Computational Characterisation of Genomic CRISPR-Cas systems in Archaea and Bacteria

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt dem Rat

der Technischen Fakultät
der Albert-Ludwigs-Universität Freiburg

2017

von

M.Sc. Informatik
Omer Salem Alkhnbashi

Dekan
Prof. Dr. Oliver Paul
Microsystems Materials
Department of Microsystems Engineering
University of Freiburg

Vorsitz
Prof. Dr. Matthias Teschner
Computer Graphics
Department of Computer Science
University of Freiburg

Beisitz
Prof. Dr. Gerald Urban
Laboratory for Sensors
Department of Microsystems Engineering
University of Freiburg

Gutachter
Prof. Dr. Rolf Backofen
Bioinformatics
Department of Computer Science
University of Freiburg

Gutachter
Prof. Dr. Wolfgang R. Hess
Genetics & Exp. Bioinformatics
Institute for Biology III
University of Freiburg

Datum der Promotion:
24.03.2017

Acknowledgements

After an intensive period of time, today is the day: writing this note of thanks is the final touch on my PhD thesis. It has been a period of intense learning for me, not only in the scientific arena, but also on a personal level as well. Writing this thesis has had a big impact on me, and I would like to reflect on the people who have supported and helped me so much throughout this period.

First and foremost I would like to express my deepest gratitude to Prof. Dr. Rolf Backofen for his tremendous support and for all the opportunities he granted me to conduct my research under his supervision. I also would like to express my gratitude to Prof. Dr. Wolfgang R. Hess for his interest in my thesis and his kindness to review it, and to Prof. Dr. Matthias Teschner and Prof. Dr. Gerald Urban for being a part of my PhD examination committee.

I also would like to extend my gratitude to Dominic Rose and Sita J. Saunders for helping me in the beginning of my research work. Their expert guidance to understand the bioinformatics techniques were invaluable.

Moreover, I would like to thank my current and former lab members for their various help and for providing an excellent working environment. I especially thank Christina Otto, Kousik Kundu and Robert Kleinkauf for their generous help, and Monika Degen-Hellmuth for her excellent support in various administration stuffs.

I am also very thankful to Anika Erxleben, Martin Mann and Sita J. Saunders for proofreading this thesis, and I owe special thanks to Florian Eggenhofer for his contribution in writing the "Zusammenfassung" part of this thesis.

Last but not least, I would like to express my gratitude to my parents for always lending me a sympathetic ear and for their wise counsel, and to my wife who has always been there for me. I would like to say thank you father, mother, and wife. I would not have finished this journey without your endless love and support.

Contents

Abstract

Recently, it has been discovered that archaeal and bacterial organisms harbour an adaptive immune system against invading viruses and plasmids called CRISPR-Cas. CRISPR-Cas stands for (Clustered Regularly Interspaced Short Palindromic Repeats) and its associated (Cas) proteins. The CRISPR-Cas system acts via three major phases: (1) *adaptation*, in which a new spacer is acquired from invading DNA into the CRISPR array; (2) *biogenesis of crRNAs*, where generally the complete CRISPR array is transcribed into a long precursor of CRISPR RNA (pre-crRNA) that is subsequently processed into small mature crRNAs and; (3) *interference*, where crRNAs, together with Cas proteins, form an interference complex, which forms to base pairs with matching sequences in foreign DNA and subsequently cleaves the foreign DNA.

This thesis addresses the major computational challenges in the field of CRISPR-Cas research, namely, the classification and structural analysis of these systems. In the first part, we provide the first automated comprehensive classification of CRISPR repeats based on sequence and structural similarity. In this work, we compiled the largest dataset of CRISPR repeats to date and performed comprehensive independent clustering analysis to determine conserved sequence families; potential structure motifs for endoribonucleases; and evolutionary relationships. Our methods are well-suited for identifying many characteristics of CRISPR-Cas systems, e.g. cleavage sites, patterns of RNA structure motifs and sequence conservation, the link between evolution of the CRISPR array and associated Cas subtypes, and the horizontal transfer of such systems. Furthermore, we developed a web server called *CRISPRmap*, which provides both a quick and detailed insight into repeat conservation and diversity of archaeal and bacterial systems.

In the second part, we present the novel method, *CRISPRstrand*, that accurately predicts the crRNA-encoding strand of CRISPR loci by predicting the correct orientation of repeats based on an advanced machine learning approach. Both the repeat sequence and mutation information were encoded and processed by an efficient graph kernel to learn higher-order correlations. The model was trained and tested on curated data comprising 45,000 CRISPR arrays and yielded a remarkable performance of 0.95 AUC ROC (area under the curve of the receiver operator characteristic). In addition, we show that accurate orientation information greatly improved detection of conserved repeat sequence families and structure motifs.

In the third part of this thesis, we introduce the novel method, *CRISPRleader*, to successfully detect leader sequences by focusing on the consensus repeat of the adjacent CRISPR array and weak upstream conservation signals. The *CRISPRleader* tool was applied to the analysis of a comprehensive genomic database and identified several characteristic properties of leader sequences specific to archaea and bacteria, ranging from distinctive sizes to preferential indel localisation. *CRISPRleader* provides a full annotation of the CRISPR array, its strand orientation, as well as the conserved core leader boundary which can be uploaded a genome browser of choice. In addition, it outputs reader-friendly HTML pages for conserved leader clusters from our database.

In the fourth part, we present a very comprehensive CRISPR-Cas classification which classifies more than 4,000 archaeal and bacterial CRISPR-Cas systems into two classes (Class I-II), five Types and sixteen subtypes. Our classification is based on Cas protein similarities involved in the interference phase. We construct the first automated classifier using prior information on the association between sequence PSSMs and CRISPR-Cas loci and the corresponding classification of the effector modules. The classifier achieved 0.998 accuracy, which means that only 4 loci out of 1,942 were assigned incorrect subtypes. The classifier method is accurate and fast, capable of analysing more than 20,000 Cas proteins in five minutes.

In the final part of this thesis, we introduce evidence, which shows that the contextual sequence surrounding a CRISPR repeat instance can lead to structure formations that inhibit stable folding of the hairpin motif. Structure accuracy calculations of the hairpin motif explained the vast majority of analysed cleavage reactions making this a good measure of structure stability and for predicting successful cleavage events. The influence of surrounding sequences might partially explain variations in crRNA abundances and should be considered when designing artificial CRISPR arrays for applications. Furthermore, we computed the average base-pair probabilities of repeats that are cleaved and not cleaved with the surrounding spacer sequences. The results show that there are many more base pairs in the surrounding context among uncleaved fragments which lead to form stable structures with their surrounding context, whereas cleaved fragments have fewer base pairs with their the surrounding context.

In summary, this thesis provides novel and accurate methods for computational characterisation and analysis of CRISPR-Cas systems that help to understand the vast variety of CRISPR-Cas systems in nature.

Zusammenfassung

Archea und Bakterien besitzen ein, kürzlich entdecktes, adaptives Immunsystem gegen eindringende Viren und Plasmide, welches als CRISPR-Cas System bezeichnet wird. CRISPR-Cas steht abgekürzt für Clustered Regularly Interspaced Short Palindramoic Repeats und die assozierten Cas-Proteine. Der Wirkungsmechanismus des CRISPR-Cas Systems kann in 3 Phasen unterteilt werden:(1) Adaption, in welcher ein neuer Spacer von eingedrungener DNA in das CRISPR Array aufgenommen wird(2) Biogenese von crRNAs, wobei im Allgemeinen das gesamte CRISPR Array in eine lange Vorstufe der CRISPR RNA (pre-crRNA) transkribiert und anschließend in kürzere gereifte crRNAs prozessiert wird.(3) Interferenz, bei der crRNAs, gemeinsam mit Cas Proteinen, einen Interferenz-Komplex bilden, welcher mit passenden Sequenzen aus der Fremd-DNA basenpaart und diese danach spaltet.

Diese Arbeit thematisiert bedeutende bioinformatische Herausforderungen im Feld der CRISPR-Cas Forschung, konkret, die Klassifikation und strukturelle Analyse dieser Systeme. Im ersten Teil stellen wir die erste umfassende, automatisierte Klassifizierung von CRISPR Repeats, basierend auf Sequenz- und Struktursimilarität, vor. Im Zuge dessen haben wir den, bis dahin, größten Datensatz von CRISPR Repeats zusammengestellt und damit eine unabhängige Analyse der Gruppierung durchgeführt, um konservierte Familiensequenzen, z.B. potentielle Strukturmotive für Endonukleasen und evolutionäre Zusammenhänge zu bestimmen. Unsere Methoden sind gut geeignet um viele Charakteristika von CRISPR-Cas Systemen, z.B., Schnittstellen, Muster von RNA Strukturmotiven und Sequenzkonservierung, die Verbindung der Evolution des CRISPR Arrays mit seinen assozierten Cas-Subtypen und der horizontale Gentransfer von solchen Systemen zu analysieren. Darüberhinaus haben wir einen Webserver mit dem Namen *CRISPRmap* entwickelt, welcher sowohl einen schnellen, als auch detailierten Einblick in Repeat Konservierung und Diversität in Bakterien und Archea gibt.

Im zweiten Teil präsentieren wir eine neue Methode, *CRISPRstrand*, die akkurat den crRNS-codierenden Strang von CRISPR Loci bestimmt. Die Vorhersage der korrekten Orientierung der Repeats, basiert auf einem Ansatz für maschinelles Lernen. Sowohl die Repeat Sequenz- und Mutations-Information wurden von einem effizienten Graph-Kernel kodiert und prozessiert um Korrelationen höherer Ordnung zu erlernen. Das Model wurde mit kuratierten Daten, welche 45,000 CRISPR Arrays umfassten, trainiert und getestet was eine

Leistung von bemerkenswerten 0.95 AUC (Fläche unter der Kurve bei der Grenzwertoptimierungs) ROC erbrachte. Zusätzlich konnten wir zeigen, dass die bekannte Orientierung der Repeats die Identifizierung von konservierten Sequenzfamilien und Strukturmotifen erheblich vereinfacht.

Im dritten Teil der Arbeit stellen wir eine weitere neue Methode, *CRISPRleader*, vor. Diese kann erfolgreich CRISPR leader Sequenzen, durch Berücksichtigung des Consensus-Repeats der benachbarten CRISPR Arrays und schwachen vorgelagerten Konservierungssignalen, identifizieren. *CRISPRleader* wurde auf eine umfassende genomische Datenbank angewandt um zu charakterisieren, welche Eigenschaften, z.B., markante Länge oder Bevorzugung von Insertions,-Deletionsstellen für Bakterien und Archea spezifisch sind. *CRISPRleader* bietet eine vollständige Annotierung des CRISPR Arrays, seine Orientierung, als auch die Grenze des konservierten Leader-Kerns, welche bei einem Genom-Browser der Wahl hochgeladen werden kann. Des Weiteren werden übersichtliche HTML-Seiten, für die konservierten Leader Gruppen aus unserer Datenbank, generiert.

Im vierten Teil der Arbeit präsentieren wir eine umfassende CRISPR-Cas Klassifikation für mehr als 4000 Systeme aus Archea und Bakterien, welche in zwei Klassen (Class I-II), fünf Typen und sechzehn Subtypen gegliedert ist. Unsere Gruppierung basiert auf den Ähnlichkeiten von Cas-Proteinen, welche an der Interferenz-Phase beteiligt sind. Wir konstruierten den ersten automatischen Klassifizierer, welcher eine Genauigkeit von 0.998 erreichte, was nur vier inkorrekten Subtyp-Zuweisungen für 1942 Loci entspricht.

Im abschließenden Teil dieser Arbeit legen wir Beweise dafür vor, dass der Sequenzkontext, der eine CRISPR Instanz umgibt, zu Strukturen führen kann, welche die stabile Bildung eines Haarnadel-Motiv inhibieren. Strukturgenauigkeitsberechnungen für das Haarnadelmotiv erklären die überwältigende Mehrheit von analysierten Spaltungsreaktionen, was dies zu einem guten Maß für Strukturstabilität und die erfolgreiche Vorhersage von Spaltungsereignissen macht. Der Einfluss der umgebenden Sequenzen könnte teilweise Variationen in crRNA Häufigkeit erklären und sollte für das Design von synthetischen CRISPR Arrays berücksichtigt werden. Die Resultate zeigen, dass ungeschnitten Fragmente wesentlich mehr Basenpaare mit dem umgebenden Regionen bilden, welche zu stabilen Strukturen führen, als geschnittene Fragmente, welche wesentlich weniger Basenpaare ausbilden.

Zusammengefasst bietet diese Arbeit neue und genaue Methoden für die bioinformatische Charakterisierung von CRISPR-Cas Systemen, welche helfen die gewaltige Variabilität dieser Systeme in der Natur zu verstehen.

List of publications

This thesis is based on the following publications:

[P1] Sita J. Lange*, **Omer S. Alkhnbashi***, Dominic Rose*, Sebastian Will, and Rolf Backofen. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Research*, 2013.

[P2] **Omer S. Alkhnbashi***, Fabrizio Costa, Shiraz A. Shah, Roger A. Garrett, Sita J. Saunders and Rolf Backofen. CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, 2014.

[P3] **Omer S. Alkhnbashi***, Shiraz A. Shah, Roger A. Garrett, Sita J. Saunders, Fabrizio Costa and Rolf Backofen. CRISPRleader: Characterizing leader sequences of CRISPR loci. *Bioinformatics*, 2016.

[P4] Kira S. Makarova, Yuri I. Wolf, **Omer S. Alkhnbashi**, Fabrizio Costa, Shiraz A. Shah, Sita J. Saunders, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Daniel H. Haft, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Rebecca M. Terns, Michael P. Terns, Malcolm F. White, Alexander F. Yakunin, Roger A. Garrett, John van der Oost, Rolf Backofen and Eugene V. Kooni. An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*, 2015.

[P5] Viktoria Reimann*,**Omer S. Alkhnbashi***, Sita J. Saunders, Ingeborg Scholz, Stephanie Hein, Rolf Backofen and Wolfgang R. Hess. Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs. *Nucleic Acids Research*, 2016.

* Joint first authors

Further publications:

1. Lisa-Katharina Maier, **Omer S. Alkhnbashi**, Rolf Backofen and Anita Marchfelder. CRISPR AND SALTY immune response in haloarchaea. *Nucleic Acids and Molecular Biology.* 2016.

2. Vanessa Tripp, Roman Martin, **Omer S. Alkhnbashi**, Rolf Backofen and Lennart Randau. Plasticity of archaeal C/D box sRNA biogenesis. *Molecular Microbiology.* 2016.

3. Simon D. B. Cass, Karina A. Haas, Britta Stoll, **Omer S. Alkhnbashi**, Kundan Sharma, Henning Urlaub, Rolf Backofen, Anita Marchfelder, and Edward L. Bolt. The role of Cas8 in type I CRISPR interference. *Bioscience* 2015.

Chapter 1

Introduction

1.1 Motivation

Archaeal and bacterial genomes are potential targets to foreign genetic elements such as viruses. Both archaea and bacteria have developed a newly discovered immune system called CRISPR-Cas. CRISPR-Cas is an acronym for Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), in combination with associated (cas) genes [1]. The CRISPR-Cas system consists of three main parts: (1) a CRISPR array of short repeated sequences (repeats) interspaced by short variable sequences (spacers); (2) the leader sequence that is usually transcribed as one transcript and located upstream of the CRISPR array; and (3) a set of *cas* genes encoding for CRISPR-associated proteins called Cas proteins. The defence mechanism is divided into three main stages: (1) *adaptation*, which is the selection of fragments (or protospacers) of genetic material from a virus or plasmid and incorporation of their reverse complement sequences (spacers) into the host's genome at active CRISPR loci; (2) *biogenesis of crRNAs*, where the CRISPR RNA is transcribed and processed into mature crRNA containing a single spacer flanked by repeat-sequence handles at either the 5'-end or on both ends; and (3) *interference*, where invader DNA [2] or RNA [3, 4] is degraded at the respective protospacers, guided by the crRNA and a highly specific complex of Cas proteins such as Cmr [3, 4] or Cascade [5, 6]. The targeting complex differentiates real protospacers from other complementary sequences in many systems by a protospacer adjacent motif (PAM). The discovery of CRISPR-Cas systems almost 10 years ago rapidly changed our perception of the archaeal and bacterial immune systems. In 2007, it was first shown that new spacers were inserted into the CRISPR array in *Streptococcus thermophilus*, which confirmed that CRISPR-Cas is an adaptive immune system against viruses and phages. Since then, the CRISPR-Cas system has become one of very widely topic in molecular biology, synthetic biology and genetic engineering. Despite the bioinformatics studies having made significant contributions within the field since initial discovery, there are still many key components of CRISPR-Cas systems that need further investigation.

Hence, it is very important to study and understand the CRISPR-Cas systems. As seen in the history of CRISPR research, bioinformatics tools play a major role here. Examples

of problems that can be investigated computationally are the determination of the specific differences between the CRISPR-Cas systems from archaeal and bacterial sources, repeat-spacer sequences that are required for processing the mature crRNA, prediction of the transcribed strand of CRISPR arrays, determination of CRISPR leader sequences, and classification of Cas proteins.

1.2 Overview of this thesis

In this thesis, my contribution is a major one as a first author in 4 out of 5 publications [7–11]. In addition, I have done an important contribution in the fifth publication. Nowadays, teamwork is the main key for successful and accomplished research. This implies that the scientists have to collaborate and discuss together more than ever before. Therefore, I collaborated with internal and external scientists during my PhD study. Thus, I have used "we" instead of "I" throughout the thesis to reflect this collaborative nature of this work.

The rest of the thesis is organised as follows. Chapter 2, describes the biological background and computational techniques that are used throughout the thesis. It contains also a detailed description about the CRISPR-Cas adaptive immune system and an overview over machine learning approaches that will help the reader to understand the current work. In chapter 3, we provide an overview, discussion and results for the five publications that form the basis for this thesis. Chapter 4 draws a conclusion for the entire work presented in this thesis, and chapter 5 contains the statement of contributions for each publication along with the publication itself.

Chapter 2

Biological and Computational Background

This chapter provides biological and computational background related to this thesis. First, we give a brief overview about the three major types of biological macromolecules in any living cell and the flow of genetic information. Second, we discuss general concepts of adaptive immunity against bacteriophages, archaeal viruses, and plasmids in archaea and bacteria. In addition, the CRISPR-Cas system and the roles which it plays in archaea and bacteria is described. Finally, we conclude the chapter with an overview about the machine learning concepts that are used in this thesis.

2.1 The Central Dogma of Molecular Biology

The Central Dogma of Molecular Biology explains the flow of genetic information in living cells from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) to protein. Cells use a system called genetic code to translate the sequence of nucleotides of an RNA molecule into a specific sequence of amino acids called proteins. Once the genetic information has passed into protein it cannot flow back again. Generally, the flow of genetic information transfer from nucleic acid to nucleic acid, or from nucleic acid to protein is possible, but transfer protein to nucleic acid, or from protein to protein is not possible. Information means the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein [12]. The flow of genetic information as well as DNA, RNA and proteins that is needed for this thesis will be introduced in this section.

2.1.1 Deoxyribonucleic acid (DNA)

DNA is the type of molecule that carries the genetic information used for growth, development and functioning of all living organisms (excluding the RNA viruses) [13, 14]. Every living cell contains double stranded DNA, in which two strands form a double helix. Each strand consists of many individual units called nucleotides, which are stored the genetic information. One nucleotide contains one out of four different nitrogenous bases: adenine (abbreviated 'A'), thymine (abbreviated 'T'), guanine (abbreviated 'G'), and cytosine (ab-

breviated 'C'). Adenine always base-pairs with thymine to form two hydrogen bonds while guanine base-pairs with cytosine to form three hydrogen bonds.

In eukaryotes (e.g. *Human*), DNA is located in the nucleus, while in prokaryotes (*Bacteria*), DNA is located directly within the cellular cytoplasm. Both eukaryotic and prokaryotic cells use the genetic information to synthesise proteins. The regions of DNA that carry the genetic information that code for proteins and functional RNAs are called genes. Other sections of DNA that are not expressed into proteins are called intergenic. In fact, most of the DNA does not code for proteins. These regions often still have some non-coding functions that play an important role within the cell. However, DNA does not only shape the structure and function of living organisms but is also the source of inheritance. For example, in the reproduction process of the organisms a copy of their DNA is passed on to their progenitor.

The flow of genetic information within a cell is described in the "Central dogma of molecular biology" [12]. The first step of the flow of genetic information process is called transcription process (see Figure 2.1). Generally speaking, DNA is transcribed into messenger RNA (mRNA) by making a copy of a gene into an mRNA. In fact, a specific enzyme called RNA polymerase uses only one strand of the DNA as a template to produce the corresponding mRNA. During the process of transcription, RNA polymerase plays an important regulatory role in gene expression. RNA polymerase binds to specific sequences within a gene, which are called promoters and begins to unwind the double stranded DNA. Next, one of the strands will be used as the template strand to generate the mRNA. The other strand is called the nontemplate strand. RNA polymerase initiates mRNA synthesis at the start codon (e.g. AUG in *Escherichia coli*) and then moves along the gene synthesising the mRNA from 5'-end to 3'-end. Once the RNA polymerase reaches the end of the gene, mRNA elongation is terminated. Under normal conditions, the RNA polymerase is then detached from the gene and the DNA returns to its normal state (double helix). At the end, an RNA has been produced that carries the information encoded in the gene, some of these RNAs will be further processed in the second step of gene expression called translation.

2.1.2 Ribonucleic acid (RNA)

RNA is the second major macromolecule besides DNA and protein, and is also essential for any living cell. RNA is very similar to DNA. However, it is generally a single-stranded nucleic acid molecule that folds onto itself. RNA is assembled as a single series of nucleotides, which are covalently bound with each other. The RNA nucleotide bases are slightly different from DNA nucleotide bases. In RNA, the nitrogenous are: adenine (A), uracil (U), guanine (G), and cytosine (C). Some viruses store their genetic information in RNA instead of DNA [13, 14]. In the cell, RNA molecules can be divided into two major groups: (1) those that are involved in protein synthesis, i.e., messenger RNA (mRNA), ribosomal (rRNA), and transfer RNA (tRNA) ; and (2) non-coding RNAs (ncRNA), which are not involved in protein synthesis, but can regulate gene expression.

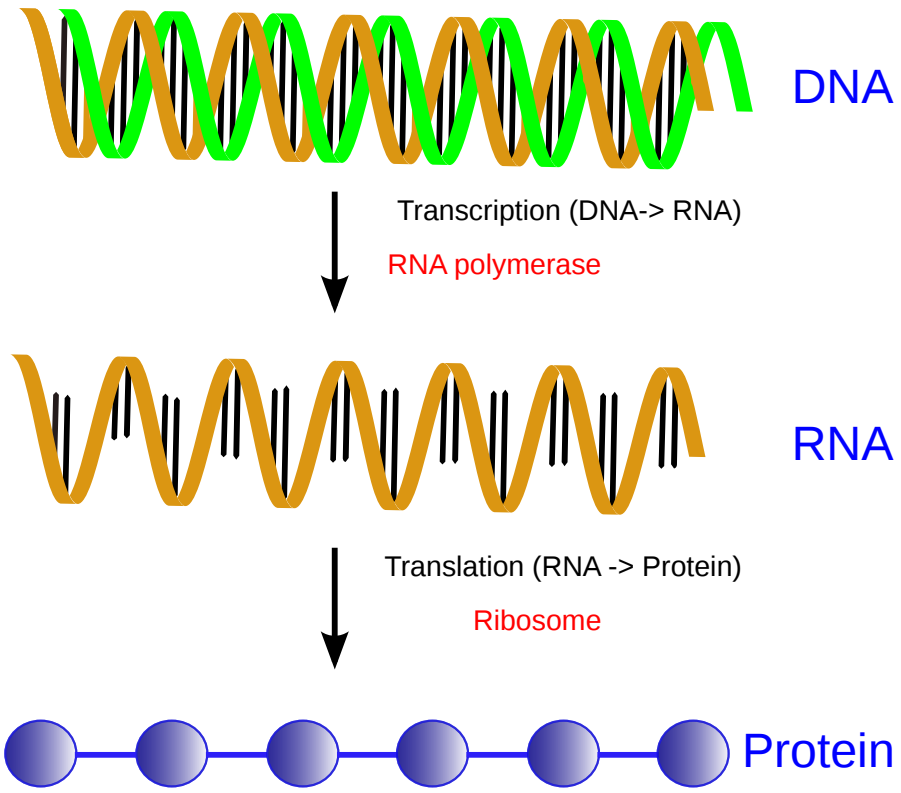


Figure 2.1: Overview over the Central Dogma. Shows the flow of genetic information process of protein synthesis. Shows the process taken DNA and making mRNA, then converting mRNA into the amino acids to make a protein.

Although the genetic information encoding proteins is stored in the DNA, DNA itself is not directly involved in protein synthesis. In fact, the cell uses an intermediate RNA molecule (mRNA) that is then used in the protein synthesis. The translation process (or protein synthesis) is the second step of the flow of genetic information (see Figure 2.1). In prokaryotic and eukaryotic cells the transcription process occurs in different locations in the cell: in the cytoplasm for prokaryotes and in the nucleus for eukaryotes. In contrast, the translation process is slightly different between prokaryotes and eukaryotes but it is located in the cytoplasm for both. In general, eukaryotic RNA needs further processing before the translation is started (known as RNA processing), while RNA in prokaryotes does not require any further processing.

As mentioned earlier, there are three different classes of RNA molecules that are involved in the translation process (mRNA, rRNA, and tRNA). During translation, a complex molecular machinery known as a ribosome is involved. The ribosome consists of two major subunits, the small ribosomal subunit and the large ribosomal subunit, which both consist of a combination of ribosomal RNA (rRNA) and proteins. During the translation process the mRNA will be used as a template that encodes for a specific protein. For protein synthesis always, three nucleotides of the mRNA (triplet) encode for one amino acid, which is known

as a codon. A transfer RNA (tRNA) with the complementary sequence to the codon (also known as anticodon) is charged with the corresponding amino acid and binds to the tRNA to transfer the amino acid to the growing peptide chain. There are many different tRNAs and each tRNA is covalently linked to a certain amino acid that corresponds to the anticodon of the tRNA. The arrangement of the nucleotides into these codons is called the reading frame. As mentioned above, RNA contains four different nucleotides (A, U, C, and G) and each codon contains three nucleotides. Therefore, all 21 amino acids are encoded by 64 (4^3) different codons. The initiation starts when the small ribosomal subunit reaches a specific group of nucleotides known as a start codon (e.g. 5'-AUG-3' in *Escherichia coli*, which encodes for the methionine amino acid). In general, the small ribosomal subunit binds to the mRNA. Once the tRNA binds to the start codon, the large ribosomal subunit joins to complete the translation initiation complex. Next, elongation starts, once the ribosome complex has been formed and the tRNA that carries the methionine in specific location in the ribosome (P-site). After that, the new tRNA that carries the next amino acid can bind into the next specific location in the ribosome (A-side). The P-site is a specialised binding pocket within the ribosome, where the formation of the peptide-bond between amino acids occurs. Once the P-site and A-side are filled with certain tRNAs, a special enzyme catalyses the formation of the peptide bond between the amino acids in the P-site and A-side. When the peptide bond has been formed, the first tRNA detaches and leaves the P-site and the ribosome complex moves three nucleotides along the mRNA in order to make space for new a tRNAs. By adding new amino acids, the peptide chain keeps growing until a stop codon is reached. As tRNAs enter and exit the ribosome, the polypeptide chain is growing. This process will repeat again and again until a stop codon is reached and the polypeptide chain is complete. The final stage of the translation process is called termination. It happens, when the small ribosomal subunit reaches a specific sequence of nucleotides known as a stop codon (either 5'- UAA, UGA, or UAG -3' in *Escherichia coli*). At the end, the two ribosomal subunits separate and release both the mRNA and the peptide chain.

Non-coding RNA (ncRNA)

Non-coding RNA is a term for RNA that does not encode a protein. For a long time, it has been assumed that the regions of the DNA that do not code for proteins have no function, and therefore it was called junk DNA regions. Nowadays it is accepted that ncRNAs play a crucial role in cell processes beyond just transmitting information. This coincides with the fact that the majority of DNA that does not code for proteins is still transcribed and regulated. There are different classes of ncRNAs in organisms. Some ncRNAs are involved in the regulation of gene expression, i.e. microRNA [15]. Another class of ncRNA is involved in generic functions in cells, such as rRNA and tRNA, which play a role in mRNA translation. snoRNAs (small nucleolar RNAs) guide chemical modifications of rRNAs and snRNAs (small nuclear RNAs), which for example are involved in splicing [16]. Further classes of ncRNAs are involved in housekeeping functions, 4.5S RNA for instance [16]. Many ncRNAs

are well characterised. However, there is a huge potential for discovering new families of RNAs and to better characterise the ones we already know.

Structure of the RNA

Non-coding RNAs, e.g. tRNAs and rRNA, realise their biological function by forming elaborate structures, while mRNAs often reside in a more open-chain conformation in order to enable the process of translation. The structure of an RNA is generally very important for its function. RNA is a single-stranded molecule and, under certain conditions in the cell, it can fold back upon itself into a complex structure. It may, therefore, have different structures that can have varying functions. Since structural properties play an important role for the functionality of RNAs, it is important to understand the RNA structure for the analysis of its function in cells. In general, RNA structure can be divided into different levels: the first level is the single strand of RNA with its nucleotide sequence, the second level is called secondary structure, which forms a set of canonical base pairs, and the finally tertiary structure that comprises the full 3D structure information of the molecule. A secondary structure implies different stem and loop structures (see Figure 2.2). The regions of consecutively paired bases are called the stems while unpaired regions form the loops. The stem-loop is the primary structure of RNA, much the same way as the alpha helix is the primary structure of the protein. RNA secondary structure can form three different loops: (1) a hairpin loop, a loop which is adjoint to one base-pairs only; (2) an internal loop, a region in which unpaired bases are found on one or both sides between base-pairs, and; (3) multi-loops, a region that is adjacent to more than three base-pairs (see Figure 2.2).

2.1.3 Proteins

Protein is the third major type of biological macromolecules, which is extremely important for any living cell. Unlike DNA and RNA, proteins are polymers of amino acids, which are linked together by amide bonds known as peptide bonds to form a polypeptide chain. Amino acids as the building blocks of proteins contain an amino group and a carboxylic acid group conserved through all types. The so-called "R-group" (R for residue) can have different properties, which makes it an important part in the amino acid because they strangely influence protein structure and function. There are twenty-one types of amino acids found in proteins, which are grouped based on properties of their characteristic side chain residue: for example, polar (e.g. serine), nonpolar (e.g. glycine), and ionic (e.g. lysine). In fact, cells use amino acids in order to build thousands of different proteins for structural and regulation functions. Proteins are described on four levels of structures: (1) primary structure, which is a linear sequence of amino acids; (2) secondary structure, which is the repeating fold of the polypeptide chain (alpha-helix and beta-sheets are the most common types of secondary structures in proteins); (3) tertiary structure, which is the spatial position of all atoms in the molecule (3D structure); and (4) quaternary structure, which is the arrangement of

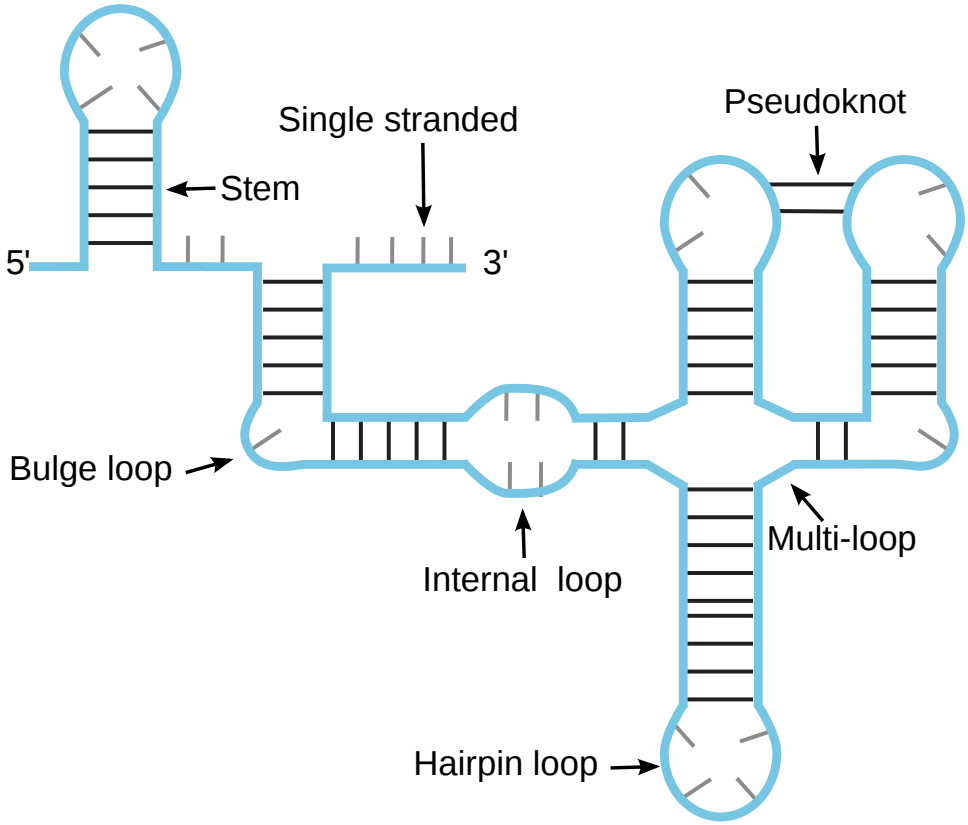


Figure 2.2: Visualisation of secondary structure elements. Blue lines in the figure indicate the backbone, black lines show base pairs and grey lines unpaired nucleotides.

the individual protein subunits, which occurs only when the protein has more than one polypeptide chain (e.g hemoglobin).

2.2 Defence Systems in Archaea and Bacteria

Both archaea and bacteria are regularly subjected to foreign genetic elements from phages and plasmids. Over the last decades, numerous archaeal viruses have been classified into new families based on their diverse morphologies [17]. Bacteriophages on the other hand, are less diverse but more abundant, far outnumbering their hosts and thus being the most ubiquitous biological entity on earth [13]. The differences between the virospheres of archaea and bacteria may stem from the major differences in life cycles and selection pressures characteristic for either host. Nevertheless, both archaea and bacteria seem to expend considerable resources in defending themselves against the viruses that infect them. To this end, both host types have developed and shared similar defence systems such as CRISPR-Cas, restriction modification, toxin-antitoxin (TA) systems and others [1].

2.3 Restriction Modification Systems

R-M (Restriction Modification) systems have been identified in 90% of sequenced archaeal and bacterial genomes [18]. The systems have two distinct enzymatic activities: (1) a restriction endonuclease that cleaves DNA at a specific recognition site, and (2) a DNA methyltransferase that methylates DNA at the same site and thus prevents cleavage by the cognate restriction enzyme [18]. This prevents foreign genetic elements from infecting the host because foreign DNA, upon entering the host, is not yet methylated, and thus selectively cleaved by the restriction enzyme. In general, R-M systems have been classified into four main classes (I-IV) on the basis of subunit composition, ATP(GTP) requirement and cleavage mechanism [19–21]. Type II systems have become the best known and used R-M system widely in molecular biology which consists of two independent proteins with enzymatic activities, a restriction endonuclease and a DNA methyltransferase [18].

