Skolkovo Institute of Science and Technology

COMPUTATIONAL APPROACHES FOR DISCOVERY OF NOVEL CRISPR-Cas SYSTEMS

Doctoral Thesis

by

SERGEY SHMAKOV

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Professor Konstantin Severinov

Co-advisor
Doctor Eugene Koonin

Moscow - 2017

Abstract

CRISPR-Cas systems are diverse adaptive immunity systems in bacteria and archaea that can incorporate fragments of foreign DNA (spacers) into host genome CRISPR array, thus granting protection from future invasions of mobile elements that contain sequence complementary to the acquired spacer. Discovery and characterization CRISOR-Cas systems not only provided better understanding of defense systems in prokaryotes, but also was a cornerstone for revolution in genome editing. Enrolling effector complexes of CRISPR-Cas class 2 systems - sequence specific nucleases such as Cas9 - greatly improved efficiency and eased genome editing.

In this study, we covered genomic and metagenomic data to perform exhaustive search for new CRISPR-Cas systems. We designed a computational pipeline for the discovery of novel class 2 effector complexes variants. This pipeline uses core features of CRISPR-Cas systems such as Cas1 and CRISPR arrays to identify loci of potential novel or known CRISPR systems and applies heuristic filtering to create a list of potential class 2 effector complexes candidates. By applying this pipeline, we discovered six novel CRISPR-Cas subtypes: Type V-B, Type V-C, Type V-U, Type IV-A, Type IV-B, Type IV-C in complete and draft bacterial and archaeal genomes. We describe the diverse properties of these new systems and offer that they can be implemented for the development of versatile genome editing and regulation tools. In this work, we present a comprehensive census of class 2 types and class 2 subtypes for available genomic data. The census shows almost complete absence of class 2 CRISPR-Cas arrays in archaea and dominance of Type II among all identifiable class 2 systems. Finally, we outline plausible evolutionary scenarios for the independent origin of class 2 CRISPR-Cas systems from mobile genetic elements, and propose amended classification and nomenclature of CRISPR-Cas.s

Headings

Introduction

Significance of the Work

CRISPR-Cas (CRISPR: Clustered Regular Interspaced Short Palindromic Repeats, CRISPR-Cas: CRISPR Associated) are diverse adaptive immune systems of archaea and bacteria (Makarova *et al.*, 2006; Barrangou *et al.*, 2007; Barrangou, 2013; Marraffini, 2015; Mohanraju *et al.*, 2016). These systems recently attracted much attention due to their unique, 'Lamarkian' mode of action (Koonin and Wolf, 2016) that retains memory (spacers) from past infections and provides specific resistance to these infections via an RNA-guided process that has been successfully used to create genome editing tools (Ran, Hsu, Wright, *et al.*, 2013). The structural features and mechanisms of CRISPR-Cas are described in detail in several recent reviews (Barrangou, 2013; van der Oost *et al.*, 2014; Marraffini, 2015; Mohanraju *et al.*, 2016).

The CRISPR-Cas systems show extreme diversity of Cas protein composition and of genomic loci architecture (Makarova, Haft, *et al.*, 2011; 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 recognizes and degrades target sequences. 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. CRISPR-Cas operation can be described in terms of adaptation, crRNAs biogenesis and interference stages. During adaptation the Cas1–Cas2 protein complex (which, in some cases, contains additional subunits) excises a segment of the target DNA (known as the proto spacer) and inserts it between the repeats at the 5′ end of a CRISPR array, yielding a new spacer. In the expression and processing stage (crRNA biogenesis), a CRISPR array, together with the spacers, is transcribed into a long transcript known as the pre-CRISPR RNA (pre-crRNA) and is processed by a distinct complex of Cas proteins (which, in some cases, involves additional proteins and RNA molecules) into mature small CRISPR RNAs (crRNAs). In the interference stage the effector module proteins, guided by crRNA, target and cleave invading nucleic acids (Makarova, Haft, *et al.*, 2011; Makarova, Wolf and Koonin, 2013b). To avoid self-targeting CRISPR-Cas systems use a protospacer (target sequence complementary to a spacer) adjacent motif (PAM), consisting of several nucleotides at one of protospacer sides. This motif and its location are distinct for different CRISPR-Cas systems and is recognized by the effector complex.

The latest classification of CRISPR-Cas systems divides them into two classes, 5 types and 16 subtypes, based on the architecture of the effector modules (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.

The effector complexes of class 1 systems consist of 4–7 Cas protein subunits in an uneven stoichiometry, as exemplified by the CRISPR-associated complex for antiviral defense (Cascade) of type I systems (Brouns *et al.*, 2008; Jore *et al.*, 2011; Jackson, Golden, *et al.*, 2014; Beloglazova *et al.*, 2015), and the Csm–Cmr complexes of type III systems (Rouillon *et al.*, 2013; Staals *et al.*, 2014; Osawa *et al.*, 2015; Taylor *et al.*, 2015). For Class 2 the effector protein of type II CRISPR-Cas systems is Cas9, a large multi domain nuclease that varies 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; Sapranauskas *et al.*, 2011; Gasiunas *et al.*, 2012; Jinek *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 an 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).

CRISPR-Cas effectors, such as the crRNA guided nuclease protein Cas9 (Gasiunas *et al.*, 2012), have significantly improved the efficiency of genome editing due to their relative ease of use and high specificity (Chen *et al.*, 2014). However, properties of existing effector complexes are not optimal (Slaymaker *et al.*, 2016), so there is a place and a need for discovery of new protein families to enrich the capabilities of existing CRISPR-Cas tools. The recently discovered Cpf1 effector complex (Zetsche *et al.*, 2015), which has recently been deployed for genome editing due of its unique properties (Zetsche *et al.*, 2017), shows that searches of genomic data can indeed yield new genomic editors with superior properties. Such work has been carried out with complete genome sequences (Makarova *et al.*, 2015) but incomplete genomes and metagenomes have not yet been investigated, so additional CRISPR-Cas systems remain to be found (Makarova *et al.*, 2015). The mode of action of Class 2 CRISPR-Cas systems has been quite well described (Mohanraju *et al.*, 2016), but the evolution of Class 2 remains unclear. Novel variants may have different properties that can shed light on defense systems in prokaryotes, virus-host interactions, and the evolution and functioning of CRISPR-Cas systems, in addition to offering new tools for deployment in experimental research and medicine.

Project Objectives

The main goal of this project was diversity assessment of Class 2 CRISPR-Cas systems - systems with a large (> 500aa) single-subunit effector complex - and the discovery of novel Class 2 genes and protein variants among bacteria and archaea through a thorough study of available genomic and metagenomic data. To achieve these goal, a computational approach was designed that uses essential components of CRISPR-Cas systems to search for associated elements.

Two components were analyzed in genomic data and used as seeds to get new CRISPR-Cas variants:

1 – Cas1: this component was chosen because it is the most conserved gene among CRISPR-Cas gene families and there are high quality profiles available (Takeuchi *et al.*, 2012; Makarova and Koonin, 2015) that can be used for protein search. The fact that phylogeny of Cas1 correlates with CRISPR-Cas types (Makarova and Koonin, 2015) was considered a useful property as it could help to identify novel systems. The Cas1 component is essential for CRISPR-Cas systems, since, in order to be adaptive, a CRISPR-Cas system needs to incorporate new spacers and this is the key function of Cas1. Thus, most CRISPR-Cas loci include this gene (Makarova, Haft, *et al.*, 2011; Makarova *et al.*, 2015) and *cas1* genes with associated genes coding for unknown proteins were the target of the search.

2 – CRISPR array: this is a hallmark component of CRISPR-Cas systems and is used as a storage (or database) for spacers (Mojica *et al.*, 2005). The majority of CRISPR-Cas loci have CRISPR arrays located adjacent to *cas* genes (Makarova *et al.*, 2015), so novel systems may be identified by association with CRISPR arrays.

There are number of standalone *cas1* genes and CRISPR arrays (Makarova *et al.*, 2015) without any neighboring *cas* genes, which threatens to complicate the detection procedure by increasing noise in the neighborhood gene pool. Nonetheless, the majority of *cas1* genes and CRISPR arrays are localized with other *cas* genes, helping to reveal unknown *cas* genes via a principle of ‘guilt by association’.

The following tasks were set for the purposes of the study:

1. Identify *cas1* or CRISPR locations in bacterial or archaeal genomes and use them as seeds. The search would use following software tools: PSI-BLAST (Altschul *et al.*, 1990) to detect Cas1; CRISPRFinder (Grissa, Vergnaud and Pourcel, 2007) and PILER-CR (Edgar, 2007) to find CRISPR arrays; and available data sets such as GenBank (Benson *et al.*, 2013) and the Whole Genome Shotgun projects database (WGS)

(‘Database resources of the National Center for Biotechnology Information’, 2016) as a dataset.

2. Identify Open Reading Frames (ORFs) around seeds using GenMark (Besemer, Lomsadze and Borodovsky, 2001) and annotate them using CDD (Marchler-Bauer *et al.*, 2011, 2017) protein profiles and RPS-BLAST (Marchler-Bauer *et al.*, 2013) for further filtering procedures.
3. Identify the type of CRISPR-Cas system type for *cas* genes around seeds (if any), using the classification approach proposed by K.S. Makarova (Makarova and Koonin, 2015) to filter out known systems or select systems with an incomplete effector complex.
4. Cluster all proteins using the UCLUST software tool (Edgar, 2010) to group proteins and assess diversity in the protein group.
5. Manually analyze protein clusters using protein domain search tools such as: HHpred (Söding *et al.*, 2006) and CD-Search (Marchler-Bauer *et al.*, 2017).
6. Select and bioinformatically characterize a list of candidates by defining protein domains and selecting domains required for a CRISPR-Cas effector complex (such as nuclease domains or DNA/RNA binding domains).

Steps 1-6 will be used to compile a list of candidates that can be used for experimental validation in order to detect putative novel CRISPR-Cas systems and their genes. Steps 1-4 will be sufficient for a comprehensive census of CRISPR-Cas systems in available datasets.

Novelty and Practical Use

This is the first work to study diversity of CRISPR-Cas systems in a wide range of prokaryotic genomic data, which were available in 2016. Previous works were carried out using a limited dataset (Makarova *et al.*, 2015). Our work led to the discovery of novel CRISPR-Cas systems. These include Type V-B, Type V-C and tentative subgroup Type V-U containing RuvC nuclease domain (DNA targeting for Type V-B was experimentally validated (Shmakov *et al.*, 2015), Type VI contains Type VI-A, Type VI-B and Type VI-C; uniquely, these systems target RNA, experimentally validated for Type VI-A (Abudayyeh *et al.*, 2016) and Type VI-B (Smargon *et al.*, 2017)).

The CRISPR-Cas systems discovered in the course of this work can be used for various applications, such as genome editing, where specificity for DNA or RNA is needed. The novel Type V systems are distinguished from the well-studied CRISPR-Cas Type II (Cas9) by having a different protein architecture, and are also distinguished from the recently characterized Type V-A (Cpf1) (Zetsche *et al.*, 2015) by their reliance on tracrRNA (transactivating crRNA). RNase domains in Type VI could be used to target or track RNA molecules. The small size of the predicted effector proteins in Type V-U means that they could be packed into smaller delivery vehicles (viral capsids) in genome editing. These systems also have different sequence specificity in interference and PAM sequence motif.

Recent research showed the applicability of Cas13a (Type VI-A) from *Leptotrichia wadei* for very specific nucleic acid detection (Gootenberg *et al.*, 2017). There is potential for detecting various viral strands and pathogenic bacteria, and for tumor cell identification using this novel method. More exciting application of new systems are likely to follow soon

Personal Contribution

The greater part of the research was carried out by the author. The computational discovery pipeline was designed to prospect microbial genome sequence diversity so that previously undetected CRISPR-Cas variants could be identified. Various bacterial and archaeal databases were analyzed to detect CRISPR arrays and *cas1* genes (seeds). Loci around seeds were annotated and analyzed for the presence of unknown genes that are associated with CRISPR-

Cas components such as *cas1* and CRISPR arrays. Candidates for novel CRISPR-Cas systems were selected, initially characterized and proposed for experimental validation.

Material for Defense

1. Design of a computational discovery pipeline for Class 2 CRISPR effectors
2. Discovery of six novel CRISPR-Cas systems
3. An updated classification of CRISPR-Cas systems, including six new subtypes
4. A comprehensive census of Class 2 systems in prokaryotic databases
5. Hypothesis of the possible origins of CRISPR-Cas Class 2 systems
6. Possible applications of the discovered CRISPR-Cas systems

Authenticity and Validation of Results

The results were presented at a scientific conference and published in international peer-reviewed journals.

Publications:

1. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, Gootenberg JS, Makarova KS, Wolf YI, Severinov K, Zhang F, Koonin EV. Diversity and evolution of Class 2 CRISPR-Cas systems. // Nature Reviews Microbiology. 2017. 15(3):169-182
2. Smargon AA, Cox DB, Pyzocha NK, Zheng K, Slaymaker IM, Gootenberg JS, Abudayyeh OA, Essletzbichler P, Shmakov S, Makarova KS, Koonin EV, Zhang F. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially

Regulated by Accessory Proteins Csx27 and Csx28. // Molecular cell. 2017. 65(4):618-630.e7.

3. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DB, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. // Science. 2016. 353(6299):aaf5573.
4. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, Zhang F, Koonin EV. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. // Molecular Cell 2015. 60(3):385-97.

Poster presentations were made at the Genome Engineering 4.0 (USA, May 2016), CRISPR 2016 (Israel, May 2016) and CRISPR 2017 (Bozeman, USA, June 2017) international conferences

Review of Literature

CRISPR-Cas systems

Mode of action

CRISPR (clustered regularly interspaced short palindromic repeat)-Cas (CRISPR ASsociated proteins) are adaptive immune systems of archaea and bacteria (Makarova *et al.*, 2006; Barrangou *et al.*, 2007; Barrangou, 2013; Marraffini, 2015; Mohanraju *et al.*, 2016). About 90% of archaea and 40% of bacteria have this defense system (Bhaya, Davison and Barrangou, 2011; Makarova *et al.*, 2015). These systems provide defense against viral DNA (Makarova *et al.*, 2006; Barrangou *et al.*, 2007) and RNA (Hale *et al.*, 2014) in a three-stage process: adaptation, biogenesis of crRNA and interference. A schematic representation of these processes is given in Figure 1 and they have been quite thoroughly explained in recent reviews (Marraffini, 2015; Mohanraju *et al.*, 2016).

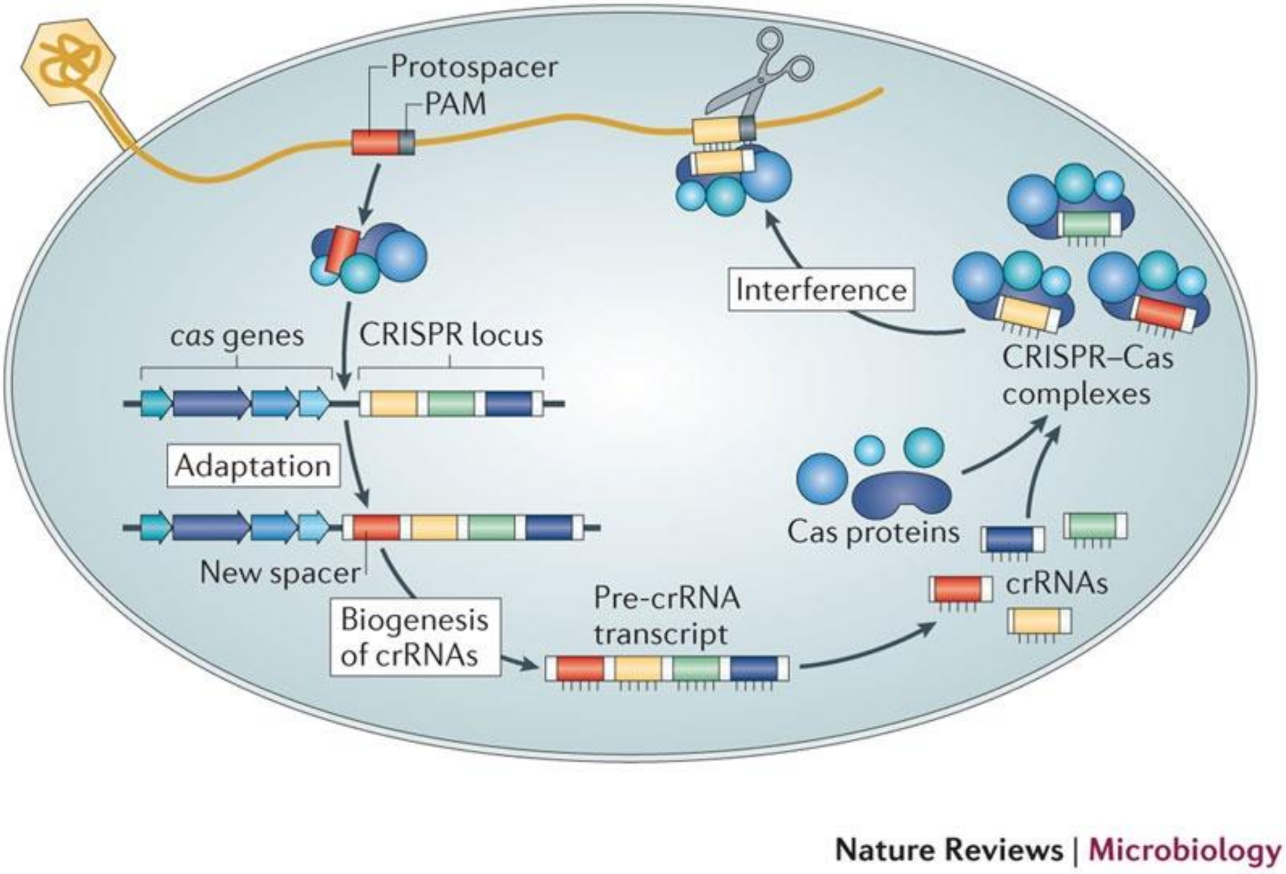


Figure 1. Schematics of CRISPR-Cas system function (reproduced with permission from Nature Reviews Microbiology (Samson *et al.*, 2013)). Three main functions of CRISPR-Cas systems are shown: Adaptation (acquisition of new spacers in the CRISPR locus), Biogenesis of crRNAs (expression and maturation of crRNA), and CRISPR-Cas interference (recognition and cleavage of foreign DNA or RNA).

Adaptation is the process of incorporation of new spacers into a CRISPR array. Cas1 and Cas2 proteins (the core of an adaptation module) form a structure of two Cas1 dimers and one Cas2 dimer (Nuñez *et al.*, 2014), and this complex binds to a protospacer (Nuñez, Harrington, *et al.*, 2015; Wang *et al.*, 2015). Recent studies suggest that protospacers come from the leftovers of double-strand break (DSB) repair machinery (exonuclease RecBCD complex (Levy *et al.*, 2015)) action, which provides single-stranded DNA pieces that are located between chi (GCTGGTGG motif) sites on a chromosome or a plasmid (Dillingham and Kowalczykowski, 2008). The mechanism of insertion of a new spacer into a CRISPR array varies between different CRISPR-Cas types (Amitai and Sorek, 2016).

For Type I-E systems, the Cas1-Cas2 complex alone is sufficient for integration of spacers (Yosef, Goren and Qimron, 2012). The complex works as an integrase by providing nicks at ends at the first repeat (the repeat closest to a leader, a conserved upstream AT rich sequence) in the CRISPR array (Nuñez, Lee, *et al.*, 2015). For Type II-A systems, Cas9 is also necessary for adaptation. The presence of the effector complex is essential and probably responsible for PAM specificity of the adaptation process (Heler *et al.*, 2015; Wei, Terns and Terns, 2015). New spacers are preferentially incorporated at the leader-end presumably because the leader is recognized by the adaptation complex (Hille and Charpentier, 2016,Barrangou *et al.*, 2007; Garneau *et al.*, 2010; Jinek *et al.*, 2012).

Two modes of adaptation are observed in Type I CRISPR-Cas systems: naïve (non-primed) and primed. Naïve adaptation requires only the adaptation module to acquire new spacers. In cases where the CRISPR array already contains a spacer similar to the sequence of the invading phage, adaptation becomes much more efficient (Datsenko *et al.*, 2012; Swarts *et al.*, 2012; Savitskaya *et al.*, 2013; Shmakov *et al.*, 2014). This mode is called ‘primed adaptation’. Cascade effector, Cas3 nuclease-helicase that cleaves the target sequence, and the adaptation module proteins are all needed for this process (Datsenko *et al.*, 2012; Swarts *et al.*, 2012; Savitskaya *et al.*, 2013). Primed adaptation has been observed in Type I-B (Li *et al.*, 2014) and Type I-E (Datsenko *et al.*, 2012; Swarts *et al.*, 2012; Savitskaya *et al.*, 2013; Shmakov *et al.*, 2014), and Type I-F (Richter *et al.*, 2014) systems.

Biogenesis of crRNAs is a process of expression and maturation of crRNA. At this stage a CRISPR array is transcribed into pre-crRNA, which is then sliced into smaller pieces between 30 and 65 nt in length each containing a spacer and parts of repeats on one or on both ends (Brouns *et al.*, 2008; Charpentier *et al.*, 2015). The mechanism of crRNA generation varies between CRISPR-Cas systems. Systems with the *cas6* gene coding for metal-independent endoribonuclease that cleaves repeat sequences in RNA transcript (Carte *et al.*, 2008; Marraffini and Sontheimer, 2010a) have crRNAs that consists of an 8 nt part of a repeat on 5’ ends, the spacer itself and part of a repeat on 3’ ends, which may form a hairpin in some systems (Jore *et al.*, 2011). In some systems Cas6 may remain associated with the transcript after cleavage, or may be associated with effector complexes (Haurwitz *et al.*, 2010; Wiedenheft, Lander, *et al.*, 2011; Rollins *et al.*, 2015). Cas6 can work in-trans with CRISPR arrays of different systems located in the same genome (Majumdar *et al.*, 2015). In some systems Cas5d substitutes Cas6 to carry out the same function (Nam *et al.*, 2012).

Systems that have a single subunit effector complex (Cas9 or Cpf1) process the CRISPR array transcript using effector complex. Systems with Cas9 carry out biogenesis of crRNA using host RNase III and tracrRNA (trans acting CRISPR RNA) encoded close to CRISPR loci. These short RNAs have partial complementarity to repeats and form duplexes with a repeat sequences of pre-crRNA (Deltcheva *et al.*, 2011; Jinek *et al.*, 2012; Charpentier *et al.*, 2015). After pre-crRNA cleavage, the complex of tracrRNA and crRNA, binds Cas9 and causes a conformational change that allows Cas9 to seek and cleave target DNA (Deltcheva *et al.*, 2011; Gasiunas *et al.*, 2012; Jinek *et al.*, 2012). Systems with Cpf1 lack tracrRNA; instead their effectors possess an RNase activity and are able to process the pre-crRNA creating crRNA with a spacer and a part of the repeat that forms the 5’ hairpin (Fonfara *et al.*, 2016).

Interference is a process of introducing a break into foreign DNA or RNA by an effector complex bounded with crRNA, with subsequent degradation of the target (Brouns *et al.*, 2008; Garneau *et al.*, 2010). To avoid self-targeting (cleavage of CRISPR array) CRISPR-Cas systems that cleave DNA, in addition to spacer-protospacer complementarity, check for a valid PAM (Protospacer Adjacent Motif) at the target site. It has been shown that there are special parts of a spacer, namely the Seed region (an 8-nt spacer region close to the PAM), and that complementarity in the Seed region is most important for protospacer recognition (Semenova *et al.*, 2011; Wiedenheft, van Duijn, *et al.*, 2011; Sternberg *et al.*, 2014). Details of the recognition, cleavage and degradation of a target DNA or RNA are different in different effector complexes.

CRISPR-Cas Type I effector complexes, which consist of various Cas proteins in different systems (Makarova *et al.*, 2015), seek for a PAM sequence and then melt DNA at the seed position of the crRNA guide and form an initial, short R-loop with crRNA (Jore *et al.*, 2011; Redding *et al.*, 2015). In case of mismatch in the seed position, the R-loop formation stalls and this halts the interference by forbidding Cas3 docking (Blosser *et al.*, 2015). It may instead trigger primed adaptation by accepting the Cas1-Cas2, Cas3 complex (Redding *et al.*, 2015). In case of a match, formation of an R-loop extending over entire spacer-protospacer length is followed by Cas3 docking to the complex and DNA degradation (Makarova *et al.*, 2006; Brouns *et al.*, 2008; Rutkauskas *et al.*, 2015).

CRISPR-Cas Type III effector complexes can target DNA with Csm proteins (for Type III-A CRISPR-Cas systems) (Marraffini and Sontheimer, 2008) and RNA with Cmr proteins (for Type III-B systems) (Hale *et al.*, 2009; Staals *et al.*, 2013; Zebec *et al.*, 2014). Csm and Cmr are transcription-dependent nucleases (Staals *et al.*, 2014; Tamulaitis *et al.*, 2014); they recognize mRNA by complementarity with crRNA, then cleave RNA and/or transcribed DNA (Deng *et al.*, 2013; Goldberg *et al.*, 2014; Peng *et al.*, 2015; Samai *et al.*, 2015; Elmore *et al.*, 2016; Estrella, Kuo and Bailey, 2016). Csm3 or Cmr4 (Type I Cas7 analogs) build a protein backbone along crRNA spacer path (Taylor *et al.*, 2015) and cleave the target RNA into 6-nt pieces. Binding of a Cmr complex to complementary DNA activates Cas10 cleaving activity (Spilman *et al.*, 2013; Samai *et al.*, 2015; Taylor *et al.*, 2015; Elmore *et al.*, 2016; Estrella, Kuo and Bailey, 2016). The main distinguishing feature of Type III is that it does not require PAM for interference. It was proposed that, in order to avoid self-immunity, Cmr-related complexes rely on recognition of CRISPR repeats that block self-interference (Marraffini and Sontheimer, 2010b; van der Oost *et al.*, 2014). However, another study shows that some Type-III systems need an RNA PAM (Elmore *et al.*, 2016).

CRISPR-Cas Type II (*cas9* gene) and Type V (*cpf1* gene) are single-subunit crRNA effector complexes. Cas9 protein with crRNA forms a crRNP (CRISPR ribonucleoprotein complex) complex, which is responsible for recognition and degradation of target DNA (Gasiunas *et al.*, 2012). These actions take place in different parts of the effector complex

(Nishimasu *et al.*, 2014). A PAM at 3' of the protospacer is needed for melting upstream DNA that allows R-loop formation and cleavage in case of spacer-protospacer complementarity (Sternberg *et al.*, 2014) (recently a single-stranded targeting Cas9 was found (Zhang *et al.*, 2015), which does not require a PAM). Match in the spacer seed region (12nt closer to PAM), which forces further DNA melting and crRNA-DNA heteroduplex formation, activates Cas9 nuclease sites (HNH and RuvC) by triggering conformational change in the protein, allowing Cas9 to introduce a blunt-end double-strand break close to the 3' end of the protospacer (Nishimasu *et al.*, 2014; Sternberg *et al.*, 2015). Cpf1 is another single-subunit effector complex that introduced DSB into a DNA target. Bioinformatical approaches show that Cpf1 lacks an HNH nuclease domain and a second nuclease domain or that the cleavage mechanism has not yet been characterized. Its structure was resolved recently (Yamano *et al.*, 2016). A unique feature of Cpf1, compared with Cas9, is that it lacks tracrRNA and introduces staggered double-stranded breaks outside the protospacer (Zetsche *et al.*, 2015; Yamano *et al.*, 2016).

