance.

B) Direct repeat sequence similarity matrix based on multiple sequence alignment of several Cas13a and Cas13b direct repeat sequences. Clustering is shown based on Euclidean distance

C) Clustering of the Cas13 cleavage activity base preferences of dinucleotide motif reporters.

10.4.11 Figure 5.S11

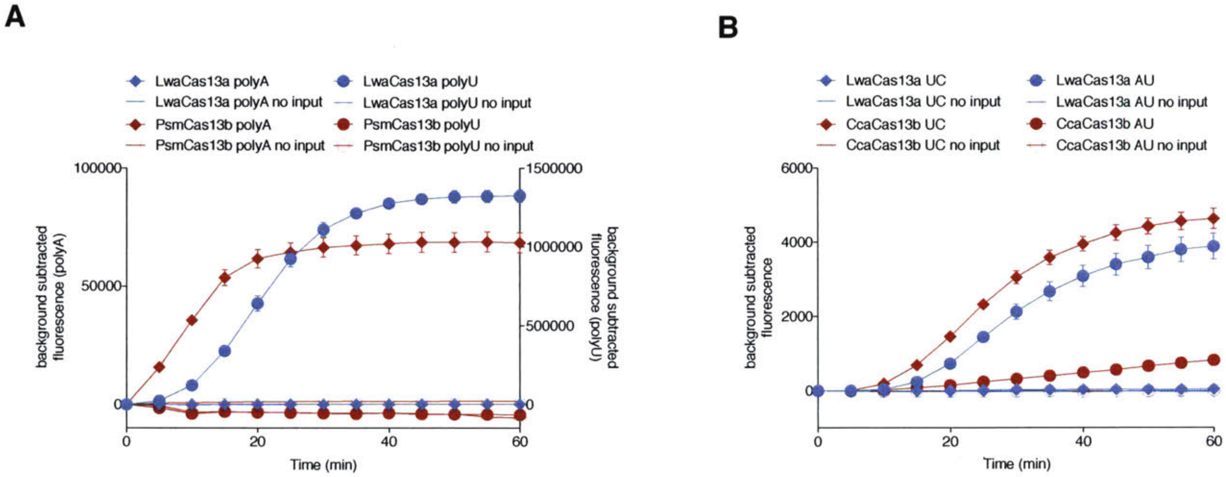


Figure 5.S11: Kinetics of cleavage activity for Cas13 enzymes with orthogonal cleavage preferences

A) Orthogonal base preferences of PsmCas13b and LwaCas13a targeting ssRNA 1 with either a U_6 or A_6 reporter.

B) Orthogonal base preferences of CcaCas13b and LwaCas13a targeting DENV RNA with either a AU or UC reporter.

10.4.12 Figure 5.S12

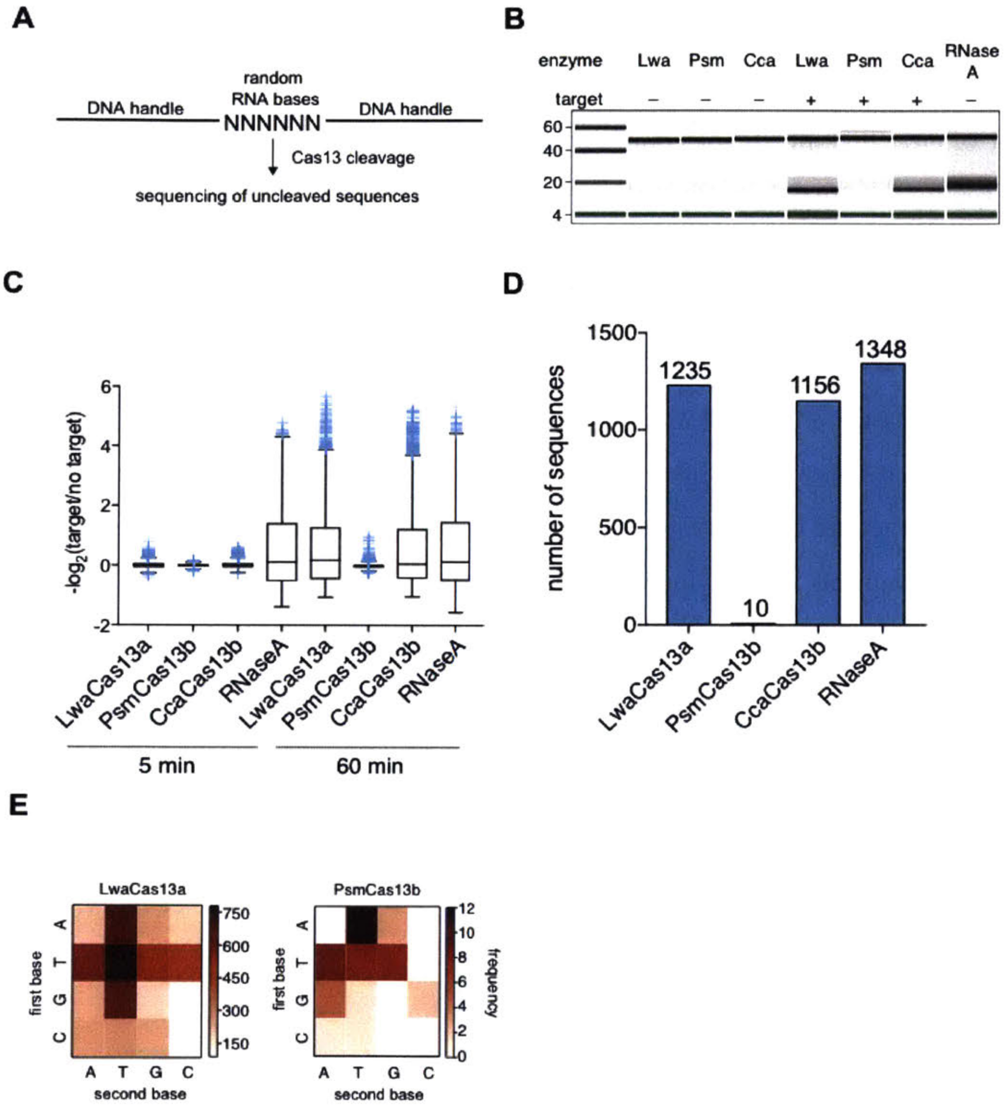


Figure 5.S12: Random motif cleavage screen for testing Cas13 base preferences

A) Schematic of cleavage motif preference discovery screen for comparing random motif prefences.

B) Bioanalyzer traces for LwaCas13a-, PsmCas13b-, CcaCas13b-, and RNase A-treated library samples showing changes in library size after RNase activity. Cas13 orthologs are targeting

DENV ssRNA and cleave the random motif-library due to collateral cleavage. Marker standards are shown in the first lane.

C) Box plots showing motif distribution of target to no-target ratios for LwaCas13a, PsmCas13b, CcaCas13b, and RNase A at 5 minute and 60 minute timepoints. RNase A ratios were compared to the average of the three Cas13 no-target conditions. Ratios are also an average of two cleavage reaction replicates.

D) Number of enriched motifs for LwaCas13a, PsmCas13b, CcaCas13b, and RNase A at the 60 minute timepoint. Enrichment motif was calculated as motifs above -$\log_2$(target/no target) thresholds of either 1 (LwaCas13a, CcaCas13b, and RNase A) or 0.5 (PsmCas13b). A threshold of 1 corresponds to at least 50% depletion while a threshold of 0.5 corresponds to at least 30% depletion.

E) Preferred two-base motifs for LwaCas13a and PsmCas13b. Values represented in the heatmap are the the counts of each two-base across all depleted motifs. Motifs are considered depleted if the -log2(target/no target) value is above 1.0 in the LwaCas13a condition or 0.5 in the PsmCas13b condition. In the -log2(target/no target) value, target and no target denote the frequency of a motif in the target and no target conditions, respectively.

10.4.13 Figure 5.S13

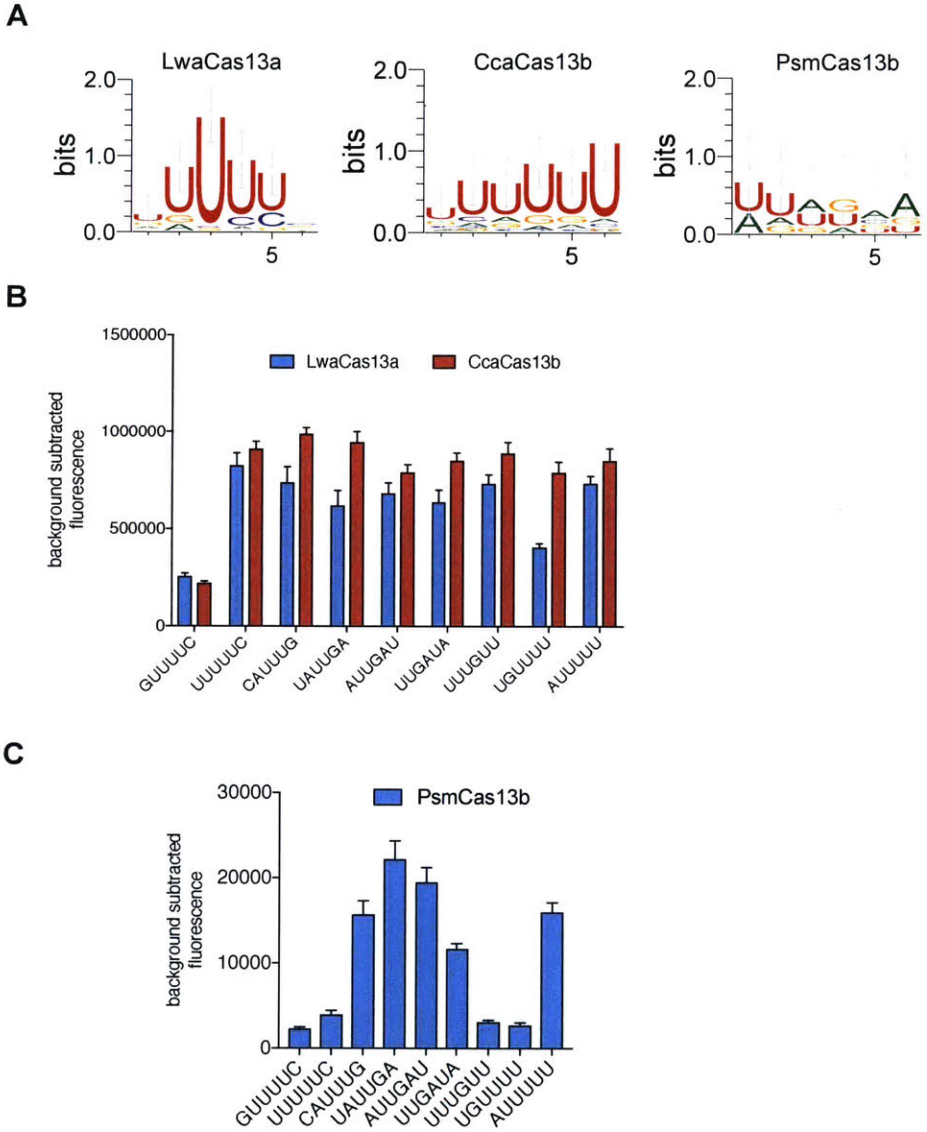


Figure 5.S13: Motifs and orthogonal sequences from random motif cleavage screen

A) Sequence logos generated from enriched motifs for LwaCas13a, PsmCas13b, and CcaCas13b. LwaCas13a and CcaCas13b show a strong U preference as would be expected, while PsmCas13b shows a unique preference for A bases across the motif, which is consistent with homopolymer collateral activity preferences.

B) Collateral activity of LwaCas13a and CcaCas13b targeting DENV ssRNA on most depleted motifs from the RNA collateral motif screen.

C) Collateral activity of PsmCas13b targeting DENV ssRNA on most depleted motifs from the RNA collateral motif screen.

10.4.14 Figure 5.S14

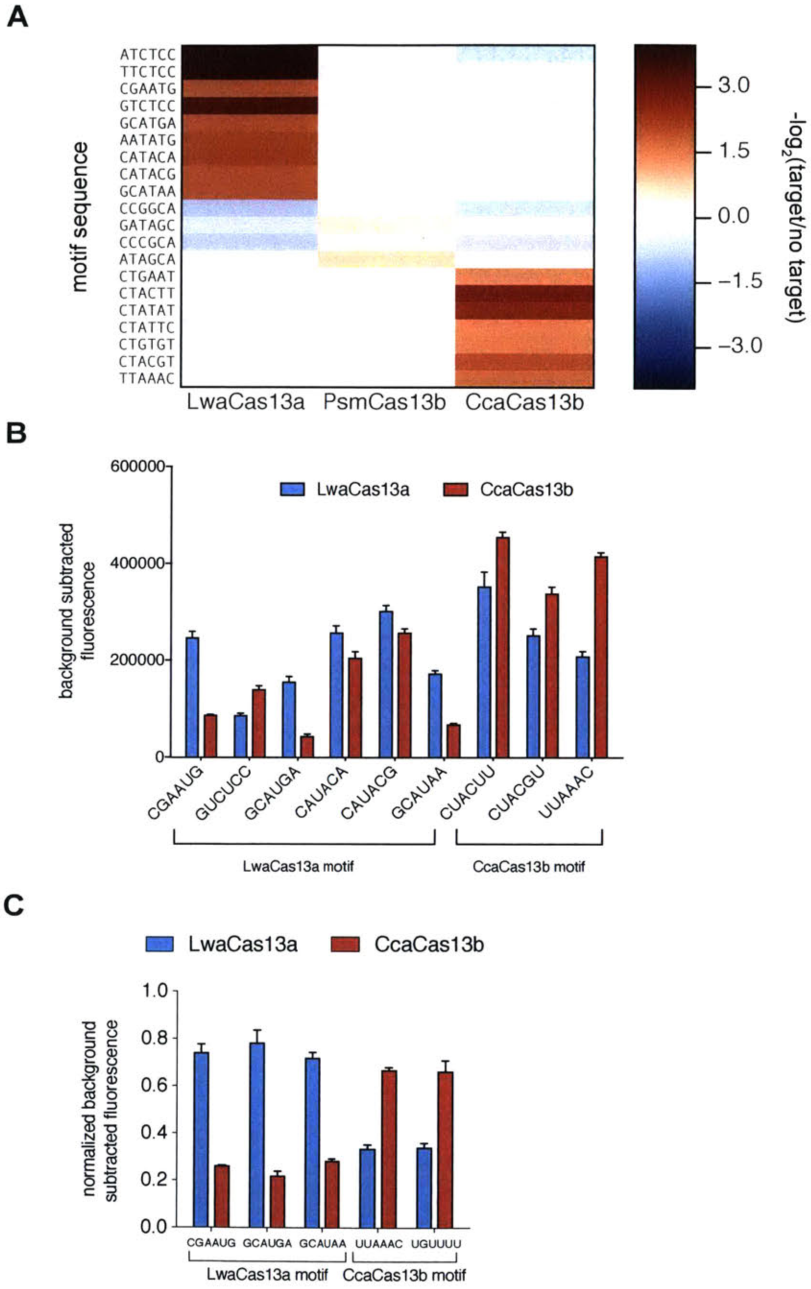


Figure 5.S14: Comparison of top collateral activity motifs from the RNA motif collateral activity screens.

A) Heatmap showing the orthogonal motif preferences of LwaCas13a, PsmCas13b, and CcaCas13b. Values represented in the heatmap are the -log$_2$(target/no target) value of each shown motif. In the -log$_2$(target/no target) value, target and no target denote the frequency of a motif in the target and no target conditions, respectively.

B) Cleavage activity of LwaCas13a and CcaCas13b on top orthogonal motifs derived from the motif preference discovery screen

C) Collateral activity of LwaCas13a and CcaCas13b targeting DENV ssRNA on top orthogonal RNA motifs.

10.4.15 Figure 5.S15

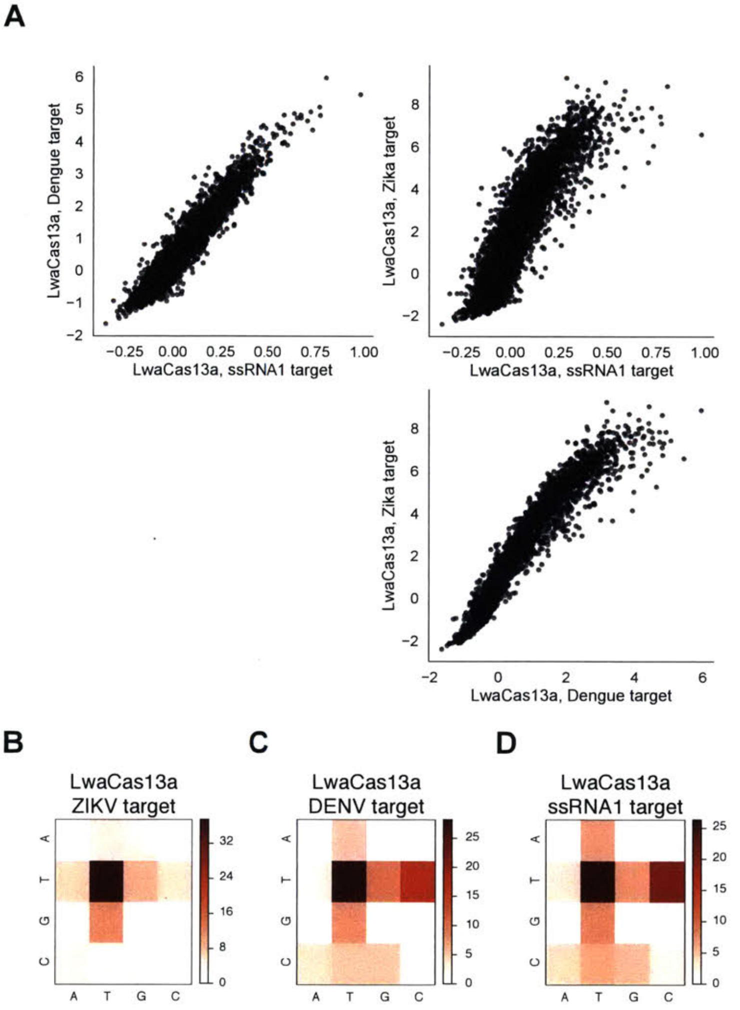


Figure 5.S15: Comparison of random motif library screen on different targets and enzymes

A) Pair-wise comparison of enrichment scores for different activating targets with LwaCas13a.

B) Heatmaps showing two-base preference for LwaCas13a with the ZIKV ssRNA target as determined by the random motif library cleavage screen. Values represented in the heatmap are the the counts of each 2-base across all depleted motifs. Motifs are considered depleted if the -$\log_2$(target/no target) value is above 1.0.

C) Heatmaps showing two-base preference for LwaCas13a with the DENV ssRNA target as determined by the random motif library cleavage screen. Values represented in the heatmap are the the counts of each 2-base across all depleted motifs. Motifs are considered depleted if the -$\log_2$(target/no target) value is above 1.0.

D) Heatmaps showing two-base preference for LwaCas13a with the ssRNA1 target as determined by the random motif library cleavage screen. Values represented in the heatmap are the the counts of each 2-base across all depleted motifs. Motifs are considered depleted if the -$\log_2$(target/no target) value is above 1.0.

10.4.16 Figure 5.S16

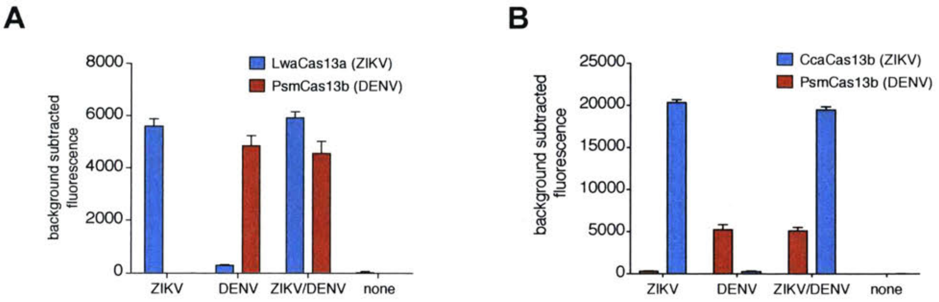


Figure 5.S16: Multiplexed detection of ZIKV ssRNA and DENV ssRNA targets.

A) In-sample multiplexed detection of 20 nM ZIKV and DENV ssRNA targets with LwaCas13a and PsmCas13b collateral activity.

B) In-sample multiplexed detection of 20 pM ZIKV and DENV ssRNA targets with CcaCas13a and PsmCas13b collateral activity.

10.4.17 Figure 5.S17

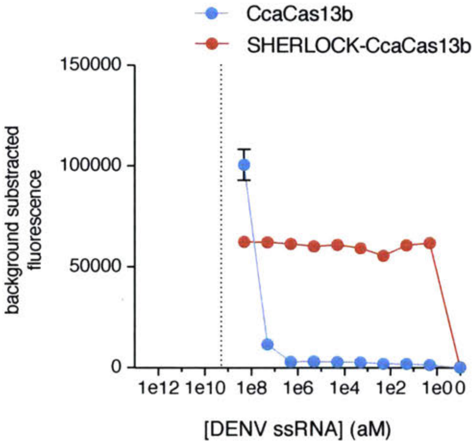


Figure 5.S17: Attomolar detection of CcaCas13b-SHERLOCK

Comparison of collateral activity and pre-amplification enhanced collateral (SHERLOCK) of CcaCas13b.

10.4.18 Figure 5.S18

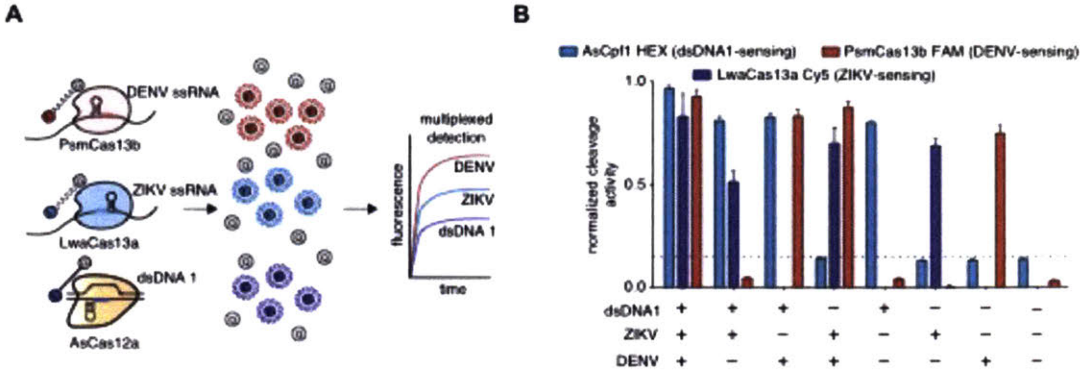


Figure 5.S18: Triplex detection using orthogonal CRISPR enzymes

I) Schematic of in-sample 3 channel multiplexing using orthogonal Cas13 and Cas12a enzymes.

J) In-sample multiplexed detection of ZIKV ssRNA, DENV ssRNA, and dsDNA 1 with LwaCas13a, PsmCas13b, and Cas12a.

