

MIT Open Access Articles

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

The MIT Faculty has made this article openly available. **Please share**
how this access benefits you. Your story matters.

Citation: Shmakov, Sergey et al. "Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems." *Molecular Cell* 60, 3 (November 2015): 385–397 © 2015 Elsevier

As Published: <http://dx.doi.org/10.1016/J.MOLCEL.2015.10.008>

Publisher: Elsevier

Persistent URL: <http://hdl.handle.net/1721.1/112723>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-NonCommercial-NoDerivs License





Published in final edited form as:

Mol Cell. 2015 November 5; 60(3): 385–397. doi:10.1016/j.molcel.2015.10.008.

Discovery and functional characterization of diverse Class 2 CRISPR-Cas systems

Sergey Shmakov^{1,2,#}, Omar O. Abudayyeh^{3,#}, Kira S. Makarova², Yuri I Wolf², Jonathan S. Gootenberg³, Ekaterina Semenova⁴, Leonid Minakhin⁴, Julia Joung³, Silvana Konermann³, Konstantin Severinov^{1,4,5}, Feng Zhang^{3,*}, and Eugene V. Koonin^{2,*}

¹ Skolkovo Institute of Science and Technology, Skolkovo, 143025, Russia

² National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda, MD 20894, USA

³ Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA, McGovern Institute for Brain Research, Department of Brain and Cognitive Science, Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

⁴ Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

⁵ Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, 123182, Russia

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.

* Authors to whom correspondence should be addressed: koonin@ncbi.nlm.nih.gov; zhang@broadinstitute.org.

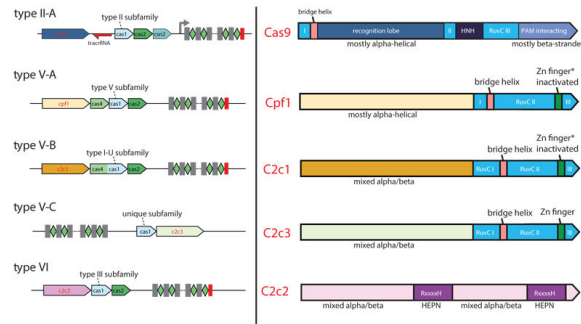
These authors made equal contributions to this work

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author contributions

EVK and FZ conceived the study; KSM, YIW and EVK designed the computational analyses; SS, KSM, YIW and FZ performed the computational analyses; OOA, JSG, FZ and KS designed the experiments; OOA, JSG, JL, SK, ES and LM performed the experiments; OOA, JSG and FZ analyzed the results; OOA, KSM, FZ and EVK wrote the manuscript that was read, edited and approved by all authors.

Graphical abstract



Keywords

CRISPR-Cas adaptive immunity; Cas9; Cpf1; crRNA; tracrRNA; PAM; RuvC-like endonuclease; HEPN domain; computational discovery pipeline; RNA-seq

INTRODUCTION

CRISPR (clustered regularly interspaced short palindromic repeat)-Cas (CRISPR-Associated proteins) are adaptive immune systems of archaea and bacteria (Marraffini and Sontheimer, 2010; Koonin and Makarova, 2013; Barrangou and Marraffini, 2014). 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 (Cong et al., 2013; Mali et al., 2013; Cho et al., 2013; Jiang et al., 2013; Hwang et al., 2013; Jinek et al., 2013). The CRISPR-Cas systems show extreme diversity of Cas protein composition as well as genomic loci architecture (Makarova et al., 2011b; Makarova et al., 2015).

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 1A) (Makarova et al., 2015). 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 (Figure 1A).

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; Garneau et al., 2010; Deltcheva et al., 2011; Sapranasauskas et al., 2011; Jinek et al., 2012; Gasiunas et al., 2012). 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 (Schunder et al., 2013; Makarova et al., 2015), has been shown to be an RNA-guided endonuclease that cleaves the target DNA via a staggered cut (Zetsche et al., 2015). Based on their unique domain architecture, the Cpf1-containing systems have been categorized as type V CRISPR-Cas (Makarova et al., 2015).