CRISPR-Cas discovery timeline

CRISPR-Cas systems were discovered quite recently and quickly acquired popularity and underwent quite thorough characterization. The first observations of unusual repetitive structures appeared 30 years ago and there has been a boom since 2012-2013, when CRISPR-Cas began to be applied to genome editing tasks (see Figure 2)

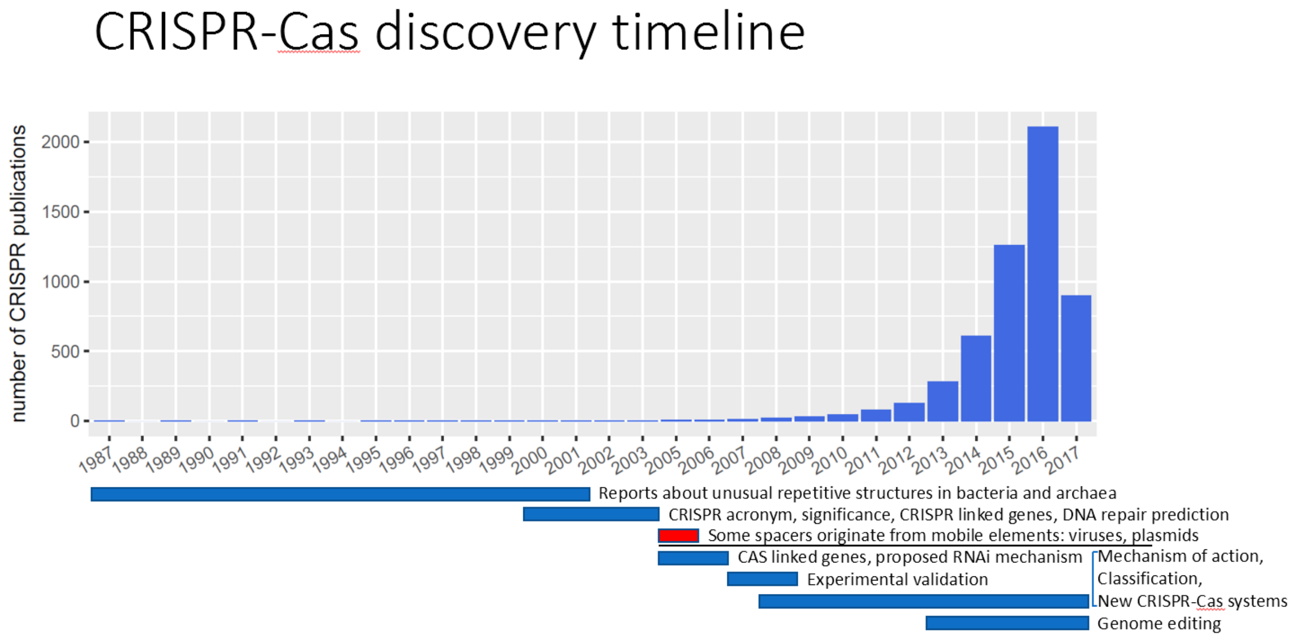


Figure 2. Number of publications per year mentioning CRISPR. The histogram shows the number of papers published per year with the keyword 'CRISPR', according to NCBI Pubmed and manual tag assignment. The timeline for different periods is shown below the histogram, with the most significant results described in text. The period shown in red pinpoints the crucial discovery that CRISPR-Cas is a prokaryotic immunity system.

Clusters of repeats separated by spacers were first observed in *Escherichia coli* in 1987 during study of the *iap* gene (Ishino *et al.*, 1987) and CRISPR array structure (the actual term was not used at the time) was described in 1989 (Nakata, Amemura and Makino, 1989). However, no functions for these loci were proposed at the time. The same nucleotide structures were later observed in different bacteria and archaea (Hoe *et al.*, no date; Hermans *et al.*, 1991; Mojica *et al.*, 1995; Bult *et al.*, 1996), but the observations did not arouse any great interest. The abundance and significance of CRISPR arrays were shown in 2000 (Mojica *et al.*, 2000). Two independent studies were made in 2002, one concerning an abundant gene family that was supposed to have DNA repair functionality (Makarova *et al.*, 2002) and another showing linkage of these genes to CRISPR arrays (Jansen *et al.*, 2002). The latter study also introduced the 'CRISPR' acronym. Increasing volumes of prokaryotic sequence data led to the crucial breakthrough in 2005, when studies detected the similarity of spacers to phage and plasmid sequences and proposed that this similarity is what gives immunity to infections (Bolotin *et al.*, 2005; Mojica *et al.*, 2005; Pourcel, Salvignol and Vergnaud, 2005). Some 45 protein families linked to CRISPR were observed at the time (Haft *et al.*, 2005). A year later, in 2006, the RNA interference mechanism was proposed as the mode of action of CRISPR-Cas systems (Makarova *et al.*, 2006). Evidence and proof of the action of CRISPR-Cas as an immune system followed in 2007 (Barrangou *et al.*, 2007). Later studies described the organization and details of CRISPR-Cas action: that it is RNA mediated (Brouns *et al.*, 2008), that it targets DNA (Marraffini and Sontheimer, 2008) and RNA (Hale *et al.*, 2009), that the Protospacer Adjacent Motif is required for some systems (Mojica *et al.*, 2009). Details of pre-crRNA transcription and processing were also revealed (Haurwitz *et al.*, 2010).

After 2010 came a period of unification of knowledge and further characterization of CRISPR-Cas building blocks. The first classification of CRISPR-Cas systems was provided in 2011 (Makarova, Haft, *et al.*, 2011). In the same year cryo-electron microscopy showed how Cascade complex building blocks are arranged with crRNA in crRNP (Wiedenheft, Lander, *et al.*, 2011). In 2011 Type II systems were further characterized by discovery of the role of RNase III in processing of pre-crRNA and tracrRNA (Deltcheva *et al.*, 2011). Different modes of adaptation were described in 2012 (Datsenko *et al.*, 2012), improving the understanding of kinetics of infection of cells with CRISPR-Cas systems.

Studies were carried out in 2012 that sparked a revolution in genome editing based on programming of Cas9 to enable cleavage of designated DNA (Jinek *et al.*, 2012; Qi *et al.*, 2012). This was followed in 2013 by works showing genome editing in human (Cong *et al.*, 2013; Jinek *et al.*, 2013; Mali *et al.*, 2013), bacteria (Gasiunas *et al.*, 2012; Jinek *et al.*, 2012) and yeast cells (DiCarlo *et al.*, 2013). These advances resulted in the first clinical trials of CRISPR-Cas genetically modified cells, which were approved in 2016 (Cyranoski, 2016; Reardon, 2016).

CRISPR-Cas classification

CRISPR-Cas is one of the actors in a never-ending ‘arms race’ between bacterial/archaeal hosts and viruses, which forces great diversity in genes, CRISPR-Cas loci composition and mechanisms of action among CRISPR-Cas systems (Makarova *et al.*, 2015). However, these systems share common features and classification has recently been made based on these features (Makarova, Haft, *et al.*, 2011; Makarova *et al.*, 2015). There are two levels of classification: by class, and by type/subtype (Makarova *et al.*, 2015).

Separation of CRISPR-Cas systems by classes is based on structure of the effector complex (see Figure 3).

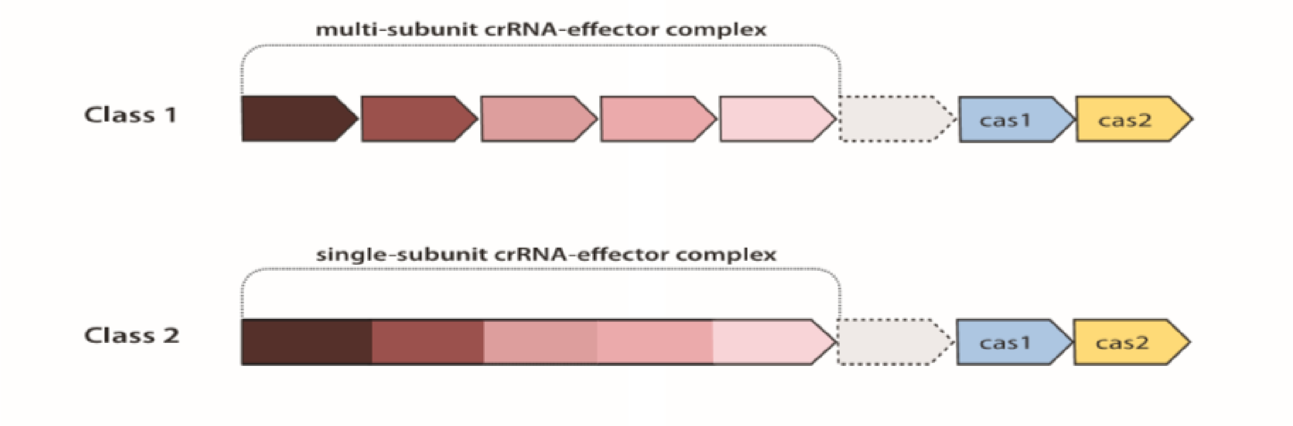

multi-subunit crRNA-effector complex
Class 1
cas1
cas2
single-subunit crRNA-effector complex
Class 2
cas1
cas2

Figure 3. Separation of CRISPR-Cas systems into two classes (reproduced with permission from Science (Mohanraju *et al.*, 2016)). Two classes of CRISPR-Cas systems are shown: Class 1 contains systems with multiprotein effector complexes, while Class 2 contains single-protein effector complexes. Proteins of effector complexes are shown in reddish colors, and the gradations of red represent different biochemical functions. Accessory proteins are shown with broken outlines. The adaptation module (*cas1*, *cas2*) genes are located at the end of the loci.

Class 1 CRISPR–Cas systems, which have multi-subunit crRNA-effector complexes (Type I, Type III, Type IV), are most common in bacteria and archaea (including in all hyperthermophiles), comprising ~90% of all identified CRISPR–Cas loci (Makarova *et al.*, 2015). Functional roles are split between proteins of the effector complexes. The remaining ~10% of CRISPR–Cas loci belong to Class 2 CRISPR–Cas systems (which use Cas9 and Cpf1 effector proteins). These systems are found almost exclusively in bacteria (a few instances of Cpf1 are found in archaea) and have not been identified in hyperthermophiles (Chylinski *et al.*, 2014; Makarova *et al.*, 2015). All activities (except for adaptation) of the Class 2 systems are processed by a single protein (Cas9, Cpf1).

The second classification level for CRISPR-Cas systems is classification by types, which uses signature genes and gene architecture to distinguish 5 types and 16 subtypes of CRISPR-Cas systems (see Figure 4).

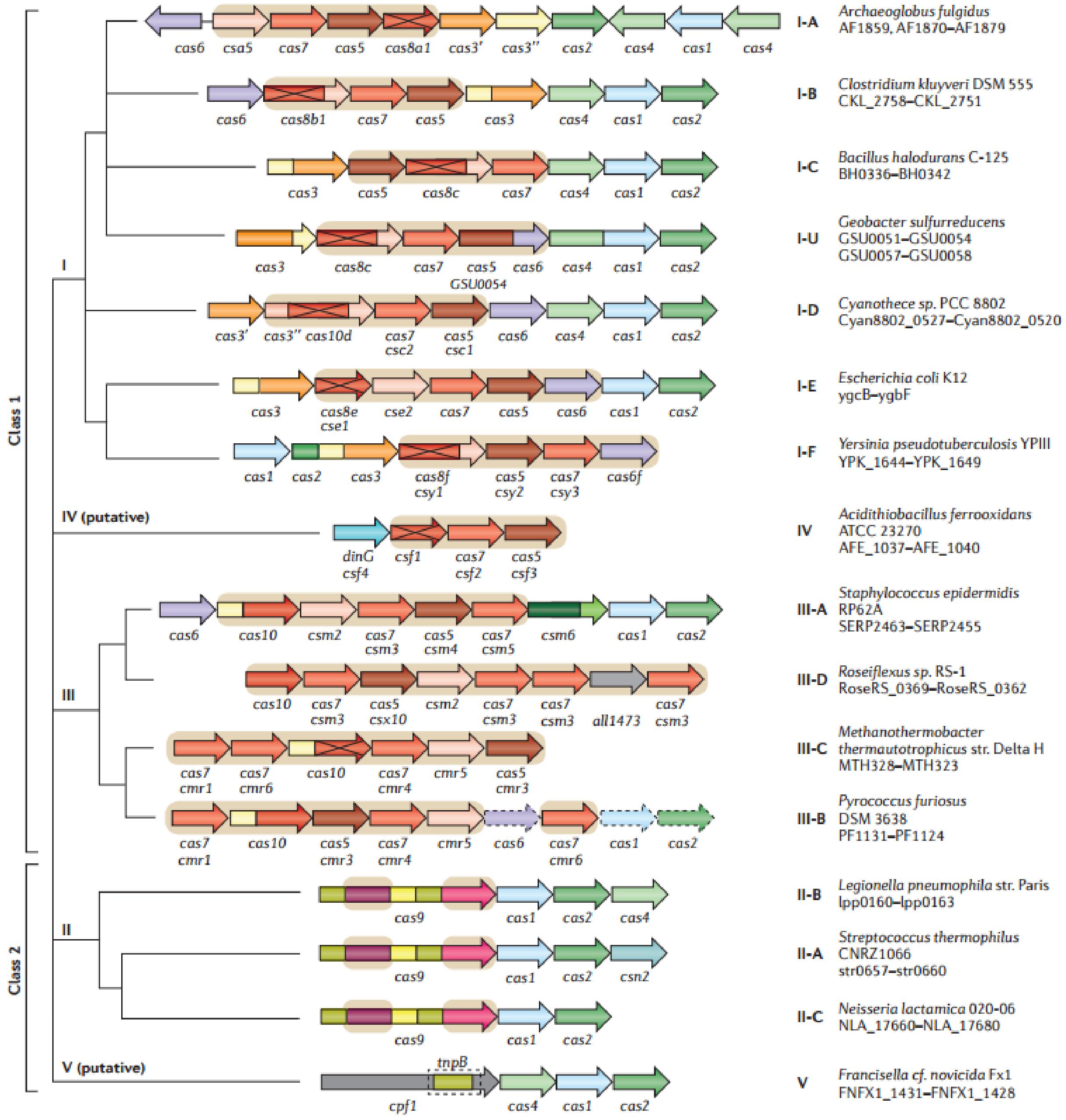

Class 1
Class 2
I
IV (putative)
III
II
V (putative)
cas6 csa5 cas7 cas5 cas8a1 cas3′ cas3″ cas2 cas4 cas1 cas4
I-A
Archaeoglobus fulgidus
AF1859, AF1870–AF1879
cas6 cas8b1 cas7 cas5 cas3 cas4 cas1 cas2
I-B
Clostridium kluyveri DSM 555
CKL_2758–CKL_2751
cas3 cas5 cas8c cas7 cas4 cas1 cas2
I-C
Bacillus halodurans C-125
BH0336–BH0342
cas3 cas8c cas7 cas5 cas6 cas4 cas1 cas2
GSU0054
I-U
Geobacter sulfurreducens
GSU0051–GSU0054
GSU0057–GSU0058
cas3′ cas3″ cas10d cas7 csc2 cas5 csc1 cas6 cas4 cas1 cas2
I-D
Cyanothece sp. PCC 8802
Cyan8802_0527–Cyan8802_0520
cas3 cas8e cse1 cse2 cas7 cas5 cas6 cas1 cas2
I-E
Escherichia coli K12
ygcB–ygbF
cas1 cas2 cas3 cas8f csy1 cas5 csy2 cas7 csy3 cas6f
I-F
Yersinia pseudotuberculosis YPIII
YPK_1644–YPK_1649
dinG csf4 csf1 cas7 csf2 cas5 csf3
IV
Acidithiobacillus ferrooxidans
ATCC 23270
AFE_1037–AFE_1040
cas6 cas10 csm2 cas7 csm3 cas5 csm4 cas7 csm5 csm6 cas1 cas2
III-A
Staphylococcus epidermidis
RP62A
SERP2463–SERP2455
cas10 cas7 csm3 cas5 csx10 csm2 cas7 csm3 cas7 csm3 all1473 cas7 csm3
III-D
Roseiflexus sp. RS-1
RoseRS_0369–RoseRS_0362
cas7 cmr1 cas7 cmr6 cas10 cas7 cmr4 cmr5 cas5 cmr3
III-C
Methanothermobacter
thermautotrophicus str. Delta H
MTH328–MTH323
cas7 cmr1 cas10 cas5 cmr3 cas7 cmr4 cmr5 cas6 cas7 cmr6 cas1 cas2
III-B
Pyrococcus furiosus
DSM 3638
PF1131–PF1124
cas9 cas1 cas2 cas4
II-B
Legionella pneumophila str. Paris
lpp0160–lpp0163
cas9 cas1 cas2 csn2
II-A
Streptococcus thermophilus
CNRZ1066
str0657–str0660
cas9 cas1 cas2
II-C
Neisseria lactamica 020-06
NLA_17660–NLA_17680
tnpB
cpf1 cas4 cas1 cas2
V
Francisella cf. novicida Fx1
FNFX1_1431–FNFX1_1428

Figure 4. CRISPR-Cas classification by types and subtypes (reproduced with permission from Nature Reviews Microbiology (Makarova *et al.*, 2015)). Figure 4 shows two classes of CRISPR-Cas, consisting of five types and 16 subtypes. Loci are shown for each subtype, with genes represented by arrows. Colors represent homologous genes. Systematic gene names are shown under the arrows with the common gene name (for example, *cas7* and *cmr1* for Type III-B). Genes marked with a cross have inactivated large subunits of their effector complexes. Genes or domains involved in interference are shown on a brown background. Genes that are not present in all systems have broken outlines.

Class 1 CRISPR-Cas systems are divided into Type I, Type III and Type IV with 12 subtypes. The signature of Type I systems is the *cas3* gene, coding for a single-stranded DNA (ssDNA)-stimulated superfamily 2 helicase, which is responsible for unwinding double stranded DNA or RNA-DNA duplexes (Sinkunas *et al.*, 2011; Gong *et al.*, 2014; Huo *et al.*, 2014). Cas3 may contain an HD nuclease domain (or this domain might be located in an protein encoded by an adjacent gene), which is responsible for target DNA cleavage (Mulepati and Bailey, 2011; Sinkunas *et al.*, 2011). Seven subtypes of Type I have been identified: Type I-A to Type I-F and Type I-U (U stands for uncharacterized; the mechanism of action for this system remains unknown (Makarova *et al.*, 2015)). All subtypes shown in Figure 4 are defined by unique combinations of Cas proteins. In most cases, Type I system genes are located in one operon in the genome, except for Type I-A and Type I-B (Vestergaard, Garrett and Shah, 2014). The phylogenetic tree of the Cas3 protein reflects classification of Type I (Jackson, Lavin, *et al.*, 2014), and a more detailed study has led to the proposal of a polyphyletic origin of these systems (Makarova *et al.*, 2015).

CRISPR-Cas Type III systems are characterized by the presence of Cas10, a multidomain protein, which contains an RNA recognition domain (Palm domain) and is often fused with an HD nuclease domain (different from the HD domain in Type I (Makarova *et al.*, 2006)). This protein is the largest of the Type III Cas proteins and is very diverse (Makarova *et al.*, 2015). Cas5 and several Cas7 proteins are also present Type III systems. There are four Type III subtypes – Type III-A to Type III-D – with distinct gene sets (see Figure 4). All of these systems have been found to co-transcriptionally target DNA (Marraffini and Sontheimer, 2008; Deng *et al.*, 2013; Goldberg *et al.*, 2014; Peng *et al.*, 2015; Samai *et al.*, 2015) and RNA (Hale *et al.*, 2009, 2012; Spilman *et al.*, 2013; Staals *et al.*, 2014; Tamulaitis *et al.*, 2014; Samai *et al.*, 2015). Phylogenetic analysis of *cas10* gives results that are compatible with the classification of Makarova *et al.*, 2015. In some cases Type III systems lack an adaptation module and it was proposed that these systems may use crRNAs produced from systems CRISPR arrays (Makarova *et al.*, 2015).

Type IV is a putative CRISPR-Cas system that has not been functionally characterized yet. This system lacks *cas1* and *cas2* genes. In many cases there is no CRISPR array nearby in the genome or there is even no CRISPR array in the genome at all. A Type IV locus usually encodes Cas5, Cas7 and the signature Csf1 protein (see Figure 4) (Makarova *et al.*, 2015). It is also predicted that it might rely on crRNAs produced from CRISPR arrays of other systems

Class 2 Type II CRISPR-Cas systems are specified by the *cas9* gene, which is a single subunit effector. Type II systems are simplest in terms of gene count: there is only *cas9* and the adaptation module (*cas1*, *cas2* and *cas4*). Cas9 targets DNA with two different nuclease domains, HNH and RuvC (Jinek *et al.*, 2012), and is involved in adaptation (Heler *et al.*, 2015; Wei, Terns and Terns, 2015). The hallmark of the Type II system is tracrRNA – a short RNA that is encoded inside the CRISPR-Cas locus, and forms a crRNA-tracrRNA complex via complementary to repeat sequence. It is needed to process pre-crRNA and becomes part of the complex with Cas9 (Chylinski, Le Rhun and Charpentier, 2013; Briner *et al.*, 2014; Chylinski *et al.*, 2014). It was proposed, based on sequence similarity, that Cas9 originated from an IscB transposable element (Chylinski *et al.*, 2014). This complicates detection of Type II systems since an adaptation module must be found near a *cas9*-like gene before it can be asserted that it is part of a CRISPR-Cas system. There are three Type II subtypes – Type II-A, Type II-B and Type II-C (the most abundant CRISPR-Cas system in bacteria (Makarova *et al.*, 2015)) – and these systems have distinct loci compositions (see Figure 4). According to sequence homology, Type II-A and Type II-B are monophyletic, while Type II-C have a distinct origin (Chylinski *et al.*, 2014; Makarova *et al.*, 2015).

Class 2 Type V is a CRISPR-Cas system that has been recently discovered and characterized (Schunder *et al.*, 2013; Zetsche *et al.*, 2015; Fonfara *et al.*, 2016) with *cpf1* as a signature gene. The usual locus composition consists of *cpf1* and an adaptation module (*cas1* and *cas2*) (see Figure 4). Class 2 Type V stands apart from Type II Cas9 systems for several reasons: it has different origin (it is homologous to proteins from the IS605 transposon family (Makarova *et al.*, 2015)); it has different protein structure (it has a RuvC nuclease domain, but lacks an HNH domain (Makarova *et al.*, 2015; Yamano *et al.*, 2016)). Also it does not use tracrRNA and thus processes pre-crRNA differently (Fonfara *et al.*, 2016). Unlike Type II, this system was also found in archaea (Vestergaard, Garrett and Shah, 2014).

Origins of CRISPR-Cas genes

Modular structure of CRISPR-Cas systems and study of these modules suggests an explanation for the origin of CRISPR-*cas* genes. The most conserved and most abundant is the adaptation module that consists of *cas1*, *cas2* and (for some systems) *cas4*, which, upon merging with *cas10*, gave birth to the ancestral Class 1 CRISPR-Cas system (Mohanraju *et al.*, 2016) (see Figure 5).

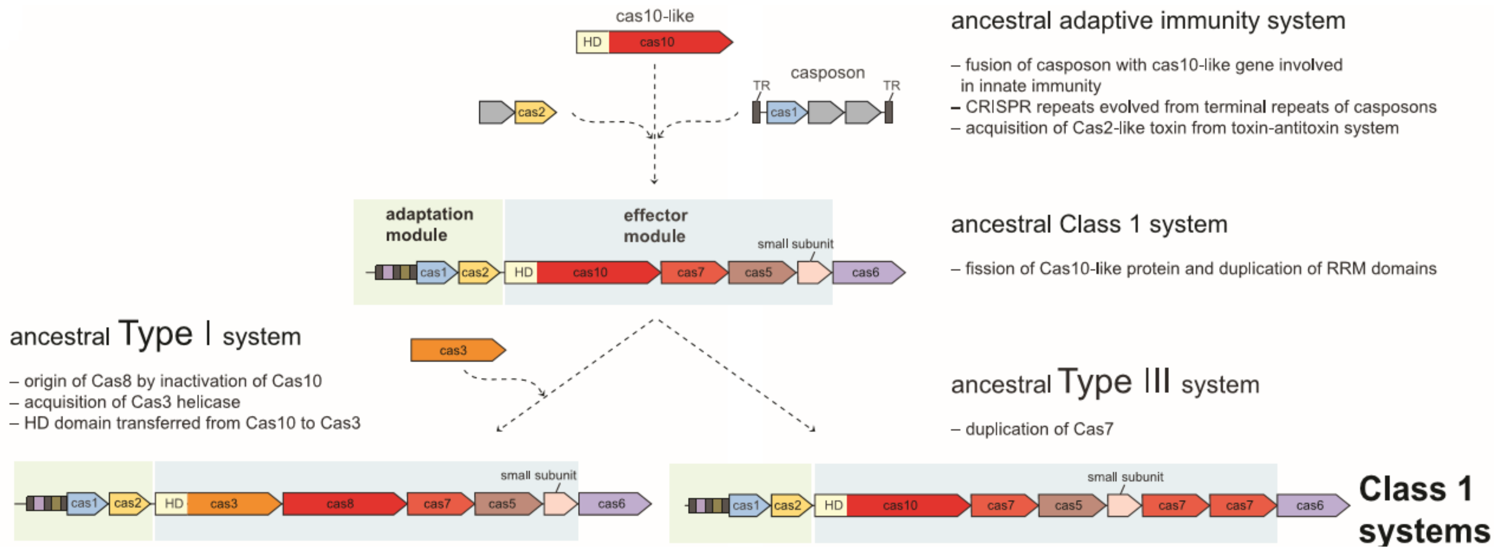

cas10-like
HD
cas10
cas2
casposon
TR
cas1
TR
ancestral adaptive immunity system
– fusion of casposon with cas10-like gene involved in innate immunity
– CRISPR repeats evolved from terminal repeats of casposons
– acquisition of Cas2-like toxin from toxin-antitoxin system
adaptation module
effector module
small subunit
cas1
cas2
HD
cas10
cas7
cas5
cas6
ancestral Class 1 system
– fission of Cas10-like protein and duplication of RRM domains
ancestral Type I system
– origin of Cas8 by inactivation of Cas10
– acquisition of Cas3 helicase
– HD domain transferred from Cas10 to Cas3
cas3
ancestral Type III system
– duplication of Cas7
small subunit
cas1
cas2
HD
cas3
cas8
cas7
cas5
cas6
small subunit
cas1
cas2
HD
cas10
cas7
cas5
cas7
cas7
cas6
Class 1 systems

Figure 5. Suggested evolutionary scenario for the origin of CRISPR-Cas systems (reproduced with permission from Science, adapted from (Mohanraju *et al.*, 2016)). This figure shows the latest theory as to the possible origin of Class 1 CRISPR-Cas systems through the merging of various genetic components. Genes are represented by arrows; genes on a gray background are thought to have been present in original loci but were lost during evolution of CRISPR-Cas systems. Green background shows a CRISPR-Cas adaptation module and blue background represents effector complex genes. TR stands for terminal repeats, TS for terminal sequences and HD for an HD-family endonuclease.

It is suggested that the *cas1* gene, which has nuclease and integrase functions (their origin is determined by homology with casposons, which are self-synthesizing transposons (Krupovic *et al.*, 2014, 2016; Hickman and Dyda, 2015)), should have been inserted close to the *cas10* homolog, which contained RNA binding nuclease domains and also served as an immunity system (Mohanraju *et al.*, 2016). It is proposed that the CRISPR array-like structures are remnants of inverted terminal repeats of the transposon (Koonin and Krupovic, 2015). The *cas2* gene, which originated (as detected by homology) from a toxin-antitoxin complex (Makarova, Aravind, *et al.*, 2011; Makarova, Wolf and Koonin, 2013b), was either in the casposon or in the immunity locus (the target of the casposon). It is proposed that *cas10* originated from a merger between a Cas10-like nuclease and one or more RRM (RNA Recognition Motif) folds of a polymerase of cyclase proteins, which evolved together with *cas1* and *cas2* and gave rise to an ancestral CRISPR-Cas system (Makarova, Aravind, *et al.*, 2011; Makarova, Wolf and Koonin, 2013b) (see Figure 5).

Class 1 CRISPR-Cas systems, are the most abundant in both bacteria and archaea (Makarova *et al.*, 2015), suggesting that the ancestral CRISPR-Cas system had the same architecture (Mohanraju *et al.*, 2016). CRISPR-Cas Type I and Type III were derived by recombination and influx of new genes from mobile elements into the ancestral system. The suggested scenario for the Type I system is: inactivation of *cas10* (origin of *cas8*) and acquisition of Cas3-like helicase with transfer of the HD nuclease domain. The Type III system, which arose from the same causes (recombination and influx of new elements from defense islands (Makarova, Wolf, *et al.*, 2011)), have duplicated the *cas7* gene in the locus.

It has been suggested that Class II systems (all subtypes of Type II and Type V) are the result of replacement of the Class 1 effector complex by nuclease proteins that originate from various mobile genetic elements (Mohanraju *et al.*, 2016). It was shown in recent studies that the *cas9* gene has parts, which are homologous to IscB transposases (Kapitonov, Makarova and Koonin, 2015), which have RuvC and HNH nuclease domains, whereas Type V have RuvC and do not have a detectable HNH domain or close homology to *cas9* or IscB (Zetsche *et al.*, 2015).