10.4.19 Figure 5.S19

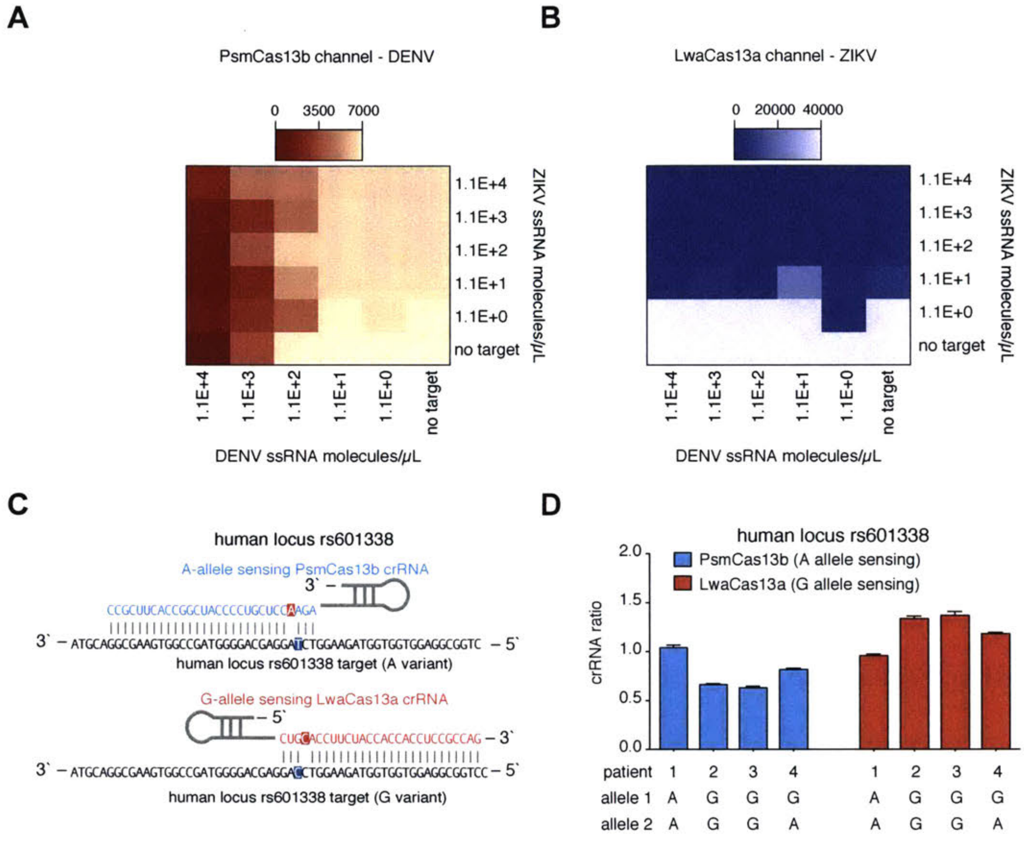


Figure 5.S19: In-sample multiplexed RNA detection of ZIKV ssRNA and DENV ssRNA targets and human genotyping.

A) In-sample multiplexed RPA and collateral detection at decreasing concentrations of ZIKV and DENV ssRNA targets with PsmCas13b.

B) In-sample multiplexed RPA and collateral detection at decreasing concentrations of ZIKV and DENV ssRNA targets with LwaCas13a.

C) Schematic of crRNA design and target sequences for multiplexed genotyping at rs601338 with LwaCas13a and PsmCas13b.

D) Multiplexed genotyping with human samples at rs601338 with LwaCas13a and PsmCas13b.

10.4.20 Figure 5.S20

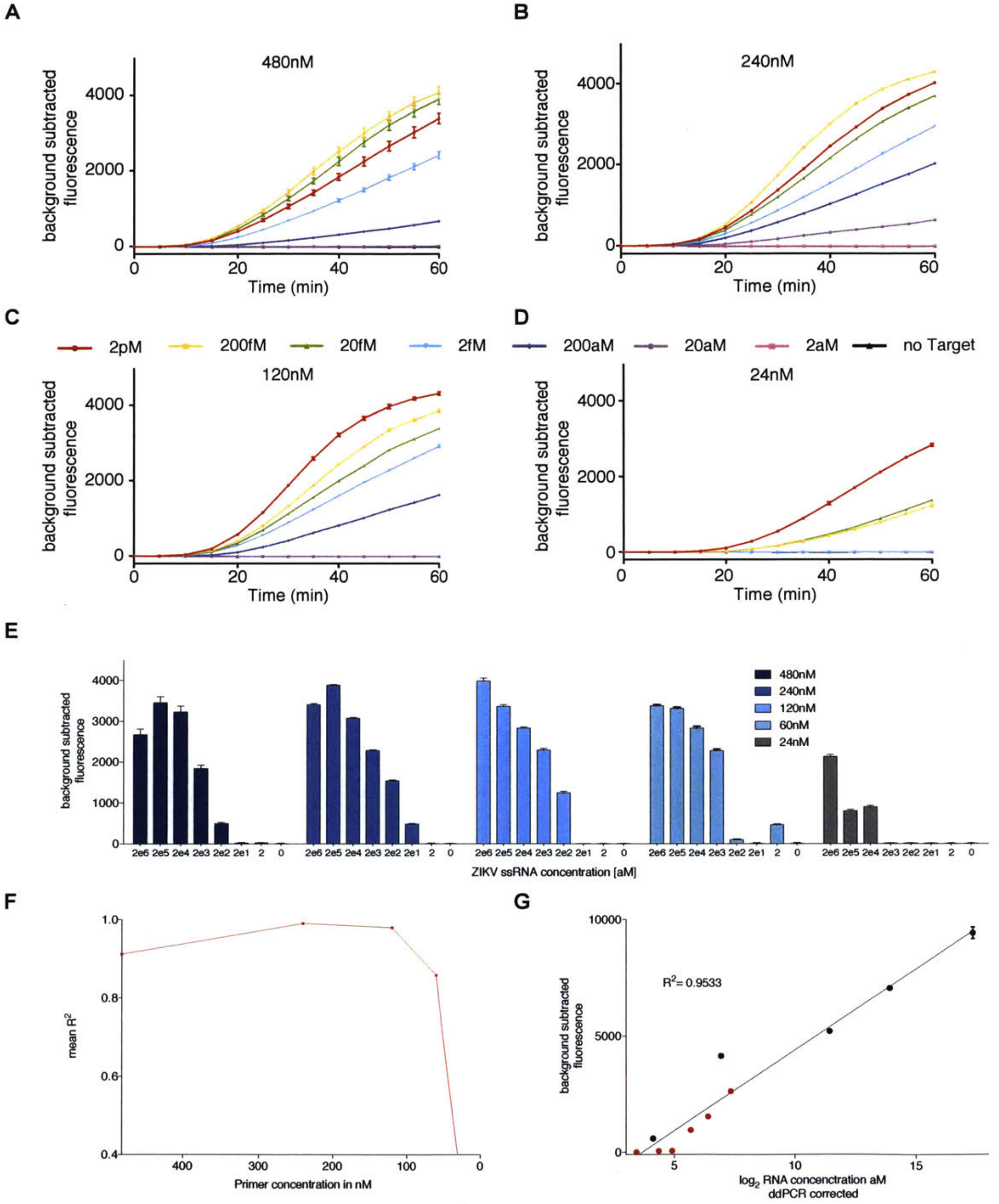


Figure 5.S20: Optimizing primer concentration for quantitative SHERLOCK

A) SHERLOCK kinetic curves of LwaCas13a incubated with ZIKV ssRNA targets of different concentration and a complementary crRNA at an RPA primer concentration of 480nM.

B) SHERLOCK kinetic curves of LwaCas13a incubated with ZIKV ssRNA targets of different concentration and a complementary crRNA at an RPA primer concentration of 240nM.

C) SHERLOCK kinetic curves of LwaCas13a incubated with ZIKV ssRNA targets of different concentration and a complementary crRNA at an RPA primer concentration of 120nM.

D) SHERLOCK kinetic curves of LwaCas13a incubated with ZIKV ssRNA targets of different concentration and a complementary crRNA at an RPA primer concentration of 24nM.

E) SHERLOCK detection of ZIKV ssRNA of different concentrations at with four different RPA primer concentrations: 480nM, 240nM, 120nM, 60nM, and 24nM.

F) The mean R^2 correlation between background subtracted fluorescence of SHERLOCK and the ZIKV ssRNA target RNA concentration at different RPA primer concentrations.

G) Quantitative SHERLOCK detection of ZIKV ssRNA targets at different concentrations in a 10-fold dilution series (black points) and 2-fold dilution series (red points). An RPA primer concentration of 240nM was used.

10.4.21 Figure 5.S21

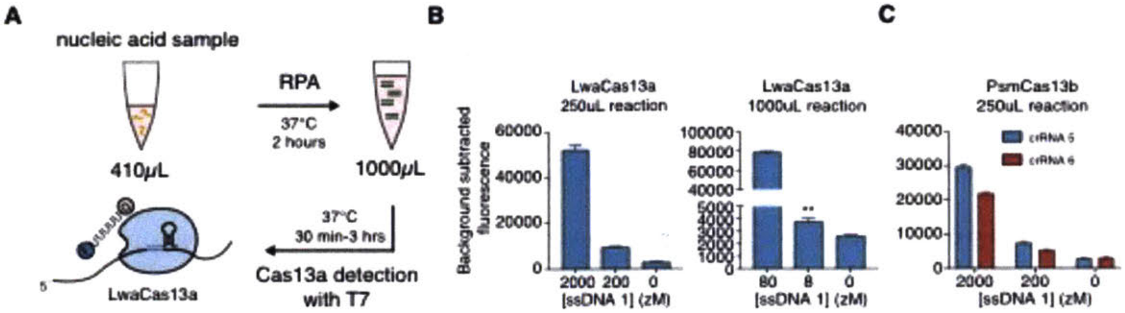


Figure 5.S21: Large volume SHERLOCK reactions with sub-attomolar sensitivity

A) Schematic of large reactions for increased sensitivity single molecule detection.

B) Single molecule SHERLOCK detection with LwaCas13a in large reaction volumes for increased sensitivity targeting ssRNA target 1. For 250μL reaction volumes, 100μL of sample input is used and for 1000μL reaction volumes, 540μL of sample input is used.

C) Single molecule SHERLOCK detection with PsmCas13b in large reaction volumes for increased sensitivity targeting ssRNA target 1. For 250μL reaction volumes, 100μL of sample input is used.

10.4.22 Figure 5.S22

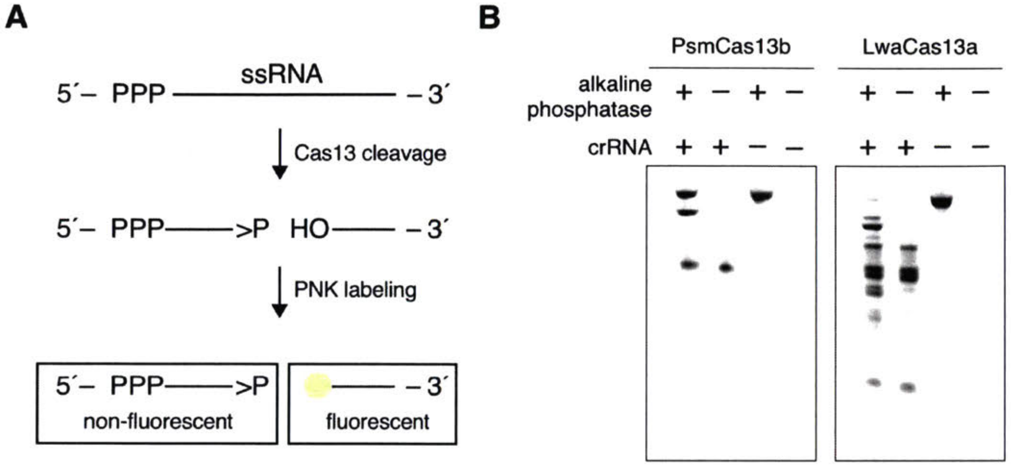


Figure 5.S22: Profiling of cleavage ends generated by LwaCas13a and PsmCas13b

A) Schematic for detection of 2,3 cyclic phosphate ends via PNK labeling and gel electrophoresis.

B) Electrophoresis gel demonstrating 2,3 cyclic phosphate ends generated by LwaCas13a or PsmCas13b cleavage of ssRNA target 2 or 3 (homopolymer loops). The Cas13 enzyme is incubated with the appropriate crRNA targeting the ssRNA target and the cleavage products are 5' labeled with a dye IR800 with or without alkaline phosphatase treatment.

10.4.23 Figure 5.S23

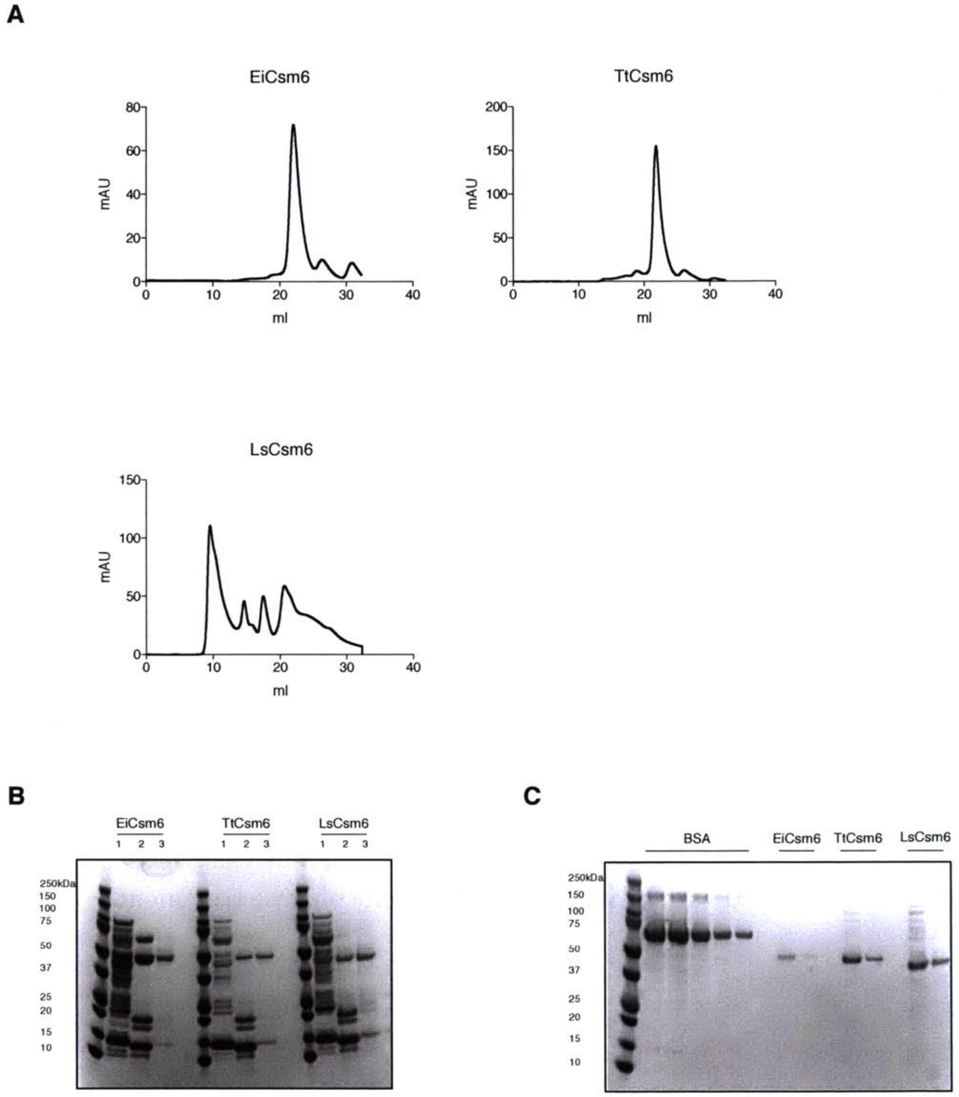


Figure 5.S23: Protein purification of Csm6 orthologs

A) Chromatograms of size exclusion chromatography for EiCsm6, TtCsm6, LsCsm6 and SaCsm6 used in this study. Measured UV absorbance (mAU) is shown against the elution volume (ml).

B) SDS-PAGE gel of EiCsm6, TtCsm6 and LsCsm6 fractions prior to size exclusion chromatography. Fractions show the bacterial lysate supernatant (1) after streptactin incubation, streptactin resins after cleavage with SUMO protease (2), as well as released, untagged Csm6 protein (3).

C) Final SDS-PAGE of concentrated Csm6 proteins after size exclusion chromatography. BSA standard curve (left) is used to quantify Csm6 proteins (right). Five dilutions of BSA and two dilutions of EiCsm6, TtCsm6 and LsCsm6 are shown.

10.4.24 Figure 5.S24

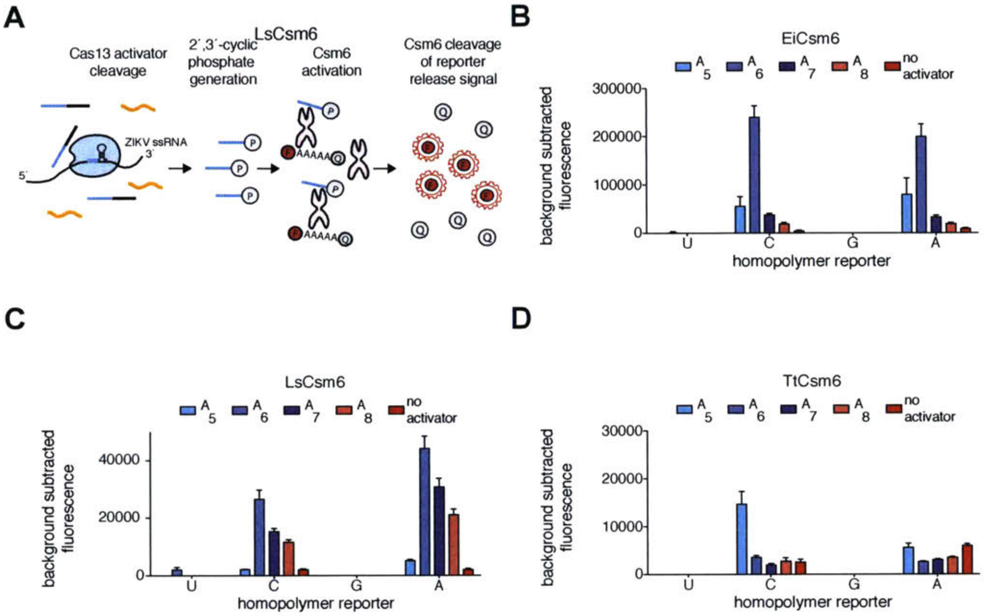


Figure 5.S24: Base preference and activation of Csm6 orthologs

A) Schematic for Csm6-mediated positive feedback in a SHERLOCK reaction.

B) Activation of EiCsm6 by 2′,3′-cyclic phosphate-terminated adenine oligomers of different lengths. Csm6 cleavage is measured using an RNA reporter consisting of A, C, G, or U homopolymer with ends labeled with a fluorophore and quencher.

C) Base preference of LsCsm6 cleavage activated by 2′,3′-cyclic phosphate-terminated adenine oligomers.

D) Base preference of TtCsm6 cleavage activated by 2′,3′-cyclic phosphate-terminated adenine oligomers.

10.4.25 Figure 5.S25

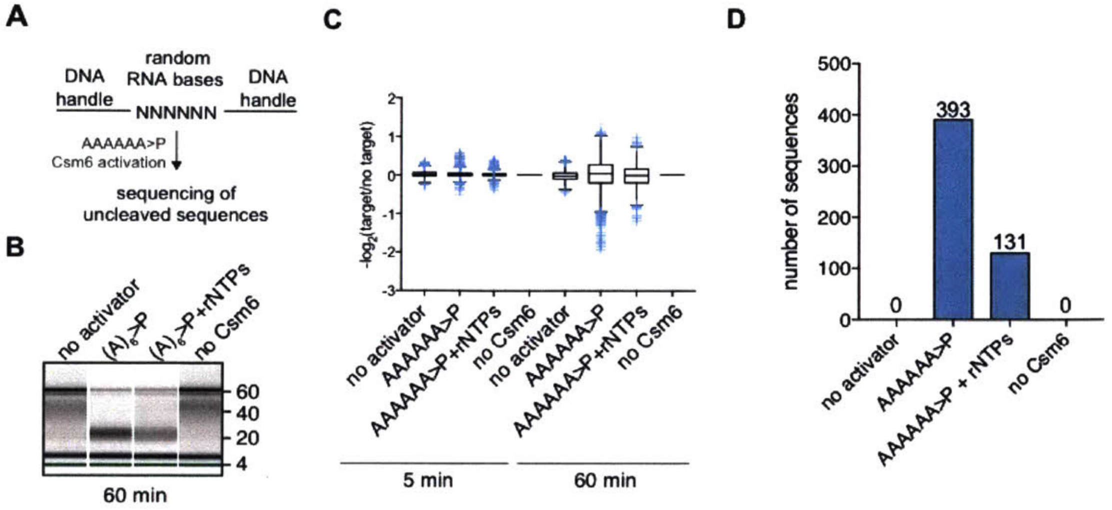


Figure 5.S25: Size analysis and representation of various motifs after Csm6 cleavage.

A) Schematic of cleavage motif preference discovery screen for Csm6 orthologs.

B) Bioanalyzer traces for EiCsm6 samples showing changes in library size after RNase activity that is activator dependent.

C) Box plots showing motif distribution of target to non-target motif ratios for Csm6, Csm6 with activator, Csm6 with activator and rNTPs, or background library at 5 minute and 60 minute timepoints.

D) Number of depleted motifs for Csm6, Csm6 with activator, Csm6 with activator and rNTPs, or background library at the 60 minute timepoint.

10.4.26 Figure 5.S26

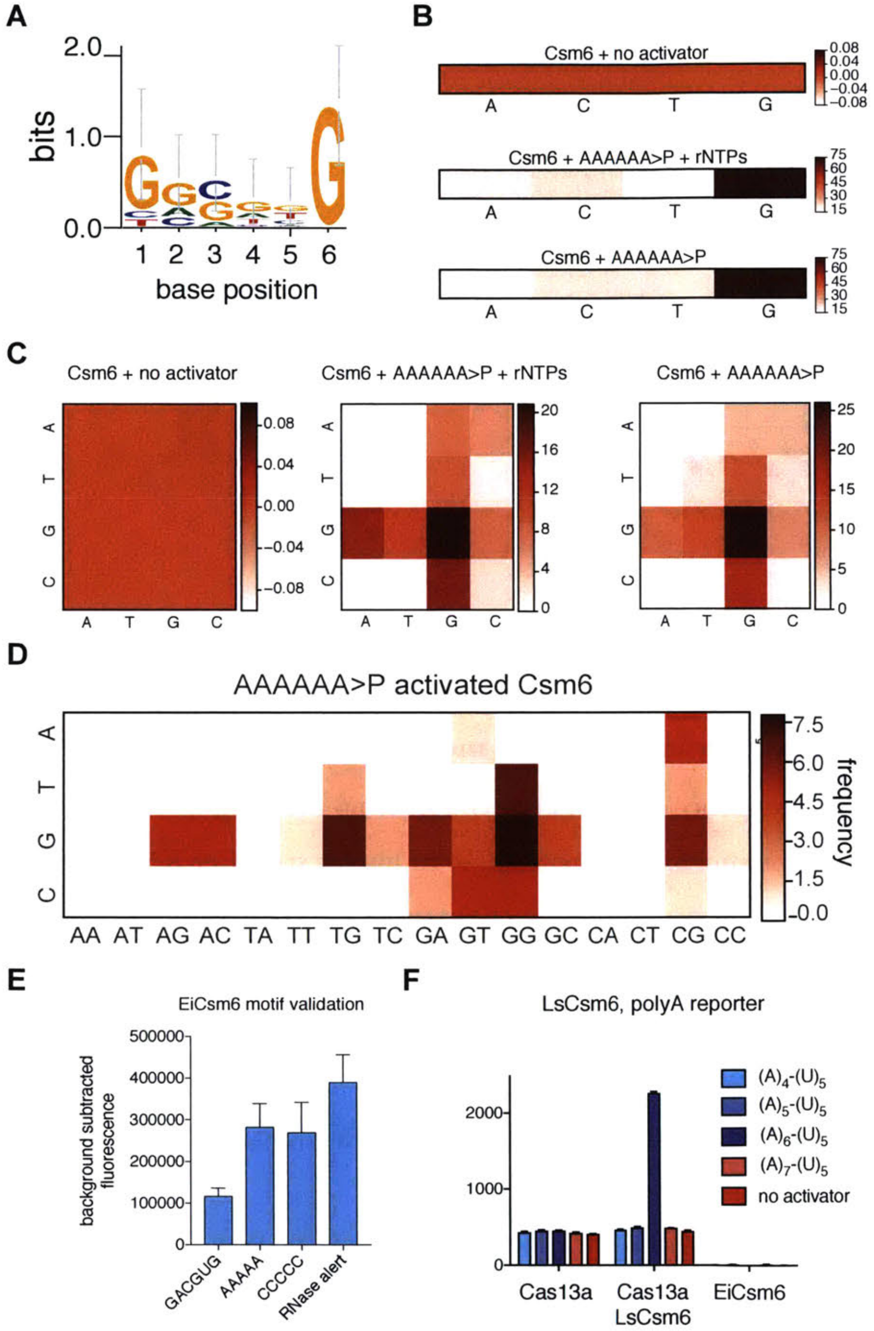


Figure 5.S26: Single- and two-base preferences of Csm6 conditions determined by random motif library screen.

A) Sequence logo of preferred sequence motif for EiCsm6 cleavage activity.

B) Heatmaps showing single base preferences for Csm6, Csm6 with activator, and Csm6 with activator and rNTPs at the 60 minute timepoint as determined by the random motif library cleavage screen. Values represented in the heatmap are the the counts of each single-base across all depleted motifs. Motifs are considered depleted if the -log$_2$(target/no target) value is above 0.5. In the -log$_2$(target/no target) value, target and no target denote the frequency of a motif in the target and no target conditions, respectively.

