HHS Public Access

Author manuscript

Mol Cell. Author manuscript; available in PMC 2019 April 19.

Published in final edited form as:

Mol Cell. 2018 April 19; 70(2): 327–339.e5. doi:10.1016/j.molcel.2018.02.028.

Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL domain-containing accessory protein

Winston X. Yan[1,*], **Shaorong Chong**[1], **Huaibin Zhang**[1], **Kira S. Makarova**[2], **Eugene V. Koonin**[2], **David R. Cheng**[1,*,†], and **David A. Scott**[1,*,†]

[1]Arbor Biotechnologies, Cambridge, MA 02139, USA

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Abstract

Bacterial class 2 CRISPR-Cas systems utilize a single RNA-guided protein effector to mitigate viral infection. We aggregated genomic data from multiple sources and constructed an expanded database of predicted class 2 CRISPR-Cas systems. A search for novel RNA targeting systems identified subtype VI-D, encoding dual HEPN-domain containing Cas13d effectors and putative WYL-domain containing accessory proteins (WYL1 and WYL-b1–5). The median size of Cas13d proteins is 190 to 300 amino acids smaller than that of Cas13a-c. Despite their small size, Cas13d orthologs from *Eubacterium siraeum* (Es) and *Ruminococcus sp.* (Rsp) are active in both CRISPR RNA processing and target as well as collateral RNA cleavage, with no target-flanking sequence requirements. The RspWYL1 protein stimulates RNA cleavage by both EsCas13d and RspCas13d, demonstrating a common regulatory mechanism for divergent Cas13d orthologs. The small size, minimal targeting constraints, and modular regulation of Cas13d effectors further expands the CRISPR toolkit for RNA-manipulation and detection.

eTOC

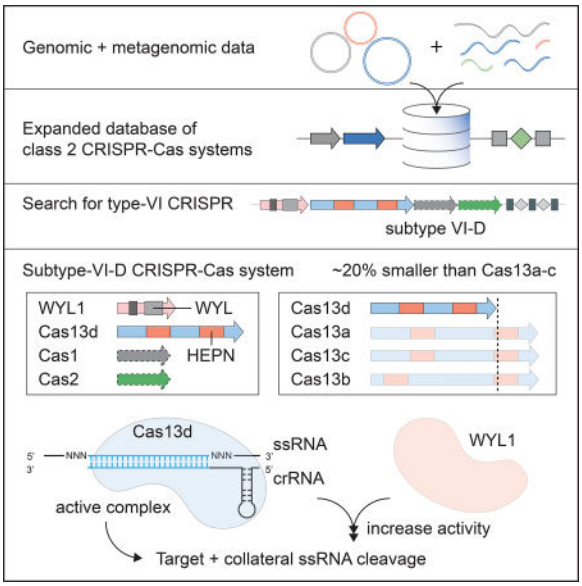
†Correspondence: dscott@arbor.bio; dcheng@arbor.bio.
*These authors contributed equally to this work
Lead contact: D.A.S., dscott@arbor.bio

AUTHOR CONTRIBUTIONS

W.X.Y., D.R.C., and D.A.S., with input from S.C. and H.Z., conceived and designed the study. D.R.C. and D.A.S. designed and implemented the computational search. W.X.Y., D.A.S., S.C., and H.Z. designed and performed all experiments. D.A.S., D.R.C., and W.X.Y., with assistance from S.C. and H.Z., analyzed all data. K.S.M., and E.V.K. provided input on annotation, classification, and naming of Cas13d and WYL1, and WYL-b1 through WYL-b5. W.X.Y., D.R.C., and D.A.S. wrote the manuscript with input from E.V.K. and help from all authors.

DECLARATION OF COMPETING INTERESTS

D.R.C, S.C., D.A.S., W.X.Y., and H.Z. declare competing financial interests as employees and as stockholders of Arbor Biotechnologies.



Compiling an expanded database of predicted class 2 CRISPR-Cas systems, Yan et al. identify and characterize subtype VI-D. Cas13d is a RNA-guided RNase effector with polyphyletic WYL-domain accessory proteins. One WYL1 ortholog enhances activity of divergent Cas13d orthologues. The small effector size and modular enhancement further expand RNA modification capabilities.

INTRODUCTION

CRISPR-Cas systems provide adaptive immunity against viruses and other mobile genetic elements (MGE) in archaea and bacteria (Barrangou and Horvath, 2017; Mohanraju et al., 2016). The CRISPR-Cas-mediated defensive response includes three stages: 1) adaptation, during which the Cas1–Cas2 protein complex (sometimes containing additional subunits) inserts a segment of the target DNA (protospacer) between the repeats at the 5’end of a CRISPR array via a copy-and-paste mechanism to create a new spacer, 2) expression and processing, when the CRISPR array is transcribed into a single, long transcript, the precursor CRISPR (pre-cr)RNA that is then processed by a distinct complex of Cas proteins or an external RNase into mature CRISPR (cr)RNAs, and 3) interference, when a Cas protein complex (in many cases, a modified processing complex) employs the crRNA as a guide to cleave the target DNA or RNA (Amitai and Sorek, 2016; Jackson et al., 2017; Marraffini, 2015; van der Oost et al., 2014).

Having evolved in a relentless arms race with MGE elements, CRISPR-Cas systems exhibit diverse defense activities, including numerous mechanisms of interference using RNA-guided DNA and RNA cleavage, as well as the associated regulatory systems to control these activities (Mohanraju et al., 2016; Shmakov et al., 2017a). The diversity of the CRISPR-Cas systems is currently organized into two classes that differ in the architecture of their effector modules that are involved in interference and, in most CRISPR-Cas variants, also in the crRNA maturation. In class 1 systems, the effector module is a complex of multiple Cas proteins, whereas in class 2 systems, it is represented by a single multi-domain

effector protein (Makarova et al., 2015). The relative simplicity of the organization of class 2 systems has enabled their widespread adoption as a toolkit for genome editing and other applications (Cong et al., 2013; Gasiunas et al., 2012; Hsu et al., 2014; Jinek et al., 2012; Mali et al., 2013; Wright et al., 2016).

Class 2 CRISPR-Cas systems currently include 3 distinct types, II, V and VI, which are subdivided into 11 subtypes. The CRISPR-Cas types within class 2 share homologous nuclease domains, namely RuvC family DNase domains in types II and V, and HEPN RNase domain in type VI, but otherwise, show low sequence conservation and considerable functional versatility (Shmakov et al., 2017a). The recent progress in expanding the diversity of class 2 CRISPR-Cas systems (Abudayyeh et al., 2016; Burstein et al., 2017; East-Seletsky et al., 2016, 2017; Fonfara et al., 2016; Shmakov et al., 2015; Smargon et al., 2017; Zetsche et al., 2015) has uncovered not only new effector proteins, but also novel fundamental molecular mechanisms of activity and regulation (Koonin et al., 2017). The continued discovery of new proteins and mechanisms within rapidly growing repositories of genomic and metagenomic sequencing data necessitates a systematic, scalable, and efficient approach combining computational and experimental screening to identify new diverse activities and their parameters. We describe here the initial use of such an approach towards the expansion and characterization of an additional type VI CRISPR-Cas subtype.

Type VI CRISPR-Cas systems are the only known dedicated RNA-targeting immune systems in prokaryotes (Abudayyeh et al., 2016; East-Seletsky et al., 2016, 2017; Shmakov et al., 2015; Smargon et al., 2017). Type VI-A and VI-B systems have been characterized in detail and shown to possess the crRNA-dependent target cleavage activity and a non-specific, collateral RNase activity that is stimulated by target recognition and cleavage (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Smargon et al., 2017). Both of these activities are mediated by the two HEPN domains contained in type VI effectors known as Cas13a and Cas13b (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Liu et al., 2017a; Smargon et al., 2017). In addition to these RNase activities, the known Cas13 proteins also catalyze the processing of the pre-crRNA into their mature form independent of divalent cations (East-Seletsky et al., 2016; Knott et al., 2017). Apart from these commonalities, Cas13a and Cas13b show substantial differences. In particular, Cas13b appears to be a more robust nuclease than Cas13a (Cox et al., 2017), and furthermore, unlike Cas13a, the activity of Cas13b is either upregulated or downregulated by distinct accessory proteins encoded in type VI-B loci (Smargon et al., 2017). Although only discovered recently, both Cas13a and Cas13b have already been harnessed as highly efficient tools for RNA targeting (Abudayyeh et al., 2016, 2017) and editing (Cox et al., 2017) as well as sensitive nucleic acid detection (East-Seletsky et al., 2016; Gootenberg et al., 2017).

Considering the functional flexibility of type VI CRISPR-Cas systems and their success as molecular tools, we sought to characterize additional RNA-targeting programmable nucleases. Through systematic processing of genomic and metagenomic sequences, we created a large database of putative CRISPR-Cas systems. Searching this database for effector proteins containing two HEPN domains identified a distinct variant of type VI CRISPR-Cas, subtype VI-D, which is distantly related to the previously characterized type VI-A systems (and even more distant from subtypes VI-B and VI-C). We demonstrated

robust target cleavage and collateral RNase activities of the subtype VI-D effectors (Cas13d) along with their ability to process pre-crRNA. Additionally, Cas13d effectors have a strong co-occurrence with polyphyletic WYL-domain containing accessory proteins. We show that for one sub-branch, designated WYL1, a single protein ortholog positively modulates target and collateral RNAse activity of divergent Cas13d orthologues. Overall, the notably smaller effector size of Cas13d and modular regulatory activity of WYL1 make this system a particularly attractive candidate for diverse RNA manipulation applications.

RESULTS

Building an expanded database of CRISPR-Cas systems

We developed a computational pipeline to produce an expanded dataase of class 2 CRISPR-Cas systems from genomic and metagenomic sources. We define a candidate class 2 CRISPR-Cas system to comprise: 1) a cluster of homologous putative effector proteins encoded in different contigs, 2) a CRISPR array associated with each putative effector, and optionally, 3) putative accessory proteins that stably co-occur with the putative effector. Aggregating and processing a collection of more than 10 Tb of prokaryotic genomic and metagenomic sequence data from multiple sources, our pipeline produced a database of 293,985 putative CRISPR-Cas systems. One important difference from previously reported computational pipelines (Shmakov et al., 2015, 2017a; Smargon et al., 2017) is that we perform minimal filtering (e.g., imposing a minimum size on putative effector) in the intermediate stages of the search in order to expand the range for potential discovery of novel CRISPR-Cas systems. As such, the resulting database of putative CRISPR-Cas loci includes all previously characterized class 2 CRISPR-Cas systems, but also contains a considerable amount of noise, such as degraded, non-functional CRISPR-Cas loci. We developed additional heuristics to perform parametric searches (e.g., by effector size or predicted effector function) through the database of putative systems, and ranked the results by their potential to be functional CRISPR-Cas systems (see Methods).