Application of CRISPR-Cas systems

CRISPR-Cas systems have been found to be a highly efficient (Chen *et al.*, 2014) and easy-to-use programmable tool for genome editing of prokaryotes and various eukaryotes including plants and animals. The techniques of genome editing with CRISPR-Cas systems have been described in numerous reviews (Ran, Hsu, Wright, *et al.*, 2013; Hsu, Lander and Zhang, 2014; Barrangou and Doudna, 2016; Mohanraju *et al.*, 2016). The first genome editing with a CRISPR-Cas Type II effector complex was shown on human cells in 2013 (Cho *et al.*, 2013; Cong *et al.*, 2013; Jinek *et al.*, 2013; Mali *et al.*, 2013). In these studies, Cas9 was shown to be a tool that can be programmed by sgRNA (single guide RNA, which is an artificial RNA construct containing: a hairpin (handle for Cas9), a terminator sequence and a spacer sequence) to make double stranded breaks into DNA, which can be used to introduce indels using non-homologous end joining DNA repair machinery or for the insertion of new DNA material by providing templates for homology directed repair (see Figure 6). Although Cas9 proved to be efficient, it has limitations, due to PAM and/or target sequence specificity. But several variants of Cas9 recognizing different PAMs have been characterized and used (Deveau *et al.*, 2008; Horvath *et al.*, 2008; Zhang *et al.*, 2013). New artificially modified variants of Cas9, which have improved specificity due to reduction of interaction with non-specific DNA sites, have also been introduced (Kleinstiver, Pattanayak, *et al.*, 2016; Slaymaker *et al.*, 2016). Another workaround for increasing specificity or reducing off-target events has been achieved by creating Cas9 dimer (Mali, Esvelt and Church, 2013; Ran, Hsu, Lin, *et al.*, 2013), a protein, in which the nuclease sites have been mutated so that each dimer can introduce only one nick. Cas9 dimer requires two sites to achieve recognition, so it increases the sequence specificity of the complex.

Genome editing operations that are often performed with CRISPR-Cas effector complexes (Bortesi and Fischer, no date) include: gene knockouts by introduction of a double-strand break (DSB) with Cas9, causing frame shifts arising after non-homologous-ends-joining (NHEJ)

DNA repair; insertion of new DNA material with NHEJ after introduction of a DSB; insertion of new material after inducing a DSB with homologous recombination (HR) by adding a DNA template for HR; and gene modification by template with HR. These operations are shown in Figure 6.

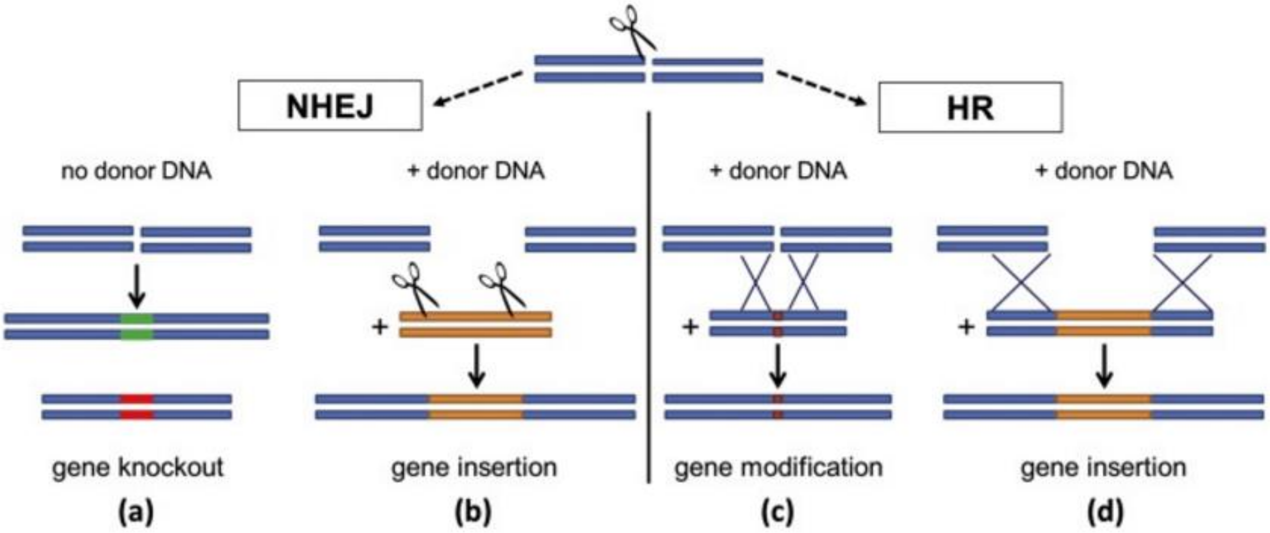


Figure 6. Common techniques of genome editing with Cas9 (reproduced with permission from Biotechnology Advances (Bortesi and Fischer, no date). Genome editing with Cas9 (represented by scissors) is shown in the Figure. Four scenarios are described: a) gene knockout by indels frameshift; b) insertion of a new gene; c) gene modification by means of a template; d) gene insertion from template DNA with homologous recombination

The general approach for genome editing includes the following operations:

1. Search for a sequence or sequences with a valid protospacer adjacent motif, which can be targeted using the selected variant of Cas9.
2. Minimizing off-target effects of the effector complex (by selecting a unique sequence in the genome and a sequence composition that is most preferable for the chosen Cas9 (Doench *et al.*, 2016)).
3. Design and synthesis of required sgRNA.

4. Making the Cas9/sgRNA construct.
5. Delivery of the complex into the cell or group of cells.
6. Validation of results.

Genome editing is not the sole application of CRISPR-Cas effector complexes. Inactivated variants of Cas9 or dead Cas9 (dCas9), which have disrupted HNN and RuvC nuclease domains and are therefore unable to introduce DSB, can be used for site-specific, non-nuclease activities, such as: activation of transcription or transcription repression (Bikard *et al.*, 2013; Gilbert *et al.*, 2013; Konermann *et al.*, 2013; Qi *et al.*, 2013); as a fluorescent label (Chen *et al.*, 2013); or for recruiting histone modification proteins (Hilton *et al.*, 2015; Kearns *et al.*, 2015). Recent studies show potential for using CRISPR-Cas effector complexes to create logic circuits for activation/repression cascades (Kiani *et al.*, 2014; Nissim *et al.*, 2014), as well as AND (logical conjunction) circuits for detection of bladder cancer cells by activation of luciferase (Liu *et al.*, 2014). Sequence-specific anti-microbial techniques have been introduced, based on the delivery of Cas9 by bacteriophages that destroy antibiotic resistance plasmids (Bikard *et al.*, 2014), and antiviral systems that can suppress hepatitis B or HIV-1 have also been described (Ebina *et al.*, 2013; Hu *et al.*, 2014; Ramanan *et al.*, 2015). Another approach shows that CRISPR-Cas effector complexes can be used for loss-of-function genetic screening for positive and negative selection in mammalian cells: a large pool of sgRNA is produced, targeting the regions of interest, then a large pool of mutants is generated by introducing Cas9 (Gilbert *et al.*, 2014; Shalem *et al.*, 2014; Wang *et al.*, 2014; Konermann *et al.*, 2015).

The recent discovery of Cpf1 (Zetsche *et al.*, 2015) – another CRISPR-Cas Class 2 protein – showed that there is a place for new effector complexes in genome editing or other applications of sequence specific nucleases. Cpfl has the same or better off-target properties (D. Kim *et al.*, 2016; Kleinstiver, Tsai, *et al.*, 2016) and it can be used in a simplified process of genome editing. It does not require tracrRNA (Zetsche *et al.*, 2015), so the complexity of effector complexes is reduced; it is able to process its own CRISPR array (Zetsche *et al.*, 2015), which makes multiple targeting easier (there is no need to provide multiple sgRNA genes; only one CRISPR array with required spacers is needed); and Cpf1 nuclease creates sticky ends (Zetsche *et al.*, 2015), which can be used to insert new DNA more efficiently (Mohanraju *et al.*, 2016).

The problems that remain to be solved for the application of CRISPR-Cas effectors include the efficient and tissue-specific delivery of CRISPR-Cas effectors as well as ethical questions of genome editing applications (Mohanraju *et al.*, 2016). Advances in delivery technology, which include the use of smaller Cas9 (Ran *et al.*, 2015), delivery by nanoparticles (Platt *et al.*, 2014), delivery by electroporation (Qin *et al.*, 2015) and delivery by micropinocytosis (D'Astolfo *et al.*, 2015), show the need for new site-specific nucleases. CRISPR-Cas class 1 effector complexes, which are not widely used for genome editing, may attract future attention for this reason.

Materials and Methods

Dataset

A search for novel CRISPR-Cas systems was carried out on various prokaryotic data sets. In the first part of the study, the search for *cas1* associated proteins used WGS and NT NCBI databases (‘Database resources of the National Center for Biotechnology Information’, 2016). For the CRISPR associated protein search in the second part of the study a separate prokaryotic database was assembled. Archaeal and bacterial genome sequences were downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/) in March 2016. For incompletely annotated genomes (coding density less than 0.6 coding DNA sequence per kbp) the existing open reading frames annotation was discarded and replaced with annotation by Meta-GeneMark (Besemer, Lomsadze and Borodovsky, 2001) with the standard model MetaGeneMark_v1.mod (Heuristic model for genetic code 11 and GC 30). Altogether the database includes 4,961 completely sequenced and assembled genomes and 43,599 partially sequenced genomes (altogether represented by 6,342,452 contigs, composed from 33,803 unique taxonomic group and 12,528 unique species, coding 182,301,555 proteins).

Pipeline for Annotation of CRISPR-Cas Loci

The pipeline takes a list of locations (coordinates in the corresponding nucleotide sequence) of the seed features (*cas1* or CRISPR) as input. Two types of seeds were used: locations of *cas1* genes in the NCBI, NR and WGS database (‘Database resources of the National Center for Biotechnology Information’, 2016) and locations of CRISPR arrays in the WGS and prokaryotic genome database. CRISPR and *cas1* seed sets were not merged and were used separately. TBLASTN searches (Altschul *et al.*, 1997) with E-value cutoff of 0.01 and low complexity filtering turned off were run against NR and WGS with the Cas1 profiles (Makarova and Koonin, 2015) as queries, resulting in the identification of 20,766 loci. The CRISPRfinder (Grissa, Vergnaud and Pourcel, 2007) and PILER-CR (Edgar, 2007) programs were used with default parameters to identify CRISPR arrays in the WGS database (47,174 loci found) and in the prokaryotic genome database (45,373 loci found). Sequences including up to 10 kbp upstream and downstream from the seed features were extracted.

Open Reading Frame (ORF) annotation was performed using Meta-GeneMark (Zhu, Lomsadze and Borodovsky, 2010) with the MetaGeneMark_v1.mod standard model (Heuristic model for genetic code 11 and GC 30). All ORFs were further annotated using RPS-BLAST (Marchler-Bauer *et al.,* 2002) searches with 30,953 profiles (COG, pfam, cd) from the NCBI CDD database (Marchler-Bauer *et al.*, 2013; 'Database resources of the National Center for Biotechnology Information', 2016) and 217 custom Cas protein profiles 10. The CRISPR-Cas system (sub)type identification for all loci was performed using procedures, which have been previously described (Makarova and Koonin, 2015; Makarova *et al.,* 2015).

For *cas1* seeds from NR and WGS databases, 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 with composition based-statistics and low complexity filtering turned off. Each non-redundant protein identified in this search was searched against the WGS database using the TBLASTN program (Altschul et al., 1997). The HHpred program was used with default parameters to identify remote sequence similarity using as the queries all proteins identified in the BLAST searches (Soding et al., 2006). Multiple sequence alignments were constructed using MUSCLE (Edgar, 2004) and MAFFT (Katoh and Standley, 2013).

Later, a clustering approach was implemented for *cas1* and CRISPR seeds (see Clustering). Potential candidates were selected out of all permissive clusters constructed from proteins from the seed loci using the size threshold (> 500aa) and distance to seed (genes closest to the seed were preferred); the selection of candidates was limited to those that were located within 4 genes from the seed; clusters that contained more homologs outside the seed loci than in those loci were discarded. Additional prediction of protein domains was performed using the CD-search (Edgar, 2010) and HHpred (Söding *et al.*, 2006).

The identified candidates were used as queries for a PSI-BLAST search against the NCBI NR and NCBI WGS databases for the *cas1* seeds, and NCBI WGS and prokaryotic databases for the CRISPR seeds in order to obtain additional loci that were added to the seed list. The evaluation procedure was then repeated until convergence occurred.

Clustering and Phylogenetic Analysis

To construct a non-redundant, representative sequence set, sequences were clustered using the NCBI BLASTCLUST program (Wheeler and Bhagwat, 2007) (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) with the sequence identity threshold of 90% and length coverage threshold of 0.9. The longest sequence was selected to represent each cluster. Permissive clustering of sequences was performed using UCLUST (Edgar, 2010), with sequence similarity threshold of 0.3.

Multiple alignments of protein sequences were constructed using MUSCLE (Edgar, 2004) and MAFFT (Katoh and Standley, 2013) programs. Sites with the gap character fraction values > 0.5 and homogeneity < 0.1 (Yutin *et al.*, 2008) were removed from the alignment. Phylogenetic analysis was performed using the FastTree program (Price, Dehal and Arkin, 2010), with the WAG evolutionary model and the discrete gamma model with 20 rate categories.

Relationships within diverse sequence families were established using the following procedure: initial sequence clusters were obtained using UCLUST (Edgar, 2010) with a sequence similarity threshold of 0.5; sequences were aligned within clusters using MUSCLE (Edgar, 2004). Then cluster-to-cluster similarity scores were obtained using HHSEARCH (Söding, 2005) (including trivial clusters consisting of a single sequence each) and a UPGMA dendrogram was constructed from the pairwise similarity scores. Highly similar clusters (pairwise score to self-score ratio >0.1) were aligned to each other using HHALIGN (Söding, 2005) and the procedure was repeated iteratively. At the last step, sequence-based trees were reconstructed from the cluster alignments using the FastTree program (Price, Dehal and Arkin, 2010) as described above and rooted by mid-point; these trees were grafted to the tips of the profile similarity-based UPGMA dendrogram.

Protospacer Analysis

The initial pool of 488,437 spacers in the CRISPR arrays was reduced to 268,409 unique sequences. The MEGABLAST program (Zhang *et al.*, no date) (word size 18) was used to search for protospacers in the virus subset of the NR database (TaxID:10239) and the prokaryotic genome database. The maximum number of mismatches for a spacer with length l was limited to $\max(0,\ l-22)$. All MEGABLAST hits that target CRISPR arrays as well as all eukaryotic virus sequences were discarded. This procedure produced 63,939 hits to prokaryotic genomes and 5,095 hits to prokaryotic viruses. The 33,480 ORFs that contained or intersected with the detected protospacers were used as BLASTP queries to search the virus database. All ORFs with strong hits (e-value <10-6) were classified as originating from (pro)viruses.

Synteny Analysis of Subtype V-U loci

Protein sequences encoded by genes in the vicinity (±3 genes) of the Type V-U effector genes were extracted and clustered using UCLUST (Edgar, 2010) with similarity threshold of 0.3. Genes were annotated by the cluster IDs; each locus was represented as a set of genes and unordered gene pairs. A weighted Jackard similarity coefficient was calculated for all pairs of loci as previously described (Makarova *et al.*, 2015), a locus similarity graph was constructed with a similarity threshold of 0.61 (e-0.5), and connected components (subsets of highly similar loci) were extracted.

Analysis of Selection in the Evolution of Class 2 Effector Genes

Nucleotide and protein sequences of the effector genes were collected; clusters of identical protein sequences were reduced to a single representative; remaining sequences were clustered using UCLUST (Edgar, 2010) with a similarity threshold of 0.67. The sequences from each cluster were aligned, and a phylogenetic tree was constructed as described above and rooted using a modified midpoint procedure. Sub-alignments of protein sequences, corresponding to sub-trees with mean depth <0.1, were extracted and converted to the nucleotide sequence alignments. Pairwise dN, dS and dN/dS values were obtained using the codeml program of the PAML package (Yang, 2007). Sequence pairs with $0.0002 \leq$ dN ≤ 1.0 and $0.0002 \leq$ dS ≤ 1.0 were selected, and the dN/dS values were calculated.

Results and Discussion

Part 1: Novel lass 2 CRISPR-Cas Systems

The computational pipeline for the discovery of Class 2 CRISPR–Cas loci

We developed a computational pipeline for the systematic detection of Class 2 CRISPR–Cas systems (see Figure 7).

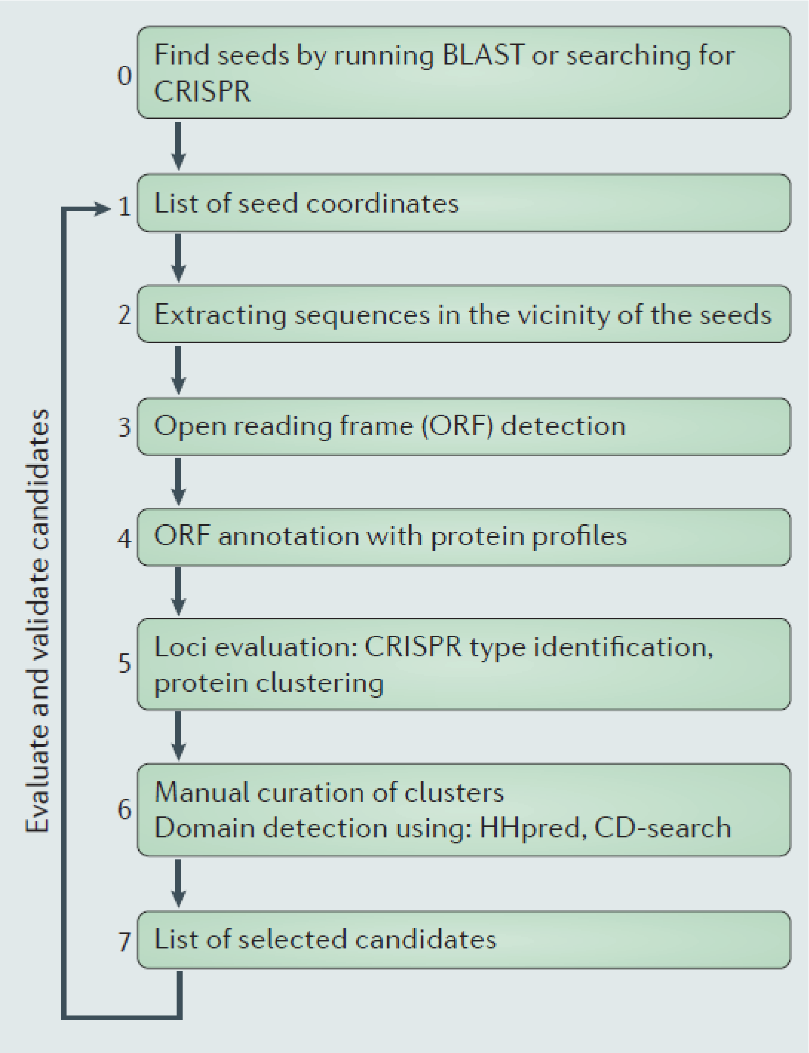

0 Find seeds by running BLAST or searching for CRISPR
1 List of seed coordinates
2 Extracting sequences in the vicinity of the seeds
3 Open reading frame (ORF) detection
4 ORF annotation with protein profiles
5 Loci evaluation: CRISPR type identification, protein clustering
6 Manual curation of clusters
Domain detection using: HHpred, CD-search
7 List of selected candidates
Evaluate and validate candidates

Figure 7. CRISPR effector discovery pipeline for search of CRISPR or Cas1 associated proteins. The computational pipeline for the discovery of Class 2 CRISPR–Cas loci is shown. The actions performed in the study are described in the text below.

The procedure begins with the identification of a Seed that signifies the likely presence of a CRISPR–Cas locus in a given nucleotide sequence (see Figure 7; the steps in the procedure are numbered in the order in which they occur). In this study, new CRISPR-Cas systems were discovered by searching the current sequence databases (see Dataset in methods). *cas1* was used as the seed, as it is the most common Cas protein in CRISPR–Cas systems and is most highly conserved at the sequence level (Takeuchi *et al.*, 2012). To ensure the maximum sensitivity of detection, the search was carried out by comparing a Cas1 sequence profile with translated genomic and metagenomic sequences. After the *cas1* genes were detected, their respective ‘neighborhoods’ were examined for the presence of other *cas* genes by searching for Cas proteins using ~400 previously developed profiles and applying the criteria for the classification of the CRISPR–Cas loci (Makarova *et al.*, 2015). In a complementary approach, to extend the search to non-autonomous CRISPR–Cas systems, the same procedure was repeated using the CRISPR array as the seed. To ensure that the CRISPR array was detected at a high level of sensitivity, the predictions that were made using the Piler-CR (Edgar, 2007) and CRISPRFinder (Grissa, Vergnaud and Pourcel, 2007) methods were pooled and taken as the final CRISPR set (see Figure 7). This procedure yielded 47,174 CRISPR arrays, which is more than twice the number of *cas1* genes that were detected, reflecting the fact that many CRISPR–Cas loci lack the adaptation module and that numerous ‘orphan’ arrays, some of which seem to be functional, also exist (Almendros *et al.*, 2016).

All loci that were assigned to known CRISPR–Cas subtypes through the Cas protein profile search were discarded from the subsequent analysis, since the aim of the search was to discover new subtypes. Among the remaining *cas1* and CRISPR neighborhoods, those that encoded large proteins (>500 amino acids) were analyzed in detail, given that Cas9 and Cpf1 are large proteins (typically >1000 amino acids) and that their protein structures suggest that this large size is required to accommodate the complex between CRISPR RNA (crRNA) and target DNA (Nishimasu *et al.*, 2014, 2015; Fonfara *et al.*, 2016). The sequences of such large proteins were then screened for known protein domains using sensitive profile-based methods, such as HHpred (Söding *et al.*, 2006), secondary structure prediction and manual examination of multiple alignments (see Methods). Based on the premise that Class 2 effector proteins contain nuclease domains, even if they are distantly related or unrelated to known families of nucleases, the proteins that contain domains that are deemed irrelevant in the context of the CRISPR–Cas function (for example, membrane transporters or metabolic enzymes) were discarded. The retained proteins either contained readily identifiable, or completely unknown nuclease domains. The sequences of these proteins were then analyzed using the most sensitive methods for domain detection, such as HHpred (Söding *et al.*, 2006), with a curated multiple alignment of the respective protein sequences that were used as the query. The use of sensitive methods is essential because proteins that are involved in antiviral defense, and the Cas proteins in particular, typically evolve extremely fast (Takeuchi *et al.*, 2012; Makarova, Wolf and Koonin, 2013a).

It is to be noted that this procedure for the discovery of Class 2 CRISPR–Cas systems should, at least in principle, be exhaustive, because all loci that contain a gene that encodes a large protein (that is, a putative Class 2 effector) in the vicinity of *cas1* and/or CRISPR are analyzed in detail. The assumption of the structural requirements for a Class 2 effector, which underlie the protein size cut-off that is used, and the precision of *cas1* and CRISPR detection, are the only limitations of this approach.

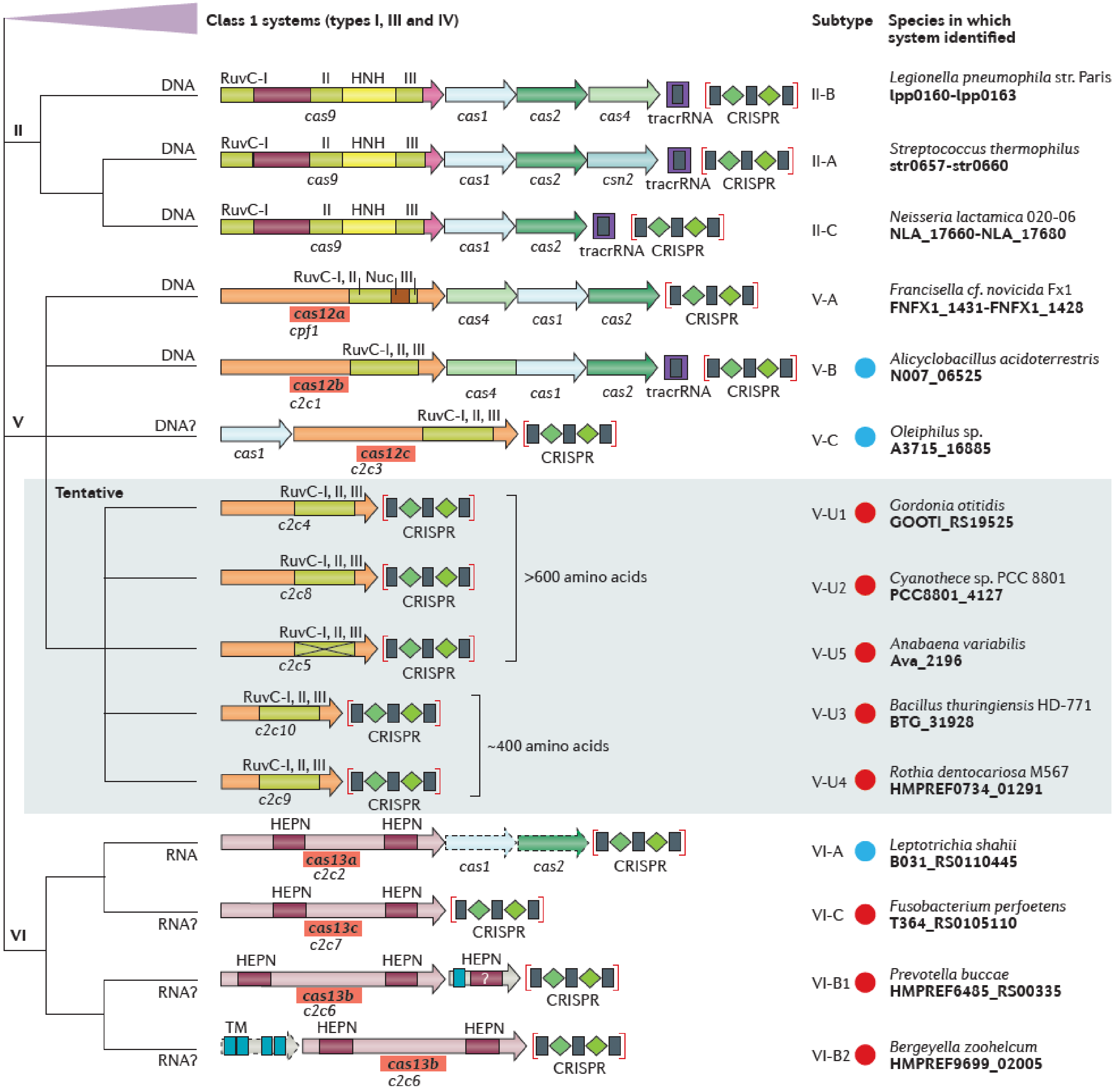

Class 1 systems (types I, III and IV)
Subtype
Species in which system identified
II
DNA
RuvC-I II HNH III
cas9
cas1
cas2
cas4
tracrRNA
CRISPR
II-B
Legionella pneumophila str. Paris
lpp0160-lpp0163
DNA
RuvC-I II HNH III
cas9
cas1
cas2
csn2
tracrRNA
CRISPR
II-A
Streptococcus thermophilus
str0657-str0660
DNA
RuvC-I II HNH III
cas9
cas1
cas2
tracrRNA
CRISPR
II-C
Neisseria lactamica 020-06
NLA_17660-NLA_17680
V
DNA
RuvC-I, II, Nuc, III
cas12a
cpf1
cas4
cas1
cas2
CRISPR
V-A
Francisella cf. novicida Fx1
FNFX1_1431-FNFX1_1428
DNA
RuvC-I, II, III
cas12b
c2c1
cas4
cas1
cas2
tracrRNA
CRISPR
V-B
Alicyclobacillus acidoterrestris
N007_06525
DNA?
cas1
RuvC-I, II, III
cas12c
c2c3
CRISPR
V-C
Oleiphilus sp.
A3715_16885
Tentative
RuvC-I, II, III
c2c4
CRISPR
V-U1
Gordonia otitidis
GOOTI_RS19525
RuvC-I, II, III
c2c8
CRISPR
>600 amino acids
V-U2
Cyanothece sp. PCC 8801
PCC8801_4127
RuvC-I, II, III
c2c5
CRISPR
V-U5
Anabaena variabilis
Ava_2196
RuvC-I, II, III
c2c10
CRISPR
~400 amino acids
V-U3
Bacillus thuringiensis HD-771
BTG_31928
RuvC-I, II, III
c2c9
CRISPR
V-U4
Rothia dentocariosa M567
HMPREF0734_01291
VI
RNA
HEPN
HEPN
cas13a
c2c2
cas1
cas2
CRISPR
VI-A
Leptotrichia shahii
B031_RS0110445
RNA?
HEPN
HEPN
cas13c
c2c7
CRISPR
VI-C
Fusobacterium perfoetens
T364_RS0105110
RNA?
HEPN
HEPN
HEPN
?
cas13b
c2c6
CRISPR
VI-B1
Prevotella buccae
HMPREF6485_RS00335
RNA?
TM
HEPN
HEPN
cas13b
c2c6
CRISPR
VI-B2
Bergeyella zoohelcum
HMPREF9699_02005

Figure 8. The updated classification scheme for Class 2 CRISPR–Cas systems. The class 1 systems are collapsed; all other systems shown are Class 2 systems. New Class 2 systems that were discovered using the computational pipeline in this study (see Figure 8) are indicated, with blue circles for those that were identified by association with *cas1* and red circles for those that were identified by association with CRISPR. For each Class 2 system subtype, as well as for the five distinct variants of the provisional V‑uncharacterized (V-U) subtype, the locus organization and the domain architecture of the effector and accessory proteins are schematically shown. RuvC-I, RuvC-II and RuvC-III are the three distinct motifs that contribute to the nuclease catalytic center; numerals in the figure correspond to the respective RuvC motif. The portions of Cas9 proteins that roughly correspond to the recognition lobe and the protospacer-adjacent motif (PAM)-interacting domain are shown by maroon and pink shapes, respectively. The proposed new systematic gene names are shown in bold type in red boxes. Provisional gene names for effector protein candidates are shown below the respective shapes as follows: C2c1–10, Class 2 candidate proteins 1–10; for subtype V‑A, the previously introduced vernacular Cpf1 is indicated. For subtype VI‑A, *cas1* and *cas2* are shown with broken outlines to indicate that only some of these loci include the adaptation module. For the V-U5 variant, the inactivation of the RuvC-like nuclease domain is indicated by a cross. The specific strains of bacteria in which these systems were identified and locus tags for the respective protein-coding genes are also indicated. The abbreviation TM indicates a predicted transmembrane helix. The predicted type of target, namely DNA or RNA, is indicated for each subtype. A question mark next to the target indicates that the activity is only predicted and has not been demonstrated experimentally. The target is not indicated for the type V‑U systems because their RNA-guided interference capacity is questionable, which is additionally emphasized by shading.