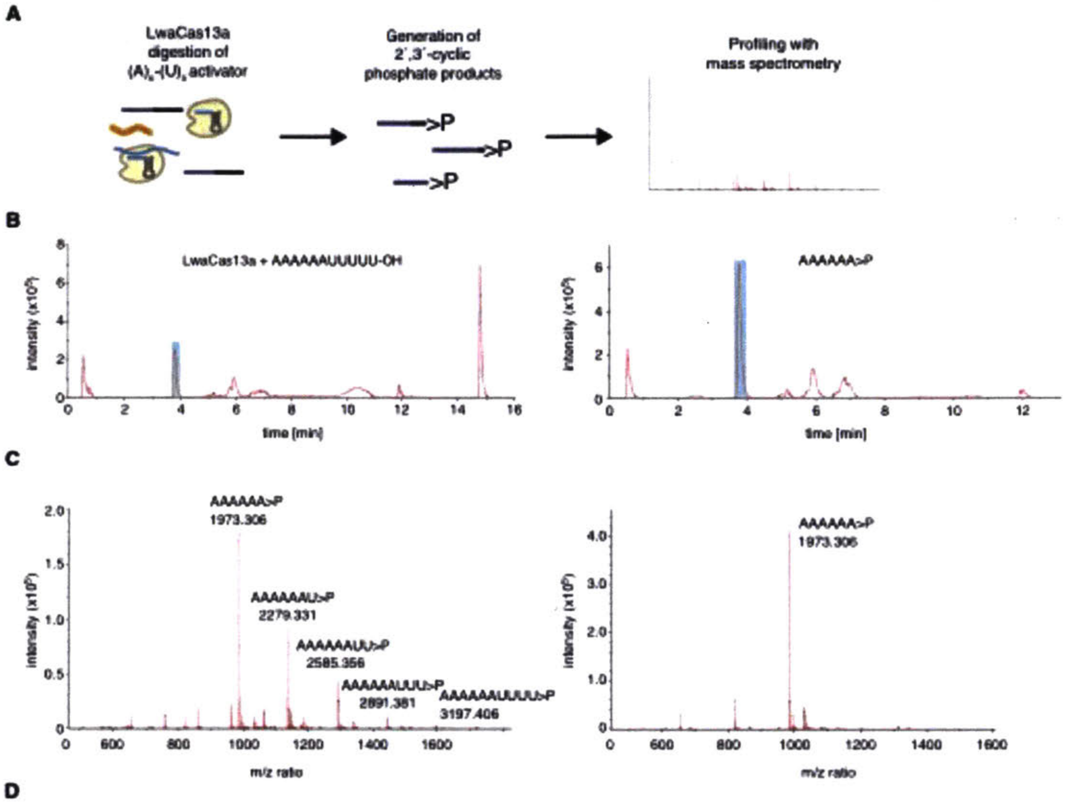
C) Heatmaps showing two-base preferences for Csm6, Csm6 with activator, and Csm6 with activator and rNTPs at the 60 minute timepoint as determined by the random motif library cleavage screen. Values represented in the heatmap are the the counts of each two-base across all depleted motifs. Motifs are considered depleted if the -log$_2$(target/no target) value is above 0.5. In the -log$_2$(target/no target) value, target and no target denote the frequency of a motif in the target and no target conditions, respectively.

D) Heatmap of preferred 3-base motifs for EiCsm6 cleavage activity. Values represented in the heatmap are the the counts of each 3-base across all depleted motifs. Motifs are considered depleted if the -log$_2$(target/no target) value is above 0.5. In the -log$_2$(target/no target) value, target and no target denote the frequency of a motif in the target and no target conditions, respectively.

E) Cleavage activity of EiCsm6 on top reporter sequences derived from the random motif library screen.

F) Activation of LsCsm6 by LwaCas13a cleavage of adenine-uridine activators with different length adenine tracts. LwaCas13a is targeting synthetic DENV ssRNA.

10.4.27 Figure 5.S27



D

LwaCas13a + AAAAAAUUUUU-OH

Cmpt .	Mass	Molecule	Abund.	Abd.[%]	Std.Dev.
A	1973.306	[M -H]-	196313	100	0.014056
B	2279.331	[M -H]-	107038	54.52	0.003627
C	2585.356	[M -H]-	62759	31.97	0.005015
D	2891.381	[M -H]-	32484	16.55	0.006336
E	3197.406	[M -H]-	19313	9.84	0.007767
F	3441.476	[M -H]-	14841	7.56	0.00678
G	1995.287	[M -H]-	13024	6.63	0.004689
H	1644.254	[M -H]-	11876	6.05	0.001244
I	2071.281	[M -H]-	11697	5.96	0.005988
J	2377.307	[M -H]-	10195	5.19	0.002957
K	2301.313	[M -H]-	7095	3.61	0.003939
L	2683.332	[M -H]-	7360	3.75	0.008262
M	2607.338	[M -H]-	5076	2.59	0.00323

AAAAAA>P

Cmpt .	Mass	Molecule	Abund.	Abd.[%]	Std.Dev.
A	1973.306	[M -H]-	439056	100	0.003272
B	1644.254	[M -H]-	65030	14.81	0.001781
C	2063.374	[M -H]-	49525	11.28	0.003049
D	1995.287	[M -H]-	29564	6.73	0.003413
E	2071.282	[M -H]-	27554	6.28	0.003287

Figure 5.S27: Mass spectrometry analysis of cleavage ends from LwaCas13a.

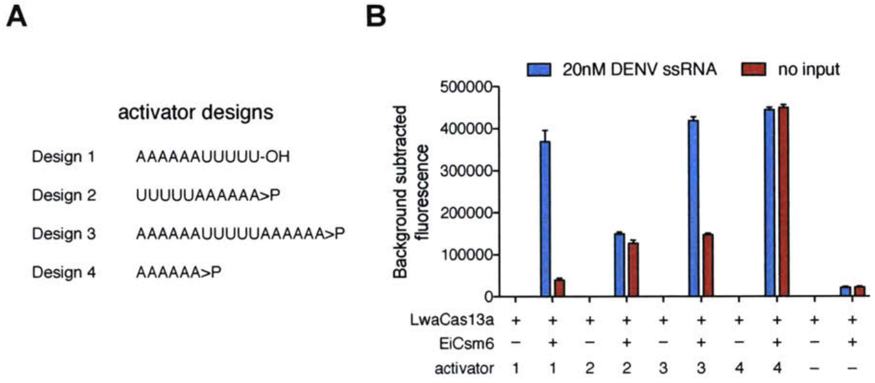
A) Schematic of LwaCas13a digestion and mass spectrometric analysis to verify cleavage products.

B) Mass spectrometry analysis of digestion products from LwaCas13a collateral cleavage (left) or 2,3 cyclic phosphate activator alone (right). Dominant peaks are labeled with mass and corresponding structure.

C) Chromatographic traces showing elution profiles for LwaCas13a-digested activator (top) or 2,3 cyclic phosphate activator (bottom). Blue highlighted peaks were analyzed for mass spectrometry in Fig. 5.

D) Table of abundances for different compounds detected by mass spectrometry in LwaCas13a-digested activator (left) or 2,3 cyclic phosphate activator (right) samples.

10.4.28 Figure 5.S28



activator designs

Design 1	AAAAAAUUUUU-OH
Design 2	UUUUUAAAAAA>P
Design 3	AAAAAAUUUUUAAAAAA>P
Design 4	AAAAAA>P

Figure 5.S28: Effect of reporter and activator optimizations on Csm6-enhancement of LwaCas13a activity.

A) Schematic of different activator designs for Csm6 enhancement of Cas13a activity.

B) Performance of EiCsm6 enhancement of LwaCas13a detection for different activator designs.

10.4.29 Figure 5.S29

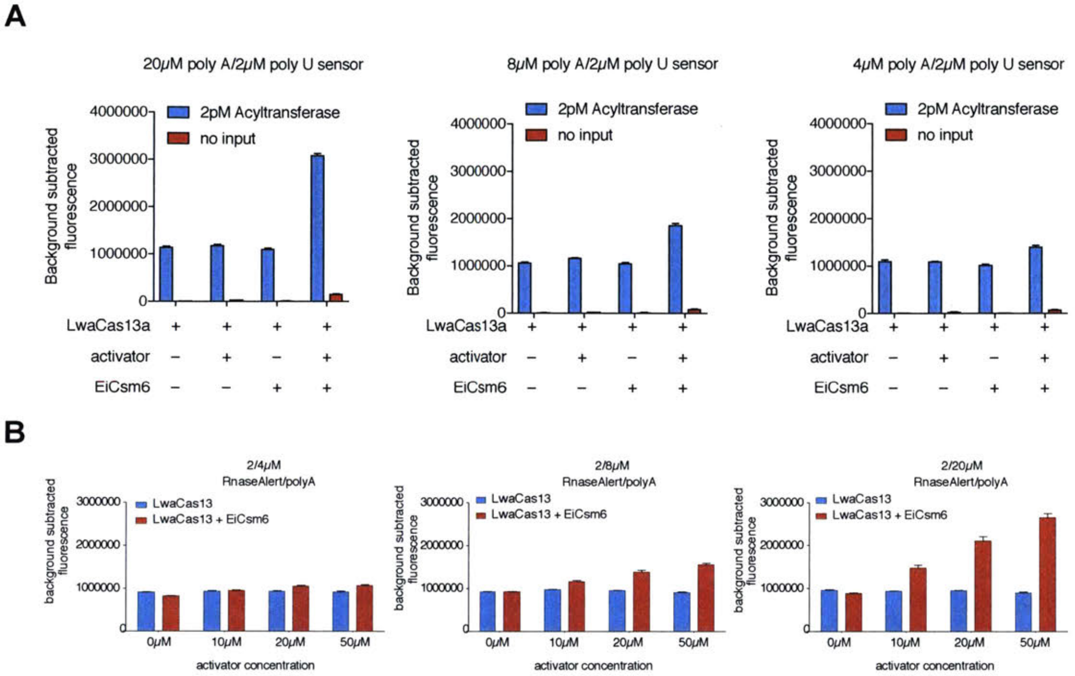


Figure 5.S29: Effect of reporter and activator concentrations on Csm6-enhancement of LwaCas13a activity.

A) EiCsm6 enhancement of LwaCas13a detection at various ratios of poly A and poly U reporters.

B) EiCsm6 enhancement of LwaCas13a detection at various concentrations of $(A)_6$-$(U)_5$ activator.

10.4.30 Figure 5.S30

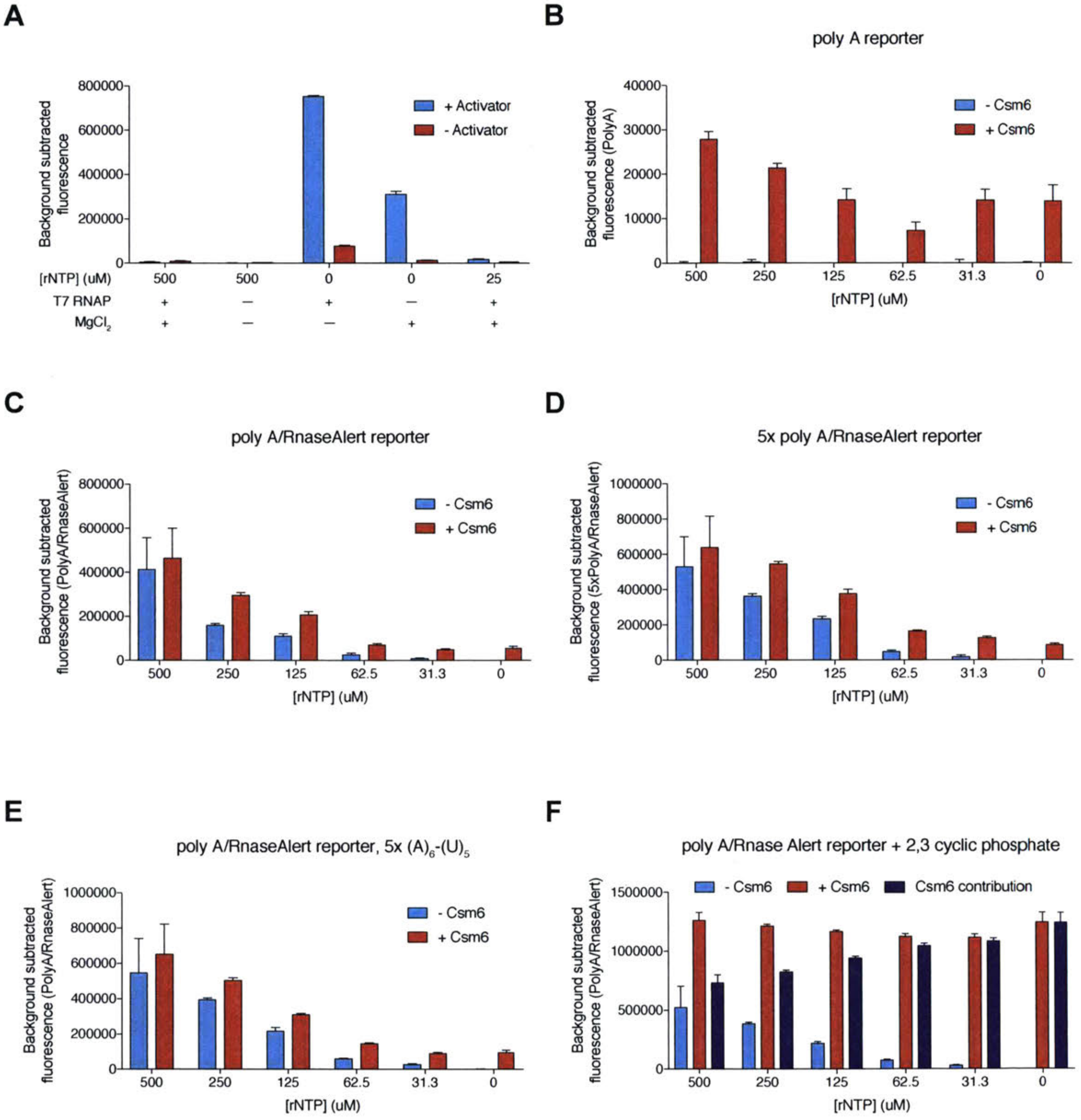


Figure 5.S30: Effect of *in vitro* transcription components on Csm6 activity.

A) EiCsm6 activity in the presence of IVT components, with and without 2,3 cyclic phosphate activator. Components include 3mM additional MgCl2, 1mM rNTP mix, 30U T7 polymerase

B) EiCsm6 and LwaCas13a activity with $(A)_6$-$(U)_5$ activator and poly-A reporter in the presence of various concentrations of ribonucleotides

C) Combined EiCsm6 and LwaCas13a activity with $(A)_6$-$(U)_5$ activator and poly-A/RNaseAlert reporter combination in the presence of various concentrations of ribonucleotides

D) Combined EiCsm6 and LwaCas13a activity with $(A)_6$-$(U)_5$ activator and poly-A/5x RNaseAlert reporter combination in the presence of various concentrations of ribonucleotides

E) Combined EiCsm6 and LwaCas13a activity with 5x$(A)_6$-$(U)_5$ activator and poly-A/RNaseAlert reporter combination in the presence of various concentrations of ribonucleotides

F) Combined EiCsm6 and LwaCas13a activity with cyclic phosphate activator and poly-A/RNaseAlert reporter combination in the presence of various concentrations of ribonucleotides

10.4.31 Figure 5.S31

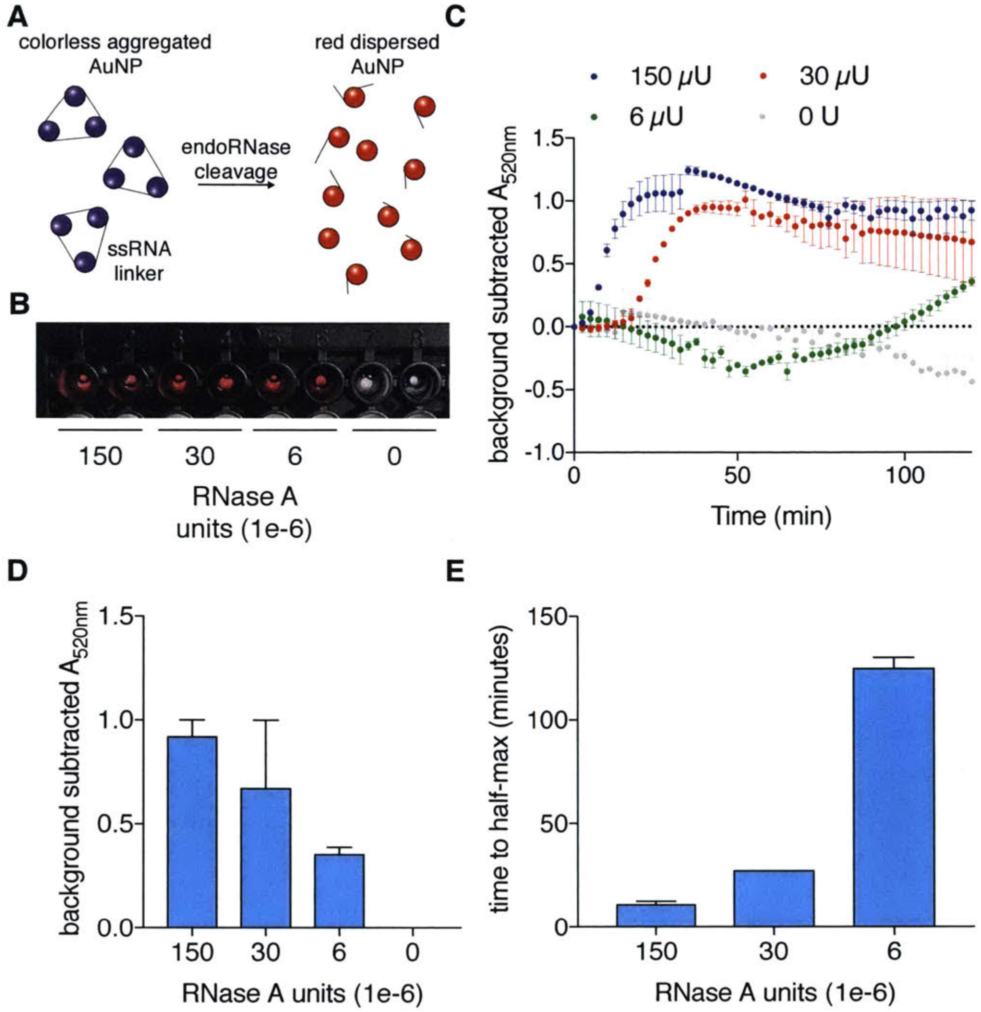


Figure 5.S31: Colorimetric detection of RNase activity with gold nanoparticle aggregation.

A) Schematic of gold-nanoparticle based colorimetric readout for RNase activity. In the absence of RNase activity, RNA linkers aggregate gold nanoparticles, leading to loss of red color. Cleavage of RNA linkers releases nanoparticles and results in a red color change.

B) Image of colorimetric reporters after 120 minutes of RNase digestion at various units of RNase A.

C) Kinetics at 520nm absorbance of AuNP colorimetric reporters with digestion at various unit concentrations of RNase A.

D) The 520nm absorbance of AuNP colorimetric reporters after 120 minutes of digestion at various unit concentrations of RNase A.

E) Time to half-A_{520} maximum of AuNP colorimetric reporters with digestion at various unit concentrations of RNase A.

10.4.32 Figure 5.S32

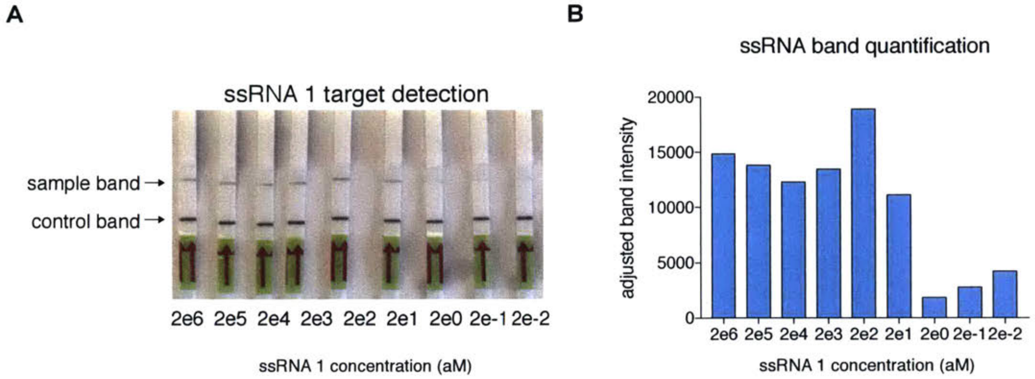


Figure 5.S32: SHERLOCK lateral flow detection of ssRNA 1

A) Detection of ssRNA 1 using lateral flow SHERLOCK at various concentrations.

B) Quantitation of band intensity from detection in (A).

10.4.33 Figure 5.S33

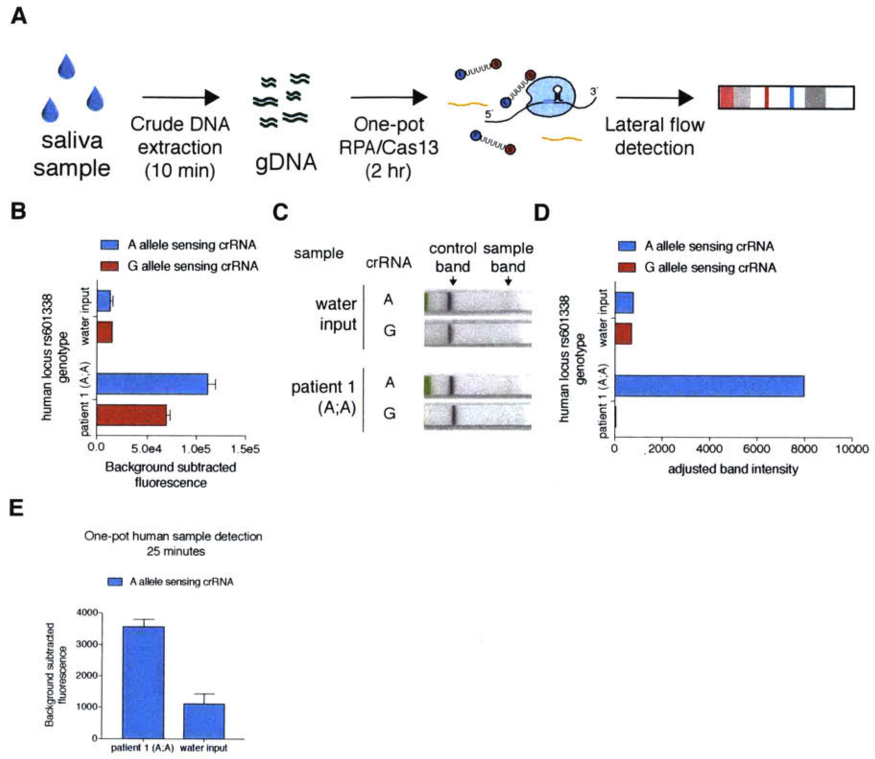


Figure 5.S33: One-pot lateral-flow genotyping of genomic DNA from saliva

A) Schematic for rapid extraction and one-pot detection of genomic DNA from patient saliva.

B) Detection of rs601338 genotypes in from crude genomic DNA extraction compared to water input.

C) Lateral-flow detection of rs601338 genotypes in from crude genomic DNA extraction.

D) Quantitation of band intensity from detection in (C)

E) Detection of patient DNA in 25 minutes from crude saliva.