Although Class 2 systems are less common than Class 1 systems (Chylinski et al., 2014; Makarova et al., 2015), 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 systems provides evidence to suggest Class 2 systems originated by combination of Class 1 adaptation modules with effector proteins derived from different mobile elements.

RESULTS

Computational prediction of candidate newly identified 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 *casI* gene (Makarova et al., 2011b; Makarova et al., 2015) and the Cas1 sequence is the most conserved among all Cas proteins (Takeuchi et al., 2012), we used *casI* 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 (Makarova et al., 2011b; Fonfara et al., 2014; Chylinski et al., 2013; Chylinski et al., 2014; Makarova et al., 2015). To identify additional 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 (Table S1). 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 *casI* 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 1C and Figure S1; Table S1).

The first group (Figure 1C and Figure 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. The C2c1 loci encode a Cas1-Cas4 fusion, Cas2, and a large putative effector protein, and are typically adjacent to a CRISPR array. 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 1C and Figure S1B). The third group, denoted C2c2 (Class 2 candidate 2), was identified in 21 genomes from five major bacterial taxa: α -proteobacteria, *Bacilli*, *Clostridia*, *Fusobacteria*, and *Bacteroidetes*. 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 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 (Table S1). 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 (Table S1). The CRISPR arrays of the C2c2 loci are also highly heterogeneous (repeat length ranging from 35 to 39 bp) and unstructured (Table S1).

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 (Table S1). 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 S2)

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., 2015), 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 2A and Figure 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 2A and Figure 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 2A; Figures S4 and 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 α/β conformation (Figures S4 and 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 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.

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 (Grynberg et al., 2003; Anantharaman et al., 2013). Additionally, a conserved glutamate embedded in a strongly predicted long α -helix and corresponding to the similar motif of HEPN domains was identified (Figure 2B; Figure S6). The HEPN superfamily includes small (~150 aa) α -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 α -helical secondary structure of each putative domain is consistent with the HEPN fold (Figure 2B; Figure S6). Outside of the two HEPN domains, the C2c2 sequence is predicted to adopt a mixed α/β structure without discernible similarity to any known protein folds (Figure S6). Given the unique predicted effector of C2c2, these systems qualify as a putative type VI CRISPR-Cas.

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

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 (Deltcheva et al., 2011; Chylinski et al., 2013; Chylinski et al., 2014). 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 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 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 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 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 S1 and S7A; Table S1). 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., 2015). Effectively identical patterns of RNA expression and processing were observed when the Aac C2c1 locus was expressed in the heterologous *E. coli* system (Figure 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 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 3E) that all had the sequence NNNNTTN (Figure 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 3G).

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 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 4B).

Because RNA-seq experiments identified putative tracrRNA transcripts of variable size (Figure 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 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 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 3F) and found that linear DNA molecules containing protospacer-1 preceded by TTT, TTA, and TTC PAMs but not GGA were efficiently cleaved (Figure 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 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 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 S7C) that displayed secondary structure and repeat-anti-repeat base-pairing similar to the putative Aac tracrRNA (Figure 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 S7E).

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 5A). In contrast to the C2c1 loci, although this C2c2 locus contains a predicted tracrRNA (Figure S1C), we did not observe its expression (Figure 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 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 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 S7F, G). An additional, uncharacterized small RNA was expressed in the vicinity of the CRISPR array in *L. shahii* (Figure S7F) but not in *E. coli* cells (Figure S7G). *In silico* folding of the crRNA predicted secondary structure that was highly similar to that in *L. seeligeri* (Figure 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.