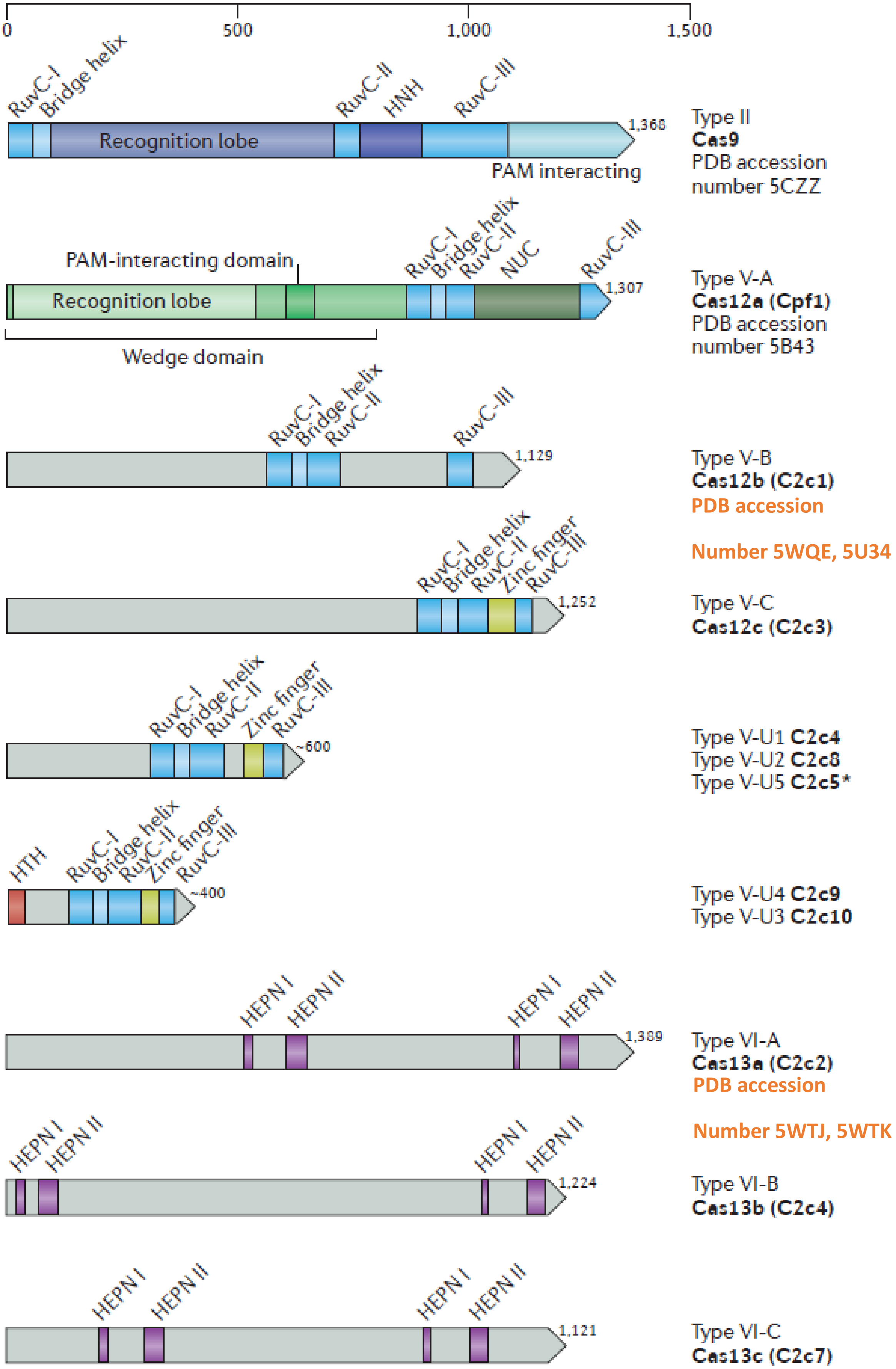

0
500
1,000
1,500
RuvC-I
Bridge helix
RuvC-II
HNH
RuvC-III
Recognition lobe
1,368
PAM interacting
Type II
Cas9
PDB accession
number 5CZZ
PAM-interacting domain
RuvC-I
Bridge helix
RuvC-II
NUC
RuvC-III
Recognition lobe
1,307
Wedge domain
Type V-A
Cas12a (Cpf1)
PDB accession
number 5B43
RuvC-I
Bridge helix
RuvC-II
RuvC-III
1,129
Type V-B
Cas12b (C2c1)
PDB accession
Number 5WQE, 5U34
RuvC-I
Bridge helix
RuvC-II
Zinc finger
RuvC-III
1,252
Type V-C
Cas12c (C2c3)
RuvC-I
Bridge helix
RuvC-II
Zinc finger
RuvC-III
~600
Type V-U1 C2c4
Type V-U2 C2c8
Type V-U5 C2c5*
HTH
RuvC-I
Bridge helix
RuvC-II
Zinc finger
RuvC-III
~400
Type V-U4 C2c9
Type V-U3 C2c10
HEPN I
HEPN II
HEPN I
HEPN II
1,389
Type VI-A
Cas13a (C2c2)
PDB accession
Number 5WTJ, 5WTK
HEPN I
HEPN II
HEPN I
HEPN II
1,224
Type VI-B
Cas13b (C2c4)
HEPN I
HEPN II
HEPN I
HEPN II
1,121
Type VI-C
Cas13c (C2c7)

Figure 9. The domain architecture of Class 2 CRISPR effector proteins. For the type II and subtype V‑A effectors, the crystal structures (indicated here by their RCSB Protein Data Bank (PDB) accession numbers (5CZZ and 5B43, respectively)) are available and the corresponding domain architectures are shown in detail for novel proteins (PDB numbers shown in orange). For the remainder of the proteins, the grey areas indicate structurally and functionally uncharacterized portions. RuvC-I, RuvC-II and RuvC-III, as well as higher eukaryotes and prokaryotes nucleotide-binding I (HEPN I) and HEPN II, denote the catalytic motifs of the respective nuclease domains of the CRISPR effectors. The bridge helix corresponds to an arginine-rich region that follows the RuvC-I motif. Other domains shown in the figure are denoted as follows: PAM interacting, protospacer-adjacent motif (PAM)-interacting domain; HNH, HNH family endonuclease domain, zinc finger domain with a CXXC..CXXC motif (dots represent the variable distance between the two pairs of cysteines); HTH, putative DNA-binding helix–turn–helix domain; NUC, nuclease domain. The proteins and domains are shown approximately to scale. For each protein, the corresponding number of amino acids is indicated, and a ruler is shown on top of the figure to guide the eye. For the functionally characterized full-length effectors, the proposed new nomenclature (Cas12 and Cas13) is indicated, whereas only the provisional names are indicated for the uncharacterized putative effectors of type V‑uncharacterized (V-U). When, and if, functional evidence of a bona fide CRISPR response is reported for these effectors, they should be referred to as Cas12 proteins with the corresponding specifying letters. The putative V-U1, V-U2 and V-U5 effectors are larger than the typical TnpB proteins, whereas the V-U3 and V-U4 effectors are in the characteristic size range of TnpB. The asterisk at C2c5 indicates that this putative effector protein contains replacements of the catalytic residues of the RuvC-like nuclease domain and lacks the zinc finger.

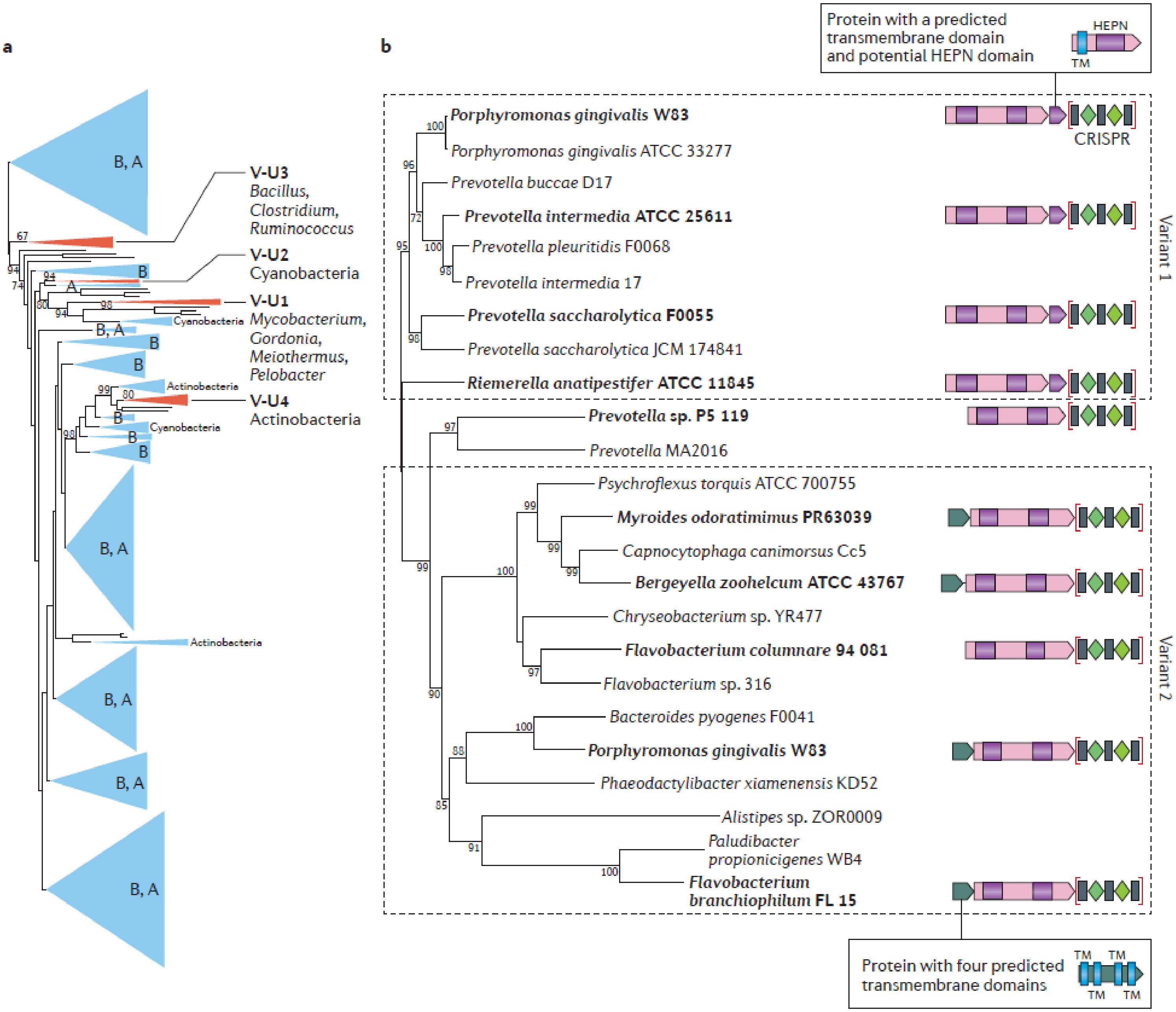
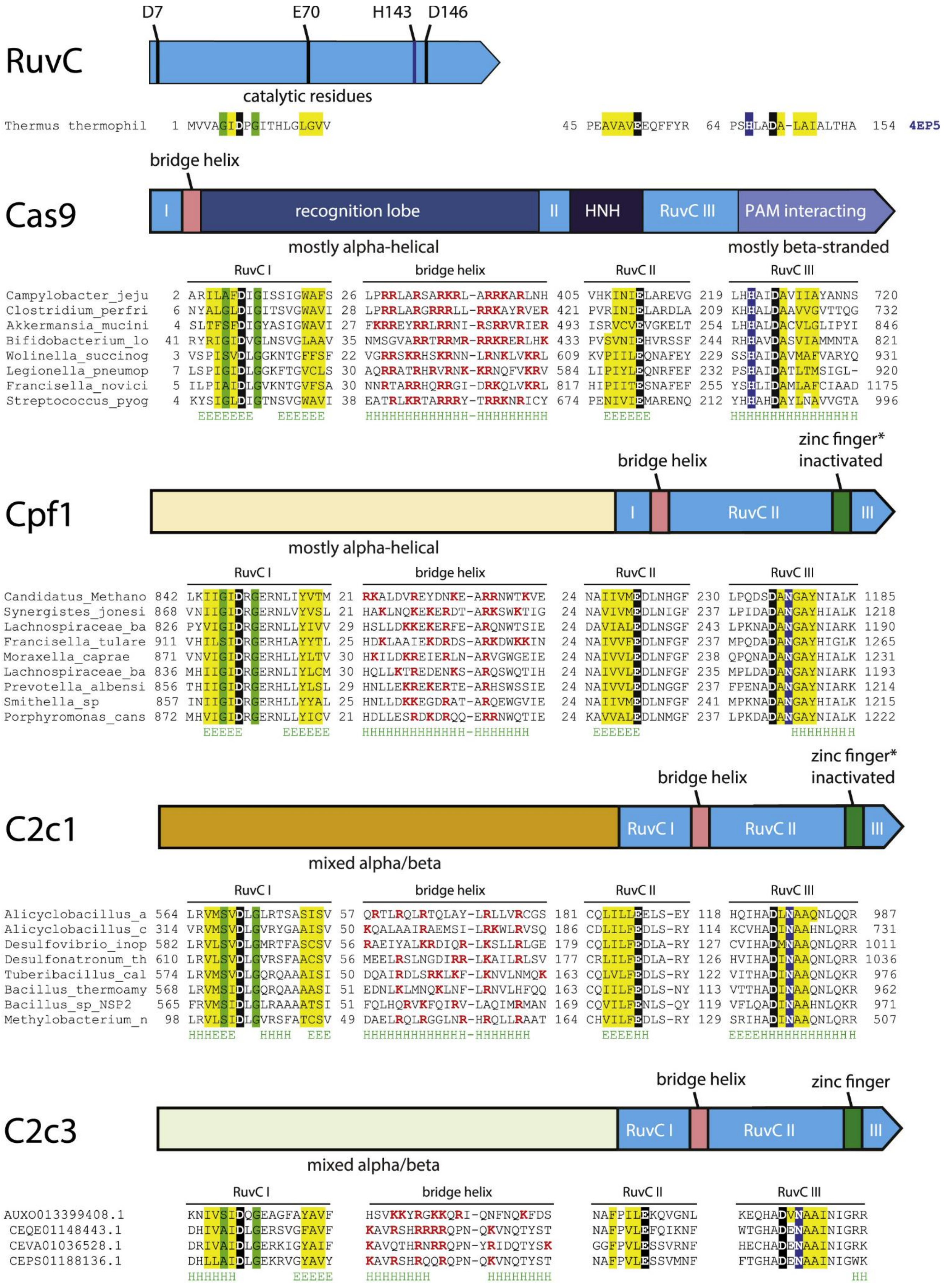

a
B, A
67
V-U3
Bacillus,
Clostridium,
Ruminococcus
94
74
94
B
V-U2
Cyanobacteria
A
80
98
V-U1
94
Cyanobacteria
Mycobacterium,
Gordonia,
Meiothermus,
Pelobacter
B, A
B
B
99
Actinobacteria
80
V-U4
Actinobacteria
B
Cyanobacteria
98
B
B
B, A
Actinobacteria
B, A
B, A
B, A
b
Protein with a predicted
transmembrane domain
and potential HEPN domain
HEPN
TM
100
Porphyromonas gingivalis W83
CRISPR
Porphyromonas gingivalis ATCC 33277
96
Prevotella buccae D17
72
Prevotella intermedia ATCC 25611
95
100
Prevotella pleuritidis F0068
98
Prevotella intermedia 17
Prevotella saccharolytica F0055
98
Prevotella saccharolytica JCM 174841
Riemerella anatipestifer ATCC 11845
Variant 1
97
Prevotella sp. P5 119
Prevotella MA2016
Psychroflexus torquis ATCC 700755
99
Myroides odoratimimus PR63039
99
Capnocytophaga canimorsus Cc5
99
99
Bergeyella zoohelcum ATCC 43767
100
Chryseobacterium sp. YR477
Flavobacterium columnare 94 081
97
Flavobacterium sp. 316
90
Variant 2
Bacteroides pyogenes F0041
100
88
Porphyromonas gingivalis W83
Phaeodactylibacter xiamenensis KD52
85
Alistipes sp. ZOR0009
91
Paludibacter
propionicigenes WB4
100
Flavobacterium
branchiophilum FL 15
Protein with four predicted
transmembrane domains
TM
TM
TM
TM

Figure 10. Phylogenies of the type V and type VI‑B effectors. a) A maximum-likelihood phylogenetic tree of TnpB nucleases, including the putative type V‑uncharacterized (V-U) effectors that have a predicted active RuvC domain (see Methods). The major subtrees of transposon-encoded TnpB proteins are collapsed and indicated by triangles; some of these large groups include tnpB genes that are adjacent to CRISPR arrays, but these do not show evolutionary stability and thus cannot be identified as effectors. The four distinct evolutionarily stable groups of CRISPR-associated TnpB assigned to subtype V‑U are shown by red triangles. Altogether, the tree includes 1,770 unique TnpB sequences, 403 of which are TnpB proteins that are encoded next to TnpA (autonomous transposons); 168 of these *tnpB* genes are adjacent to CRISPR arrays and, of these, 49 are assigned to four variants of subtype V‑U (none of these belong to autonomous transposons). In the subtrees that include the subtype V‑U variants, bootstrap values (percentages) are shown for those subtrees that include the distinct V‑U variants. For each type V‑U variant, the bacterial taxa that harbor the majority of the respective loci are indicated. Dominant bacterial or archaeal lineages, if any, are indicated in the triangles. For the complete tree and accession numbers of all sequences, see Supplementary Information Box 2 (parts c and h).

b) Phylogenetic tree of the subtype VI‑B Cas13b effector proteins. The tree was constructed as in part a, and the bootstrap values that are larger than 70% are indicated. The organization of typical Cas13b loci for selected representatives (specifically those that are shown in bold) is schematically shown on the right. Variant 1 and variant 2 correspond to the two major branches of the tree and differ with respect to the domain architectures of the second smaller protein encoded in the locus; the domain architectures of these putative accessory proteins are shown above (for variant 1) and below (for variant 2) the respective loci schematics. The CRISPR arrays are shown schematically in brackets. TM indicates a predicted transmembrane domain, shown by blue boxes. Higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domains are shown as maroon boxes. A, diverse archaea; B, diverse bacteria.

Subtypes V-B and V-C identified using a cas1 seed: large multidomain effectors

The distinctive feature of type II and type V CRISPR–Cas sequences is the presence of a RuvC-like nuclease domain in their multidomain effector proteins (Makarova *et al.*, 2015). In the type II effector Cas9, the RuvC-like domain contains an inserted HNH nuclease domain (See Figures 8, 9). Other than the RuvC-like domain, the effector proteins of the three type V subtypes do not share any detectable sequence similarity to each other or to Cas9. However, the only crystal structures of Class 2 effectors that are available (at the time of the study), specifically those of Cas9 and Cpf1, reveal a common structural framework (see Figure 9) (Dong *et al.*, 2016; Yamano *et al.*, 2016). The structures of the putative, large, type V effectors that were discovered using the *cas1* seed, namely those of the subtype V-C, are unsolved, but the subtype V-B effector C2c1 was solved (Yang *et al.*, 2016; Liu *et al.*, 2017) and shown experimentally to have robust interference activity (Shmakov *et al.*, 2015). All of the class V effectors that have been identified to date share a similar, large size (typically, 1,000–1,300 amino acid residues) and a single common domain – the RuvC-like endonuclease domain (see Figure 9) – although the sequence similarity between the effector proteins of different subtypes is extremely low. It is likely that all type V effectors adopt similar bilobed structures that hold the crRNA and target DNA together, although the effector proteins of different subtypes do not seem to be directly related.



```
Thermus thermophil    1 MVVAGIDPGITHLGLGVV                                       45 PEAVAVEEQFFYR  64 PSHLADA-LAIALTHA  154 4EP5
```

```
                              RuvC I                    bridge helix                     RuvC II                  RuvC III
Campylobacter_jeju    2 ARILAFDIGISSIGWAFS 26 LPRRLARSARKRL-ARRKARLNH 405 VHKINIELAREVG 219 LHHAIDAVIIAYANNS   720
Clostridium_perfri    6 NYALGLDIGITSVGWAVI 28 LPRRLARGRRRLL-RRKAYRVER 421 PVRINIELARDLA 209 KHHALDAAVVGVTTQG   732
Akkermansia_mucini    4 SLTFSFDIGYASIGWAVI 27 FKRREYRRLRRNI-RSRRVRIER 493 ISRVCVEVGKELT 254 LHHALDACVLGLIPYI   846
Bifidobacterium_lo   41 RYRIGIDVGLNSVGLAAV 35 NMSGVARRTRRMR-RRKRERLHK 433 PVSVNIEHVRSSF 244 RHHAVDASVIAMMNTA   821
Wolinella_succinog    3 VSPISVDLGGKNTGFFSF 22 VGRRSKRHSKRNN-LRNKLVKRL 609 KVPIILEQNAFEY 229 SSHAIDAVMAFVARYQ   931
Legionella_pneumop    7 LSPIGIDLGGKFTGVCLS 30 AQRRATRHRVRNK-KRNQFVKRV 584 LIPIYLEQNRFEF 232 PSHAIDATLTMSIGL-   920
Francisella_novici    5 ILPIAIDLGVKNTGVFSA 30 NNRTARRHQRRGI-DRKQLVKRL 817 HIPIITESNAFEF 255 YSHLIDAMLAFCIAAD  1175
Streptococcus_pyog    4 KYSIGLDIGTNSVGWAVI 38 EATRLKRTARRRY-TRRKNRICY 674 PENIVIEMARENQ 212 YHHAHDAYLNAVVGTA   996
                          EEEEEEE   EEEEEE     HHHHHHHHHHHHH-HHHHHHHHH       EEEEEE           HHHHHHHHHHHHHHH
```

```
                               RuvC I                    bridge helix                     RuvC II                  RuvC III
Candidatus_Methano 842 LKIIGIDRGERNLIYVTM 21 RKALDVREYDNKE-ARRNWTKVE  24 NAIIVMEDLNHGF 230 LPQDSDANGAYNIALK 1185
Synergistes_jonesi 868 VNIIGIDRGERNLVYVSL 21 HAKLNQKEKERDT-ARKSWKTIG  24 NAVIVMEDLNIGF 237 LPIDADANGAYHIALK 1218
Lachnospiraceae_ba 826 PYVIGIDRGERNLLYIVV 29 HSLLDKKEKERFE-ARQNWTSIE  24 DAVIALEDLNSGF 243 LPKNADANGAYNIARK 1190
Francisella_tulare 911 VHILSIDRGERHLAYYTL 25 HDKLAAIEKDRDS-ARKDWKKIN  24 NAIVVFEDLNFGF 237 MPQDADANGAYHIGLK 1265
Moraxella_caprae   871 VNVIGIDRGERHLLYLTV 30 HKILDKREIERLN-ARVGWGEIE  24 NAIVVLEDLNFGF 238 QPQNADANGAYHIALK 1231
Lachnospiraceae_ba 836 MHIIGIDRGERNLIYLCM 30 HQLLKTREDENKS-ARQSWQTIH  24 NAIVVLEDLNFGF 235 MPLDADANGAYNIARK 1193
Prevotella_albensi 856 THIIGIDRGERHLLYLSL 29 HNLLEKREKERTE-ARHSWSSIE  24 NAIVVLEDLNGGF 237 FPENADANGAYNIARK 1214
Smithella_sp       857 INIIGIDRGERHLLYYAL 25 HNLLDKKEGDRAT-ARQEWGVIE  24 NAIIVMEDLNFGF 241 MPKNADANGAYHIALK 1215
Porphyromonas_cans 872 MHVIGIDRGERNLLYICV 21 HDLLESRDKDRQQ-ERRNWQTIE  24 KAVVALEDLNMGF 237 LPKDADANGAYNIALK 1222
                          EEEEE      EEEEEE     HHHHHHHHHHHHH-HHHHHHH        EEEEEE                   HHHHHHHH
```

```
                               RuvC I                    bridge helix                     RuvC II                  RuvC III
Alicyclobacillus_a 564 LRVMSVDLGLRTSASISV 57 QRTLRQLRTQLAY-LRLLVRCGS 181 CQLILLEELS-EY 118 HQIHADLNAAQNLQQR   987
Alicyclobacillus_c 314 VRVMSVDLGVRYGAAISV 50 KQALAAIRAEMSI-LRKWLRVSQ 186 CDLILFEDLS-RY 114 KCVHADINAAHNLQRR   731
Desulfovibrio_inop 582 LRVLSVDLGMRTFASCSV 56 RAEIYALKRDIQR-LKSLLRLGE 179 CQLILFEDLA-RY 127 CVIHADMNAAQNLQRR  1011
Desulfonatronum_th 610 LRVLSVDLGVRSFAACSV 56 MEELRSLNGDIRR-LKAILRLSV 177 CRLILFEDLA-RY 126 HVIHADINAAQNLQRR  1036
Tuberibacillus_cal 574 LRVMSVDLGQRQAAAISI 50 DQAIRDLSRKLKF-LKNVLNMQK 163 CQLVLFEDLS-RY 122 VITHADINAAQNLQKR   976
Bacillus_thermoamy 568 LRVMSIDLGQRQAAAASI 51 EDNLKLMNQKLNF-LRNVLHFQQ 163 CQIILFEDLS-NY 113 VTTHADINAAQNLQKR   962
Bacillus_sp_NSP2   565 FRVMSIDLGLRAAAATSI 51 FQLHQRVKFQIRV-LAQIMRMAN 169 CQVILFENLS-QY 119 VFLQADINAAHNLQKR   971
Methylobacterium_n  98 LRVLSIDLGVRSFATCSV 49 DAELRQLRGGLNR-HRQLLRAAT 164 CHVILFEDLS-RY 129 SRIHADINAAQNLQRR   507
                       HHHEEE    HHHH  EEE     HHHHHHHHHHHHH-HHHHHHH          EEEEHH           EEEEHHHHHHHHHHHH
```

```
                                RuvC I                    bridge helix                 RuvC II               RuvC III
AUXO013399408.1        KNIVSIDQGEAGFAYAVF     HSVKKYRGKKQRI-QNFNQKFDS     NAFPILEKQVGNL     KEQHADVNAAINIGRR
 CEQE01148443.1        DHIVAIDLGERSVGFAVF     KAVRSHRRRRQPN-QKVNQTYST     NAFPVLEFQIKNF     WTGHADENAAINIGRR
 CEVA01036528.1        DRIVAIDLGERKIGYAIF     KAVQTHRNRRQPN-YRIDQTYSK     GGFPVLESSVRNF     HECHADENAAINIGRK
 CEPS01188136.1        DHLLAIDLGEKRVGYAVY     KAVRSHRQQRQPN-QKVNQTYST     NAFPVLESSVMNF     FTGHADENAAINIGWK
                       HHHHHH        EEEEE     HHHHHHHH         HHHHHHHH                                     HH
```

Figure 11. Domain Architectures and Conserved Motifs of the Class 2 Effector Proteins. Types II and V: TnpB-derived nucleases. The top panel shows the RuvC nuclease from Thermus thermophilus (PDB: 4EP5) with the catalytic amino acid residues denoted. An alignment of the conserved motifs in selected representatives of the respective protein family (a single sequence for RuvC) is shown underneath each domain architecture. 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.