10.4.34 Figure 5.S34

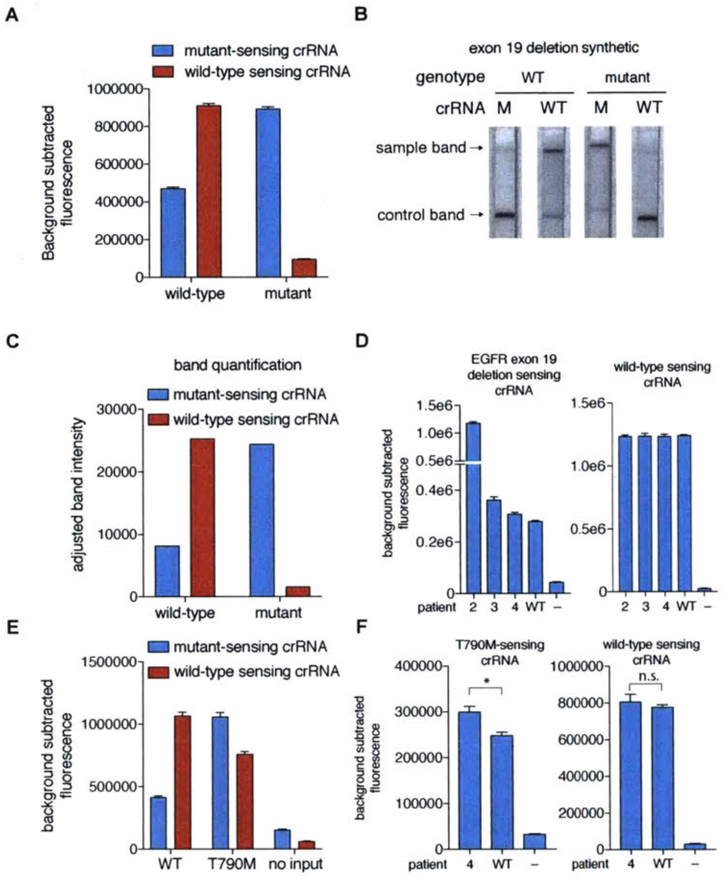


Figure S34: SHERLOCK lateral flow detection of synthetic cfDNA samples

A) Detection of EGFR exon 19 deletion mutation in synthetic DNA samples with either exon 19 deletion or WT genotype using LwaCas13a.

B) Lateral-flow detection of EGFR exon 19 deletion mutation in synthetic DNA samples with either exon 19 deletion or WT genotype using LwaCas13a.

C) Quantitation of band intensity from detection in (B).

D) Detection of EGFR exon 19 deletion mutation in 4 patient cfDNA samples with either exon 19 deletion or WT genotype using LwaCas13a.

E) Detection of EGFR T790M deletion mutation in synthetic DNA samples with either T790M or WT genotype using LwaCas13a.

Detection of EGFR T790M deletion mutation in patient cfDNA samples with either T790M or WT genotype using LwaCas13a. (*, $p < 0.05$; n.s., not significant; bars represent mean ± s.e.m.). In this case, patient 4's T790M allelic fraction, as verified by targeted sequencing, was 0.6%. We were still able to see significant detection of this low allelic fraction due to the sensitivity and specificity of SHERLOCKv2, agreeing with our previous results showing that we could detect greater than 0.1% allelic fraction samples(Gootenberg et al., 2017c). Additionally, because the Bsu polymerase in RPA has a minimum error rate of 10^{-5} errors per base incorporated per cycle(Chen, 2014), we can expect about 0.02% of amplicons to contain an error at the mutation we are trying to sense. Because spurious signal will only be detected if the correct mutation is formed on a wild-type amplicon, then only 0.0067% of amplicons will have a mutation that causes spurious detection of the mutation. As most patients do not have below 0.01% allelic fraction of cfDNA mutations, this error rate is acceptable.

10.4.35 Figure 5.S35

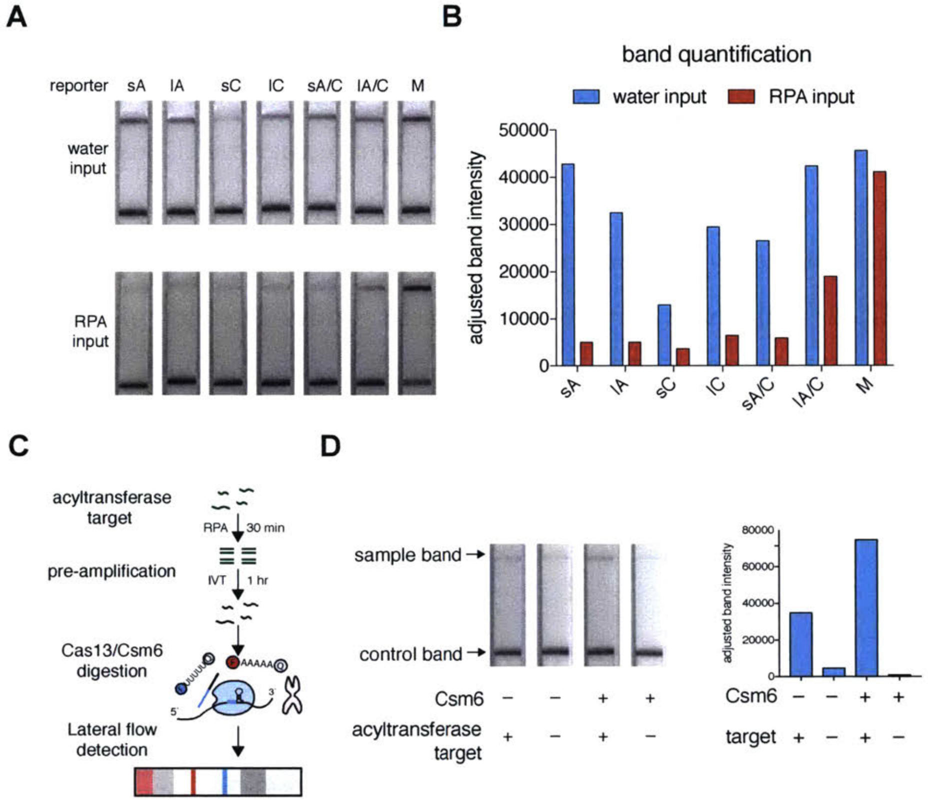


Figure 5.S35: Lateral flow Csm6-enhanced SHERLOCK with different reporter combinations

A) Lateral-flow detection of Csm6-enhanced SHERLOCK with various reporter designs. sA: short poly-A sensor; lA: long poly A sensor; sC: short poly C sensor; lC: long poly C sensor; sA/C: short poly-A/C sensor; lA/C: long poly-A/C sensor; M: mixed RNase alert-like sensor.

B) Quantitation of band intensity from detection in (A)

C) Schematic of lateral flow readout of EiCsm6-enhanced LwaCas13a SHERLOCK detection of acyltransferase ssDNA with separate RPA and IVT steps

D) EiCsm6-enhanced lateral flow SHERLOCK of *P. aeruoginosa* acyltransferase gene in combination with LwaCas13a. Band intensity quantitation is shown to the right.

10.4.36 Figure 5.S36

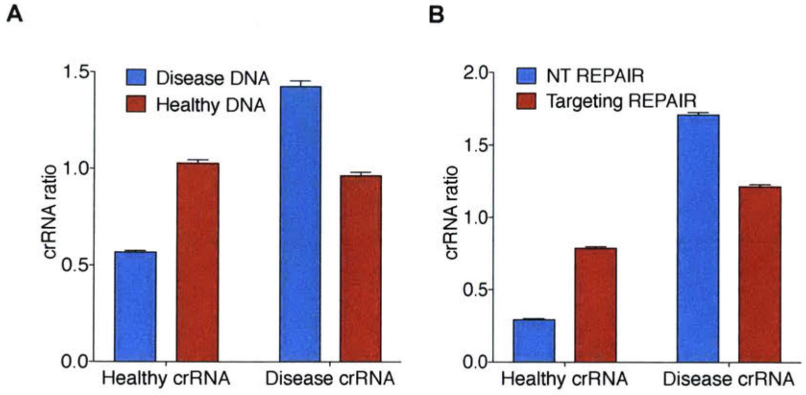


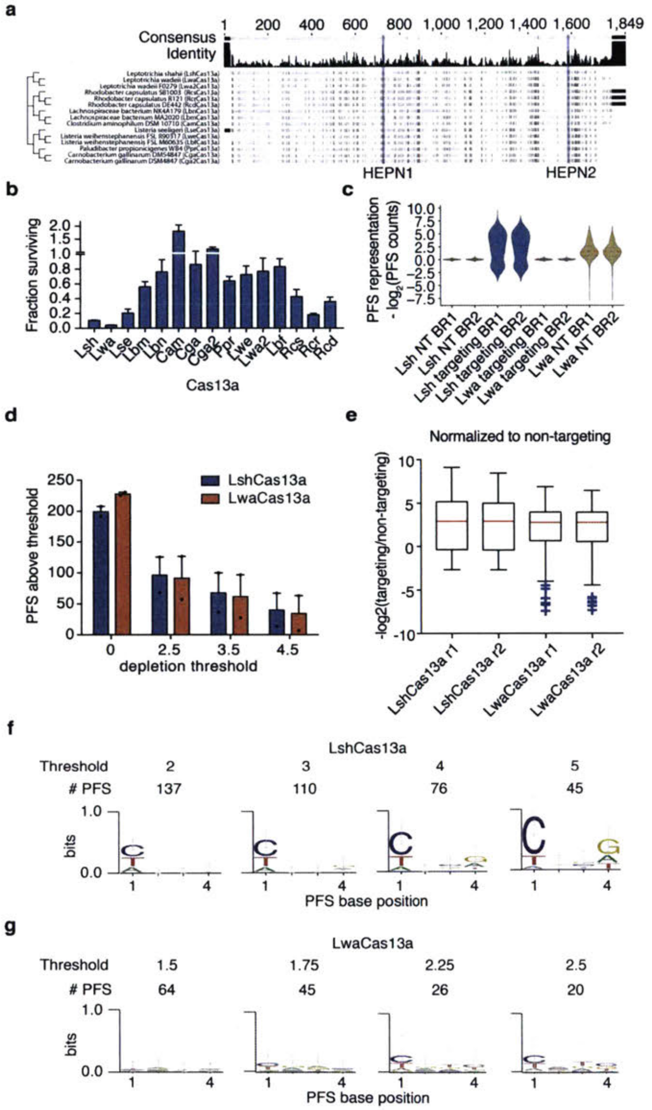
Figure 5.S36: Non-multiplexed theranostic detection of mutations and REPAIR editing

A) Detection of *APC* alleles from healthy- and disease-simulated samples with LwaCas13a.

B) Detection with LwaCas13a of editing correction at the *APC* alleles from REPAIR targeting and non-targeting samples.

10.5 Chapter 6 Supplementary Figures

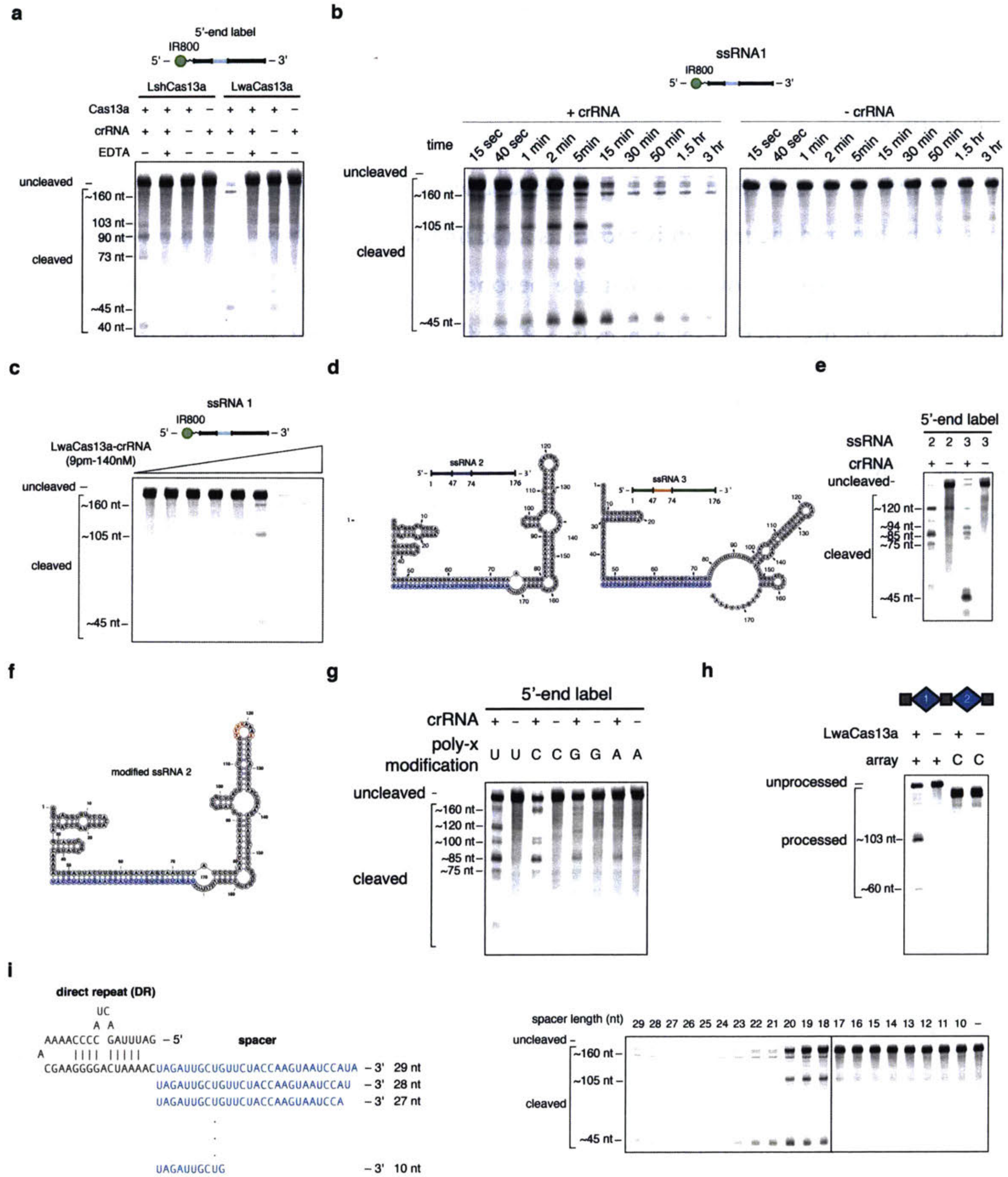
10.5.1 Figure 6.1



Extended Data Figure 6.1 | Evaluation of LwaCas13a PFS preferences and comparisons to LshCas13a.

a, Sequence comparison tree of the fifteen Cas13a orthologs evaluated in this study. **b,** Ratios of *in vivo* activity from Fig. 1B. **c,** Distributions of PFS enrichment for LshCas13a and LwaCas13a in targeting and non-targeting samples. The 25th and 75th percentiles are shown as grey dotted lines and the median is shown as a red dotted line. The minimum and maximum are marked by the ends of the distribution. Each distribution represents 976 PFS sequences (n = 976). **d,** Number of LshCas13a and LwaCas13a PFS sequences above depletion threshold for varying depletion thresholds. Values are mean ± SEM with n = 2. **e,** Distributions of PFS enrichment for LshCas13a and LwaCas13a in targeting samples, normalized to non-targeting samples. The 25th and 75th percentiles are marked by the ends of the box and the median is shown as a red line within the box. Whiskers denote 1.5 times the interquartile range. +, outliers that are beyond the 1.5 times the interquartile range. Each distribution represents 976 PFS sequences (n = 976). **f,** Sequence logos and counts for remaining PFS sequences after LshCas13a cleavage at varying enrichment cutoff thresholds. **g,** Sequence logos and counts for remaining PFS sequences after LwaCas13a cleavage at varying enrichment cutoff thresholds.

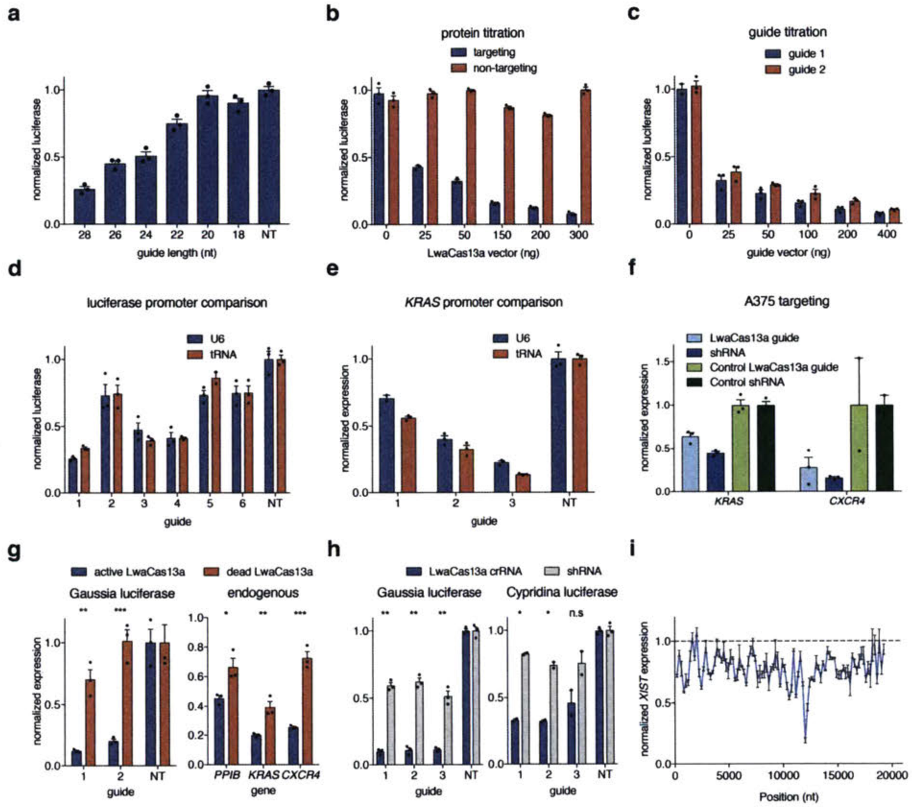
10.5.2 Figure 6.2



Extended Data Fig. 6.2 | Biochemical characterization of LwaCas13a RNA cleavage activity. **a,** Gel electrophoresis comparison of LwaCas13a and LshCas13a RNase activity on ssRNA 1. **b,** Gel electrophoresis of ssRNA1 after incubation with LwaCas13a with or without crRNA 1 for varying amounts of times. **c,** Gel electrophoresis of ssRNA 1 after incubation with varying amounts of

LwaCas13a-crRNA complex. **d,** Sequence and structure of ssRNA 4 and ssRNA 5. crRNA spacer sequence is highlighted in blue. **e,** Gel electrophoresis of ssRNA 4 and ssRNA 5 after incubation with LwaCas13a and crRNA 1. **f,** Sequence and structure of ssRNA 4 with sites of poly-x modifications highlighted in red. crRNA spacer sequence is highlighted in blue. **g,** Gel electrophoresis of ssRNA 4 with each of 4 possible poly-x modifications incubated with LwaCas13a and crRNA 1. **h,** Gel electrophoresis of pre-crRNA from the *L. wadei* CRISPR-Cas locus showing LwaCas13a processing activity. **i,** Cleavage efficiency of ssRNA 1 for crRNA spacer truncations after incubation with LwaCas13a.

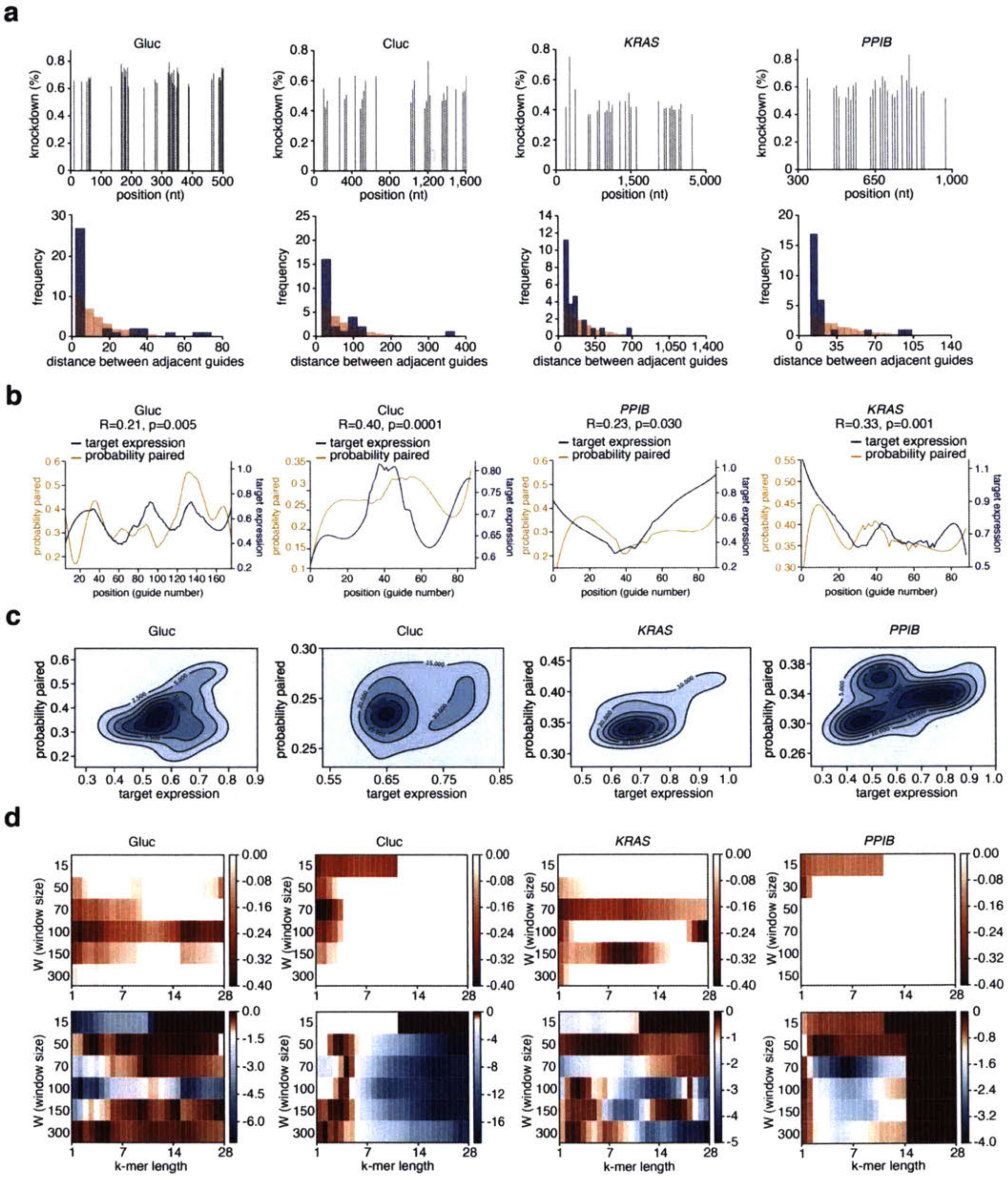
10.5.3 Figure 6.3



Extended Data Fig. 6.3 | Engineering and optimization of LwaCas13a for mammalian knockdown.

a, Knockdown of Gluc transcript by LwaCas13a and Gluc guide 1 spacers of varying length. **b,** Knockdown of Gluc transcript with Gluc guide 1 and varying amounts of transfected LwaCas13a plasmid. **c**, Knockdown of Gluc transcript by LwaCas13a and varying amounts of transfected Gluc guide 1 and 2 plasmid (n = 2 or 3). **d**, Knockdown of Gluc transcript using guides expressed from either U6 or tRNA^{Val} promoters (n = 2 or 3). **e,** Knockdown of *KRAS* transcript using guides expressed from either U6 or $tRNA^{Val}$ promoters (n = 2 or 3). **f,** Knockdown of *KRAS* and *CXC4* transcripts by LwaCas13a using guides transfected in A375 cells with position-matched shRNA comparisons (n = 2 or 3). **g,** Knockdown of Gluc transcript and endogenous transcripts *PPIB, KRAS,* and *CXCR4* with active and catalytically inactive LwaCas13a. **h,** Validation of the top three guides from the arrayed knockdown Gluc and Cluc screens with shRNA comparisons (n = 2 or 3). **i,** Arrayed knockdown screen of 93 guides evenly tiled across the *XIST* transcript. All values are mean ± SEM with n = 3, unless otherwise noted (n represents the number of transfection replicates). $^{**}p < 0.01$; $^{*}p < 0.05$. ns = not significant. A two-tailed student's T-test was used for comparisons.