Genomic survey of type VI RNA-targeting CRISPR-Cas systems

To expand the repertoire of Cas nucleases for RNA manipulation and sensing, we searched our database for type VI CRISPR-Cas systems with effector proteins containing two HEPN-domains each (2-HEPN proteins). In addition to the previously identified 2-HEPN proteins, Cas13a, Cas13b and Cas13c, we detected a group of 2-HEPN proteins distantly related to Cas13a (effectors of type VI-A), primarily in *Eubacterium* and *Ruminococcus*, which we denote Cas13d. The amino acid sequences of Cas13d proteins show less than 8% identity to the most similar Cas13a sequences; nevertheless, statistically significant sequence similarity between Cas13d and Cas13a can be demonstrated using PSI-BLAST initiated with a profile made from the multiple alignment of Cas13a (E-value = 0.002). This significant similarity is primarily due to the conservation of the HEPN domain sequences between Cas13a and Cas13d, whereas the remaining portions of the protein sequences in the two families are highly divergent; in particular, Cas13d proteins lack a counterpart to the Helical-1 domain of Cas13a (Figure S1B). Two representatives of subtype VI-D have been detected in our previous searches for type VI CRISPR-Cas systems but have not been identified as a distinct group or explored in detail (Shmakov et al., 2017a). Phylogenetic analysis of the Cas13

proteins clearly shows that Cas13a and Cas13d form strongly supported clades (Figure S1A). Additionally, Cas13d effectors are notably smaller than previously characterized class 2 CRISPR effectors, with a median size of 928 aa. For comparison, this median size is 190 aa (17%) less than that of Cas13c, more than 200 aa (18%) less than that of Cas13b, and more than 300 aa (26%) less than that of Cas13a (Figure 1C). Taken together, these lines of evidence suggest that this distinct group of class 2 CRISPR-Cas systems are best classified as subtype VI-D, with the effector denoted Cas13d (Figure 1B).

We found that 77% of Cas13d genes occur adjacent to CRISPR arrays, and for 19%, the adaptation module (Cas1 and Cas2 genes) is present in the vicinity (Figure 1A), suggesting that many subtype VI-D loci encode CRISPR-Cas systems that are active in both adaptation and interference. Phylogenetic analysis indicates that Cas1 proteins associated with subtype VI-D are monophyletic and, in accord with previous observations on other type VI systems, are affiliated with the type II-A clade (Data File S1). Thus, in the case of type VI, the adaptation module seems to have co-evolved with the effector module.

The CRISPR arrays adjacent to Cas13d genes contain 198 spacers total, of which 182 are unique. A BLASTN (Camacho et al., 2009) search of the unique spacer sequences against a database comprising known phages and NCBI prokaryotic sequences revealed 7 spacers with significant hits (defined as E-value < 0.0001, alignment length at least 24, 0 gaps, and no more than one mismatch). One spacer, from *Ruminococcus flavefaciens FD-1*, showed significant matches against the Arthrobacter dsDNA phage Gordon (alignment length=28, 1 mismatch) and against a putative prophage region in an uncultured *Flavonifractor* sequence (alignment length=24, 0 mismatches). A different spacer, from a gut metagenome sequence, resulted in a significant match against a putative prophage region in *Bacillus soli* (alignment length=24, 0 mismatches). The remaining 5 spacer matches targeted ORFs in prokaryotic sequences, but were not classified as being in prophage regions (see Methods). The presence of spacers homologous to DNA phage genomic sequences in an RNA-targeting CRISPR-Cas system might appear unexpected but is in line with similar observations on type VI-A and type VI-B systems (Smargon et al., 2017). Presumably, type VI systems abrogate the reproduction of DNA phages by cleaving phage mRNAs, but the mechanistic details of the antivirus activity of these systems remain to be characterized experimentally.

Examination of the additional genes in the vicinity of Cas13d led to the identification in most of the VI-D loci of potential accessory proteins containing WYL domains (so denoted after three amino acids that were conserved in the originally identified group of these domains) and additionally, helix-turn-helix (HTH) or ribbon-helix-helix (RHH) DNA-binding domains (Figure S2A). The WYL-domain proteins are often associated with various microbial defense systems (Makarova et al., 2014), and one of these proteins has been shown to negatively regulate a subtype I-D CRISPR-Cas system in *Synechococcus* (Hein et al., 2013). The WYL-domain proteins contained in subtype VI-D loci fall into 6 strongly supported branches of the broader phylogenetic tree of WYL-domain proteins (Data File S2). The branch we denote WYL1 is a single WYL-domain protein associated primarily with *Ruminococcus*. Multiple sequence alignment of WYL1 shows an N-terminal RHH domain, as well as a pattern of primarily hydrophobic conserved residues, including an invariant tyrosine-leucine doublet corresponding to the original WYL motif (Figure S2B).

Other VI-D loci contain duplicated genes encoding WYL-domain proteins, as in *Ruminococcus flavefaciens*, or a fusion of two WYL-domain proteins, as in *Eubacterium sp*. Although a substantial majority of the VI-D loci encode WYL-domain proteins, phylogenetic analysis indicates that these CRISPR-associated WYL proteins are scattered among different branches of the WYL family tree, i.e. are polyphyletic (Figure S2A). Thus, the VI-D CRISPR-Cas systems appear to have acquired WYL-domain proteins on several independent occasions, suggesting a role for this protein in modulating the CRISPR-Cas function.

Synthesis of a minimal system for Cas13d experimental evaluation

Having identified the minimal suite of subtype VI-D CRISPR-Cas system components, we selected two loci for functional validation, those from *Eubacterium siraeum* DSM 15702 (EsCas13d) and *Ruminococcus sp*. N15.MGS-57 (RspCas13d). RspCas13d is a member of the largest subgroup of Cas13d proteins which contains 13 of the 31 unique members of the family and shows co-conservation with a putative WYL1 accessory protein (Figures 1A, S2B). In contrast, there are no WYL-domain proteins (or other putative accessory proteins) encoded within 3kb of the EsCas13d effector.

To test the activity of subtype VI-D CRISPR-Cas, we designed and synthesized minimal systems containing RspCas13d or EsCas13d into the pET28a(+) vector. The synthesized *Ruminococcus sp*. RspCas13d system included RspCas13d and RspWYL1, codon optimized for *E. coli* expression under the control of a *lac* promoter and separated by an *E. coli* ribosome binding sequence (Figure 2A). Following the open reading frames for RspCas13d and RspWYL1, we included an acceptor site for a CRISPR array library driven by a J23119 promoter. The *Eubacterium siraeum* system was prepared similarly but included no gene for a WYL-domain containing protein.

Accelerated functional screening for Cas13d

To accelerate functional screening of subtype VI-D systems, we developed a strategy to derive the following functional information in a single screen: 1) crRNA expression direction and processing, 2) nucleic acid substrate type, and 3) targeting requirements such as protospacer adjacent motif (PAM), protospacer flanking sequence (PFS), or target secondary structure. We designed minimal CRISPR array libraries consisting of two consensus direct repeats, each flanking a unique natural-length spacer sequence targeting either the pACYC184 vector or an absent GFP sequence as a negative control. The CRISPR array libraries for EsCas13d and RspCas13d systems consisted of 4549 and 3972 pACYC184-targeting spacers respectively, in addition to 452 and 450 spacers targeting the GFP negative control sequence, respectively. We also designed a bidirectional array library cloning strategy to test both possible CRISPR array expression directions in parallel. The CRISPR array libraries for RspCas13d and EsCas13d were cloned into acceptor sites on respective subtype VI-D expression plasmids such that each plasmid contained a single library element and orientation (Figure 2A). The resulting plasmid libraries were transformed with pACYC184 into Stbl3 *E. coli* using electroporation, yielding a maximum of one plasmid library element per cell. Transformed *E. coli* cells were plated on bioassay plates containing Kanamycin (selecting for the library plasmid), Chloramphenicol (CAM;

selecting for intact pACYC184 CAM expression), and Tetracycline (TET; selecting for intact pACYC184 TET expression), such that interruption of pACYC184 plasmid DNA or antibiotic resistance gene expression by the CRISPR-Cas system results in bacterial cell death. Screens were harvested 12h after plating, and plasmid DNA was extracted (Figure 2B). We PCR amplified the CRISPR array region of the input plasmid library prior to transformation and the output plasmid library after bacterial selection on antibiotic plates. To identify specific parameters resulting in enzymatic activity and bacterial cell death, we used next generation sequencing (NGS) to quantify and compare the representation of individual CRISPR arrays (i.e., repeat-spacer-repeat) in the PCR of the input and output plasmid libraries. We define the array depletion ratio as the normalized output read count divided by the normalized input read count. An array was considered to be strongly depleted if the depletion ratio was less than 0.1 (more than 10-fold depletion). When calculating the array depletion ratio across biological replicates, we took the maximum depletion ratio value for a given CRISPR array across all experiments (i.e. a strongly depleted array must be strongly depleted in all biological replicates). We generated a matrix including array depletion ratios and the following features for each spacer target: target strand, transcript targeting, ORI targeting, target sequence motifs, flanking sequence motifs, and target secondary structure. We investigated the degree to which different features in this matrix explained target depletion for RspCas13d and EsCas13d systems, thereby yielding a broad survey of functional parameters within a single screen.

Distribution of bacterial screening targets indicates that Cas13d targets ssRNA transcripts

To identify the targeted substrate for Cas13d, we first identified a set of minimal CRISPR arrays that were strongly depleted in 2 screen biological replicates. For both RspCas13d and EsCas13d systems, these strongly depleted arrays primarily targeted pACYC184, with minimal depletion of the negative control (Figures S3A, S3B). We observed 1119 and 806 strongly depleted arrays for the RspCas13d and EsCas13d systems, respectively (Figure 2C). The spatial distribution and strand preference of the strongly depleted target sites along pACYC184 (Figures 2D, S3C) indicate a preference for transcript targeting, suggesting that Cas13d targets single-stranded RNA transcripts. Additionally, the presence of strongly depleted targets within the non-coding region of pACYC184 between the Tet and CAM ORFs corresponds to the extension of RNA transcripts coding for these genes beyond the end of the open reading frame. These results indicate that targeting of non-essential regions of transcripts might trigger additional catalytic activities of Cas13d enzymes resulting in toxicity and cell death. Such a result is consistent with the activation of non-specific collateral RNA cleavage activity following target RNA hybridization by the Cas13d-crRNA complex, as reported for previously characterized Cas13 enzymes (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Smargon et al., 2017).

Lack of PFS for Cas13d and a new model for analysis of sequence constraints

Previous RNA targeting CRISPR-Cas systems from subtypes VI-A-C have shown varying dependence on a protospacer flanking sequence (PFS) for efficient RNA targeting (Abudayyeh et al., 2016, 2017; Cox et al., 2017; East-Seletsky et al., 2016, 2017; Gootenberg et al., 2017; Smargon et al., 2017). Here we present evidence that RspCas13d and EsCas13d have no such flanking sequence requirements. For each enzyme, WebLogos

(Crooks et al., 2004) show that at each of 30 positions before and after the target sequences for strongly depleted arrays the nucleotide frequencies do not appreciably differ from a uniform distribution (Figure S4B). To further investigate possible flanking sequence requirements, we developed a combinatorial model to search for up to 3 nucleotide locations distributed across the target or flanking sequences that might explain the observed strongly depleted arrays. We calculate bit score to measure the degree to which the selected locations correspond to strongly biased outcomes (e.g. all hits or all non-hits) (see Methods). For CRISPR-Cas systems with known PAM or PFS requirements, such as BzCas13b, high bit scores for targeting requirements of length 2 or 3 within 15 nt flanks of the target were obtained, and accurately recapitulate the location of the known PFS (Figure 3A). Conversely, for RspCas13d and EsCas13d, our analysis shows no evidence of flanking or spacer sequences contributing to the targeting efficiency of strongly depleted arrays (Figure 3A).