The TnpB homology regions of C2c1 and C2c3 contain the three catalytic motifs of the RuvC-like nuclease (Aravind, Makarova and Koonin, 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) (see Figures 9, 11; Supplementary information S1 and S4). 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 (Supplementary information S1 and S4). 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 (see Figure 11). Accordingly, it was proposed 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.

The C2c1 system from *Alicyclobacillus acidoterrestris* ATCC 49025 (Aac) was experimentally characterized by Feng Zhang lab (Shmakov *et al.*, 2015). 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. It was also identified that an abundant 79-nt small RNA is encoded between the *cas2* gene and the CRISPR array and transcribed in the same orientation as the CRISPR array. 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.

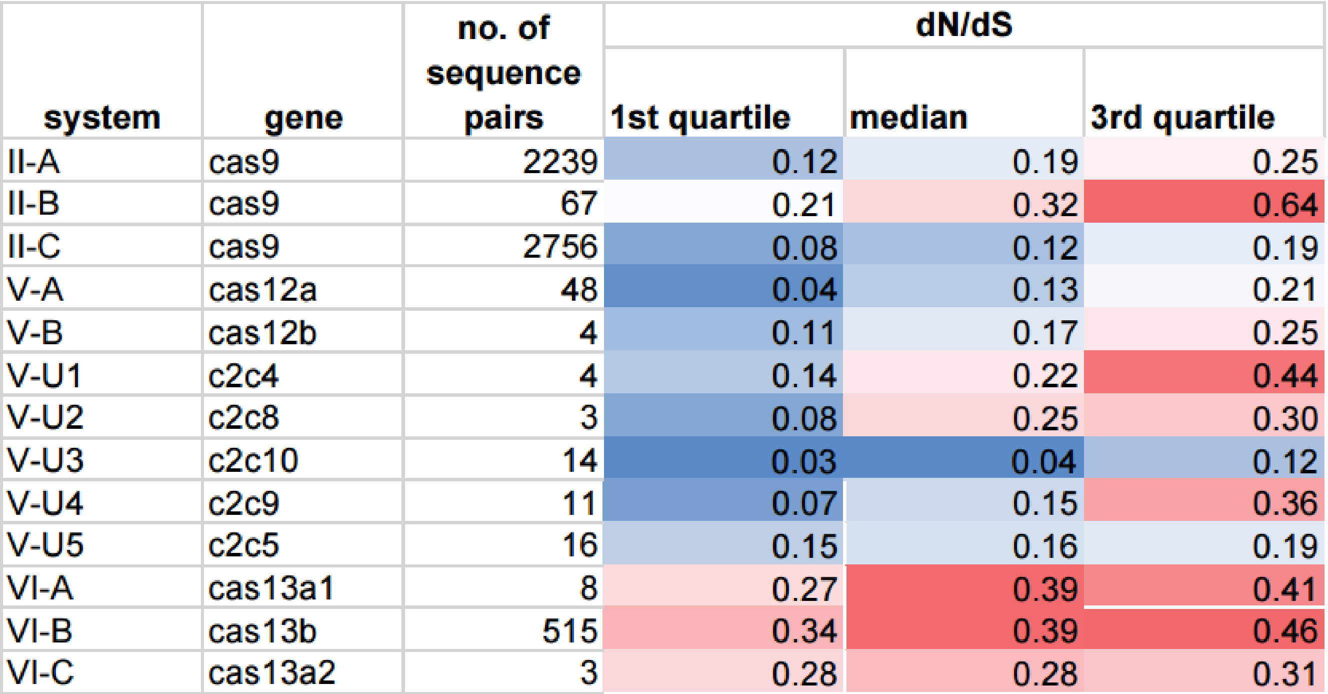
The search for homologues of the type II and type V effectors showed that the RuvC-like nuclease domains are related to TnpB proteins, an extremely abundant but poorly characterized family of nucleases that are encoded by many autonomous (i.e., which encode an active transposase, denoted TnpA, and mediate their own transposition) and even more numerous non-autonomous (i.e. which consist solely of the *tnpB* gene and rely on transposases from other elements for transposition) bacterial and archaeal transposons (Bao and Jurka, 2013; Pasternak *et al.*, 2013; Kapitonov, Makarova and Koonin, 2015) (see Figure 10a). In addition to the RuvC-like nuclease domain, TnpB proteins contain a predicted, positively charged, long α-helix that seems to be the counterpart to the bridge helix, which is a common feature of Cas9 and Cpf1 (see Figures 9, 11). Thus, similar to the Class 2 effectors, the TnpB proteins are predicted to bind to RNA. Moreover, it has been reported that a TnpB protein from the haloarchaeon *Halobacterium salinarum* binds to short overlapping sense transcripts of its own gene (Gomes-Filho *et al.*, 2015). Biochemical and biological characterization of TnpB should shed light on the evolution of the functions of Class 2 CRISPR–Cas effectors.

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

Subtype V-U identified using a CRISPR seed: small putative effectors

The search for CRISPR–Cas loci that lack the adaptation module (that is, loci that were identified with a CRISPR seed but not with a *cas1* seed; see Figure 7) yielded several additional variants of putative type V systems (Figure 8, 9, 10a) that might help to explain how CRISPR–Cas effectors evolved from TnpB. The putative effector proteins of these loci, which we have provisionally assigned to subtype V-U (where the 'U' stands for 'uncharacterized'; see below), share two features that distinguish them from type II and type V effectors that are found at CRISPR–Cas loci containing Cas1 (see Figure 8). First, these proteins are much smaller than Class 2 effectors that contain Cas1, comprising between ~500 amino acids (only slightly larger than the typical size of TnpB) and ~700 amino acids (between the size of TnpB and the typical size of the bona fide Class 2 effectors). Second, these putative effectors show a higher level of similarity to TnpB proteins than the larger type I and type V effectors (see Supplementary information S3). In particular, three groups of TnpB homologues, which are included here in subtype V-U (denoted V-U1, V-U2 and VU-5), showed evolutionary stability in terms of sequence conservation, consistent association with CRISPR arrays and presence in distinct groups of bacteria (see Figures 8, 9; see below). A more detailed examination showed that, within each of these groups, the respective loci in closely related bacterial genomes were genuinely orthologous, as indicated by the gene synteny conservation.

In view of the identification of these smaller CRISPR-associated TnpB homologues, we ran the pipeline (see Figure 7) removing the requirement for the minimal length of the protein adjacent to the CRISPR array, and examined the results for the presence of additional TnpB homologues. Numerous CRISPR-associated TnpB homologues were detected in the size range that is typical of the transposon-encoded TnpB, that is, ~400 amino acids (Supplementary information S2 (box), part a). Most of these loci were not evolutionarily conserved and were thus of questionable functional relevance. However, two distinct groups of such smaller CRISPR-associated TnpB (V-U3 and V-U4) were additionally detected, having characteristics similar to the three subtype V-U groups with intermediately sized CRISPR-associated TnpB (see Figures 8, 9; Table 1). Notably, the genes for the putative effectors of subtype V-U showed signs of purifying selection on protein sequences (as indicated by the low values of the non-synonymous to synonymous nucleotide substitutions, dN/dS; see Table 1), which was found to be particularly strong for the subtype V-U3 group (Supplementary information S2 (box), part b, Table 1). Taken together, these observations imply that the respective TnpB homologues have CRISPR-dependent functions and, in our view, justify the designation of the respective loci as subtype V-U.



system	gene	no. of sequence pairs	dN/dS		
			1st quartile	median	3rd quartile
II-A	cas9	2239	0.12	0.19	0.25
II-B	cas9	67	0.21	0.32	0.64
II-C	cas9	2756	0.08	0.12	0.19
V-A	cas12a	48	0.04	0.13	0.21
V-B	cas12b	4	0.11	0.17	0.25
V-U1	c2c4	4	0.14	0.22	0.44
V-U2	c2c8	3	0.08	0.25	0.30
V-U3	c2c10	14	0.03	0.04	0.12
V-U4	c2c9	11	0.07	0.15	0.36
V-U5	c2c5	16	0.15	0.16	0.19
VI-A	cas13a1	8	0.27	0.39	0.41
VI-B	cas13b	515	0.34	0.39	0.46
VI-C	cas13a2	3	0.28	0.28	0.31

Table 1. Strength of purifying selection for Class 2 effector protein families. The three quartiles of the distribution of the dN/dS ratio, estimated for sequence pairs with 0.0002 < dN < 1.0 and 0.0002 < dS < 1.0 (see Methods for details) are given. The background color highlights the scale from low (blue) to high (red) values.

For the larger bona fide type V effectors, low sequence conservation precluded reliable phylogenetic analysis, whereas a robust tree could be constructed for the smaller CRISPR-associated homologues, together with the typical transposon-encoded TnpB (see Methods and Supplementary information S2 (box), part c). The topology of this tree indicated that four of the five distinct variants of subtype V-U (hereafter referred to as subtypes V-U1, V-U2, V-U3, V-U4 and V-U5) originated from different TnpB families (see Figure 10a), which is in agreement with the hypothesis of the independent evolution of different Class 2 subtype effectors from transposon-encoded nucleases. The fifth variant (subtype V-U5), which is found in various cyanobacteria, consists of diverged TnpB homologues that have several mutations in the catalytic motifs of their RuvC-like domain and was accordingly not included in the phylogeny here. Of the five stable variants, subtype V-U1 is found in diverse bacteria, whereas the remaining subtypes are largely limited in their spread to particular bacterial taxa (see Figure 10a; Supplementary Information S2 (box), part d). We further extended this evolutionary analysis to all putative type V effectors by building a cluster dendrogram based on the distances that were derived from profile-to-profile comparisons of the respective protein sequences (see Methods). The results suggest that the effectors of each of the identified subtypes, as well as the five distinct variants in subtype V-U, originated independently from different TnpB families (see Figure 12).

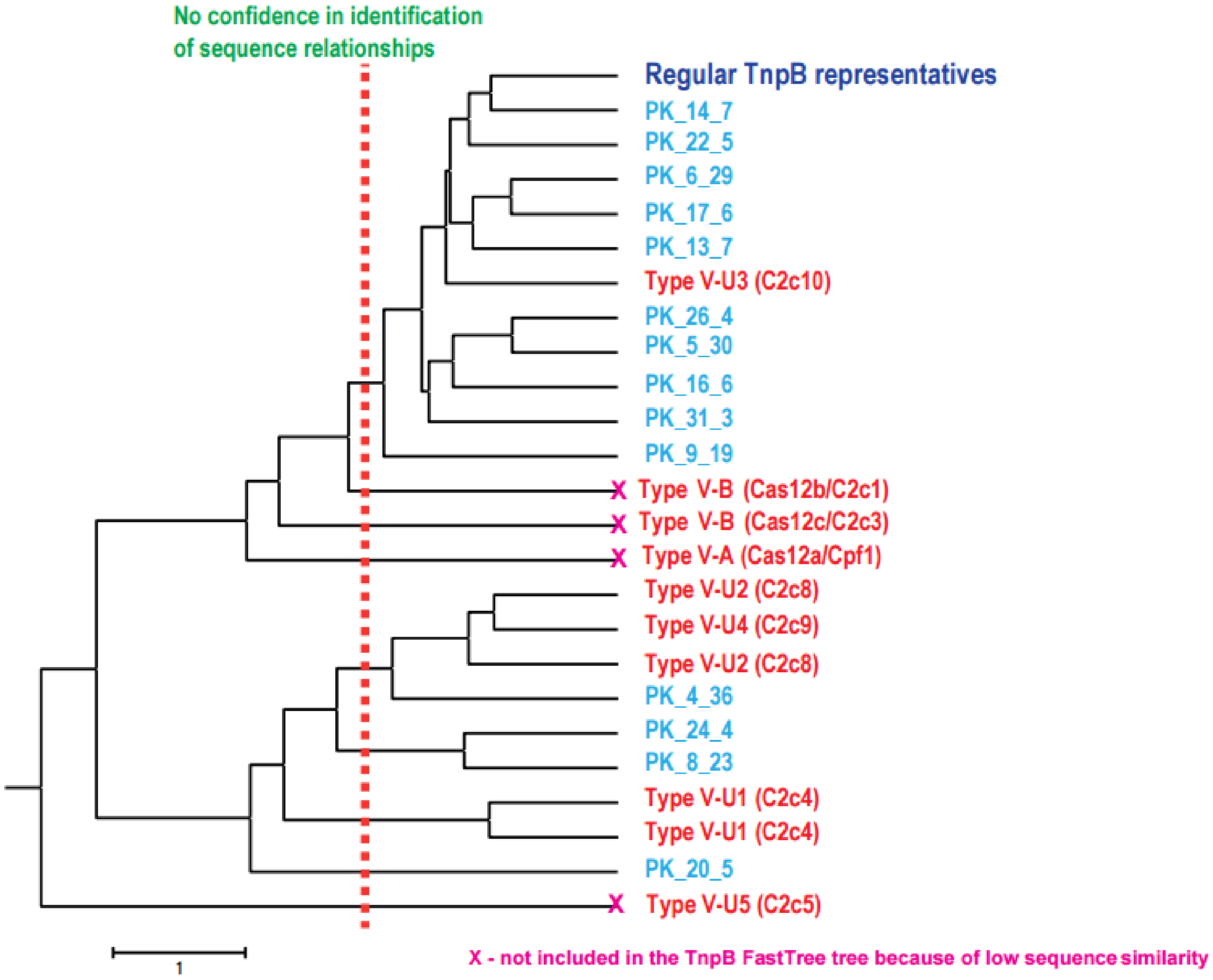


Figure 12. UPGMA dendrogram of protein family profile similarity. Profiles were built for distinct subfamilies of type V systems (red) and TnpB family (blue). The profiles correspond to the clusters, information for which is provided in the supplementary information S2 (box, part h). The profile dendrogram was built on the basis of a similarity score matrix obtained using the HHalign program (see details in the Supplementary Methods). The dotted line indicates the arbitrary similarity cutoff ~2 (in distance units shown by the scale bar below the tree) which, empirically, corresponds to the limit of confident identification of relationships between groups of sequences (i.e. the groups to the right of the line are considered to be confidently identified.)

The subtype V-U TnpB-like proteins are too small to adopt a bilobed structure of sufficient size to accommodate the crRNA–target DNA complex, as the typical Class 2 effectors do, and, therefore, are unlikely to function in that capacity without additional partners. Furthermore, the subtype V-U loci lack any additional *cas* genes (see Figure 8), which, together with the above structural considerations, suggests that any prediction of their fully fledged CRISPR activity should be cautious. Nevertheless, the evolutionarily stable association of at least five distinct subtype V-U variants with CRISPR arrays implies that at least some of these proteins do carry out CRISPR-dependent biological functions. Such functions might involve a typical CRISPR response that is aided by Cas proteins from other loci and/or by additional non-Cas proteins. Remarkably, the CRISPR arrays that are associated with group V-U3, which is mostly found in *bacilli* and *clostridia*, contain several spacers that match the genomic sequences of bacteriophages that infect these bacteria (Supplementary information S2 (box), part e). Furthermore, the sets of spacers in each subtype V-U group were completely different, even between closely related bacterial genomes (Supplementary information S2 (box), part e), which implies active spacer turnover. The diversity of the spacers and the presence of the phage-specific spacers in group V-U3 imply that at least some subtype V-U variants are functional CRISPR–Cas systems that are engaged in anti-phage adaptive immunity. Many of the complete genomes that contain group V-U3 and group V-U4 loci lack any additional CRISPR–Cas systems (Supplementary information S2 (box), part f), which makes it puzzling as to how these systems acquire their spacers. Alternatively, some of the subtype V-U systems might have distinct regulatory roles that do not require the formation of a ternary complex with the crRNA and the DNA target; indeed, several non-defense functions of CRISPR–Cas have been described (Westra, Buckling and Fineran, 2014). This possibility is particularly plausible for the V-U5 variant, which seems to encompass a catalytically inactive TnpB homologue (see Figure 9, denoted C2c5*; Supplementary information S3 (box)). Furthermore, in genomes that contain the group V-U2 and group V-U5 loci, along with other CRISPR–Cas systems, the CRISPR sequences that are associated with the former loci are unique (Supplementary information S2 (box), part f), which suggests that these subtype V-U systems have distinct functions.

Subtypes VI-A, VI-B and VI-C identified using a cas1 and CRISPR seeds: RNA-targeting CRISPR–Cas multidomain effectors

The signature of type VI systems is the presence of an effector protein that contains two HEPN domains (see Figures 8, 9). HEPN domains (Higher Eukaryotes and Prokaryotes Nucleotide-binding) are common in various defense systems, among which those that have been experimentally characterized, such as the toxins of numerous prokaryotic toxin–antitoxin systems or eukaryotic RNase L, all have RNase activity (Grynberg, Erlandsen and Godzik, 2003; Anantharaman *et al.*, 2013; Makarova *et al.*, 2014). Therefore, the first putative type VI effector, denoted C2c2, found by association with *cas*1, was predicted to function as an RNA-guided RNase.

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 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 (see Figure 13).

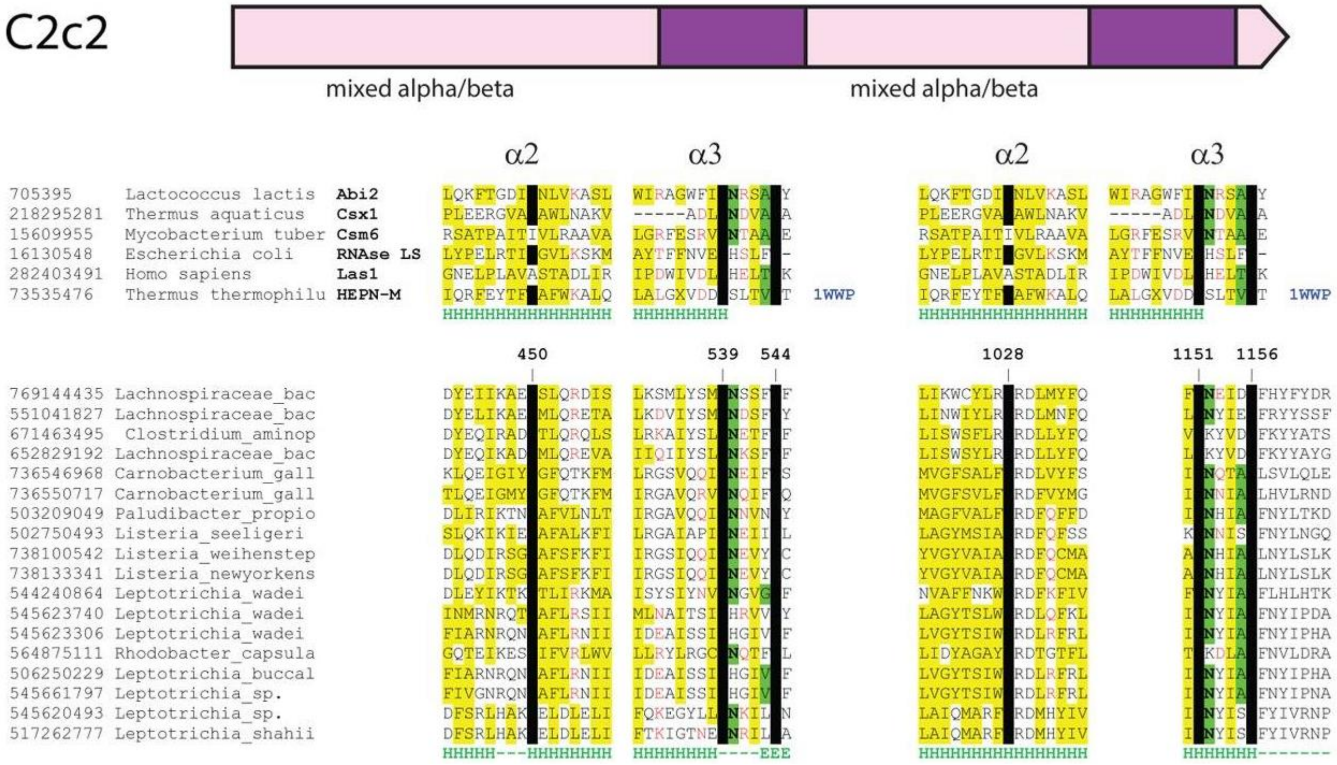

C2c2
mixed alpha/beta
mixed alpha/beta
α2
α3
α2
α3
705395 Lactococcus lactis Abi2 LQKFTGDI NLVKASL WIRAGWFI NRSA Y LQKFTGDI NLVKASL WIRAGWFI NRSA Y
218295281 Thermus aquaticus Csx1 PLEERGVA AWLNAKV -----ADL NDVA A PLEERGVA AWLNAKV -----ADL NDVA A
15609955 Mycobacterium tuber Csm6 RSATPAITIVLRAAVA LGRFESRV NTAA E RSATPAITIVLRAAVA LGRFESRV NTAA E
16130548 Escherichia coli RNAse LS LYPELRTI GVLKSKM AYTFFNVE HSLF - LYPELRTI GVLKSKM AYTFFNVE HSLF -
282403491 Homo sapiens Las1 GNELPLAVASTADLIR IPDWIVDL HELT K GNELPLAVASTADLIR IPDWIVDL HELT K
73535476 Thermus thermophilu HEPN-M IQRFEYTF AFWKALQ LALGXVDD SLTV T 1WWP IQRFEYTF AFWKALQ LALGXVDD SLTV T 1WWP
HHHHHHHHHHHHHHHH HHHHHHHHH HHHHHHHHHHHHHHHH HHHHHHHHH
450 539 544 1028 1151 1156
769144435 Lachnospiraceae_bac DYEIIKAE SLQRDIS LKSMLYSM NSSF F LIKWCYLR RDLMYFQ F NEID FHYFYDR
551041827 Lachnospiraceae_bac DYELIKAE MLQRETA LKDVIYSM NDSF Y LINWIYLR RDLMNFQ L NYIE FRYYSSF
671463495 Clostridium_aminop DYEQIRAD TLQRQLS LRKAIYSL NETF F LISWSFLR RDLLYFQ V KYVD FKYYATS
652829192 Lachnospiraceae_bac DYEQIKAD MLQREVA IIQIIYSL NKSF F LISWSYLR RDLLYFQ L KYVD FKYYAYG
736546968 Carnobacterium_gall KLQEIGIY GFQTKFM LRGSVQQI NEIF S MVGFSALF RDLVYFS I NQTA LSVLQLE
736550717 Carnobacterium_gall TLQEIGMY GFQTKFM IRGAVQRV NQIF Q MVGFSVLF RDFVYMG I NNIA LHVLRND
503209049 Paludibacter_propio DLIRIKTN AFVLNLT IRGAVQQI NNVN Y MAGFVALF RDFQFFD I NHIA FNYLTKD
502750493 Listeria_seeligeri SLQKIKIE AFALKFI LRGAIAPI NEII L LAGYMSIA RDFQFSS K NNIS FNYLNGQ
738100542 Listeria_weihenstep DLQDIRSG AFSFKFI IRGSIQQI NEVY C YVGYVAIA RDFQCMA A NHIA LNYLSLK
738133341 Listeria_newyorkens DLQDIRSG AFSFKFI IRGSIQQI NEVY C YVGYVAIA RDFQCMA A NHIA LNYLSLK
544240864 Leptotrichia_wadei DLEYIKTK TLIRKMA ISYSIYNV NGVG F NVAFFNKW RDFKFIV F NYIA FLHLHTK
545623740 Leptotrichia_wadei INMRNRQT AFLRSII MLNAITSI HRVV Y LAGYTSLW RDLQFKL I NYIA FNYIPDA
545623306 Leptotrichia_wadei FIARNRQN AFLRNII IDEAISSI HGIV F LVGYTSIW RDLRFRL I NYIA FNYIPHA
564875111 Rhodobacter_capsula GQTEIKES IFVRLWV LLRYLRGC NQTF L LIDYAGAY RDTGTFL T KDLA FNVLDRA
506250229 Leptotrichia_buccal FIARNRQN AFLRNII IDEAISSI HGIV F LVGYTSIW RDLRFRL I NYIA FNYIPHA
545661797 Leptotrichia_sp. FIVGNRQN AFLRNII IDEAISSI HGIV F LVGYTSIW RDLRFRL I NYIA FNYIPNA
545620493 Leptotrichia_sp. DFSRLHAK ELDLELI FQKEGYLL NKIL N LAIQMARF RDMHYIV I NYIS FYIVRNP
517262777 Leptotrichia_shahii DFSRLHAK ELDLELI FTKIGTNE NRIL A LAIQMARF RDMHYIV I NYIS FYIVRNP
HHHHH---HHHHHHHH HHHHHHHH----EEE HHHHHHHHHHHHHHHH HHHHHHH-------

Figure 13. 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 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.

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(Marchler-Bauer *et al.*, 2002) database using the HHpred program (Söding *et al.*, 2006) 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 (see Figure 13). Outside of the two HEPN domains, the C2c2 sequence is predicted (see Methods) to adopt a mixed α/β structure without discernible similarity to any known protein folds (Supplementary information S5). Given the unique predicted effector of C2c2, these systems were qualified as a type VI CRISPR-Cas.

Subsequently, RNA targeting prediction was experimentally validated, and the type VI effectors were shown to protect against the RNA bacteriophage MS2 (Abudayyeh *et al.*, 2016). In addition, a novel feature of C2c2 is that, once primed with the cognate target RNA, the effector becomes a promiscuous RNase that has a toxic, growth-inhibitory effect on bacteria. These findings demonstrate a coupling between adaptive immunity and programmed cell death (or dormancy induction) that was previously predicted through comparative genomic analysis (Makarova *et al.*, 2012) and mathematical modelling (Iranzo *et al.*, 2015). More recently, the C2c2 protein was shown to mediate not only interference but also the processing of pre-crRNA (East-Seletsky *et al.*, 2016).

The search for CRISPR–Cas loci using the CRISPR seed identified two additional large putative effectors that contained two HEPN domains and which we assigned to subtype VI-B and subtype VI-C, respectively (accordingly, the C2c2‑encoding loci became subtype VI-A). This classification of the type VI systems into separate subtypes is justified by the extremely low sequence similarity between the three groups of effectors, which is practically limited to the catalytic motif of the HEPN domain, the different positions of the HEPN domains with the large protein sequences, and the additional features of the locus architecture in the case of subtype VI-B (see Figures 8, 9 Supplementary information S2 (box), part d). Specifically, the two distinct variants of subtype VI-B (variants VI‑B1 and VI‑B2) both encode additional proteins that contain predicted transmembrane domains; VI‑B1 encodes four of these and VI‑B2 encodes one (see Figure 10b; Supplementary information S2 (box), part d). Phylogenetic analysis of the effector proteins suggests that the VI‑B1 and VI‑B2 variants diverged during evolution in accordance with the distinct architectures of the associated predicted membrane proteins (see Figure 10b; Supplementary information S2 (box), part d). VI‑B1 systems that contain several transmembrane domains might localize to membranes and could thus include membrane-associated RNA-targeting systems, which would be a novel feature in the biology of CRISPR–Cas. Furthermore, the single-transmembrane protein of variant VI‑B2 encompasses an additional HEPN domain, which is the third one in the type VI system (see Figure 10b; Supplementary information S2 (box), part d, and Supplementary information S6 (figure)).

Type VI-B was experimentally characterized (Smargon *et al.*, 2017) and shown to be functional and possess RNase activity. It was shown that VI-B1 and VI-B2 are able to regulate RNA interference (VI-B1 represses it and VI-B2 enhances it).