2.4 CRISPR-Cas Adaptive Immune Systems

Recently, a new defence mechanism called CRISPR-Cas system (Clustered Regularly Interspaced Short Palindromic Repeats) and its associated (Cas) proteins was discovered, which protects archaea and bacteria from foreign invading genetic elements in the form of phages and plasmids [1, 22, 23]. The CRISPR-Cas adaptive immune systems consist of three main components (*CRISPR array, CRISPR leader and Cas proteins*) and act through three consecutive phases (*adaptation, expression and interference*), see Figure 2.3.

The CRISPR-Cas adaptive immune system is present in most archaea and in many bacteria, mostly on chromosomes, but also on plasmids [1, 24, 25]. CRISPR-Cas works by incorporating short foreign sequences as spacers in between the repeats of the CRISPR locus on the genome. The CRISPR locus thus acts as a library of past infections and is transcribed to produce a long pre-CRISPR RNA (pre-crRNA). Specific Cas proteins then cleave the pre-crRNA into small individual CRISPR RNAs (crRNAs) that contain only a single spacer flanked by parts of the repeats [26]. Finally, individual crRNAs are subsequently bound by a complex of Cas proteins and are used to target either foreign double stranded DNA or single stranded RNA for degradation upon any subsequent infection by a matching virus or plasmid [3, 5] (see Figure 2.3).

2.4.1 The CRISPR array

The CRISPR array (or CRISPR locus) is composed of very similar sequences called repeats that are separated by similarly-sized sequences called spacers. In general, the CRISPR array consists of a few to dozens of repeat-spacer units, while some CRISPR arrays contain up to hundreds of repeat-spacer units. The number of CRISPR arrays varies between archaeal and bacterial genomes. According to CRISPR databases [7, 8, 27] the archaon *Methanotorris igneus Kol 5* contains the highest number of CRISPR arrays (25 arrays in the chromosome)

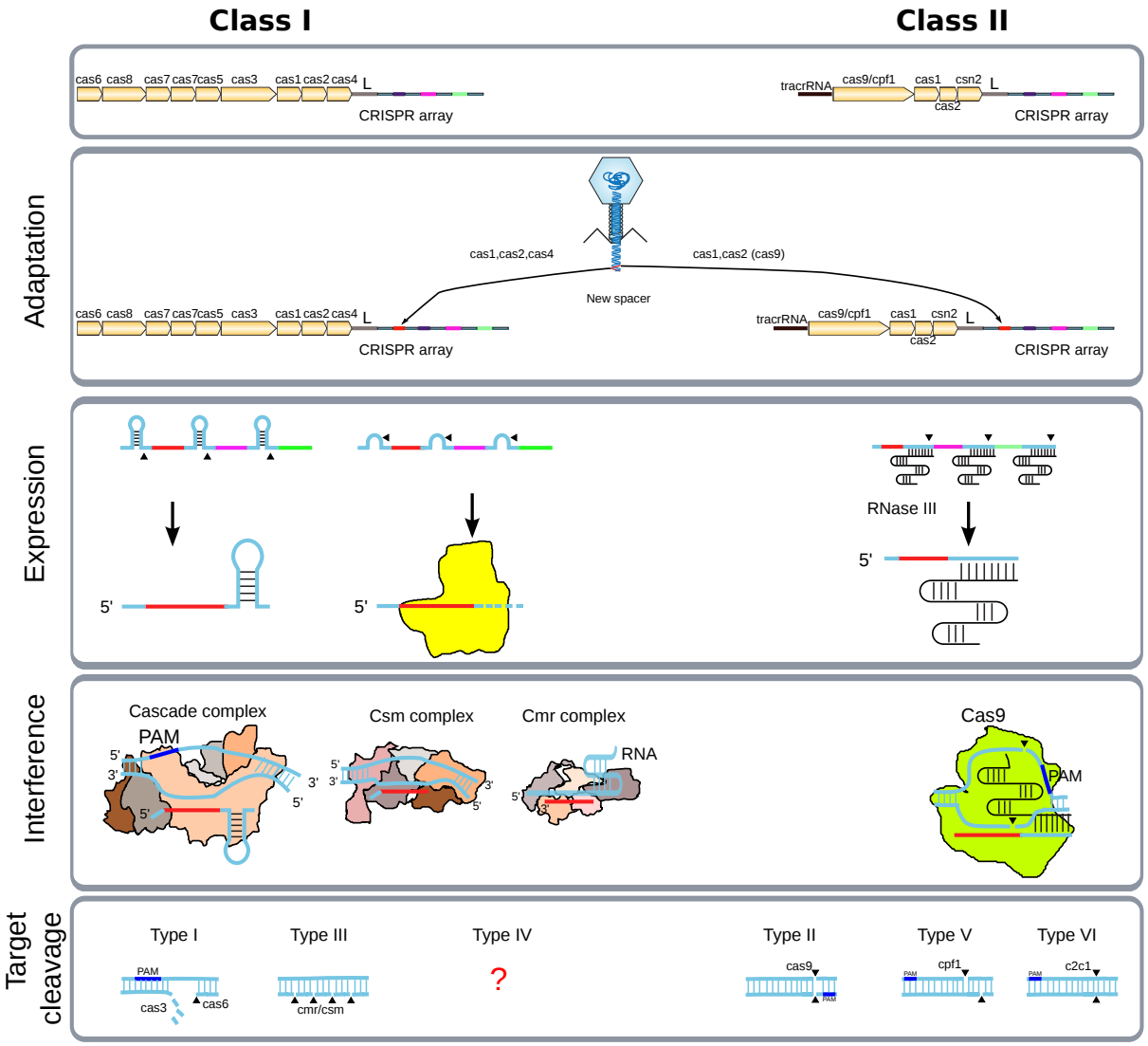


Figure 2.3: Overview of the mechanism of Class I and Class II CRISPR-Cas adaptive immune system. The CRISPR-Cas system consists of three components (1) a CRISPR array of identical repeat sequences (blue rectangles) that are separated by the so-called spacers (coloured rectangles) (2) upstream of the CRISPR array, the leader sequence (grey rectangle) that is usually transcribed as one transcript (3) and a set of CRISPR-associated *cas* genes (yellow arrows) that encode the Cascade machinery and sequence-specific nucleases. CRISPR-Cas systems are classified into tow major classes: Class I systems contain multi-protein effector complexes, whereas Class II systems contain only a big single protein [10]. In the *adaptation* stage (top), Cas1 and Cas2 (and may be Cas4 and Cas9) copy and paste invader DNA sequences as novel spacers at the leader end of the CRISPR array. During the *expression* stage (center), the Cas machinery transcribes CRISPR arrays and generates mature small crRNAs. During the *interference* stage (bottom), guide RNAs direct the Cas machinery toward complementary DNA flanked by PAM sequences and drive sequence-specific cleavage of target DNA.

while in bacteria, *Nocardiopsis alba ATCC BAA-2165* ranks the highest with 27 arrays in the chromosome. Furthermore, the bacterium *Haliangium ochraceum DSM 14365* has nine CRISPR arrays, where one of these arrays is considered to be the largest CRISPR array

(588 of repeat-spacer units) found to date [28]. Interestingly, CRISPR arrays are not located on chromosomes and plasmids only, but also in free viral genomes and in small conjugative plasmids, such as pNOB8 and pKEF from *Sulfolobus* [28–32].

In archaea and bacteria genomes, the CRISPR arrays can be automatically detected using a computational tools such as web-server based (CRISPRFinder) or command-line executable programs (CRT and PILER-CR) [33–35]. These tools have generally good performance in predicting CRISPR arrays and have different structural features and output formats. However, all these programs do not determine the strand from which crRNAs are processed. Knowledge of the correct orientation is crucial for many tasks, including the classification of CRISPR repeats, the detection of leader regions, the identification of target sites (protospacers) on invading genetic elements and the characterisation of protospacer-adjacent motifs (PAM). The orientation dilemma is one of the main subjects of this thesis. It will be discussed in chapter 3 in more detail.

CRISPR Repeats

In 1987, CRISPR repeats were discovered in *Escherichia coli K12* [36], in which five highly homologous sequences of 29 nucleotides were found that were separated by non-homologous sequences of the length 32 nucleotides. Later several studies identified CRISPR repeats in some more organisms, for example *Mycobacterium tuberculosis*, *Haloferax mediterranei*, *Haloferax volcanii*, *Anabaena* and *Streptococcus pyogenes* [37–42]. Those studies observed a few features of CRISPR repeats like partial palindromicity, local sequence motifs (GAAAN) and hairpin-structure motifs. However, the biological roles and functions remained unknown. Moreover, those studies have proposed that repeat sequences can be targets for DNA binding proteins.

Although within a single CRISPR array repeat sequences are mostly very well conserved, they may contain a few mutations, specifically at the 3' end of the array. Repeat sequences are in the range of 19-48 nucleotides [7, 8, 28]. The sequence similarities vary between repeats from different CRISPR arrays in the same or different genomes. An early comparative study of CRISPR repeat diversity yielded 12 main clusters with specific sequence characteristics; only a subset folded into characteristic hairpin structure motifs [43].

Most recently, we presented a major reevaluation of CRISPR conservation [7, 8] on a much larger data set of 4,719 CRISPR repeats, where 24 conserved repeat sequence families were identified together with a total of 18 potential structural motifs. The repeat clusters were further classified into six superclasses, some of which showed strong biases to specific CRISPR subtypes and to certain bacterial or archaeal phyla [7, 8]. For further details see section 3.2 and *Publication 1.*

CRISPR Spacers

In any CRISPR array, spacer sequences are located between the repeat sequences and are completely variable in sequence. Spacer sequences are between 20 and 72 nucleotides in length. In 2005 it was first proposed that CRISPR spacers originate from extrachromosomal elements [44, 45]. Normally spacer sequences match the foreign DNA sequences of phages and viruses, but some also match the host genome (self-targeting spacers) [45, 46].

The content of spacers in CRISPR arrays is very flexible and only distinct strains of a given species tend to have identical spacers inside their arrays. Furthermore, these conserved spacers are normally found towards the end of the array, maintaining their relative order. Closely associated strains have accumulated new unique spacers close to the leader and this accretion of new spacers produces a distinct chronological record of invading genetic elements. This feature also forms the foundation for new strain typing strategy [1]. Several studies [47–50] have shown that some CRISPR loci create dynamic structures. Deletions happen, and evidences were found for duplication of repeat-spacer elements and of recombination events happening between CRISPR loci with similar repeats. Furthermore, stimulating CRISPR systems with vector borne-protospacers having Protospacer Adjacent Motifs (PAM) and sustained under selection, produces deletions in CRISPR loci which compass the matching spacer and occasionally whole CRISPR loci [51].

2.4.2 CRISPR leaders

The second element of the CRISPR-Cas system is called the *leader* sequence. Leaders are located upstream of the CRISPR arrays at the 5'-end, and are between 40 bp in some bacteria to a few hundred bp in some hyperthermophilic archaea. Leaders are non-coding regions and contain large portions of low complexity sequence, with limited sequence conservation [9, 52]. However, several studies have shown that leaders carry the transcription initiation signal for the CRISPR array [26, 48, 53], and they contain the signals for CRISPR-Cas adaptation that involves insertion of small fragments of DNA excised from invading genetic elements [54–56].

Although most CRISPR loci carry a leader region, it has been found that a few loci lack leaders. These leaderless loci do not acquire new spacers, which complies with the lack of CRISPR-Cas adaptation signals [26, 48]. However, they do still yield processed CRISPR RNAs (crRNAs), presumably as a result of leaky transcription of upstream genes or promoters taken up randomly in spacers [57, 58]. Moreover, in some crenarchaea, leaderless CRISPR loci are also characterised by being strikingly conserved. Whereas CRISPR loci with leaders undergo spacer acquisition and are also susceptible to extensive deletions and rearrangements, leaderless CRISPR loci are both resistant to CRISPR adaptation and deletions, and this is why they are structurally invariant [26, 48, 51]. Since the leaders are AT-rich nucleotides and have low sequence conservation, there were no comprehensive studies addressing their diversity. We will tackle this problem in the next chapter (*Publication*

3 [9]).

2.4.3 Protospacer Adjacent Motifs (PAM)

The Protospacer Adjacent Motif is a short motif which consists of 2-7 nucleotides. Although the PAM motif is very small, it plays a critical role in CRISPR mediated immunity by allowing self versus non-self discrimination. In general, PAMs are located immediately adjacent or up to a couple of positions either upstream (e.g. TYPE I, III, V and VI) or downstream (e.g. TYPE II) from the target DNA (protospacer) depending on CRISPR-Cas type and the organism [52, 59]. Initial evidence for PAM motif involvement in adaptation process was presented in 2007 (spacer acquisition) [60]. Later experimental evidences from different studies have shown that the PAM motif is necessary for the interference process [51, 61, 62]. Other studies were done to target the plasmid protospacers in *Sulfolobus* and supported the role that PAM motif is essential for interference, where the *CCN* PAM motif was changed to *GGN*, *GAN* or *TTN* resulting in a lack of interference process [48, 52, 63]. Furthermore, the authors provided evidence showing some correlation between PAM motifs and the spacer acquisition process in *Sulfolobus*.

Since PAM motifs are located in the direct vicinity of protospacers, determining PAM motifs is a straightforward task if the origin of spacer sequences in CRISPR arrays can be determined in sequenced viruses and plasmids. Although recent advances in sequencing technology (High Throughput Data) provide massive information (sequences) from many living organisms in a multitude of environments, the difficulty in discerning virus and plasmid sequences from host sequences makes detecting PAM motifs difficult task. Recently a study predicted PAM motifs and examined their potential functional diversity [64]. Based on diverse experimental evidence on recognition of PAM motifs and cognate DNA strands during adaptation and interference processes, the authors suggested that these motifs can be further discriminated into: (1) a spacer acquisition motif (SAM) for the acquisition process and (2) a target interference motif (TIM) for the interference process. Moreover, they show that predicted PAM motifs have a close correlation to the CRISPR-Cas subtype, which extends and is consistent with earlier studies [48, 52, 63].

2.4.4 CRISPR-associated (Cas) proteins

Generally, CRISPR loci are associated with specific sets of genes encoding proteins called Cas proteins. Cas proteins are the final component of the CRISPR-Cas system that are essential for the multistep mechanism of defence against foreign genetic elements. Initially, four Cas proteins (Cas1, Cas2, Cas3 and Cas4) were identified from *cas* genes near CRISPR arrays that were otherwise absent from genomes lacking CRISPR arrays [25]. Moreover, they observed that if a genome contains several multiple CRISPR arrays with the same consensus repeat sequence, then *cas* genes were associated with only one of them. Later, it was found that Cas proteins are highly diverse and occur only in specific combinations of

what was termed CRISPR-Cas *subtypes*, with the exception of Cas1 and Cas2 proteins which seem to be very well conserved among all CRISPR-Cas subtypes. 45 Cas protein families were identified and classified into 8 main subtypes based on gene synteny and phylogeny of the highly conserved Cas1 protein [65].

The Cas1 protein has been shown to be an α-helical nuclease and a metal-dependent DNA-specific endonuclease that produces double-stranded DNA fragments [66]. Other studies also have used Cas1 phylogeny as a guide for CRISPR-Cas system classification [10, 67–69], due to the fact that the Cas1 protein is universal and evolves slower than other Cas proteins [70]. Recently though, a new family of transposons was defined, the so-called *casposons* which are self-synthesizing mobile genetic elements that rely on Cas1 proteins for integration [71]. The Cas2 protein is the other CRISPR-Cas protein that is universally present along with Cas1 in genomes that have a functional CRISPR-Cas system and together they can be considered hallmarks of the system. The Cas2 proteins are typically small proteins (approximately 120 amino acids) and predicted to be nucleases [72]. Cas2 proteins have been observed to have several conserved sequence motifs located close to the N-terminal β-strand that are suggested to possess nuclease activity [66]. Moreover, Cas2 proteins have some structure and sequence similarity to the VapD toxin subunit of one of the experimentally characterised toxin-antitoxin (TA) systems, which is proposed to be a functional link between CRISPR-Cas and TA systems [66, 73, 74]. Both Cas1 and Cas2 proteins along with possible the additional Cas4 protein form a complex which is essential for integrating new spacers from invading DNA elements [59, 68] (adaptation phase, details later in section 2.4.5).

As mentioned above, the syntenies and types of Cas proteins seem to be enormously diverse, and this tendency increases fast with even more complexity as the number of genomes annotations is increasing. Therefore, multiple classifications of Cas proteins have been established. An early classification was introduced in [65]. The authors identified 45 Cas proteins in the vicinity of CRISPR loci in many archaea and bacteria genomes. Based on multiple sequence alignment and hidden Markov models they built 45 Cas protein families. These models identified family members with sensitivity and selectivity and classify key regulators of development and show evidence that CRISPR systems are more extensive, more complex, and more heterogeneous than it was assumed before that study. Basically, the classification identified eight different Cas subtypes based on gene synteny and diversity of Cas proteins. These Cas subtypes proteins were named: Ecoil, Ypest, Nmeni, Dvulg, Tneap, Hmari, Apern, and Mtube that are implied from the organism where the particular subtype was first found. In addition, they showed that the Repair Associated Mysterious Proteins (RAMP) module that had been classified as a repair system in a previous study [75] is linked to the CRISPR-Cas system and renamed as Repeat Associated Mysterious Proteins (RAMPs). As more and more genomes were published, it became apparent that the subtypes defined by [65] were neither comprehensive nor very accurate. This was not the fault of the original authors as they only had a limited set of genomes to work from when the clas-

sification scheme was first devised. Makarova *et al.*. published their own classification [66] but it had several limitations. The result was that individual lab working experimentally on CRISPR systems from different organisms abandoned any official classification attempts and devised their own systems, further adding to the general sense of confusion. In 2011, after some years of confusion within the community, consensus was regained and a new classification of CRISPR-Cas systems was proposed [68, 69]. It was based on multiple criteria, including the repeats and the evolutionary relationship of Cas proteins, but in practice, with the exception of the addition of a single new subtype, it wasn't much different from the original classification proposed by Haft *et al.* [65] six years prior. Crucially though, the authors of the study had gained the acceptance of all major players within the field before publication, and this made the new classification widely influential as opposed to earlier classification schemes. New Cas protein families were devised namely Cas7, Cas8, Cas9, and Cas10 proteins, by unifying older protein families which were previously thought to be separate. Also, an additional hierarchy was added above the subtype level, called the Type, and three such Types were introduced, namely Types I, II, and III. These major types were further divided into subtypes, mostly corresponding to the ones devised earlier by [65]. The old Apern, Dvulg, Ecoli and Ypest were renamed I-A, I-C, I-E and I-F respectively. Hmari and Tneap were merged into I-B, Nmeni split into II-A and II-B while Mtube and RAMP were renamed III-A and III-B. I-D was the only new subtype not building on the 2005 classification. A system of "signature genes" was introduced to aid identification of the types and subtypes. However, it didn't work in practice and the identification of subtypes in the labs by members of the community remained as obscure as it had been in the Haft era. However, all the work the authors had put into ensuring a consensus prior to publication kept any widespread confusion at bay and kept the field united, in spite of the new system not working in practice. In 2014 one of the few remaining critical groups of the 2011 classification published their results for archaeal CRISPR-Cas systems in [76]. Based on an idea of separating the classification of the proteins responsible for adaptation from those responsible for interference they were able to refine the 2011 classification, and split subtype I-B back into two separate subtypes, denoted I-B and I-G. III-B was split into two subtypes, III-B and III-C and III-A was split into III-A and III-D. Furthermore, candidates for what would later become Types IV and V were proposed. Additionally, the authors identified 12 new accessory protein families and found evidence for widespread exchange between adaptation and interference modules across subtypes. Finally, the authors highlighted the problems with the "signature gene" system for subtype identification as mandated by the 2011 classification, and proposed an alternative of aggregate clustering of interference module components, which was certainly accurate but complicated and prove too difficult to be implement in practice.

In [10], we presented a very comprehensive CRISPR-Cas classification, which classifies archaeal and bacterial CRISPR-Cas systems into two main classes (class I and II). These classes are subdivided into six Types, which are further divided into nineteen subtypes.

Currently, this classification is widely accepted and will be discussed in this thesis in section 3.5. (*Publication 5*).

2.4.5 Mechanism of defence via CRISPR-Cas systems

As mentioned before, CRISPR-Cas systems are composed of three major elements namely: CRISPR array, leader sequence, and Cas proteins. These elements are involved in different phases of CRISPR-Cas mediated immunity. The mechanism of the CRISPR-Cas system is divided into three main phases that are described below: (1) *adaptation*, which is the selection of fragment of genetic material (protospacer) from a plasmid or virus and its insertion into the CRISPR array yielding a new spacer; (2) *expression*, where the CRISPR array is transcribed and processed into mature crRNA; and (3) *interference*, where the invader DNA or RNA is targeted and degraded by a Cas-crRNA ribonucleoprotein complex. Each stage (*adaptation*, *expression*, and *interference*) is associated with a specific group of Cas proteins. These main stages are illustrated in Figure 2.3.

Adaptation stage

The adaptation process is the first step that occurs in CRISPR-Cas mediated immunity, where genetic material from an invading DNA called protospacer is taken up and integrated into the CRISPR array forming a new spacer at the leader-proximal end. This process allows the host organism to memorise the invader and is a prerequisite for the subsequent expression and interference stages that neutralise any re-invading nucleic acids. The ability to acquire new spacers has been experimentally demonstrated for different CRISPR types in several model organisms, such as Type I-A (*Sulfolobus islandicus* and *Sulfolobus solfataricus* [55, 77]), I-B (*Haloarcula hispanica* [78]), I-E (*E. coli* [54, 56]), I-F (*Pectobacterium atrosepticum* and *Pseudomonas aeruginosa* [79, 80]), and II-A (*S. pyogenes* and *S. thermophilus* [60, 81]).

In brief, spacer acquisition starts after infection by a foreign genetic element. New spacers with sequences identical to those of an invading genetic element are incorporated into CRISPR array by duplicating the first repeat of the array. Despite the fact that the spacer acquisition is poorly understood, experimental studies have been demonstrated that Cas1 and Cas2 proteins and in some cases also Cas4 protein are obligatory for the spacer acquisition at the host CRISPR array [53, 54, 59, 68, 72, 82]. The PAM motif, which is normally located in a protospacer flanking region plays critical role in determining the orientation of inserted spacers and also for discrimination between self/non-self targeting [26, 48, 61, 63, 83, 84]. Moreover, the *adaptation* process can be further separated into two distinct types [85]: (1) naïve acquisition, which occurs during infection with a phage that has not previously been encountered in acquisition of a new spacer; and (2) primed acquisition, when the invader has been previously encountered. In primed space acquisition, the spacer acquisition is coupled with the interference machinery. This accelerates the acquisition of the subsequent spacer. The advantage of this type of acquisition is to increase

the resistance toward recurring invasions.

Biogenesis of crRNAs (the expression stage)

The *expression* stage can be subdivided into two stages: (1) CRISPR array transcription and regulation, and (2) crRNA processing. Both stages are obligatory for the subsequent interference process (see Interference stage below). Transcription initiation of the CRISPR array starts upstream of the first repeat sequence inside the leader sequence region and terminates downstream from the CRISPR array. The CRISPR array is transcribed into long precursors of CRISPR RNA (pre-crRNA) that are subsequently matured by processing into small CRISPR RNAs called crRNAs. These crRNAs contain a single targeting spacer flanked by a part of CRISPR repeat that can accommodate small hairpin structure motifs [86] (see Figure 2.3).

The *expression* process was the first process in the CRISPR-Cas system to undergo experimental studies [87, 88]. Unlike the *adaptation* stage, the *expression* stage is different among the various CRISPR-Cas Types. The key factor in the processing of pre-crRNA for Type I and III systems is the Cas6 protein family that is involved either in cleaving at a small hairpin inside the repeat or within the repeat sequence without requiring a hairpin (unstructured repeat) [53, 89–93]. In some cases in Type I systems, the Cas6 protein is a part of the ribonucleoprotein complex of Cas proteins known as the Cascade complex (CRISPR-associated complex for antiviral defense), while for Type III systems Cas6 usually does not belong to the complex. The Cas6 protein is an endoribonuclease necessary for crRNA production whereas the additional Cas proteins that form the Cascade complex are needed for crRNA stability [53, 93–96]. Overall, all subtypes of Type I and III systems use a Cas6 protein for pre-crRNA processing except subtype I-C which utilises a modified version of Cas5, namely Cas5d [6].

In contrast, in Type II systems a totally different process is employed for crRNA biogenesis pathway. It depends on the Cas9 protein, the housekeeping endoribonuclease RNase III, a small RNA and tracrRNA (trans-activating crRNA) that is required for the processing of the pre-crRNA [97, 98]. The tracrRNA interacts with each of the repeat sequences of the pre-crRNA to generate RNA duplex, which is stabilised by the Cas9 protein. Further, the RNA duplex is processed and cleaved by the endoribonuclease RNase III [97, 98] whereas another maturation, which leads to the mature small crRNA, is performed by an unknown enzyme [98]. In addition, an interesting study in the recently defined subtype II-C system of *Neisseria meningitidis* shows that crRNA guides are transcribed from promoters located within the repeats of the CRISPR array, indicating that the RNase III is not involved in the crRNA biogenesis pathway [99].

Interference stage

The *interference* process is the last main stage that occurs in the CRISPR-Cas mediated immunity, where crRNA guides associate with the Cas interference complex to specifically cleave matching invasive nucleic acids. Invading nucleic acids are identified by base-pairing interactions between the crRNA spacer sequence and a complementary sequence from the invader. Both the *expression* and *interference* stages occur differently in each of the CRISPR systems [2, 53, 100]. Type I and III systems utilise "Cascade" complexes to perform the target degradation. On the other hand, Type II systems require only the Cas9 protein in order to employ target recognition and degradation of nucleic acids [97, 98]. The PAM motif plays an important role also in the *interference* stage. In Types I and II systems, the PAM motif that is located either upstream or downstream of the invading foreign DNA (protospacer sequence), is used for discriminating of non-self from self targeting [51, 101–104], while for Type III systems, the self versus non-self targeting is achieved through additional hybridisation, or lack thereof, within the repeat portion of the crRNA [105].

As mentioned previously, CRISPR-Cas proteins are highly diverse. They are key factors during all stages of immunisation, in which different groups of Cas proteins are involved in the different steps of the processing of CRISPR-Cas system. Furthermore, different CRISPR-Cas subtypes (e.g. I-A, I-C, I-E, etc) within the same CRISPR Type (e.g. Type I) have different groups of Cas proteins, which implies variations in the composition of the Cascade complex in the *interference* stage. For instance, the subtype I-A Cascade complex is composed of multiple copies of Cas7 proteins comprising the backbone of the complex, the large Cas3 protein with helicase functionality, a single-stranded DNA known as Cas3 protein, as well as a subcomplex of Cas5 and Cas8 bound to each other [106, 107]. Subtype III-B Cmr-α complex consists of many copies of Cas7 paralogs, i.e. three subunits of Cmr1, Cmr4, and Cmr6 known as RAMPs, small subunits of Cmr5, and large subunits of Cas10 protein that contains a HD phosphohydrolase domain and palm domain [69, 108, 109]. The Cascade complex in Type I and III CRISPR-Cas systems contains large number of Cas proteins, which have the ability to affect foreign double-stranded DNA (dsDNA) and single-stranded (ss) RNA. On the other hand, the Type II system interference complex is again used composed of only the Cas9 protein, which has the ability to generate crRNA, target invading DNA for degradation [97, 98].

2.5 CRISPR-Cas-based gene editing

Genome editing or gene editing is an approach in which DNA sequence is directly inserted, replaced or deleted in the genome of a living organism [110]. Meganucleases and Zinc-Finger Nucleases (ZFNs) have been the first approaches developed for gene editing [111, 112]. Although both approaches have made a critical improvement in the gene editing field, they had several disadvantages. For example, they are very expensive, difficult to engineer and require robots. A few years later, a new approach called Transcription Activator-Like

Effector Nucleases (TALENs) was developed [113, 114]. On one hand, TALENs is similar to ZFNs in which both use DNA binding motifs to cleave the genome at specific site [113, 114]. On the other hand, TALENs relies on the modularity of the TALE subunits, which makes it less expensive, faster and better in comparison to ZFNs. Nowadays, the type II CRISPR-Cas systems have become a very efficient tool for gene editing. In type II systems, Cas9 protein associates with a dual-RNA guide structure that consists of a crRNA and tracrRNA to cleave double-stranded DNAs complementary to the 5' terminal guide segment in crRNA. Unlike ZFNs and TALENs, CRISPR-Cas system uses single guide RNA (sgRNA) to make a DNA double-strand breaks (DSBs) on a targeted location of genome instead of modular DNA-binding proteins. CRISPR-Cas has many advantages when compared to other methods for genome editing. It is efficient, simple, inexpensive, and has much more efficient multiplexed mutations. To date, a rapid series of studies showed the CRISPR-Cas efficient genome editing in a variety of species and cell types, including human cell lines, bacteria, yeast, zebrafish, mouse, fruit fly, roundworm, rat, common crops, pig, and monkey [115, 116].

2.6 Machine Learning Techniques

Over the past two decades, Machine Learning has become one of the backbones of information technology. The tremendous growing of data is the main motivation to believe that smart data analysis will become even more prevalent as a necessary element for technological advancement [117]. Machine learning is "an adaptive process that enables computers to learn from experience, learn by example, and learn by analogy" [118]. Learning abilities are vital for automatically improving the performance of the system over time on the basis of previous results. Typically, a basic model of machine learning consists of three mechanisms: (i) the learning component, in charge of improving its performance; (ii) the critical component, which tells the learning component how the algorithm performs and (iii) the problem generator, responsible for developing actions that can lead to new or informative experiences [119].

Machine learning processes comprise several aspects of learning, such as the acquisition of new declarative knowledge, the transformation of new knowledge into general effective representations, and the finding of new facts and theories through observation and experimentation [120]. Since the commencement of the computer era, researchers have been struggling to embed such abilities in computers. Solving this problem has been a most challenging and captivating long-range aim in artificial intelligence (AI) [120].

Machine learning classically can be divided into three phases: (1) analysis of a training set of cases and generation of a set of rules from the training set; (2) verification of the rules by human experts or automatic knowledge based components and (3) use of the validated rules in responding to some new testing datasets [119]. Machine learning has many applications. For instance, it can be used to automate the process of designing a good search engine [117]. Another application is collaborative filtering. Giant e-commerce websites such

as Amazon or Netflix use this information broadly to induce customers to purchase other related goods. Automated translation of documents is another example. The optimum expectation is a fully understanding a text before translating it using a set of rules crafted by a computational linguist well experienced in the two languages, we would like to translate [117]. This is a rather strenuous mission, particularly assuming that text is not always grammatically correct, nor is the document understanding part itself a trivial one. Other applications (listed in the book [117]) that need and take advantage of learning are speech recognition (annotating an audio sequence with text), the recognition of handwriting (annotating a sequence of strokes with text), track-pads of computers, the detection of failure in jet engines, avatar behaviour in computer games, direct marketing (companies use prior purchase behaviour to estimate whether customers might be willing to purchase even more) and floor cleaning robots.

One large and important application area for machine learning approaches is the field of bioinformatics. Biological data are large in volume. Traditional computer science techniques and algorithms fail to solve complex biological problems of the real world [118]. However, there are new computational approaches based on machine learning that can overcome the limitations of the traditional techniques. Machine learning can excerpt the description of the hidden situation and then create rules that match the expert's behaviour. Moreover, Information systems sometimes produce results different from the desired ones. This is due to unknown properties or functions of inputs during the design of systems. This situation always arises in the biological world because of the complexities and mysteries of organismal life. However, with its capability of dynamic improvement, machine learning can deal with this problem. In terms of molecular biology, new data and concepts are generated daily, and those new data update or replace the old ones. Machine learning can be adapted to a changing environment. This aids system designers, as they do not need to redesign systems whenever the environment changes [118]. Last but not least, missing and noisy data is a well-known problem of biological data. Conventional computer techniques are incapable of handling this. Machine learning techniques are able to treat missing and noisy data. Finally, with developments in biotechnology, enormous volumes of biological data are generated. In addition, it is possible that important hidden relationships and correlations exist in the data. Machine learning methods are designed to handle very large data sets and can be used to extract such relationships.

Generally, there are three types of machine learning algorithms: (1) Supervised learning, (2) Unsupervised learning, and (3) Reinforcement Learning. In this section, we will give a brief overview about only Supervised and Unsupervised learning techniques that are used in this thesis.

2.6.1 Supervised learning

Supervised learning is the type of machine learning tasks that infers a function from training data. The training data require a set of training instances. In supervised learning, each

instance is a pair containing an input object and a desired output value (supervisory signal). A supervised learning algorithm is fed with the training data, analyses it, and produces an inferred function (a classifier) when output is continuous. The output function should forecast the correct output value for any valid input object. This requires the learning algorithm to extract useful representations from training data in a reasonable way. To solve a problem using supervised learning method, the typical steps are (1) determine the type of training data; (2) gather a training set; (3) determine the input feature representation of the learned function; (4) determine the structure of the learned function and corresponding learning algorithm; (5) complete the design and run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters; (6) evaluate the accuracy of the learned function after parameter adjustment and learning.

Supervised learning has limitations such as the trade-off between bias and variance [121]. For instance, when there are several good training data sets, a learning algorithm might be biased for a particular input if it was trained on these data sets, or it might have high variance for a specific input if it predicts different output values when trained on different training sets. Then, the prediction error of the classifier is defined as the sum of the bias and the variance of the learning algorithm [122]. Usually, there is a trade-off between bias and variance. However, a learning algorithm with low bias should be "flexible" so that it is fitting the data well. But if the learning algorithm is too flexible, it will be suitable for each training data set in a different way and hereafter have high variance. A key feature of numerous supervised learning methods is that they are capable of adjusting this trade-off between bias and variance either manually or through model selection.

2.6.2 Unsupervised learning

Unsupervised learning overall has a long and illustrious history. Some early inspirations were given by Horace Barlow, who wanted ways of characterising neural codes, Donald MacKay (1956), who used a cybernetic-theoretic approach, and David Marr (1970), who made an early unsupervised learning hypothesise about the aim of learning in his model of the neocortex [123].

Unsupervised learning is a type of machine learning algorithm used to extract useful information from data sets without labelled responses. Unsupervised learning is concerned about how systems could learn to identify specific input patterns in a way that reflects the statistical structure of the whole collection of input patterns [123]. The unsupervised learning problems can be classified as clustering and association problems. Clustering problems are the type where you want to determine the inherent groupings in the data, such as grouping of similar protein domains [124]. The second category is association, which involves problems where you want to find out rules that define large portions of your data [125].

The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are

obtained using a measure of similarity which is defined using metrics such as euclidean or probabilistic distance. The common clustering algorithms comprise: (a) hierarchical clustering, which develops a multilevel hierarchy of clusters by generating a cluster tree; (b) k-means clustering, which divides data into k distinct clusters based on distance to the centroid of a cluster; (c) gaussian mixture models, which models clusters as a combination of multivariate normal density components.

2.7 The K-Nearest Neighbour Algorithm (KNN)

K-Nearest Neighbour (KNN) is a supervised learning algorithm that has applications ranging from vision to proteins, computational geometry, graphs and so on [126]. KNN is a non parametric lazy learning algorithm. It means that it does not make any assumptions on the underlying data distribution [126]. This feature has an advantages in real life because most of the practical data does not follow the classic theoretical assumptions made (gaussian mixtures, linearly separable, etc). It is also a lazy algorithm, which means that it does not use the training data points to do any generalisation [126]. Lack of generalisation means that KNN keeps all the training data during the testing phase. KNN makes a decision based on the entire training data set.