Explaining strongly depleted arrays for RspCas13d and EsCas13d

Cumulatively, transcript targeting explained 86% and 66% of the strongly depleted arrays for RspCas13d and EsCas13d, respectively (Figure 3B). Accordingly, little if any targeting was observed for the ORF template strand. Non-coding and origin of replication (ORI) targeting correspond to actively transcribed regions of the ORI and the extension of coding transcripts into the intergenic region, as corroborated by RNA sequencing of Stbl3 *E. coli* containing pACYC184 (Figures 2D, S3C). Secondary structure analysis of the transcripts further enhanced the explanation of targeting for Cas13d. We predicted RNA secondary structure (Lorenz et al., 2011) for all sub-sequences within 30nt of transcript target sites, and found that sequences with no predicted stable secondary structure corresponded to a higher percentage of strongly depleted targets (Figure S4C). Accordingly, we selected several sub-sequence ranges around the target site (Figure S4C), and defined a minimal secondary structure targeting requirement to be satisfied if the target site exhibited no predicted stable secondary structure for any of the selected sequence ranges. Among the transcript target sites that satisfy the minimal secondary structure requirement, we can explain 93% and 84% of all strongly depleted arrays for RspCas13d and EsCas13d, respectively (Figure 3C). Together, our results indicate that RspCas13d and EsCas13d are RNA-targeting effectors with no flanking sequence requirements and a preference for minimal secondary structure for RNA targeting in *E. coli*.

RspCas13d and EsCas13d process pre-crRNAs

Previously characterized type VI CRISPR effectors (Cas13a-c) demonstrate self-processing of pre-crRNAs to form the mature crRNA-enzyme complex capable of targeting ssRNA (Abudayyeh et al., 2016; East-Seletsky et al., 2016, 2017; Smargon et al., 2017). To investigate crRNA biogenesis for subtype VI-D CRISPR-Cas systems, we purified and sequenced small RNAs from *E. coli* expressing EsCas13d and the minimal CRISPR array library from the bacterial screen (Figure 4A). We analyzed the pre-crRNA processing in the screen output samples for the direct repeat orientation that demonstrated successful targeting of pACYC184 and identified a mature 53nt crRNA consisting of a 5’ direct repeat truncated by 6nt (Figure 4A). The most common spacer length observed for EsCas13d was 23nt, with length variation between 20nt and 30nt (length of the native spacer for EsCas13d).

Multiple sequence alignment of Cas13d with Cas13a shows a large gap in the middle of the Cas13d sequences (Figure S1B). Notably, the missing sequence region, which is largely responsible for the reduced size of Cas13d compared to Cas13a, includes the Cas13a Helical-1 domain that has been shown to be responsible for crRNA processing (Liu et al., 2017a). To assess whether EsCas13d and RspCas13d are capable of autonomous pre-crRNA processing, we purified recombinant versions of the EsCas13d and RspCas13d effectors (Figure S5A). The effectors were then incubated with their respective *in vitro* transcribed pre-crRNAs consisting of a minimal CRISPR array with the repeat-spacer-repeat construction used in the bacterial screening library, but with a single spacer instead of a library. We found that EsCas13d and RspCas13d effectors process pre-crRNAs to form mature crRNAs in the absence of any accessory proteins (Figure 4B). Performing next-generation sequencing of the *in vitro* cleaved RNA fragments enabled the exact identification of the processing intermediates and mature crRNA (Figure 4C) visualized by denaturing gel. For both EsCas13d and RspCas13d, sequencing the mature crRNA corroborated the 6nt truncation from the 5' end of the first direct repeat found in the *in vivo* small RNA sequencing. For the 3' end, 6 nt of the second direct repeat remained attached to the 3' end of the spacer, yielding a total product of 66nt consistent with the mature crRNA visualized by denaturing gel. The difference between the well-defined 3' end of the mature crRNA forms observed *in vitro* versus the various lengths identified *in vivo* may be the result of further truncation *in vivo* by endogenous RNases following the initial pre-crRNA cleavage. We next examined the dependence of pre-crRNA cleavage on divalent metal ions. We observed that the generation of mature crRNA for both EsCas13d and RspCas13d is substantially inhibited by the addition of EDTA (Figures 4B, S5B–C), while Cas13a from Leptotrichia wadei (LwaCas13a) is still able to generate mature crRNAs in the presence of EDTA (Figure S5D). This dependence of Cas13d on divalent cations to generate mature crRNA is a notable functional distinction from Cas13a crRNA processing (East-Seletsky et al., 2016; Knott et al., 2017).

RspCas13d and EsCas13d activity using strongly depleted arrays

We next sought to biochemically validate the RNA-guided ssRNA cleavage activities of the Cas13d enzymes observed in our bacterial screens. Cas13a-c have been shown to mediate cleavage of ssRNA targets specified by the crRNA spacer (on-target RNA cleavage). Hybridization of the Cas13-crRNA complex to a target RNA also activates non-targeted ssRNA cleavage (collateral RNA cleavage) (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Liu et al., 2017b, 2017a; Smargon et al., 2017). To test both of these activities for RspCas13d and EsCas13d, we identified spacer sequences for several strongly depleted arrays from bacterial screens for each CRISPR-Cas system and generated pre-crRNAs with the repeat-spacer-repeat arrangement for each effector. We targeted EsCas13d and RspCas13d enzyme-crRNA complexes to 130nt ssRNA substrates containing target sequences complementary to the crRNA spacer and demonstrated targeted RNA cleavage activity for both enzymes (Figure 5A). To evaluate the collateral RNA cleavage activity, identical reactions were prepared and supplemented with 800nt fluorescent body-labeled ssRNA fragments that did not contain the target sequence. Both EsCas13d and RspCas13d showed substantial collateral activity that occurs with the target cleavage (Figure 5B). We further demonstrated that both EsCas13d and RspCas13d show robust targeted and collateral

RNA cleavage activities across a wide dose range of enzyme-crRNA complex, and in a target-specific manner (Figure S6).

Modulation of Cas13d activity by a WYL domain-containing accessory protein

Putative accessory proteins containing WYL domains and additional predicted DNA-binding domains are present in the great majority of the subtype VI-D loci (Figure 1A). We initially synthesized and screened the predicted minimal CRISPR-Cas system for RspCas13d including both the RspCas13d effector and RspWYL1 accessory protein. To investigate the modulation of Cas13d by WYL1, we screened both the RspCas13d effector and RspWYL1 accessory protein separately. Comparison of screening results for RspCas13d effector alone versus the RspCas13d system, including RspWYL1, shows that RspCas13d targeted RNA cleavage is increased in the presence of RspWYL1 (Figure 6A–B). Bacterial screening with RspWYL1 alone yielded a minimal number of hits, indicating that RspWYL1 has no individual activity (Figure 6C). Cumulatively, these results suggest that RspCas13d enzymatic activity is modulated either directly or indirectly by WYL1.

We further investigated whether WYL1 could modulate RspCas13d *in vitro* by purifying recombinant RspWYL1 for use in ssRNA cleavage biochemical assays. To enable high resolution of enhanced or decreased complex activity in the presence of WYL, we selected doses of Cas13d-crRNA complex resulting in approximately 50% cleavage of the target substrates based on a dose titration curve (Figure S6). We pre-incubated Cas13d-crRNA with no RspWYL1, an equimolar ratio of RspWYL1 to Cas13d, or a molar excess of RspWYL1 over Cas13d, and the resulting samples were incubated with target and collateral ssRNA under the same conditions as in the target cleavage assays. We observed that RspWYL1 increases both the targeted and collateral ssRNA cleavage activity of RspCas13d in a dose-dependent manner, with a molar excess of RspWYL1 yielding the greatest increase in Cas13d activity (Figures 7A, S7).

Given that subtype VI-D CRISPR-Cas systems appear to have acquired WYL-domain containing accessory proteins on multiple, independent occasions (Figure S2a), we tested the specificity of RspWYL1 in modulating the cleavage activity of orthologous Cas13d effectors. We observed that RspWYL1 enhanced the targeted and collateral ssRNA nuclease activities of EsCas13d to a similar extent as observed for RspCas13d (Figure 7B). Thus, the effect of WYL1 orthologs appear not to be limited to their native effectors but instead reflect a modular regulatory mechanism for Cas13d effectors.

DISCUSSION

The structural and functional diversity of class 2 CRISPR-Cas systems provides versatility and flexibility of genome editing and nucleic acid detection tools. In this work, we investigated a previously uncharacterized subtype, VI-D, with the Cas13d effector that differs from the previously characterized Cas13a and Cas13b in at least three significant respects: a substantially smaller size, the lack of appreciable sequence constraints on the target flanking sequences, and an alternative mechanism for crRNA processing. The Cas13d effectors showed robust target cleavage and collateral RNase activity, which could make them promising candidates for new, highly efficient molecular tools that would be

particularly suitable for *in vivo* delivery due to the comparatively small size of these proteins.

Apart from potential applications, the small size of Cas13d poses an interesting, fundamental question: Cas13d is capable of efficient pre-crRNA processing yielding mature crRNAs but lacks a counterpart to the domain that has been shown to be responsible for processing in Cas13a (Liu et al., 2017a). The sequence of this domain is poorly conserved within the Cas13a family itself which, together with the present results on Cas13d, implies substantial structural flexibility of the pre-crRNA processing activity. Furthermore, unlike the cases of Cas13a and Cas13b (East-Seletsky et al., 2016; Knott et al., 2017), the pre-crRNA processing activity of Cas13d depends on divalent cations, suggestive of a distinct biochemical mechanism. Indeed, EsCas13d and RspCas13d may have different mechanisms and dependencies of crRNA maturation, as they appear to generate different intermediate forms of pre-crRNA in response to the addition of EDTA (Figure 4B), even if neither generates mature crRNA in this condition. Subsequent experiments aimed at the detailed characterization of the processing activity of Cas13d should shed light on the mechanistic diversity of the dual RNase activity of the type VI effectors.

An additional interesting observation made in the course of this work is the stimulation of the Cas13d effector activity by the WYL and RHH domain-containing accessory protein WYL1. Previously, it has been shown that the RNA cleavage activity of Cas13b2 is stimulated by Csx28 (Smargon et al., 2017). However, the Csx28 protein lacks the WYL domain and is unrelated to WYL1 or any other putative accessory proteins in the subtype VI-D systems, emphasizing the diversity of the mechanisms of effector modulation in type VI CRISPR-Cas systems. Previous work on the WYL-domain containing CRISPR accessory proteins for a subtype I-D system has demonstrated negative regulation via transcriptional repression (Hein et al., 2013). Thus, the WYL-domain containing accessory protein WYL1 either has the opposite, activating effect on the target transcription, or else, directly modulates nuclease activity suggesting a transcription-independent effect. Recently, it has been shown that a WYL domain in a Pif1 helicase (not associated with CRISPR-Cas) binds ssDNA with a high affinity and stimulates the helicase activity (Andis et al., 2018). Further biochemical experiments should shed light on the mechanism of Cas13d regulation by WYL-domain containing accessory proteins.

So far, focused searches for class 2 CRISPR-Cas systems in genomic and metagenomic databases have revealed only 3 types (II, V and VI) with distinct domain architectures of the effector nucleases. However, within each type, there are multiple subtypes and variants that show substantial structural and functional flexibility. Investigation of the properties of these diverse CRISPR-Cas systems provides for better understanding of the biology of microbial adaptive immunity and for development of new tools for molecular biology.