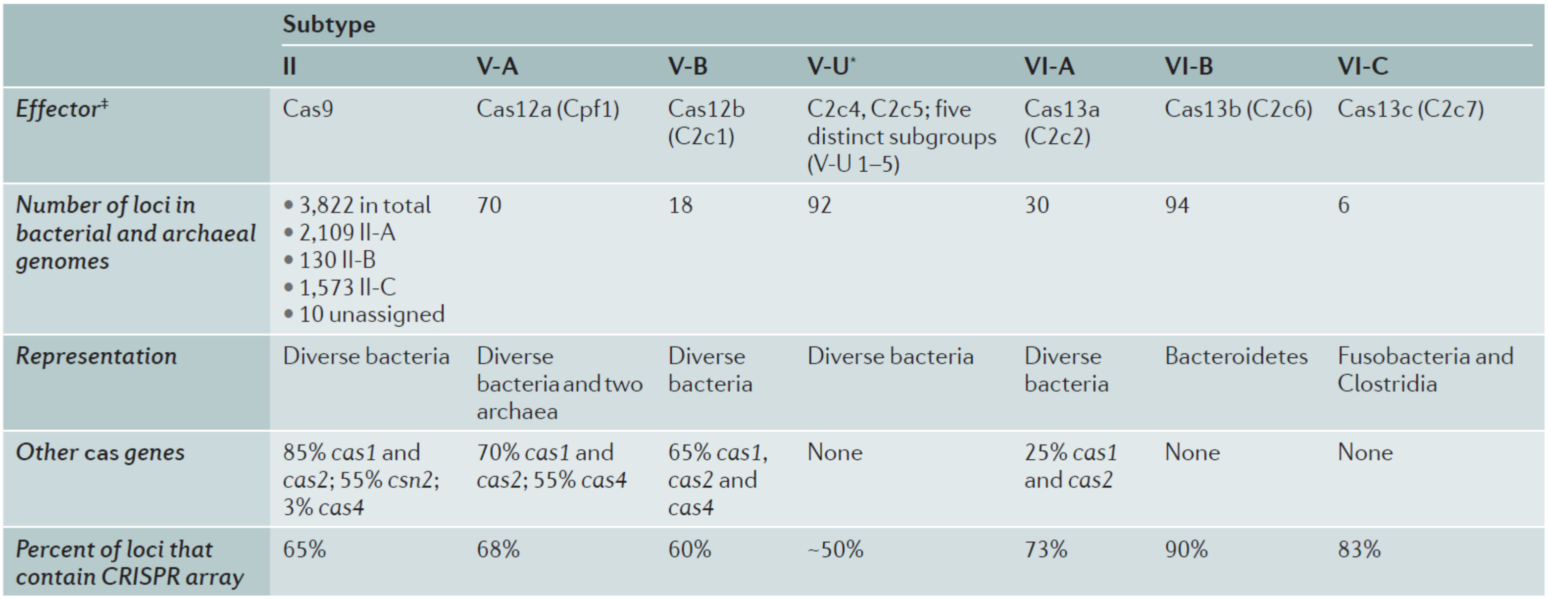
Given that all of the putative type VI effectors that have been discovered so far are similar in size to the active Class 2 effectors of subtype VI-A, even the loci that lack Cas1 are likely to be functional CRISPR–Cas systems that rely on adaptation modules from other loci in the same genome. Moreover, given that RNA viruses only represent a minor part of the prokaryotic virome (Koonin, Dolja and Krupovic, 2015), type VI systems might primarily elicit toxin activity in response to the active transcription of foreign DNA. This mechanism might not be limited to type VI systems, given the presence of HEPN domains in poorly characterized Cas proteins in many other CRISPR–Cas systems. Indeed, the RNase activity of the HEPN containing Csm6 and Csx1 proteins in type III systems has been demonstrated (Niewoehner and Jinek, 2016; Sheppard *et al.*, 2016), although their functions in the CRISPR response remain to be studied.

Part 2: Class 2 Census and Amended Classification

Comprehensive census of Class 2 CRISPR–Cas loci in bacteria and archaea

Comprehensive census of Class 2 types and subtypes in the current set of complete bacterial and archaeal genomes was provided in our study. To this end, we constructed sequence profiles for the effectors of all identified Class 2 subtypes (two separate profiles were used for the variants V-U1, V-U2 and V-U5; the V-U3 and V-U4 variants were not included in the census because, in database searches, they cannot be readily distinguished from transposon-encoded TnpB) and compared these with the proteins that are encoded in the 4,961 completely sequenced prokaryotic genomes and 43,599 partial prokaryotic genomes that are available from the National Center for Biotechnology Information (NCBI) database ('Database resources of the National Center for Biotechnology Information', 2016) (see Methods). This procedure should detect almost all instances of each effector, including highly diverged variants.

The neighborhoods of the respective genes were then examined for the presence of CRISPR arrays and additional *cas* genes, as described previously (Makarova *et al.*, 2015). The most remarkable observation is the substantial dominance of type II among the Class 2 systems. It is represented in about 8% of bacterial genomes (see Table 2). Both type V and type VI are less abundant by more than an order of magnitude, which is in agreement with the expectation that the CRISPR–Cas types and subtypes that remain to be discovered are rare variants (Makarova *et al.*, 2015).



	Subtype						
	II	V-A	V-B	V-U*	VI-A	VI-B	VI-C
Effector‡	Cas9	Cas12a (Cpf1)	Cas12b (C2c1)	C2c4, C2c5; five distinct subgroups (V-U 1–5)	Cas13a (C2c2)	Cas13b (C2c6)	Cas13c (C2c7)
Number of loci in bacterial and archaeal genomes	• 3,822 in total • 2,109 II-A • 130 II-B • 1,573 II-C • 10 unassigned	70	18	92	30	94	6
Representation	Diverse bacteria	Diverse bacteria and two archaea	Diverse bacteria	Diverse bacteria	Diverse bacteria	Bacteroidetes	Fusobacteria and Clostridia
Other cas *genes*	85% *cas1* and *cas2*; 55% *csn2*; 3% *cas4*	70% *cas1* and *cas2*; 55% *cas4*	65% *cas1*, *cas2* and *cas4*	None	25% *cas1* and *cas2*	None	None
Percent of loci that contain CRISPR array	65%	68%	60%	~50%	73%	90%	83%

Table 2. A comprehensive census of Class 2 CRISPR–Cas systems in bacterial and archaeal genomes.

*The subtype V‑uncharacterized (V-U) loci were originally identified on the basis of the adjacency of *tnpB* genes to CRISPR arrays and the evolutionary conservation of this association. Then, this putative subtype of Class 2 CRISPR–Cas systems was expanded by searching for homologues of the respective effector proteins, irrespective of their adjacency to CRISPR arrays. Hence, only about half of the V‑U loci include CRISPR.

‡Both the proposed systematic Cas names and the provisional vernacular names are used for the effectors, with the exception of type II effectors, which have only systematic names, and type V‑U effectors, to which a systematic name has so far not been assigned

An intriguing question is whether the type II CRISPR–Cas system provides a substantial fitness advantage, perhaps being more efficient in defense and/or incurring a lower cost than other Class 2 variants.

Most of the Class 2 subtypes are represented in taxonomically diverse bacteria, and, furthermore, for type II and subtype V-A, the effector tree topologies differ from the topology of the species tree (Chylinski *et al.*, 2014; Zetsche *et al.*, 2015). These observations indicate that horizontal gene transfer might be a key process in the evolution of CRISPR–Cas. However, it is notable that the relatively abundant subtype VI-B seems to be restricted to the phylum *Bacteroidetes*, which perhaps reflects a unique aspect of the biology of these bacteria. Similarly, the V-U5 variant, which contains an inactivated TnpB homologue, is limited to cyanobacteria (see above), and could be involved in a distinct cyanobacterial regulatory pathway. As has been previously noted (Makarova, Haft, *et al.*, 2011; Makarova *et al.*, 2015), and is emphasized by this expansion of the diversity of Class 2 systems, apart from the identification of subtype V-A in mesophilic archaea in two instances, Class 2 systems are unique to bacteria. The exclusion of Class 2 systems from archaea, particularly from hyperthermophiles in which class 1 systems are ubiquitous, implies that there is a major functional distinction between the two classes of CRISPR–Cas system, the nature of which remains enigmatic.

Amended classification of Class 2 CRISPR–cas systems

The systematic search for novel Class 2 CRISPR–Cas loci described here led to a major expansion of the known diversity of these systems. Instead of the two types and four subtypes that were included in the latest classification (Makarova *et al.*, 2015), there are now three types and at least 10 subtypes (see Figure 8). Some uncertainty remains, owing to the lack of functional data on subtype V-U, but it seems likely that evolutionarily stable and apparently functional variants that are currently grouped into this provisional subtype, particularly V-U3, will eventually be ‘upgraded’ to subtypes in their own right. The functional characterization of V-U variants will provide a more precise classification, although it is likely that many V-U loci do not encode typical active CRISPR–Cas systems. Given the comprehensive nature of the search described here (see Figure 7), it is expected that the new variants will be extremely rare or restricted in their spread to particular groups of bacteria and archaea that are not adequately represented in current sequence databases.

The expansion of the CRISPR–Cas classification calls for a corresponding change to the nomenclature, in which at least the experimentally characterized effectors and their homologues are given new names that correspond to numbered Cas proteins (see Figure 8; table 2). Thus, the type V effectors would become Cas12a, Cas12b and Cas12c, and those of type VI would become Cas13a, Cas13b and Cas13c (numerical continuity with Cas9 is not possible because Cas10 and Cas11 are already used for other proteins) (Makarova *et al.*, 2015). Putative subtype V-U refrained from renaming it's effectors until functional evidence of a bona fide CRISPR response for these effectors is reported, at which time it was proposed that they be referred to as Cas12 proteins.

Part 3: Evolutionary origins of novel CRISPR-Cas systems

In an extension of the previous hypothesis on the independent origins of the effectors in different types and subtypes of Class 2 CRISPR–Cas systems, we use the findings on incomplete type V loci to propose a more specific evolutionary scenario (see Figure 14).

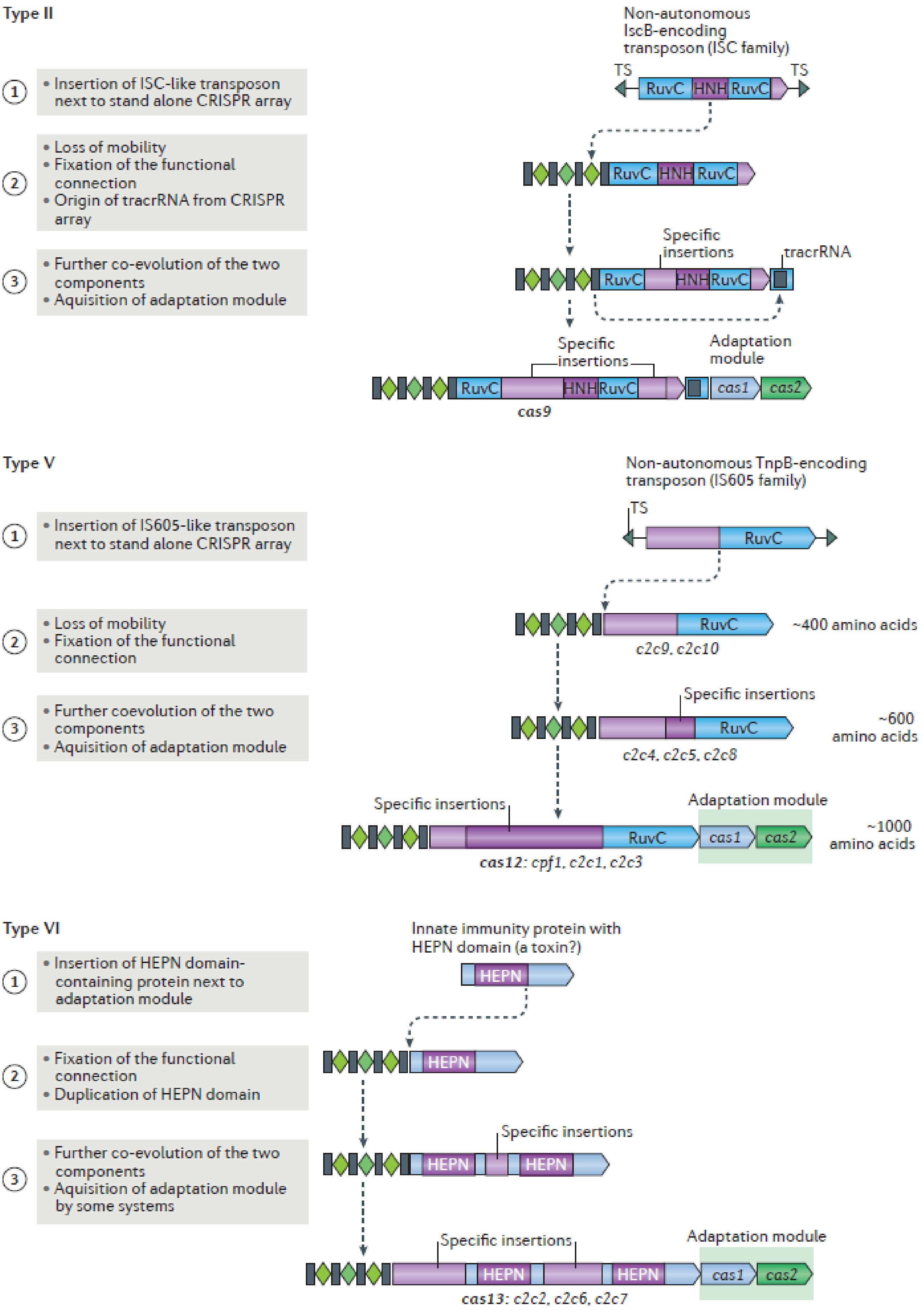

Type II
1
• Insertion of ISC-like transposon next to stand alone CRISPR array
2
• Loss of mobility
• Fixation of the functional connection
• Origin of tracrRNA from CRISPR array
3
• Further co-evolution of the two components
• Aquisition of adaptation module
Non-autonomous IscB-encoding transposon (ISC family)
TS
TS
RuvC
HNH
RuvC
Specific insertions
tracrRNA
Specific insertions
Adaptation module
cas1
cas2
cas9
Type V
1
• Insertion of IS605-like transposon next to stand alone CRISPR array
2
• Loss of mobility
• Fixation of the functional connection
3
• Further coevolution of the two components
• Aquisition of adaptation module
Non-autonomous TnpB-encoding transposon (IS605 family)
TS
RuvC
~400 amino acids
c2c9, c2c10
Specific insertions
~600 amino acids
c2c4, c2c5, c2c8
Specific insertions
Adaptation module
~1000 amino acids
cas12: cpf1, c2c1, c2c3
Type VI
1
• Insertion of HEPN domain-containing protein next to adaptation module
2
• Fixation of the functional connection
• Duplication of HEPN domain
3
• Further co-evolution of the two components
• Aquisition of adaptation module by some systems
Innate immunity protein with HEPN domain (a toxin?)
HEPN
Specific insertions
Specific insertions
Adaptation module
cas13: c2c2, c2c6, c2c7

Figure 14. Possible routes of evolution for Class 2 CRISPR–Cas systems. The figure depicts the three-step pathway of the evolutionary 'maturation' of type II, type V and type VI CRISPR–Cas systems. The systematic and/or provisional gene names are indicated below the respective 'mature' effector protein schematics and the proposed intermediate forms of type V systems. The first step involves the random insertion of a TnpB-encoding or insertion of sequences of Cas9-like protein B (IscB)-encoding transposon or a higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domain RNase-encoding gene next to a CRISPR cassette for type II, type V and type VI systems, respectively. During the second step, the functional connection between this protein and the CRISPR array is established and co-evolution begins, in particular, in the form of the accumulation of specific insertions that facilitate CRISPR RNA (crRNA) binding. For type V systems, the intermediate forms that correspond to the first and second step are identified as different type V-uncharacterized (V-U) variants. Additional components of the system could have originated during the second step, such as *trans*-acting CRISPR RNA (tracrRNA) in the case of type II systems. During the third step, further insertions lead to increased specificity of crRNA and target binding, and enable interactions with accessory proteins, such as Csn2 for type II-A and a protein with predicted transmembrane (TM) domains for type VI-B. The adaptation module is only inserted into some of the Class 2 CRISPR–Cas loci during the third step. TS stands for target site.

As discussed above, at least five distinct variants within subtype V-U show a substantial degree of evolutionary stability and consistent association with CRISPR arrays, and typically contain TnpB homologues that are intermediate in size between the compact transposon-encoded TnpB proteins and the large Class 2 effectors (see Figure 9, 10b). These groups of TnpB homologues might represent intermediate stages in independent pathways to the emergence of new CRISPR–Cas variants. The other CRISPR–*tnpB* associations are not evolutionarily conserved and are likely to result from more or less random insertions of *tnpB* genes next to CRISPR arrays; some of these loci could represent the earliest stages of the evolution of CRISPR–Cas systems.

All subtype V-U loci lack adaptation modules, which suggests that the first stage of the evolution of new Class 2 CRISPR–Cas systems involves the random insertion of a TnpB-encoding element next to an orphan CRISPR array (see Figure 14). At the next stage of evolution, the association between CRISPR and a TnpB derivative would become fixed in the microbial population, conceivably owing to the emergence of a novel function, the exact nature of which is not yet understood. This would be accompanied by an increase in the size of the protein through internal duplications and/or the insertion of additional domains (see Figure 14). The final stages include further growth of the effector protein, resulting in the typical bilobed structure, and, in some cases, its association with an adaptation module through recombination with a different CRISPR–Cas locus (see Figure 14). Compatible with this scenario, the Cas1 proteins of different subtypes of type II and of type V are homologous to different subtypes of type I (see Figure 15). The fact that no subtype V-U loci contain *cas1* and *cas2* genes, whereas many of the loci that encode typical large effector proteins do contain them, strongly suggests that the adaptation modules came last.

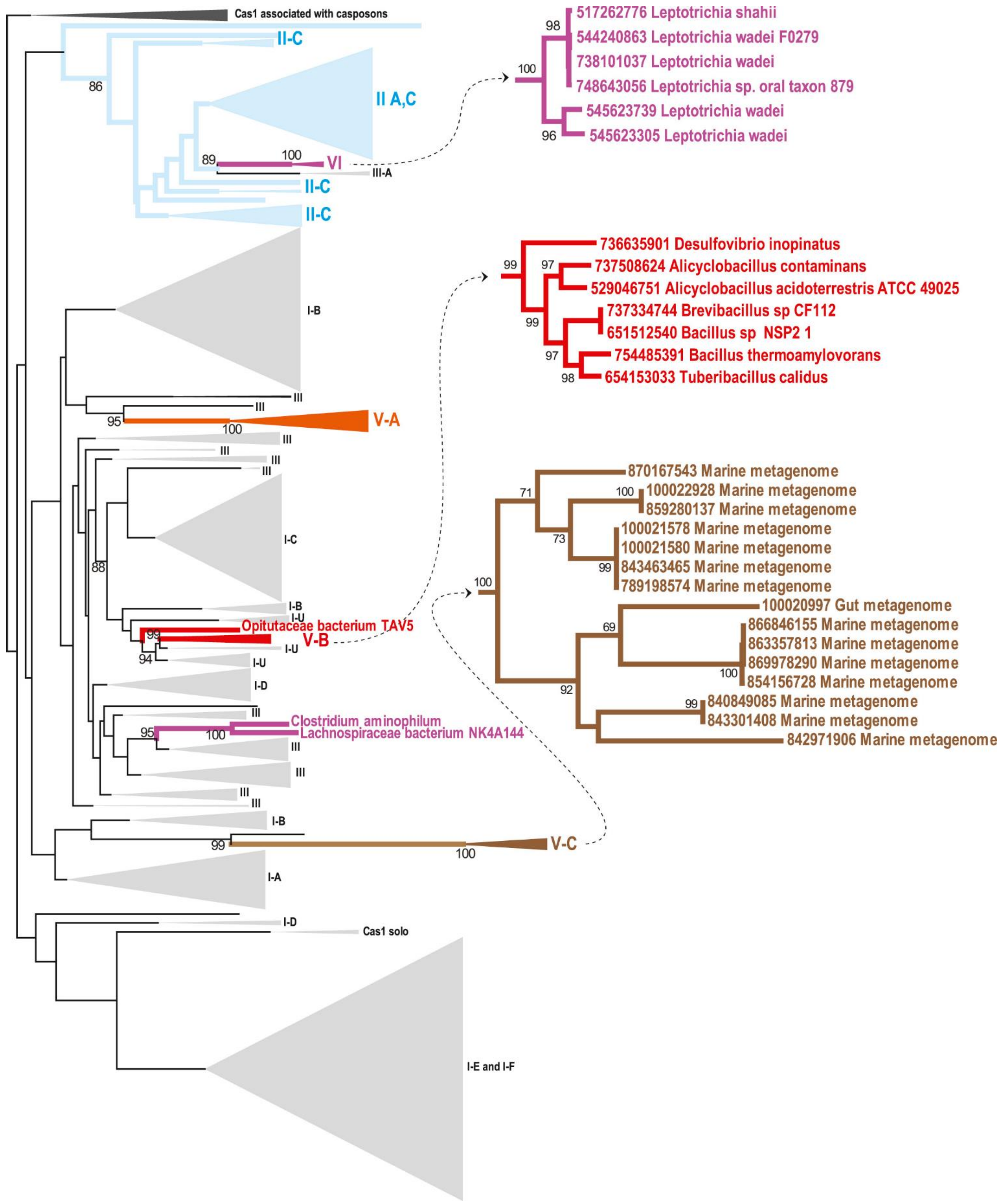

Cas1 associated with casposons
II-C
86
II A,C
89
100
VI
III-A
II-C
II-C
517262776 Leptotrichia shahii
98
544240863 Leptotrichia wadei F0279
738101037 Leptotrichia wadei
100
748643056 Leptotrichia sp. oral taxon 879
545623739 Leptotrichia wadei
96
545623305 Leptotrichia wadei
I-B
736635901 Desulfovibrio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

Figure 15. 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 is available at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Class2/. See also Methods.

The above scenario might be challenged in regard to the directionality of evolution: it is conceivable that the transposon-encoded TnpB protein actually evolved from Class 2 effectors. However, the scenario in which transposon-encoded TnpB is the ancestral form (see Figure 14) seems much more likely for several reasons. First, TnpB-encoding transposons (autonomous and non-autonomous, including some that have lost mobility) are far more abundant across a broad range of bacteria and archaea than Class 2 CRISPR–Cas systems, which are relatively rare and limited in their spread to a subset of bacterial phyla (see comprehensive census; table 2; Supplementary information S2 (box), part d). Second, and perhaps more important, the Class 2 effectors are much larger and more complex than TnpB proteins, which makes them unlikely ancestral forms. Third, the TnpB proteins are encoded in transposons, which, through their mobility, are well suited to move into the vicinity of CRISPR arrays; by contrast, CRISPR–Cas systems lack active mobility mechanisms. Finally, the observations that are reported here on the phylogeny of TnpB, in which the CRISPR-associated variants are lodged among the transposon-encoded proteins (see Figure 10a), imply the ancestral status of TnpB.

Hypothetically, a similar scenario could apply to the type VI systems (see Figure 14). A comprehensive database search for HEPN domain-containing proteins that are encoded in the vicinity of CRISPR arrays failed to identify any evolutionarily stable configurations that might have been analogous to subtype V-U, whereas it detected numerous members of the HEPN-containing Cas protein families, Csm6 and Csx1 (Supplementary information S2 (box), part g). So it seems possible that, during evolution, type VI systems recruited one of the HEPN-containing Cas proteins, which was followed by duplication of the HEPN domain and further expansion of the protein to the typical size of a Class 2 effector (see Figure 14). However, the possibility that type VI effectors originate directly from HEPN-containing toxins cannot be ruled out; further screening of new genomes and metagenomes for likely ancestors of the two HEPN domain proteins should establish the origin of type VI effectors.

Part 4: Possible Applications for Novel CRISPR-Cas Systems

Most applications of CRISPR systems have focused on the programmable DNA-targeting activity of Cas9. The cleavage activity of Cas9 can be harnessed for genome editing, including gene knockout and precise editing through homology-directed repair. Catalytically inactive (‘dead’) variants of Cas9 have been used for transcriptional control (Chavez *et al.*, 2016), epigenetic modulation (Thakore *et al.*, 2016) and imaging (Chen *et al.*, 2013; Knight *et al.*, 2015; Nelles *et al.*, 2016). All of these advances notwithstanding, Cas9 has its limitations, due to the potential for off-target effects, challenges that are associated with delivery and the difficulty of targeting RNA rather than DNA. Thus, alternative tools for CRISPR-mediated editing are in high demand.

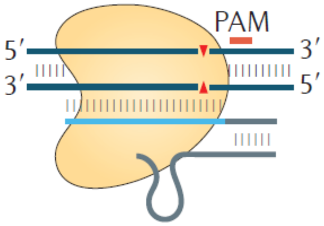
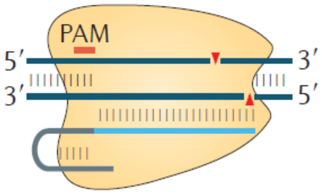
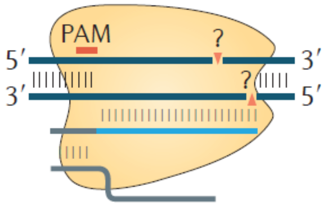
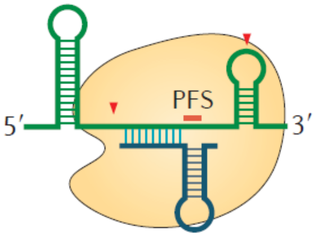
		Nuclease domains	tracrRNA	PAM	Substrate	Cleavage pattern
Type II Cas9		RuvC and HNH	Yes	3′, GC-rich	dsDNA	Blunt ends
Type V-A Cas12a (Cpf1)		RuvC and Nuc	No	5′, AT-rich	dsDNA	Staggered ends, 5′overhangs
Type V-B Cas12b (C2c1)		RuvC	Yes	5′, AT-rich	dsDNA	Staggered seven-nucleotide cut of target DNA
Type VI-A Cas13a (C2c2)		2 HEPN domains	No	5′, non-G PFS	ssRNA	Cleaves ssRNA near uracil and collateral activity

Figure 16. Functional diversity of the experimentally characterized Class 2 CRISPR–Cas systems. For each type of the Class 2 CRISPR–Cas systems (and two subtypes in the case of type V), a schematic of the complex between the effector protein, the target, crRNA and, in the case of type II and type V-B systems, *trans*-acting CRISPR RNA (tracrRNA), is shown. The position of the protospacer adjacent motif (PAM) or the protospacer flanking site (PFS) is indicated by a red bar. The small red triangles show the position of the cut, or cuts, in the target DNA or RNA molecule: dsDNA, double-stranded DNA; ssRNA, single-stranded RNA.

Although functional characterization of the Class 2 subtypes is far from complete, even at this stage, remarkable functional diversity is apparent. The manifestations of this diversity include different targets (dsDNA for type II and type V, but RNA for type VI), the requirement for tracrRNA (type II and subtype V-B require it, but not subtype V-A or type VI), the sequence of the PAM and the type of cut that is introduced into the target nucleic acid (see Figure 16). This functional diversity is a major incentive for further characterization of different Class 2 systems, as it creates opportunities for the enhancement and expansion of the capabilities of the genome editing toolbox for research, biotechnology and medicine (Hsu, Lander and Zhang, 2014). The use of Cas12a (better known as Cpf1) from the type V-A family of effectors has already yielded simpler, single RNA-guided and more specific enzymes than Cas9 for genome-editing applications (Zetsche *et al.*, 2015; Fonfara *et al.*, 2016; Hur *et al.*, 2016; Kleinstiver, Tsai, *et al.*, 2016; Li, Zhao and Wang, 2016; Y. Kim *et al.*, 2016), as well as offering an alternative PAM that would facilitate genome editing in AT-rich genomes, such as the genome of *Plasmodium falciparum*.

The continued exploration of CRISPR effector diversity, such as the recently characterized type VI-A effector Cas13a (previously known as C2c2) (Abudayyeh *et al.*, 2016), also opens the door for the development of new RNA-guided RNA-targeting technologies that enable the perturbation, modulation, modification and monitoring of specific RNA transcripts in cells. The development of an efficient programmable RNA-binding protein (for example, of a ‘dead’ Cas13a that has mutated HEPN domains) could rapidly advance our existing understanding of RNA biology. Such a tool would enable the sensing of different cellular states, the manipulation of translation, and tracking of RNA levels and localization in live cells. Although Cas9 has been modified to provide some RNA-targeting capabilities (O’Connell *et al.*, 2014), this system requires the delivery of chemically modified DNA, which limits its use for many applications, including genome-wide screening or virus delivery.

Upon binding to a complementary RNA target, Cas13a engages both specific and nonspecific RNase activities, and induces growth inhibition in *Escherichia coli* (Abudayyeh *et al.*, 2016). This feature complicates the use of Cas13a for specific RNA knockdown, but could potentially be harnessed for other applications, such as the selective ablation of cell types based on expression profiles. It remains to be investigated whether the nonspecific RNase activity of

Cas13a can be inactivated independently of its target-specific activity and whether other type VI effectors, such as Cas13b, have similar properties. Further mining of CRISPR–Cas systems, and, more broadly, of the diversity of bacterial and archaeal defence systems and of mobile genetic elements, is expected to enable new applications in biotechnology. In particular, programmable integrases or transposases that have yet to be discovered would be powerful tools for targeted genomic integration and rearrangement.

Recent study shows one example of an application for the discovered Cas13a (Gootenberg *et al.*, 2017). In this study it was shown that C2c2 can be used for CRISPR-based diagnostics to detect DNA or RNA at the attomolar level and single-base mismatch level. This method was able to detect specific strains of eukaryotic viruses and specific tumors, and to distinguish pathogenic bacteria. These properties can be used in a field-based application to detect various viruses or pathogens, DNA/RNA quantitation etc.