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., 2015). In the phylogenetic tree of Cas1, putative subtype V-B (C2c1) is largely monophyletic and confidently clusters with type I-U (Figure 6; see also Supplemental Experimental Procedures). 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 6). In the Cas1 tree, the type V-C branch is inside subtype I-B, although the position of such a fast evolving group should be taken with caution. The type VI Cas1 proteins are distributed between 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. 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* (Norais et al., 2013; Chylinski et al., 2014). Thus, the adaptation modules of these newly identified Class 2 CRISPR-Cas systems apparently come from different variants of Class 1.

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., 2015). 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., 2009; Hale et al., 2014; 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., 2015), 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 (Makarova et al., 2011b; Chylinski et al., 2014) 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) (Figure 2). 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 (Figure S3). 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 (Nishimasu et al., 2014; Anders 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 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 (Krupovic et al., 2014; Koonin and Krupovic, 2015). 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 6).

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 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 newly identified 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 C2c1 nuclease in human cell lysate described here, and Cpf1-mediated genome editing in human cells (Zetsche et al., 2015). In addition, the

discovery of new variants will provide direct tests of the modular scenario of the evolution of CRISPR-Cas systems (Figure 7) and shed further light on the function of these diverse systems.

EXPERIMENTAL PROCEDURES

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., 2015) 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., 2015) 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., 2015) 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 (Kato 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.

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 Herculanase II polymerase (Agilent Technologies).

RNA-sequencing analysis

The prepared cDNA libraries were sequenced on a 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.

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 (see Supplemental Experimental Procedures). 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, $>4 \times 10^6$ 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.

PAM validation

Sequences corresponding to both PAMs and 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).

***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).

For additional details see Supplemental Experimental Procedures

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank R. Macrae and F. A. Ran for critical reading of the manuscript. S.S. is supported by the graduate program of Skoltech Data-Intensive Biomedicine and Biotechnology Center for Research, Education, and Innovation. K.S.M., Y.I.W., and E.V.K are supported by the intramural program of the US Department of Health and Human services (to the national Library of Medicine). J.S.G. is supported by a D.O.E. Computational Science Graduate Fellowship. F.Z. is supported by the NIMH (5DP1-MH100706 and 1R01-MH110049), NIDDK (5R01DK097768-03), the Poitras Center, Vallee, Simons, Paul G. Allen, and New York Stem Cell Foundations, David R. Cheng, Tom Harriman, and Bob Metcalfe. F.Z. is a New York Stem Cell Foundation Robertson Investigator. K.S. is supported by an NIH grant GM10407, Russian Science Foundation grant 14-14-00988, and by Skoltech.

REFERENCES

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 1997; 25:3389–3402. [PubMed: 9254694]
- Anantharaman V, Makarova KS, Burroughs AM, Koonin EV, Aravind L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct*. 2013; 8:15. [PubMed: 23768067]
- Anders C, Niewoehner O, Duerst A, Jinek M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*. 2014; 513:569–573. [PubMed: 25079318]
- Aravind L, Makarova KS, Koonin EV. SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res*. 2000; 28:3417–3432. [PubMed: 10982859]
- Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 2007; 315:1709–1712. [PubMed: 17379808]
- Barrangou R, Marraffini LA. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Molecular cell*. 2014; 54:234–244. [PubMed: 24766887]
- Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*. 2001; 29:2607–2618. [PubMed: 11410670]
- Biswas A, Fineran PC, Brown CM. Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. *Bioinformatics*. 2014; 30:1805–1813. [PubMed: 24578404]