STAR METHODS

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contacts, David Scott (dscott@arbor.bio) and David Cheng

(dcheng@arbor.bio). The authors plan to make the reagents widely available to the academic community through Addgene subject to a MTA.

Experimental Model and Subject Details

Endura electrocompetent E. coli—*E. coli* were electroporated according to the manufacturer's protocols. After mixing 25uL of thawed cells with DNA, the *E. coli* were electroporated with a Bio-Rad Gene Pulser Xcell (Bio-rad) using a 1.0mm cuvette at settings of 10 uF, 600 Ohms, and 1800 Volts. 975 uL of Recovery Media (Lucigen) were added directly after the pulse, which were then shaken for 1 hour at 37°C at 250 rpm.

NEB 5-alpha Competent E. coli (High Efficiency)—Following transformation and outgrowth according to the manufacturer's protocols, the *E. coli* were plated onto LB agar with appropriate antibiotic selection and grown at 37°C overnight.

NEB NiCo21(DE3)—Expression vectors for protein purification (Key Resources Table) were grown in the *E. coli* T7 expression strain, NiCo21(DE3) (New England Biolabs). 1mL of overnight culture was inoculated into 1 liter of Luria-Bertani broth growth media (10g/L tryptone, 5 g/L yeast extract, 5g/L NaCl, Sigma) supplemented with 50 μg/mL Kanamycin. Cells were grown at 37°C to a cell density of 0.5–0.8 OD600. Protein expression was then induced by supplementing with IPTG to a final concentration of 0.2 mM and the culture continued to grow for 14–18 hours at 20°C.

Method Details

Pipeline for Class 2 CRISPR-Cas loci identification—Genome and metagenome sequences were downloaded from NCBI (Benson et al., 2013; Pruitt et al., 2012), NCBI whole genome sequencing (WGS), and DOE JGI Integrated Microbial Genomes (Markowitz et al., 2012). Proteins were predicted (Meta-GeneMark (Zhu et al., 2010) using the standard model MetaGeneMark_v1.mod, and Prodigal (Hyatt et al., 2010) in anon mode) on all contigs at least 5kb in length, and de-duplicated in favor of pre-existing annotations to construct a complete protein database. CRISPR arrays were identified and protein sequences for ORFs located within +/− 10kb from CRISPR arrays were grouped into CRISPR-proximal protein clusters. Clusters of fewer than 4 proteins, or comprising proteins from fewer than 3 contigs were discarded. Each of these remaining protein clusters were considered to be a putative effector of a CRISPR-Cas system. In addition to the CRISPR array and putative effector protein, many CRISPR-Cas systems also include additional proteins which enable adaptation, crRNA processing, and defense. Potential additional CRISPR-Cas system components associated with each of the predicted effectors were identified as clusters of protein-coding genes with high effector co-occurrence, and CRISPR enrichment or CRISPR representation of at least 15%.

Effector co-occurrence was calculated as the percentage of loci containing the effector that also contain the potential co-occurring protein. The high co-occurrence threshold was a function of the cohesiveness of the effector cluster (more homogenous clusters requiring a higher threshold). The CRISPR enrichment was calculated as follows: 1) Up to 20 unique proteins were sampled from each protein cluster, and UBLAST (Edgar, 2010) was used to

generate a rank ordered list of proteins by E-value from the complete protein database. 2) An E-value threshold was imposed to recover at least 50% of the members of the cluster. 3) CRISPR enrichment was calculated by dividing the number of CRISPR-proximal proteins below the E-value threshold by the total number of proteins below the threshold. CRISPR representation was calculated as the percentage of effector-proximal proteins in a CRISPR-proximal protein cluster. All clustering operations were performed using mmseqs2 (Steinegger and Söding, 2017).

This information was incorporated into a database of (predicted) CRISPR-Cas systems, each composed of: 1) a CRISPR array, 2) a putative effector, and optionally, 3) clusters of potential co-acting proteins. For functional characterization of this database of candidate CRISPR-Cas systems, we constructed multiple sequence alignment for each family of putative effectors using MAFFT (Katoh and Standley, 2013) and conducted an HMM search using HMMer (Eddy, 2011) against protein family databases Pfam (Finn et al., 2014) and Uniprot (Bateman et al., 2017), as well as a BLASTN search of CRISPR spacer sequences against a reference set of phages. This analysis led to the detection of protein families corresponding to all previously identified class 2 CRISPR-Cas systems, indicating a minimal false negative rate. To identify novel class 2 CRISPR-Cas systems, features included above for the prediction of the functions of putative CRISPR-Cas systems were used to rank candidate families for follow-up functional evaluation.

Phylogenetic analysis—Maximum likelihood trees were constructed using FastTree (Price et al., 2010). For the phylogenetic analysis of Cas1, all Cas1 sequences that were assigned to Type II and Type VI-A in the course of previous work (Shmakov et al., 2017a) were used, and all Cas1 sequences associated with Cas13d were added. Altogether, a set of 817 Cas1 sequences was employed for phylogenetic analysis (Data File S1).

For the WYL family analysis, in addition to automatically identified WYL proteins, we used PSI-BLAST (Altschul et al., 1997) to search over a local set of NCBI sourced proteins using RspWYL1 as a query. The results with E-value 0.01 or lower were added to the set of WYL proteins. Proteins smaller than 150 aa were discarded from the data set, and UCLUST (Edgar, 2010) with identity threshold 0.90 was used to obtain a non-redundant set. We then added all WYL proteins identified in the vicinity of Cas13d genes to form a set of 3908 WYL sequences for phylogenetic analysis.

Multiple alignment and phylogeny of protein sequences were constructed as described previously (Peters et al., 2017). Briefly, the sequences were clustered by similarity, and for each cluster, a multiple alignment was built using MUSCLE (Edgar, 2004). Alignments were combined into larger aligned clusters by HHalign (Yu et al., 2015) if the resulting score between the two alignments was higher than the threshold; otherwise, the scores were recorded in a similarity matrix. The matrix was used to reconstruct a UPGMA tree. For each cluster, the alignment was filtered as follows: the alignment positions with the gap character fraction values of 0.5 and homogeneity values of 0.1 or less were removed. The remaining positions were used for tree reconstruction using FastTree with the WAG evolutionary model and the discrete gamma model with 20 rate categories. The same program was used to compute SH (Shimodaira-Hasegawa)-like node support values (Data Files S2, S3).

Spacer Analysis—Spacer sequences from CRISPR arrays within 3kb of Cas13d effectors were extracted. In the case of multiple contigs containing the same Cas13d sequence (eg duplicated locus), only the contig containing the longest CRISPR array was used. Subsequent spacer analysis closely follows the method described previously (Shmakov et al., 2017b). Briefly, the resulting 198 spacers were de-duplicated by comparison of direct and reverse complement sequences, to produce a set of 182 unique spacers. A BLASTN search with the command line parameters - word_size 7 -gapopen 5 -gapextend 2 -reward 1 -penalty -3 was performed for the unique spacer set against a database comprising the virus and prokaryotic sequences in NCBI. To identify prophage regions, (i) all ORFs within 3kb of prokaryotic matches were collected; (ii) a PSI-BLAST search was conducted against the proteins extracted from the virus part of NCBI, using the command line parameters -seg no -evalue 0.000001 -dbsize 20000000; (iii) a spacer hit was classified as prophage if it overlapped with an ORF with a viral match, or if two or more ORFs with viral matches were identified within the neighborhood of the spacer hit.

DNA synthesis & effector library cloning—The *E. coli* codon-optimized genes representing the minimal CRISPR effectors and accessory proteins were synthesized (Genscript) into a custom expression system derived from the pET-28a(+) (EMD-Millipore). Briefly, the *Ruminococcus sp*. synthesis product included Cas13d and WYL1 codon optimized for *E. coli* expression under the control of a Lac promoter and separated by an *E. coli* ribosome binding sequence. Following the open reading frames for Cas13d and WYL1, we included an acceptor site for a CRISPR array library driven by a J23119 promoter (Registry of Standard Biological Parts: http://parts.igem.org/Part:BBa_J23119). Our *Eubacterium siraeum* system was similarly constructed, but with only the effector protein.

In tandem with the effector gene synthesis, we first computationally designed an oligonucleotide library synthesis (OLS) pool containing “repeat-spacer-repeat” sequences, where “repeat” represents the consensus direct repeat sequence found in the CRISPR array associated with the effector, and “spacer” represents sequences tiling the pACYC184 plasmid. The spacer length was determined by the mode of the spacer lengths found in the endogenous CRISPR array. The repeat-spacer-repeat sequence was appended with restriction sites enabling the bi-directional cloning of the fragment into the aforementioned CRISPR array library acceptor site, as well as unique PCR priming sites to enable specific amplification of a specific repeat-spacer-repeat library from a larger pool. The library synthesis was performed by Agilent Genomics.

We next cloned the repeat-spacer-repeat library into the plasmid containing the minimal engineered locus using the Golden Gate assembly method. In brief, we first amplified each repeat-spacer-repeat from the OLS pool (Agilent Genomics) using unique PCR primers, and pre-linearized the plasmid backbone using BsaI to reduce potential background. Both DNA fragments were purified with Ampure XP (Beckman Coulter) prior to addition to Golden Gate Assembly Master Mix (New England Biolabs) and incubated as per manufacturer’s instructions. We further purified and concentrated the Golden Gate reaction to enable maximum transformation efficiency in the subsequent steps of the bacterial screen.

Bacterial screening for effector activity—The plasmid library containing the distinct repeat-spacer-repeat elements and Cas proteins was electroporated into Endura electrocompetent *E. coli* (Lucigen) using a Gene Pulser Xcell (Bio-rad) following the protocol recommended by Lucigen. The library was either co-transformed with purified pACYC184 plasmid, or directly transformed into pACYC184-containing Endura electrocompetent *E. coli* (Lucigen), plated onto agar containing Chloramphenicol (Fisher), Tetracycline (Alfa Aesar), and Kanamycin (Alfa Aesar) in BioAssay dishes (Thermo Fisher), and incubated for 10–12h. After estimation of approximate colony count to ensure sufficient library representation on the bacterial plate, the bacteria were harvested and DNA plasmid extracted using a QIAprep Spin Miniprep Kit (Qiagen) to create the 'output library'. By performing a PCR using custom primers containing barcodes and sites compatible with Illumina sequencing chemistry, we generated a barcoded next generation sequencing library from both the pre-transformation 'input library' and the post-harvest 'output library', which were then pooled and loaded onto a Nextseq 550 (Illumina) to evaluate the effectors. At least two independent biological replicates were performed for each screen to ensure consistency.

Bacterial screen sequencing analysis—Next generation sequencing data for screen input and output libraries were demultiplexed using Illumina bcl2fastq. Reads in resulting fastq files for each sample contained the CRISPR array elements for the screening plasmid library. The direct repeat sequence of the CRISPR array was used to determine the array orientation, and the spacer sequence was mapped to the source plasmid pACYC184 or negative control sequence (GFP) to determine the corresponding target. For each sample, the total number of reads for each unique array element (r_a) in a given plasmid library was counted and normalized as follows: (r_a+1)/total reads for all library array elements. The depletion score was calculated by dividing normalized output reads for a given array element by normalized input reads.