KNN assumes that the data is in a feature-space. More precisely, the data points have to be in a metric space. Hence, the data can be represented by scalar multi-dimensional vectors. Since the points are in feature-space, they have a notion of distance [126]. KNN can be employed for both classification and regression predictive cases. However, it is more extensively used in classification problems in the industry [127].

The KNN algorithm finds the k samples in the training dataset that are most similar to the point that we want to classify. The class label of the new data point is then selected by a majority vote among its k nearest neighbours [128]. The KNN algorithm is straightforward and can be explained by the following steps: (1) Choose the number of k and a distance metric; (2) Find the k nearest neighbours of the sample that we want to classify; (3) Assign the class label by majority vote.

2.8 Kernel

A kernel is defined as a continuous and symmetric function K that takes two arguments x and x'. K maps them symmetrically to a real value that represents similarity between them. $K : X \times X \rightarrow |\mathbb{R}$
$\forall x, x' \in X : K(x, x') = K(x', x)$

In this section, we will give a brief introduction about two types of kernel approaches. Special cases that are used in this thesis: (i) *Graph kernel approach* (see section 3.3 and *Publication 2*) and (ii) *String kernel approach* (see Section 3.4 and *Publication 3*).

2.8.1 Graph kernel approach

Machine learning and data mining in the last decade have contributed considerably in facilitating the processing of all manner of input data regardless of the type formats and data size. Data structures such as sequences, trees, and graphs are becoming possible to process. In supervised learning methods, one advantage is that a linear model with worthy generalisation features could be expanded to a non-linear model using a kernel [129, 130]. Numerous kernel functions are used (e.g. Gaussian, polynomial, and others) in e.g. support vector machines to map the input data into a very high-dimensional feature space. Here, we focus on graph kernels that process entities encoded as graphs. They use dot products to calculate a similarity measure among graphs. Though there are numerous types of graph kernels, we have detail in the following a fast kernel, specifically, Neighbourhood Subgraph Pairwise Distance Kernel (`NSPDK`), which is lately presented by [131], due to its suitability for large sets of sparse graphs with discrete vertex and edge labels.

Notation and definitions

Consider a graph G = (V,E), where V denotes the set of vertices and E denotes the set of edges. Every edge of E is associated with a set of two elements of V. A graph with labelled vertices and edges is called labelled graph. The function ℓ maps vertices/edges to the set of label symbols. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are isomorphic, if there is a bijection $\phi : V_1 \to V_2$. This is represented by $G_1 \cong G_2$. Normally, an isomorphism is a structure-preserving bijection. Therefore, when two labelled graphs have isomorphism, we call them isomorphic.

Graph kernel

The `NSPDK` is an instance of a decomposition kernel [132] where all the potential "parts", due to a given relation, are extracted. In this situation, each part is a pair of different subgraphs, which are recognised as "neighbourhood" subgraphs. At this point, the central goal is to decompose a graph into a small neighbourhood subgraphs of increasing order radii $r < r_{max}$. A neighbourhood subgraph considering of all nodes (and connecting edges) within distance r around a root node u are denoted by N_r^u. If the roots of all pairs of such subgraphs are at a distance (d) not exceeding d_{max} ($d < d_{max}$), we consider them as individual features. The portion of features shared between two graphs is defined as the resemblance notion. The formal kernel definition is described here. The relation between neighbourhood subgraphs is defined as:

$$R_{r,d} = \{(N_r^v(G), N_r^u(G), G) : d(u, v) = d\}, \tag{2.1}$$

where $R_{r,d}$ (neighbourhood pair relation) recognises a pair of neighbourhood subgraphs of radius r, which has root distance exactly equal to d. On this relation $R_{r,d}$, a decomposition

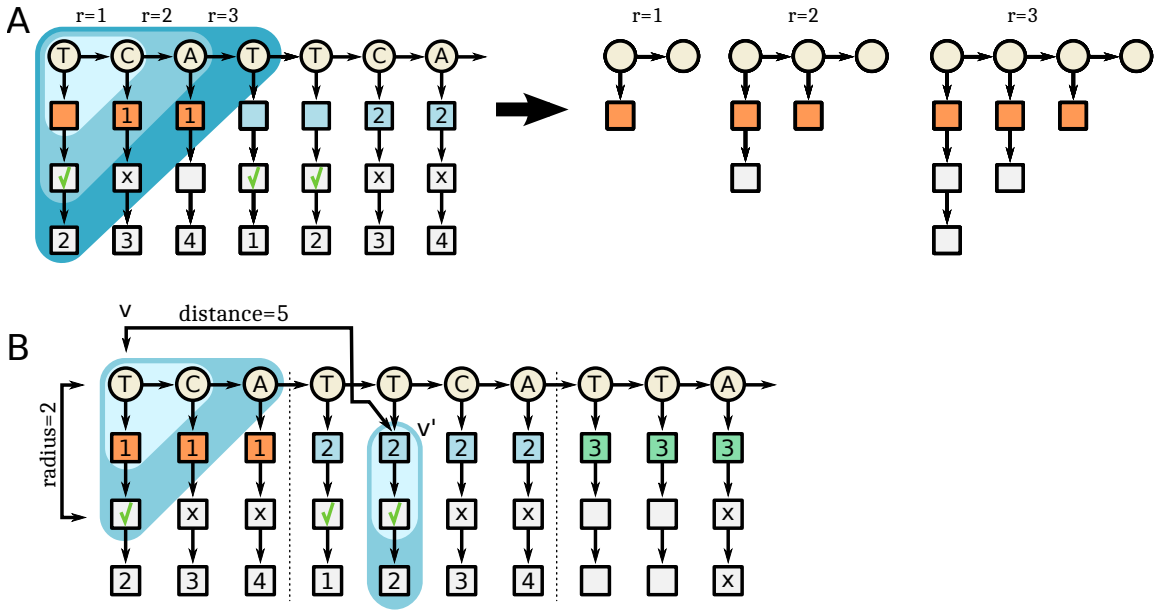


Figure 2.4: (A) NSPDK features for Distance (d) = 0 and Radius (r) = 1, 2, 3 relative to a given root vertex highlighted in Light Cyan, Spary, and Viking respectively. The directed property of the graph allows to induce features that can differentiate strand directions. (B) Example of feature for $r = 2$ and $d = 5$ capable to capture the dependency between two nucleotides at relative distance 5. The sequence information that is not contained in the neighbourhoods is ignored; the effect is equivalent to a *don't care* pattern. The figure adopted from *Publication 2*.

kernel $\kappa_{r,d}$ is defined as:

$$\kappa_{r,d}(G, G') = \sum_{\substack{A,B \in R_{r,d}^{-1} \\ A',B' \in R_{r,d}^{-1}}} \xi(A \cong A') \cdot \xi(B \cong B'), \tag{2.2}$$

where $R_{r,d}^{-1}$ is the inverse of $R_{r,d}$, which designates all the potential pairs of neighbourhood subgraphs of radius r and the root distance d that occur in the graph G, and the pointer function and the isomorphism between graphs are signified by ξ and $\cong$ correspondingly. The isomorphism check is achieved by the method Graph invariant, described below in this section. The case of nucleotide sequences are considered as NSPDK features (see Figure 2.4). The non-normalised NSPDK is defined as:

$$K(G, G') = \sum_{r} \sum_{d} \kappa_{r,d}(G, G'). \tag{2.3}$$

to increase the efficacy, we can impose an upper bound on the radius and distance parameters:

$$K_{r_{max},d_{max}}(G, G') = \sum_{r=0}^{r_{max}} \sum_{d=0}^{d_{max}} \kappa_{r,d}(G, G'). \tag{2.4}$$

Lastly, a normalised version of $\kappa_{r,d}$ is

$$\hat{\kappa}_{r,d}(G, G') = \frac{\kappa_{r,d}(G, G')}{\sqrt{\kappa_{r,d}(G, G)\kappa_{r,d}(G', G')}}. \tag{2.5}$$

This guarantees that the graph properties persuaded by all values of radii and distances are correspondingly weighted regardless of the feature space dimensionality.

Graph invariant

For solving of the graph isomorphism problem (GIP), it is unknown whether polynomial algorithms are available [133]. Yet, for special graph classes, polynomial algorithms do exist [134]. Limited algorithms that are exponential were established to resolve the GIP earlier [135, 136]. However, the precise isomorphism test is computationally very expensive and often not needed. Thus, Costa *et al.* developed a solution, similar to [137], where the particular isomorphism examination is substituted by presenting an effectual graph invariant computation [131]. The key concept is to generate a typical string from two isomorphic graphs by effective graph serialisation method. Thus, the string could be mapped into a code by a repeated hashing technique. Consequently, the isomorphic exam could be effortlessly replaced by an equivalence test between the codes of two graphs. This entire process works in two main stages: (1) construction of a graph invariant coding $\mathcal{L}^g(G)$ and (2) use of a standard hash function $H(\mathcal{L}^g(G)) \to \mathbb{N}$ to obtain the wanted identifier. Notice that overall, the process is influenced by the probability for a collision between two non-isomorphic graphs. This could occur either due to the non-isomorphic graphs own the same encodings or due to a collision presented by the hashing method even though they have dissimilar encodings.

In the computation of a graph invariant, the graph encoding $\mathcal{L}^g(G)$ was gained by recognising two label functions: (1) for the vertices ($\mathcal{L}^v$) and (2) for the edges ($\mathcal{L}^e$). The function $\mathcal{L}^v(v)$ maps a vertex v to a lexicographically sorted sequence of pairs comprising a topological distance and a vertex label all $u \in G$. By combining the original edge label and the new vertex labels, the new edge label, $\mathcal{L}^e(uv)$ is created for an edge $v \to u$. After that, the sorted lexicographically series is allocated to the graph G by $\mathcal{L}^g(G)$. Finally, a construction based hashing technique, developed by Damgård [138] is utilised to map the adaptable length data into numerous lists of integer codes [131]. The graph invariant computing procedure is portrayed in Figure 2.5.

2.8.2 String kernel approach

A string kernel is a method that performs the computational operation of strings in a high-dimensional, implicit feature space without ever computing the definite coordinates of the string in that space, but rather by basically computing the inner products between the

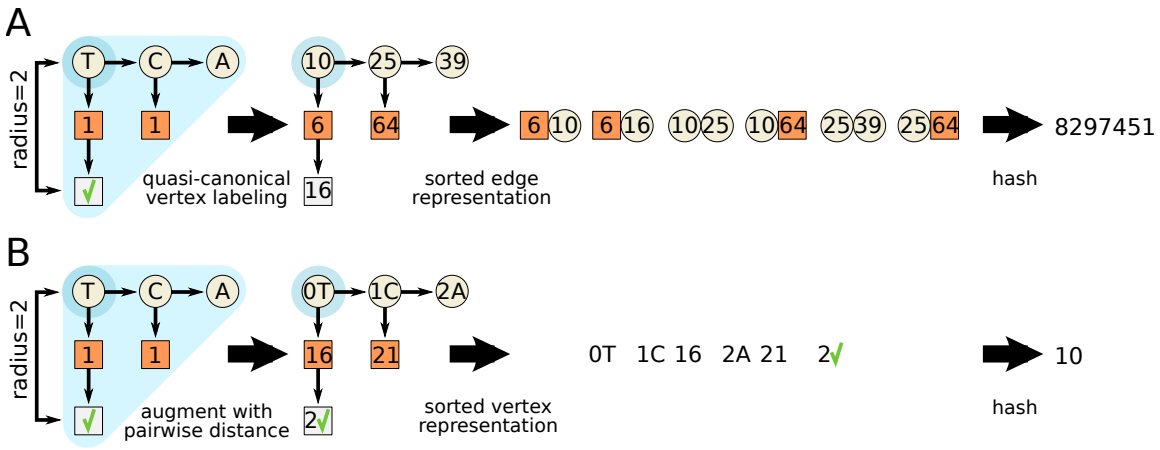


Figure 2.5: Graph invariant computation for rooted graphs. (A) an integer code is obtained from the sorted list of edges by hashing technique. (B) a vertex quasi-canonical label is computed. Here, the root vertex T is converted to an integer code 10. The figure adopted from *Publication 2*

images of pairs of strings in the feature space, which is known as kernel trick. The inner product computed by the kernel method could be utilised to identify a similarity notion. When normalised, the kernel matches pairs of strings into the interval $[0, 1]$ where 1 refers to the two strings are indistinguishable (for the kernel) and 0 refers to strings that do not share any similarity. Common string kernels are based on the notion of *k-mers* (which are substrings of length k). The *k-mer* kernel (also called spectral kernel in [139]) between s and s', $K(s, s')$, is the number of *k-mers* that are identical between two strings s and s'. A normalised kernel calculates the portion of identical *k-mers* w.r.t. the overall number of *k-mers* present in the two strings s and s', often as the quantity: $K(s, s')/\sqrt{K(s, s) \cdot K(s', s')}$. Since the occurrence of *k-mers* is exponentially less probable w.r.t. their length k, there is a slight advantage in considering large *k-mers* (e.g. $k > 10$) when working with biological sequences. Small *k-mers* on the other hand might not enable an adequate discriminative power. To reduce these problems, a notion of "approximate match" was presented in [140], where the insertion, deletion or mismatch of up to m components of the *k-mers* is accepted when computing the correspondences. Practically however, these inexact techniques have high run-times and are not always effective in increasing the discriminative power.

Chapter 3

Bioinformatic analysis of CRISPR-Cas systems—classification and structure analysis

This chapter contains an overview, results and conclusion of each publication in this PhD thesis. We start by briefly showing the bioinformatics-based effort to discover new functions and stay at the forefront of the CRISPR-Cas system research. Then, we show the three major areas that are the subject of this thesis: namely, classification and sequence-structure analysis of CRISPR arrays (repeats and spacers), detection and analysis of leader sequences, and classification of CRISPR-Cas proteins.

3.1 The Role of Bioinformatics in CRISPR-Cas System Research

The signature architecture of repeat-spacer units in the CRISPR array was first described in *Escherichia coli* [36]. Later, aided by bioinformatics tools, Francisco Mojica showed that CRISPR arrays not only exist in *Escherichia coli*, but also in most archaeal and many bacterial genomes [141]. Subsequently, bioinformatics analyses revealed matches of spacers to bacteriophages, which lead to the correct hypothesis that CRISPR-Cas systems act as an acquired immune defence system [44, 45, 47]. Later, another bioinformatics study on spacer matches successfully predicted CRISPR-Cas systems to target primarily DNA rather than RNA [63]. Thus, both the discovery and initial hypotheses were borne out of bioinformatics analyses. Bioinformatics studies within the field have progressed along two tracks, with analyses of protein coding genes on the one hand and prediction of nucleic acid interactions on the other hand. Since the initial discovery, much of effort has gone into the classification of CRISPR systems through a computational discovery of novel *cas* genes. Bioinformatics-based searches and genomic-context analyses were performed to devise several iterations of classification and annotation schemes [10, 25, 65, 66, 68]. Each iteration has spawned new generations of biochemical studies that focused on the freshly discovered proteins and eventually resolved their role in CRISPR-Cas immunity. Initially, 4 *cas* genes were identi-

fied [25], whereas in the latest classification system [10], there are more than 30 core genes described. We have also observed a continuously growing list of accessory proteins, which lists now at least 12 families [76]. Thus, we can assume that there are more to be discovered in the future. The analysis of metagenomic data provides a potentially rich source of new *cas* genes and CRISPR-Cas types. As for nucleic acid bioinformatics, analyses of RNA structure and sequence was performed on the repeats of CRISPR arrays to cluster these into functional groups [7, 43]. Detection of weak DNA sequence similarity signals has previously aided the identification of CRISPR spacer matches and now also resulted in a new method for accurate identification of CRISPR leaders [9]. Bioinformatics studies of Cas proteins and CRISPR nucleic acids has helped to define the vast variety of these systems in nature and has guided the search for molecular mechanisms. Recent examples include the definition of Type V and VI systems [10, 76].

3.2 CRISPRmap: An automated classification of repeat conservation in prokaryotic adaptive immune systems

This section summarises the work from *Publication 1*, which was published in the Nucleic Acids Research journal.

3.2.1 Overview

Normally, the annotation of new CRISPR-Cas elements is based on a direct search for CRISPR arrays (repeat-spacer units), which is the only element that is present in all CRISPR-Cas systems. The CRISPR array can be identified easily through programs such as CRISPRFinder [33] or CRT [34] and this is in contrast to leader sequences and Cas proteins (see Subsection 2.4.1). In the CRISPR-Cas system, the CRISPR repeat can be treated as the central regulatory element. The CRISPR repeat acts as a binding template for Cas proteins and is relevant in all three identified CRISPR-Cas phases of immunity: *adaptation*, *expression*, and *interference* (see Subsection 2.4.5). Therefore, it is important to have a systematic CRISPR repeat-based classification in order to properly evaluate the CRISPR-Cas immune systems from perspective of function, diversity and phylogeny.

We provide the first automated comprehensive classification system of CRISPR repeat based on sequence and structure similarity. This is a major advance over previous classification attempts, which were strongly limited by their reliance on non-trivial manual annotation and inspection of CRISPR associated Cas proteins (*Publication 1* [7]). Previous clustering approaches are based on pairwise similarities: in order to find a biologically meaningful clustering, similarities between repeats should reflect conserved binding mechanisms. The binding affinity of Cas proteins is not only affected by the repeat sequence: a small hairpin structure is a key binding motif for Cas endoribonucleases in several systems [11, 53, 90, 91, 142–144]. To correctly identify these structure motifs, our clustering

is the first that is based not only on sequence but also on structure similarities. This approach is well-established for the identification and characterisation of structured ncRNA [145–148]. The conservation of ncRNAs structure is considered more important than sequence for the biological function [149, 150]. Although many of the CRISPR repeats form hairpin-structure motifs, no structure-based clustering existed so far. In 2007, around of 300 bacterial and archaeal CRISPR repeat sequences were classified into twelve main classes [43]. Although structure motifs were identified in six classes, the underlying clustering was based purely on sequence similarity.

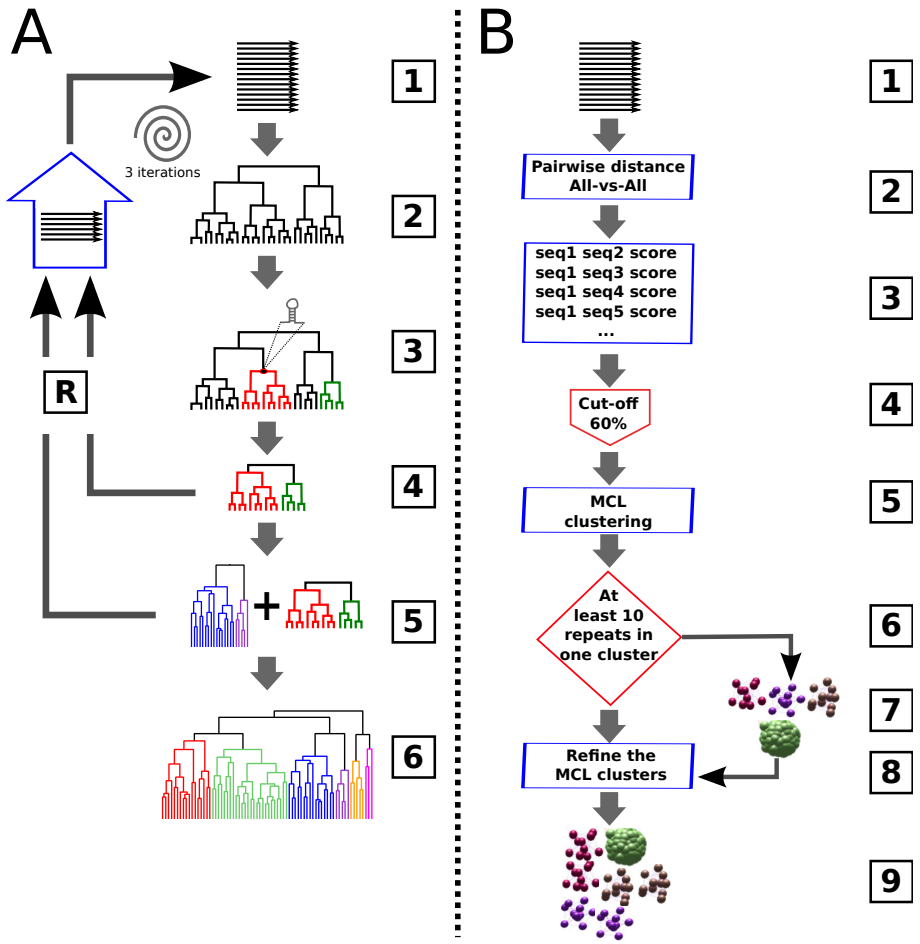


Figure 3.1: Overview of the two independent approaches to characterise the sequence and structure motif to which the Cas protein binds. (A) Pipeline for identifying conserved structure motifs. (1) Pool of repeats with predicted orientation. (2) Hierarchical clustering of all repeats in the pool. (3) Selection of subtrees with CRISPR-like consensus structures. (4) Creation of a supertree with only repeats fitting to the identified consensus structures. (5) Merging the supertree from current iteration with the trees from a previous iteration. (6) Final cluster tree containing 33 structure motifs. (B) Pipeline for identifying conserved sequence families. (1) Pool of repeats with predicted orientation. (2) and (3) Pairwise similarity with global sequence alignment. (4) Reasonable cut-off chosen to represent a significant similarity. (5) Clustering into related families using Markov clustering. (6) Considered clusters with at least 10 repeats. (7) and (8) Reassigning repeats to families to which they are sufficiently similar. (9) Final 40 sequence families. The figure is adopted from *Publication 1*.

3.2.2 Discussion and Results summary

To provide a complete overview of the conservation of both structured and unstructured CRISPR repeats. We compiled the largest dataset of CRISPR repeats at that time (> 3,500 CRISPR repeat sequences in around 2,500 genomes), and performed comprehensive, independent clustering analyses to determine conserved sequence families, potential structure motifs for endoribonucleases; and evolutionary relationships. The results of our independent clustering approaches are as follows: (1) thirty-three structure motifs were identified based on sequence and structure alignments using LocARNA [146, 148]; (2) forty conserved sequence families were identified based on Markov clustering (MCL) [151] (see Figure 3.1). Briefly, we generated a hierarchical representation that portrayed the relationships between structure motifs and sequence families classes and also between individual CRISPR repeats. The resulting hierarchical CRISPRmap tree provides both a quick and detailed insight into CRISPR repeat conservation and diversity of CRISPR-Cas systems and reveals unexplored regions. We enable practical access to our data via an easy-to-use web server, CRISPRmap, where users can identify relative positions of both published and unpublished sequences on the map of CRISPR sequence and structure conservation.

Furthermore, we grouped our CRISPR repeats database into six major superclasses that share common conservation patterns. Superclasses corresponding to archaeal CRISPR repeats contain well-conserved sequence families, and half of the repeats have structure motifs associated. The archaeal structure motifs, however, are less stable than those associated with bacteria.We found that the superclasses correlate well with Cas1 protein evolution, which was already known to be linked to repeat evolution. In addition to the CRISPR repeat structure motifs and sequence families, we annotated taxonomic phyla, Cas1 protein sequences homology clusters, and Cas subtype annotations. (see Figure 3.2).

We used all published repeat structures at the time to validated our result. We observed from the literature that cleavage by Cas6-like endoribonucleases (during crRNA maturation) happens either at the 3' base of the hairpin motif, or within the double-stranded region of the hairpin stem, and usually below a $C \rightarrow G$ base pair [11, 53, 90, 91, 142–144, 152]. The cleavage product is an 8-nt-long repeat tag at the 5' end of the mature crRNA (5' tag), which corresponds to the last 8 nt from the 3' end of the repeat sequence. Some exceptions to the 8-nt length have been observed in several organisms [6, 11, 98, 144, 152, 153]. We examined potential cleavage sites on our structure motifs based on published observations [6, 11, 53, 91, 142–144, 152]. Eleven structure motifs out of thirty-three structure motifs contain a potential cleavage site between two base pairs in the conserved stem of the motif of which seven are below a $C \rightarrow G$ base pair. The remaining thirteen structure motifs have a potential cleavage site at the 3' base of the conserved stem.

To conclude, we performed a comprehensive, independent clustering and identified a novel set of 33 potential structure motifs and 40 conserved sequence families based on the largest data set of CRISPR repeat sequences available. We show conclusively that our methods

3.2. CRISPRmap: An automated classification of repeat conservation in prokaryotic adaptive immune systems

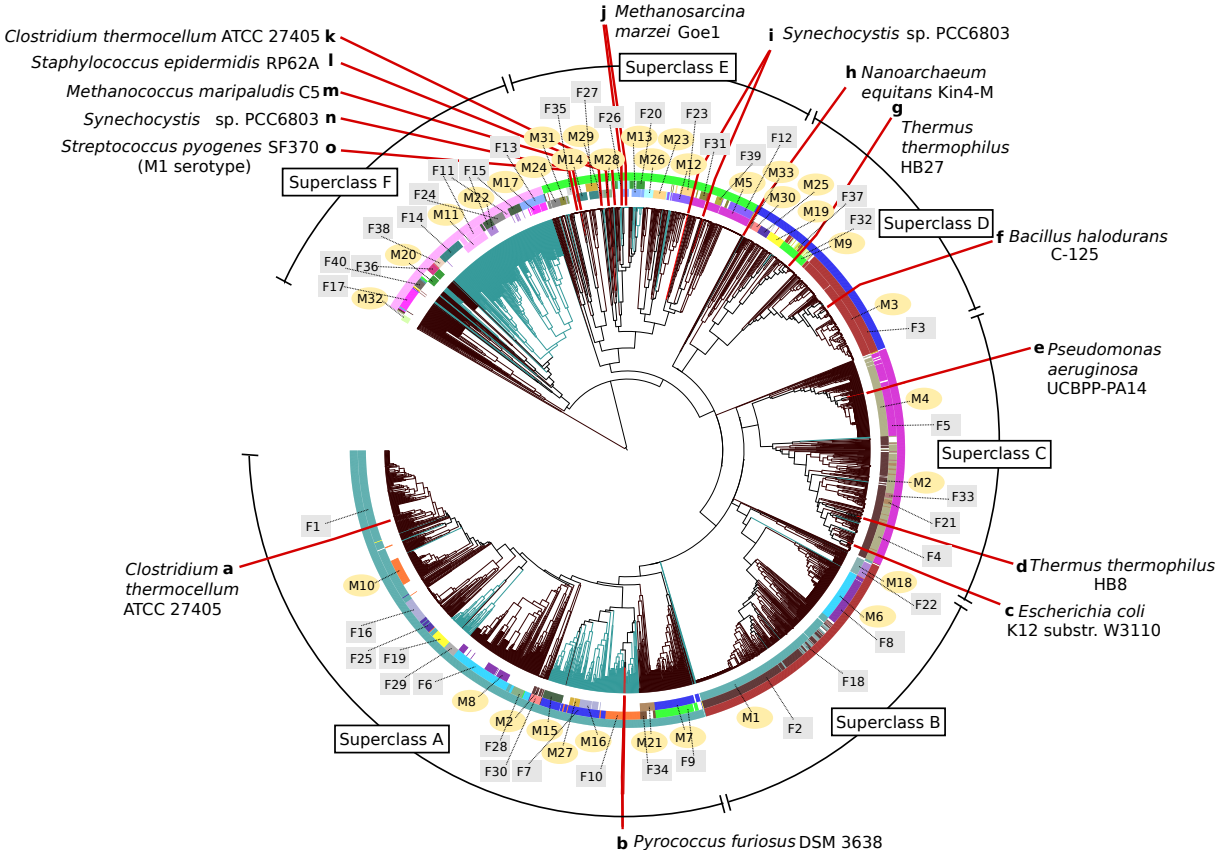


Figure 3.2: The CRISPRmap hierarchical tree: a map of repeat sequence and structure conservation. The hierarchical tree is generated with respect to repeat sequence and structure pairwise similarity and the branches are coloured according to their occurrence in the domains bacteria (dark brown) or archaea (blue-green). The rings annotate the conserved structure motifs (inner), sequence families (middle), and the superclass (outer). Motifs and families are marked and highlighted with yellow circles, and grey squares, respectively. Finally, **a**–**o** mark locations of published CRISPR-Cas systems for which experimental evidence of the processing mechanism exists. The figure is adopted from *Publication 1*.

are accurate and suitable for identifying the most important characteristics of CRISPR-Cas systems, such as cleavage sites, patterns of RNA structure motifs and sequence conservation. Additionally, our methods are also suitable to measure the link between evolution of the CRISPR array and associated Cas subtypes, and the horizontal transfer of such systems. Given the mapping of published CRISPR repeats to our sequence families and structure motifs (Figure 3.2), we assume that our results for members of those same families and motifs are correct. Finally, we developed a web server called CRISPRmap that enables individualised mappings (as in Figure 3.2) and allows scientists to access all our data. CRISPRmap helps researchers to identify the location of both published and unpublished CRISPR repeats on the map of sequence and structure conservation as well as to find potential CRISPR-Cas system that are highly divergent from the rest.

3.3 CRISPRstrand: Predicting repeat orientations that determine the strand from which crRNAs are processed at CRISPR loci

Prediction of the CRISPR orientation (i.e., the strand from which mature crRNAs are derived) is a non-trivial problem. This section is based on the work from *Publication 2*, which was published in the Bioinformatics journal.

3.3.1 Motivation

CRISPR arrays consist of several repetitive DNA sequences called repeats interspaced by stretches of variable length sequences called spacers (see Subsection 2.4.1). The mechanism of CRISPR-Cas system is divided into three phases: *adaptation*, *biogenesis of crRNAs*, and *interference*. CRISPR arrays are transcribed and processed into mature RNA species (crRNAs), which is guided by specific Cas-protein interference complexes to their target, leading to cleavage of invading nucleic acids carrying matching sequences (see Subsection 2.4.5). Although existing bioinformatics tools can recognise CRISPR arrays by their characteristic repeat-spacer architecture, they provide limited insight into the orientation of the CRISPR array. Knowledge of the correct orientation of CRISPR arrays is crucial for many tasks including classification of CRISPR subtypes, based on sequence and structure conservation, the location and sequence of the leader region, the identities of target sites (i.e. protospacers) on invading genetic elements, and the locations and identities of protospacer-adjacent motifs (PAMs).

CRISPR repeats are unique in that they play essential roles in each of the three main steps of the different immune responses. During adaptation, they are recognised and duplicated during de novo spacer insertion. In addition, the repeat transcripts are specifically cleaved during the processing reactions and, moreover, recognition of the repeat sequence at the 5'-end of the crRNA is essential for the interference reactions (see Subsection 2.4.5). Thus, despite their relatively short lengths, each repeat carries essential structure and/or sequence motifs that are recognised by enzymes or proteins complexes. However, the repeats are very heterogeneous. They occur in lengths between 19 and 48 nt, and show considerable sequence diversity.

The knowledge of orientation is crucial since only this information allows one to determine the actually active transcript. Various criteria have been used. For example, when a detectable CRISPR leader sequence is present, it defines the repeat orientation. However, these leaders tend to vary extensively in sequence and size, and are often difficult to detect. Moreover, many published CRISPR loci appear to lack leaders. Another criterion is the direction of transcription, with transcription initiation generally occurring predominantly from the leader end of the CRISPR array. However, to date very few systems have been studied experimentally.

In 2014, the first bioinformatic tool to predict the orientation of CRISPR arrays was introduced in [154]. In their model, a linear predictor was developed based on series of

3.3. CRISPRstrand: Predicting repeat orientations that determine the strand from which crRNAs are processed at CRISPR loci

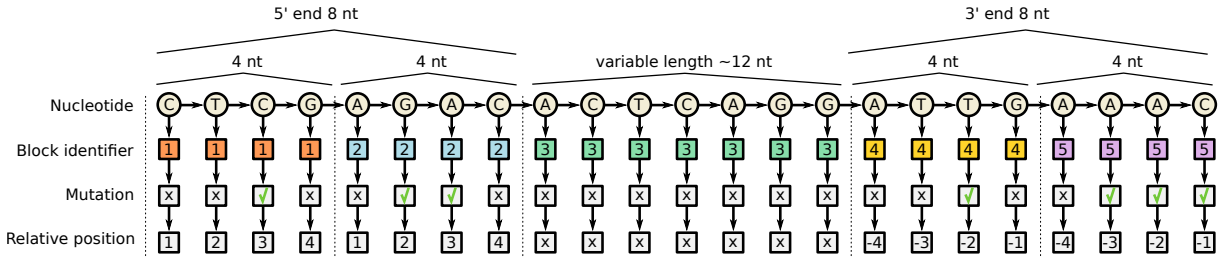


Figure 3.3: Graph encoding the consensus repeat sequence. The consensus nucleotide information is represented as a path graph, and additional information is modelled as a chain of additional vertices. The terminal parts of the repeat are marked with block identifiers. The figure is taken from *Publication 2*.

features: (1) the existence of an ATTGAAAN motif in CRISPR repeats; (2) a higher A or T content in the flanking regions of CRISPR arrays, nucleotide composition within the CRISPR array; (3) the presence of mutations in specific parts of the array; and (4) the tendency to fold into a secondary structure. Each feature is considered as an independent predictor and is given a weight proportional to its estimated precision. The final prediction is computed as the weighted combination of each predictor.

3.3.2 Discussion and Results summary

CRISPR repeat-orientation provides valuable information for detecting leader sequences and classifying CRISPR repeat conservation. Furthermore, it helps identify target sites (protospacers) on invading genetic elements and characterise PAM motifs. We compiled a comprehensive dataset of CRISPR arrays from published archaeal and bacterial genomes (> 4,700 CRISPR arrays in around 4899 genomes). We developed *CRISPRstrand* a tool that uses a graph-kernel machine learning approach to determine the crRNA encoding strand at a CRISPR locus. There are two core ideas underlying our algorithm. The first one is to use a combinatorial technique to extract a very large number of features. The second idea is to encode our knowledge about the problem as a directed graph with discrete labels. The first idea allows a predictive system to be very accurate and to express complex discriminative decisions; the second idea allows a natural and flexible encoding of background knowledge. We encoded all our intuitions and knowledge on the relevant signals in a graph data structure. The reason for this choice is two-fold: First, we want an easy and natural way to inject different types of information in the problem solution, and second, we want to exploit efficient techniques existing in the machine learning literature to automatically construct a large number of derived features to improve the accuracy of predictive models (see Figure 3.3).

We carefully compared our method (*CRISPRstrand*) with the state-of-the-art tool [154]. Both methods were applied to the same test data set, filtered for decreasing levels of sequence identity w.r.t. the training set. We report the comparative AUC ROC (area under the curve of the receiver operator characteristic) performance and observe that *CRISPRstrand* offers

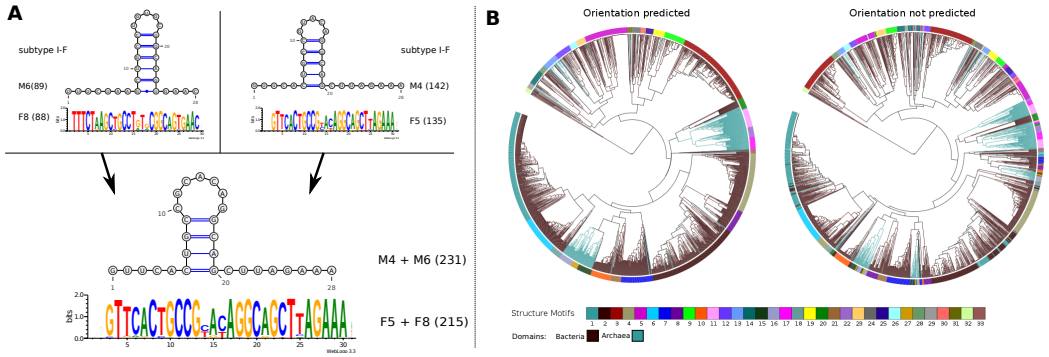


Figure 3.4: A: Given the novel predicted orientation Family 5 with Family 8 and Motif 4 with Motif 6 could be merged. **B:** The 33 structural motifs from [7] are clustered (1) with the orientation prediction; (2) without orientation prediction. The figure is taken from *Publication 2*.

a substantial improvement, both in prediction performance and in generalisation capacity, with a less pronounced degradation as the sequence identity decreases. Furthermore, we measured the runtime for both methods on 956 CRISPR arrays (average length 28 nucleotides). The classification task was completed in 59 seconds by *CRISPRstrand* and in 37 minutes by the state-of-the-art predictive model [154]. We report that the state-of-the-art tool failed to make any prediction in 98 cases out of 948, whereas *CRISPRstrand* achieved an AUC ROC of 0.89 on the same instances.