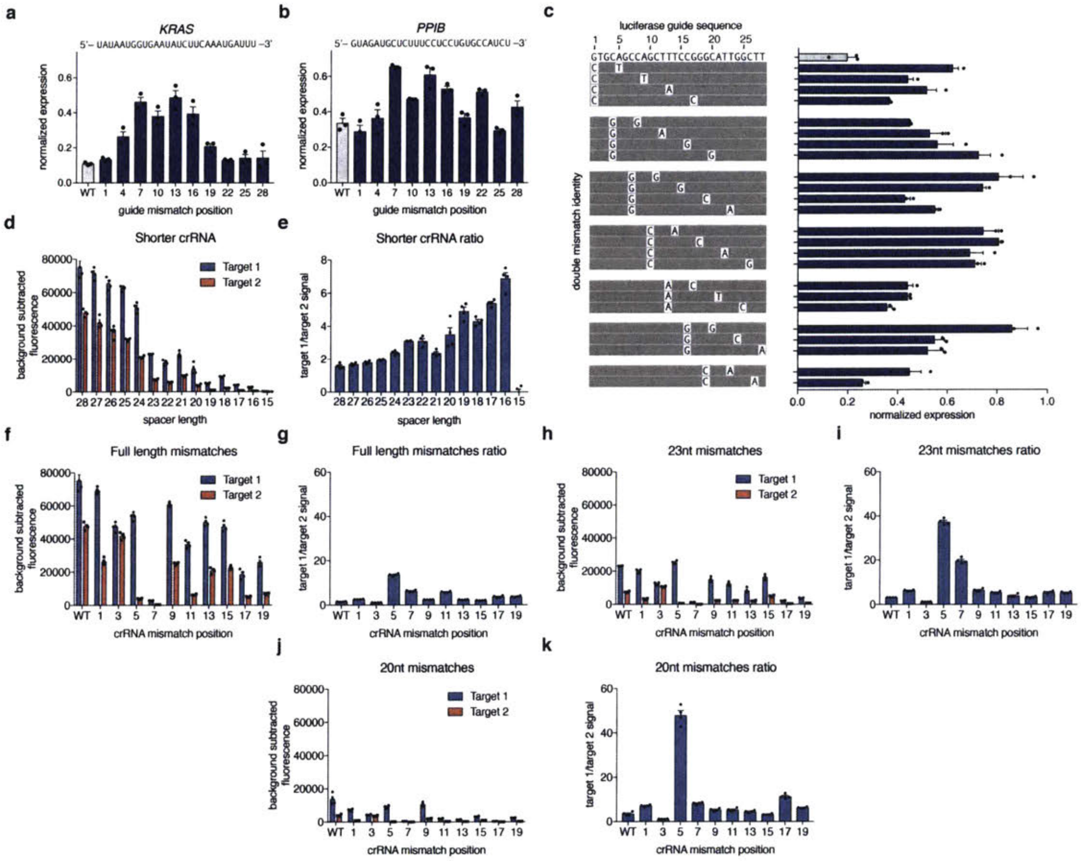
10.5.4 Figure 6.4



Extended Data Fig. 6.4 | LwaCas13a targeting efficiency is influenced by accessibility along the transcript.

a, *First row:* Top knockdown guides are plotted by position along target transcript. Top knockdown guides are defined as top 20% of guides for Gluc and top 30% of guides for Cluc, *KRAS,* and *PPIB. Second row*: Histograms for the pairwise distance between adjacent top guides for each transcript (blue) compared to a random null-distribution (red). Inset shows the cumulative frequency curves for these histograms. A shift of the blue curve (actual measured distances) to the left of the red curve (null distribution of distances) indicates that guides are closer together than expected by chance. **b,** Gluc, Cluc, *PPIB,* and *KRAS* knockdown partially correlates with target accessibility as measured by predicted folding of the transcript. The correlation was computed using a Pearson's correlation coefficient and two-tailed significance test. **c,** Kernel density estimation plots depicting the correlation between target accessibility (probability of a region being base-paired) and target expression after knockdown by LwaCas13a. **d,** *First row:* Correlations between target expression and target accessibility (probability of a region being base-paired) measured at different window sizes (W) and for different k-mer lengths. *Second row:* P-values for the correlations between target expression and target accessibility (probability of a region being base-paired) measured at different window sizes (W) and for different k-mer lengths. The color scale is designed such that p-values > 0.05 are shades of red and p-values < 0.05 are shades of blue.

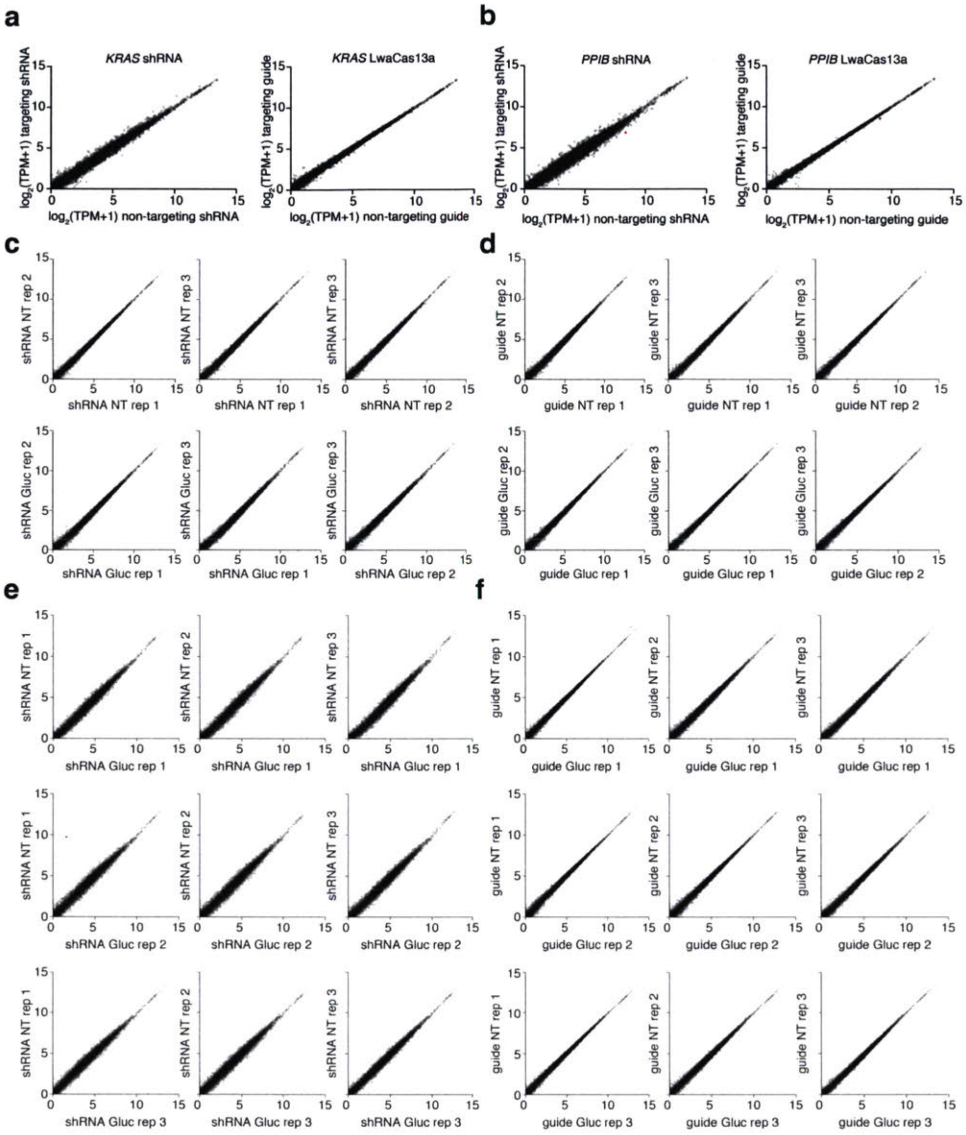
10.5.5 Figure 6.5



Extended Data Fig. 6.5 | Detailed evaluation of LwaCas13a sensitivity to mismatches in the guide:target duplex at varying spacer lengths.

a, Knockdown of *KRAS* evaluated with guides containing single mismatches at varying positions across the spacer sequence (n = 2 or 3). **b,** Knockdown of *PPIB* evaluated with guides containing single mismatches at varying positions across the spacer sequence (n = 2 or 3). **c,** Knockdown of Gluc evaluated with guides containing non-consecutive double mismatches at varying positions across the spacer sequence. The wild-type sequence is shown at the top with mismatch identities shown below. **d,** Collateral cleavage activity on ssRNA 1 and 2 for varying spacer lengths. **e,** Specificity ratios of guide tested in (**d**). Specificity ratios are calculated as the ratio of the on-target RNA (ssRNA 1) collateral cleavage to the off-target RNA (ssRNA 2) collateral cleavage. **f,** Collateral cleavage activity on ssRNA 1 and 2 for 28-nt spacer crRNA with synthetic mismatches tiled along the spacer. **g,** Specificity ratios, as defined in (**e**), of crRNA tested in (**f**). **h,** Collateral cleavage activity on ssRNA 1 and 2 for 23-nt spacer crRNA with synthetic mismatches tiled along the spacer. **i,** Specificity ratios, as defined in (**e**), of crRNA tested in (**h**). **j,** Collateral cleavage activity on ssRNA 1 and 2 for 20-nt spacer crRNA with synthetic mismatches tiled along the spacer. **k,** Specificity ratios, as defined in (**e**), of crRNA tested in (**j**). For (**a-c**), all values are mean ± SEM with n = 3, unless otherwise noted (n represents the number of transfection replicates). For (**d-k**), all values are mean ± SEM with n = 4 (n represents the number of technical replicates).

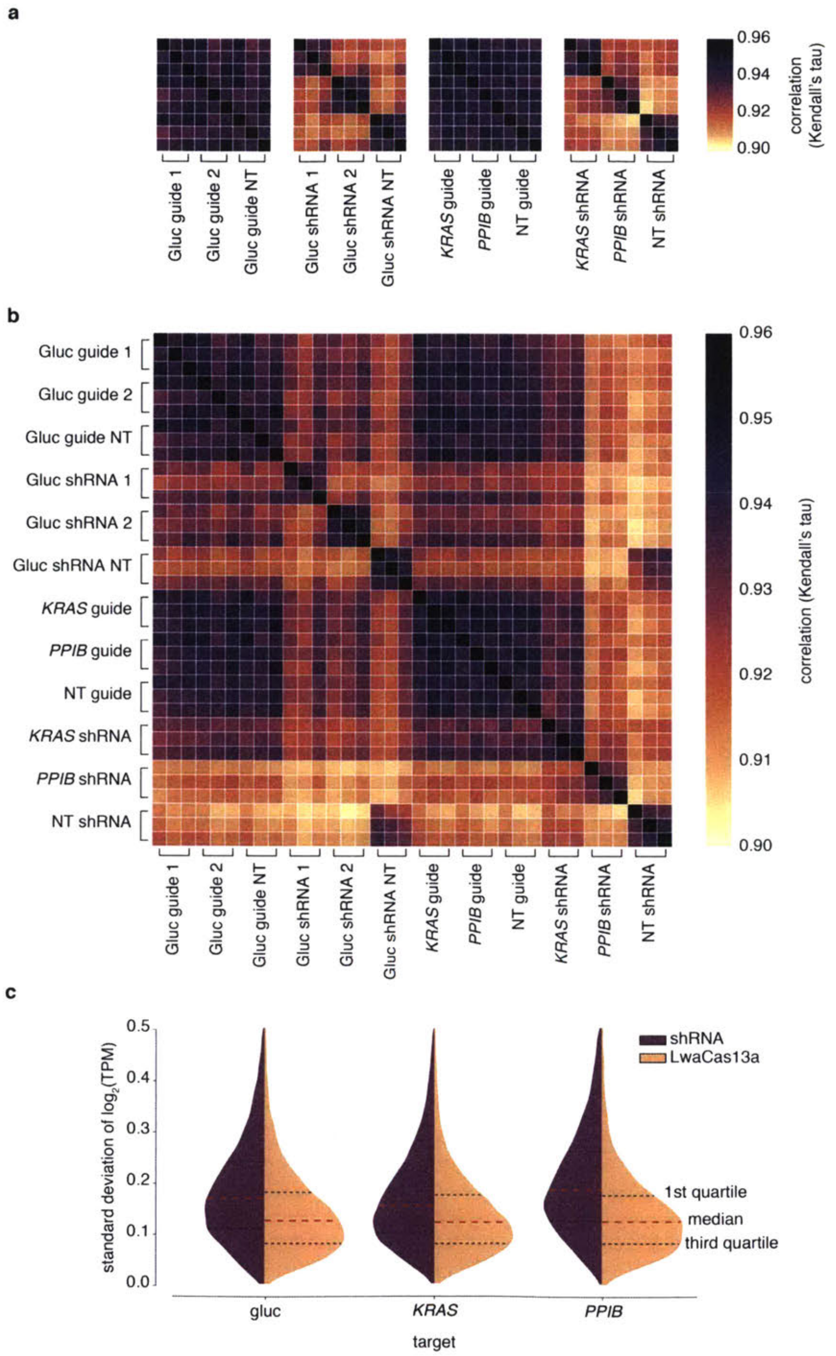
10.5.6 Figure 6.6



Extended Data Fig. 6.6 | LwaCas13a is more specific than shRNA knockdown for endogenous targets.

a, *Left:* Expression levels in log2(transcripts per million (TPM)) values of all genes detected in RNA-seq libraries of non-targeting shRNA-transfected control (x-axis) compared to *KRAS*-targeting shRNA (y-axis). Shown is the mean of three biological replicates. The *KRAS* transcript data point is colored in red. *Right:* Expression levels in log2(transcripts per million (TPM)) values of all genes detected in RNA-seq libraries of non-targeting LwaCas13a-guide-transfected control (x-axis) compared to *KRAS*-targeting LwaCas13a-guide (y-axis). Shown is the mean of three biological replicates. The *KRAS* transcript data point is colored in red. **b,** *Left:* Expression levels in $\log_2$(transcripts per million (TPM)) values of all genes detected in RNA-seq libraries of non-targeting shRNA-transfected control (x-axis) compared to *PPIB*-targeting shRNA (y-axis). Shown is the mean of three biological replicates. The *PPIB* transcript data point is colored in red. *Right:* Expression levels in $\log_2$(transcripts per million (TPM)) values of all genes detected in RNA-seq libraries of non-targeting LwaCas13a-guide-transfected control (x-axis) compared to *PPIB*-targeting LwaCas13a-guide (y-axis). Shown is the mean of three biological replicates. The *PPIB* transcript data point is colored in red. **c,** Comparisons of individual replicates of non-targeting shRNA conditions (first row) and Gluc-targeting shRNA conditions (second row). **d,** Comparisons of individual replicates of non-targeting guide conditions (first row) and Gluc-targeting guide conditions (second row). **e,** Pairwise comparisons of individual replicates of non-targeting shRNA conditions against the Gluc-targeting shRNA conditions. **f,** Pairwise comparisons of individual replicates of non-targeting guide conditions against the Gluc-targeting guide conditions.

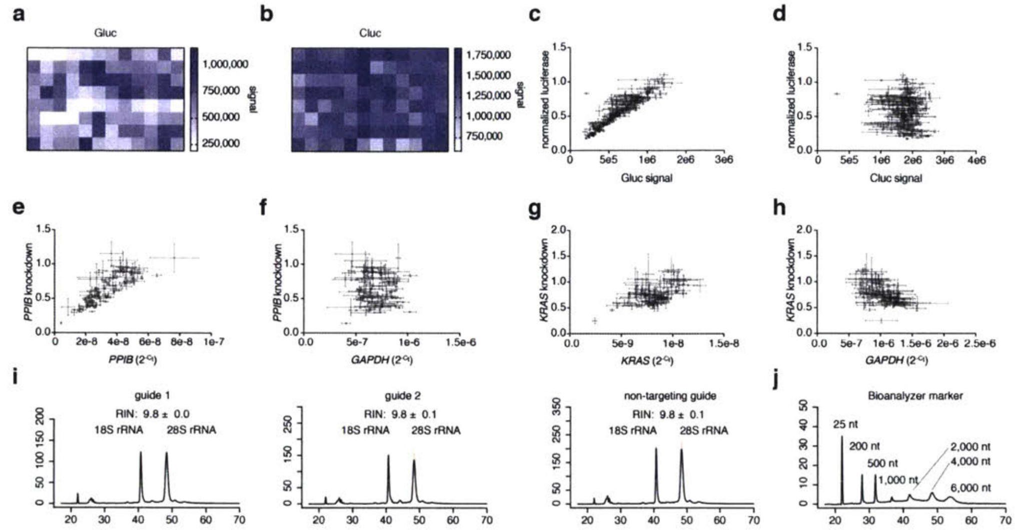
10.5.7 Figure 6.7



Extended Data Fig. 6.7 | Detailed analysis of LwaCas13a and RNAi knockdown variability (standard deviation) across all samples.

a, Heatmap of correlations (Kendall's tau) for $\log_2$(transcripts per million (TPM+1)) values of all genes detected in RNA-seq libraries between targeting and non-targeting replicates for shRNA or guide targeting either luciferase reporters or endogenous genes. **b,** Heatmap of correlations (Kendall's tau) for $\log_2$(transcripts per million (TPM+1)) values of all genes detected in RNA-seq libraries between all replicates and perturbations. **c,** Distributions of standard deviations for $\log_2$(transcripts per million (TPM+1)) values of all genes detected in RNA-seq libraries among targeting and non-targeting replicates for each gene targeted by either shRNA or guide.

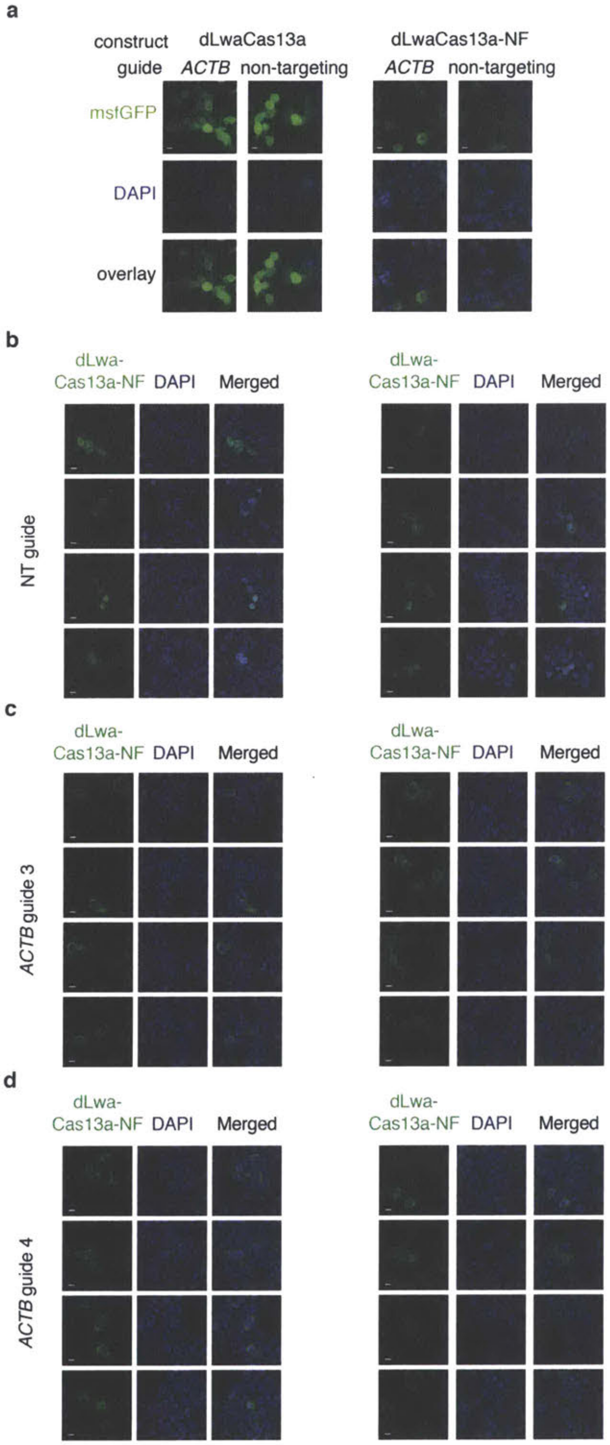
10.5.8 Figure 6.8



Extended Data Fig. 6.8 | LwaCas13a knockdown is specific to the targeted transcript with no activity on a measured off-target transcript.

a, Heatmap of absolute Gluc signal for first 96 spacers tiling Gluc. **b,** Heatmap of absolute Cluc signal for first 96 spacers tiling Gluc. **c,** Relationship between absolute Gluc signal and normalized luciferase for Gluc tiling guides. **d,** Relationship between absolute Cluc signal and normalized luciferase for Gluc tiling guides. **e,** Relationship between *PPIB* 2^{-Ct} levels and *PPIB* knockdown for *PPIB* tiling guides. **f,** Relationship between *GAPDH* 2^{-Ct} levels and *PPIB* knockdown for *PPIB* tiling guides. **g,** Relationship between *KRAS* 2^{-Ct} levels and *KRAS* knockdown for *KRAS* guides. **h,** Relationship between *GAPDH* 2^{-Ct} levels and *KRAS* knockdown for *KRAS* guides. **i,** Bioanalyzer traces of total RNA isolated from cells transfected with Gluc-targeting guides 1 and 2 or non-targeting guide from the experiment with active LwaCas13a in Extended Data Fig. 3g. The RNA-integrity number (RIN) is shown and 18S rRNA and 28S rRNA peaks are labeled above. A student's t-test shows no significant difference for the RIN between either of the targeting conditions and the non-targeting condition. The curves are shown as a mean of three replicates and the shaded areas in light red around the curves show the s.e.m. **j,** The Bioanalyzer trace for the RNA ladder with peak sizes labeled above.

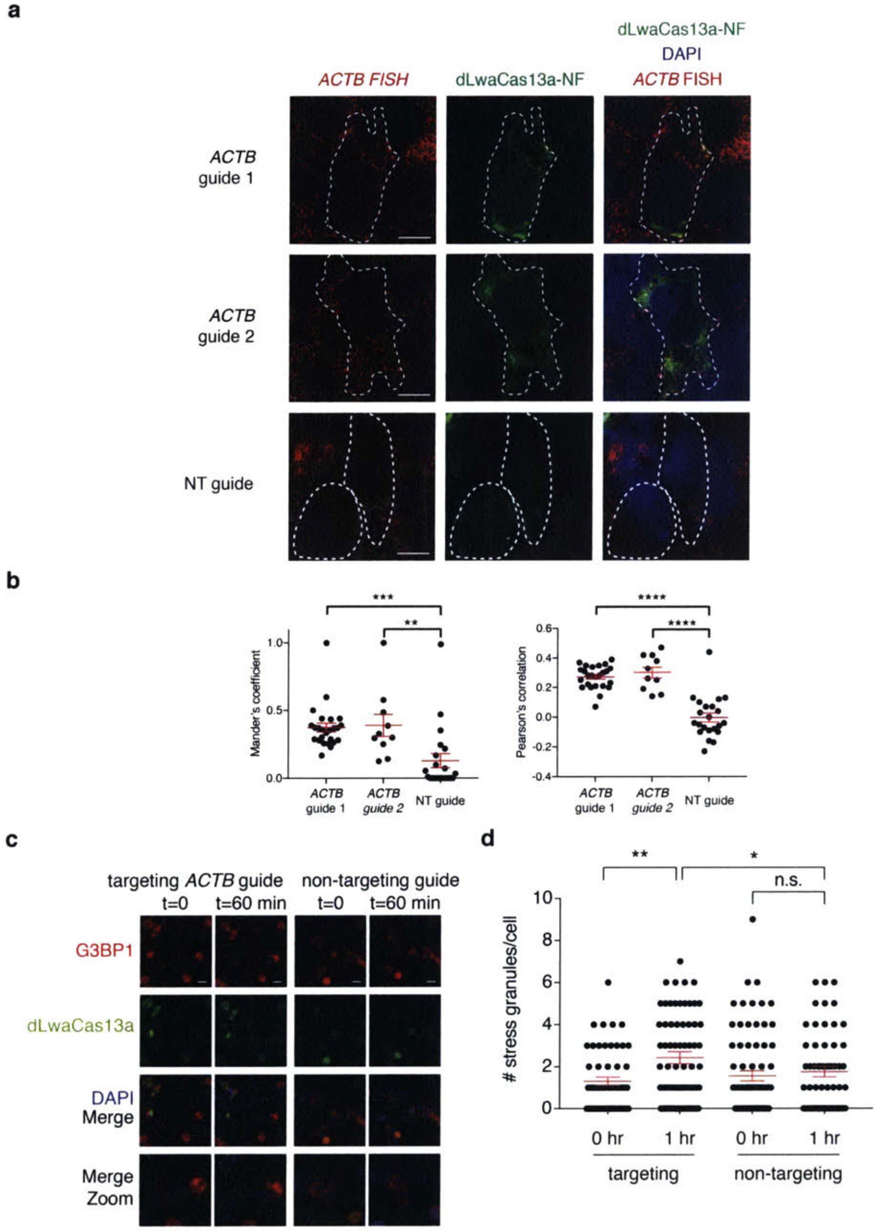
10.5.9 Figure 6.9



Extended Data Fig. 6.9 | dLwaCas13a-NF can be used for ACTB imaging.

a, Comparison between localization of dLwaCas13-GFP and dLwaCas13a-GFP-KRAB (dLwaCas13a-NF) constructs for imaging *ACTB*. Scale bars, 10µm **b**, Additional fields of view of dLwaCas13a-NF delivered with a non-targeting guide. Scale bars, 10µm. **c,** Additional fields of view of dLwaCas13a-NF delivered with *ACTB* guide 3. Scale bars, 10µm. **d,** Additional fields of view of dLwaCas13a-NF delivered with *ACTB* guide 4. Scale bars, 10µm.