PFS and sequence constraint determination—We want to determine if a subset of nucleotide positions in the region of the targeting area can explain strongly depleted targets. To this end, we define a targeting requirement to comprise a set of locations relative to a target sequence (Figure S4a) and the corresponding nucleotide sequences at those locations. For a given targeting requirement, we define the hit ratio (hr) as the ratio of the number of strongly depleted CRISPR arrays to the total number of library targets satisfying the requirement. When searching for a PAM or PFS of length k, we consider $\binom{n}{k}$ potential targeting requirement locations, where n = spacer length + 2 · flank length. The bit score for a potential targeting requirement is calculated as $bitscore = \Sigma\ -hr\log(hr)$ over all nucleotide sequences at the specified targeting requirement locations.

Effector and accessory protein purification—The effector or accessory protein expression construct was transformed into an *E. coli* T7 expression strain, NiCo21(DE3) (New England Biolabs) and grown as described in the Experimental Models and Subject Details section of the STAR methods. The cells were harvested by centrifugation and cell paste was resuspended in 80 ml of freshly prepared Lysis Buffer (50 mM Hepes pH 7.6, 0.5M NaCl, 10 mM imidazole, 14 mM 2-mercaptoethanol and 5% glycerol) supplemented

with protease inhibitors (cOmplete, EDTA-free, Roche Diagnostics Corporation). The resuspended cells were broken by passing through a cell disruptor (Constant System Limited). Lysate was cleared by centrifugation twice at 28,000g for 30 min each. The clarified lysate was applied to a 5 ml HisTrap FF chromatography column (GE Life Sciences). Protein purification was performed via FPLC (AKTA Pure, GE Healthcare Life Sciences). After washing with Lysis Buffer, protein was eluted with a gradient of 10 mM to 250 mM of imidazole. Fractions containing protein of the expected size were pooled, concentrated in Vivaspin 20 ultrafiltration unit (Sartorius) and either used directly for biochemical assays or frozen at −80°C for storage. Protein purity was determined by SDS-PAGE analysis and protein concentration was determined by Qubit protein assay kit (Thermo Fisher).

crRNA and substrate RNA preparation—DNA oligo templates for crRNA and substrate RNA *in vitro* transcription were ordered from IDT (Tables S1b). Templates for crRNAs were annealed to a short T7 primer (final concentrations 4μM) and incubated with T7 RNA polymerase overnight at 37°C using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). Annealing was performed by incubating T7 primer with templates for 2 minutes at 95°C foll owed by a −5°C/s ramp down to 23°C. Templates for substrate RNA were PCR amplified to yield dsDNA and then incubated with T7 RNA polymerase at 37°C overnight using the same T7 Quick High Yield RNA Synthesis kit. After *in vitro* transcription, samples were treated with DNase I (Zymo Research) and then purified using RNA Clean & Concentrator kit (Zymo Research).

5’ end labeling was accomplished using the 5’ end labeling kit (VectorLabs) and with a IR800 dye-maleimide probe (LI-COR Biosciences). Body labeling of RNA was performed during *in vitro* transcription using the HiScribe T7 Quick High Yield RNA Synthesis kit (New England Biolabs). The *in vitro* transcription reactions contained 2.5 mM Fluorescein-12-UTP (Sigma Aldrich). Labeled RNA was purified to remove excess dyes using RNA Clean & Concentrator kit (Zymo Research). The RNA concentration was measured on Nanodrop 2000 (Thermo Fisher).

Pre-crRNA processing assays—Pre-crRNA cleavage assays were performed at 37°C in processing buffer (20 mM Tris pH8.0, 50 mM KCl, 1 mM EDTA, 10mM MgCl2, and 100 ug/ml BSA) unless otherwise indicated, with a final reaction concentration of 200nM of pre-crRNA and varying enzyme concentrations and EDTA as indicated. Reactions were incubated for 30 minutes, and quenched with the addition of 1ug/uL of proteinase K (Ambion) incubated for 10 minutes at 37°C. Afterwards, 50mM of EDTA was added to the reaction, which was then mixed with equal parts 2× TBE-Urea Sample Buffer (Invitrogen) prior to denaturing at 65C for 3 minutes. Samples were analyzed by denaturing gel electrophoresis on 15% TBE-Urea gels (Invitrogen) and stained using SYBR Gold nucleic acid stain (Invitrogen) for 10–20 minutes prior to imaging on a Gel Doc EZ (Biorad).

RNA sequencing—Sequencing of in vitro cleaved pre-crRNA began with performing and quenching the cleavage assays as described above. The reactions were then column purified using a RNA Clean and Concentrator-5 kit (Zymo Research). The RNA samples were then PNK treated for 3 hours without ATP to enrich for 3’-P ends, after which ATP was added

and the reaction incubated for another hour to enrich for 5'-OH ends. The samples were then column purified, incubated with RNA 5' polyphosphatase (Lucigen) and column purified again prior to preparation for next-generation sequencing using the NEBNext Multiplex Small RNA Library Prep Set for Illumina (New England Biolabs). The library was paired-end sequenced on a Nextseq 550 (Illumina), and the resulting paired end alignments were analyzed using Geneious 11.0.2 (Biomatters).

Sequencing the small RNA from the *in vivo* bacterial screen began by extracting total RNA from harvested screen bacteria using the Direct-zol RNA MiniPrep Plus w/TRI Reagent (Zymo Research). Ribosomal RNA was removed using a Ribo-Zero rRNA Removal Kit for Bacteria, followed by cleanup using a RNA Clean and Concentrator-5 kit. The resultant ribosomal RNA depleted total RNA was treated with T4 PNK, RNA 5' polyphosphatase, prepared for sequencing using the NEBNext Small RNA Library Prep Set, and analyzed as described above.

Target cleavage assays—Target cleavage assays were performed at 37°C in cl eavage buffer (20 mM HEPES pH 7.1, 50 mM KCl, 5 mM MgCl2 and 5% glycerol). Cas13-crRNA complex formation was performed in cleavage buffer by incubating a 2:1 molar ratio of protein to crRNA at 37°C for 5 minutes, and RspWYL1 was added to the Cas13-crRNA pre-incubation according to the experimental conditions. For the cleavage reactions at different Cas13 concentrations, the pre-formed Cas13-crRNA complexes were diluted on ice, keeping the Cas13-crRNA ratio constant at 2:1. The 5' IR800 labeled target ssRNA and/or additional unlabeled and fluorescent body-labeled ssRNAs were then added to the pre-formed complex and incubated at 37°C for 30 minutes. The final concentration of short substrate RNAs was 100nM and the fluorescent body-labeled ssRNA for collateral effect visualization was 50nM, unless otherwise indicated. Reactions were quenched by adding 1ug/uL of proteinase K (Ambion) and incubating for 10 minutes at 37°C. Afterwards, 50mM of EDTA was added to the reaction, which was then mixed with equal parts 2× TBE-Urea Sample Buffer (Invitrogen) prior to denaturing at 65°C for 3 minutes. Samples were analyzed by denaturing gel electrophoresis on 6% or 15% TBE-Urea gels (Invitrogen). Fluorescence images were obtained using a Gel Doc EZ (Biorad), and near-infrared images were obtained using an Odyssey CLx scanner (LI-COR Biosciences). Afterwards, the gels were stained for 10–20 minutes using SYBR Gold nucleic acid stain (Invitrogen) and imaged on the Gel Doc EZ to verify the results from the fluorescence and IR images.

QUANTIFICATION AND STATISTICAL ANALYSIS

Bacterial screen sequencing analysis—To analyze primary screening results (Figures 2, 6, S3), we calculated an inverted depletion score as normalized input reads/normalized output reads for each repeat-spacer-repeat. Note that in this formulation, a score of 1 represents no change in relative representation of a repeat-spacer-repeat element, and 10 represents a normalized 10-fold decrease in representation. For the effector deletion condition, the primary screening experiment was performed using pET28a(+) vectors that contain a repeat-spacer-repeat cloned from a library of N elements, with N = 10,002 for EsCas13d, and N = 8,844 for RspCas13d. Mean and standard deviations of inverted depletion scores were calculated as μ = 1.01 and 1.04; σ = 0.44 and 0.46 for two biological

replicates of EsCas13d, and $\mu = 1.03$ and 1.04; $\sigma = 0.58$ and 0.69 for two biological replicates of RspCas13d. Setting a minimum inverted depletion score threshold of 10 (maximum depletion score of 0.1 as defined in the main text) represents a deviation greater than 10 standard deviations from negative control conditions.

DATA AND SOFTWARE AVAILABILITY

Data have been deposited in the following resources: Next-Generation Sequencing for bacterial DNA-sequencing and RNA-sequencing of *E. coli* primary screens, and RNA-sequencing of *in vitro* pre-crRNA processing: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434567

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the team at Arbor Biotechnologies for supporting this work. K.S.M. and E.V.K. are supported by the intramural program of the U.S. Department of Health and Human services (to the National Library of Medicine). A patent application has been filed related to this work, and authors plan to make the reagents widely available to the academic community through Addgene.

References

Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DBT, Shmakov S, Makarova KS, Semenova E, Minakhin L, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science. 2016; 353:aaf5573. [PubMed: 27256883]

Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, Belanto JJ, Verdine V, Cox DBT, Kellner MJ, Regev A, et al. RNA targeting with CRISPR–Cas13. Nature. 2017; 550:280–284. [PubMed: 28976959]

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. Nat. Rev. Microbiol. 2016; 14:67–76. [PubMed: 26751509]

Andis NM, Sausen CW, Alladin A, Bochman ML. The WYL Domain of the PIF1 Helicase from the Thermophilic Bacterium Thermotoga elfii is an Accessory Single-Stranded DNA Binding Module. Biochemistry (Mosc.). 2018

Barrangou R, Horvath P. A decade of discovery: CRISPR functions and applications. Nat. Microbiol. 2017; 2:17092. [PubMed: 28581505]

Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45:D158–D169. [PubMed: 27899622]

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2013; 41:D36–42. [PubMed: 23193287]

Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, Doudna JA, Banfield JF. New CRISPR-Cas systems from uncultivated microbes. Nature. 2017; 542:237–241. [PubMed: 28005056]

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. [PubMed: 20003500]

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013; 339:819–823. [PubMed: 23287718]

Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, Zhang F. RNA editing with CRISPR-Cas13. Science. 2017; 358:1019–1027. [PubMed: 29070703]

Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–1190. [PubMed: 15173120]

East-Seletsky A, O'Connell MR, Knight SC, Burstein D, Cate JHD, Tjian R, Doudna JA. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. Nature. 2016; 538:270–273. [PubMed: 27669025]

East-Seletsky A, O'Connell MR, Burstein D, Knott GJ, Doudna JA. RNA Targeting by Functionally Orthogonal Type VI-A CRISPR-Cas Enzymes. Mol. Cell. 2017; 66:373–383.e3. [PubMed: 28475872]

Eddy SR. Accelerated Profile HMM Searches. PLoS Comput. Biol. 2011; 7:e1002195. [PubMed: 22039361]

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26:2460–2461. [PubMed: 20709691]

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–D230. [PubMed: 24288371]

Fonfara I, Richter H, Bratovič M, Le Rhun A, Charpentier E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature. 2016; 532:517–521. [PubMed: 27096362]

Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. 2012; 109:E2579–E2586. [PubMed: 22949671]

Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia N, Daringer NM, Freije CA, et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science. 2017; 356:438–442. [PubMed: 28408723]

Hein S, Scholz I, Voß B, Hess WR. Adaptation and modification of three CRISPR loci in two closely related cyanobacteria. RNA Biol. 2013; 10:852–864. [PubMed: 23535141]

Hsu PD, Lander ES, Zhang F. Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell. 2014; 157:1262–1278. [PubMed: 24906146]

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010; 11:119. [PubMed: 20211023]

Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJJ. CRISPR-Cas: Adapting to change. Science. 2017; 356

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science. 2012; 337:816–821. [PubMed: 22745249]

Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 2013; 30:772–780. [PubMed: 23329690]

Knott GJ, East-Seletsky A, Cofsky JC, Holton JM, Charles E, O'Connell MR, Doudna JA. Guide-bound structures of an RNA-targeting A-cleaving CRISPR–Cas13a enzyme. Nat. Struct. Mol. Biol. 2017; 24:825–833. [PubMed: 28892041]

Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. 2017; 37:67–78. [PubMed: 28605718]

Liu L, Li X, Wang J, Wang M, Chen P, Yin M, Li J, Sheng G, Wang Y. Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. Cell. 2017a; 168:121–134.e12. [PubMed: 28086085]

Liu L, Li X, Ma J, Li Z, You L, Wang J, Wang M, Zhang X, Wang Y. The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. Cell. 2017b; 170:714–726.e10. [PubMed: 28757251]

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol. Biol. 2011; 6:26. [PubMed: 22115189]

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

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. 2015; 13:722–736. [PubMed: 26411297]

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. Science. 2013; 339:823–826. [PubMed: 23287722]

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012; 40:D115–122. [PubMed: 22194640]

Marraffini LA. CRISPR-Cas immunity in prokaryotes. Nature. 2015; 526:55–61. [PubMed: 26432244]

Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, van der Oost J. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science. 2016; 353:aad5147. [PubMed: 27493190]

van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat. Rev. Microbiol. 2014; 12:479–492. [PubMed: 24909109]

Peters JE, Makarova KS, Shmakov S, Koonin EV. Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proc. Natl. Acad. Sci. U. S. A. 2017; 114:E7358–E7366. [PubMed: 28811374]

Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010; 5:e9490. [PubMed: 20224823]

Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012; 40:D130–135. [PubMed: 22121212]

Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, et al. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol. Cell. 2015; 60:385–397. [PubMed: 26593719]

Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, Gootenberg JS, Makarova KS, Wolf YI, et al. Diversity and evolution of class 2 CRISPR-Cas systems. Nat. Rev. Microbiol. 2017a; 15:169–182. [PubMed: 28111461]

Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. MBio. 2017b; 8:e01397–17. [PubMed: 28928211]

Smargon AA, Cox DBT, Pyzocha NK, Zheng K, Slaymaker IM, Gootenberg JS, Abudayyeh OA, Essletzbichler P, Shmakov S, Makarova KS, et al. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. Mol. Cell. 2017; 65:618–630.e7. [PubMed: 28065598]

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. 2017

Wright AV, Nuñez JK, Doudna JA. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell. 2016; 164:29–44. [PubMed: 26771484]

Yu J, Picord G, Tuffery P, Guerois R. HHalign-Kbest: exploring sub-optimal alignments for remote homology comparative modeling. Bioinforma. Oxf. Engl. 2015; 31:3850–3852.

Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell. 2015; 163:759–771. [PubMed: 26422227]

Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010; 38:e132–e132. [PubMed: 20403810]

Highlights

- Type VI-D is a CRISPR-Cas system with a Cas13d effector and a WYL-domain accessory [84 char]
- Cas13d is a RNA-guided RNase approx. 20% smaller than Cas13a-c effectors [74]
- WYL1 positively modulates Cas13d target and collateral RNase activity [71]
- Cas13d has minimal sequence and secondary structure requirements for targeting [81]

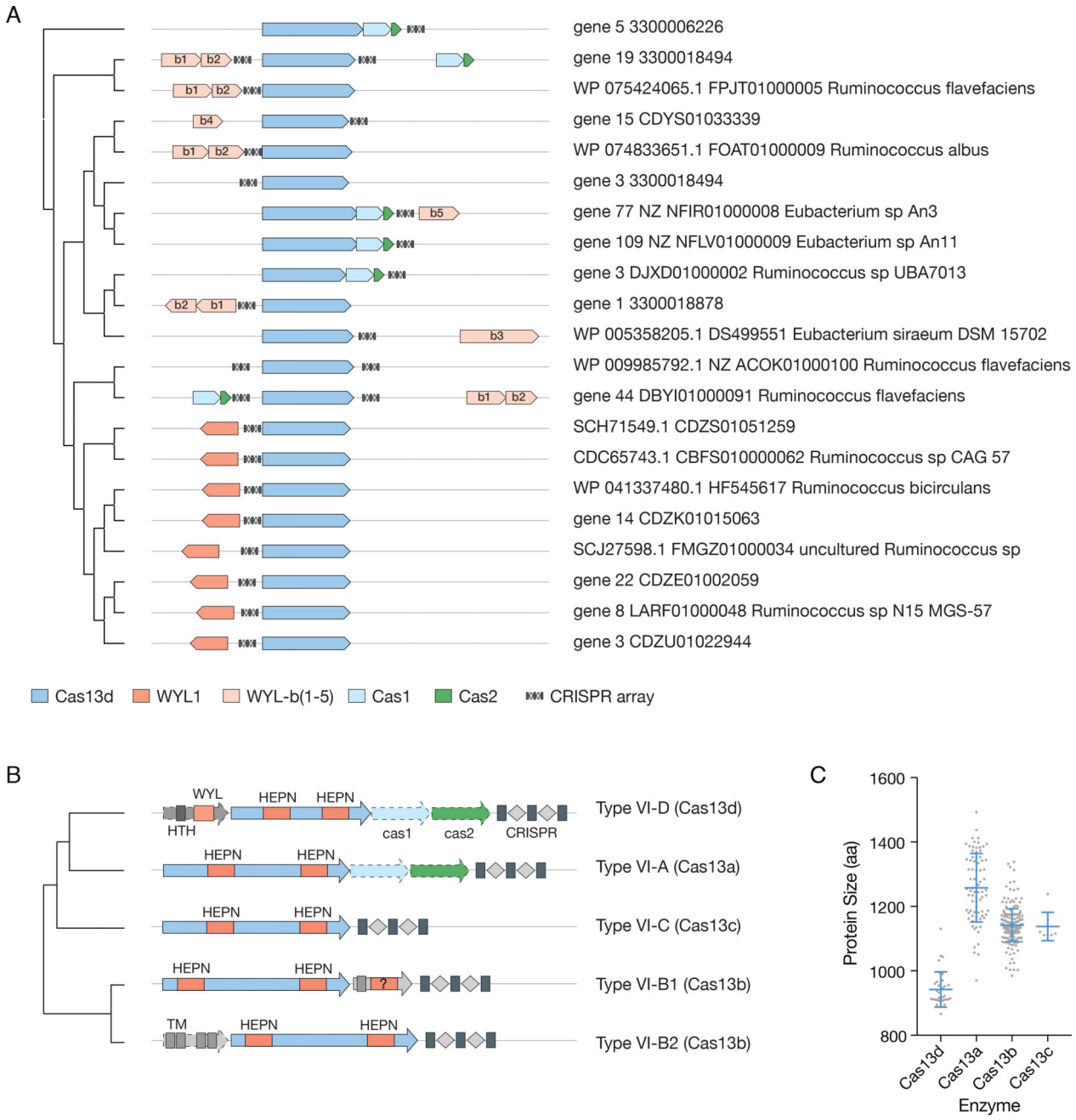


Figure 1. Subtype VI-D CRISPR-Cas systems and diversity of type VI subtypes

(A) Schematic representation of a maximum likelihood tree topology for a selected subset of Cas13d, with the genomic arrangement of the genes encoding predicted protein components of subtype VI-D system components shown to the right. Each locus sequence is identified by a protein accession or gene number, with the species name provided where available. Key proteins and CRISPR arrays are color-coded as follows: blue – Cas13d, light orange – WYL domain containing protein, light blue – Cas1, green – Cas2, dark gray/gray – CRISPR array. (B) Schematic tree comparing the different type VI subtype locus structures. Gene arrows are shown roughly proportional to size. Labels denote the following, HTH – helix-turn-helix

domain, WYL – WYL domain, HEPN – HEPN nuclease domain, TM – transmembrane domains of Csx27–28. Key proteins and CRISPR arrays are color-coded as follows: blue – Cas13d, gray – Csx accessory proteins (differentiated by colored domains), light blue – Cas1, green – Cas2, dark gray/gray – CRISPR array. Figure S1 contains the sequence alignment and phylogeny of Cas13d compared to Cas13a, and Figure S2 the same analysis for the WYL-domain proteins. (C) Size comparison for Cas13 proteins from the 4 type VI subtypes; error bars specify the mean and standard deviation.