Finally, we have applied *CRISPRstrand* to identify the transcribed strand for the set of 3,527 repeats available from [7] and for the novel set of 4,719 individual CRISPR arrays. We have identified 536 repeats with incorrect orientation out of 3527 repeats from [7]. Next, we ran our *CRISPRmap* cluster pipeline, retrieving 29 potential structural motifs and 37 conserved sequences families. As shown in Figure 3.4, the orientation of F8 and M6 was incorrect. Using corrected orientations, we could merge F8 with F6, and M5 with M6. Overall, in Figure 3.4, we show how the cluster quality can be significantly improved when we can make use of a better orientation prediction.

In conclusion, we developed and tested *CRISPRstrand*, an accurate tool to predict the transcribed strand of CRISPR arrays based on an advanced machine learning approach. *CRISPRstrand* is motivated by recent findings and encodes the most relevant information in the form of a graph structure that can be efficiently processed with graph kernel methods. In addition, we showed that accurate orientation information greatly improved detection of conserved repeat structure motifs and sequence families. Furthermore, we achieved up to 0.95 AUC ROC, compare to about 0.88 AUC ROC for the the state-of-the-art tool [154]. Finally, we integrated CRISPRstrand predictions into CRISPRmap web server of CRISPR conservation to improve the accuracy of the previously published classification of CRISPR repeats, and resulted in: (i) a comprehensive dataset with >4500 consensus repeats; (ii) the most recent classification of Cas subtypes based on Cas-protein occurrences for archaea[76]; and (iii) an improved annotation Cas subtypes for bacteria respecting the rules published in[68].

3.4 CRISPRleader: Characterising leader sequences of CRISPR loci

Here, we focus on the detecting and characterising CRISPR leader sequences to improve our understanding of the adaptation phase. This section concludes the work from *Publication 3*, which was published in the Bioinformatics journal.

3.4.1 Motivation

The first phase of CRISPR-Cas immunity is called adaptation (see Subsection 2.4.5), in which small DNA fragments are excised from genetic elements and are inserted into a CRISPR array generally adjacent to its so-called leader sequence at one end of the array (see Subsection 2.4.2). It has been shown that transcription initiation and adaptation signals of the CRISPR array are located within the leader. However, apart from promoters, there is very little knowledge of sequence or structural motifs or their possible functions. Leader properties have mainly been characterised through transcriptional initiation data from single organisms but large-scale characterisation of leaders has remained challenging due to their low level of sequence conservation.

The existence of adaptation signals in the leader region is also supported by evolutionary studies. Despite their relatively low sequence conservation, sequence clustering studies for the *Sulfolobales* have shown that the leaders tend to coevolve with its CRISPR repeat, the adaptation module (Cas1, Cas2 and Cas4) and the protospacer-adjacent motif (PAM) [52]. Experimental support for this inference was provided by studies on the *E. coli* type I-E system [56]. Interestingly, leader sequences also carry conserved sequence motifs, currently of unknown function [155]. The latter are possibly involved in aligning multiple RNA polymerase complexes for CRISPR transcription, or in assembling Cas proteins adjacent to the CRISPR adaptation site [26, 63, 84, 156, 157].

In general, CRISPR arrays are preceded by the leader region and leader sequences are thus directly adjacent to new spacer integration sites. The leaders vary in size, extending from 47 base-pairs in some bacteria to a few hundred base-pairs in some hyperthermophilic archaea, and they tend to exhibit long regions of low complexity sequence, with limited sequence conservation [52]. Due to their limited sequence conservation, even between very similar archaea and bacteria, no bioinformatic tool exists that can automatically annotate leaders and define their boundaries.

Although most CRISPR-Cas loci carry a leader region, a few experimentally characterised loci lack a leader. These leaderless loci do not acquire new spacers, which is consistent with lack of CRISPR-Cas adaptation signals [26, 48]. However, they do still yield processed CRISPR RNAs (crRNAs), presumably as a result of transcription from promoters taken up randomly in spacers [57, 58]. Moreover, in some crenarchaea, leaderless CRISPR loci are also characterised by their invariant DNA structure. Whereas CRISPR loci with leaders undergo spacer acquisition and are susceptible to periodic deletions and rearrangements,

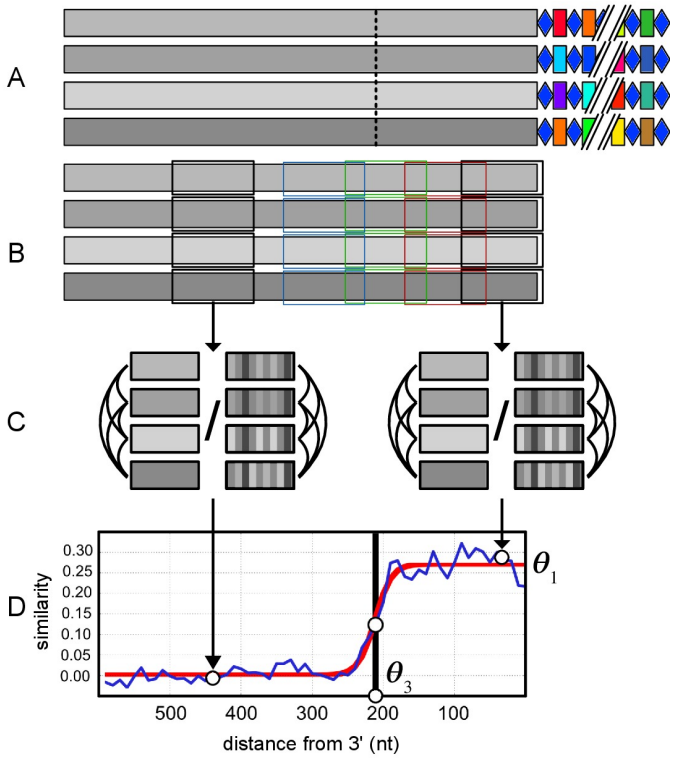


Figure 3.5: Leader boundary identification: (A) Leader sequences are clustered together according to the similarity between the associated repeat sequences; the 3' end of the sequences in a cluster is aligned w.r.t. the first CRISPR repeat (B) Shifting windows spanning the same positions are extracted. (C) The average pairwise similarity between all subsequences in a window is computed using the proposed string kernel; the same procedure is applied to shuffled sequences to compute the log odds ratio (D) A saturating function is fitted to distinguish the highly conserved region from the non-conserved one; the point of maximum slope θ_3 is returned as leader boundary. The figure is taken from *Publication 3*.

resulting from recombination between direct repeats. Leaderless CRISPR loci are both resistant to CRISPR adaptation and structurally invariant [26, 48, 51].

3.4.2 Discussion and Results summary

To improve our understanding of the *adaptation* stage, we studied leader sequences in more detail. We developed *CRISPRleader*, an efficient approach to determining CRISPR leader boundaries by focusing on leader sequence conservation within groupings based on the similarity of the repeats in the adjacent CRISPR arrays. *CRISPRleader* utilises a string-kernel technique that can capture more information than traditional sequence alignments and is especially capable of detecting a collection of local motifs.

We have used a comprehensive data set of CRISPR loci of archeal and bacterial genomes downloaded from CRISPRmap webserver [7, 8]. As mentioned above, the leader sequence co-evolves with CRISPR repeats, with the Cas1 protein and with the PAM motif. To make use of this evolutionary information, we introduce the notion of a leader cluster. We group

together leader sequences not based primarily on their sequence similarity, but rather based on the similarity of the associated repeat sequence. By doing so, we overcome the problem of the limited sequence similarity of leaders. To group repeat sequences we follow the approach presented in CRISPRmap [7, 8]. In more detail, given a CRISPR array, we first computed the consensus-repeat sequence by aligning all repeat sequences without gaps and then take, for each position, the most frequent nucleotide. We have defined the similarity between two consensus repeat sequences as the global pairwise alignment score computed. To obtain coherent sets, we then apply Markov Clustering [151]. In *CRISPRmap*, it was found that better results can be obtained if the similarity matrix is thresholded, i.e. if we set to 0 the similarity value for pairs of repeat sequences that are not sufficiently similar. We optimised these parameters to guarantee that archaea and bacteria are always placed in distinct clusters. Due to evolution-related adaptation signals, the leader sequence will likely exhibit detectable signals of sequence conservation (see Figure 3.5).

In our work, we showed that the region of sequence conservation in leaders tends to extend further upstream of the CRISPR locus when similar leaders are compared within the same genome or between closely related strains of the same species. In contrast, when comparing similar leaders across different species, the conserved regions end closer to the CRISPR locus. Thus, we define the former as the extended leader and the latter as the core leader (see Figure 3.6). The sequence conservation in the CRISPR-distal regions of the extended leader is likely to have resulted from relatively recent duplication events. The core leader, on the other hand, tends to be well conserved, even for divergent hosts, which implies that only the core region is of special functional significance. In addition, we find numerous archaeal leader clusters that are shared between several species and genera, but seldomly cross the order boundary. In contrast, bacterial leader clusters are much more diverse taxonomically.

To conclude, we developed an efficient tool for determining the CRISPR leader boundaries called *CRISPRleader*. *CRISPRleader* provides a full annotation of the CRISPR array, its strand orientation as well as conserved core leader boundaries that can be uploaded to any genome browser. In addition, we provided reader-friendly HTML pages for conserved leader clusters. We analysed 1,426 archaeal and bacterial genomes using *CRISPRleader* and identified several characteristic properties of the leader sequences. Results show that although an extended region can be conserved between few very closely related species or CRISPR loci, generally, a smaller core leader region, directly adjacent to the CRISPR locus, is conserved between more distantly-related species. Furthermore, we identified core leaders from 770 archaeal and 2,224 bacterial CRISPR loci and observed significant differences between leader clusters. First, core leaders tend to be longer in archaea than in bacteria. Second, leader clusters in archaea are more homogeneous in terms of phyla than in bacteria. This may reflect the fact that archaea have survived primarily in low-energy environments, which are often quite isolated (e.g. solfataric fields or hypersaline lakes) such that genetic exchange is much more limited than for most bacteria. Third, bacteria exhibit more indels

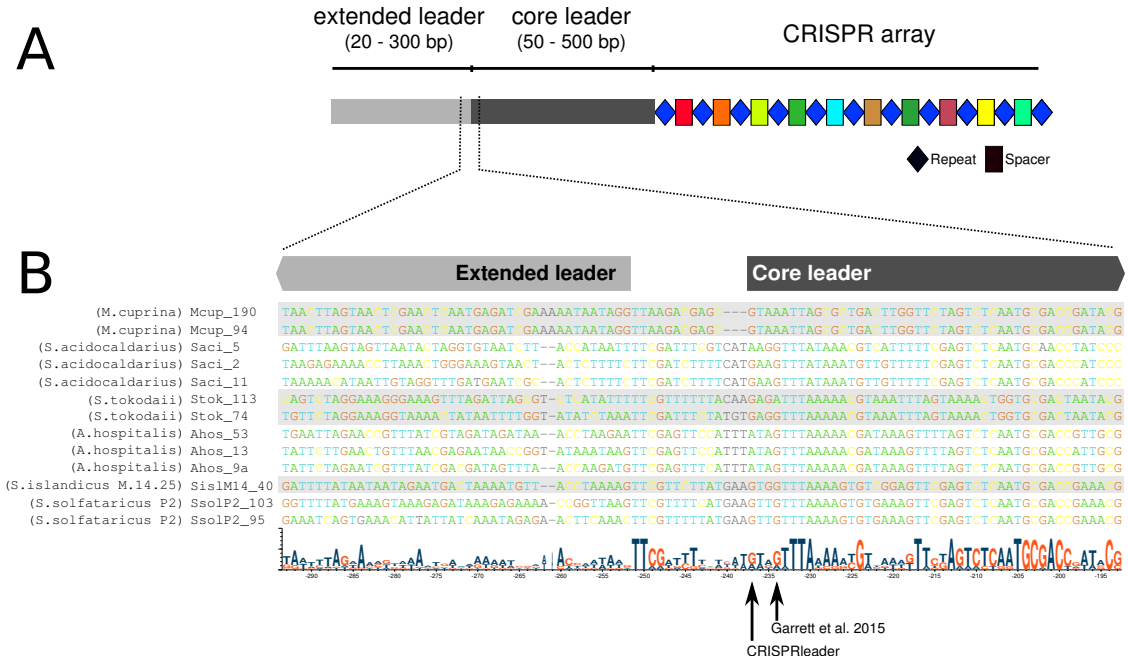


Figure 3.6: (A) Schematic view of the elements of a CRISPR array showing the repeats (blue diamonds) and spacers (coloured rectangles) of a CRISPR array and the leader region, which we separate into a core and an extended leader. The *core* leader is generally conserved across different host species and is shorter than the *extended* leader which is normally only conserved between multiple leader copies in the same genome. (B) Sequences correspond to a cluster of related leaders shared between species of the genera Acidianus, Metalosphaera and Sulfolobus. Each leader is identified by the number of repeats in the adjacent CRISPR. *CRISPRleader* predicts the length of the core leader, since the extended leader is assumed to be functionally less important. In the bottom we provide and example of a leader alignment to show a detailed view at the junction between the core and extended leader. Here it is possible to see how the extended part is only conserved between multiple copies in the same organism. In contrast, the core part is conserved across all of the different hosts, is underlined by the sequence logo below. The leader boundary predicted by *CRISPRleader* and the boundary determined by expert inspection are indicated by black arrows at the bottom. The figure is taken from *Publication 3*.

in the CRISPR-proximal region of the core leaders than archaea. This core leader region has been shown to be important for CRISPR transcription and CRISPR-Cas adaptation and may be readily inactivated, or modulated, by indel activity, possibly triggered by an invader to circumvent targeting. Finally, we showed that both archaea and bacteria (1) have leader sequences and repeats that tend to coevolve with the Cas1 protein more broadly than previously believed, i.e., irrespectively of the systems subtype; and (2) leaderless CRISPR loci tend to be much smaller than loci with a leader present. This is possibly indicative of a displacement event from the leader-distal ends of other CRISPR loci. Leaderless CRISPR loci have been shown not to undergo adaptation but can still contribute to crRNA-directed interference. The lack of adaptation is also consistent with their smaller size.

3.5 Evolutionary classification of CRISPR-Cas systems of archaeal and bacterial adaptive immunity

The problem we will tackle here is the classification and automated annotation of the CRISPR subtypes. This section summarises the work from *Publication 4*, which was published in the journal Nature Reviews Microbiology.

3.5.1 Motivation

In general, the CRISPR locus is flanked by highly diverse *cas* (Clustered-associated) genes that encode Cas proteins (see Subsection 2.4.4). Most of the Cas proteins evolve rapidly, which complicates their classification into families [65, 66]. The diversity of the Cas protein sequences is matched by the remarkable variation in the genomic architecture of CRISPR-Cas loci. Thus, consistent annotation of the Cas proteins and classification of the different CRISPR-Cas systems is a major challenge [68]. Nevertheless, such classification is essential for expedient and robust characterisation of the CRISPR-Cas loci in new genomes for facilitating further progress of CRISPR research.

Owing to the very diverse sets of Cas proteins that exist, classification of CRISPR-Cas systems based on Cas proteins has been defined to be a difficult problem to tackle. Although Cas proteins and CRISPR arrays are very important elements of the CRISPR-Cas system and generally are located close together, there is no evidence showed for a one to one relationship between Cas proteins and the type of CRISPR repeat that recognised by those proteins. Furthermore, CRISPR-Cas systems are deeply affected by modularity and interchangeability of Cas proteins [59]. Therefore, classification of CRISPR-Cas systems based on Cas proteins is a very important and difficult issue that needs to be solved to investigate the differences between the CRISPR-Cas systems from archaeal and bacterial.

3.5.2 Discussion and Results

The classification of CRISPR-Cas proteins is based on comparative analysis of the CRISPR-Cas loci of all available genomic data. Identification of the CRISPR-Cas proteins is a non-trivial task due to high sequence variability. For this reason, we have developed a library of profiles, or more precisely, position-specific scoring matrices (PSSM) [159] for all known protein families associated with CRISPR-Cas systems. Then, the Cas loci were identified based on several criteria. Next, we separated Cas proteins related to interference from Cas proteins related to adaptation, since the current published subtyping is mostly based on the interference process. We based our classification on a new similarity between two sets of interference proteins, which is defined as the average of pairwise normalised similarity between all protein pairings between the two sets (see Figure 3.7). Clustering based on this similarity is able to faithfully reproduce the existing manual CRISPR annotation.

Given the rapid pace of microbial genome sequencing, we constructed a classifier tool for

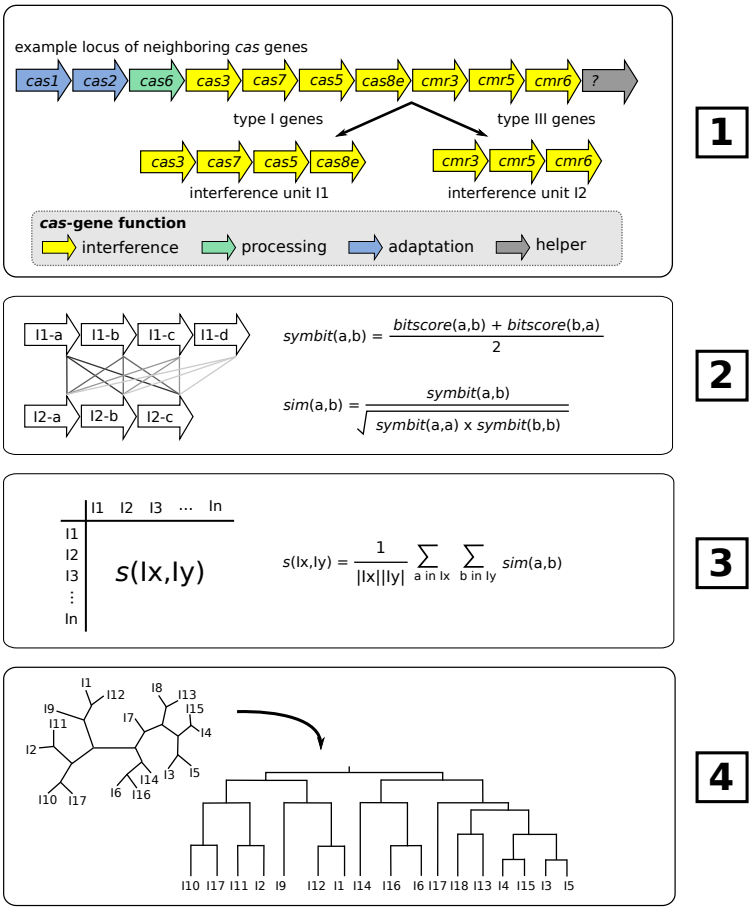


Figure 3.7: Clustering of the CRISPR-Cas crRNA-effector modules by protein sequence similarity. Given variability of the gene composition and architecture of the effector modules and the lack of a universal marker suitable for phylogenetic analysis, we perform the following steps: (1) Select only genes encoding interference proteins and split into separate interference units (encoding CASCADE complexes or just Cas9). (2) Perform a pairwise protein sequence comparison with FASTA [158]; make bit scores symmetrical; and normalise resulting bit scores. (3) Calculate similarity matrix for all pairs of interference units. The matrix is symmetric. (4) Calculate unrooted tree with the Rapid Neighbour-Joining method. Root tree so that interference units of types I, II and III, IV, V come together in the tree. The figure is adopted from *Publication 4*.

automated annotation of *cas* loci subtypes for newly sequenced genomes, which would be valuable for numerous researchers, especially those that do not specialise in CRISPR-Cas systems. Although a careful inspection of combined features is required for accurate subtype annotation, we show that an automated annotation method based on protein-sequence similarity of effector modules (i.e., the similarity notion underlying the dendrogram, see Figure 3.7) is able to faithfully reproduce the manual subtype classification. Using prior information on the association between and CRISPR-Cas loci and the corresponding classification of the effector modules, we developed a k-nearest neighbour classifier that achieved 0.998 accuracy, i.e. only 4 loci out of 1942 were assigned an incorrect subtypes. These results are stable w.r.t. the number of neighbours, although the specific accuracy depends on the level of sequence similarity of the analysed Cas proteins to those available in the

3.5. Evolutionary classification of CRISPR-Cas systems of archaeal and bacterial adaptive immunity

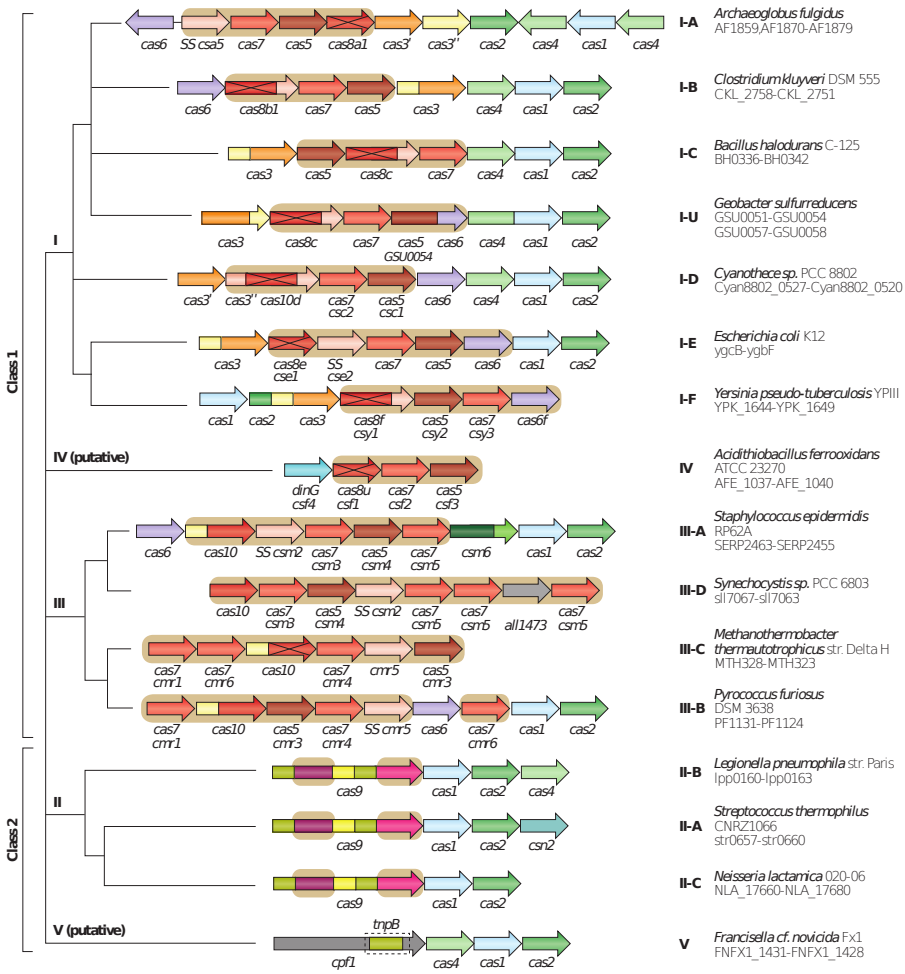


Figure 3.8: Architectures of the genomic loci for the subtypes of CRISPR-Cas systems. Classic operon organisation are shown for every type and subtype of CRISPR-Cas system. The particular gene locus tag names are presented for each genome. The color-coded homologous genes are recognised by a family name. The small subunit is programmed by either csm2, cmr5, cse2 or csa5; no all-inclusive name was proposed to cooperatively define the gene family to date. Inactivation of the respective catalytic sites was presented by crosses through genes encoding the large subunit (Cas8 or Cas10 family members). Genes and gene regions encoding elements of the interference module (CRISPR RNA (crRNA)-effector complexes or Cas9 proteins) are decorated with a beige background. The adaptation module (cas1 and cas2) and cas6 are expendable in subtypes III-A and III-B; especially, they are seldom existing in subtype III-B (dashed lines). Dark green signifies the CARF domain. Gene regions coloured cream characterise the HD nuclease domain; the HD domain in Cas10 is different from that of Cas3 and Cas3". Also coloured are the areas of cas9 that unevenly correspond to the RuvC-like nuclease (lime green), yellow for HNH nuclease, purple recognition lobe, and protospacer adjacent motif (PAM)-interacting domains (pink). The areas of cpf1 apart from the RuvC-like domain are functionally uncategorised and are presented in grey, as is the functionally uncategorised all1473 gene in subtype III-D. This figure is adopted from *Publication 4*.

modelling phase, and predictably drops when the variants are only distantly related to the existing subtypes.

The adaptation module (Cas1, Cas2 and Cas4) evolved, to a large extent, independently of

the operational modules (in particular, crRNA-effector complexes) of CRISPR-Cas systems. This is in agreement with the hypothesis of the origin of the system as the result of the integration of a casposon-like mobile element next to an operon encoding a stand-alone effector complex. The dynamic, modular evolution of CRISPR-Cas is also manifested at the level of the architecture of Cas loci and the combination of different families of CRISPR arrays with different Cas system. However, a complementary trend is the frequent horizontal transfer of complete CRISPR-Cas loci, which confers a degree of coherence to these systems and ensures that there is almost no congruence between the evolution of CRISPR-Cas and the species phylogeny. The dynamic and modular evolution of CRISPR-Cas is manifested also at the level of the Cas loci architecture and the combination of different classes of CRISPR arrays with Cas loci.

To conclude, we introduced two classes of CRISPR-Cas systems as a new, top level of classification and define two putative new types and five new subtypes within these classes, resulting in a total of five types and 16 subtypes (see Figure 3.8). We employ this classification to analyse the evolutionary relationships between CRISPR-Cas loci using several measures. The results of this analysis highlight pronounced modularity as an emerging trend in the evolution of CRISPR-Cas systems. Finally, we demonstrate the potential for automated annotation of CRISPR-Cas loci by developing a computational approach that uses the new classification to assign CRISPR-Cas system subtype with high precision.

3.6 Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs in cyanobacteria

This section summarises the work from Publication 5, which was published in the Nucleic Acids Research journal

3.6.1 Motivation

Cyanobacterial model *Synechocystis* sp. PCC6803 has one main chromosomal and seven extrachromosomal elements (plasmids). Only one plasmid encodes a CRISPR-Cas system. The plasmid *pSYSA_M* of *Synechocystis* sp. PCC 6803 consists of three CRISPR loci, which are located in the forward strand namely: CRISPR1, CRISPR2 and CRISPR3. According to the current classification [10], CRISPR1 is classified as subtype I-D system, whereas CRISPR2 and CRISPR3 are classified III-D and III-B systems respectively. CRISPR1 consists of 17 repeat-spacer units, 55 repeat-spacer units for CRISPR2, and 37 repeat-spacer units in CRISPR3 according to CRISPRmap [7, 8].

Upstream of the CRISPR1 and CRISPR2 locus, three Cas proteins were identified that have high sequence similarity to the Cas6 endoribonuclease family: *slr7014*, *slr7068* and *sll7075* [11, 144]. Based on their location in the genome, we called them Cas6-1, Cas6-2a and Cas6-2b respectively. As already mentioned, CRISPR locus is transcribed into pre-crRNA and then are processed crRNAs which are comprised of a single invader-targeting sequence and a short segment of the repeat sequences. In Type I and III, generally Cas6 proteins are involved in the processing of pre-crRNA into crRNA fragments (see Subsection 2.4.5). Cas6 proteins have the ability to bind and cleave pre-crRNA containing CRISPR repeats that are able to form hairpin structures or unstructured repeats [53, 66, 68, 89, 92, 93, 160].

In 2013, RNA-seq analysis of *Synechocystis* sp. PCC 6803 determined a putative cleavage site within the repeat sequences of CRISPR1 and CRISPR2 [144]. The authors showed that the cleavage site generates in both cases intermediate crRNAs with a length of 68-83 nt consisting of a single spacer sequence as well as an 8-nt-repeat handle at the 5'-end and a 29 nt repeat handle at the 3'-end. Furthermore, they showed that the mature crRNAs is possibly shorter according to *vivo* data from northern hybridisations and RNA-seq. In CRISPR1, the length of mature crRNAs are 39 and 45 nt and 36 or 37 nt for CRISPR2 [144]. Thus, in a second, so far uncharacterised step the crRNA intermediates are processed further into the mature crRNAs. Such a further processing by an unknown trimming nuclease that removes 3' portions of the crRNA is also known from several Type III and at least one subtype I-A system [4, 100, 107].

3.6.2 Discussion and Results summary

We determined, biochemically, that Cas6-1 from the subtype I-D system and Cas6-2a from the subtype III-D system are the endoribonucleases that process crRNA for CRISPR1 and

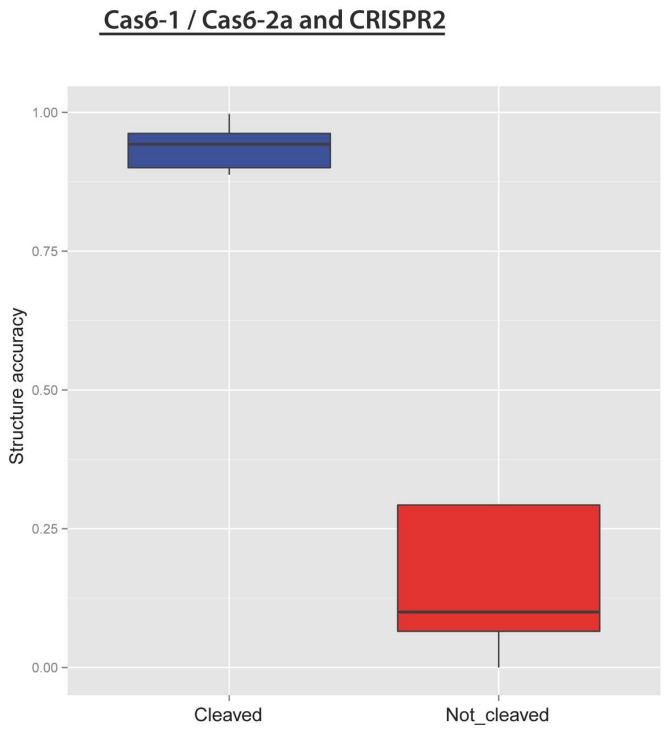


Figure 3.9: Boxplot: The structure stability of the functional CRISPR2 hairpin motif measured as the base pair accuracy (y-axis), is compared between repeat instances that were cleaved (blue) and not cleaved (red) by Cas6-1 and Cas6-2a in the *in vitro* experiments. Clearly, high base pair accuracies correspond to successful cleavage events, whereas low base pair accuracies explain repeats that were not cleaved. For both enzymes only 3 out of 25 experimental notices were not clarified by the base pair accuracy. The figure is taken from *Publication 5*.

CRISPR2, respectively. For both enzymes, we identified a promiscuity to process not only their cognate CRISPR1 or CRISPR2 transcripts, but further, to cleave the transcripts from the other locus as well *in vitro*. This is completely in contrast to the *in vivo* specificity of these enzymes found in the analysis of deletion mutants [144]. In addition, cleavage of the non-cognate substrates was less efficient and not all possible cleavage sites were recognised.

We characterised the ectopic Cas6-1 mediated processing of CRISPR2 transcripts by systematic substrate variation. In addition, we tested whether Cas6-2a could possibly mediate processing of CRISPR1 transcripts as well. For CRISPR1 and CRISPR2 arrays, we generated nine RNA fragments with a different number of repeats and spacers (see Figure 3.10A). These fragments I-IX were incubated *in vitro* in the presence or absence of Cas6-1 or Cas6-2a and resulting cleavage fragments were analysed by denaturing gel electrophoresis. Detected fragment sizes were consistent with the expected lengths when assuming a cleavage 8 nt upstream of the 3' end of each repeat instance in the CRISPR1 or CRISPR2 fragments. Both enzymes delivered very similar patterns for the respective substrates, suggesting that the identical sites were recognised and cleaved. However, for both enzymes, we recognised that for CRISPR1 all but for CRISPR2 not all theoretically possible fragments were observed, consistent with the idea that they could generate some but not all of the theoretically possi-

ble products. The presence of potential contaminating RNase activities in the preparations is considered very low because there was no RNA processing or degradation in parallel incubations with empty-vector mock preparations.

In [144], the authors suggested that adjacent spacer sequences can influence the formation of the repeat-structure motif. For this reason, we provide bioinformatics analysis of the nine *in vitro* experiments, and test whether surrounding sequence context influences Cas6-1 and Cas6-2a cleavage of CRISPR1 and CRISPR2 transcripts. For each fragment, we calculated the accuracies of the local functional repeat motifs, each representing a subsequence of the original CRISPR1 or CRISPR2 array with a different number of repeats and spacers (see Figure 3.9). Results showed clearly that the accuracy of the functional local repeat structure is significantly lower for the non-cleaved compared to the cleaved fragments. Furthermore, we chose the CRISPR2 repeats R6 and R7 within fragment VIII as an example. We observed that repeat R7 was cleaved in this fragment, whereas repeat R6 was not cleaved as part of the same fragment VIII (see Figure 3.10A), indicated by the lack of the 123, 76 and 67 nt fragments for CRISPR2-VIII. In order to detect the effect for all cleaved and non-cleaved fragments, we averaged the dot plots (plus a context of 35 nt) of repeats that are cleaved or not cleaved in the fragments of CRISPR1 and CRISPR2. Results showed that among uncleaved fragments there are many more base pairs in the surrounding context than among cleaved fragments, whereas the base pairing of the functional motif has a much higher average probability for the cleaved fragments.

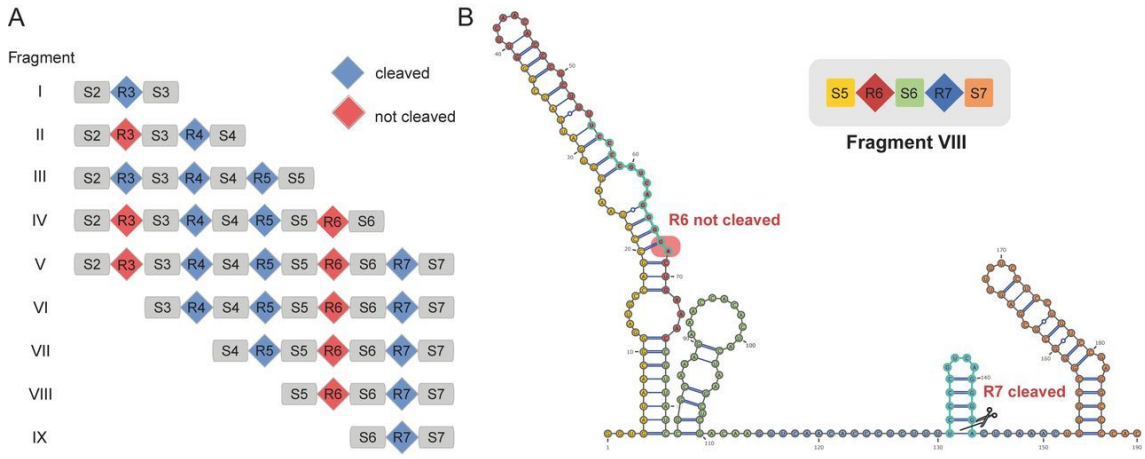


Figure 3.10: Methodical analysis of CRISPR2 cleavage by Cas6-1. (A) Overview of full length CRISPR2 transcripts and positions of cleavage by Cas6-1 as determined by the experiment. (B) Prediction of a global MFE structure to determine the most probable structure for the complete CRISPR2 fragment VIII. We have indicated the positions covered by the local functional repeat structure in turquoise and the remaining repeat sequence in red (R6) or blue (R7). Spacers are coloured in yellow (S5), green (S6) or orange (S7). The local functional repeat structure is formed in the cleaved repeat R7, whereas the associated position is blocked by other stems in the non-cleaved repeat R6 of fragment VIII. The figure is taken from *Publication 5*.