- Chang SS, Kang DH. Alicyclobacillus spp. in the fruit juice industry: history, characteristics, and current isolation/detection procedures. *Crit Rev Microbiol*. 2004; 30:55–74. [PubMed: 15239380]
- Cho SW, Kim S, Kim JM, Kim JS. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology*. 2013; 31:230–232.
- Chylinski K, Le Rhun A, Charpentier E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA biology*. 2013; 10:726–737. [PubMed: 23563642]
- Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type II CRISPR-Cas systems. *Nucleic acids research*. 2014; 42:6091–6105. [PubMed: 24728998]
- 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]
- Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*. 2011; 471:602–607. [PubMed: 21455174]
- Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res*. 2015; 43:W389–394. [PubMed: 25883141]
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*. 2004; 32:1792–1797. [PubMed: 15034147]
- Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*. 2007; 8:18. [PubMed: 17239253]
- Fonfara I, Le Rhun A, Chylinski K, Makarova KS, Lecrivain AL, Bzdrenga J, Koonin EV, Charpentier E. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic acids research*. 2014; 42:2577–2590. [PubMed: 24270795]
- Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. 2010; 468:67–71. [PubMed: 21048762]
- Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*. 2012; 109:E2579–2586. [PubMed: 22949671]
- Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*. 2007; 35:W52–57. [PubMed: 17537822]
- Grynberg M, Erlandsen H, Godzik A. HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins. *Trends in biochemical sciences*. 2003; 28:224–226. [PubMed: 12765831]
- Hale CR, Coccozaki A, Li H, Terns RM, Terns MP. Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex. *Genes Dev*. 2014; 28:2432–2443. [PubMed: 25367038]
- Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*. 2009; 139:945–956. [PubMed: 19945378]
- Heidrich N, Dugar G, Vogel J, Sharma CM. Investigating CRISPR RNA Biogenesis and Function Using RNA-seq. *Methods Mol Biol*. 2015; 1311:1–21. [PubMed: 25981463]
- Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology*. 2013; 31:227–229.
- Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology*. 2013; 31:233–239.
- 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]
- Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. *eLife*. 2013; 2:e00471. [PubMed: 23386978]
- 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]

- Koonin EV, Krupovic M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nature reviews Genetics*. 2015; 16:184–192.
- Koonin EV, Makarova KS. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA biology*. 2013; 10:679–686. [PubMed: 23439366]
- Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biology*. 2014; 12:36. [PubMed: 24884953]
- Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic acids research*. 2013; 41:8034–8044. [PubMed: 23863837]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–1760. [PubMed: 19451168]
- Majumdar S, Zhao P, Pfister NT, Compton M, Olson S, Glover CV 3rd, Wells L, Graveley BR, Terns RM, Terns MP. Three CRISPR-Cas immune effector complexes coexist in *Pyrococcus furiosus*. *RNA*. 2015; 21:1147–1158. [PubMed: 25904135]
- Makarova KS, Aravind L, Wolf YI, Koonin EV. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct*. 2011a; 6:38. [PubMed: 21756346]
- Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 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*. 2006; 1:7. [PubMed: 16545108]
- Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, et al. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol*. 2011b; 9:467–477. [PubMed: 21552286]
- Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. *Methods Mol Biol*. 2015; 1311:47–75. [PubMed: 25981466]
- Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, et al. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol*. 2015
- Makarova KS, Wolf YI, Koonin EV. The basic building blocks and evolution of CRISPR-cas systems. *Biochem Soc Trans*. 2013; 41:1392–1400. [PubMed: 24256226]
- Mali P, Esvelt KM, Church GM. Cas9 as a versatile tool for engineering biology. *Nature methods*. 2013; 10:957–963. [PubMed: 24076990]
- Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*. 2002; 30:281–283. [PubMed: 11752315]
- Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, et al. CDD: conserved domains and protein three-dimensional structure. *Nucleic acids research*. 2013; 41:D348–352. [PubMed: 23197659]
- Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews Genetics*. 2010; 11:181–190.
- Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*. 2009; 155:733–740. [PubMed: 19246744]
- Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA. Database indexing for production MegaBLAST searches. *Bioinformatics*. 2008; 24:1757–1764. [PubMed: 18567917]
- Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*. 2014; 156:935–949. [PubMed: 24529477]
- Norais C, Moisan A, Gaspin C, Clouet-d'Orval B. Diversity of CRISPR systems in the euryarchaeal Pyrococcales. *RNA Biol*. 2013; 10:659–670. [PubMed: 23422322]