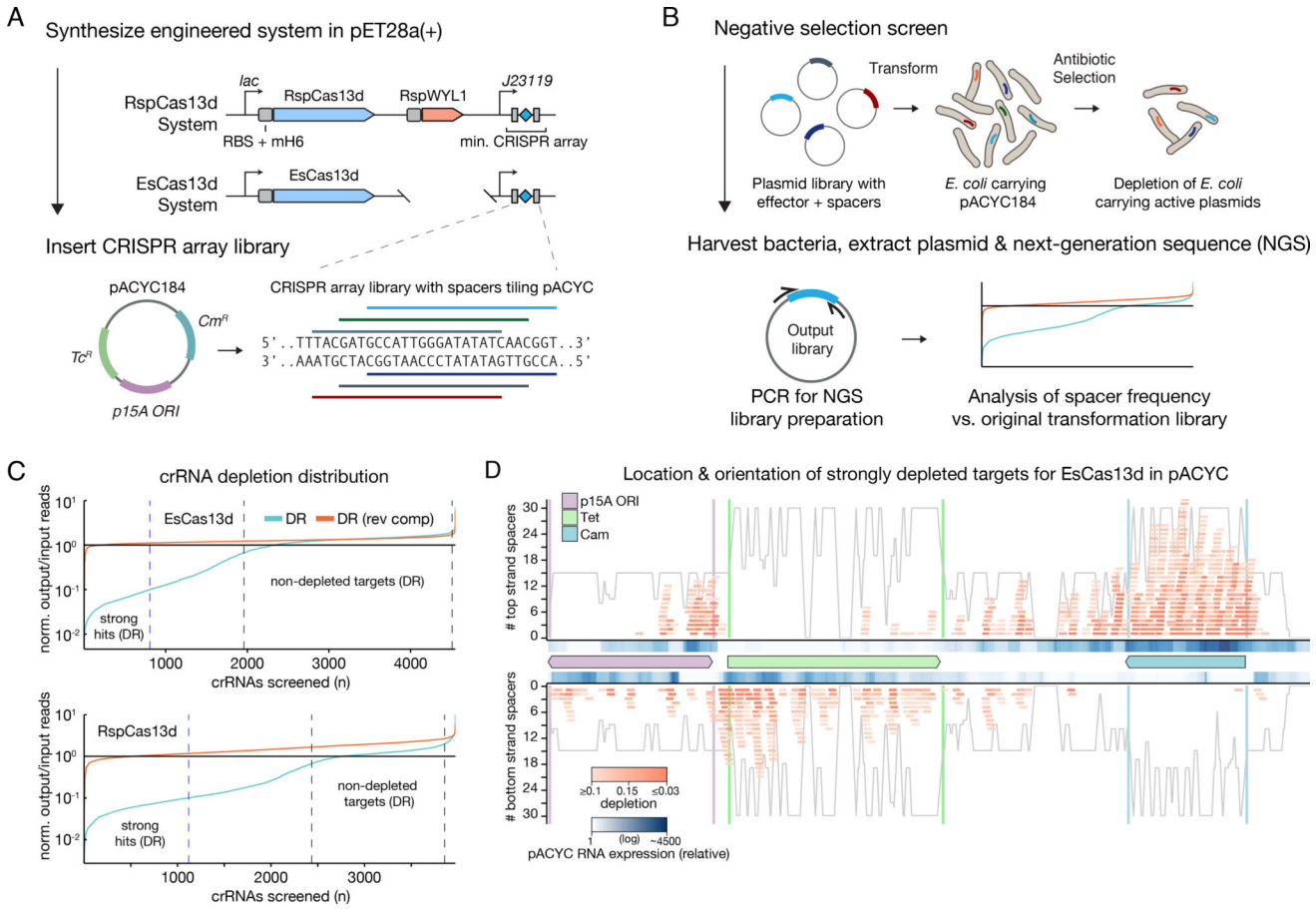


Figure 2. Synthesis of minimal CRISPR-Cas system and bacterial screening

(A) Design & synthesis of minimal engineered CRISPR-Cas systems for the Rsp and Es type VI-D CRISPR loci with a spacer library tiling pACYC184 (referred to as RspCas13d and EsCas13d systems). (B) Schematic of the bacterial negative selection screen used to evaluate functional parameters of RspCas13d and EsCas13d systems. (C) The distribution and magnitude of crRNA depletion, as calculated by normalized sequencing reads from the screen output divided by normalized reads from the pre-transformation screen input library for each crRNA spacer and orientation. Blue and orange represent both possible direct repeat (DR) orientations cloned into the screening library. The blue dashed lines demarcate the intersection of the ranked screen hits with the depletion fraction of 0.1, below which we define as strongly depleted. See Figures S3A and S3B for the negative control data. (D) Distribution of strongly depleted targets of the active DR orientation over the strands and genetic features of the pACYC184 plasmid for EsCas13d. Gray outlines represent the total number of spacers (y-axis) targeting a location, while red bars depict the locations of strongly depleted spacers with heatmap color proportional to magnitude of depletion. Directional expression data for pACYC184 is plotted as a heatmap in blue under the corresponding strand. The identical analysis on the RspCas13d bacterial screening data is presented in Figure S3C.

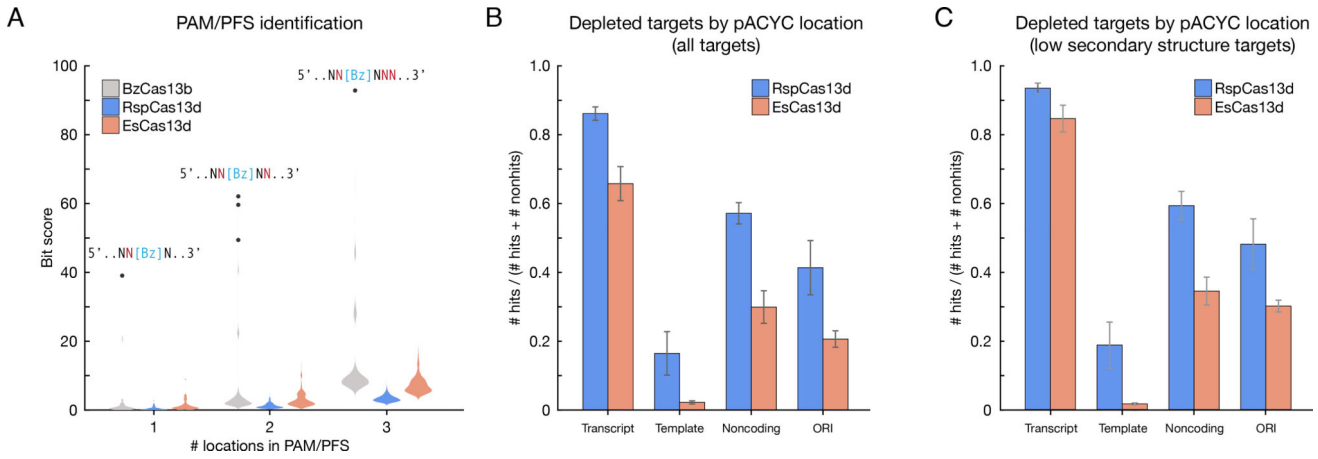


Figure 3. Cas13d activity is PFS independent, with increased efficiency for targeting sites with minimal secondary structure

(A) Violin plots of bit scores of all possible PFS targeting rules of up to length 3 involving the target site and +/− 15nt flanking region, for BzCas13b, RspCas13d, and EsCas13d systems. Dots represent data points outside of the discernable density of the violin plot. We find that these dots accurately recapitulate the known PFS positions of BzCas13b, as shown above the dots. (B & C) Breakdown of fraction of hits for RspCas13d and EsCas13d systems according to features of the plasmid for (B) all targets and (C) targets that have low secondary structure; error bars represent the upper and lower bound of two screen biological replicates. See also Figure S4.

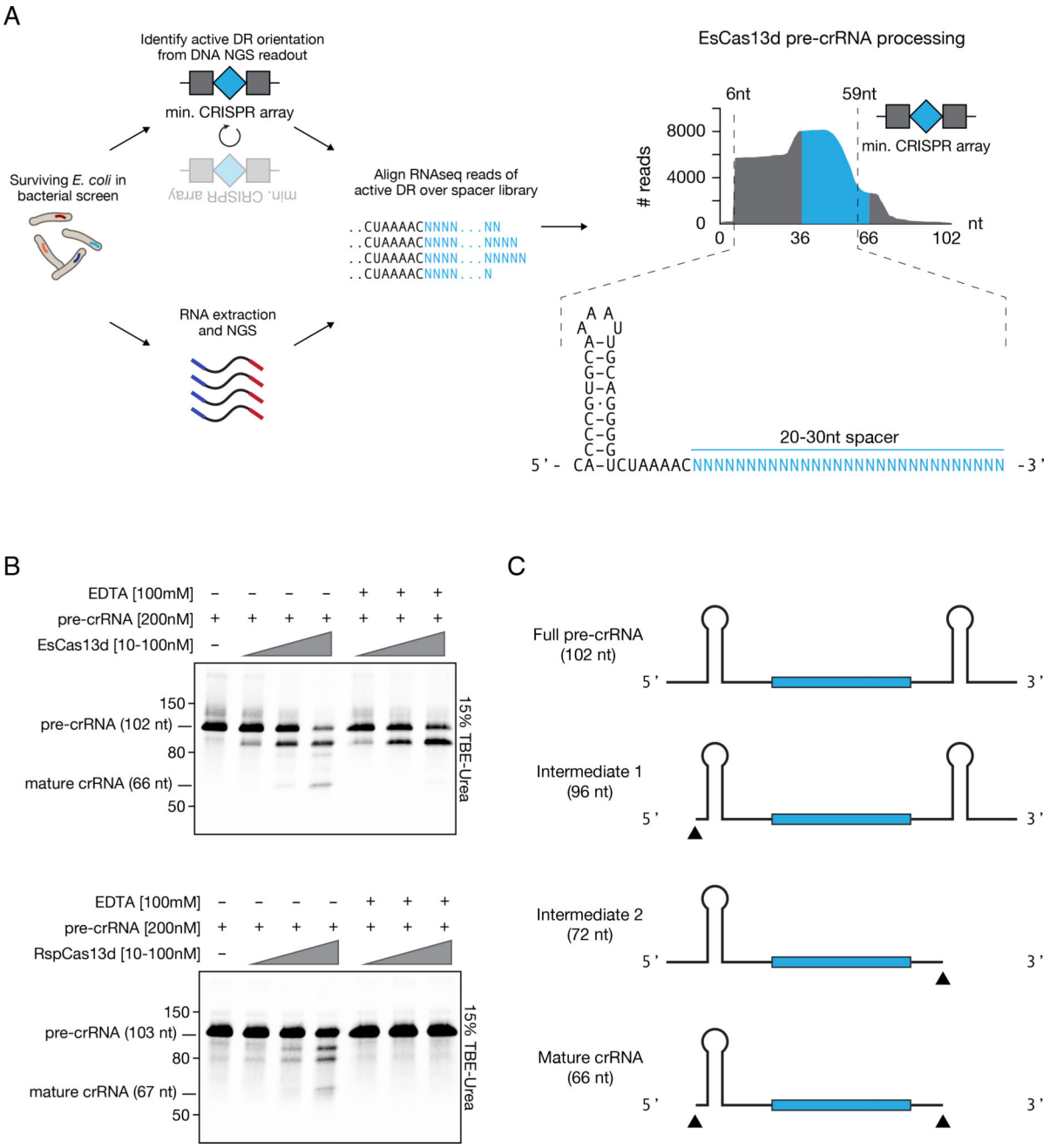


Figure 4. Cas13d effectors process precursor crRNA into mature crRNAs

(A) Schematic of the RNA extraction from bacterial screen, next-generation sequencing (NGS), and alignment to determine the mature crRNA for EsCas13d. Predicted EsCas13d pre-crRNA secondary structure is shown. (B) Representative denaturing gels depicting Cas13d mediated cleavage of their cognate pre-crRNAs over a dose titration of effector concentration. The dependence of Cas13d crRNA biogenesis on divalent metal cations was evaluated with the introduction of 100mM EDTA to the standard reaction conditions. See also Figure S5 and Table S1. (C) Schematic representing the major products identified from next-generation sequencing of *in vitro* cleaved RNA fragments from the pre-crRNA

processing with EsCas13d and RspCas13d. The black line represents the direct repeats and associated secondary structure, blue box the full-length spacer, and filled triangle the cleavage sites. The lengths described are for processed EsCas13d crRNAs, with RspCas13d having one extra nucleotide due to the 31nt natural length spacer used for instead of 30. Not depicted are the 3–4 nt at the 5' end of the pre-crRNA from T7 *in vitro* transcription; accounting for them provides better matches to the approximate sizes determined on the denaturing gel in panel B.