In conclusion, a first study provided biochemical analysis of pre-crRNA processing by Cas6 proteins in cyanobacteria and in a subtype I-D CRISPR-Cas system. We showed that Cas6-1 and Cas6-2a enzymes are able to cleavage *in vitro* transcript RNAs that are derived from CRISPR1 and CRISPR2. We address a promiscuity of both enzymes to process *in vitro* not only their cognate transcripts, but also the respective non-cognate precursors, whereas they are specific *in vivo*. Furthermore, while most of the repeats serving as substrates were cleaved *in vitro*, some were not. Finally, based on RNA structure predictions, we showed that the context sequence surrounding a repeat can interfere with its stable folding. Structure accuracy calculations of the hairpin motif explained the vast majority of analysed cleavage reactions making this a good measure of structure stability and for predicting successful cleavage events. The influence of surrounding sequences might partially explain variations in crRNA abundances and should be considered when designing artificial CRISPR arrays.

Chapter 4

Conclusion

In this thesis, we have developed novel approaches that address the problem of automated characterisation and structural analysis of CRISPR-Cas systems. These methods were provided in a set of applications to support the analysis and characterisation of CRISPR-Cas systems. Specifically, for elucidating CRISPR sequence and structure properties and their evolutions, and the automated annotation of associated Cas proteins and elements.

In the first part, we provided a very comprehensive analysis of CRISPR structure and sequence conservation based on all publicly available CRISPR repeat sequences. In particular, we clustered the CRISPR repeats into classes based on sequence and structure similarity and eventually to recognise binding motifs and patterns of associated Cas proteins. In addition, we verified our methods by comparing results with CRISPR-Cas systems where the crRNA maturation mechanism has been characterised experimentally. The published structure was consistent with our classified structure motifs and subtype annotations. Furthermore, we showed evidence of horizontal transfer of CRISPR-Cas systems between archaeal and bacterial genomes. Finally, we developed a web server called *CRISPRmap* which provides an easy access to our data and useful resource for investigating the conservation and diversity of repeat sequences in CRISPR-Cas systems.

In the second part, we presented a highly flexible approach to accurately predict the transcribed strand of CRISPR loci. The novel method enabled us to encode the most relevant information in the form of a graph structure which can be efficiently processed with graph kernel methods. In addition, we showed that accurate orientation information greatly improved detection of conserved repeat sequence families and structure motifs. Furthermore, *CRISPRstrand* predictions were integrated into our *CRISPRmap* web server of CRISPR conservation, which was updated later to version 2.0. Finally, *CRISPRstrand* is fast, accurate and can be easily integrated into existing pipelines. In future work, we will employ it to enhance the identification of novel targets (protospacers), PAM motifs and the investigation of regulatory motifs in the leader sequences of CRISPR arrays.

In the third part, we developed a novel method *CRISPRleader*, an efficient approach to determine CRISPR leader boundaries by focusing on leader sequence conservation within groups based on the similarity of the repeats in the adjacent CRISPR arrays. *CRISPRleader*

utilised a string-kernel technique that can capture more information than traditional sequence alignments and is especially capable of detecting a collection of local motifs. We built specialised HMM models for each of the 51 and the 144 CRISPR-leader clusters from archaea and bacteria, respectively. In addition, we defined leaderless CRISPR arrays based on a few criteria. Our results demonstrated that 13% of 980 archaeal CRISPR loci, and 24% of 2852 bacterial loci, could be considered leaderless. Furthermore, we showed that in both archaea and bacteria, leader sequences and repeats tend to coevolve with the Cas1 protein more broadly than previously believed, i.e. irrespectively of the system's subtype. Finally, *CRISPRleader* accepts either a complete or partial genome sequence as input and provides a full annotation of CRISPR arrays, and their strand locations as well as conserved core leader boundaries which can be uploaded to any genome browser.

In the fourth part, we provided a very comprehensive CRISPR-Cas classification which classifies archaeal and bacterial CRISPR-Cas systems into two main classes (class I and II) based on interference module similarities. These classes are subdivided into six Types which are further divided into sixteen subtypes. In particular, we presented a method for annotation of locus subtypes that relies on a novel similarity notion for the interference modules. Our method is achieved by nearest neighbour classification, which yields highly consistent results with respect to the current subtype classification. Finally, we investigated the evolutionary modularity of adaptation and interference modules.

In the last part, we presented evidence that the context surrounding a repeat instance can influence stable structure motif and therefore can sequester the cleavage reaction. In particular, we performed the analysis on a series of *in vitro* experiments. The results suggested the successful cleavage to depend on the stable formation of a hairpin motif, which was similar for the two CRISPR loci studied. In addition, we showed that the sequences of adjacent spacers can lead to alternative structures that the inhibit structure motif and thus are incompatible with the cleavage reaction. Finally, the influence of surrounding sequences might lead to variations in crRNA abundances and therefore we proposed this to be taken into account when designing artificial CRISPR arrays.

Publications

This chapter contains the publications on which this thesis is based. They are listed here in a chronological order that is not the same order in which they are presented in Chapter 3. As all these publications are joint works with other authors, I provide a statement of contribution for each one of them to show the level of my contribution to the corresponding publication.

CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems

Sita J. Lange, **Omer S. Alkhnbashi**, Dominic Rose, Sebastian Will and Rolf Backofen. **Nucleic Acids Research Journal, 2013**, doi:10.1093/nar/gkt606.

Personal contribution

I have done a major contribution in this project, which led to share a first authorship with Sita J. Lange and Dominic Rose. I performed the software implementation, data acquisition, and some data analysis. Furthermore, I helped in designeing and implementing the CRISPRmap webserver.

Omer S. Alkhnbashi

The following co-authors confirm the above stated contribution.

Sita J. Lange

Dominic Rose

Sebastian Will

Rolf Backofen

Nucleic Acids Research Advance Access published July 17, 2013

Nucleic Acids Research, 2013, 1–11
doi:10.1093/nar/gkt606

CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems

Sita J. Lange[1], Omer S. Alkhnbashi[1], Dominic Rose[1], Sebastian Will[1] and Rolf Backofen[1,2,3,4,*]

[1]Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany, [2]ZBSA Centre for Biological Systems Analysis, Albert-Ludwigs-University Freiburg, Habsburgerstr. 49, 79104 Freiburg, Germany, [3]BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, Albert-Ludwigs-University Freiburg, Germany and [4]Center for non-coding RNA in Technology and Health, University of Copenhagen, Gronnegardsvej 3, DK-1870 Frederiksberg C, Denmark

Received May 7, 2013; Revised June 7, 2013; Accepted June 17, 2013

ABSTRACT

Central to Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-Cas systems are repeated RNA sequences that serve as Cas-protein–binding templates. Classification is based on the architectural composition of associated Cas proteins, considering repeat evolution is essential to complete the picture. We compiled the largest data set of CRISPRs to date, performed comprehensive, independent clustering analyses and identified a novel set of 40 conserved sequence families and 33 potential structure motifs for Cas-endoribonucleases with some distinct conservation patterns. Evolutionary relationships are presented as a hierarchical map of sequence and structure similarities for both a quick and detailed insight into the diversity of CRISPR-Cas systems. In a comparison with Cas-subtypes, I-C, I-E, I-F and type II were strongly coupled and the remaining type I and type III subtypes were loosely coupled to repeat and Cas1 evolution, respectively. Subtypes with a strong link to CRISPR evolution were almost exclusive to bacteria; nevertheless, we identified rare examples of potential horizontal transfer of I-C and I-E systems into archaeal organisms. Our easy-to-use web server provides an automated assignment of newly sequenced CRISPRs to our classification system and enables more informed choices on future hypotheses in CRISPR-Cas research: http://rna.informatik.uni-freiburg.de/CRISPRmap.

INTRODUCTION

Acquired immunity in prokaryotes is directed by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and their associated (Cas) proteins. This CRISPR-Cas system, present in many bacteria and most archaea, recognises and subsequently degrades exogenous genetic elements [for recent reviews see (1–3)]. The adaptive immune response is divided into three main phases: (i) 'Adaptation', the selection of short target segments (protospacers) from foreign DNA and the incorporation of their reverse complement sequence (spacers) into the organism's active CRISPR locus between directly repeated sequences (repeats); (ii) 'crRNA maturation', expression of the CRISPR RNA (a leader followed by an array of repeat-spacer units) and subsequent processing of the transcript into mature RNA species, called crRNA; and (iii) 'target interference', invader DNA (4) or RNA (5,6) degradation at the respective protospacer, guided by the crRNA and a highly specific complex of Cas proteins such as Cmr (5) or Cascade (7).

CRISPR arrays are associated with diverse sets of Cas proteins. Therefore, several global classification systems of Cas subtypes have been introduced (8–10). In the literature, CRISPR-Cas systems are frequently characterised solely by the associated Cas-protein subtypes and relationships between repeats are rarely considered. Although this division into Cas-subtypes is generally effective, an accurate Cas-protein-based classification is complicated: First, CRISPR loci may include novel, chimeric, mixed subtypes or *cas* genes that are missing entirely (10–14). Second, it is not always obvious which *cas* genes are specific to a repeat-spacer array or Cas proteins could be

*To whom correspondence should be addressed. Tel: +49 761 2037460; Fax: +49 761 2037462; Email: backofen@informatik.uni-freiburg.de

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

shared between arrays (13). Finally, many of the cas genes belong to extremely diverse families (8,10).

We provide a comprehensive classification of all publicly available CRISPRs that is based solely on the sequence and structure evolution of repeats. The repeat-spacer array is the only element to be present in all systems and CRISPR-Cas systems are identified first by the existence of such an array. In contrast to the annotation of *cas* genes, repeat-spacer arrays are easily identified by programs such as CRISPRFinder (15) or CRT (16). The repeat is the central regulatory element in the CRISPR-Cas system, as it serves as a binding template for Cas proteins in all three phases of immunity. For these reasons, a systematic repeat-based classification is fundamental to further understand the function, diversity and phylogeny of CRISPR-Cas immune systems.

All clustering approaches are based on pairwise similarities: similarities between repeats are assumed to reflect conserved binding motifs and mechanisms. The binding affinity of Cas proteins is not only affected by the repeat sequence: a small hairpin structure is a key binding motif for Cas endoribonucleases in several systems (17–25). To correctly identify these structure motifs, our clustering is the first that is based not only on sequence—but also on structure—similarities. This approach is well-established for the identification and characterisation of structured non-coding RNA (ncRNA) (26–29). For these ncRNAs, the conservation of structure is often more important than sequence for the biological function (30,31). Although CRISPRs are partially structured ncRNAs, no structure-based clustering exists. To our knowledge, the only CRISPR-specific classification was performed on 349 bacterial and archaeal repeats in 2007 (32). Although structure motifs were identified, the underlying clustering was based purely on sequence and not structure similarity. An analysis of the archaeal domain, also based on only sequence similarities, was done more recently (12).

To provide a complete overview of the conservation of both unstructured and structured CRISPRs, we performed an independent sequence-based clustering to identify conserved sequence families. In addition, we combined identified structure motifs and sequence families with a hierarchical representation of sequence and structure similarities to generate a map that directly reflects relationships between classes and individual CRISPRs. This hierarchical CRISPRmap tree enables a fast comparison between CRISPRs of interest and previously published systems. Automated access to our data via an easy-to-use web server allows users to identify relative positions of both published and unpublished sequences. CRISPRmap is a valuable resource to elucidate and generalise functional mechanisms of CRISPR-Cas immunity.

We rigorously analysed clustering results and observed the following: First, identified structure motifs and automated Cas subtype annotations are consistent with experimentally verified work (18–23). Second, cleavage sites in relation to the structure motifs could be inferred from common features observed in the many articles on crRNA maturation (13,17–21,23–25,33–36). Third, sequence families exhibit varying patterns of repeat sequence conservation. Fourth, some type I and both type III Cas subtypes do not correlate with repeat and Cas1 evolution. Finally, examples of horizontal transfer events of CRISPR-Cas systems between bacteria and archaea are identified, supported by CRISPR conservation and Cas homology.

MATERIALS AND METHODS

CRISPR data

Repeats from all publicly available genome sequences

All currently available genome sequences were downloaded from the NCBI server (http://www.ncbi.nlm.nih.gov/) and the CRISPR databases, CRISPI (37) and CRISPRdb (38) (August 2012). Redundant genomes were removed. We predicted CRISPRs using the two most common programs, CRISPRFinder (15) and CRT (16). For both tools, we used parameters that corresponded to at least three repeats within an array; repeat and spacer lengths were set to 18–58 nt. Although repeats within one array are largely identical, they can contain some mutations, especially toward the 3′-end of the array. Thus, we used a single representative repeat of a CRISPR array by calculating the consensus sequence of all repeat occurrences. Finally, we merged the results from both programs and the CRISPR databases to form a non-redundant set of >3500 consensus repeats, which we refer to as REPEATS. Table 1 gives a summary of our REPEATS data set. The results from CRISPRFinder and CRT give no information on the correct strand orientation. Therefore, we predict the repeat orientation within our clustering approach.

Set of repeats from Kunin et al. *2007 (32)*

We downloaded the data set from the supplementary material of (32) and refer to it as REPEATS_{Kunin}. This data set contains 271 bacterial and 78 archaeal sequences (349 in total). The orientations were predicted by the authors using previously published sequence features.

Set of archaeal repeats from Shah and Garrett 2011 (11)

We received 378 archaeal repeat sequences from Shah and Garrett that were the basis for the results in (11). The repeat orientations were manually verified by Shah and Garrett. We refer to this data set as REPEATS_{Shah}.

Table 1. Summary of our REPEATS data set including all publicly available CRISPR arrays

Data descriptor	Archaea	Bacteria
Genomes	279	2289
Genomes with CRISPRs (%)	177 (63)	877 (38)
Plasmids	41	1286
Plasmids with CRISPRs (%)	14 (34)	76 (6)
CRISPRs	643	2884
Repeats per array (median)	3–190 (15)	3–1371 (12)
Repeat lengths (median)	20–44 (29)	19–48 (30)
Spacer lengths (median)	20–50 (38)	19–70 (35)

Identifying conserved structure motifs

Our procedure for identifying conserved, local, hairpin-structure motifs (referred to as structure motifs) in all CRISPRs involves a complex multi-faceted workflow.

Step 1—Pool of unstructured repeats

The procedure starts with a pool, P_u, of repeats that have not been assigned to a structure motif. Initially P_u contains our entire REPEATS data set. The orientation of each repeat is predicted by a graph-kernel-based machine-learning model (39), slightly modified to work on directed graphs. We trained the model on the REPEATS_{Shah} data (using the 253 repeats that had <95% similarity to ones in REPEATS_{Kunin}). Each repeat sequence is given as a directed graph, i.e., the nucleotides are represented by nodes, which are linked by directed edges indicating the particular orientation. To test the performance of our model, we applied it to the REPEATS_{Kunin} data. Overall, we achieved a performance of 0.68 for the area under the receiver operating curve (ROC) with feature parameters radius $r = 1$ and distance $D = 2$. Because we did not achieve a perfect orientation prediction (mostly due to insufficient training data), we addressed this issue throughout our clustering process. Nonetheless, the model ensures that at least the majority of sequences are in the correct orientation for the first clustering steps.

Step 2—Generating a hierarchical cluster tree reflecting sequence and structure similarity

A hierarchical cluster tree T_i for the current iteration i is generated from all sequences in P_u using RNAclust (27). RNAclust uses a hierarchical clustering algorithm [UPGMA (40)] based on similarities calculated with a sequence-and-structure alignment program, LocARNA (27,29). Thus, the relationships in the resulting binary tree not only reflect sequence, but also structure similarity. For each node of the cluster tree, there exists a sequence-structure alignment with the respective predicted consensus structure as given by LocARNA.

Step 3—Selecting subtrees with CRISPR-like consensus structures

Starting from the root node in T_i, each child node is traversed in hierarchical order until a CRISPR-like hairpin consensus structure is found at a certain node t. The consensus structure is *local* in the sense that it does not cover the entire repeat sequences. All repeats descending from node t are considered to form a candidate structure motif, $Motif(t,T_i)$, if the following requirements, derived from published repeat structures (17–23), are met: First, the consensus structure of $Motif(t,T_i)$ is a hairpin with a stack of at least 4 bp and no bulges or internal loops. Second, at least 10 repeat sequences fit to the consensus structure of the motif candidate. All repeats that do not fit to the consensus structure are removed from $Motif(t,T_i)$. Third, the two child nodes of t must have compatible consensus structures. This means at least 75% of the base pairs must overlap with the consensus structure at t. The remaining child nodes of t are assigned to $Motif(t,T_i)$ and the procedure is repeated until all nodes in T_i have either been checked for—or have been assigned to—a structure motif.

Step 4—Supertree of only structured repeats

All repeats that have not been assigned to a structure motif are removed from the tree and are put back into the pool of unassigned repeats P_u. All other repeats, which form one of the consensus structures, are put into a set P_s. From this set P_s, a *supertree*, $ST(i)$, is generated by repeating Steps 2 and 3. Again repeats that do not conform to the criteria are removed and put back into the unassigned pool P_u. This reclustering ensures the robustness of identified motifs.

Step 5—Merging supertrees

In one RNAclust run, we identify conserved structures of repeat sequences that are neighbouring in the cluster tree T_i. To locate more distantly related repeat sequences that can still form a common consensus structure, we repeat the clustering with the remaining sequences in the pool P_u. Consequently, Steps 2–4 are repeated for three iterations, resulting in three separate supertrees (ST_1, ST_2 and ST_3) that are merged into one supertree, $ST_{1,2,3}$. Merging starts with ST_1. Because it is the result of the first iteration, it includes the largest and most well-conserved structure motifs. Each structure motif of the supertrees ST_2 and ST_3 is merged with ST_1, one at a time. Due to the orientation uncertainty, we also attempt to merge the reverse complement sequences of the whole structure motif. Merging occurs by repeating Steps 2–4 and we use the orientation that results in the fewest number of repeat sequences being lost to P_u in the merging process.

Step 6—Final cluster tree with structure motifs

We perform a last post-processing step to produce the final cluster tree with the structure motifs. For each structure motif, we calculate the consensus structure of the reverse complement repeat sequences. *GU* base pairs cannot form in the reverse complement orientation; therefore, we consider the orientation with the most stable consensus structure to be correct. We also check whether the reverse complement of a motif can be merged with another existing motif. Two biological features were used to check the orientations of entire motifs: the conserved 3′-end of repeats, *AUUGAAA(C/G)* and a majority of *A* instead of *U* nucleotides for archaeal sequences—as observed in the manually verified orientations in REPEATS_{Shah}. If any changes were made in the orientation, Steps 2–4 are repeated. Note that changes to the input set can lead to changes in the resulting tree; therefore, our repeated runs of RNAclust ensure that most of the noise is removed and we only include stable structure motifs in our final result.

Improving the orientation of repeats

The identification of conserved structure motifs gives some evidence on the likely orientation of the repeats involved. For the unassigned repeats, however, we had no information to deduce the correct orientation. Therefore, we merged all structured repeats with the REPEATS_{Shah} data and retrained our prediction model; we excluded repeats $\geq$95 % similarity with the test data.

Again, we tested our model on the $REPEATS_{Kunin}$ data and achieved a substantial improvement with an area under the ROC of 0.82 in comparison with 0.68 previously. We subsequently used our retrained model to predict the correct orientation of the repeats remaining in the unassigned pool P_u. Even if some orientations are still incorrect, this step ensures that the repeat orientations in our REPEATS data are consistent. To add the sequences that were previously in the incorrect orientation, we repeated Steps 1–6 with the improved orientation predictions.

Clustering of repeat sequences into conserved sequence families

Repeat sequences were clustered into related families based on global sequence similarity using Markov clustering (MCL) (41,42). The MCL method is a popular method for clustering biological sequence data and was applied previously to CRISPRs (11,32). First, we calculated pairwise similarities with the Needleman–Wunsch alignment algorithm (43). These similarities (i.e., percent identities) were plotted (Supplementary Figure S1) and a reasonable cutoff of 65% identity was chosen to represent a significant similarity. Similarities below this value were explicitly set to zero to reduce noise. We ran the MCL program (downloaded from http://micans.org/mcl/) with an inflation parameter $I = 2.5$. This parameter gave a good balance between the number of sequences assigned to a family and the conservation within a family. Only clusters with at least 10 repeat sequences were considered as a *conserved* sequence family.

We supplemented the Markov clustering with sequence profiles generated by CLUSTAL W (v. 1.83) (44). We used these profiles to reassign repeats to families to which they are sufficiently similar, as follows: Let $sim(F, r)$ be the profile score of a repeat r compared with the profile of the family F, where $r \notin F$. For each family, the minimum F_{min} and maximum F_{max} profile similarity was determined by removing each sequence from the family, recalculating the profile for the remaining sequences and determining the similarity score of the respective repeat to the profile. A repeat r was then assigned to a sequence family F if $sim(F, r) \geq F_{min}$ and the distance between $sim(F, r)$ and F_{max} is the minimum for all families. In total, 73 sequences were reassigned by the sequence profiles. The sequence conservation did not change significantly, but we were able to identify those few repeats that where missed by the MCL algorithm.

For each family, we generated sequence logos (Supplementary Figure S10 and Supplementary Tables S2–S19) by creating a multiple sequence alignment with the MAFFT program (45), version 6.4. The multiple sequence alignment was converted into a logo by WebLogo version 3 (46).

Cas gene and Cas-subtype annotations

Annotations of all cas genes

Subtype-independent annotation of *cas* genes was performed on the entire chromosome or plasmid that harbours the respective CRISPR array. We applied the TIGRFAM models from Haft *et al.* (8,47) in combination with HMMER (48), but used the more recent *cas* gene names from Makarova *et al.* (10). A *cas* gene was annotated when one of its respective models was found with an E-value ≤ 0.001. On our web server Web site, we offer a full table of *cas* gene annotations for each repeat, giving the minimum distance of that gene to the CRISPR array. For each sequence family and structure motif, we identified single *cas* genes that were associated with the majority of CRISPRs in the respective class (categories 50–69%, 70–89% and 90–100%); all *cas* genes on the entire chromosome or plasmid with the CRISPR were considered. Results are given in summary in Supplementary Tables S2–S19 and in full on the web server.

Cas subtype annotation from Makarova et al. 2011 (10)

The automatic annotation of subtypes is tricky owing to the fact that genes of multiple subtypes can be present in the genome, subtypes are often incomplete and it is not known if the *cas* genes must be within a certain distance of the CRISPR array. However, in many published CRISPR-Cas systems, the *cas* genes are located either directly upstream or downstream of the array (10). We used the following procedure that enabled a suitable trade-off between strictness and completeness of the annotations. We first compiled a list of signature *cas* genes that were unique to each type and subtype from (10). For each repeat, i.e. CRISPR array locus, we identified the closest subtype signature and then noted the distance of the respective type signature, if available. We plotted the distance of subtype and type signatures and determined a clear peak (at 14.5 kb) in their distances to their respective CRISPR array (Supplementary Figure S3). We considered a cutoff of 180 kb to represent a suitable distance from the CRISPR array; this cutoff corresponds to the 70th percentile of distances of the subtype signatures. A repeat is assigned to a subtype if both subtype and type signatures are within this distance. Note that with this approach, not all *cas* genes have to be present (or annotated).

Clustering of Cas1 proteins

Cas1 protein sequences were assigned to the closest CRISPRs if they were within 180 kb of the array (see Supplementary Figure S3 for cutoff explanation). These Cas1 proteins were again clustered using MCL (41,42) with default parameters. Here, pairwise sequence similarities were calculated with the local Smith–Waterman alignment algorithm (49) and percent identities <40% were set to zero to reduce noise. Only clusters with at least 10 proteins were considered.

The CRISPRmap cluster tree

The tree was generated by RNAclust (27) and visualised with iTOL (50). In the tree, we see relationships based on sequence *and* structure similarity; however, when the repeat is unstructured, only the sequence similarity is considered. The encircling rings correspond to the following annotations (displayed as selected by the user): structure motifs, sequence families, Cas subtypes, phyla (taxonomy)

and superclasses. The branches are coloured according to whether the CRISPRs were from bacteria or archaea.

Web server input: adding new sequences

The user of our CRISPRmap web server can enter up to 300 CRISPR sequences in FASTA format and indicate whether the correct orientation is unknown and requires prediction. We use a multi-step procedure that has been optimised for speed to assign the given repeats to our structure motifs and sequence families. Further details are given in the Supplementary Methods S1.2.

RESULTS AND DISCUSSION

All available CRISPR sequences from bacteria and archaea

We obtained >3500 consensus repeat sequences from predicted CRISPR arrays in ~2500 available genomes. This data set, referred to as REPEATS (see Table 1), is the most complete set of CRISPRs to date. We compared the REPEATS data set to previous work in Supplementary Figure S5.

Structure motifs and sequence families

We performed a comprehensive search for both conserved sequence families and small CRISPR-like hairpin motifs, using *independent* approaches to allow for both structured and unstructured repeats. First, we partitioned CRISPRs into sequence families using Markov clustering, as in previous studies (11,32); in addition, we applied sequence profiles to refine the Markov clusters. We identified 40 conserved families. The mean pairwise sequence identity of 68–96% (avg. 82%) reflects a high level of sequence conservation. Second, independent to identified sequence families, we searched for conserved structure motifs using sequence-and-structure alignments. Structure motif candidates were constrained to be reminiscent of those previously published (17,19,20,22–25). More specifically, 33 small hairpin (or stem-loop) motifs with at least 4 bp and no bulges were identified. Their sequence conservation was generally lower than for sequence families: mean pairwise sequence identities between 47 and 94% with an average of 69% (compared with 82%). Sequence families and structure motifs were numbered according to size, starting with the largest clusters; the smallest cluster size was 10. Summary tables with sequence logos for families, secondary structures for motifs, mappings between families and motifs and annotations are available in the Supplementary Material; full alignments are available on the CRISPRmap web server.

To provide further support for our secondary structure predictions, we evaluated the motifs using the general ncRNA predictor, RNAz (51). Although RNAz is not specifically trained for CRISPR elements, it classified 79% (26 out of 33) of our motifs as structured ncRNAs with an SVM-RNA-class probability >0.6 (22 motifs even achieved >0.9, a clear indication that these motifs are evolutionary conserved). Compared with other ncRNA classes, RNAz only exhibits such promising sensitivities on some of the classical ncRNAs (52,53), for example, transfer RNAs or microRNAs, which are known for their distinct and well-defined secondary structures (54,55).

In total, out of all CRISPRs in our REPEATS data set, 64% were assigned to a conserved sequence family and 51% were assigned to a structure motif; 26% of repeats remained unassigned to either a family or motif, i.e. showed no conservation with available CRISPRs.

A detailed visual map of CRISPR conservation

As a visual *map* of both bacterial and archaeal CRISPR domains, we combined our categorisation into repeat families and motifs with a hierarchical tree based on sequence-and-structure similarities (see non-hierarchical, sequence-similarity-based visualisation in Supplementary Figure S9). This CRISPRmap tree details relationships between individual repeats and whole families and motifs (Figure 1).

In addition to the repeat families and motifs, we annotated taxonomic phyla, Cas1 sequence homology clusters, and Cas subtype annotations (8,10); the branches are coloured according to whether the CRISPRs stem from bacteria or archaea. We show one possible view of the CRISPRmap tree with structure-motifs, sequence-families and superclass classifications and the domain in Figure 1. Further views and annotation data are available in the supplementary material and on our CRISPRmap web server.

In summary, the CRISPRmap tree was designed to provide a visual overview of CRISPR conservation and to aid in the understanding of CRISPR-Cas diversity.

The CRISPRmap tree is divided into six superclasses

Based on sequence-and-structure similarities and the tree topology, the REPEATS data set could be broadly grouped into six major superclasses (Figure 2). The superclasses, labelled A–F, are ordered according to generally decreasing conservation. The following information is quickly observed in the CRISPRmap tree (Figure 1): Superclass A contains highly conserved CRISPRs on the sequence level, but only a few small structure motifs. Superclasses B–C contain sequence families that roughly correspond to one structure motif each; the same is true for half of superclass D. The other half of superclass D and superclass E contain little sequence conservation, but many small conserved motifs. Archaeal CRISPRs in both superclasses A and F contain well-conserved sequence families and we find motifs for about half; however, these are less stable than the bacterial motifs in superclasses B–D (Supplementary Tables S2–S19). The bacterial repeats in superclass F are divergent. We included arrays with at least three repeat instances to ensure that our data set was complete. Many arrays with up to five repeat instances, however, show little conservation (Supplementary Figure S8): roughly 50% were not assigned to sequence families or structure motifs and most are in this diverse part of superclass F. In addition to array size, we marked repeats or spacers with unusual lengths on the CRISPRmap tree in Supplementary Figure S8. Some of

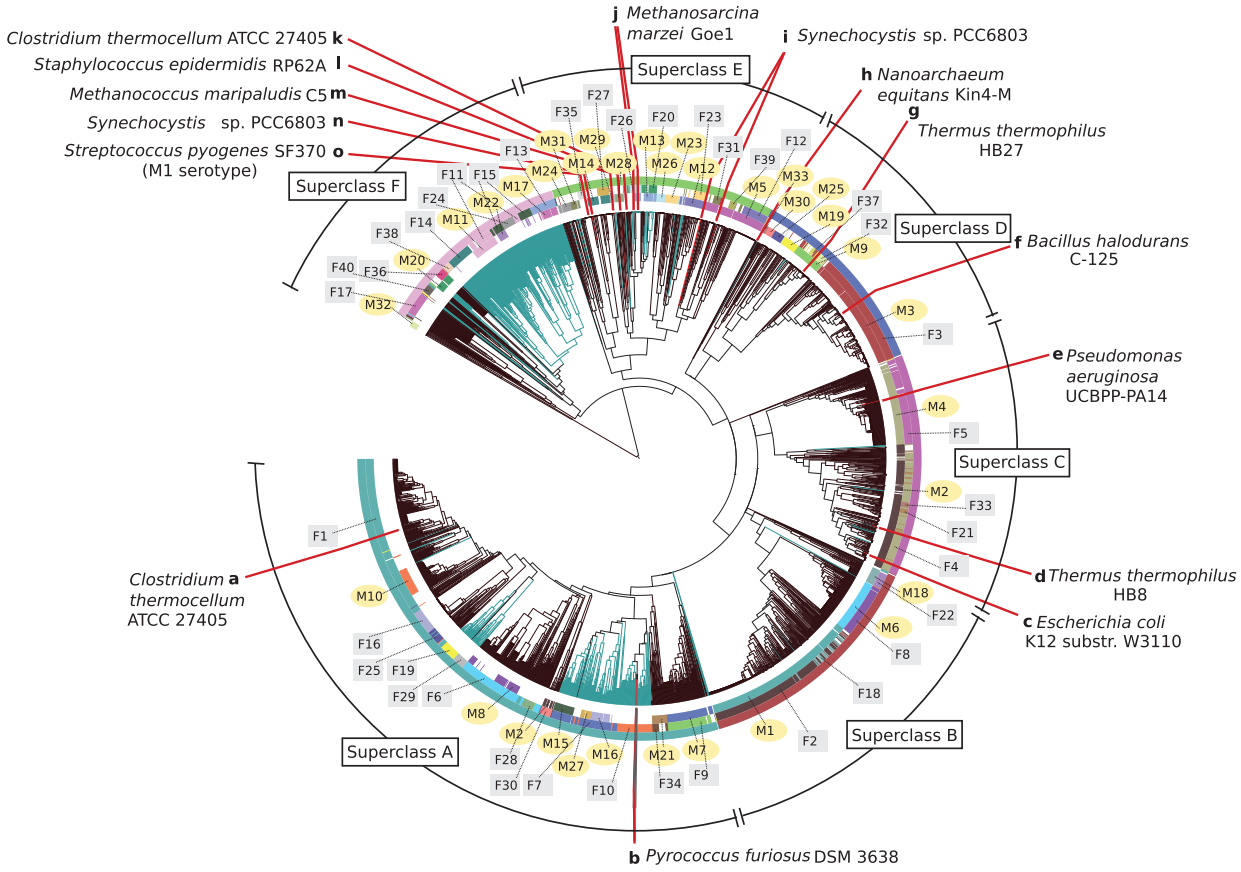


Figure 1. The CRISPRmap tree: a map of repeat sequence and structure conservation. The hierarchical tree is generated with respect to repeat sequence and structure pairwise similarity and the branches are coloured according to their occurrence in the domains bacteria (dark brown) or archaea (blue-green). The rings annotate the conserved structure motifs (inner), sequence families (middle) and the superclass (outer). Motifs and families are marked and highlighted with yellow circles, and grey squares, respectively. Finally, we marked locations of published CRISPR-Cas systems for which experimental evidence of the processing mechanism exists (13,17–25,33–36,51). A summary for these published systems is given in Supplementary Table S20. Repeats that show no conservation, i.e. were not assigned to either a sequence family or structure motif, were removed to clarify the visualisation.

the really short arrays, especially those with unusual repeat and/or spacer lengths are unlikely to contain functional CRISPRs.

We summarised subsequent annotations and clustering results to give a brief overview of each superclass in Figure 2; more details are given in the following results. In the CRISPRmap tree views (e.g., Figure 1), the superclass is always annotated in the outermost ring.

Structure motifs fit to known cleavage sites

Most sequence families and structure motifs are associated with either bacterial or archaeal CRISPRs: only four motifs (M11, M20, M29 and M31) and one family (F20) contain a significant mixture of both domains. Bacterial CRISPRs are more structured in general than those from archaea. Although structured motifs were identified for both domains, the longer, more thermodynamically stable hairpins—associated with Cas subtypes I-C, I-E and I-F—belonged almost exclusively to bacterial CRISPRs in superclasses B–D (Supplementary Figure S10A–C and Supplementary Tables S6–S11). To add to the stability of such short hairpin motifs, 65% of base pairs are *G*s paired to *C*s. In a closer inspection, we observed that 94% of *GC* base pairs were orientated with the *G* toward the 3′-end (Supplementary Tables S2–S19). Such consecutive $C \rightarrow G$ base pairs form a 3′ *G* side to the stem, which might be important for crRNA processing due to sequence specificity in this region (20,22,23).