- Peng W, Feng M, Feng X, Liang YX, She Q. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic acids research*. 2015; 43:406–417. [PubMed: 25505143]
- Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010; 5:e9490. [PubMed: 20224823]
- Sampson TR, Saroj SD, Llewellyn AC, Tzeng YL, Weiss DS. A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature*. 2013; 497:254–257. [PubMed: 23584588]
- Sapranaukas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic acids research*. 2011; 39:9275–9282. [PubMed: 21813460]
- Schunder E, Rydzewski K, Grunow R, Heuner K. First indication for a functional CRISPR/Cas system in *Francisella tularensis*. *Int J Med Microbiol*. 2013; 303:51–60. [PubMed: 23333731]
- Soding J, Remmert M, Biegert A, Lupas AN. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. *Nucleic acids research*. 2006; 34:W374–378. [PubMed: 16845029]
- Stoddard BL. Homing endonuclease structure and function. *Q Rev Biophys*. 2005; 38:49–95. [PubMed: 16336743]
- Takeuchi N, Wolf YI, Makarova KS, Koonin EV. Nature and intensity of selection pressure on CRISPR-associated genes. *Journal of bacteriology*. 2012; 194:1216–1225. [PubMed: 22178975]
- Zetsche B, Gootenberg J, Abudayyeh O, Slaymaker I, Makarova K, Volz S, Joung J, Essletzbichler P, Van der Oost J, Regev A, et al. Cpf1 is a single RNA-guided endonuclease of a novel Class 2 CRISPR-Cas system. *Cell*. 2015 in press.
- Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, Vogel J, Sontheimer EJ. Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Molecular cell*. 2013; 50:488–503. [PubMed: 23706818]
- Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 2003; 31:3406–3415. [PubMed: 12824337]

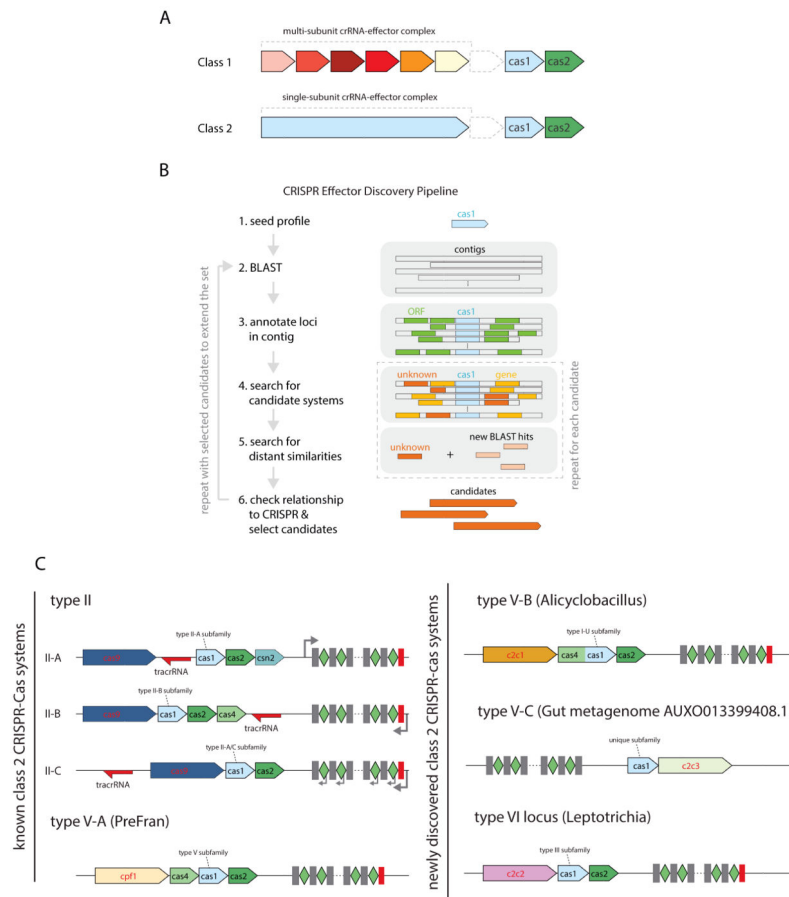


Figure 1. Prediction of candidate newly identified Class 2 CRISPR-Cas systems

(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; 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 Figures S1 and S2, and Table S1.