Conclusion

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

The new Class 2 variants show some unprecedented functional features; for example, subtype V-B requires tracrRNA, while V-A does not require it, whereas other variants, such as subtype VI-A (and probably all type VI systems), exclusively target RNA and seem to induce a toxic response in bacterial cells. Subtype V-U is expected to show even more unusual properties. This functional diversity offers potential for the development of new, versatile genome-editing and regulation tools. We provide indications that different Class 2 types and subtypes independently originate from mobile elements that encode diverse TnpB proteins (type II and type V) and from HEPN domain-containing proteins (type VI) that ultimately originate from mRNA-cleaving toxins. Notwithstanding this remarkable diversity, we believe that the computational pipeline that is applied here provides for a nearly exhaustive identification of Class 2 systems. Additional variants that remain to be found will be either extremely rare or confined to bacterial phyla that are currently unknown or poorly sampled. However, as shown by the example of type VI, despite the rarity and/or narrow spread of such variants, their biological features could be of major interest and potential value for new applications.

Research results

1. A computational pipeline was created to search for CRISPR-Cas Class 2 systems in genomic and metagenomic prokaryotic databases.

2. Six novel CRISPR-Cas Class 2 systems, including a collection of unknown V-U variants, were found: subtypes V-B, V-C, V-U with RuvC nuclease domains; and VI-A, VI-B, VI-C with HEPN RNase domains. These systems were computationally and experimentally characterized. Three subtypes were experimentally validated by collaborators and the results matched the predictions: V-B was shown to be a tracrRNA dependent CRISPR-Cas effector complex that provides defense against DNA targets (Shmakov *et al.*, 2015); VI-A and VI-B were shown to provide defense against RNA viruses (Abudayyeh *et al.*, 2016; Smargon *et al.*, 2017).

3. Amended classification of CRISPR-Cas systems was proposed, to include the six discovered subtypes.

4. A comprehensive census for CRISPR-Cas Class 2 systems was provided, showing the prevalence of Class 2 systems among bacteria and archaea in up-to-date known genomic sequences.

5. Possible evolutionary origins of Class 2 systems were described, showing the path from transposons to mature (or possibly immature for Type V-U) CRISPR-Cas systems.

6. Possible biotechnological applications, including genome editing, were proposed for the novel CRISPR-Cas systems.

References

Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B. T., Shmakov, S., Makarova, K. S., Semenova, E., Minakhin, L., Severinov, K., Regev, A., Lander, E. S., Koonin, E. V and Zhang, F. (2016) ‘C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector.’, *Science (New York, N.Y.)*, 353(6299), p. aaf5573. doi: 10.1126/science.aaf5573.

Almendros, C., Guzmán, N. M., García-Martínez, J. and Mojica, F. J. M. (2016) ‘Anti-cas spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems.’, *Nature microbiology*, 1(8), p. 16081. doi: 10.1038/nmicrobiol.2016.81.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) ‘Basic local alignment search tool.’, *Journal of molecular biology*, 215(3), pp. 403–10. doi: 10.1016/S0022-2836(05)80360-2.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) ‘Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.’, *Nucleic acids research*, 25(17), pp. 3389–402. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9254694.

Amitai, G. and Sorek, R. (2016) ‘CRISPR-Cas adaptation: insights into the mechanism of action.’, *Nature reviews. Microbiology*, 14(2), pp. 67–76. doi: 10.1038/nrmicro.2015.14.

Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V and Aravind, L. (2013) ‘Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing.’, *Biology direct*, 8, p. 15. doi: 10.1186/1745-6150-8-15.

Aravind, L., Makarova, K. S. and Koonin, E. V (2000) ‘SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.’, *Nucleic acids research*, 28(18), pp. 3417–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10982859.

Bao, W. and Jurka, J. (2013) ‘Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements.’, *Mobile DNA*, 4(1), p. 12. doi: 10.1186/1759-8753-4-12.

Barrangou, R. (2013) ‘CRISPR-Cas systems and RNA-guided interference.’, *Wiley interdisciplinary reviews. RNA*, 4(3), pp. 267–78. doi: 10.1002/wrna.1159.

Barrangou, R. and Doudna, J. A. (2016) ‘Applications of CRISPR technologies in research and beyond.’, *Nature biotechnology*, 34(9), pp. 933–941. doi: 10.1038/nbt.3659.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. and Horvath, P. (2007) ‘CRISPR provides acquired resistance against viruses in prokaryotes.’, *Science (New York, N.Y.)*, 315(5819), pp. 1709–12. doi: 10.1126/science.1138140.

Beloglazova, N., Kuznedelov, K., Flick, R., Datsenko, K. A., Brown, G., Popovic, A., Lemak, S., Semenova, E., Severinov, K. and Yakunin, A. F. (2015) ‘CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli.’, *Nucleic acids research*, 43(1), pp. 530–43. doi: 10.1093/nar/gku1285.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2013) ‘GenBank.’, *Nucleic acids research*, 41(Database issue), pp. D36-42. doi: 10.1093/nar/gks1195.

Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) ‘GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.’, *Nucleic acids research*, 29(12), pp. 2607–18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11410670.

Bhaya, D., Davison, M. and Barrangou, R. (2011) ‘CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation.’, *Annual review of genetics*, 45, pp. 273–97. doi: 10.1146/annurev-genet-110410-132430.

Bikard, D., Euler, C. W., Jiang, W., Nussenzweig, P. M., Goldberg, G. W., Duportet, X., Fischetti, V. A. and Marraffini, L. A. (2014) ‘Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials.’, *Nature biotechnology*, 32(11), pp. 1146–50. doi: 10.1038/nbt.3043.

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F. and Marraffini, L. A. (2013) ‘Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system.’, *Nucleic acids research*, 41(15), pp. 7429–37. doi: 10.1093/nar/gkt520.

Blosser, T. R., Loeff, L., Westra, E. R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S. J. J. and Joo, C. (2015) ‘Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex.’, *Molecular cell*, 58(1), pp. 60–70. doi: 10.1016/j.molcel.2015.01.028.

Bolotin, A., Quinquis, B., Sorokin, A. and Ehrlich, S. D. (2005) ‘Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.’, *Microbiology (Reading, England)*, 151(Pt 8), pp. 2551–61. doi: 10.1099/mic.0.28048-0.

Bortesi, L. and Fischer, R. (no date) ‘The CRISPR/Cas9 system for plant genome editing and beyond.’, *Biotechnology advances*, 33(1), pp. 41–52. doi: 10.1016/j.biotechadv.2014.12.006.

Briner, A. E., Donohoue, P. D., Gomaa, A. A., Selle, K., Slorach, E. M., Nye, C. H., Haurwitz, R. E., Beisel, C. L., May, A. P. and Barrangou, R. (2014) ‘Guide RNA functional modules direct Cas9 activity and orthogonality.’, *Molecular cell*, 56(2), pp. 333–9. doi: 10.1016/j.molcel.2014.09.019.

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V and van der Oost, J. (2008) ‘Small CRISPR RNAs guide antiviral defense in prokaryotes.’, *Science (New York, N.Y.)*, 321(5891), pp. 960–4. doi: 10.1126/science.1159689.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. and Venter, J. C. (1996) ‘Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii.’, *Science (New York, N.Y.)*, 273(5278), pp. 1058–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8688087.

Carte, J., Wang, R., Li, H., Terns, R. M. and Terns, M. P. (2008) ‘Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes.’, *Genes & development*, 22(24), pp. 3489–96. doi: 10.1101/gad.1742908.

Charpentier, E., Richter, H., van der Oost, J. and White, M. F. (2015) ‘Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity.’, *FEMS microbiology reviews*, 39(3), pp. 428–41. doi: 10.1093/femsre/fuv023.

Chavez, A., Tuttle, M., Pruitt, B. W., Ewen-Campen, B., Chari, R., Ter-Ovanesyan, D., Haque, S. J., Cecchi, R. J., Kowal, E. J. K., Buchthal, J., Housden, B. E., Perrimon, N., Collins, J. J. and Church, G. (2016) ‘Comparison of Cas9 activators in multiple species.’, *Nature methods*, 13(7), pp. 563–7. doi: 10.1038/nmeth.3871.

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S. and Huang, B. (2013) ‘Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system.’, *Cell*, 155(7), pp. 1479–91. doi: 10.1016/j.cell.2013.12.001.

Chen, L., Tang, L., Xiang, H., Jin, L., Li, Q., Dong, Y., Wang, W. and Zhang, G. (2014) ‘Advances in genome editing technology and its promising application in evolutionary and ecological studies.’, *GigaScience*, 3, p. 24. doi: 10.1186/2047-217X-3-24.

Cho, S. W., Kim, S., Kim, J. M. and Kim, J.-S. (2013) ‘Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease.’, *Nature biotechnology*, 31(3), pp. 230–2. doi: 10.1038/nbt.2507.

Chylinski, K., Makarova, K. S., Charpentier, E. and Koonin, E. V (2014) ‘Classification and evolution of type II CRISPR-Cas systems.’, *Nucleic acids research*, 42(10), pp. 6091–105. doi: 10.1093/nar/gku241.

Chylinski, K., Le Rhun, A. and Charpentier, E. (2013) ‘The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems.’, *RNA biology*, 10(5), pp. 726–37. doi: 10.4161/rna.24321.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A. and Zhang, F. (2013) ‘Multiplex genome engineering using CRISPR/Cas systems.’, *Science (New York, N.Y.)*, 339(6121), pp. 819–23. doi: 10.1126/science.1231143.

Cyranoski, D. (2016) ‘CRISPR gene-editing tested in a person for the first time’, *Nature*, 539(7630), pp. 479–479. doi: 10.1038/nature.2016.20988.

D’Astolfo, D. S., Pagliero, R. J., Pras, A., Karthaus, W. R., Clevers, H., Prasad, V., Lebbink, R. J., Rehmann, H. and Geijsen, N. (2015) ‘Efficient intracellular delivery of native proteins.’, *Cell*, 161(3), pp. 674–90. doi: 10.1016/j.cell.2015.03.028.

‘Database resources of the National Center for Biotechnology Information’ (2016) *Nucleic Acids Research*, 44(D1), pp. D7–D19. doi: 10.1093/nar/gkv1290.

Datsenko, K. A., Pougach, K., Tikhonov, A., Wanner, B. L., Severinov, K. and Semenova, E. (2012) ‘Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system.’, *Nature communications*, 3, p. 945. doi: 10.1038/ncomms1937.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J. and Charpentier, E. (2011) ‘CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.’, *Nature*, 471(7340), pp. 602–7. doi: 10.1038/nature09886.

Deng, L., Garrett, R. A., Shah, S. A., Peng, X. and She, Q. (2013) ‘A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus.’, *Molecular microbiology*, 87(5), pp. 1088–99. doi: 10.1111/mmi.12152.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P. and Moineau, S. (2008) ‘Phage response to CRISPR-encoded resistance in Streptococcus thermophilus.’, *Journal of bacteriology*, 190(4), pp. 1390–400. doi: 10.1128/JB.01412-07.

DiCarlo, J. E., Norville, J. E., Mali, P., Rios, X., Aach, J. and Church, G. M. (2013) ‘Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems.’, *Nucleic acids research*, 41(7), pp. 4336–43. doi: 10.1093/nar/gkt135.

Dillingham, M. S. and Kowalczykowski, S. C. (2008) ‘RecBCD enzyme and the repair of double-stranded DNA breaks.', *Microbiology and molecular biology reviews : MMBR*, 72(4), p. 642–71, Table of Contents. doi: 10.1128/MMBR.00020-08.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J. and Root, D. E. (2016) 'Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9.', *Nature biotechnology*, 34(2), pp. 184–91. doi: 10.1038/nbt.3437.

Dong, D., Ren, K., Qiu, X., Zheng, J., Guo, M., Guan, X., Liu, H., Li, N., Zhang, B., Yang, D., Ma, C., Wang, S., Wu, D., Ma, Y., Fan, S., Wang, J., Gao, N. and Huang, Z. (2016) 'The crystal structure of Cpf1 in complex with CRISPR RNA.', *Nature*, 532(7600), pp. 522–6. doi: 10.1038/nature17944.

East-Seletsky, A., O'Connell, M. R., Knight, S. C., Burstein, D., Cate, J. H. D., Tjian, R. and Doudna, J. A. (2016) 'Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection.', *Nature*, 538(7624), pp. 270–273. doi: 10.1038/nature19802.

Ebina, H., Misawa, N., Kanemura, Y. and Koyanagi, Y. (2013) 'Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus.', *Scientific reports*, 3, p. 2510. doi: 10.1038/srep02510.

Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.', *Nucleic acids research*, 32(5), pp. 1792–7. doi: 10.1093/nar/gkh340.

Edgar, R. C. (2007) 'PILER-CR: fast and accurate identification of CRISPR repeats.', *BMC bioinformatics*, 8, p. 18. doi: 10.1186/1471-2105-8-18.

Edgar, R. C. (2010) 'Search and clustering orders of magnitude faster than BLAST.', *Bioinformatics (Oxford, England)*, 26(19), pp. 2460–1. doi: 10.1093/bioinformatics/btq461.

Elmore, J. R., Sheppard, N. F., Ramia, N., Deighan, T., Li, H., Terns, R. M. and Terns, M. P. (2016) 'Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system.', *Genes & development*, 30(4), pp. 447–59. doi: 10.1101/gad.272153.115.

Estrella, M. A., Kuo, F.-T. and Bailey, S. (2016) 'RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex.', *Genes & development*, 30(4), pp. 460–70. doi:

10.1101/gad.273722.115.

Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. and Charpentier, E. (2016) ‘The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA.’, *Nature*, 532(7600), pp. 517–21. doi: 10.1038/nature17945.

Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H. and Moineau, S. (2010) ‘The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA.’, *Nature*, 468(7320), pp. 67–71. doi: 10.1038/nature09523.

Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. (2012) ‘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*, 109(39), pp. E2579-86. doi: 10.1073/pnas.1208507109.

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M. and Weissman, J. S. (2014) ‘Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation.’, *Cell*, 159(3), pp. 647–61. doi: 10.1016/j.cell.2014.09.029.

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S. and Qi, L. S. (2013) ‘CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes.’, *Cell*, 154(2), pp. 442–51. doi: 10.1016/j.cell.2013.06.044.

Goldberg, G. W., Jiang, W., Bikard, D. and Marraffini, L. A. (2014) ‘Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting.’, *Nature*, 514(7524), pp. 633–7. doi: 10.1038/nature13637.

Gomes-Filho, J. V., Zaramela, L. S., Italiani, V. C. da S., Baliga, N. S., Vêncio, R. Z. N. and Koide, T. (2015) ‘Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea.’, *RNA biology*, 12(5), pp. 490–500. doi: 10.1080/15476286.2015.1019998.

Gong, B., Shin, M., Sun, J., Jung, C.-H., Bolt, E. L., van der Oost, J. and Kim, J.-S. (2014) ‘Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3.’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(46), pp. 16359–64. doi: 10.1073/pnas.1410806111.

Gootenberg, J. S., Abudayyeh, O. O., Lee, J. W., Essletzbichler, P., Dy, A. J., Joung, J., Verdine, V., Donghia, N., Daringer, N. M., Freije, C. A., Myhrvold, C., Bhattacharyya, R. P., Livny, J., Regev, A., Koonin, E. V, Hung, D. T., Sabeti, P. C., Collins, J. J. and Zhang, F. (2017) ‘Nucleic acid detection with CRISPR-Cas13a/C2c2.’, *Science (New York, N.Y.)*. doi: 10.1126/science.aam9321.

Grissa, I., Vergnaud, G. and Pourcel, C. (2007) ‘CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.’, *Nucleic acids research*, 35(Web Server issue), pp. W52-7. doi: 10.1093/nar/gkm360.

Grynberg, M., Erlandsen, H. and Godzik, A. (2003) ‘HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins.’, *Trends in biochemical sciences*, 28(5), pp. 224–6. doi: 10.1016/S0968-0004(03)00060-4.

Haft, D. H., Selengut, J., Mongodin, E. F. and Nelson, K. E. (2005) ‘A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.’, *PLoS computational biology*, 1(6), p. e60. doi: 10.1371/journal.pcbi.0010060.

Hale, C. R., Cocozaki, A., Li, H., Terns, R. M. and Terns, M. P. (2014) ‘Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex.’, *Genes & development*, 28(21), pp. 2432–43. doi: 10.1101/gad.250712.114.

Hale, C. R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A. M., Glover, C. V. C., Graveley, B. R., Terns, R. M. and Terns, M. P. (2012) ‘Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs.’, *Molecular cell*, 45(3), pp. 292–302. doi: 10.1016/j.molcel.2011.10.023.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M. and Terns, M. P. (2009) ‘RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex.’, *Cell*, 139(5), pp. 945–56. doi: 10.1016/j.cell.2009.07.040.

Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. and Doudna, J. A. (2010) ‘Sequence- and structure-specific RNA processing by a CRISPR endonuclease.’, *Science (New York, N.Y.)*, 329(5997), pp. 1355–8. doi: 10.1126/science.1192272.

Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D. and Marraffini, L. A. (2015) ‘Cas9 specifies functional viral targets during CRISPR-Cas adaptation.’, *Nature*, 519(7542), pp. 199–202. doi: 10.1038/nature14245.

Hermans, P. W., van Soolingen, D., Bik, E. M., de Haas, P. E., Dale, J. W. and van Embden, J. D. (1991) ‘Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains.’, *Infection and immunity*, 59(8), pp. 2695–705. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1649798.

Hickman, A. B. and Dyda, F. (2015) ‘The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications.’, *Nucleic acids research*, 43(22), pp. 10576–87. doi: 10.1093/nar/gkv1180.

Hille, F. and Charpentier, E. (2016) ‘CRISPR-Cas: biology, mechanisms and relevance’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1707), p. 20150496. doi: 10.1098/rstb.2015.0496.

Hilton, I. B., D’Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E. and Gersbach, C. A. (2015) ‘Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers.’, *Nature biotechnology*, 33(5), pp. 510–7. doi: 10.1038/nbt.3199.

Hoe, N., Nakashima, K., Grigsby, D., Pan, X., Dou, S. J., Naidich, S., Garcia, M., Kahn, E., Bergmire-Sweat, D. and Musser, J. M. (no date) ‘Rapid molecular genetic subtyping of serotype M1 group A Streptococcus strains.’, *Emerging infectious diseases*, 5(2), pp. 254–63. doi: 10.3201/eid0502.990210.

Horvath, P., Romero, D. A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) 'Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus.', *Journal of bacteriology*, 190(4), pp. 1401–12. doi: 10.1128/JB.01415-07.

Hsu, P. D., Lander, E. S. and Zhang, F. (2014) 'Development and applications of CRISPR-Cas9 for genome engineering.', *Cell*, 157(6), pp. 1262–78. doi: 10.1016/j.cell.2014.05.010.

Hu, W., Kaminski, R., Yang, F., Zhang, Y., Cosentino, L., Li, F., Luo, B., Alvarez-Carbonell, D., Garcia-Mesa, Y., Karn, J., Mo, X. and Khalili, K. (2014) 'RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection.', *Proceedings of the National Academy of Sciences of the United States of America*, 111(31), pp. 11461–6. doi: 10.1073/pnas.1405186111.

Huo, Y., Nam, K. H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M. D., Zhou, S., Rajashankar, K., Kurinov, I., Zhang, R. and Ke, A. (2014) 'Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation.', *Nature structural & molecular biology*, 21(9), pp. 771–7. doi: 10.1038/nsmb.2875.

Hur, J. K., Kim, K., Been, K. W., Baek, G., Ye, S., Hur, J. W., Ryu, S.-M., Lee, Y. S. and Kim, J.-S. (2016) 'Targeted mutagenesis in mice by electroporation of Cpf1 ribonucleoproteins.', *Nature biotechnology*, 34(8), pp. 807–8. doi: 10.1038/nbt.3596.

Iranzo, J., Lobkovsky, A. E., Wolf, Y. I. and Koonin, E. V (2015) 'Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems.', *BMC evolutionary biology*, 15, p. 43. doi: 10.1186/s12862-015-0324-2.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A. (1987) 'Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.', *Journal of bacteriology*, 169(12), pp. 5429–33. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3316184.

Jackson, R. N., Golden, S. M., van Erp, P. B. G., Carter, J., Westra, E. R., Brouns, S. J. J., van der Oost, J., Terwilliger, T. C., Read, R. J. and Wiedenheft, B. (2014) 'Structural biology.

Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli.', *Science (New York, N.Y.)*, 345(6203), pp. 1473–9. doi: 10.1126/science.1256328.

Jackson, R. N., Lavin, M., Carter, J. and Wiedenheft, B. (2014) 'Fitting CRISPR-associated Cas3 into the helicase family tree.', *Current opinion in structural biology*, 24, pp. 106–14. doi: 10.1016/j.sbi.2014.01.001.

Jansen, R., van Embden, J. D. A., Gaastra, W. and Schouls, L. M. (2002) 'Identification of a novel family of sequence repeats among prokaryotes.', *Omics : a journal of integrative biology*, 6(1), pp. 23–33. doi: 10.1089/15362310252780816.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. and Charpentier, E. (2012) 'A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.', *Science (New York, N.Y.)*, 337(6096), pp. 816–21. doi: 10.1126/science.1225829.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E. and Doudna, J. (2013) 'RNA-programmed genome editing in human cells.', *eLife*, 2, p. e00471. doi: 10.7554/eLife.00471.

Jore, M. M., Lundgren, M., van Duijn, E., Bultema, J. B., Westra, E. R., Waghmare, S. P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., Beijer, M. R., Barendregt, A., Zhou, K., Snijders, A. P. L., Dickman, M. J., Doudna, J. A., Boekema, E. J., Heck, A. J. R., van der Oost, J. and Brouns, S. J. J. (2011) 'Structural basis for CRISPR RNA-guided DNA recognition by Cascade.', *Nature structural & molecular biology*, 18(5), pp. 529–36. doi: 10.1038/nsmb.2019.

Kapitonov, V. V, Makarova, K. S. and Koonin, E. V (2015) 'ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs.', *Journal of bacteriology*, 198(5), pp. 797–807. doi: 10.1128/JB.00783-15.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability.', *Molecular biology and evolution*, 30(4), pp. 772–80. doi: 10.1093/molbev/mst010.

Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M. and Maehr, R. (2015) 'Functional annotation of native enhancers with a Cas9-histone demethylase fusion.', *Nature methods*, 12(5), pp. 401–3. doi: 10.1038/nmeth.3325.

Kiani, S., Beal, J., Ebrahimkhani, M. R., Huh, J., Hall, R. N., Xie, Z., Li, Y. and Weiss, R. (2014) 'CRISPR transcriptional repression devices and layered circuits in mammalian cells.', *Nature methods*, 11(7), pp. 723–6. doi: 10.1038/nmeth.2969.

Kim, D., Kim, J., Hur, J. K., Been, K. W., Yoon, S.-H. and Kim, J.-S. (2016) 'Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells.', *Nature biotechnology*, 34(8), pp. 863–8. doi: 10.1038/nbt.3609.

Kim, Y., Cheong, S.-A., Lee, J. G., Lee, S.-W., Lee, M. S., Baek, I.-J. and Sung, Y. H. (2016) 'Generation of knockout mice by Cpf1-mediated gene targeting.', *Nature biotechnology*, 34(8), pp. 808–10. doi: 10.1038/nbt.3614.

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z. and Joung, J. K. (2016) 'High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects.', *Nature*, 529(7587), pp. 490–5. doi: 10.1038/nature16526.

Kleinstiver, B. P., Tsai, S. Q., Prew, M. S., Nguyen, N. T., Welch, M. M., Lopez, J. M., McCaw, Z. R., Aryee, M. J. and Joung, J. K. (2016) 'Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells.', *Nature biotechnology*, 34(8), pp. 869–74. doi: 10.1038/nbt.3620.

Knight, S. C., Xie, L., Deng, W., Guglielmi, B., Witkowsky, L. B., Bosanac, L., Zhang, E. T., El Beheiry, M., Masson, J.-B., Dahan, M., Liu, Z., Doudna, J. A. and Tjian, R. (2015) 'Dynamics of CRISPR-Cas9 genome interrogation in living cells.', *Science (New York, N.Y.)*, 350(6262), pp. 823–6. doi: 10.1126/science.aac6572.

Konermann, S., Brigham, M. D., Trevino, A. E., Hsu, P. D., Heidenreich, M., Cong, L., Platt, R. J., Scott, D. A., Church, G. M. and Zhang, F. (2013) 'Optical control of mammalian endogenous transcription and epigenetic states.', *Nature*, 500(7463), pp. 472–6. doi: 10.1038/nature12466.

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu,

P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O. and Zhang, F. (2015) ‘Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex.’, *Nature*, 517(7536), pp. 583–8. doi: 10.1038/nature14136.

Koonin, E. V, Dolja, V. V and Krupovic, M. (2015) ‘Origins and evolution of viruses of eukaryotes: The ultimate modularity.’, *Virology*, 479–480, pp. 2–25. doi: 10.1016/j.virol.2015.02.039.

Koonin, E. V and Krupovic, M. (2015) ‘Evolution of adaptive immunity from transposable elements combined with innate immune systems.’, *Nature reviews. Genetics*, 16(3), pp. 184–92. doi: 10.1038/nrg3859.

Koonin, E. V and Wolf, Y. I. (2016) ‘Just how Lamarckian is CRISPR-Cas immunity: the continuum of evolvability mechanisms.’, *Biology direct*, 11(1), p. 9. doi: 10.1186/s13062-016-0111-z.

Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. and Koonin, E. V (2014) ‘Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity.’, *BMC biology*, 12, p. 36. doi: 10.1186/1741-7007-12-36.

Krupovic, M., Shmakov, S., Makarova, K. S., Forterre, P. and Koonin, E. V (2016) ‘Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity.’, *Genome biology and evolution*, 8(2), pp. 375–86. doi: 10.1093/gbe/evw006.

Levy, A., Goren, M. G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U. and Sorek, R. (2015) ‘CRISPR adaptation biases explain preference for acquisition of foreign DNA.’, *Nature*, 520(7548), pp. 505–10. doi: 10.1038/nature14302.

Li, M., Wang, R., Zhao, D. and Xiang, H. (2014) ‘Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process.’, *Nucleic acids research*, 42(4), pp. 2483–92. doi: 10.1093/nar/gkt1154.

Li, S.-Y., Zhao, G.-P. and Wang, J. (2016) ‘C-Brick: A New Standard for Assembly of Biological Parts Using Cpf1.’, *ACS synthetic biology*, 5(12), pp. 1383–1388. doi:

10.1021/acssynbio.6b00114.

Liu, L., Chen, P., Wang, M., Li, X., Wang, J., Yin, M. and Wang, Y. (2017) 'C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism.', *Molecular cell*, 65(2), pp. 310–322. doi: 10.1016/j.molcel.2016.11.040.

Liu, Y., Zeng, Y., Liu, L., Zhuang, C., Fu, X., Huang, W. and Cai, Z. (2014) 'Synthesizing AND gate genetic circuits based on CRISPR-Cas9 for identification of bladder cancer cells.', *Nature communications*, 5, p. 5393. doi: 10.1038/ncomms6393.

Majumdar, S., Zhao, P., Pfister, N. T., Compton, M., Olson, S., Glover, C. V. C., Wells, L., Graveley, B. R., Terns, R. M. and Terns, M. P. (2015) 'Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus.', *RNA (New York, N.Y.)*, 21(6), pp. 1147–58. doi: 10.1261/rna.049130.114.

Makarova, K. S., Anantharaman, V., Aravind, L. and Koonin, E. V (2012) 'Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes.', *Biology direct*, 7, p. 40. doi: 10.1186/1745-6150-7-40.

Makarova, K. S., Anantharaman, V., Grishin, N. V, Koonin, E. V and Aravind, L. (2014) 'CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems.', *Frontiers in genetics*, 5, p. 102. doi: 10.3389/fgene.2014.00102.

Makarova, K. S., Aravind, L., Grishin, N. V, Rogozin, I. B. and Koonin, E. V (2002) 'A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.', *Nucleic acids research*, 30(2), pp. 482–96. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11788711.

Makarova, K. S., Aravind, L., Wolf, Y. I. and Koonin, E. V (2011) 'Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems.', *Biology direct*, 6, p. 38. doi: 10.1186/1745-6150-6-38.

Makarova, K. S., Grishin, N. V, Shabalina, S. A., Wolf, Y. I. and Koonin, E. V (2006) 'A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.', *Biology direct*, 1, p. 7. doi: 10.1186/1745-6150-1-7.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., van der Oost, J. and Koonin, E. V (2011) 'Evolution and classification of the CRISPR-Cas systems.', *Nature reviews. Microbiology*, 9(6), pp. 467–77. doi: 10.1038/nrmicro2577.

Makarova, K. S. and Koonin, E. V (2015) 'Annotation and Classification of CRISPR-Cas Systems.', *Methods in molecular biology (Clifton, N.J.)*, 1311, pp. 47–75. doi: 10.1007/978-1-4939-2687-9_4.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., Backofen, R. and Koonin, E. V (2015) 'An updated evolutionary classification of CRISPR-Cas systems.', *Nature reviews. Microbiology*, 13(11), pp. 722–36. doi: 10.1038/nrmicro3569.