10.5.10 Figure 6.10



Extended Data Fig. 10 | dLwaCas13a-NF can image stress granule formation in living cells. a, Representative images from RNA FISH of the *ACTB* transcript in dLwaCas13a-NF-expressing cells with corresponding *ACTB*-targeting and non-targeting guides. Cell outline is shown with a dashed line. Scale bars, 10μm **b,** Overall signal overlap between *ACTB* RNA FISH signal and dLwaCas13a-NF quantified by the Mander's overlap coefficient (*left*) and Pearson's correlation (*right*). Correlations and signal overlap are calculated pixel-by-pixel on a per cell basis. n = 10-25 cells per condition. ****p < 0.0001; ***p < 0.001; **p < 0.01. A two-tailed student's T-test was used for comparisons. **c,** Representative images from live-cell analysis of stress granule formation in response to 400 uM sodium arsenite treatment. Scale bars, 20μm **d,** Quantitation of stress granule formation in response to sodium arsenite treatment. Quantitation is based on overlapping dLwaCas13a-NF and *G3BP1* puncta. n = 54-72 cells per condition. All values are mean ± SEM. ****p < 0.0001; ***p < 0.001; **p < 0.01; *p < 0.05. ns = not significant. A two-tailed student's T-test was used for comparisons.

10.6 Chapter 7 Supplementary Figures

10.6.1 Figure 7.S1

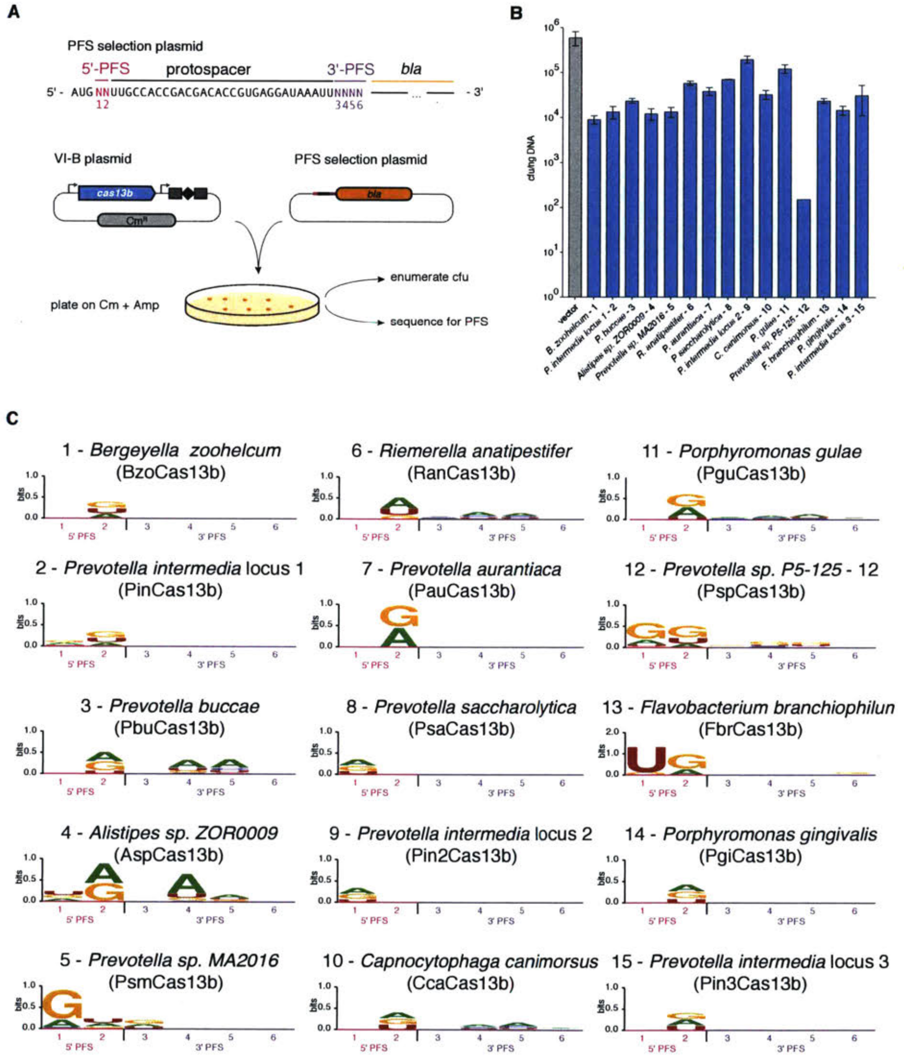


Figure 7.S1: Bacterial screening of Cas13b orthologs for *in vivo* efficiency and PFS determination.

A) Schematic of bacterial assay for determining the PFS of Cas13b orthologs. Cas13b orthologs with beta-lactamase targeting spacers are co-transformed with beta-lactamase expression plasmids containing randomized PFS sequences and subjected to dual antibiotic selection. PFS sequences that are depleted during co-transformation with Cas13b suggest targeting activity and are used to infer PFS preferences.

B) Quantification of interference activity of Cas13b orthologs targeting beta-lactamase as measured by colony forming units (cfu). Values represent mean +/– S.D.

C) PFS weblogos for Cas13b orthologs as determined by depleted sequences from the bacterial assay. PFS preferences are derived from sequences depleted in the Cas13b condition relative to empty vector controls.

10.6.2 Figure 7.S2

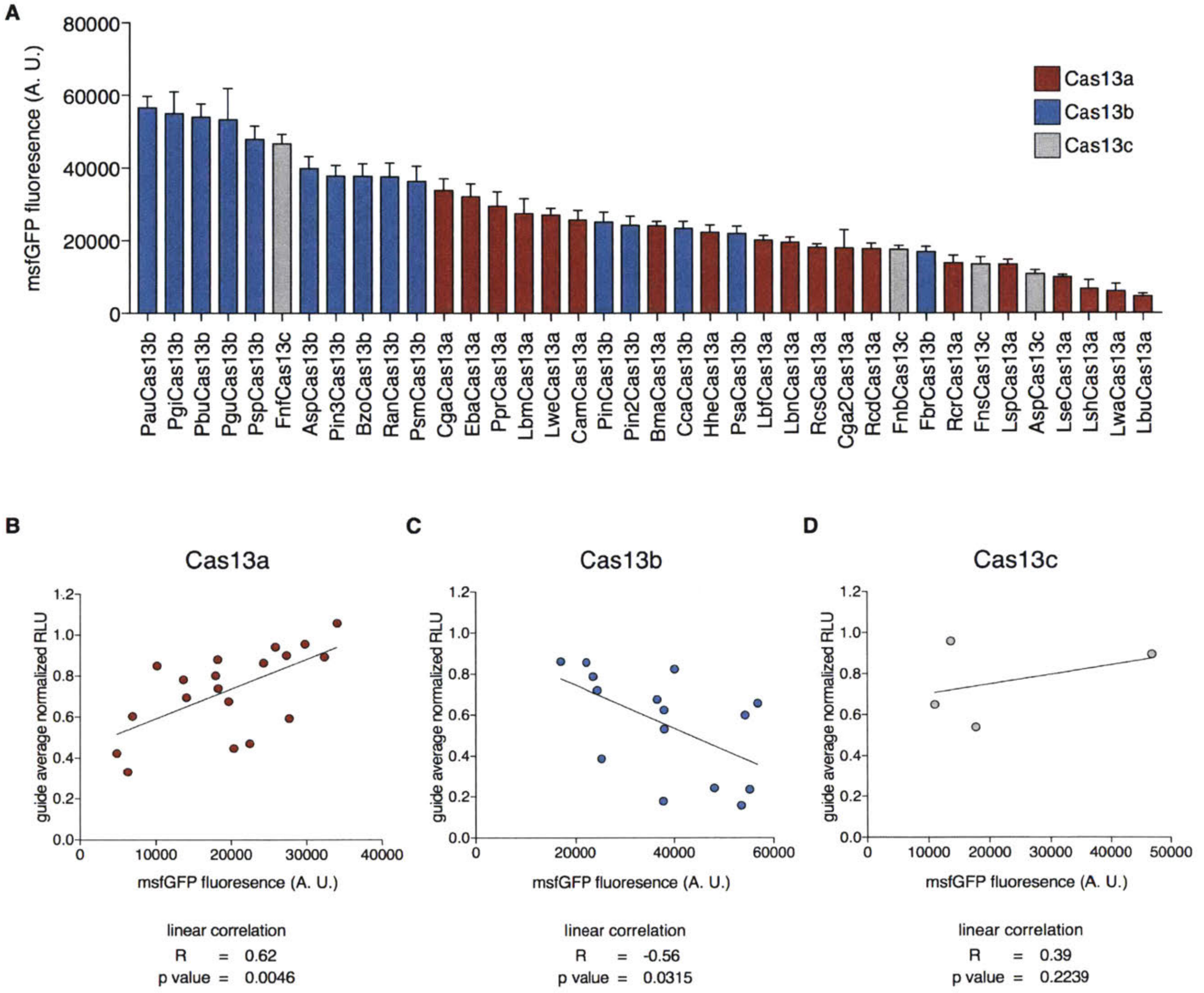


Figure 7.S2: Relative expression of Cas13 orthologs in mammalian cells and correlation of expression with interference activity.

A) Expression of Cas13 orthologs as measured by msfGFP fluoresence. Cas13 orthologs C-terminally tagged with msfGFP were transfected into HEK293FT cells and their fluorescence measured 48 hours post transfection.

B) Correlation of Cas13 expression to interference activity. The average RLU of two *Gluc* targeting guides for Cas13 orthologs, separated by subfamily, is plotted versus expression as determined by msfGFP fluoresence. The RLU for targeting guides are normalized to RLU for a non-targeting guide, whose value is set to 1. The non-targeting guide is the same as in Figure 1B for Cas13b.

10.6.3 Figure 7.S3

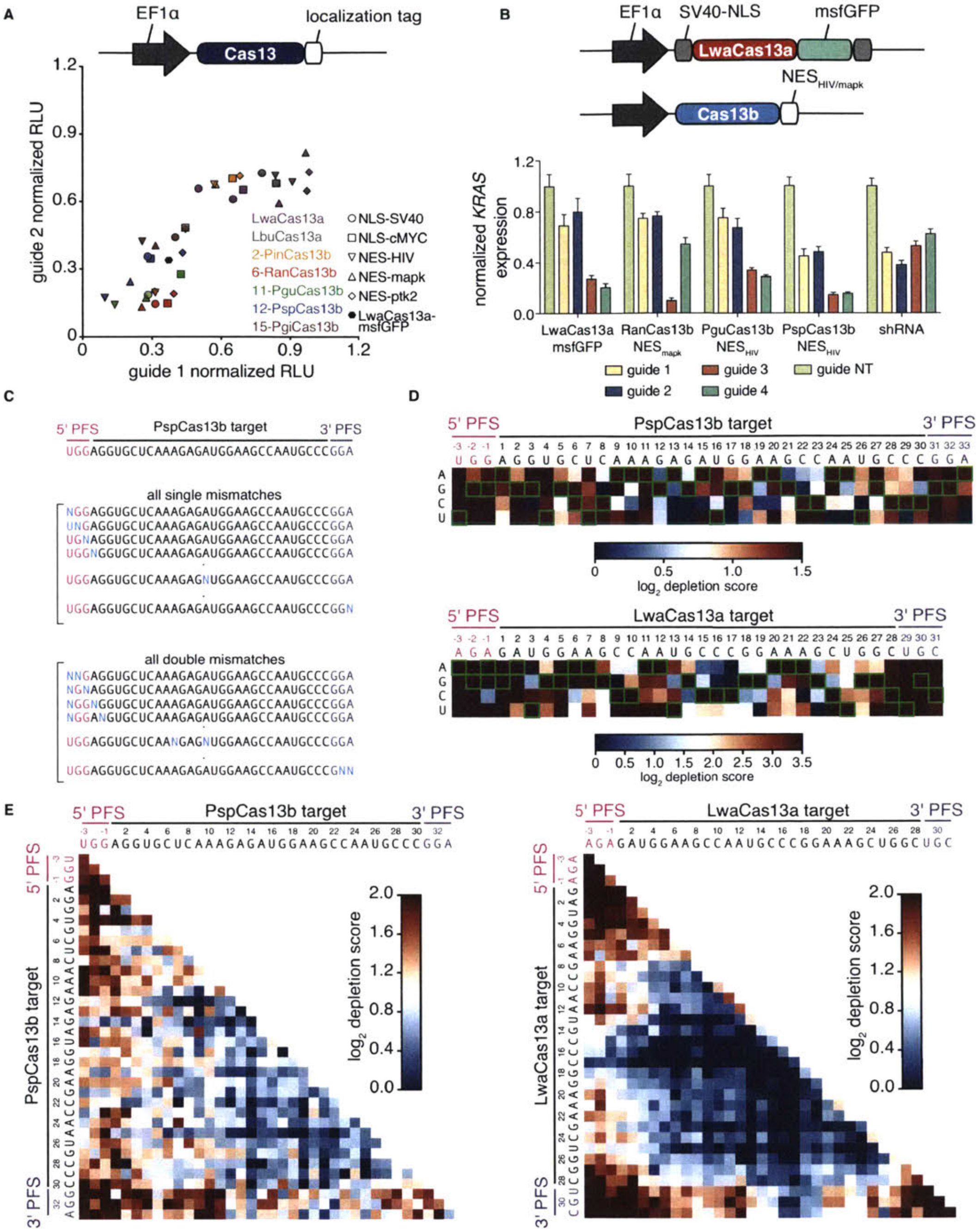


Figure 7.S3: Optimization of Cas13b knockdown and further characterization of mismatch specificity.

A) Gluc knockdown with two different guides is measured using the top two Cas13a and top four Cas13b orthologs fused to a variety of C-terminal nuclear localization and nuclear export tags.

B) Knockdown of *KRAS* is measured for LwaCas13a, RanCas13b, PguCas13b, PspCas13b and shRNA with four position-matched guides. Non-targeting guide is the same as in Figure 1B.

C) Schematic of the single and double mismatch plasmid libraries used for evaluating the specificity of LwaCas13a and PspCas13b knockdown. Every possible single and double mismatch is present in the target sequence as well as in three positions directly flanking the 5' and 3' ends of the target site.

D) The depletion levels of transcripts with the indicated single mismatches are plotted as a heatmap for both the LwaCas13a and PspCas13b conditions. The wildtype base is outlined by a green box.

The depletion levels of transcripts with the indicated double mismatches are plotted as a heatmap for both the LwaCas13a and PspCas13b conditions. Each box represents the average of all possible double mismatches for the indicated position.

10.6.4 Figure 7.S4

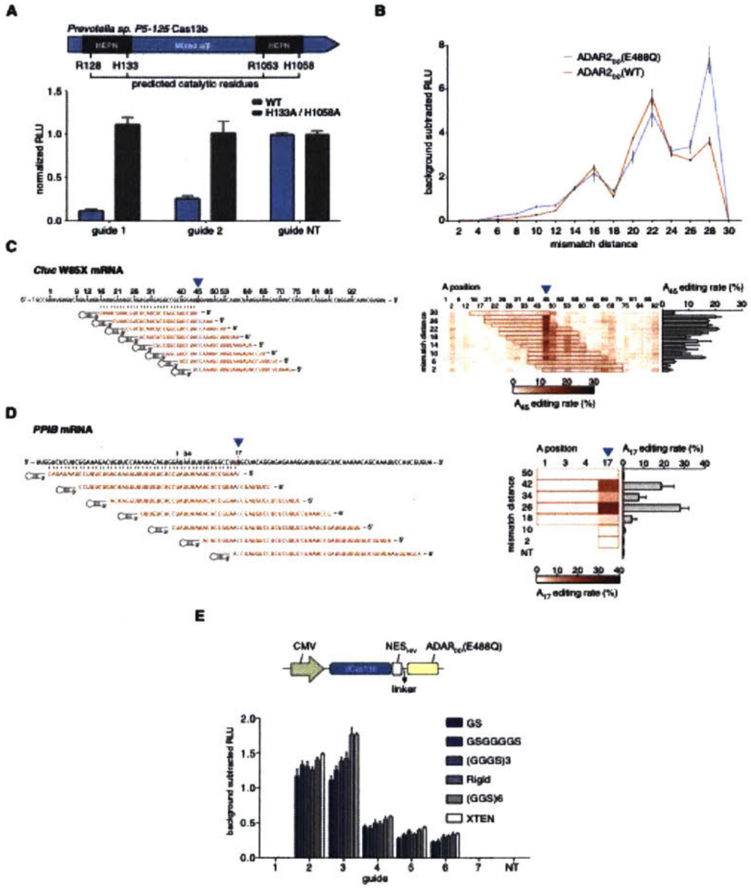


Figure 7.S4: Characterization of design parameters for REPAIRv1.

A) Knockdown efficiency of Gluc with wild-type Cas13b or catalytically inactive H133A/H1058A Cas13b (dCas13b).

B) Quantification of luciferase activity restoration by dCas13b fused to either the wild-type ADAR2 deaminase domain (ADAR2DD) or the hyperactive E488Q mutant ADAR2DD(E488Q) deaminase domain, tested with tiling *Cluc* targeting guides.

C) Guide design and sequencing quantification of A to I editing for 30-nt guides targeting *Cluc* W85X.

D) Guide design and sequencing quantification of A to I editing for 50-nt guides targeting *PPIB.*

E) Influence of linker choice on luciferase activity restoration by REPAIRv1. Values represent mean +/– S.E.M.

10.6.5 Figure 7.S5

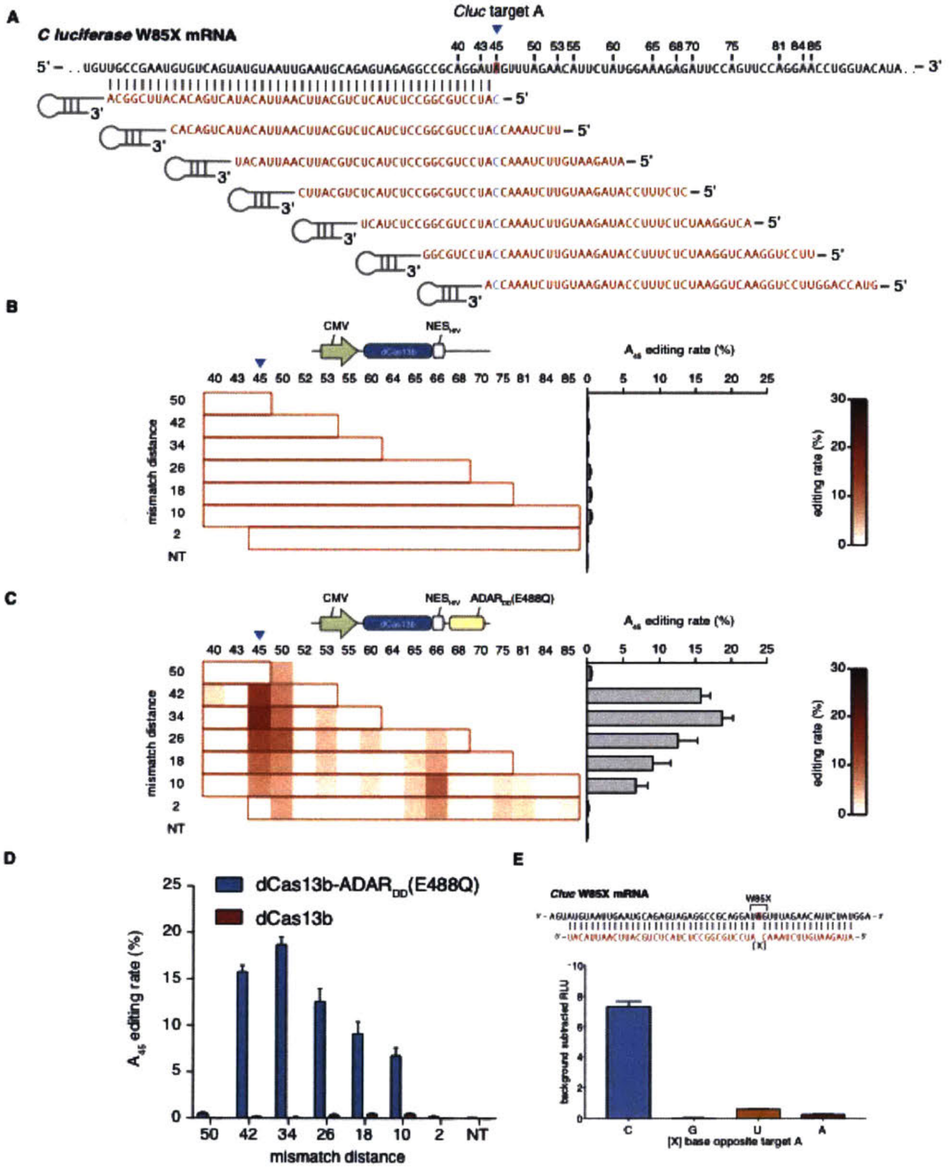


Figure 7.S5: Comparison of RNA editing activity of dCas13b and REPAIRv1.

A) Schematic of guides used to target the W85X mutation in the *Cluc* reporter.

B) Sequencing quantification of A to I editing for indicated guides transfected with dCas13b. For each guide, the region of duplex RNA is outlined in red. Values represent mean +/- S.E.M. Non-targeting guide is the same as in Fig2C.

C) Sequencing quantification of A to I editing for indicated guides transfected with REPAIRv1. For each guide, the region of duplex RNA is outlined in red. Values represent mean +/- S.E.M. Non-targeting guide is the same as in Fig2C.

D) Comparison of on-target A to I editing rates for dCas13b and dCas13b-ADAR2$_{DD}$(E488Q) for guides tested in panel B and C.

E) Influence of base identify opposite the targeted adenosine on luciferase activity restoration by REPAIRv1. Values represent mean +/– S.E.M.

10.6.6 Figure 7.S6

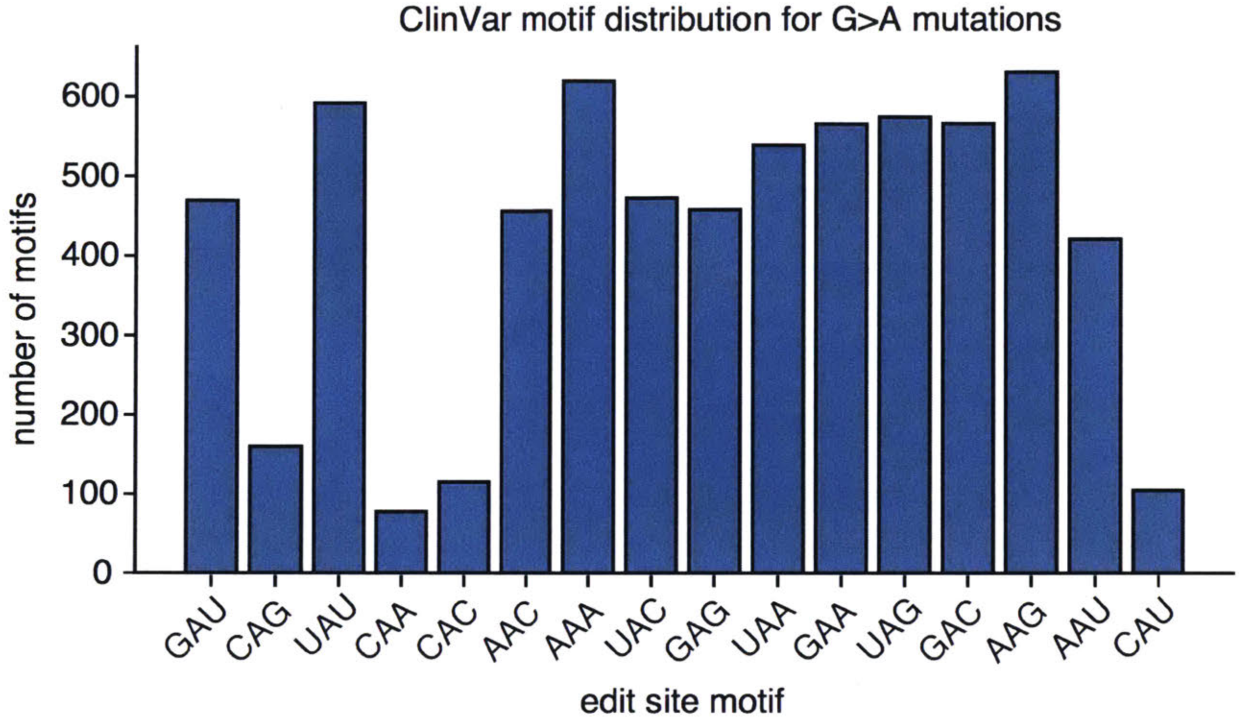


Figure 7.S6: ClinVar motif distribution for G>A mutations.