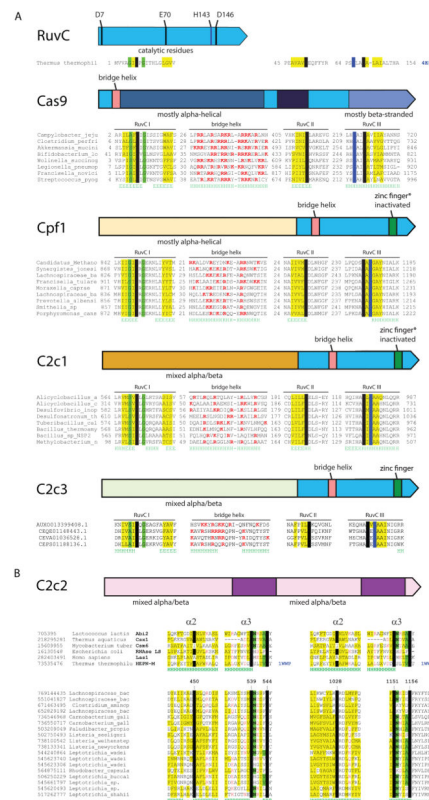


Figure 2. Domain architectures and conserved motifs of the Class 2 effector proteins

(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 α -helix and E denotes extended conformation (β -strand). The poorly conserved spacers between the alignment blocks are shown by numbers. See also Figures S3, S4 and 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). See also Figures S3 and S6.

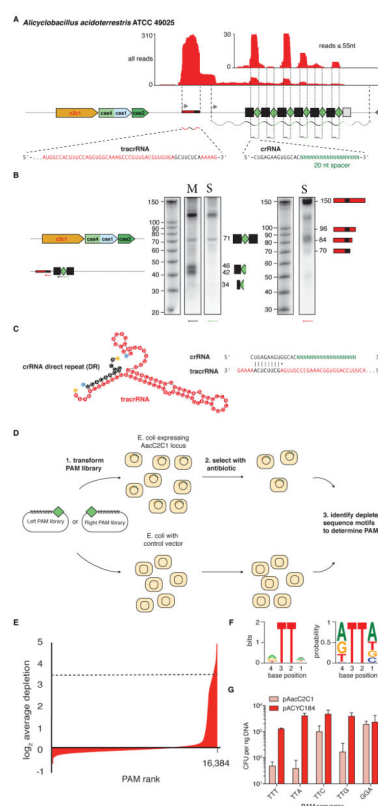


Figure 3. Functional validation of the *Alicyclobacillus acideoterrestris* C2c1 locus

(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 S7A, B and C).

(B) Northern blot of RNAs expressed from the 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 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 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.