Makarova, K. S., Wolf, Y. I. and Koonin, E. V (2013a) 'Comparative genomics of defense systems in archaea and bacteria.', *Nucleic acids research*, 41(8), pp. 4360–77. doi: 10.1093/nar/gkt157.

Makarova, K. S., Wolf, Y. I. and Koonin, E. V (2013b) 'The basic building blocks and evolution of CRISPR-CAS systems.', *Biochemical Society transactions*, 41(6), pp. 1392–400. doi: 10.1042/BST20130038.

Makarova, K. S., Wolf, Y. I., Snir, S. and Koonin, E. V (2011) 'Defense islands in bacterial and archaeal genomes and prediction of novel defense systems.', *Journal of bacteriology*, 193(21), pp. 6039–56. doi: 10.1128/JB.05535-11.

Mali, P., Esvelt, K. M. and Church, G. M. (2013) 'Cas9 as a versatile tool for engineering biology.', *Nature methods*, 10(10), pp. 957–63. doi: 10.1038/nmeth.2649.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E. and Church,

G. M. (2013) ‘RNA-guided human genome engineering via Cas9.’, *Science (New York, N.Y.)*, 339(6121), pp. 823–6. doi: 10.1126/science.1232033.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y. and Bryant, S. H. (2017) ‘CDD/SPARCLE: functional classification of proteins via subfamily domain architectures.’, *Nucleic acids research*, 45(D1), pp. D200–D203. doi: 10.1093/nar/gkw1129.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V, Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C. and Bryant, S. H. (2011) ‘CDD: a Conserved Domain Database for the functional annotation of proteins.’, *Nucleic acids research*, 39(Database issue), pp. D225-9. doi: 10.1093/nar/gkq1189.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. and Bryant, S. H. (2002) ‘CDD: a database of conserved domain alignments with links to domain three-dimensional structure.’, *Nucleic acids research*, 30(1), pp. 281–3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11752315.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D. and Bryant, S. H. (2013) ‘CDD: conserved domains and protein three-dimensional structure.’, *Nucleic acids research*, 41(Database issue), pp. D348-52. doi: 10.1093/nar/gks1243.

Marraffini, L. A. (2015) ‘CRISPR-Cas immunity in prokaryotes.’, *Nature*, 526(7571), pp. 55–61. doi: 10.1038/nature15386.

Marraffini, L. A. and Sontheimer, E. J. (2008) ‘CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA.’, *Science (New York, N.Y.)*, 322(5909), pp. 1843–5. doi: 10.1126/science.1165771.

Marraffini, L. A. and Sontheimer, E. J. (2010a) ‘CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea.’, *Nature reviews. Genetics*, 11(3), pp. 181–90. doi: 10.1038/nrg2749.

Marraffini, L. A. and Sontheimer, E. J. (2010b) ‘Self versus non-self discrimination during CRISPR RNA-directed immunity.’, *Nature*, 463(7280), pp. 568–71. doi: 10.1038/nature08703.

Mohanraju, P., Makarova, K. S., Zetsche, B., Zhang, F., Koonin, E. V and van der Oost, J. (2016) ‘Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems.’, *Science (New York, N.Y.)*, 353(6299), p. aad5147. doi: 10.1126/science.aad5147.

Mojica, F. J., Díez-Villaseñor, C., Soria, E. and Juez, G. (2000) ‘Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria.’, *Molecular microbiology*, 36(1), pp. 244–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10760181.

Mojica, F. J., Ferrer, C., Juez, G. and Rodríguez-Valera, F. (1995) ‘Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning.’, *Molecular microbiology*, 17(1), pp. 85–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7476211.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. and Almendros, C. (2009) ‘Short motif sequences determine the targets of the prokaryotic CRISPR defence system.’, *Microbiology (Reading, England)*, 155(3), pp. 733–740. doi: 10.1099/mic.0.023960-0.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. and Soria, E. (2005) ‘Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.’, *Journal of molecular evolution*, 60(2), pp. 174–82. doi: 10.1007/s00239-004-0046-3.

Mulepati, S. and Bailey, S. (2011) ‘Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3).’, *The Journal of biological chemistry*, 286(36), pp. 31896–903. doi: 10.1074/jbc.M111.270017.

Nakata, A., Amemura, M. and Makino, K. (1989) ‘Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome.’, *Journal of bacteriology*, 171(6), pp. 3553–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2656660.

Nam, K. H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M. P. and Ke, A. (2012) ‘Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system.’, *Structure (London, England : 1993)*, 20(9), pp. 1574–84. doi: 10.1016/j.str.2012.06.016.

Nelles, D. A., Fang, M. Y., O’Connell, M. R., Xu, J. L., Markmiller, S. J., Doudna, J. A. and Yeo, G. W. (2016) ‘Programmable RNA Tracking in Live Cells with CRISPR/Cas9.’, *Cell*, 165(2), pp. 488–96. doi: 10.1016/j.cell.2016.02.054.

Niewoehner, O. and Jinek, M. (2016) ‘Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6.’, *RNA (New York, N.Y.)*, 22(3), pp. 318–29. doi: 10.1261/rna.054098.115.

Nishimasu, H., Cong, L., Yan, W. X., Ran, F. A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F. and Nureki, O. (2015) ‘Crystal Structure of Staphylococcus aureus Cas9.’, *Cell*, 162(5), pp. 1113–26. doi: 10.1016/j.cell.2015.08.007.

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F. and Nureki, O. (2014) ‘Crystal structure of Cas9 in complex with guide RNA and target DNA.’, *Cell*, 156(5), pp. 935–49. doi: 10.1016/j.cell.2014.02.001.

Nissim, L., Perli, S. D., Fridkin, A., Perez-Pinera, P. and Lu, T. K. (2014) ‘Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells.’, *Molecular cell*, 54(4), pp. 698–710. doi: 10.1016/j.molcel.2014.04.022.

Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. and Doudna, J. A. (2015) ‘Foreign DNA capture during CRISPR-Cas adaptive immunity.’, *Nature*, 527(7579), pp. 535–8. doi: 10.1038/nature15760.

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V, Davies, C. W. and Doudna, J. A.

(2014) ‘Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity.’, *Nature structural & molecular biology*, 21(6), pp. 528–34. doi: 10.1038/nsmb.2820.

Nuñez, J. K., Lee, A. S. Y., Engelman, A. and Doudna, J. A. (2015) ‘Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity.’, *Nature*, 519(7542), pp. 193–8. doi: 10.1038/nature14237.

O’Connell, M. R., Oakes, B. L., Sternberg, S. H., East-Seletsky, A., Kaplan, M. and Doudna, J. A. (2014) ‘Programmable RNA recognition and cleavage by CRISPR/Cas9.’, *Nature*, 516(7530), pp. 263–6. doi: 10.1038/nature13769.

van der Oost, J., Westra, E. R., Jackson, R. N. and Wiedenheft, B. (2014) ‘Unravelling the structural and mechanistic basis of CRISPR-Cas systems.’, *Nature reviews. Microbiology*, 12(7), pp. 479–92. doi: 10.1038/nrmicro3279.

Osawa, T., Inanaga, H., Sato, C. and Numata, T. (2015) ‘Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog.’, *Molecular cell*, 58(3), pp. 418–30. doi: 10.1016/j.molcel.2015.03.018.

Pasternak, C., Dulermo, R., Ton-Hoang, B., Debuchy, R., Siguier, P., Coste, G., Chandler, M. and Sommer, S. (2013) ‘ISDra2 transposition in Deinococcus radiodurans is downregulated by TnpB.’, *Molecular microbiology*, 88(2), pp. 443–55. doi: 10.1111/mmi.12194.

Peng, W., Feng, M., Feng, X., Liang, Y. X. and She, Q. (2015) ‘An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference.’, *Nucleic acids research*, 43(1), pp. 406–17. doi: 10.1093/nar/gku1302.

Platt, R. J., Chen, S., Zhou, Y., Yim, M. J., Swiech, L., Kempton, H. R., Dahlman, J. E., Parnas, O., Eisenhaure, T. M., Jovanovic, M., Graham, D. B., Jhunjhunwala, S., Heidenreich, M., Xavier, R. J., Langer, R., Anderson, D. G., Hacohen, N., Regev, A., Feng, G., Sharp, P. A. and Zhang, F. (2014) ‘CRISPR-Cas9 knockin mice for genome editing and cancer modeling.’, *Cell*, 159(2), pp. 440–55. doi: 10.1016/j.cell.2014.09.014.

Pourcel, C., Salvignol, G. and Vergnaud, G. (2005) ‘CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.’, *Microbiology (Reading, England)*, 151(Pt 3), pp. 653–63. doi: 10.1099/mic.0.27437-0.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010) ‘FastTree 2--approximately maximum-likelihood trees for large alignments.’, *PloS one*, 5(3), p. e9490. doi: 10.1371/journal.pone.0009490.

Qi, L., Haurwitz, R. E., Shao, W., Doudna, J. A. and Arkin, A. P. (2012) ‘RNA processing enables predictable programming of gene expression.’, *Nature biotechnology*, 30(10), pp. 1002–6. doi: 10.1038/nbt.2355.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P. and Lim, W. A. (2013) ‘Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression.’, *Cell*, 152(5), pp. 1173–83. doi: 10.1016/j.cell.2013.02.022.

Qin, W., Dion, S. L., Kutny, P. M., Zhang, Y., Cheng, A. W., Jillette, N. L., Malhotra, A., Geurts, A. M., Chen, Y.-G. and Wang, H. (2015) ‘Efficient CRISPR/Cas9-Mediated Genome Editing in Mice by Zygote Electroporation of Nuclease.’, *Genetics*, 200(2), pp. 423–30. doi: 10.1534/genetics.115.176594.

Ramanan, V., Shlomai, A., Cox, D. B. T., Schwartz, R. E., Michailidis, E., Bhatta, A., Scott, D. A., Zhang, F., Rice, C. M. and Bhatia, S. N. (2015) ‘CRISPR/Cas9 cleavage of viral DNA efficiently suppresses hepatitis B virus.’, *Scientific reports*, 5, p. 10833. doi: 10.1038/srep10833.

Ran, F. A., Cong, L., Yan, W. X., Scott, D. A., Gootenberg, J. S., Kriz, A. J., Zetsche, B., Shalem, O., Wu, X., Makarova, K. S., Koonin, E. V, Sharp, P. A. and Zhang, F. (2015) ‘In vivo genome editing using Staphylococcus aureus Cas9.’, *Nature*, 520(7546), pp. 186–91. doi: 10.1038/nature14299.

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y. and Zhang, F. (2013) ‘Double nicking by RNA-guided

CRISPR Cas9 for enhanced genome editing specificity.', *Cell*, 154(6), pp. 1380–9. doi: 10.1016/j.cell.2013.08.021.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A. and Zhang, F. (2013) 'Genome engineering using the CRISPR-Cas9 system.', *Nature protocols*, 8(11), pp. 2281–308. doi: 10.1038/nprot.2013.143.

Reardon, S. (2016) 'First CRISPR clinical trial gets green light from US panel', *Nature*. doi: 10.1038/nature.2016.20137.

Redding, S., Sternberg, S. H., Marshall, M., Gibb, B., Bhat, P., Guegler, C. K., Wiedenheft, B., Doudna, J. A. and Greene, E. C. (2015) 'Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System.', *Cell*, 163(4), pp. 854–65. doi: 10.1016/j.cell.2015.10.003.

Richter, C., Dy, R. L., McKenzie, R. E., Watson, B. N. J., Taylor, C., Chang, J. T., McNeil, M. B., Staals, R. H. J. and Fineran, P. C. (2014) 'Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer.', *Nucleic acids research*, 42(13), pp. 8516–26. doi: 10.1093/nar/gku527.

Rollins, M. F., Schuman, J. T., Paulus, K., Bukhari, H. S. T. and Wiedenheft, B. (2015) 'Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa.', *Nucleic acids research*, 43(4), pp. 2216–22. doi: 10.1093/nar/gkv094.

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C. V, Spagnolo, L. and White, M. F. (2013) 'Structure of the CRISPR interference complex CSM reveals key similarities with cascade.', *Molecular cell*, 52(1), pp. 124–34. doi: 10.1016/j.molcel.2013.08.020.

Rutkauskas, M., Sinkunas, T., Songailiene, I., Tikhomirova, M. S., Siksnys, V. and Seidel, R. (2015) 'Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection.', *Cell reports*. doi: 10.1016/j.celrep.2015.01.067.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A. and Marraffini, L. A. (2015) ‘Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity.’, *Cell*, 161(5), pp. 1164–74. doi: 10.1016/j.cell.2015.04.027.

Samson, J. E., Magadán, A. H., Sabri, M. and Moineau, S. (2013) ‘Revenge of the phages: defeating bacterial defences.’, *Nature reviews. Microbiology*, 11(10), pp. 675–87. doi: 10.1038/nrmicro3096.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and Siksnys, V. (2011) ‘The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli.’, *Nucleic acids research*, 39(21), pp. 9275–82. doi: 10.1093/nar/gkr606.

Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. and Severinov, K. (2013) ‘High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli.’, *RNA biology*, 10(5), pp. 716–25. doi: 10.4161/rna.24325.

Schunder, E., Rydzewski, K., Grunow, R. and Heuner, K. (2013) ‘First indication for a functional CRISPR/Cas system in Francisella tularensis.’, *International journal of medical microbiology : IJMM*, 303(2), pp. 51–60. doi: 10.1016/j.ijmm.2012.11.004.

Semenova, E., Jore, M. M., Datsenko, K. A., Semenova, A., Westra, E. R., Wanner, B., van der Oost, J., Brouns, S. J. J. and Severinov, K. (2011) ‘Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence.’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(25), pp. 10098–103. doi: 10.1073/pnas.1104144108.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G. and Zhang, F. (2014) ‘Genome-scale CRISPR-Cas9 knockout screening in human cells.’, *Science (New York, N.Y.)*, 343(6166), pp. 84–7. doi: 10.1126/science.1247005.

Sheppard, N. F., Glover, C. V. C., Terns, R. M. and Terns, M. P. (2016) ‘The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease.’, *RNA (New York, N.Y.)*, 22(2), pp. 216–24. doi: 10.1261/rna.039842.113.

Shmakov, S., Abudayyeh, O. O., Makarova, K. S., Wolf, Y. I., Gootenberg, J. S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., Zhang, F. and Koonin, E. V (2015) ‘Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems.’, *Molecular cell*, 60(3), pp. 385–97. doi: 10.1016/j.molcel.2015.10.008.

Shmakov, S., Savitskaya, E., Semenova, E., Logacheva, M. D., Datsenko, K. A. and Severinov, K. (2014) ‘Pervasive generation of oppositely oriented spacers during CRISPR adaptation.’, *Nucleic acids research*, 42(9), pp. 5907–16. doi: 10.1093/nar/gku226.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. and Siksnys, V. (2011) ‘Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system.’, *The EMBO journal*, 30(7), pp. 1335–42. doi: 10.1038/emboj.2011.41.

Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X. and Zhang, F. (2016) ‘Rationally engineered Cas9 nucleases with improved specificity.’, *Science (New York, N.Y.)*, 351(6268), pp. 84–8. doi: 10.1126/science.aad5227.

Smargon, A. A., Cox, D. B. T., Pyzocha, N. K., Zheng, K., Slaymaker, I. M., Gootenberg, J. S., Abudayyeh, O. A., Essletzbichler, P., Shmakov, S., Makarova, K. S., Koonin, E. V and Zhang, F. (2017) ‘Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28.’, *Molecular cell*, 65(4), p. 618–630.e7. doi: 10.1016/j.molcel.2016.12.023.

Söding, J. (2005) ‘Protein homology detection by HMM-HMM comparison.’, *Bioinformatics (Oxford, England)*, 21(7), pp. 951–60. doi: 10.1093/bioinformatics/bti125.

Söding, J., Remmert, M., Biegert, A. and Lupas, A. N. (2006) ‘HHsenser: exhaustive transitive profile search using HMM-HMM comparison.’, *Nucleic acids research*, 34(Web Server issue), pp. W374-8. doi: 10.1093/nar/gkl195.

Spilman, M., Cocozaki, A., Hale, C., Shao, Y., Ramia, N., Terns, R., Terns, M., Li, H. and Stagg, S. (2013) ‘Structure of an RNA silencing complex of the CRISPR-Cas immune system.’, *Molecular cell*, 52(1), pp. 146–52. doi: 10.1016/j.molcel.2013.09.008.

Staals, R. H. J., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D. W., van Duijn, E., Barendregt, A., Vlot, M., Koehorst, J. J., Sakamoto, K., Masuda, A., Dohmae, N., Schaap, P. J., Doudna, J. A., Heck, A. J. R., Yonekura, K., van der Oost, J. and Shinkai, A. (2013) ‘Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus.’, *Molecular cell*, 52(1), pp. 135–45. doi: 10.1016/j.molcel.2013.09.013.

Staals, R. H. J., Zhu, Y., Taylor, D. W., Kornfeld, J. E., Sharma, K., Barendregt, A., Koehorst, J. J., Vlot, M., Neupane, N., Varossieau, K., Sakamoto, K., Suzuki, T., Dohmae, N., Yokoyama, S., Schaap, P. J., Urlaub, H., Heck, A. J. R., Nogales, E., Doudna, J. A., Shinkai, A. and van der Oost, J. (2014) ‘RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus.’, *Molecular cell*, 56(4), pp. 518–30. doi: 10.1016/j.molcel.2014.10.005.

Sternberg, S. H., LaFrance, B., Kaplan, M. and Doudna, J. A. (2015) ‘Conformational control of DNA target cleavage by CRISPR-Cas9.’, *Nature*, 527(7576), pp. 110–3. doi: 10.1038/nature15544.

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. and Doudna, J. A. (2014) ‘DNA interrogation by the CRISPR RNA-guided endonuclease Cas9.’, *Nature*, 507(7490), pp. 62–7. doi: 10.1038/nature13011.

Swarts, D. C., Mosterd, C., van Passel, M. W. J. and Brouns, S. J. J. (2012) ‘CRISPR interference directs strand specific spacer acquisition.’, *PloS one*, 7(4), p. e35888. doi: 10.1371/journal.pone.0035888.

Takeuchi, N., Wolf, Y. I., Makarova, K. S. and Koonin, E. V (2012) ‘Nature and intensity of selection pressure on CRISPR-associated genes.’, *Journal of bacteriology*, 194(5), pp. 1216–25. doi: 10.1128/JB.06521-11.

Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, Č., Nwokeoji, A. O., Dickman, M. J., Horvath, P. and Siksnys, V. (2014) ‘Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus.’, *Molecular cell*, 56(4), pp. 506–17. doi: 10.1016/j.molcel.2014.09.027.

Taylor, D. W., Zhu, Y., Staals, R. H. J., Kornfeld, J. E., Shinkai, A., van der Oost, J., Nogales,

E. and Doudna, J. A. (2015) ‘Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning.’, *Science (New York, N.Y.)*, 348(6234), pp. 581–5. doi: 10.1126/science.aaa4535.

Thakore, P. I., Black, J. B., Hilton, I. B. and Gersbach, C. A. (2016) ‘Editing the epigenome: technologies for programmable transcription and epigenetic modulation.’, *Nature methods*, 13(2), pp. 127–37. doi: 10.1038/nmeth.3733.

Vestergaard, G., Garrett, R. A. and Shah, S. A. (2014) ‘CRISPR adaptive immune systems of Archaea.’, *RNA biology*, 11(2), pp. 156–67. doi: 10.4161/rna.27990.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M. and Wang, Y. (2015) ‘Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems.’, *Cell*, 163(4), pp. 840–53. doi: 10.1016/j.cell.2015.10.008.

Wang, T., Wei, J. J., Sabatini, D. M. and Lander, E. S. (2014) ‘Genetic screens in human cells using the CRISPR-Cas9 system.’, *Science (New York, N.Y.)*, 343(6166), pp. 80–4. doi: 10.1126/science.1246981.

Wei, Y., Terns, R. M. and Terns, M. P. (2015) ‘Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation.’, *Genes & development*, 29(4), pp. 356–61. doi: 10.1101/gad.257550.114.

Westra, E. R., Buckling, A. and Fineran, P. C. (2014) ‘CRISPR-Cas systems: beyond adaptive immunity.’, *Nature reviews. Microbiology*, 12(5), pp. 317–26. doi: 10.1038/nrmicro3241.

Wheeler, D. and Bhagwat, M. (2007) ‘BLAST QuickStart: example-driven web-based BLAST tutorial. (BLASTCLUST)’, *Methods in molecular biology (Clifton, N.J.)*, 395, pp. 149–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17993672.

Wiedenheft, B., van Duijn, E., Bultema, J. B., Bultema, J., Waghmare, S. P., Waghmare, S., Zhou, K., Barendregt, A., Westphal, W., Heck, A. J. R., Heck, A., Boekema, E. J., Boekema, E., Dickman, M. J., Dickman, M. and Doudna, J. A. (2011) ‘RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions.’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(25), pp. 10092–7. doi: 10.1073/pnas.1102716108.

Wiedenheft, B., Lander, G. C., Zhou, K., Jore, M. M., Brouns, S. J. J., van der Oost, J., Doudna, J. A. and Nogales, E. (2011) 'Structures of the RNA-guided surveillance complex from a bacterial immune system.', *Nature*, 477(7365), pp. 486–9. doi: 10.1038/nature10402.

Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I. M., Li, Y., Fedorova, I., Nakane, T., Makarova, K. S., Koonin, E. V, Ishitani, R., Zhang, F. and Nureki, O. (2016) 'Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA.', *Cell*, 165(4), pp. 949–62. doi: 10.1016/j.cell.2016.04.003.

Yang, H., Gao, P., Rajashankar, K. R. and Patel, D. J. (2016) 'PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease.', *Cell*, 167(7), p. 1814–1828.e12. doi: 10.1016/j.cell.2016.11.053.

Yang, Z. (2007) 'PAML 4: phylogenetic analysis by maximum likelihood.', *Molecular biology and evolution*, 24(8), pp. 1586–91. doi: 10.1093/molbev/msm088.

Yosef, I., Goren, M. G. and Qimron, U. (2012) 'Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli.', *Nucleic acids research*, 40(12), pp. 5569–76. doi: 10.1093/nar/gks216.

Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I. and Koonin, E. V (2008) 'The deep archaeal roots of eukaryotes.', *Molecular biology and evolution*, 25(8), pp. 1619–30. doi: 10.1093/molbev/msn108.

Zebec, Z., Manica, A., Zhang, J., White, M. F. and Schleper, C. (2014) 'CRISPR-mediated targeted mRNA degradation in the archaeon Sulfolobus solfataricus.', *Nucleic acids research*, 42(8), pp. 5280–8. doi: 10.1093/nar/gku161.

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A., Koonin, E. V and Zhang, F. (2015) 'Cpf1 is a single RNA-guided endonuclease of a Class 2 CRISPR-Cas system.', *Cell*, 163(3), pp. 759–71. doi: 10.1016/j.cell.2015.09.038.

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E. M., Winblad, N., Choudhury, S. R., Abudayyeh, O. O., Gootenberg, J. S., Wu, W. Y., Scott, D. A., Severinov, K., van der Oost, J. and Zhang, F. (2017) 'Erratum: Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array.', *Nature biotechnology*, 35(2), p. 178. doi: 10.1038/nbt0217-178b.

Zhang, Y., Heidrich, N., Ampattu, B. J., Gunderson, C. W., Seifert, H. S., Schoen, C., Vogel, J. and Sontheimer, E. J. (2013) 'Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis.', *Molecular cell*, 50(4), pp. 488–503. doi: 10.1016/j.molcel.2013.05.001.

Zhang, Y., Rajan, R., Seifert, H. S., Mondragón, A. and Sontheimer, E. J. (2015) 'DNase H Activity of Neisseria meningitidis Cas9.', *Molecular cell*, 60(2), pp. 242–55. doi: 10.1016/j.molcel.2015.09.020.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (no date) 'A greedy algorithm for aligning DNA sequences.', *Journal of computational biology : a journal of computational molecular cell biology*, 7(1–2), pp. 203–14. doi: 10.1089/10665270050081478.

Zhu, W., Lomsadze, A. and Borodovsky, M. (2010) 'Ab initio gene identification in metagenomic sequences.', *Nucleic acids research*, 38(12), p. e132. doi: 10.1093/nar/gkq275.

Supplementary information

All ad hoc software developed in this project is unavailable due to dependencies on NCBI infrastructure used to optimize computational efforts.

Links for supplementary information are given due to big size or inappropriate format (for text document) of supplementary files. Supplementary information placed on NCBI FTP site.

Supplementary information S2

Files located on following site:
ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/CRISPRclass2NRM/

Description for files located on FTP site:

Supplementary information S2 (box, part a) (MS Excel):

Pipeline output for all protein families associated with CRISPR arrays. Protein clusters ±10 kb vicinity of CRISPR arrays, their annotation (if any) and representative sequences. All clusters are sorted by the relative frequency of genes in the CRISPR loci.

Supplementary information S2 (box, part b) (MS Excel):

Pipeline output for all protein families associated with CRISPR arrays. Protein clusters ±10 kb vicinity of CRISPR arrays, their annotation (if any) and representative sequences. All clusters are sorted by the relative frequency of genes in the CRISPR loci.

Supplementary information S2 (box, part c):

Class 2 loci. For each Class 2 effector gene the surrounding genomic locus is given. Protein-coding genes and CRISPR arrays are shown. Genes annotated in GenBank are identified with GenBank locus tags; genes annotated de novo are identified by contig IDs and gene numbers.

Supplementary information S2 (box, part d):

TnpB family FastTree in the newick format. Complete tree used for the Figure 3A is provided. Sequences are denoted by a local GI number, species name and those that are located next to CRISPR array marked by “CRISPR” prefix. More details on the sequences could be found in supplementary information S2 (box, part g).

Supplementary information S2 (box, part e) (MS Excel):

CRISPR array spacers. Unique spacers were retrieved from all CRISPR arrays in supplementary information S2 (box, part a). Similarity searches were performed using MEGABLAST (see Figure 7) Hits are annotated as follows: Spacer ID (column 1) includes contig ID where this spacer was found, CRISPR coordinates and position number of the spacer in CRISPR array separated by underscore. For the hits (column 2,3,4) coordinates of best spacer hit (contig ID, hit start position, hit end position) are shown. Brief information about hit identity and annotation is provided (column 5) as follows: “Intergenic” stands for hits that do not target ORFs or viruses. “ORF” for hits into ORF that do not have good hits in virus contigs, “Virus” for hits that targets viruses.

Supplementary information S2 (box, part f) (MS Excel):

CRISPR-Cas systems and CRISPR arrays in the genomes with Type V-U system. For each complete genome that contains at least one V-U representative all CRISPR-Cas loci, CRISPR arrays and sequences or repeats are provided. Loci are annotated according to the CRISPR-Cas system classification. VU genes are indicated.

Supplementary information S2 (box, part g) (MS Excel):

HEPN domain proteins in the CRISPR vicinity. All protein containing HEPN domains from the known families located in the vicinity of CRISPR arrays are listed. The following information is provided: gene ID and location, HEPN family, CRISPR-Cas system type (if any), sequence cluster ID.

Supplementary information S2 (box, part h) (MS Excel):

Sequences used for analysis of type V systems and TnpB family. For each sequence that was used for reconstruction of the phylogenetic tree (Figure 2 and Figure 7) and profile dendrogram (supplementary information S3) the following information is provided: TnpB sequence ID and its coordinated in the genome, cluster ID, subfamily description, Genome ID and species name and association with CRISPR array (if any).

Supplementary information S1: Multiple alignment of C2c1 protein family

Supplementary information S4: Multiple alignment of C2c3 protein family

Supplementary information S5: Multiple alignment of C2c2 protein family

Merged into one supplementary file, see S4, S5, S6 figures in:

ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Shmakov/SupplementS1_4_5.pdf

Supplementary information S3: Multiple alignment of representatives from five V-U families.

ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Shmakov/SupplementS3.pdf

Supplementary information S6 (figure): Membrane proteins associated with Cas13b genes

ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Shmakov/SupplementS6.pdf

Acknowledgements

I would like to thank my supervisors: Konstantin V. Severinov and Eugene V. Koonin who started this project and gave a most of knowledge and experience, who arranged the vital collaboration with experimental labs.

I would like to thank Kira Makarova and Yuri I. Wolf who are great and knowledgeable mentors for me in this project and who made a major impact in this work.

Would like to thank our collaborators in Broad Institute and Rutgers who did very productive and effective verification of our predictions in very short time frame.

As well I would like to express my gratitude to Skoltech staff who showed that administrative side can be very helpful and supporting, and would like to thank NCBI staff for providing and supporting computational work space for this huge project.

Would like to thank all Skoltech and NIH colleagues for their support and friendship that helping me to keep go forward.