The number of each possible triplet motif observed in the ClinVar database for all G>A mutations.

10.6.7 Figure 7.S7

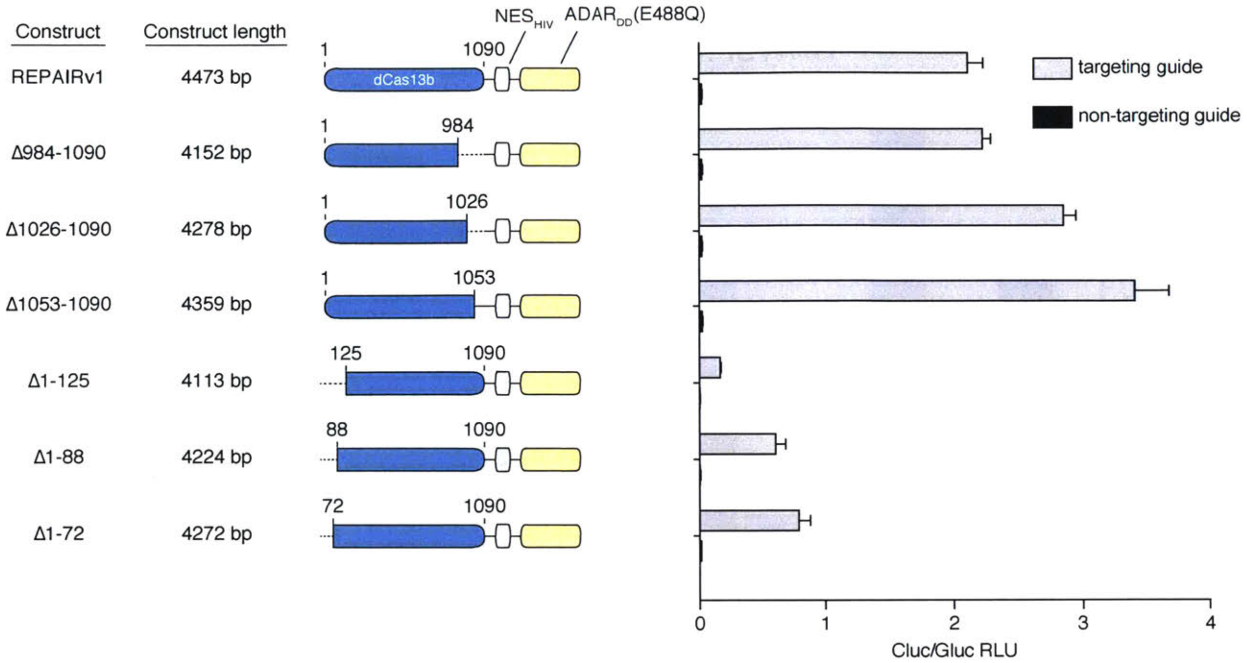


Figure 7.S7: Truncations of dCas13b support functional RNA editing.

N-terminal and C-terminal truncations of dCas13b allow for RNA editing as measured by restoration of luciferase signal for the *Cluc* W85X reporter. Values represent mean +/- S.E.M. The construct length refers to the coding sequence of the REPAIR constructs.

10.6.8 Figure 7.S8

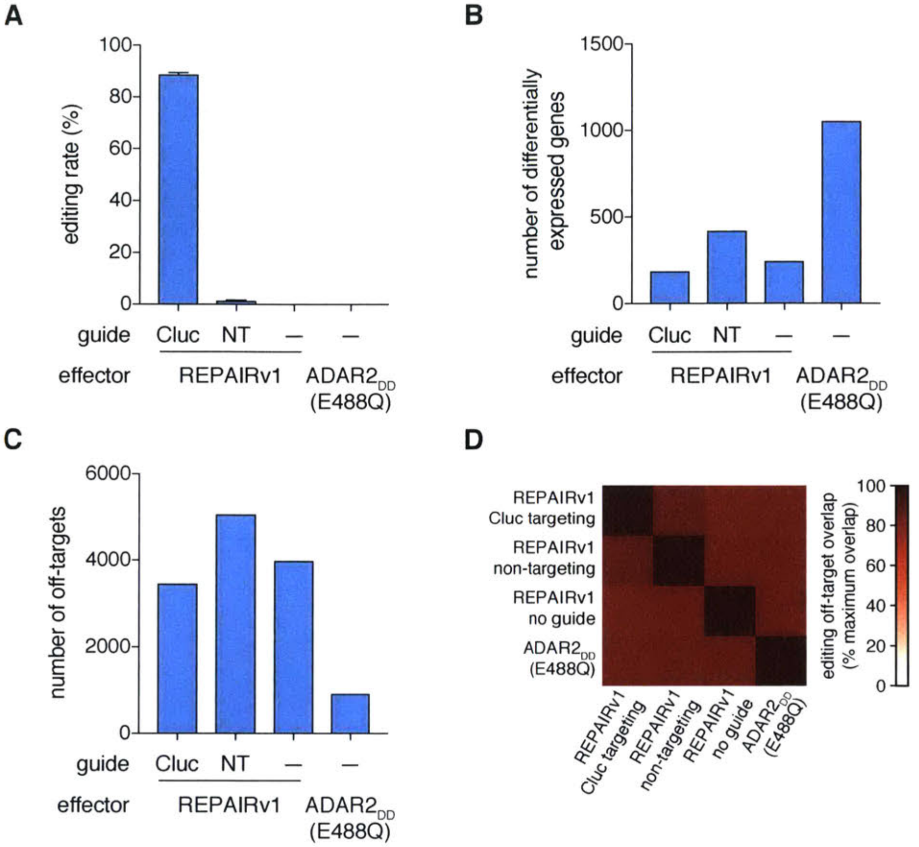


Figure 7.S8: REPAIRv1 editing activity evaluated without a guide and in comparison to ADAR2 deaminase domain alone.

A) Quantification of A to I editing of the *Cluc* W85X mutation by REPAIRv1 with and without guide as well as the ADAR2 deaminase domain only without guide. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig2C.

B) Number of differentially expressed genes in the REPAIRv1 and ADAR2DD conditions from panel A.

C) The number of significant off-targets from the REPAIRv1 and ADAR2DD conditions from panel A.

D) Overlap of off-target A to I editing events between the REPAIRv1 and ADAR2DD conditions from panel A. The values plotted are the percent of the maximum possible intersection of the two off-target data sets.

10.6.9 Figure 7.S9

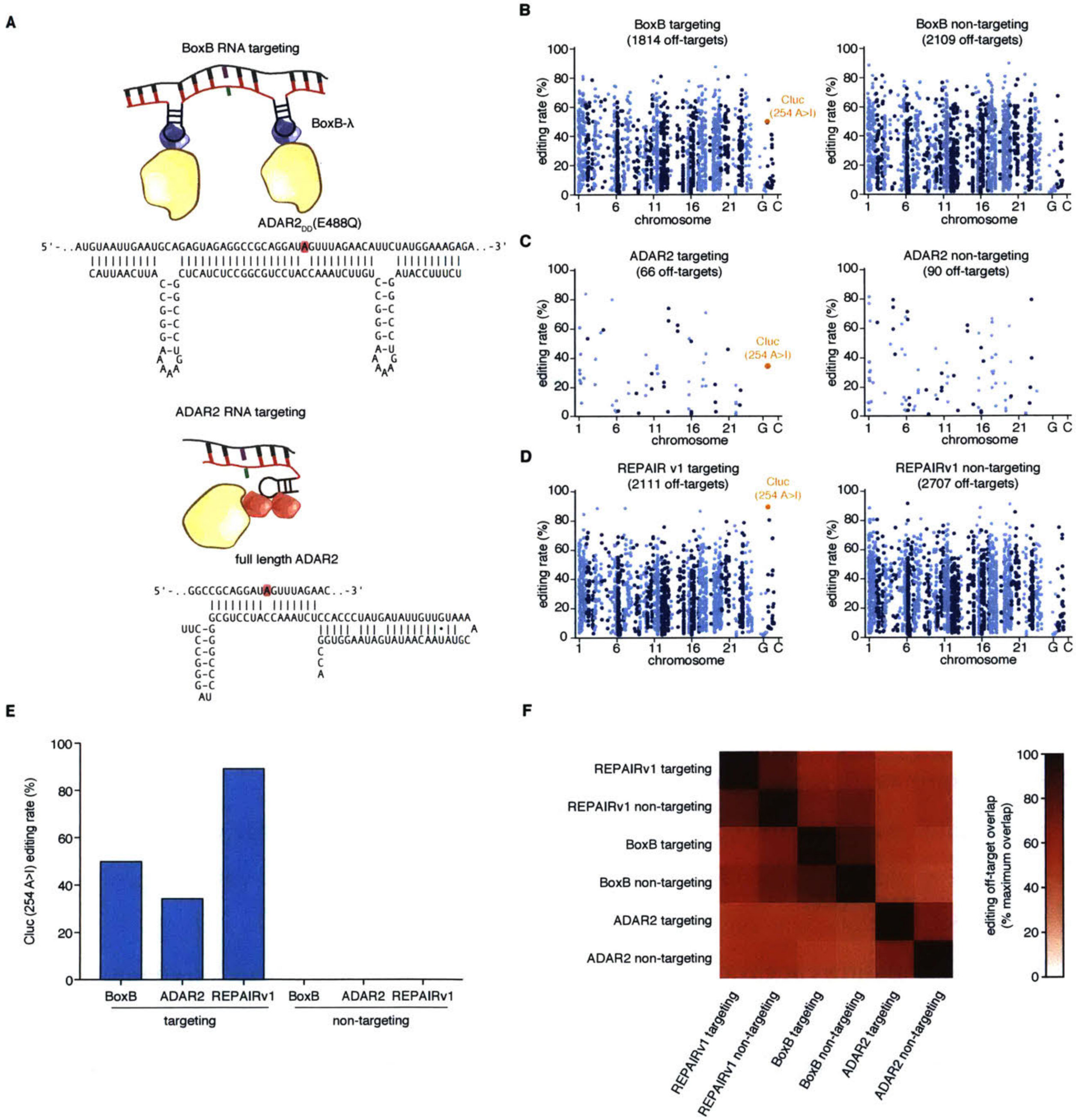


Figure 7.S9: Comparison of REPAIRv1 to other programmable ADAR systems.

A) Schematic of two programmable ADAR schemes: BoxB-based targeting (top) and full length ADAR2 targeting (bottom). For BoxB-based targeting, ADARDD(E488Q) is fused to the viral protein lambda N (BoxB-λ), and the fusion protein is recruited to target adenosines by a guide

RNA containing homology to the target site and hairpins that BoxB-λ binds to. Full length ADAR2 targeting utilizes a guide RNA with homology to the target site and a motif recognized by the double strand RNA binding domains of ADAR2.

B) Transcriptome-wide sites of significant RNA editing by BoxB-ADAR2 $_{\mathrm{DD}}$(E488Q) with a guide targeting *Cluc* and a non-targeting guide. The on-target *Cluc* site (254 A>I) is highlighted in orange.

C) Transcriptome-wide sites of significant RNA editing by full length ADAR2 with a guide targeting *Cluc* and a non-targeting guide. The on-target *Cluc* site (254 A>I) is highlighted in orange.

D) Transcriptome-wide sites of significant RNA editing by REPAIRv1 with a guide targeting *Cluc* and a non-targeting guide. The on-target *Cluc* site (254 A>I) is highlighted in orange. The non-targeting guide is the same as in Fig2C.

E) Quantification of on-target editing rate percentage for BoxB-ADAR2 $_{\mathrm{DD}}$(E488Q), ADAR2, and REPAIRv1 for targeting guides against *Cluc.*

F) Overlap of off-target sites between different targeting and non-targeting conditions for programmable ADAR systems. The values plotted are the percent of the maximum possible intersection of the two off-target data sets.

10.6.10 Figure 7.S10

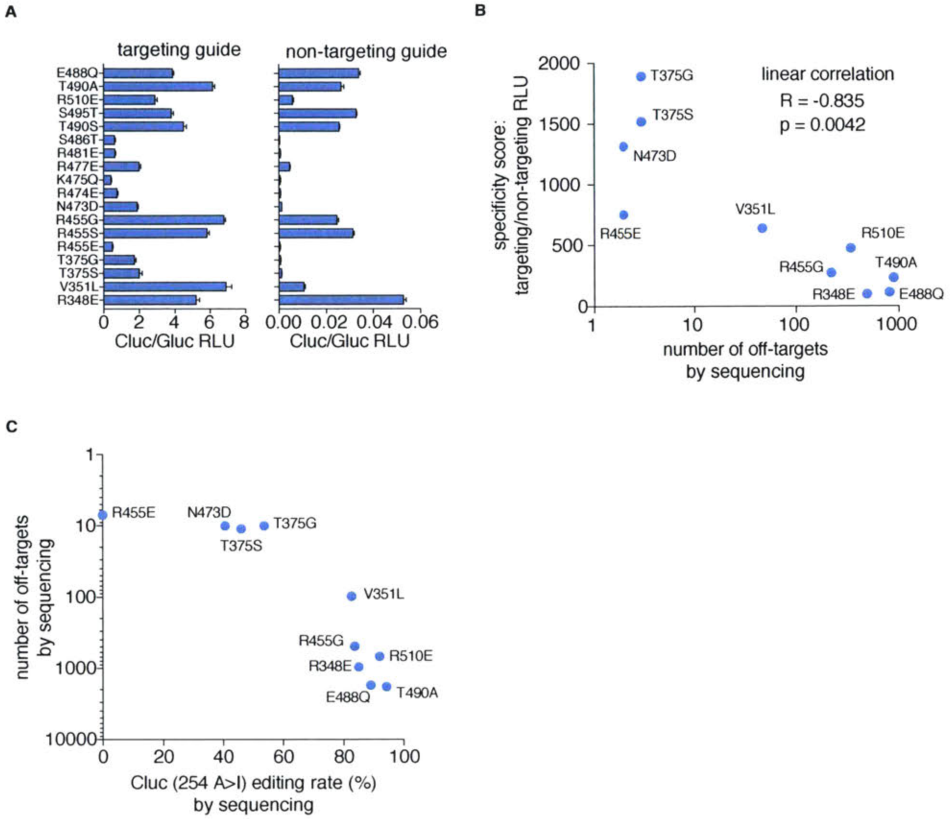


Figure 7.S10: Efficiency and specificity of dCas13b-ADAR2 $_{DD}$(E488Q) mutants.

A) Quantification of luciferase activity restoration by dCas13b-ADAR2$_{DD}$(E488Q) mutants for *Cluc*-targeting and non-targeting guides. Non-targeting guide is the same as in Fig2C.

B) Relationship between the ratio of targeting and non-targeting guide RLU and the number of RNA-editing off-targets as quantified by transcriptome-wide sequencing

C) Quantification of transcriptome-wide off-target RNA editing sites versus on-target *Cluc* editing efficiency for dCas13b-ADAR2 $_{DD}$(E488Q) mutants.

10.6.11 Figure 7.S11

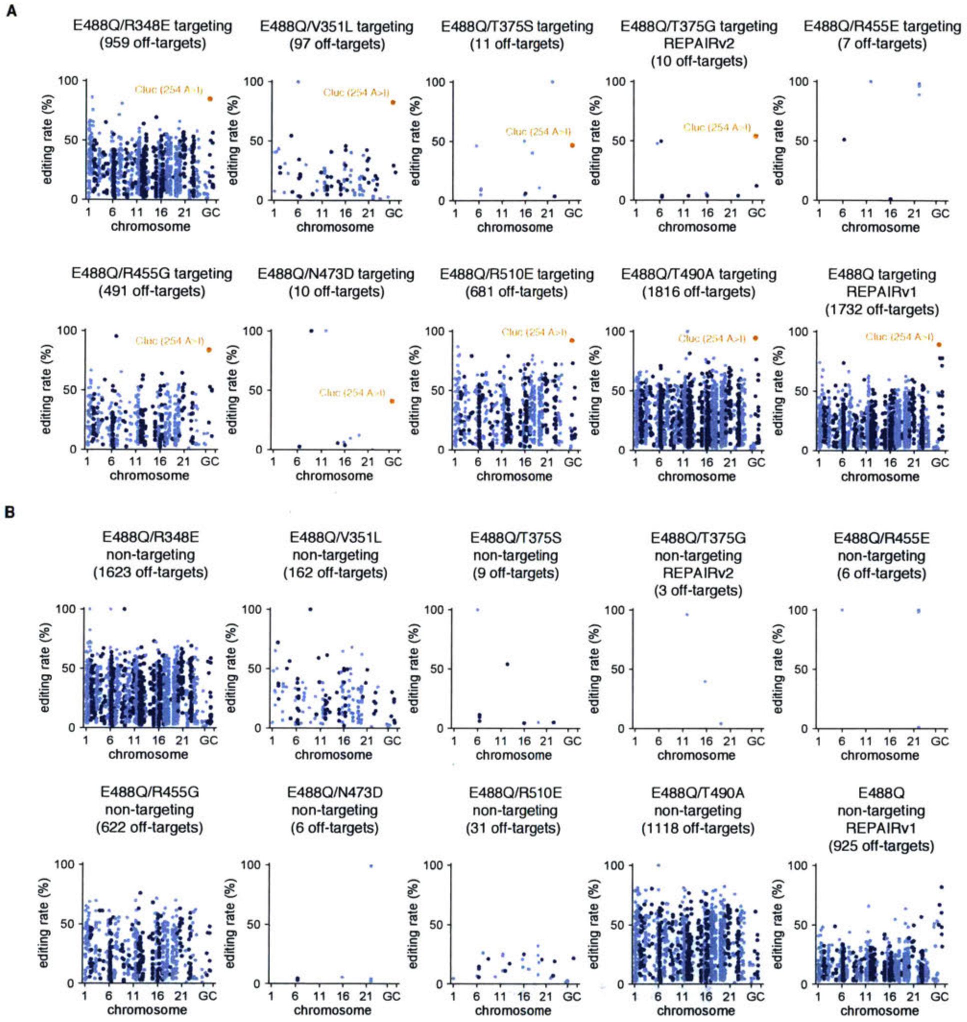


Figure 7.S11: Transcriptome-wide specificity of RNA editing by dCas13b-ADAR2 $_{DD}$(E488Q).

A) Transcriptome-wide sites of significant RNA editing by dCas13b-ADAR2 $_{DD}$(E488Q) mutants with a guide targeting *Cluc*. The on-target *Cluc* site (254 A>I) is highlighted in orange.

B) Transcriptome-wide sites of significant RNA editing by dCas13b-ADAR2 $_{DD}$(E488Q) mutants with a non-targeting guide.

10.6.12 Figure 7.S12

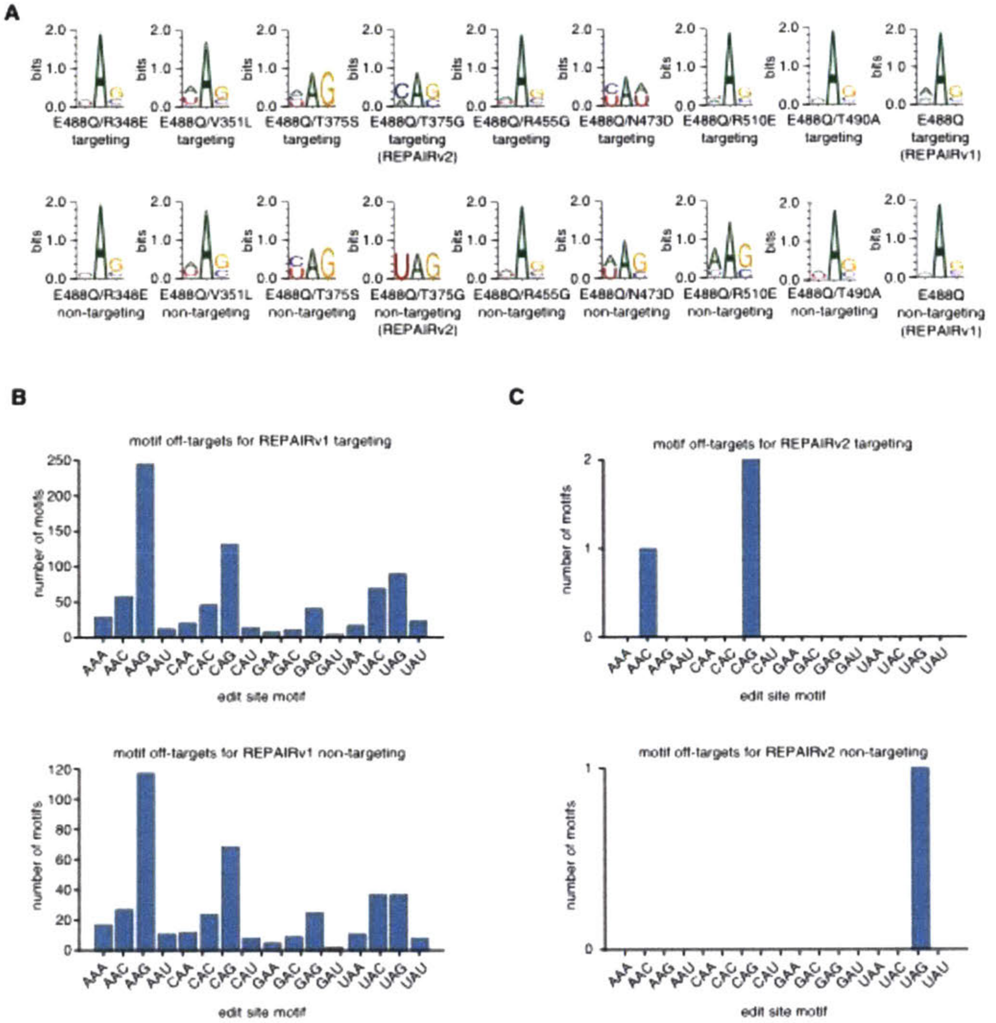


Figure 7.S12: Characterization of motif biases in the off-targets of dCas13b-ADAR2 $_{DD}$(E488Q) editing.

A) For each dCas13b-ADAR2 $_{DD}$(E488Q) mutant, the motif present across all A>I off-target edits in the transcriptome is shown.

B) The distribution of off-target A>I edits per motif identity is shown for REPAIRv1 with targeting and non-targeting guide.

C) The distribution of off-target A>I edits per motif identity is shown for REPAIRv2 with targeting and non-targeting guide.