In the literature, cleavage by known Cas6-like endoribonucleases (during crRNA maturation) occurs either at the 3′ base of the hairpin motif, or within the double-stranded region of the hairpin stem, usually below such a $C \rightarrow G$ base pair (13,17–21,23–25,33–36). The product of this cleavage is an 8-nt-long repeat tag at the 5′-end of the mature crRNA (5′ tag), which corresponds to the last eight nucleotides from the 3′-end of the repeat sequence. Some exceptions to the 8-nt length exist (23,24,35,56,57). We located potential cleavage sites on our structure motifs according to published observations (17–20,22–25). Of all 33 structure motifs, 11 contain a potential cleavage site within the conserved stem of the motif of which 7 are below a $C \rightarrow G$ base pair. Another 13 motifs have a potential cleavage site at the 3′ base of the

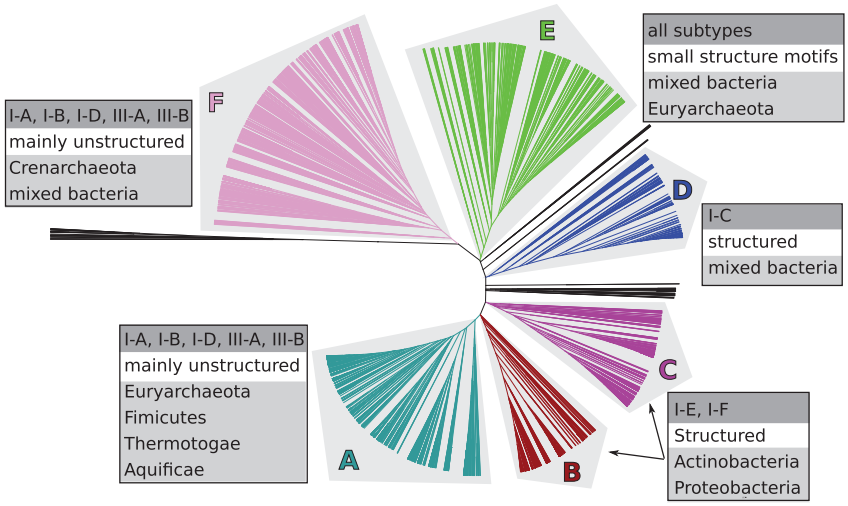


Figure 2. CRISPRs cluster into six major superclasses according to sequence and structure similarity. We summarised general results of our structure motif detection (i.e. structured or unstructured), Cas-subtype annotations (10) and taxonomic phyla beside each superclass.

conserved stem. In Figure 2, we see that Cas subtypes I-E and I-F are split across the two superclasses B and C. This split is due to exactly one repeat-structure feature: the hairpin motifs are closer to the 3′-end of the CRISPRs in superclass B, resulting in a cleavage site within the stem. In superclass C, the cleavage site is at the base of the hairpin motif. In accordance to previously mentioned literature, the cleavage sites are below a $C \rightarrow G$ base pair in both superclasses. Aside from this difference in position, the hairpin structures associated with either I-E or I-F are similar.

Sequence families exhibit variations in conservation

In a closer inspection of the family sequence logos, we see different patterns of *sequence* conservation (Supplementary Figure S10 and Supplementary Tables S2–S19). We highlight these difference using four selected examples: First, CRISPRs associated with the I-E subtype show a high conservation of *G*s and *C*s that form the base pairs of the hairpin motif. Second, CRISPRs associated with the I-F subtype are well-conserved across the entire repeat sequence and contain fewer consecutive *C*s and *G*s (Supplementary Figure S10A and B). Third, CRISPRs associated with the I-C subtype show a higher conservation at the base of the hairpin stem and in the single-stranded 5′- and 3′-ends, which suggests that the top of the stem and the hairpin loop is likely insignificant for the binding affinity (Supplementary Figure S10C); this conservation pattern is well-supported by mutation experiments in the type I-C system in *Bacillus halodurans* C-125 where crRNAs were still processed with a truncated upper stem and mutated hairpin loop, but processing was sequestered by mutations at the base of the stem or by the removal of the unpaired 3′-end (23). Fourth, in Figure 3, we marked the well-conserved 8-nt-long 5′ tag, $AUUGAAA(C/G)$, at the 3′-end of the repeats. Out of our 40 sequence families, 17 (∼40%) show a conservation of exactly this sequence tag; others contain minor deviations. Interestingly, bacterial superclasses B and C do not show this tag, whereas it is highly conserved throughout the other bacterial superclass D and in almost all archaeal families (9 out of 12). We hypothesise that these patterns of conservation give a good indication of differences in binding affinities for specific Cas proteins in the various CRISPR-Cas systems.

Sequence families and structure motifs provide independent information about evolution

Structured ncRNA families cannot be identified by sequence conservation alone because standard alignment tools fail when the pairwise sequence identity is <60% (58). We see the same tendency for structured and unstructured repeats in our data: The CRISPRmap tree shows different patterns of overlap between sequence families and structure motifs that we identified by independent clustering approaches (Figure 1). In Figure 3, we highlight two overlap patterns. First, in superclass A, the largest family, namely F1, is mainly unstructured. For a subset of these CRISPRs, however, we identified a thermodynamically stable hairpin motif (M10) with four consecutive $C \rightarrow G$ base pairs; these CRISPRs are clearly structured. Second, in superclass D, we found a conserved hairpin motif (M28), also with four consecutive $C \rightarrow G$ base pairs and a large 8-nt hairpin loop that was verified by mutational analyses in a type III-A system in *Staphylococcus epidermidis* RP62A (20); this motif does not show enough sequence conservation to be detected as a sequence family. Both M10 and M28 would not have been identified with the approach used in (32), in which consensus structures were calculated from (entire) sequence families. In addition, we observe cases where a structure motif corresponds almost fully to a sequence family, e.g. M1 with F2 and M2 with F4. Nevertheless, individual members of the sequence families cannot form the associated consensus structure: this may indicate a degenerate and non-functional CRISPR-Cas system, or one that has evolved to function with a different or no repeat structure.

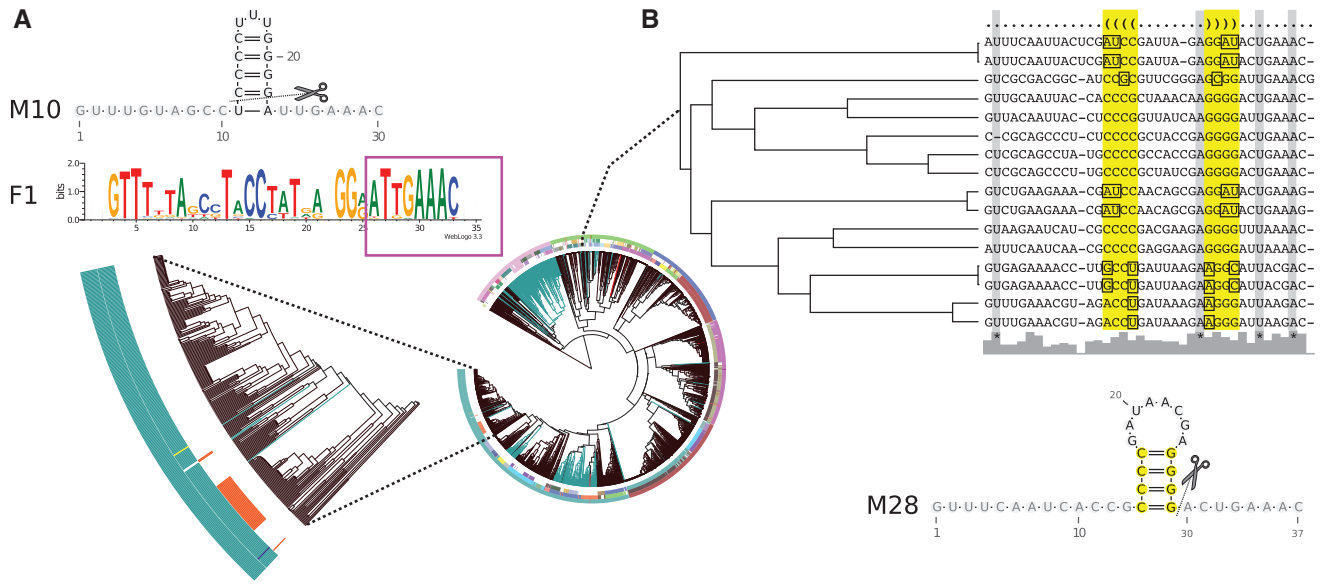


Figure 3. Highlighting the advantage of independent clustering approaches. (**A**) CRISPRs in the largest sequence family, F1, are mostly unstructured; however, for 50 CRISPRs also a conserved structure motif, M10, was identified. This indicates that subsets of conserved families can be structured. F1 contains the conserved 5′ tag, marked with the magenta box. (**B**) Structure motif M28 shows no sequence conservation, but a conserved structure (base pairs are highlighted in yellow). The many compensatory base pairs are marked in the alignment with squares. This structure has been verified via mutational analyses in (20). Potential cleavage sites are indicated as observed in the literature (13,17–21,23–25,33–36).

A subset of Cas subtypes are weakly linked to repeat and Cas1 evolution

From the literature, we already know that Cas1 is strongly linked to repeat evolution (12,59). This link could be verified for our large-scale data set (Figure 4A). We clustered associated Cas1-protein sequences and the resulting Cas1 clusters fit well with the superclasses, except superclass E (Figure 4). There are several indications that superclass E contains only partial data, e.g. conserved sequence families and structure motifs are smaller and most CRISPRs show little to no conservation; however, 50% of the CRISPRs from metagenomic data in the subsequent use-case study fall into this superclass and new conserved classes are indicated (Supplementary Figure S7).

For Cas-subtypes (10), the linkage pattern is different: subtypes I-C, I-E and I-F correlate well with repeat (and thus Cas1) conservation, whereas the remaining type I and both type III Cas subtypes are only weakly linked (Figure 4). The bacterial superclasses B, C and D contain well-defined structure motifs and sequence families (Figure 1 and Supplementary Tables S2–S19), which are associated with subtypes I-E and I-F (superclasses B and C) and I-C (half of superclass D). Superclasses A and F contain both bacterial and archaeal CRISPRs—most of which are unstructured—and although they also fit well to the Cas1 clusters, the annotated Cas subtypes are a diverse mixture of the remaining type I subtypes (I-A, I-B and I-D) and both type III subtypes (Figure 4). In superclass E, we observe a similar co-occurrence of these subtypes; however, this superclass contains all subtypes owing to aforementioned diversity and incomplete data.

There are two possible explanations for the co-occurrence of type I and type III subtypes. First, these subtypes are composed of interchangeable modules as previously mentioned for archaeal systems in (12,60). In such cases, one would expect Cas proteins from different subtypes to be able to process similar repeat sequences; two examples in the literature that support this theory is a Cas6 (Cas6b) protein that can process both type I-B systems in *Methanococcus maripaludis* C5 and *Clostridium thermocellum* ATCC 27405 (13) and two CRISPRs in *Methanosarcina marzei* Gö1 with near-identical repeats are associated with different subtypes I-B and III-B (25). Also, many sequence families and structure motifs co-occur with multiple, or a mixture of, subtypes (see Supplementary Tables S2–S19 and web server). The co-occurrence of subtypes is widespread in archaea and bacteria. In general, an exchange of protein modules would require compatible repeat sequences and structures. The only similarity observed in CRISPRs associated with mixed subtypes is the conserved 5′ tag—*AUUGAAA(C|G)*—or a slight variation. In comparison, repeats associated with the bacterial subtypes I-E and I-F do not contain this tag. Second, additional or unknown Cas proteins are required to achieve a subclassification of Cas subtypes that is more compatible with repeat conservation. Most likely, the truth lies in a combination of both explanations. Finally, subtypes I-A, I-B, I-D, III-A and III-B are more enriched in extremophiles, e.g. thermophiles (Supplementary Figure S6).

CRISPRs in Euryarchaeota are closer to bacterial systems than ones in Crenarchaeota

Ninety-seven precent of the archaeal CRISPRs originate from two phyla: 380 from Euryarchaeota and 245 from Crenarchaeota. In the CRISPRmap tree (Figure 1 and Supplementary Figure S4), we observe a clear separation of these two CRISPR groups: 60% of CRISPRs from

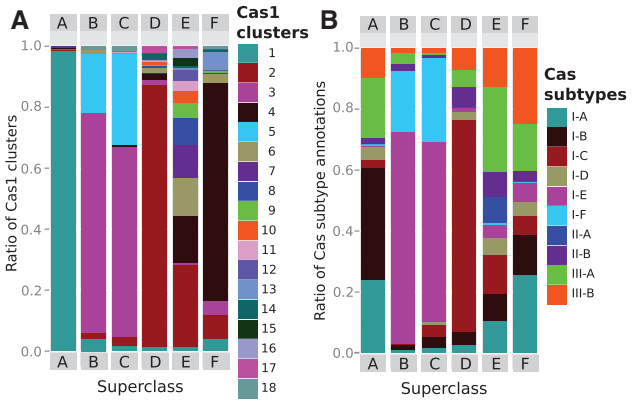


Figure 4. Relative ratios of Cas1 sequence clusters and Cas-subtype annotations per superclass. (**A**) Cas1 sequence clusters correspond well to the superclass and thus the CRISPRmap tree with the exception of superclass E; superclass E is diverse in both repeat and associated Cas1 conservation and it probably contains only partial data. (**B**) Bacterial CRISPRs that are assigned to well-defined structure motifs are associated with subtypes I-C, I-E and I-F in superclasses B–D and are strongly linked to both repeat and Cas1-sequence similarities (i.e. CRISPR evolution). Superclass A and F contain both bacterial and archaeal CRISPRs (many are unstructured), which are loosely associated with the remaining type I and both type III subtypes. These subtypes do not correspond to Cas1 and repeat evolution and are likely composed of interchangeable protein complexes or modules. The diversity of superclass E is also reflected by the mixture of all subtypes; in addition, the majority of type II CRISPRs are also located in this region.

Euryarchaeota and 96% from Crenarchaeota cluster into superclasses A and F, respectively. In superclass A, the euryarchaeal and bacterial CRISPRs are associated with Cas1 proteins that cluster into the same Cas1-cluster-1, i.e. these Cas1 sequences are evolutionarily close (Figure 4). In contrast, CRISPRs from Crenarchaeota are located almost exclusively in a subregion of superclass F and are associated with the separate Cas1-cluster-4 (Supplementary Figure S4).

Evidence of horizontal transfer

As previously mentioned, archaeal and bacterial CRISPRs are distinctly separated in the CRISPRmap tree (Figure 1). This is consistent with a rare exchange of genetic material between archaeal and bacterial systems (11,12). Nevertheless, we observed a few instances where archaeal repeats are located in a bacterial-dominated region and vice versa (see Supplementary Methods S1.4 for more details). With one exception, all cases involved a transfer of the CRISPR-Cas system from bacteria to archaea; archaea have also been shown to uptake bacterial and eukaryotic DNA as spacers (61). Supplementary Figure S11 gives examples of archaea that contain full bacterial CRISPR-Cas systems where a strong conservation of the structure motif is supported by multiple compensatory base pair mutations. In addition, not only the Cas1 proteins are conserved, but the archaeal CRISPRs are associated with the complete set of proteins from the bacterial subtypes I-C and I-E.

The transfer of genetic material between prokaryotes often occurs via plasmids; however, in Supplementary Figure S11, all horizontally transferred systems are located on chromosomes and not on plasmids. In fact, only 7% of over 1300 plasmids analysed contained a CRISPR array. Therefore, it is unlikely that the dominant mechanism of transferring CRISPR-Cas systems between organisms is via plasmids.

The CRISPRmap web server

The CRISPRmap web server enables easy access to our data and allows scientists to compare the conservation of individual repeats. Repeats are entered in FASTA format and the web server automatically assigns them to our classification system; previously unknown repeats are assigned to existing families and/or motifs, if possible. Non-conserved input sequences remain unassigned, but are still located according to their relative similarity in the tree. Furthermore, if the correct orientation of the input repeats is unknown, the user can request to predict the orientations to ensure that they are consistent with our data.

A use-case study

A valuable source of new CRISPR-Cas systems are metagenomic studies of multiple, often novel, prokaryotic organisms. Recently, a targeted search for CRISPR arrays was performed in the bacterial metagenome of different sites on the human body (62). In this study, 150 CRISPRs were identified that could potentially be used to learn more about invader patterns. We applied the CRISPRmap web server to determine the conservation of these CRISPRs at a quick glance: only 38 and 29% were assigned to our structure motifs or sequence families, respectively. Notably, 50% of the metagenomic CRISPRs were assigned to the diverse superclass E where most remained unassigned to either a structure motif or sequence family; however, in Supplementary Figure S7, many of these repeats cluster together to potentially form new classes of motifs and families. Two CRISPRs fall into the euryarchaeal region in superclass A, despite the fact that archaea are rarely associated with human microbiomes (62). These results highlight the fact that even with the large-scale analysis performed in this work, we still do not know the full extent of CRISPR-Cas diversity. Therefore, the dynamic nature of our web server—in the fact that it allows the classification of newly sequenced CRISPRs to be assigned to existing sequence families and structure motifs—is particularly useful.

CONCLUSION

We provide a comprehensive analysis of CRISPR structure and sequence conservation based on the largest data set of repeat sequences available. We show extensively that our methods are well suited to identifying many characteristics of CRISPR-Cas systems: e.g. cleavage sites, patterns of RNA structure motifs and sequence conservation, the link between evolution of CRISPRs and associated Cas subtypes and the horizontal transfer of such systems. On the one hand, specific conservation patterns can be combined with published data to make assumptions about CRISPRs belonging to the same

sequence families or structure motifs. On the other hand, the CRISPRmap overview can be used to find potentially novel CRISPR-Cas systems that are highly divergent from the rest. User-based queries on our data enable more informed choices on future hypotheses in CRISPR-Cas research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [63,64].

ACKNOWLEDGEMENTS

The authors thank Fabrizio Costa, Steffen Heyne, Martin Mann, Roger Garrett, Shiraz Shah, Rhodri Saunders, Michael Uhl and Ibrahim Kessba for their support. Conceived the methods and analyses: S.J.L., D.R., O.S.A., R.B. and S.W. Data acquisition: O.S.A. Implemented software: O.S.A., D.R. and S.J.L. Technical assistance: D.R. Designed the web server: D.R., O.S.A. and S.J.L. Analysed the data: S.J.L. and O.S.A. Planned and wrote the article: S.J.L. and R.B. All authors read and revised the final manuscript.

FUNDING

German Research Foundation (DFG) program FOR1680 'Unravelling the Prokaryotic Immune System' [BA 2168/5-1 to R.B.]; The German Academic Exchange Service (DAAD) (to O.S.A.). Funding for open access charge: Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg.

Conflict of interest statement. None declared.

REFERENCES

1. Terns,M.P. and Terns,R.M. (2011) CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.*, **14**, 321–327.
2. Al-Attar,S., Westra,E.R., van der Oost,J. and Brouns,S.J. (2011) Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. *Biol. Chem.*, **392**, 277–289.
3. Wiedenheft,B., Sternberg,S.H. and Doudna,J.A. (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, **482**, 331–338.
4. Garneau,J.E., Dupuis,M.E., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadan,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.
5. Hale,C.R., Majumdar,S., Elmore,J., Pfister,N., Compton,M., Olson,S., Resch,A.M., Glover,C.V., Graveley,B.R., Terns,R.M. *et al.* (2012) Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell*, **45**, 292–302.
6. Zhang,J., Rouillon,C., Kerou,M., Reeks,J., Brugger,K., Graham,S., Reimann,J., Cannone,G., Liu,H., Albers,S.V. *et al.* (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol. Cell*, **45**, 303–313.
7. 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. *et al.* (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.*, **18**, 529–536.
8. 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 Comput. Biol.*, **1**, e60.
9. Makarova,K.S., Aravind,L., Wolf,Y.I. and Koonin,E.V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct.*, **6**, 38.
10. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
11. Shah,S.A. and Garrett,R.A. (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol.*, **162**, 27–38.
12. Garrett,R.A., Vestergaard,G. and Shah,S.A. (2011) Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol.*, **19**, 549–556.
13. Richter,H., Zoephel,J., Schermuly,J., Maticzka,D., Backofen,R. and Randau,L. (2012) Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. *Nucleic Acids Res.*, **40**, 9887–9896.
14. Juranek,S., Eban,T., Altuvia,Y., Brown,M., Morozov,P., Tuschl,T. and Margalit,H. (2012) A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs. *RNA*, **18**, 783–794.
15. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
16. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
17. Brouns,S.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J., 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*, **321**, 960–964.
18. Gesner,E.M., Schellenberg,M.J., Garside,E.L., George,M.M. and Macmillan,A.M. (2011) Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat. Struct. Mol. Biol.*, **18**, 688–692.
19. Sashital,D.G., Jinek,M. and Doudna,J.A. (2011) An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat. Struct. Mol. Biol.*, **18**, 680–687.
20. Hatoum-Aslan,A., Maniv,I. and Marraffini,L.A. (2011) Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc. Natl Acad. Sci. USA*, **108**, 21218–21222.
21. Haurwitz,R.E., Sternberg,S.H. and Doudna,J.A. (2012) Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *EMBO J.*, **31**, 2824–2832.
22. Sternberg,S.H., Haurwitz,R.E. and Doudna,J.A. (2012) Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA*, **18**, 661–672.
23. 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*, **20**, 1574–1584.
24. Scholz,I., Lange,S.J., Hein,S., Hess,W.R. and Backofen,R. (2013) CRISPR-Cas systems in the Cyanobacterium Synechocystis sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS One*, **8**, e56470.
25. Nickel,L., Weidenbach,K., Jager,D., Backofen,R., Lange,S.J., Heidrich,N. and Schmitz,R.A. (2013) Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III. *RNA Biol.*, **10**, 5.
26. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.

27. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
28. Havgaard,J.H., Torarinsson,E. and Gorodkin,J. (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, **3**, 1896–1908.
29. Will,S., Joshi,T., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2012) LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, **18**, 900–914.
30. Gruber,A.R., Bernhart,S.H., Hofacker,I.L. and Washietl,S. (2008) Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, **9**, 122.
31. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: Wikipedia, clans and the ''decimal'' release. *Nucleic Acids Res.*, **39**, D141–D145.
32. Kunin,V., Sorek,R. and Hugenholtz,P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.*, **8**, R61.
33. 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*, **329**, 1355–1358.
34. Wang,R., Preamplume,G., Terns,M.P., Terns,R.M. and Li,H. (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure*, **19**, 257–2564.
35. Garside,E.L., Schellenberg,M.J., Gesner,E.M., Bonanno,J.B., Sauder,J.M., Burley,S.K., Almo,S.C., Mehta,G. and MacMillan,A.M. (2012) Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA*, **18**, 2020–2028.
36. Randau,L. (2012) RNA processing in the minimal organism Nanoarchaeum equitans. *Genome Biol.*, **13**, R63.
37. Rousseau,C., Gonnet,M., Le Romancer,M. and Nicolas,J. (2009) CRISPI: a CRISPR interactive database. *Bioinformatics*, **25**, 3317–3318.
38. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.
39. Costa,F. and Grave,K.D. (2010) Fast neighborhood Subgraph Pairwise distance kernel. In: *Proceedings of the 26 th International Conference on Machine Learning*. Omnipress, pp. 255–262.
40. Gronau,I. and Moran,S. (2007) Optimal implementations of UPGMA and other common clustering algorithms. *Inf. Process. Lett.*, **104**, 205–210.
41. van Dongen,S. (2000) Graph Clustering by Flow Simulation, *PhD thesis*. University of Utrecht.
42. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
43. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
44. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
45. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
46. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
47. Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Beck,E. TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res*, **41**, D387–D395.
48. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
49. Smith,T. and Waterman,M. (1981) Comparison of Biosequences. *Adv. Appl. Math.*, **2**, 482–489.
50. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W478.
51. Gruber,A.R., Findeiss,S., Washietl,S., Hofacker,I.L. and Stadler,P.F. (2010) RNAz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, 69–79.
52. Rose,D., Hackermuller,J., Washietl,S., Reiche,K., Hertel,J., Findeiss,S., Stadler,P.F. and Prohaska,S.J. (2007) Computational RNomics of drosophilids. *BMC Genomics*, **8**, 406.
53. Rose,D., Joris,J., Hackermuller,J., Reiche,K., Li,Q. and Stadler,P.F. (2008) Duplicated RNA genes in teleost fish genomes. *J. Bioinform. Comput. Biol.*, **6**, 1157–1175.
54. Kaczkowski,B., Torarinsson,E., Reiche,K., Havgaard,J.H., Stadler,P.F. and Gorodkin,J. (2009) Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics*, **25**, 291–294.
55. Giege,R., Juhling,F., Putz,J., Stadler,P., Sauter,C. and Florentz,C. (2012) Structure of transfer RNAs: similarity and variability. *Wiley Interdiscip. Rev. RNA*, **3**, 37–61.
56. Sinkunas,T., Gasiunas,G., Waghmare,S.P., Dickman,M.J., Barrangou,R., Horvath,P. and Siksnys,V. (2013) *In vitro* reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *EMBO J.*, **32**, 385–94.
57. Deltcheva,E., Chylinski,K., Sharma,C.M., Gonzales,K., Chao,Y., Pirzada,Z.A., Eckert,M.R., Vogel,J. and Charpentier,E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, **471**, 602–607.
58. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
59. Horvath,P., Coute-Monvoisin,A.C., Romero,D.A., Boyaval,P., Fremaux,C. and Barrangou,R. (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int. J. Food Microbiol.*, **131**, 62–70.
60. Shah,S.A., Erdmann,S., Mojica,F.J. and Garrett,R.A. (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.*, **10**, 5.
61. Brodt,A., Lurie-Weinberger,M.N. and Gophna,U. (2011) CRISPR loci reveal networks of gene exchange in archaea. *Biol. Direct.*, **6**, 65.
62. Rho,M., Wu,Y.W., Tang,H., Doak,T.G. and Ye,Y. (2012) Diverse CRISPRs evolving in human microbiomes. *PLoS Genet.*, **8**, e1002441.
63. Hofacker,I.L. and Stadler,P.F. (2006) Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, **22**, 1172–1176.
64. Theocharidis,A., van Dongen,S., Enright,A.J. and Freeman,T.C. (2009) Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat. Protoc.*, **4**, 1535–1550.

CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci

Omer S. Alkhnbashi, Fabrizio Costa, Shiraz A. Shah, Roger A. Garrett, Sita J. Saunders and Rolf Backofen. **Bioinformatics Journal, 2014**, doi:10.1093/bioinformatics/btu459.

Personal contribution

My contribution in this project is a major one. I conceived the method and its analysis. Furthermore, I implemented the software (CRISPRstrand), updated the CRISPRmap webserver and was involved in planning and writing the article.

Omer S. Alkhnbashi

The following co-authors confirm the above stated contribution.

Fabrizio Costa

Shiraz A. Shah

Roger A. Garrett

Sita J. Saunders

Rolf Backofen

BIOINFORMATICS

Vol. 30pages i489–i496
doi:10.1093/bioinformatics/btu459

CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci

Omer S. Alkhnbashi[1], Fabrizio Costa[1], Shiraz A. Shah[2], Roger A. Garrett[2], Sita J. Saunders[1] and Rolf Backofen[1,3,*]

[1]Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany, [2]Department of Biology, University of Copenhagen, Archaea Centre, Ole Maaloes Vej 5, DK2200 Copenhagen, Denmark and [3]BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, University of Freiburg, Germany

ABSTRACT

Motivation: The discovery of CRISPR-Cas systems almost 20 years ago rapidly changed our perception of the bacterial and archaeal immune systems. CRISPR loci consist of several repetitive DNA sequences called repeats, inter-spaced by stretches of variable length sequences called spacers. This CRISPR array is transcribed and processed into multiple mature RNA species (crRNAs). A single crRNA is integrated into an interference complex, together with CRISPR-associated (Cas) proteins, to bind and degrade invading nucleic acids. Although existing bioinformatics tools can recognize CRISPR loci by their characteristic repeat-spacer architecture, they generally output CRISPR arrays of ambiguous orientation and thus do not determine the strand from which crRNAs are processed. Knowledge of the correct orientation is crucial for many tasks, including the classification of CRISPR conservation, the detection of leader regions, the identification of target sites (protospacers) on invading genetic elements and the characterization of protospacer-adjacent motifs.

Results: We present a fast and accurate tool to determine the crRNA-encoding strand at CRISPR loci by predicting the correct orientation of repeats based on an advanced machine learning approach. Both the repeat sequence and mutation information were encoded and processed by an efficient graph kernel to learn higher-order correlations. The model was trained and tested on curated data comprising >4500 CRISPRs and yielded a remarkable performance of 0.95 AUC ROC (area under the curve of the receiver operator characteristic). In addition, we show that accurate orientation information greatly improved detection of conserved repeat sequence families and structure motifs. We integrated CRISPRstrand predictions into our CRISPRmap web server of CRISPR conservation and updated the latter to version 2.0.

Availability: CRISPRmap and CRISPRstrand are available at http://rna.informatik.uni-freiburg.de/CRISPRmap.

Contact: backofen@informatik.uni-freiburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

CRISPR-Cas immune systems of bacteria and archaea provide adaptive defence against a variety of invading genetic elements. They have been classified into three major classes: Types I, II and III, where Type II systems are confined to bacteria (Makarova *et al*., 2011a; Vestergaard *et al*., 2014). The adaptive immune response of all types is divided into three major phases: (i) adaptation, the uptake of DNA fragments from genetic elements and their insertion between consecutive repeats of a CRISPR array, generally adjacent to a leader sequence; (ii) processing of the CRISPR array transcripts within the repeats to generate small crRNAs that derive from part or all of each spacer region and (iii) interference involving targeting and cleavage of an invading genetic element, or its transcripts, by Cas protein–crRNA complexes (Barrangou and van der Oost, 2013, and Fig. 1). Whereas the adaptation phase is relatively conserved in the different CRISPR-Cas systems, significant differences occur in the processing and interference mechanisms. Thus, where Type I and III systems employ a Cas6 processing endonuclease to cleave within the repeats, the bacterial Type II system uses the host-encoded RNase III, together with a CRISPR-associated, trans-encoded tracrRNA (Deltcheva *et al*., 2011). Furthermore, the various interference complexes exhibit considerable diversity (Barrangou and van der Oost, 2013; Makarova *et al*., 2011a; Vestergaard *et al*., 2014).

We developed an efficient tool for determining the strand from which mature crRNAs are derived by focussing on the repeats at CRISPR loci. The repeats are unique within the CRISPR-Cas system because they are the only element to play a vital role in all phases of immunity (Barrangou and van der Oost, 2013). Thus, despite their relatively short lengths, each repeat carries essential structural parameters or sequence motifs that are recognized by enzymes or structural proteins involved in adaptation, crRNA biogenesis and interference. Paradoxically, however, the repeats are very heterogeneous, occurring in a range of lengths, 19–48 nt, and display considerable sequence diversity. An early comparative study of CRISPR diversity yielded 12 main clusters with specific sequence characteristics; only a subset folded into characteristic hairpin structure motifs (Kunin *et al*., 2007). More recently, a major reevaluation of CRISPR conservation was executed by Lange *et al*. (2013), on a much larger data set of 3527 CRISPRs, where 40 conserved repeat sequence families were identified together with a total of 33 potential structural motifs. The repeat clusters were further classified into six superclasses, some of which showed strong biases to specific CRISPR subtypes and to certain bacterial or archaeal phyla (Lange *et al*., 2013).

CRISPR loci are generally identified by their characteristic repeat-spacer architecture. For example CRT (Bland *et al*., 2007) and CRISPRFinder (Grissa *et al*., 2007) provide sensitive predictions of CRISPR arrays, but do not provide unambiguous orientation information. In the literature, orientation is derived

*To whom correspondence should be addressed.

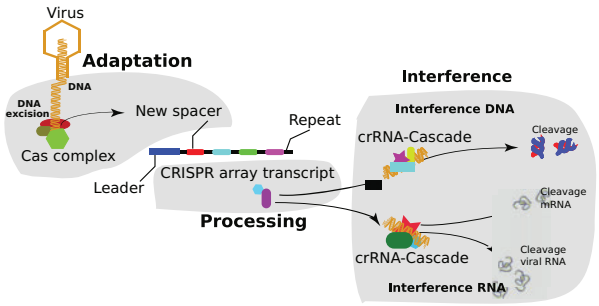


Fig. 1. The three major phases of CRISPR-Cas immune systems. First, in the adaptation phase, Cas proteins excise the protospacer sequence from foreign DNA and insert it into the repeat, adjacent to the leader at the CRISPR locus. Second, CRISPR arrays are transcribed and then processed into multiple crRNAs, each carrying a single spacer sequence and part of the adjoining repeat sequence. Third, at the interference phase, the crRNAs are assembled into different classes of protein targeting complexes (Cascades) that anneal to, and cleave, spacer matching sequences on either invading element or their transcripts

mainly by characteristic sequence motifs in the repeat, the detection of a conserved leader region in closely related CRISPR loci or by transcriptome experiments where the dominantly transcribed strand is determined. However, to date very few systems have been studied experimentally, and many large-scale studies require accurate orientation information for all available CRISPR arrays. Recently, Biswas *et al.* (2014) has presented the first tool to predict the orientation of CRISPR arrays. Their model is essentially a linear predictor based on a number of features which comprise the presence of the ATTGAAAN motif in repeats, a higher A or T content in the flanking regions of CRISPR arrays, nucleotide composition within the CRISPR array, the presence of mutations in specific parts of the array and the tendency to fold into a secondary structure. Each feature is considered as an independent predictor and is given a weight proportional to its estimated precision. The final prediction is computed as the weighted combination of each predictor.

Knowledge of the correct repeat orientation is crucial for accurate characterization of CRISPR conservation and for subsequently studying mechanisms of adaptation, CRISPR RNA processing and interference. In particular, it can help to (i) detect leader regions, currently poorly described in the literature; (ii) identify signals of transcription initiation and termination; (iii) determine the orientation of protospacers on invading genetic elements; and finally, (iv) characterize cognate protospacer-adjacent motifs (PAMs). Thus, we consider that the repeat orientation tool presented here will be of critical importance for future CRISPR-based experimental studies.

2 MATERIALS AND METHODS

We present a linear discriminative model based on graph kernels to accurately predict the orientation of the CRISPR sequence. The method first generates a sequence alignment of all repeat *instances* in the CRISPR array and outputs the consensus repeat sequence in its predicted orientation and whether it lies on the forward or reverse strand. There are two core ideas underlying our approach. The first one is to use a combinatorial technique to extract a very large number of features. The second idea is to encode our knowledge about the problem as a directed graph with discrete labels. The first idea allows a predictive system to be very accurate and to express complex discriminative decisions; the second idea allows a natural and flexible encoding of background knowledge.

2.1 Novel comprehensive identification of CRISPR loci

We extracted a comprehensive dataset of CRISPR loci from published archaeal and bacterial genomes. All genome sequences were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/). We predicted CRISPR loci using CRISPRFinder (Grissa *et al.*, 2007) and CRT (Bland *et al.*, 2007). For both tools, we used (i) default parameter values for predicted CRISPR loci and (ii) parameters that corresponded to at least two repeats within a CRISPR locus; repeat and spacer lengths were set to a range between 18 and 78 bp. We then (iii) generated a consensus repeat for each CRISPR locus exploiting the fact that repeats within a CRISPR locus are almost completely identical with some loci that carry few mutations, preferably at the start and end (see Supplementary Figure S3, Supplementary material). Because CRT does not output consensus repeats, we used the MAFFT program (Katoh *et al.*, 2002), version 6.4., to compute the multiple alignments and the Cons program from EMBOSS package (Rice *et al.*, 2000) to obtain the consensus repeat from the multiple sequence alignments. Finally, (iv) the results from both CRISPRFinder and CRT tools were merged and redundant CRISPR loci were removed. In this way, we obtained a CRISPR databases with >4700 consensus repeats, which we refer to as REPEATS (see Table 1 for details).

2.2 Datasets from literature

2.2.1 Set of repeats from Lange et al. (2013) We selected structural motifs that fit to known cleavage sites (Lange et al., 2013). Table 2 gives a summary of published CRISPR-Cas systems with experimental evidence for the processing mechanism which we refer to as REPEATS_{Lange}. This dataset contains 324 bacterial and 118 archaeal repeat sequences (442 in total).