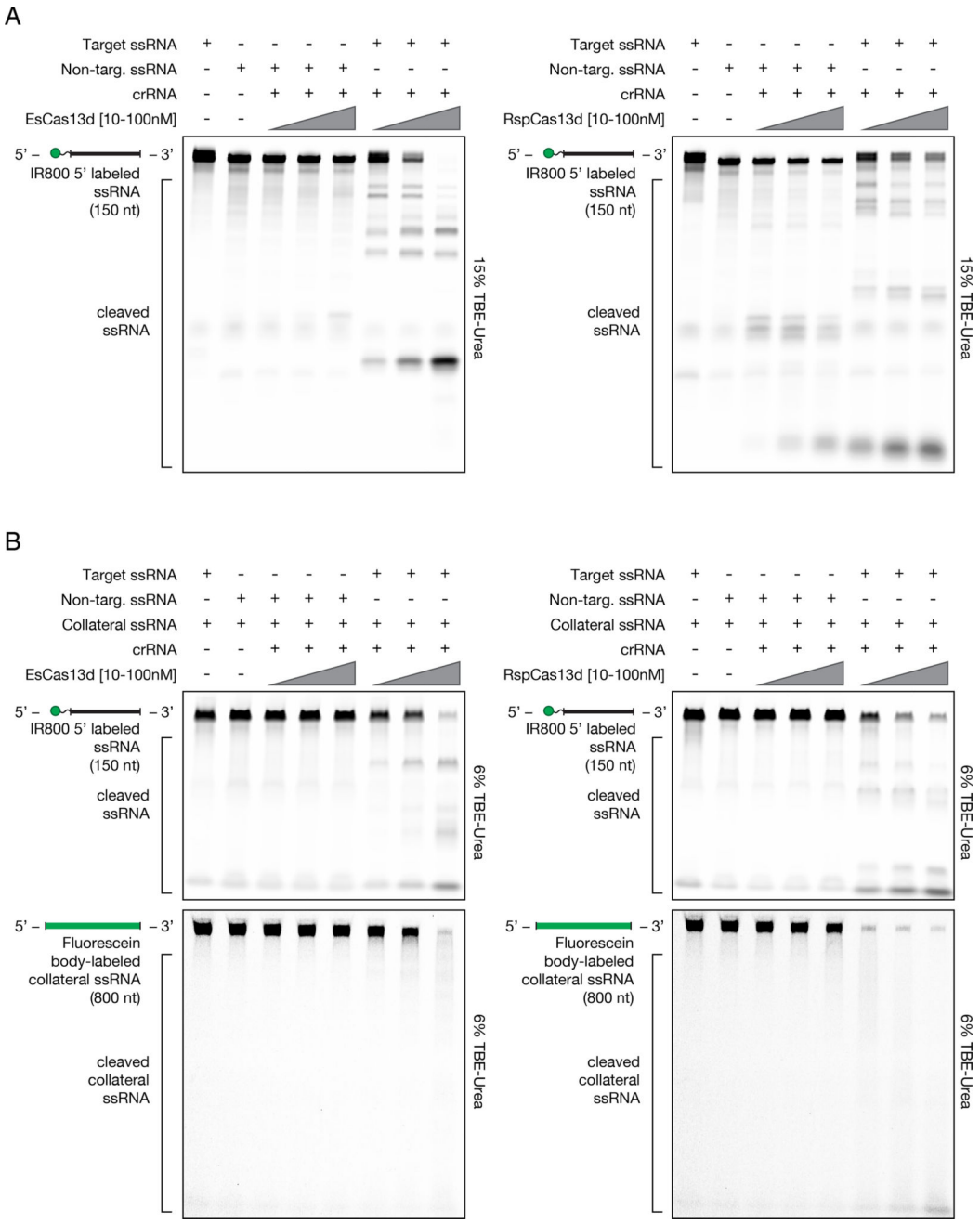


Figure 5. Cas13d is a programmable RNA-guided single stranded RNA nuclease with collateral RNase activity

(A) Representative denaturing gels displaying the targeted RNase activity of EsCas13d and RspCas13d effector proteins, with substrate RNA cleavage occurring when the crRNA matches its complementary target ssRNA. RNA substrates are 5' labeled with IRDye 800. (B) Representative denaturing gels displaying non-specific RNase activity of the Cas13d effectors upon targeted substrate recognition, demonstrated by the cleavage of fluorescein dUTP body-labeled collateral RNA upon activation of the target nuclease activity. For all reactions, EsCas13d-crRNA and RspCas13d-crRNA complexes were formed by pre-incubating Cas13d and cognate crRNA for 5 minutes at 37°C, prior to a dding target and/or

collateral ssRNA and incubating the reaction for 30 minutes. See also Figure S6 and Table S1.

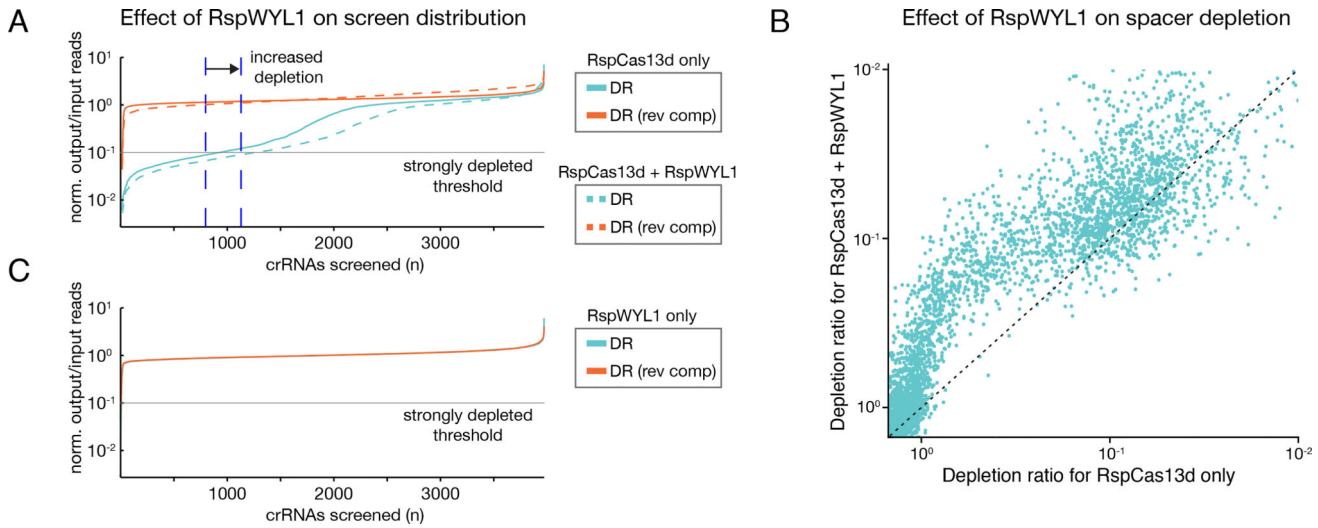


Figure 6. RspWYL1 positively modulates *in vivo* target depletion

Analysis of bacterial screens performed with RspCas13d effector only versus the RspCas13d with the RspWYL1 accessory (referred to as the RspCas13d system). (A) Comparative depletion plot of bacterial screens performed on RspCas13d only (solid line) versus RspCas13d with RspWYL1 (dotted line). The blue dashed lines demarcate the intersection of the ranked screen hits with the depletion fraction of 0.1, below which we define as strongly depleted. (B) Spacer depletion ratios for RspCas13d with and without RspWYL1. (C) Depletion plot of bacterial screens using only RspWYL1 and the repeat-spacer-repeat library associated with RspCas13d. Two biological replicates were performed for each screen.

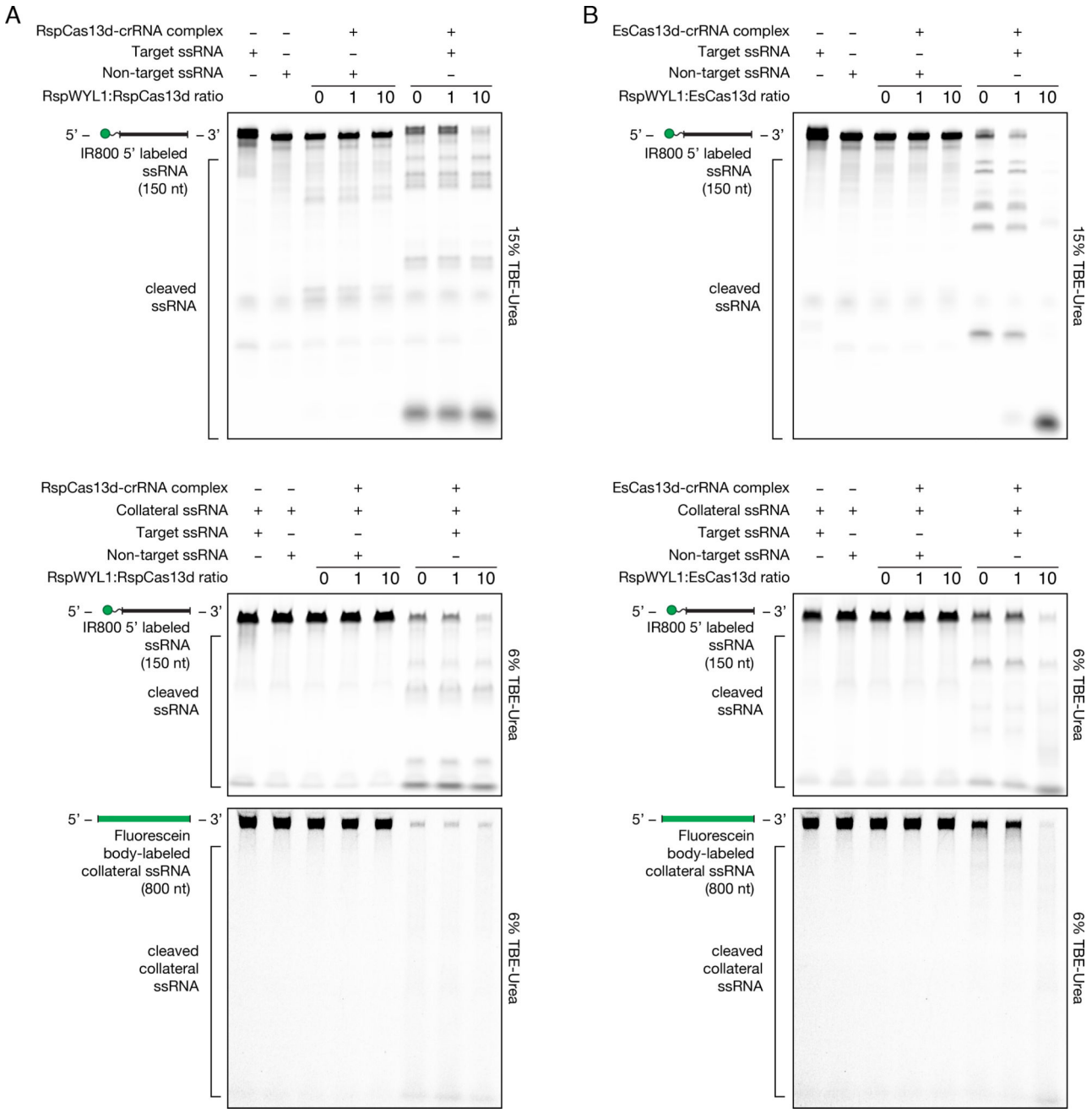


Figure 7. RspWYL1 is non-specific positive modulator of divergent Cas13d proteins

(A) Representative activity of titrating different molar ratios of purified RspWYL1 to a fixed dose of RspCas13d. The top panel is an ssRNA substrate cleavage assay, and the lower two panels evaluate the effect of RspWYL1 on collateral activity. (B) Extensibility of the activity in RspWYL1 for an orthologous effector, EsCas13d. In both of these reactions, RspWYL1 was pre-incubated along with the pre-crRNA and Cas13d effector for 5 minutes at 37°C before incubation with substrate RNA. The final concentration of Cas13d in the reaction is

33nM with a 2:1 ratio of Cas13d to pre-crRNA. See also Figure S7 for an extended dose titration of RspWYL1 with RspCas13d.