10.6.13 Figure 7.S13

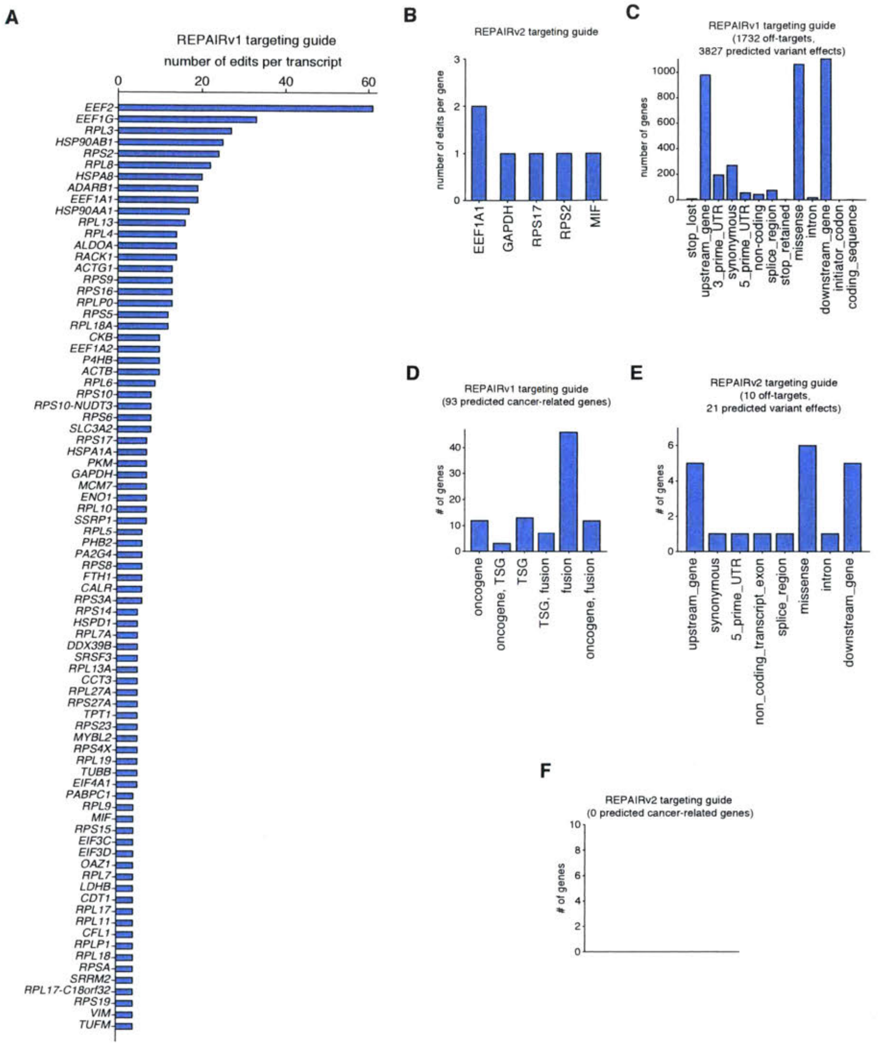


Figure 7.S13: Further characterization of REPAIRv1 and REPAIRv2 off-targets.

A) Histogram of the number of off-targets per transcript for REPAIRv1.

B) Histogram of the number of off-targets per transcript for REPAIRv2.

C) Variant effect prediction of REPAIRv1 off targets.

D) Distribution of REPAIRv1 off targets in cancer-related genes. TSG, tumor suppressor gene.

E) Variant effect prediction of REPAIRv2 off targets.

Distribution of REPAIRv2 off targets in cancer-related genes.

10.6.14 Figure 7.S14

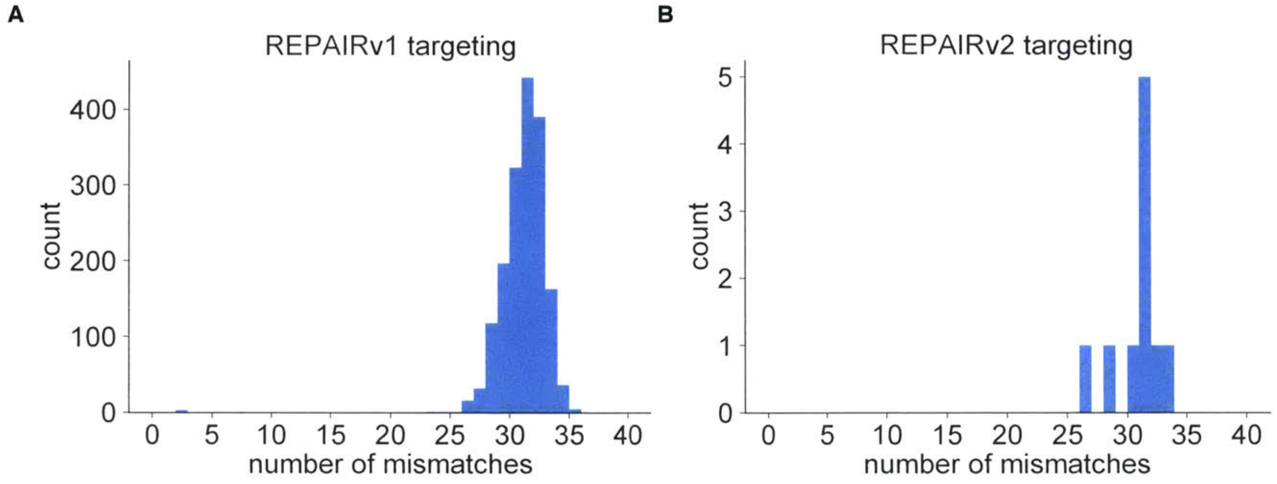


Figure 7.S14: Evaluation of off-target sequence similarity to the guide sequence.

A) Distribution of the number of mismatches (hamming distance) between the targeting guide sequence and the off-target editing sites for REPAIRv1 with a Cluc targeting guide.

B) Distribution of the number of mismatches (hamming distance) between the targeting guide sequence and the off-target editing sites for REPAIRv2 with a Cluc targeting guide.

10.6.15 Figure 7.S15

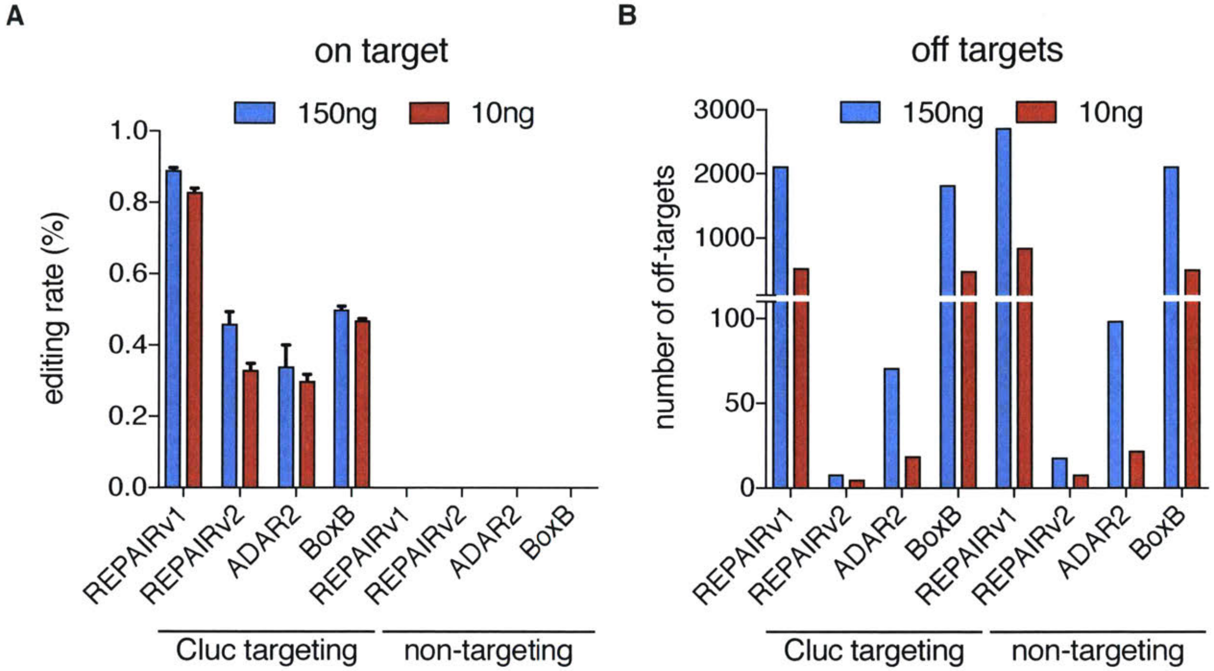


Figure 7.S15: Comparison of REPAIRv1, REPAIRv2, ADAR2 RNA targeting, and BoxB RNA targeting at two different doses of vector (150ng and 10ng effector).

A) Quantification of RNA editing activity at the *Cluc* W85X (254 A>I) on-target editing site by REPAIRv1, REPAIRv2, ADAR2 RNA targeting, and BoxB RNA targeting approaches. Each of the four methods were tested with a targeting or non-targeting guide. Values shown are the mean of the three replicates.

B) Quantification of RNA editing off-targets by REPAIRv1, REPAIRv2, ADAR2 RNA targeting, and BoxB RNA targeting approaches. Each of the four methods were tested with a targeting guide for the *Cluc* W85X (254 A>I) site or non-targeting guide. For REPAIR constructs, non-targeting guide is the same as in Fig. 2C.

10.6.16 Figure 7.S16

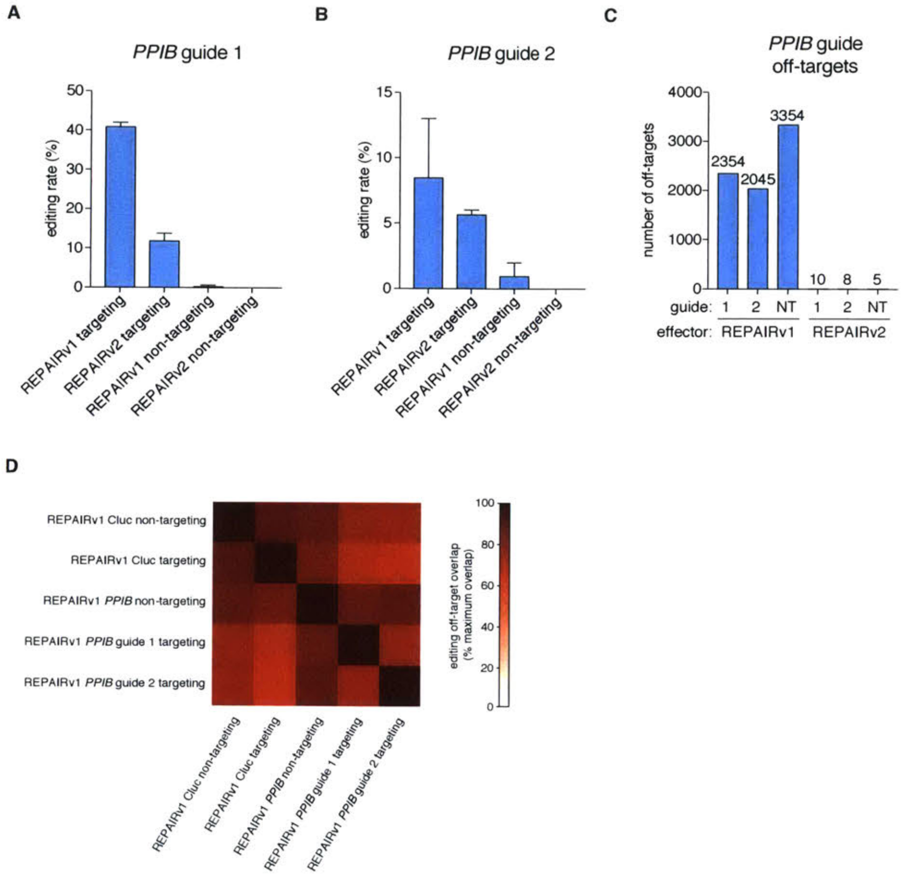


Figure 7.S16: RNA editing efficiency and genome-wide specificity of REPAIRv1 and REPAIRv2.

A) Quantification of RNA editing activity at the *PPIB* guide 1 on-target editing site by REPAIRv1, REPAIRv2 with targeting and non-targeting guides. Values represent mean +/– S.E.M.

B) Quantification of RNA editing activity at the *PPIB* guide 2 on-target editing site by REPAIRv1, REPAIRv2 with targeting and non-targeting guides. Values represent mean +/– S.E.M.

C) Quantification of RNA editing off-targets by REPAIRv1 or REPAIRv2 with *PPIB* guide 1, *PPIB* guide 2, or non-targeting guide.

D) Overlap of off-targets between REPAIRv1 for *PPIB* targeting, Cluc targeting, and non-targeting guides. The values plotted are the percent of the maximum possible intersection of the two off-target data sets.

10.6.17 Figure 7.S17

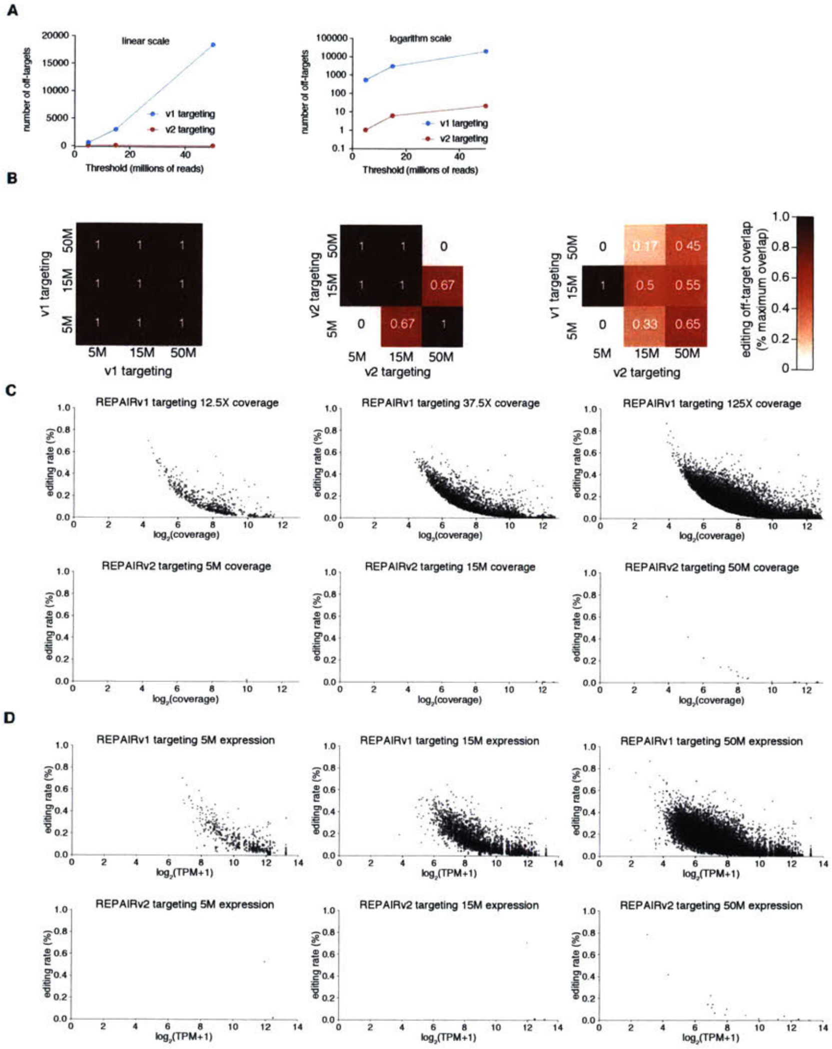


Figure 7.S17: High coverage sequencing of REPAIRv1 and REPAIRv2 off-targets.

A) Quantitation of off-target edits for REPAIRv1 and REPAIRv2 as a function of read depth with a total of 5 million reads (12.5x coverage), 15 million reads (37.5x coverage) and 50 million reads (125x coverage) per condition.

B) Overlap of off-target sites at different read depths of the following conditions: REPAIRv1 versus REPAIRv1 (left), REPAIRv2 versus REPAIRv2 (middle), and REPAIRv1 versus REPAIRv2 (right). The values plotted are the percent of the maximum possible intersection of the two off-target data sets.

C) Editing rate of off-target sites compared to the coverage (log2(number of reads)) of the off-target for REPAIRv1 and REPAIRv2 targeting conditions at different read depths.

D) Editing rate of off-target sites compared to the log2(TPM+1) of the off-target gene expression for REPAIRv1 and REPAIRv2 targeting conditions at different read depths.

10.6.18 Figure 7.S18

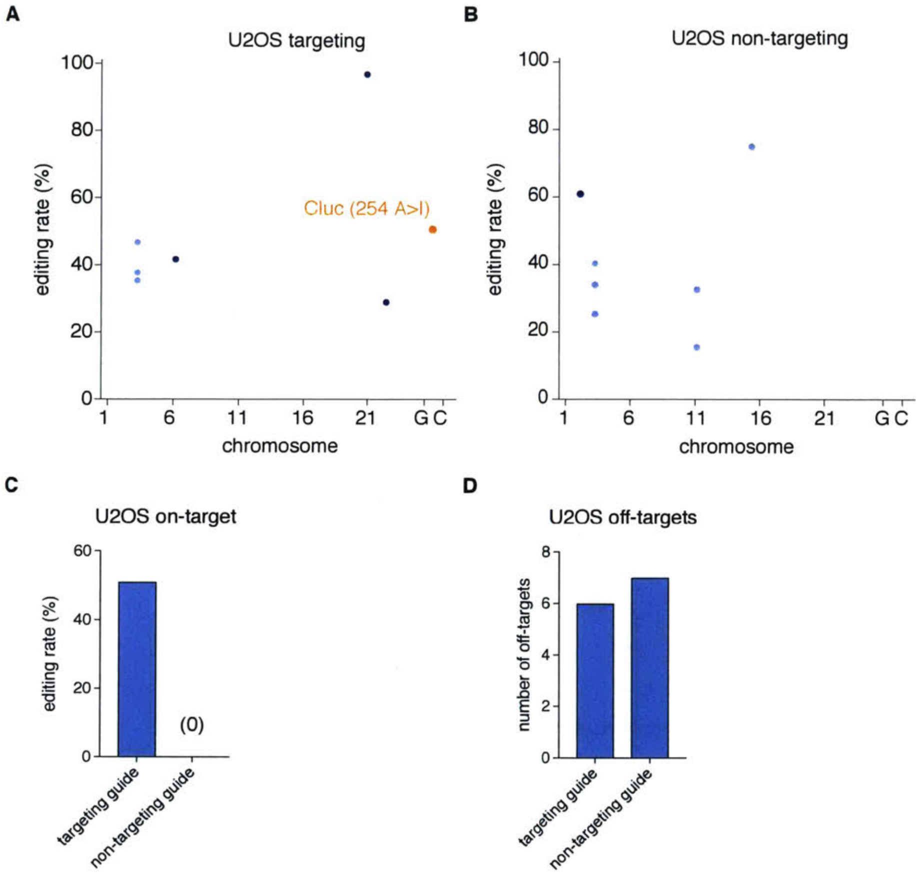


Figure 7.S18: Quantification of REPAIRv2 activity and off-targets in the U2OS cell line.

A) Transcriptome-wide sites of significant RNA editing by REPAIRv2 with a guide targeting *Cluc* in the U2OS cell line. The on-target *Cluc* site (254 A>I) is highlighted in orange.

B) Transcriptome-wide sites of significant RNA editing by REPAIRv2 with a non-targeting guide in the U2OS cell line.

C) The on-target editing rate at the *Cluc* W85X (254 A>I) by REPAIRv2 with a targeting guide or non-targeting guide in the U2OS cell line.

D) Quantification of off-targets by REPAIRv2 with a guide targeting *Cluc* or non-targeting guide in the U2OS cell line.

10.6.19 Figure 7.S19

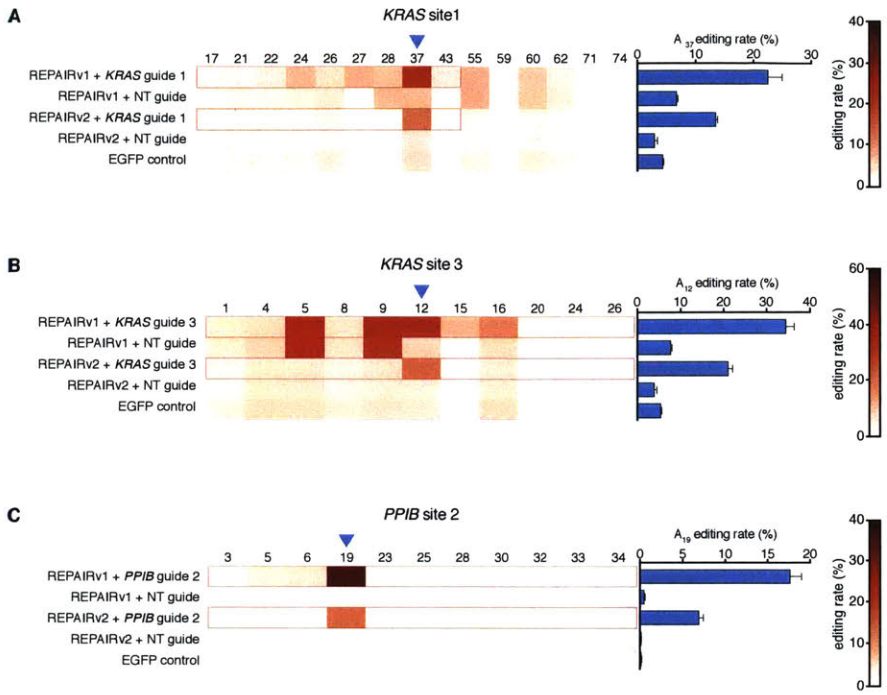


Figure 7.S19: RNA editing efficiency and specificity of REPAIRv1 and REPAIRv2.

A) Quantification of percent editing of *KRAS* with *KRAS*-targeting guide 1 at the targeted adenosine (blue triangle) and neighboring sites for REPAIRv1 and REPAIRv2. For each guide, the region of duplex RNA is outlined in red. Values represent mean +/– S.E.M. Non-targeting guide is the same as in Fig. 2C.

B) Quantification of percent editing of *KRAS* with *KRAS*-targeting guide 3 at the targeted adenosine and neighboring sites for REPAIRv1 and REPAIRv2. Non-targeting guide is the same as in Fig. 2C.

C) Quantification of percent editing of *PPIB* with *PPIB*-targeting guide 2 at the targeted adenosine and neighboring sites for REPAIRv1 and REPAIRv2. Non-targeting guide is the same as in Fig. 2C.

10.6.20 Figure 7.S20

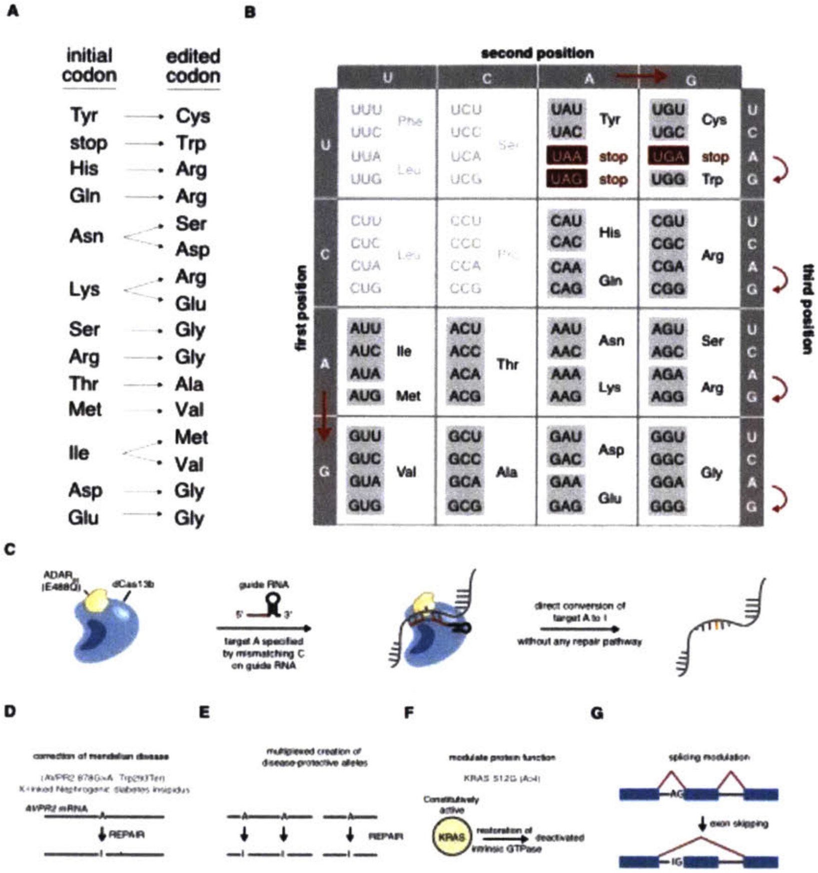


Figure 7.S20: Demonstration of all potential codon changes with an A>I RNA editor.

A) Table of all potential codon transitions enabled by A>I editing.

B) A codon table demonstrating all the potential codon transitions enabled by A>I editing. Adapted and modified based on (Watson, 2014).

C) Model of REPAIR A to I editing of a precisely encoded nucleotide via a mismatch in the guide sequence. The A to I transition is mediated by the catalytic activity of the ADAR2 deaminase domain and will be read as a guanosine by translational machinery. The base change does not rely on endogenous repair machinery and is permanent for as long as the RNA molecule exists in the cell.

D) REPAIR can be used for correction of Mendelian disease mutations.

E) REPAIR can be used for multiplexed A to I editing of multiple variants for engineering pathways or modifying disease. Multiplexed guide delivery can be achieved by delivering a single CRISPR array expression cassette since the Cas13b enzyme processes its own array.

F) REPAIR can be used for modifying protein function through amino acid changes that affect enzyme domains, such as kinases.

G) REPAIR can modulate splicing of transcripts by modifying the splice acceptor site.