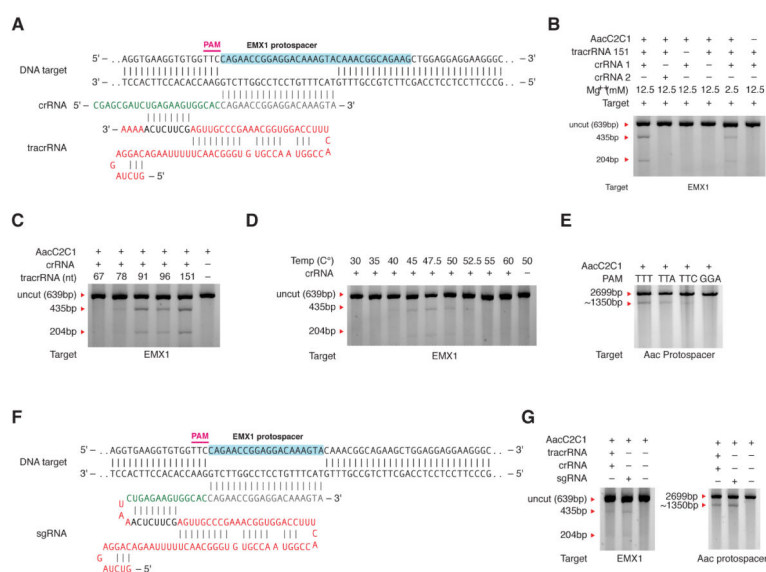


Figure 4. Characterization of the cleavage requirements of *A. acideoterrestres* C2c1 in a human cell lysate

(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 Mg²⁺ and weak cleavage in the absence of Mg²⁺.

(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 91-nt 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.