2.2.2 Set of repeats from Kunin et al. (2007) We denote the dataset originally published in Kunin *et al.* (2007) as REPEATS_{Kunin}. The dataset contains 327 bacterial and 92 archaeal repeat sequences (419 in total). The orientations were assigned by the authors using previously published sequence features.

2.2.3 Set of archaeal repeats from Shah and Garrett (2011) We denote the dataset based on the results available in (Shah and Garrett, 2011) as REPEATS_{Shah}. This dataset contains 478 archaeal repeat sequences with manually verified strand orientation.

2.3 Encoding CRISPR repeats as graphs

The features used to discriminate between the different orientation are based on available biological knowledge of CRISPR evolution and processing. During CRISPR RNA processing by Cas6-like endoribonucleases, cleavage occurs either at the 3′-end base of the hairpin motif, or within the double-stranded region of the hairpin stem, usually below a $C \rightarrow G$ base pair (Barrangou and van der Oost, 2013; Richter *et al.*, 2012; Scholz *et al.*, 2013). The product of this cleavage is an 8-nt-long AUUGAAA(N) repeat tag at the 5′-end of the mature crRNA (5′-tag), which corresponds to the last eight nucleotides from the 3′-end of the repeat sequence. Kunin *et al.* (2007) and Lange *et al.* (2013) showed that in some cases the four nucleotides AAA(N) motif can be used to identify the orientation. These observations lead to the hypothesis that the terminal region of the sequence, comprising four or eight nucleotides, plays a key role. We observed also that the mutation rate in various parts of the CRISPR locus is non-uniform, in particular the middle part of the CRISPR locus is more conserved. This finding motivated

the idea of using the presence of mutations as an additional signal to detect the predominantly transcribed strand.

We made use of this background knowledge to partition the consensus repeat into specific informative parts: we distinguish terminal regions of identical size k at both ends (as the correct orientation is unknown) and a central variable length area. The terminal sequences are further partitioned into P equally sized parts, where we expect to find key motifs. We call each part *block*. One of the main signals that we used to define the number and size of the blocks is the mutation rate, defined as the fraction of mutations per nucleotide in each block. In Supplementary Figure S3, we report the mutation rate for the CRISPR locus partitioning with $k = 8$ and $P = 2$ on a dataset of 897 CRISPR arrays (Kunin *et al.*, 2007; Shah and Garrett, 2011): each repeat is split into five adjacent regions, with terminal blocks spanning exactly 4 nucleotides and a central block spanning 12 nucleotides on average. In these settings, we observed a highly significant 4-fold and a 16-fold increase in the mutation rate in the initial 8 nucleotides and in the terminal block, respectively, as compared to the middle block. In Section 3.1, we have further validated the optimality of this partitioning with *in silico* simulations.

We encoded all our intuitions and knowledge on the relevant signals that a predictive model should be aware of in a graph data structure. The reason for this choice is 2-fold. First, we want an easy and natural way to inject different types of information in the problem solution, and, second, we want to exploit efficient techniques developed in the Machine Learning literature to automatically construct a large number of derived features to improve the accuracy of predictive models.

The graph formalism allows us, in a very natural and flexible way, to add knowledge by inserting informative entities as vertices and connecting them to the relevant parts of the current encoding via the edge notion. In our case, the information provided by the consensus sequence is modelled directly as a path graph with vertices labelled with the consensus nucleotide code (see Fig. 2). We then model the global localization information as additional vertices with a label that indicates the block identity. This reveals whether a nucleotide is located at the very beginning or just near the beginning of the sequence (and symmetrically for the opposite end). Furthermore, we consider a more fine grained localization information, identifying the specific position of a nucleotide within a block. The reason to encode an increasingly refined localization information is to allow the algorithm to choose the optimal level of detail needed in various parts of the sequence. Finally, the main piece of information is whether there is evidence of a mutation at a specific location; we model this with an additional vertex labelled with a binary code to indicate the presence of a mutation in at least one of the repeated sequences.

Table 1. Summary of our REPEATS dataset derived from all available CRISPR loci

Data statistics	Archaea		Bacteria	
Genomes (total)	309		4590	(4899)
Genomes with CRISPRs (%)	217	(70)	1409	(30)
CRISPRs on forward strand	516		1810	(2326)
CRISPRs on reverse strand	530		1859	(2389)
Repeats per array (median)	2–198	(20)	2–1371	(16)
Repeat lengths (median)	20–44	(29)	19–48	(30)
Spacer lengths (median)	20–54	(38)	19–72	(35)

Table 2. Summary of REPEATS_{Lange} dataset: published CRISPR-Cas systems with experimental evidence of the processing mechanism

Organism	Motif	Cas subtype	Summary
Escherichia coli K12	M2	I-E	Structure predicted, but stable; 8-nt-5′-tag; cleavage by **Cas6e**, biochemical experiments (Brouns *et al.*, 2008)
Thermus thermophilus HB8	M2	I-E	Structured; 8-nt-5′-tag; cleavage by **Cas6e**; crystal structure of repeat hairpin in Cas6e (Cse3) (Gesner *et al.*, 2011; Juranek *et al.*, 2012; Sashital *et al.*, 2011)
Bacillus halodurans C-125	M3	I-C	Cleavage by **Cas5d**; 11-nt-5′-tag mutational analysis of hairpin structure (Nam *et al.*, 2012)
Pseudomonas aeruginosa UCBPP-PA14	M4	I-F	Cleavage by Cas6f (Csy4); 8-nt-5′-tag; crystal structure and mutational analyses of repeat hairpin in **Cas6f** (Haurwitz *et al.*, 2010, 2012; Sternberg *et al.*, 2012)
Synechocystis sp. PCC6803	M5	I-DIII-variant	Cleavage by **Cas6**; 8-nt-5′-tag; biochemical experiments, extended structure prediction of hairpin motif (Scholz *et al.*, 2013)
Thermus thermophilus HB27	M9	I-C	Cleavage by **Cas5d**; 11-nt-5′-tag biochemical experiments (Garside *et al.*, 2012)
Methanosarcina marzei Gö1	M13	I-B III-B	Cleavage by **Cas6b**; 8-nt-5′-tag; structure probing experiment of hairpin (Nickel *et al.*, 2013)
Synechocystis sp. PCC6803	M14	III-variant	Biochemical analysis of **Cmr2** implicate its involvement in either cleavage, crRNA stabilization, or array expression regulation; 13-nt-5′-tag (Scholz *et al.*, 2013)
Staphylococcus epidermidis RP62A	M28	III-A	Cleavage by **Cas6**; 8-nt-5′-tag; hairpin structure as in M28 verified by mutational analysis and sequence specificity around cleavage site (Hatoum-Aslan *et al.*, 2011)
Methanococcus maripaludis C5	M29	I-B	Cleavage by **Cas6b**; 8-nt-5′-tag; biochemical experiments (Richter *et al.*, 2012)

Note: In particular, these are systems for which (i) the Cas endoribonuclease has been characterized and/or (ii) the repeat structure has been verified. Published results are consistent with the data of Lange *et al.* (2013).

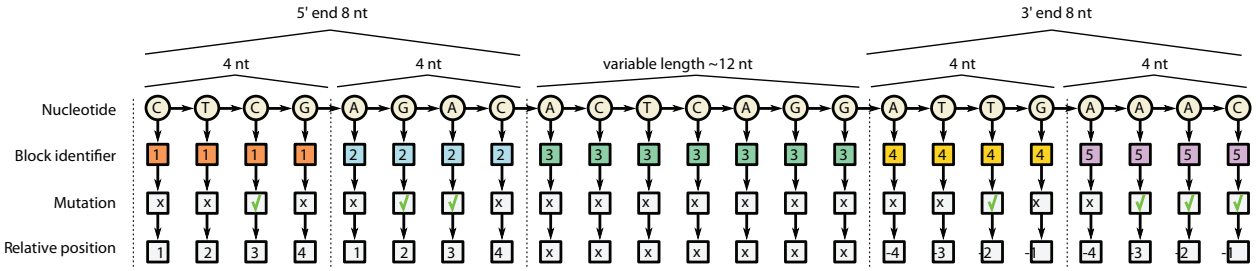


Fig. 2. Graph encoding the consensus repeat sequence. The consensus nucleotide information is represented as a path graph, and additional information is modelled as a chain of additional vertices. The terminal parts of the repeat are marked with block identifiers

The final modelling decision regards the topology of the graph, i.e. how the additional vertices, which encode the different types of information, should be connected together. We identify an order which reflects the importance of the different types of information, starting from the nucleotide type, the block ID, the mutation evidence and finally the relative position within a block. Note that the combinatorial feature generation phase is affected by the sequential order of these attributes, as the information that is ranked higher will participate in the generation of more features and will therefore be regarded as more prominent.

2.4 Predictive model and feature extraction

After having encoded domain expert knowledge as a graph, we need to process this type of structured data to induce a predictive model. We do this using the technique developed by Costa and Grave (2010), based on the notion of graph kernels. The core idea (see Supplementary Information for a formal description) is to decompose each graph in a (multi) set of fragments and use these as features, in a similar fashion to what is done in the chemoinformatics domain with the *fingerprint* technique. The resulting sparse vectors can then be processed by efficient machine learning techniques, such as the stochastic gradient descent SVM (Bottou, 2010), to yield fast and highly predictive models. The type of graph decomposition that we use is called Neighbourhood Subgraph Pairwise Distance Kernel (NSPDK), and it involves the extraction of all possible pairs of small neighbourhood subgraphs that are not too distant (see Fig. 3). Intuitively one can think about this type of decomposition as an upgrade of the concept of k-mers with gaps from the domain of strings to that of graphs. Both the extraction of the features and the training of the predictive model have linear complexity and offer therefore excellent scaling capability. More precisely, extracting all neighbourhood subgraphs is achieved with a breadth-first visit for a limited depth starting from each node, and as the graphs are sparse, it takes $O(|n * m|)$ where n is the number of nucleotides and m the number of repeat alignments.

Finally, given that one of the two strands can be the one that exhibits a characteristic pattern, we train a predictive model on both variants of each repeat sequence: one obtained from the forward strand and the other from the complementary reverse strand. The binary task is therefore to assign a positive score to the sequences that are transcribed and a negative one to the complementary strand. In the predictive phase, we enforce consistency by considering the prediction on both variants of the sequence: a strong confidence of the prediction of the forward strand should also correspond to an equally confident prediction that the reverse complementary sequence is not transcribed. To do so, we simply perform the individual predictions and then average the prediction of the forward strand with the *opposite* prediction for the reverse strand. If the resulting score is positive, then the forward strand is predicted to be transcribed, whereas the reverse strand is selected if the score is negative.

3 RESULTS AND DISCUSSION

3.1 Parameter selection

We have previously described how relevant biological knowledge was used to determine various modelling choices. The proposed model admits, however, different configurations both in the encoding part as well as in the combinatorial feature generation part. To determine the best configuration, we therefore performed extensive *in silico* simulations. More specifically, the encoding phase allows the following parametric variants: (i) choice of attribute type (i.e. whether to use the mutation information or the block identity); (ii) choice of attribute order (i.e. whether the block identifier should precede the mutation marker or vice versa); (iii) size of the terminal regions (more, equal or less than 8 nucleotides); (iv) number of blocks within the terminal regions (1, 2 or 3). The combinatorial feature construction phase is parametrized instead by the maximal radius R and distance D, where larger values for R translates in more complex features and larger values for D in an increased tolerance for larger gaps.

For each model variant, we designed a selection experiment to identify the best configuration of parameters as the one that achieves the minimum expected predictive error. Not surprisingly, results are consistent with the background knowledge that originally motivated the encoding, that is, the best model uses all attributes in the order presented in Figure 2 with terminal regions of size 8 nucleotides divided into blocks of 4 nucleotides. We observed that the actual attribute order had just a modest influence on the results (see Supplementary Table S1).

3.1.1 Choice of attribute type We estimated the expected prediction error of five different encodings, which use an increasing amount of information. We denote them with $model_i$ with $i \in \{1, 2, 3, 4, 5\}$ (consider Fig. 2 as a reference). In all cases blocks have a constant size of 4 nucleotides.

- $model_1$: nucleotide sequence only (layer 1 in Fig. 2)
- $model_2$: nucleotide sequence with additional mutation attribute for the terminal 8 nucleotides (layer 1 + 3 in Fig. 2)
- $model_3$: nucleotide sequence with additional block attribute (layer 1 + 2 in Fig. 2)
- $model_4$: nucleotide sequence with mutation and block attribute (layer 1 + 2 + 3 in Fig. 2)
- $model_5$: nucleotide sequence with block, mutation and relative position attribute (layer 1 to 4 in Fig. 2)

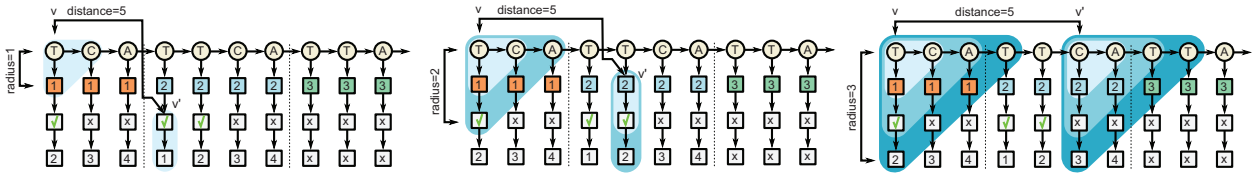


Fig. 3. The NSPDK approach extracts a large number of features taking only specific fragments into account. The procedure is parametrized by the radius R and the distance D. Each vertex is considered in turn as a root. A neighbourhood graph of radius R is extracted around each root. All possible pairs of neighbourhood graphs of the same size R are considered, provided that their respective roots are exactly at distance D. To understand the importance of the sequential order of the attributes consider the left part of the figure: here we depict a feature with radius 1 and distance 0, which will encode three pieces of information: (i) the specific dinucleotide combination, (ii) the block ID and (iii) whether a mutation is likely to occur on the first nucleotide of the dinucleotide. As we increase the maximal distance between the roots in the pair, the encoded information is further specialized. In the middle part of the figure, we show a feature that additionally includes the presence of a mutation at distance 5. When the radius is increased to 2, the specific position within the block is also considered

To evaluate the generalization capacity of the resulting predictive models, we used as training material the 442 sequences in REPEATS_{Lange} and as test material the 419 REPEATS_{Kunin} + 478 sequences REPEATS_{Shah} filtered so as to guarantee a maximal pre-specified level of *sequence identity* w.r.t. the training material. In Figure 4, we report the area under the curve for the receiver operator characteristic (AUC ROC) when the test material has pairwise sequence identity $\leq$0.95, 0.85, 0.75 and 0.65, respectively, as measured by the Needelman–Wunsch algorithm (Needleman and Wunsch, 1970).

The simulations show that the mutation information does indeed provide good discriminative features (increasing performance of 5%) and that partitioning the sequence into blocks can further improve the predictive performance (an extra 10%). Finally, the model is shown to yield ~0.85 AUC ROC when tested on sequences with only 0.65 sequence identity, indicating a reasonable generalization capacity to evolutionary distant sequences. Note that extrapolating the predictive tendency with a quadratic fit, we get a random AUC ROC of 0.53 at 25% sequence similarity, i.e. for random sequences.

3.1.2 Choice of terminal region size We validated the notion of a most informative leading part of the consensus repeat sequence via an *in silico* simulation. An encoding was created that uses only three blocks: two terminal ones of fixed size k nucleotides and a central one of variable length. We computed the average AUC ROC in a 10-fold cross validation for $k = 1, \ldots, 10$. Results shown in Supplementary Figure S1 are in striking agreement with our biological knowledge, with clear performance peaks at exactly 4 and 8 nucleotides.

3.1.3 Choice of number of blocks within the terminal regions We also validated the notion that there is an advantage in considering a finer partition of the terminal parts. We started from an encoding with terminal regions spanning 8 nucleotides and then we subdivided them into 1, 2 or 4 equal sized subparts, that is, in subparts of 8, 4 and 2 nucleotides. Once again results (shown in Supplementary Figure S2) are in agreement with the biological findings, and confirm that a subdivision in 4 nucleotide parts is indeed beneficial.

3.1.4 Combinatorial features The complexity of the derived feature representation depends on the maximum radius R and maximum distance D that are considered. Using $model_5$, we simulated all possible combinations of values $R = \{0, \ldots, 7\}$ and $D = \{1, \ldots, 7\}$ (see Supplementary Table S2) in a 10-fold cross-validated experiment on the REPEATS_{Lange} and obtained the best predictive performance with $R = 3$ and $D = 5$. Note that, unsurprisingly, the optimal size $R = 3$ is also the minimal size that allows to capture all available attributes in $model_5$.

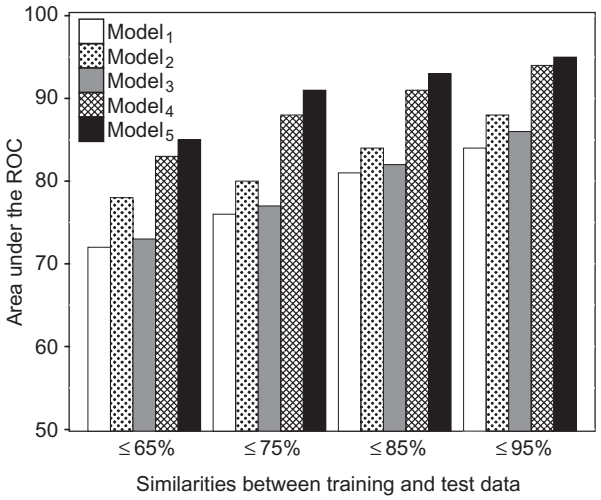


Fig. 4. AUC ROC performance comparison of the five models that encode increasing amount of information about the CRISPR arrays

3.2 Comparison with Biswas *et al.* (2014)

We used the same dataset as in Biswas *et al.* (2014) to train our model. Both methods were then applied to the REPEATS_{Shah} data set, filtered for decreasing levels of sequence identity w.r.t. the training set. In Figure 5 we report the comparative AUC ROC performance and observe that our proposal offers a substantial improvement both in prediction performance and in generalization capacity with a less pronounced degradation as the sequence identity decreases.

Finally, we measured the runtime for both approaches on 956 CRISPR repeat arrays (average length 28 nucleotides). The classification task was completed in 59 s by our approach and in 37 min by the Biswas predictive model. We report that the Biswas tool failed to make any prediction in 98 cases out of 948, while our method achieved an AUC ROC of 0.89 on the

same instances, indicating that these sequences were on average only slightly more difficult to predict.

3.3 CRISPR-Cas system annotation

We used our orientation prediction method to identify the transcribed strand for the set of 3527 repeats available from Lange *et al.* (2013) and for the novel set of 4719 individual CRISPR loci identified as described in Section 2.1. This material was finally used to update the CRISPRmap web server, which provides an automated and easy-to-use classification of all currently available and newly sequenced CRISPRs.

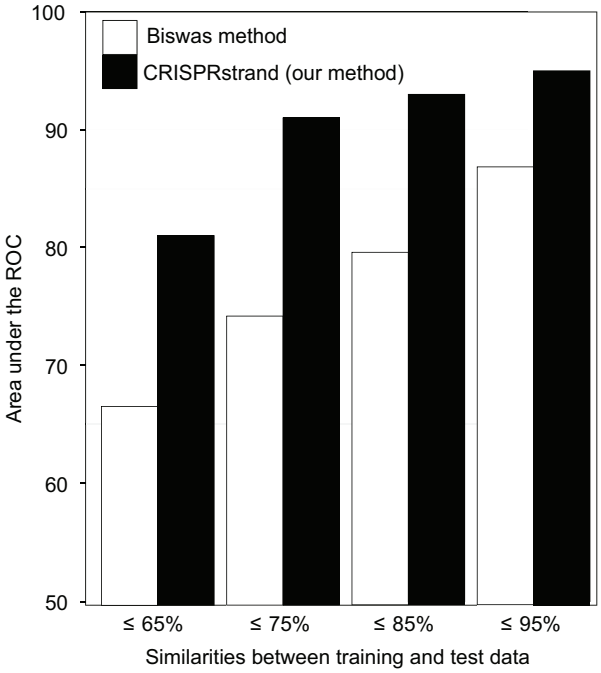


Fig. 5. Performance comparison between our method and Biswas method. The test database contains 948 CRISPR repeats

3.3.1 Re-correcting the orientation of 3527 repeats from Lange et al. (2013) Our tool was run on 3527 repeats, which were then clustered into 40 conserved sequence families, 33 potential structural motifs and 6 major superclasses. In this set, we identified 536 repeats with incorrect orientation (see Supplementary Table S12). Next we ran our cluster pipeline for three iterations, retrieving 29 potential structural motifs and 37 conserved sequences families (see Supplementary Tables S3 and S4). As shown in Figure 6, the orientation of F8 and M6 was incorrect. Using corrected orientations, we could merge F8 with F6, and M5 with M6. Overall, in Figure 6, we show how the cluster quality can be significantly improved when we can make use of a better orientation prediction.

3.3.2 Update of CRISPRmap web server to version 2.0 The database REPEATS of 4719 individual CRISPR loci was collected performing an exhaustive search for CRISPR loci within all available bacterial and archaeal genomes (see Table 1). We developed two independent clustering approaches to identify structural motifs and conserved sequence families. In both approaches, we call a cluster of structural motifs or conserved sequences a *class* if they contain CRISPR repeats which come from 10 different species (see Supplementary Tables S5–S11 and CRISPRmap). The results of our independent clustering approaches are as follows: (i) 18 structure motifs were identified based on sequence and structure alignments using LocARNA (Smith *et al.*, 2010; Will *et al.*, 2007, 2012). Structure motif candidates were constrained to be similar to those previously published (Brouns *et al.*, 2008; Hatoum-Aslan *et al.*, 2011; Nam *et al.*, 2012; Nickel *et al.*, 2013; Sashital *et al.*, 2011; Scholz *et al.*, 2013; Sternberg *et al.*, 2012). (ii) Twenty-four conserved sequence families were identified based on Markov clustering (Enright *et al.*, 2002). Full details of structure motifs and conserved sequence families are available in the Supplementary file and in full on CRISPRmap web server. We grouped all the sequences available in the REPEATS database into six major superclasses (labelled A to F) based on sequence and structure similarities and tree topology (Supplementary Figure S6). Owing to the corrected orientation, there are two main differences between superclasses from Lange *et al.* (2013) and current superclasses. First, superclasses B and C were merged together and the

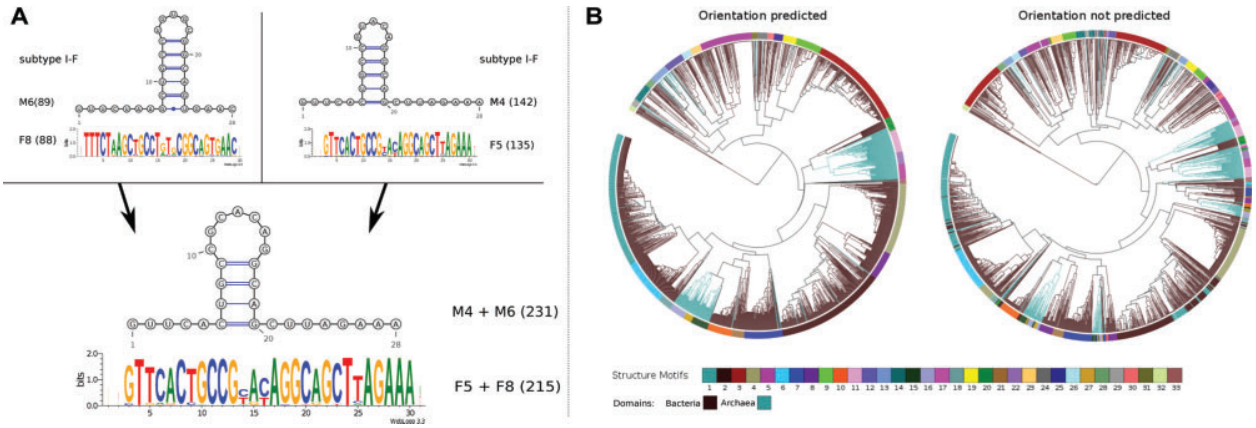


Fig. 6. (**A**) Given the novel predicted orientation Family 5 with Family 8 and Motif 4 with Motif 6 could be merged. (**B**) The 33 structural motifs from Lange *et al.* (2013) are clustered (i) with the orientation prediction; (ii) without orientation prediction

resulting new superclass was called B. Second, parts of superclasses E and F were moved to superclass D (Supplementary Figure S6).

Archaea CRISPR-Cas subtype annotation from Vestergaard et al. (2014) A very recent study has classified archaeal CRISPR-Cas systems into two main types, called Type I and Type III and 12 subtypes (Vestergaard *et al.*, 2014). We annotated all archaea CRISPR loci based on these subtypes. For genomes which became available after this study was completed, we annotated them following the procedure employed in the *cas* gene cassette study (Vestergaard *et al.*, 2014). To assign subtypes to specific CRISPR loci automatically, we first identified the distance of the closest *cas* gene cassette subtype to each CRISPR locus. Second, we plotted the distances and determined a clear peak (Supplementary Figure S9 in Supplementary Material). Finally, we used the peak as a cut-off to assign CRISPR-Cas subtypes to specific CRISPR loci.

CRISPR-Cas subtype annotation from Makarova et al. (2011) We extracted all genes from all available bacterial genomes. We then searched for all *cas* genes using a recent version of TIGRFAM models from Haft *et al.* (2005, 2013) in combination with HMMER (Eddy, 2011). A *cas* gene was annotated when one of its respective models was found with an E-value ≤ 0.0001. Next, we took the results and searched them against protein family databases CDD (Makarova *et al.*, 2011a), COG (Makarova *et al.*, 2006) and Pfam (Punta *et al.*, 2012) using RPS-Blast (Marchler-Bauer *et al.*, 2011). Then, we generated new models and supermodels from those databases. Finally, we used the new models to annotate all *cas* genes based on Makarova *et al.* (2011a,b) classification. We assigned *cas* subtype to CRISPR loci in the same way as in the previous subsection.

4 CONCLUSION

We presented a highly flexible approach to accurately predict the transcribed strand of CRISPR loci. The method is motivated by recent findings and encodes the most relevant information in the form of a graph structure that can be efficiently processed with graph kernel methods. Our tool compares favourably against a recent approach proposed in Biswas *et al.* (2014) in terms of accuracy (0.95 compared to 0.88 AUC ROC), runtimes (59 s rather than 37 min on a 1K sequences dataset) and coverage (we achieve 0.89 AUC ROC on the 10% sequences that the Biswas tool fails to classify).

Our approach was integrated in CRISPRmap (Lange *et al.*, 2013) to improve the accuracy of the previously published classification of CRISPRs, and resulted in: (i) a comprehensive dataset with >4500 consensus repeats; (ii) the most recent classification of Cas subtypes based on Cas-protein occurrences for archaea (Vestergaard *et al.*, 2014); and (iii) an improved annotation of Makarova Cas subtypes for bacteria respecting the rules published in Makarova *et al.* (2011a).

The orientation prediction approach that we have presented is fast, accurate and can be easily integrated in existing pipelines. In future work, we will employ it to ease the identification of novel targets (protospacers), PAM motifs and the investigation of regulatory motifs in the leader sequences of CRISPR arrays.

ACKNOWLEDGEMENT

The authors thank Martin Mann for his help with the webserver.

Funding: This work was funded by the German Research Foundation (DFG) program FOR1680 'Unravelling the Prokaryotic Immune System' (BA 2168/5-1 to R.B.).

Conflict of interest: none declared.

REFERENCES

Barrangou,R. and van der Oost,J. eds. (2013) *CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea*. Springer Press, Heidelburg, Germany, pp. 1-129.

Biswas,A. *et al.* (2014) Accurate computational prediction of the transcribed strand of CRISPR noncoding RNAs. *Bioinformatics*, **30**, 1805–1813.

Bland,C. *et al.* (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.

Bottou,L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent. In: *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*. Springer, pp. 177–187.

Brouns,S.J.J. *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.

Costa,F. and Grave,K.D. (2010) Fast neighborhood subgraph pairwise distance kernel. In: *Proceedings of the 26th International Conference on Machine Learning*. Omnipress, pp. 255–262.

Deltcheva,E. *et al.* (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, **471**, 602–607.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Garside,E.L. *et al.* (2012) Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA*, **18**, 2020–2028.

Gesner,E.M. *et al.* (2011) Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat. Struct. Mol. Biol.*, **18**, 688–692.

Grissa,I. *et al.* (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.

Haft,D.H. *et al.* (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.*, **1**, e60.

Haft,D.H. *et al.* (2013) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.

Hatoum-Aslan,A. *et al.* (2011) Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc. Natl Acad. Sci. USA*, **108**, 21218–21222.

Haurwitz,R.E. *et al.* (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, **329**, 1355–1358.

Haurwitz,R.E. *et al.* (2012) Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *EMBO J.*, **31**, 2824–2832.

Juranek,S. *et al.* (2012) A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs. *RNA*, **18**, 783–794.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Kunin,V. *et al.* (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.*, **8**, R61.

Lange,S.J. *et al.* (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.*, **41**, 8034–8044. SJL, OSA and DR contributed equally to this work.

Makarova,K.S. *et al.* (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct.*, **1**, 7.

Makarova,K.S. *et al.* (2011a) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.

Makarova,K.S. *et al.* (2011b) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct.*, **6**, 38.

Marchler-Bauer,A. *et al.* (2011) CDD: a conserved domain database for the functional annotation of proteins. *Database*, **39**, D225–D229.

Nam,K.H. *et al.* (2012) Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure*, **20**, 1574–1584.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Nickel,L. *et al.* (2013) Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III. *RNA Biol.*, **10**, 779–791.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Rice,P. *et al.* (2000) EMBOSS: the European Molecular Biology open software suite. *Trends Genet.*, **16**, 276–277.

Richter,H. *et al.* (2012) Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. *Nucleic Acids Res.*, **40**, 9887–9896.

Sashital,D.G. *et al.* (2011) An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat. Struct. Mol. Biol.*, **18**, 680–687.

Scholz,I. *et al.* (2013) CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS One*, **8**, e56470.

Shah,S.A. and Garrett,R.A. (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol.*, **162**, 27–38.

Smith,C. *et al.* (2010) Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA. *Nucleic Acids Res.*, **38** (**Suppl**), W373–W377.

Sternberg,S.H. *et al.* (2012) Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA*, **18**, 661–672.

Vestergaard,G. *et al.* (2014) CRISPR adaptive immune systems of Archaea. *RNA Biol.*, **11**, 157–168.

Will,S. *et al.* (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.

Will,S. *et al.* (2012) LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, **18**, 900–914.

An updated evolutionary classification of CRISPR-Cas systems

Kira S. Makarova, Yuri I. Wolf, **Omer S. Alkhnbashi**, Fabrizio Costa, Shiraz A. Shah, Sita J. Saunders, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Daniel H. Haft, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Rebecca M. Terns, Michael P. Terns, Malcolm F. White, Alexander F. Yakunin, Roger A. Garrett, John van der Oost, Rolf Backofen and Eugene V. Koonin. **Nature Reviews Microbiology Journal, 2015**, doi:10.1038/nrmicro3569.

Personal contribution

I have made an important contribution in this project. I implemented the automated clustering method and investigated the correlation between CRISPR locus and CRISPR subtypes. Furthermore, I was involved in the planning and writing the article.

Omer S. Alkhnbashi

The following co-authors confirm the above stated contribution.

Kira S. Makarova

Fabrizio Costa

Sita J. Saunders

Rolf Backofen

Eugene V. Koonin

Nature Reviews Microbiology | AOP, published online 28 September 2015; doi:10.1038/nrmicro3569

ANALYSIS

An updated evolutionary classification of CRISPR–Cas systems

Kira S. Makarova[1], Yuri I. Wolf[1], Omer S. Alkhnbashi[2], Fabrizio Costa[2], Shiraz A. Shah[3], Sita J. Saunders[2], Rodolphe Barrangou[4], Stan J. J. Brouns[5], Emmanuelle Charpentier[6], Daniel H. Haft[1], Philippe Horvath[7], Sylvain Moineau[8], Francisco J. M. Mojica[9], Rebecca M. Terns[10], Michael P. Terns[10], Malcolm F. White[11], Alexander F. Yakunin[12], Roger A. Garrett[3], John van der Oost[5], Rolf Backofen[2,13] and Eugene V. Koonin[1]

Abstract | The evolution of CRISPR–*cas* loci, which encode adaptive immune systems in archaea and bacteria, involves rapid changes, in particular numerous rearrangements of the locus architecture and horizontal transfer of complete loci or individual modules. These dynamics complicate straightforward phylogenetic classification, but here we present an approach combining the analysis of signature protein families and features of the architecture of *cas* loci that unambiguously partitions most CRISPR–*cas* loci into distinct classes, types and subtypes. The new classification retains the overall structure of the previous version but is expanded to now encompass two classes, five types and 16 subtypes. The relative stability of the classification suggests that the most prevalent variants of CRISPR–Cas systems are already known. However, the existence of rare, currently unclassifiable variants implies that additional types and subtypes remain to be characterized.

[1]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. Correspondence to E.V.K. e-mail: koonin@ncbi.nlm.nih.gov*
doi:10.1038/nrmicro3569
Published online
28 September 2015

The CRISPR–Cas modules are adaptive immune systems that are present in most archaea and many bacteria[1–5] and provide sequence-specific protection against foreign DNA or, in some cases, RNA[6]. A CRISPR locus consists of a CRISPR array, comprising short direct repeats separated by short variable DNA sequences (called 'spacers'), which is flanked by diverse *cas* genes. CRISPR–Cas immunity involves three distinct mechanistic stages: adaptation, expression and interference[7–11]. The adaptation stage involves the incorporation of fragments of foreign DNA (known as 'protospacers') from invading viruses and plasmids into the CRISPR array as new spacers. These spacers provide the sequence memory for a targeted defence against subsequent invasions by the corresponding virus or plasmid. During the expression stage, the CRISPR array is transcribed as a precursor transcript (pre-crRNA), which is processed and matured to produce CRISPR RNAs (crRNAs). During the interference stage, crRNAs, aided by Cas proteins, function as guides to specifically target and cleave the nucleic acids of cognate viruses or plasmids[7,9,12,13]. Recent studies suggest that CRISPR–Cas systems can also be used for non-defence roles, such as the regulation of collective behaviour and pathogenicity[14–16].

Numerous, highly diverse Cas proteins are involved in the different stages of CRISPR activity (BOX 1; see Supplementary information S1 (table)). Briefly, Cas1 and Cas2, which are present in most known CRISPR–Cas systems, form a complex that represents the adaptation module and is required for the insertion of spacers into CRISPR arrays[17,18]. Protospacer acquisition in many CRISPR–Cas systems requires recognition of a short protospacer adjacent motif (PAM) in the target DNA[19–22]. During the expression stage, the pre-crRNA molecule is bound to either Cas9 (which is a single, multidomain protein) or to a multisubunit complex, forming the crRNA–effector complex. The pre-crRNA is processed into crRNAs by an endonuclease subunit of the multisubunit effector complex[23] or via an alternative mechanism that involves bacterial RNase III and an additional RNA species, the tracrRNA (transactivating CRISPR RNA)[24]. Finally, at the interference stage, the mature crRNA remains bound to Cas9 or to the multisubunit crRNA–effector complex, which recognizes and cleaves the cognate DNA[10,11,25,26] or RNA[26–31].