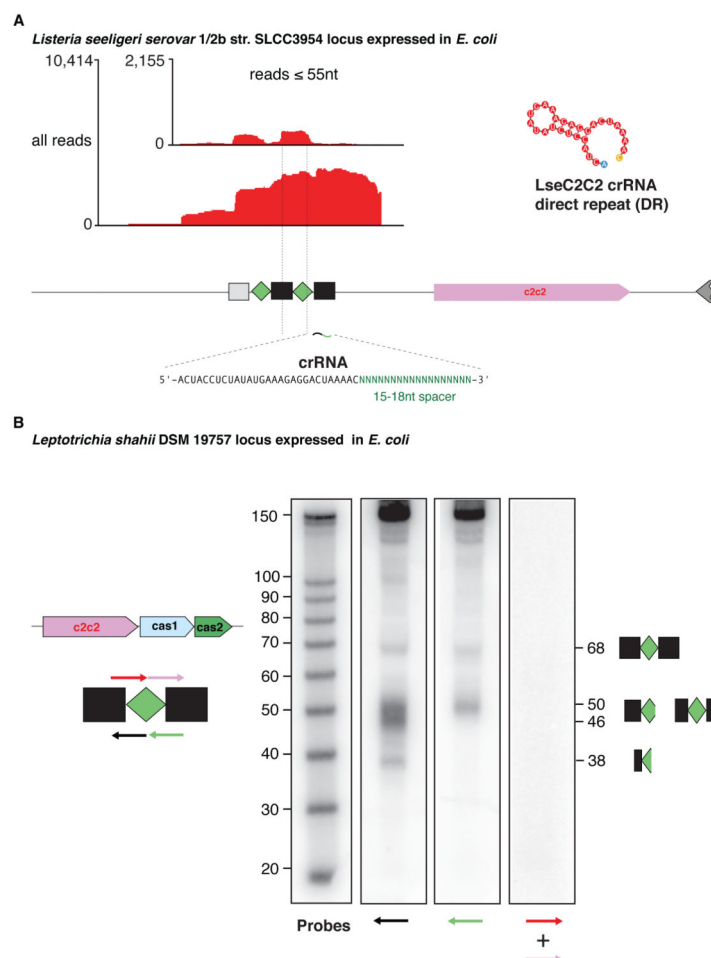


Figure 5. Expression and processing of C2c2 loci

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

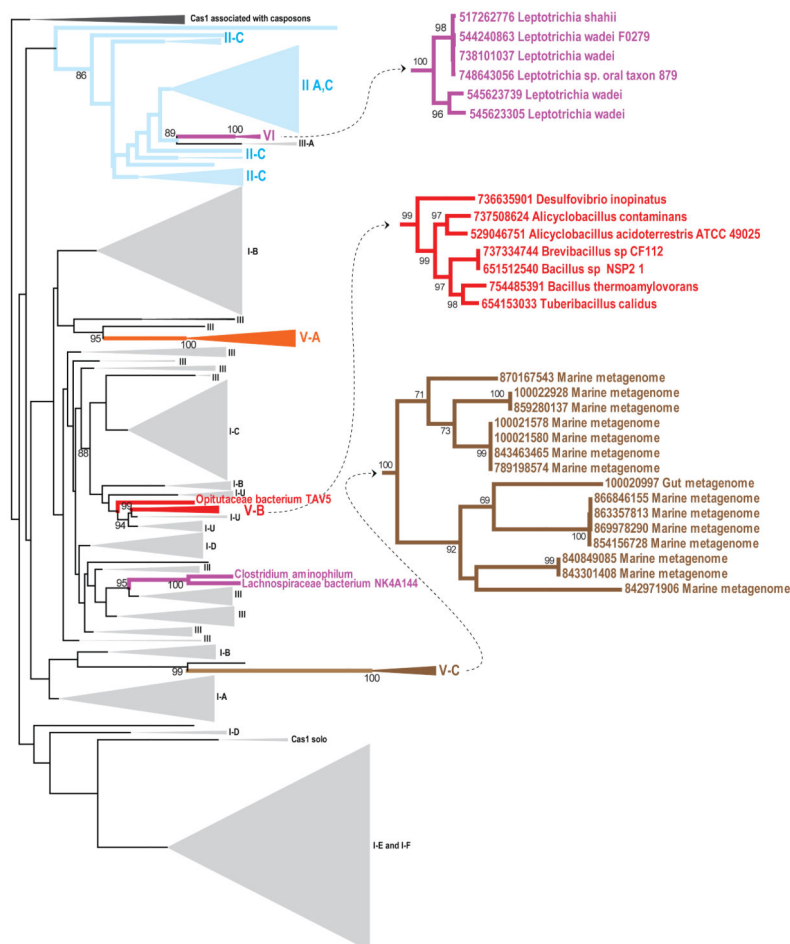


Figure 6. Phylogenetic tree of Cas1

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 newly identified (sub)types. The bootstrap support values are given as percentage points and shown only for several relevant branches. The complete tree in the Newick format with species names and bootstrap support values and the underlying alignment are available at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Class2/.

See also Supplemental Experimental Procedures.

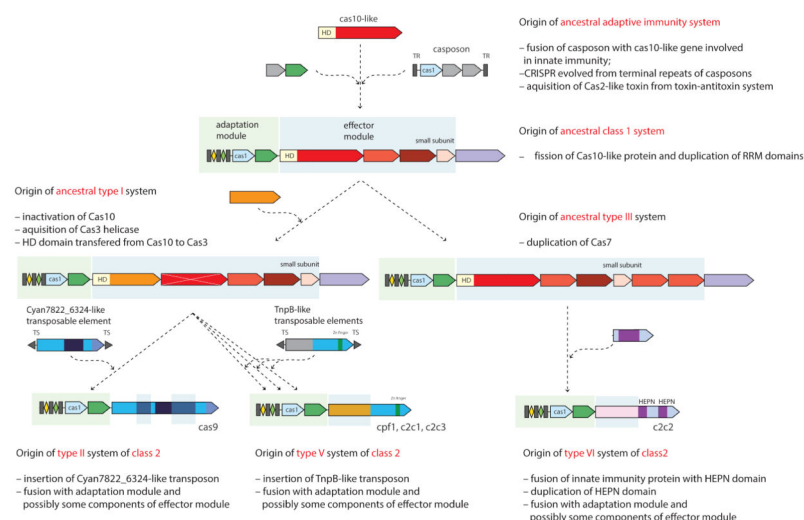


Figure 7. Evolutionary scenario for the CRISPR-Cas systems

The scenario is a synthesis of the present and previous analyses (Makarova et al., 2011a; Makarova et al., 2013; Chylinski et al., 2014; Makarova et al., 2015). 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. 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. 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.