The rapid evolution of most *cas* genes[32–34] and the remarkable variability in the genomic architecture of CRISPR–*cas* loci poses a major challenge for the

consistent annotation of Cas proteins and for the classification of CRISPR–Cas systems[13,35]. Nevertheless, a consistent classification scheme is essential for expedient and robust characterization of CRISPR–*cas* loci in new genomes, and thus important for further progress in CRISPR research. Owing to the complexity of the gene composition and genomic architecture of the CRISPR–Cas systems, any single, all-encompassing classification criterion is rendered impractical, and thus a 'polythetic' approach based on combined evidence from phylogenetic, comparative genomic and structural analysis was developed[13]. At the top of the classification hierarchy are the three main types of CRISPR–Cas systems (type I–type III). These three types are readily distinguishable by virtue of the presence of unique signature proteins: Cas3 for type I, Cas9 for type II and Cas10 for type III[13]. Within each type of CRISPR–Cas system, several subtypes have been delineated based on additional signature genes and characteristic gene arrangements[13,35]. Recently, in-depth sequence and structural analysis of the effector complexes from different variants of CRISPR–Cas systems has uncovered common principles of their organization and function[4,30,31,36–46]. In parallel, the biotechnological development of molecular components of type II CRISPR–Cas systems into a powerful new generation of genome editing and engineering tools has triggered intensive research into the functions and mechanisms of these systems, thereby advancing our understanding of the Cas proteins and associated RNAs[47,48].

In this Analysis article, we refine and extend the classification of CRISPR–*cas* loci based on a comprehensive analysis of the available genomic data. As a result of this analysis, we introduce two classes of CRISPR–Cas systems as a new, top level of classification and define two putative new types and five new subtypes within these classes, resulting in a total of five types and 16 subtypes. We employ this classification to analyse the evolutionary relationships between CRISPR–*cas* loci using several measures. The results of this analysis highlight pronounced modularity as an emerging trend in the evolution of CRISPR–Cas systems. Finally, we demonstrate the potential for automated annotation of CRISPR–*cas* loci by developing a computational approach that uses the new classification to assign CRISPR–Cas system subtype with high precision.

Author addresses

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.
[2]Bioinformatics group, Department of Computer Science, University of Freiberg, Georges-Kohler-Allee 106, 79110 Freiberg, Germany.
[3]Archaea Centre, Department of Biology, Copenhagen University, Ole Maaløes Vej 5, DK2200 Copenhagen N, Denmark
[4]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, North Carolina 27606, USA.
[5]Laboratory of Microbiology, Wageningen University, Dreijenplein 10, 6703HB Wageningen, Netherlands.
[6]Department of Regulation in Infection Biology, Helmholtz Centre for Infection Research, D-38124 Braunschweig, Germany.
[7]DuPont Nutrition and Health, BP10, Dangé-Saint-Romain 86220, France.
[8]Département de Biochimie, de Microbiologie et de Bio-informatique, Faculté des Sciences et de Génie, Groupe de Recherche en Écologie Buccale, Félix d'Hérelle Reference Center for Bacterial Viruses, Faculté de médecine dentaire, Université Laval, Québec City, Québec, Canada.
[9]Departamento de Fisiología, Genética y Microbiología. Universidad de Alicante. 03080-Alicante, Spain.
[10]Biochemistry and Molecular Biology, Genetics and Microbiology, University of Georgia, Davison Life Sciences Complex, Green Street, Athens, Georgia 30602, USA.
[11]Biomedical Sciences Research Complex, University of St Andrews, North Haugh, St Andrews, KY16 9TZ, UK.
[12]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, M5S 3E5, Canada.
[13]BIOSS Centre for Biological Signaling Studies, Cluster of Excellence, University of Freiburg, Germany.

Classification of CRISPR–*cas* loci

The classification of CRISPR–Cas systems should ideally represent the evolutionary relationships between CRISPR–*cas* loci. However, the pervasive exchange and divergence of *cas* genes and gene modules has resulted in a complex network of evolutionary relationships that cannot be readily (and cleanly) partitioned into a small number of distinct groupings (although such partitioning might be achievable for individual modules, see below). Therefore, we adopted a two-step classification approach that first identified all *cas* genes in each CRISPR–*cas* locus and then determined the signature genes and distinctive gene architectures that would allow the assignment of these loci to types and subtypes.

To robustly identify *cas* genes, which is a non-trivial task owing to high sequence variability, we developed a library of 394 position-specific scoring matrices (PSSM)[49] for all 93 known protein families associated with CRISPR–Cas systems (see Supplementary information S2 (table)). Importantly, this set included 229 PSSMs for recently characterized families that were not part of the previous CRISPR–Cas classification[13]. The PSSMs were used to search the protein sequences annotated in 2,751 complete archaeal and bacterial genomes that were available at the National Center for Biotechnology Information (NCBI) as of 1 February 2014 (see Supplementary information S3 (box) for a detailed description of the methods). A highly significant similarity threshold was used to identify bona fide *cas* genes. Genes that were located in the same genomic neighbourhood as bona fide *cas* genes (irrespectively of their proximity to a CRISPR array) and that encoded proteins with moderate similarity to Cas PSSMs were then identified as putative *cas* genes. This two-step procedure was devised to minimize the false-positive rate, while allowing the detection of diverged variants of Cas proteins.

Gene neighbourhoods around the identified *cas* genes were merged into 1,949 distinct *cas* loci from 1,302 of the 2,751 analysed genomes, including 1,694 complete loci. A *cas* locus was annotated as 'complete' if it encompassed at least the full complement of genes for the main components of the interference module (the multisubunit crRNA–effector complex or Cas9). This criterion was adopted because, although the adaptation module genes *cas1* and *cas2* are the most common *cas* genes, many otherwise complete (and hence thought to be

Box 1 | **Cas protein families and functional modules**

The Cas proteins can be divided into four distinct functional modules: adaptation (spacer acquisition); expression (crRNA processing and target binding); interference (target cleavage); and ancillary (regulatory and other CRISPR-associated functions) (FIG. 1). In recent years, a wealth of structural and functional information has accumulated for the core Cas proteins (Cas1–Cas10) (see Supplementary information S1 (table)), which allows them to be classified into these modules.

The adaptation module is largely uniform across CRISPR–Cas systems and consists of the Cas1 and Cas2 proteins, with possible additional involvement of the restriction endonuclease superfamily enzyme Cas4 (REF. 91) and, in type II systems, Cas9 (REFS 63,64). Cas1, which adopts a unique α-helical fold, is an integrase that mediates the insertion of new spacers into CRISPR arrays by cleaving specific sites within the repeats[17,89,92]. The role of Cas2, which is a homologue of the mRNA interferase toxins of numerous toxin–antitoxin systems, is less well understood[3,72,93,94]. Cas2 has been shown to form a complex with Cas1 in the *Escherichia coli* type I CRISPR–Cas system and is required for adaptation. However, although Cas2 has RNase[95] and DNase activities[96], its catalytic residues are dispensable for adaptation[17], indicating that these activities are not directly involved in this process, at least in this species.

The expression and interference modules are represented by multisubunit CRISPR RNA (crRNA)–effector complexes[36,38,39,43–46,97,98] (BOX 2) or, in type II systems, by a single large protein, Cas9 (REFS 24,25,99). In the expression stage, pre-crRNA is bound to the multisubunit crRNA–effector complex, or to Cas9, and processed into a mature crRNA in a step catalysed by an RNA endonuclease[23] (typically Cas6; in type I and type III systems) or an alternative mechanism that involves RNase III and a transactivating CRISPR RNA (tracrRNA)[24] (in type II systems). However, in at least one type II CRISPR–Cas system, that of *Neisseria meningitidis*, crRNAs with mature 5′ ends are directly transcribed from internal promoters, and crRNA processing does not occur[69].

In the interference module, the crRNA–effector complex (in type I and type III systems) or Cas9 (in type II systems) combines nuclease activity with dedicated RNA-binding domains. Target binding relies on base pair formation with the spacer region of the crRNA. Cleavage of the target is catalysed by the HD family nuclease (Cas3″ or a domain in Cas3) in type I systems[52,100], by the combined action of the Cas7 and Cas10 proteins in type III systems[26,39,46,101–104] or by Cas9 in type II systems[25]. In type I systems, the HD nuclease domain is either fused to the superfamily 2 helicase Cas3′ (REFS 50–52) or is encoded by a separate gene, *cas3″*, whereas in type III systems a distinct HD nuclease domain is fused to Cas10 and is thought to cleave single-stranded DNA during interference[105]. In type II systems, the RuvC-like nuclease (RNase H fold) domain and the HNH (McrA-like) nuclease domain of Cas9 each cleave one of the strands of the target DNA[25,106]. Remarkably, the large (~950–1,400 amino acids) multidomain Cas9 protein is required for all three of the functional steps of CRISPR-based immunity (adaptation, expression and interference) in type II systems and thus concentrates much of the CRISPR–Cas system's function in a single protein.

The ancillary module is a combination of various proteins and domains that, with the exception of Cas4, are much less common than the core Cas proteins in CRISPR–Cas systems. Aside from its putative role in adaptation, Cas4 is thought to contribute to CRISPR–Cas-coupled programmed cell death[3,94]. Other notable components of the ancillary module include: a diverse set of proteins containing the CRISPR-associated Rossmann fold (CARF) domain[35,107], which have been hypothesized to regulate CRISPR–Cas activity[107] (in many type I and type III systems); and the inactivated P-loop ATPase Csn2, which forms a homotetrameric ring that accommodates linear double-stranded DNA in the central hole (in type II systems)[108–111]. Csn2 is not required for interference but apparently has a role in spacer integration, possibly preventing damage from the double-strand break in the chromosomal DNA[6,110]. Ancillary module genes are often found outside of CRISPR–*cas* loci, but the functions of these stand-alone genes have not been characterized in depth[72,94].

functionally active) CRISPR–Cas systems lack *cas1* and *cas2* and seem to instead depend on adaptation modules from other loci in the same genome. Within the set of complete loci, 111 composite loci that contained two or more adjacent CRISPR–Cas units (each consisting of at least a full complement of essential effector complex components) were identified and split into distinct units. Each locus or unit was classified by scoring type-specific and subtype-specific PSSMs that were constructed from multiple sequence alignments of the respective signature Cas proteins (see Supplementary information S2,S4 (tables)). For some of the more diverged signature proteins, multiple PSSMs were required for a single protein to capture the entire diversity of the cognate CRISPR–Cas subtype.

Of the single-unit complete loci, 1,574 (93%) were assigned to a specific subtype or the newly defined putative types IV and V, which are not split into subtypes, eight were identified up to the type only and one remained unclassified by our procedure (a subtype I-D system operon that is adjacent to the remnants of a subtype III-B system operon disrupted by recombination).

Our analysis suggests that the CRISPR–Cas systems can be divided on the basis of the genes encoding the effector modules; that is, whether the systems have several variants of a multisubunit complex (the CRISPR-associated complex for antiviral defence (Cascade) complex, the Csm complex or the Cmr complex) or Cas9. Thus, we introduce a new, broadest level of classification of CRISPR–Cas systems, which divides them into 'class 1' and 'class 2'. Class 1 systems possess multisubunit crRNA–effector complexes, whereas in class 2 systems all functions of the effector complex are carried out by a single protein, such as Cas9. We also find evidence for two putative new types, type IV and type V, which belong to class 1 and class 2, respectively. These observations result in a new classification system in which CRISPR–Cas systems are clustered into five types, each with a distinctive composition of expression, interference and adaptation modules (FIG. 1). These five types are divided into 16 subtypes, including five new subtypes (II-C, III-C and III-D, together with the single subtypes of type IV and type V systems), as detailed below.

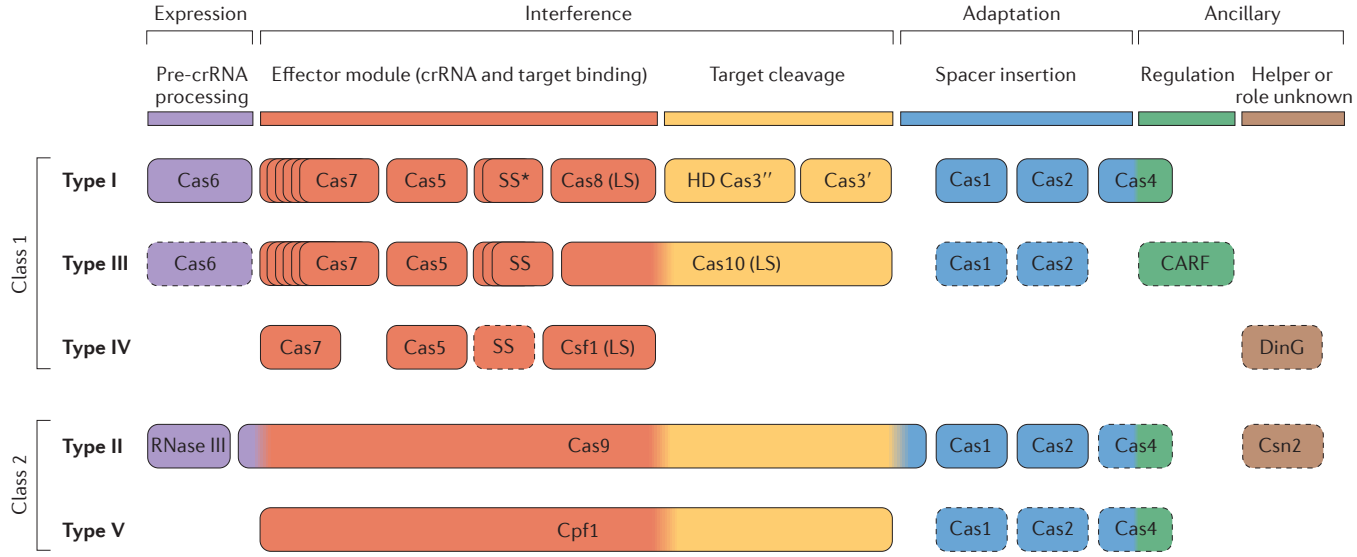


Figure 1 | **Functional classification of Cas proteins.** Protein names follow the current nomenclature and classification[13]. An asterisk indicates that the putative small subunit (SS) protein is instead fused to Cas8 (the type I system large subunit (LS)) in several type I subtypes[33]. The type III system LS and type IV system LS are Cas10 and Csf1 (a Cas8 family protein), respectively. Dispensable components are indicated by dashed outlines. Cas6 is shown with a solid outline for type I because it is dispensable in some but not most systems and by a dashed line for type III because most systems lack this gene and use the Cas6 provided *in trans* by other CRISPR–*cas* loci. The two colours for Cas4 and three colours for Cas9 reflect that these proteins contribute to different stages of the CRISPR–Cas response. The functions shown for type IV and type V system components are proposed based on homology to the cognate components of other systems, and have not yet been experimentally verified. The functional assignments for Cpf1 are tentatively inferred by analogy with Cas9 (only the RuvC (and TnpB)-like domains of the two proteins are homologous). CARF, CRISPR-associated Rossmann fold; pre-crRNA, pre-CRISPR RNA. This research was originally published in *Biochem. Soc. Trans.* Makarova K. S., Wolf Y. I., & Koonin E. V. The basic building blocks and evolution of CRISPR–Cas systems. *Biochem. Soc. Trans.* 2013; 41: 1392–1400 © The Biochemical Society.

Class 1 CRISPR–Cas systems

Class 1 CRISPR–Cas systems are defined by the presence of a multisubunit crRNA–effector complex. The class includes type I and type III CRISPR–Cas systems, as well as the putative new type IV.

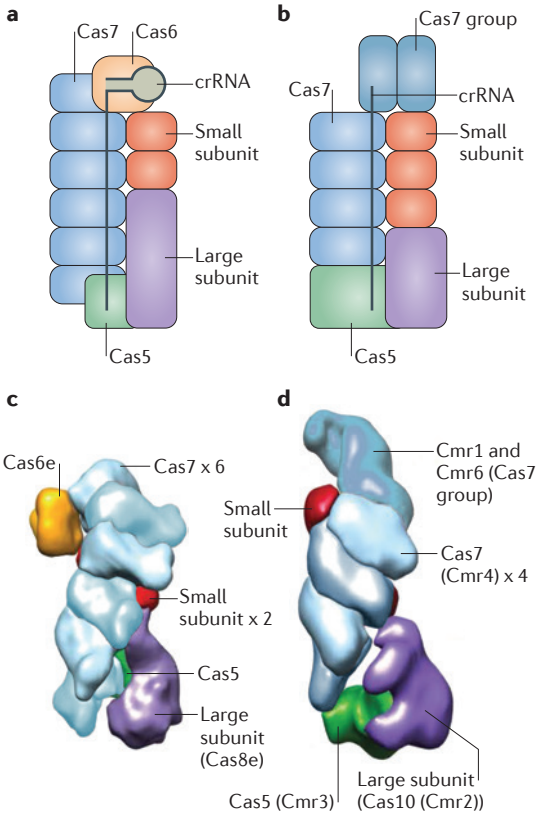
Type I CRISPR–Cas systems. All type I loci contain the signature gene *cas3* (or its variant *cas3'*), which encodes a single-stranded DNA (ssDNA)-stimulated superfamily 2 helicase with a demonstrated capacity to unwind double-stranded DNA (dsDNA) and RNA–DNA duplexes[50–52]. Often, the helicase domain is fused to a HD family endonuclease domain that is involved in the cleavage of the target DNA[50,53]. The HD domain is typically located at the amino terminus of Cas3 proteins (with the exception of subtype I-U and several subtype I-A systems, in which the HD domain is at the carboxyl terminus of Cas3) or is encoded by a separate gene (*cas3''*) that is usually adjacent to *cas3'* (FIG. 1).

Type I systems are currently divided into seven subtypes, I-A to I-F and I-U, all of which have been defined previously[13]. In the case of subtype I-U, U stands for uncharacterized because the mechanism of pre-crRNA cleavage and the architecture of the effector complex for this system remain unknown[33]. The type I-C, I-D, I-E and I-F CRISPR–Cas systems are typically encoded by a single (predicted) operon that encompasses the *cas1*, *cas2* and *cas3* genes together with the genes for the subunits of the Cascade complex (BOX 2). By contrast, many type I-A and I-B loci seem to have a different organization in which the *cas* genes are clustered in two or more (predicted) operons[35]. In most type I loci, each of the *cas* gene families is represented by a single gene.

Each type I subtype has a defined combination of signature genes and distinct features of operon organization (FIG. 2; see Supplementary information S4 (table)). Notably, *cas4* is absent in I-E and I-F systems, and *cas3* is fused to *cas2* in I-F systems. Subtypes I-E and I-F are monophyletic (that is, all systems of the respective subtype are descended from a single ancestor) in phylogenetic trees of Cas1 and Cas3, and each has one or more distinct signature genes (see Supplementary information S4,S5,S6 (table, box, box)).

Subtypes I-A, I-B and I-C seem to be descendants of the ancestral type I gene arrangement (*cas1–cas2–cas3–cas4–cas5–cas6–cas7–cas8*)[4,54]. This arrangement is preserved in subtype I-B, whereas subtypes I-A and I-C are diverged derivatives of I-B with differential gene loss and rearranged gene orders. A single signature gene for each of these subtypes could not be defined. The only protein that shows no significant sequence similarity between the subtypes is Cas8. However, the Cas8 sequence is highly diverged even within subtypes, so that consistent application of the signature gene approach would result in numerous new subtypes. For example, there are at least 10 distinct Cas8b families within subtype I-B

Box 2 | **Structural composition of multiprotein crRNA–effector complexes**



In type I and type III CRISPR–Cas systems, multiprotein CRISPR RNA (crRNA)–effector complexes mediate the processing and interference stages of the CRISPR defence system. In type I systems, this complex is known as the CRISPR-associated complex for antiviral defence (Cascade; see the figure, part **a**) complex, whereas in type III-A and type III-B systems the complexes are respectively known as Csm and Cmr (see the figure, part **b**) complexes. A common structural feature among the Cas proteins found in crRNA–effector complexes is the RNA recognition motif (RRM), a nucleic acid-binding domain that is the core fold of the extremely diverse RAMP protein superfamily[4,32,34]. The RAMPs Cas5 and Cas7 comprise the skeleton of the crRNA–effector complexes. In type I systems, Cas6 is typically the active endonuclease that is responsible for crRNA processing, and Cas5 and Cas7 are non-catalytic RNA-binding proteins; however, in type I-C systems, crRNA processing is catalysed by Cas5 (REF. 55). In type III systems, the enzyme that is responsible for processing has not been directly identified but is generally assumed to be Cas6 (REFS 38–40; however, Cas6 is not a subunit of the effector complex in these systems, and in some cases is provided *in trans* by other CRISPR–Cas loci), whereas Cas7 is involved in co-transcriptional RNA degradation during the interference stage[26].

In addition to Cas5, Cas6 and Cas7, crRNA–effector complexes typically contain two proteins that are designated, according to their size, the large subunit and the small subunit. The large subunit is present in all known type I and type III crRNA–effector complexes, whereas the small subunit is missing in some type I loci; a carboxy-terminal domain of the large subunit is predicted to functionally replace the small subunit in complexes where the small subunit is absent[33]. In type III systems, the large subunit is the putative cyclase-related enzyme encoded by *cas10*, whereas in type I systems the large subunit is encoded by diverse *cas8* genes that adopt a complex structure and show no readily detectable similarity to other proteins. Cas10 contains two cyclase-like Palm domains (a form of the RRM domain)[112,113], and the conservation of catalytic amino acid residues implies that one of these domains is active whereas the other is inactivated; the catalytic site of the active domain is required for cleavage of double-stranded DNA during interference[26], but its activity remains to be characterized in detail. Although it has been speculated that Cas8 is a highly derived homologue of Cas10 (REFS 4,33), and the similarity between the organizations of the types I and III crRNA–effector complexes is consistent with this possibility, sequence and structural comparisons fail to provide clear evidence. Some Cas8 proteins of subtype I-B have been shown to possess the single-stranded DNA-specific nuclease activity[114] required for interference[115]. However, whether such activity is a universal feature of the large subunit remains to be determined.

The small subunit proteins are encoded by *csm2* (subtypes III-A and III-D), *cmr5* (subtypes III-B and III-C), *cse2* (subtype I-E) or *csa5* (subtype I-A). They are α-helical proteins that have no detectable homologues, although a structural comparison suggests that the small subunit proteins of type I and III systems are homologous to one another[116].

Despite differences in structural details, the overall shapes and architectures of the Cascade[43,45,97], Cmr and Csm complexes[36,38,41,98,117] are remarkably similar, as can be seen from electron microscopy images of *Escherichia coli* Cascade complexes[31] (comprising Cas5, Cas6e and six Cas7 proteins, together with Cas8e as the large subunit and two Cse2 proteins as the small subunits; see the figure, part **c**) and *Thermus thermophilus* Cmr complexes[36] (comprising a Cas5 group protein known as Cmr3 and six Cas7 group proteins, namely Cmr1, Cmr6 and 4 copies of Cmr4, together with a Cas10 group protein known as Cmr2 as the large subunit and Cmr5 as the small subunit; see the figure, part **d**). This suggests that the ancestral multisubunit effector complex evolved before the divergence of type I and type III CRISPR–Cas systems.

Figure part **c** from REF. 31, Nature Publishing Group. Figure part **d** adapted with permission from REF. 36, Cell Press.

and at least 8 Cas8a families within subtype I-A (see Supplementary information S2 (table)). Thus, notwithstanding its complex evolution, we retain subtype I-B, which is best defined by the ancestral type I gene composition. The three main subdivisions within subtype I-B roughly correspond to the previously described subtypes Hmari, Tneap and Myxan[32] (see also TIGRFAM directory), and now could be defined through specific Cas8b families, Cas8b1 (for Hmari), Cas8b2 (for Tneap) and Cas8b3 (for Myxan), with a few exceptions.

A subset of subtype I-B systems defined by the presence of the *cas8b1* gene has been described as subtype I-G in the recent classification of archaeal CRISPR–Cas systems[35]. However, inclusion of bacterial CRISPR–Cas leads to increased diversity within subtype I-B so that if subtype I-G is recognized, consistency would require

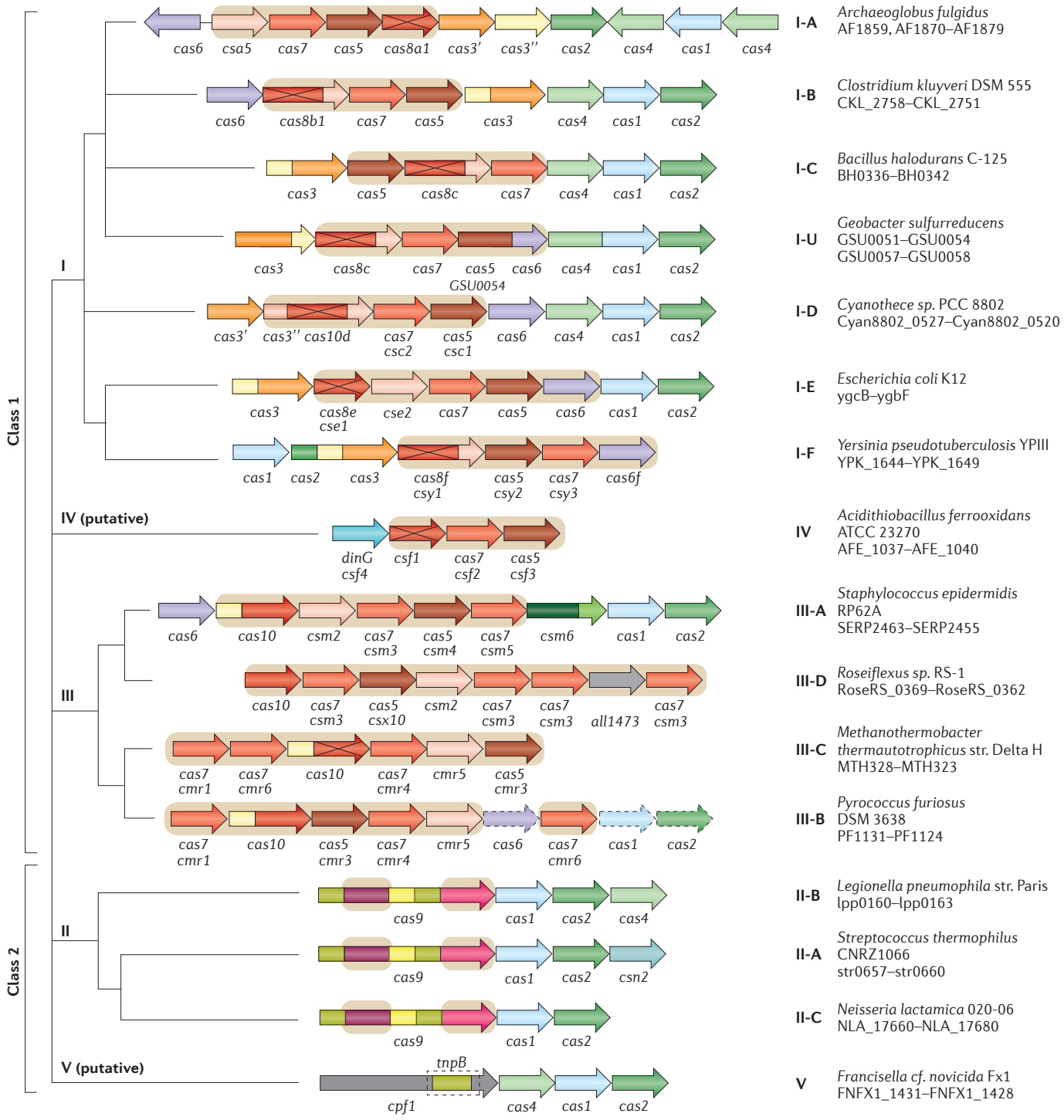


Figure 2 | **Architectures of the genomic loci for the subtypes of CRISPR–Cas systems.** Typical operon organization is shown for each CRISPR–Cas system subtype. For each representative genome, the respective gene locus tag names are indicated for each subunit. Homologous genes are colour-coded and identified by a family name. The gene names follow the classification from REF. 13. Where both a systematic name and a legacy name are commonly used, the legacy name is given under the systematic name. The small subunit is encoded by either *csm2*, *cmr5*, *cse2* or *csa5*; no all-encompassing name has been proposed to collectively describe this gene family to date. Crosses through genes encoding the large subunit (Cas8 or Cas10 family members) indicate inactivation of the respective catalytic sites. Genes and gene regions encoding components of the interference module (CRISPR RNA (crRNA)–effector complexes or Cas9 proteins) are highlighted with a beige background. The adaptation module (*cas1* and *cas2*) and *cas6* are dispensable in subtypes III-A and III-B; in particular, they are rarely present in subtype III-B (dashed lines). Dark green denotes the CARF domain. Gene regions coloured cream represent the HD nuclease domain; the HD domain in Cas10 is distinct from that of Cas3 and Cas3″. Also coloured are the regions of *cas9* that roughly correspond to the RuvC-like nuclease (lime green), HNH nuclease (yellow), recognition lobe (purple) and protospacer adjacent motif (PAM)-interacting domains (pink). The regions of *cpf1* aside from the RuvC-like domain are functionally uncharacterized and are shown in grey, as is the functionally uncharacterized *all1473* gene in subtype III-D.

splitting I-B into several subtypes. Therefore, at present, we classify these variants within subtype I-B.

Subtype I-C seems to be a derivative of subtype I-B that lacks Cas6, which seems to be functionally replaced by Cas5 (REF. 55). Subtype I-A is another derivative of subtype I-B and is typically characterized by the fission of *cas8* into two genes that encode degraded large and small subunits, respectively, as well as fission of *cas3* into *cas3′* and *cas3″*.

Subtype I-D also has several unique features, including Cas10d (instead of a Cas8 family protein) and a distinct variant of Cas3 (REF. 13) (FIG. 2; see Supplementary information S2,S4 (tables)). Subtype I-U is typified by the presence of an uncharacterized signature gene (*GSU0054*; TIGRFAM reference TIGR02165) and several other distinctive features that have been analysed in detail previously[33] (see Supplementary information S4 (table)). This group is monophyletic in the Cas3 tree and mostly monophyletic in the Cas1 tree (see Supplementary information S5,S6 (boxes)).

The phylogenetic tree of the type I signature protein Cas3′ (and the homologous region of Cas3) has been reported to accurately reflect the subtype classification[43], which is suggestive of a degree of evolutionary coherence between the phylogenies of the different genes in the operons of each subtype. However, re-analysis of the Cas3 phylogeny using a larger, more diverse sequence set (see Supplementary information S6 (box)) reveals a complex picture in which subtypes I-A, I-B and I-C are polyphyletic (that is, not descended from a common ancestor). Conceivably, this discrepancy results from a combination of accelerated evolution of many Cas3 variants and horizontal gene transfer.

In addition to the complete type I CRISPR–*cas* loci, analysis of sequenced genomes has revealed a variety of putative type I-related operons that encode effector complexes but are not associated with *cas1, cas2* or *cas3* genes and are only in some cases adjacent to CRISPR arrays (see Supplementary information S4 (table)). These solo effector complexes are often encoded on plasmids and/or associated with transposon-related genes. Many of these operons are derivatives of subtype I-F, whereas others are derivatives of subtype I-B (see Supplementary information S4,S7 (tables)). Some of the genomes that have these incomplete type I systems encode Cas1–Cas2 as parts of other CRISPR–*cas* loci but others lack these genes altogether (see Supplementary information S7 (table)). The functionality of solo effector complexes has not been investigated.

Type III CRISPR–Cas systems. All type III systems possess the signature gene *cas10*, which encodes a multidomain protein containing a Palm domain (a variant of the RNA recognition motif (RRM)) that is homologous to the core domain of numerous nucleic acid polymerases and cyclases and that is the largest subunit of type III crRNA–effector complexes (BOX 2). Cas10 proteins show extensive sequence variation among the diverse type III CRISPR–Cas systems, which means that several PSSMs are required to identify these loci. All type III loci also encode the small subunit protein (see below), one Cas5 protein and typically several paralogous Cas7 proteins (FIG. 1). Often, Cas10 is fused to an HD family nuclease domain that is distinct from the HD domains of type I CRISPR–Cas systems and, unlike the latter, contains a circular permutation of the conserved motifs of the domain[34,56].

Type III systems have been previously classified into two subtypes, III-A (previously known as Mtube subtype or Csm module) and III-B (previously known as Cmr module or RAMP module), that can be distinguished by the presence of distinct genes encoding small subunits, *csm2* (in the case of subtype III-A) and *cmr5* (in the case of subtype III-B) (FIG. 2; see Supplementary information S4 (table)). Subtype III-A loci usually contain *cas1*, *cas2* and *cas6* genes, whereas most of the III-B loci lack these genes and therefore depend on other CRISPR–Cas systems present in the same genome[4], providing strong evidence for the modularity of CRISPR–Cas systems[35] (FIG. 2). Both subtype III-A and subtype III-B CRISPR–Cas systems have been shown to co-transcriptionally target RNA[26,27,37–39,57] and DNA[26,58–61].

The composition and organization of type III CRISPR–*cas* loci are more diverse than those of type I systems — although there are fewer type III subtypes, each of these is more polymorphic than type I subtypes. This diversity is due to gene duplications and deletions, domain insertions and fusions, and the presence of additional, poorly characterized domains that could be involved either in crRNA–effector complex functions or in associated immunity. At least two type III variants (one from subtype III-A and one from subtype III-B) are common and are here upgraded to subtypes III-D and III-C, respectively, as proposed earlier for archaea[35] (FIG. 3; see Supplementary information S8 (table)). The distinctive feature of subtype III-C (previously known as MTH326-like[33]) is the apparent inactivation of the cyclase-like domain of Cas10 accompanied by extreme divergence of the sequence of this protein. Subtype III-D loci typically encode a Cas10 protein that lacks the HD domain. They also contain a distinct *cas5*-like gene known as *csx10* and often an uncharacterized gene that is homologous to *all1473* from *Nostoc sp.* PCC 7120 (REF. 33). Both of these new subtypes lack *cas1* and *cas2* genes (FIG. 2) and accordingly are predicted to recruit adaptation modules *in trans*. The phylogeny of Cas10, the signature gene of type III CRISPR–Cas, is consistent with the subtype classification, with each subtype representing a distinct clade (see Supplementary information S9 (box)).

Putative type IV CRISPR–Cas systems. Several bacterial genomes contain putative, functionally uncharacterized type IV systems, often on plasmids, as can be typified by the AFE_1037-AFE_1040 operon in *Acidithiobacillus ferrooxidans* ATCC 23270. Similar to most subtype III-B loci, this system lacks *cas1* and *cas2* genes and is often not in proximity to a CRISPR array or, in many cases, is encoded in a genome that has no detectable CRISPR arrays (it might be more appropriate to denote the respective loci Cas systems rather than CRISPR–Cas). Type IV systems encode a predicted minimal

multisubunit crRNA–effector complex that consists of a partially degraded large subunit, Csf1, Cas5 and — as a single copy — Cas7, and in some cases, a putative small subunit[33] (FIG. 1); *csf1* can serve as a signature gene for this system. The minimalist architecture of type IV loci is distinct from those of all type I and type III subtypes (FIG. 2; see Supplementary information S4 (table)), which together with the unique large subunit (Csf1) justifies their status as a new type.

There are two distinct variants of type IV CRISPR–Cas systems, one of which contains a DinG family helicase (REF. 62), and a second one that lacks DinG but typically contains a gene encoding a small α-helical protein, which is a putative small subunit [33]. Type IV systems could be mobile modules that, similar to subtype III-B systems, use crRNAs from different CRISPR arrays once these become available. This possibility is consistent with the occasional localization of type IV loci adjacent to CRISPR arrays, *cas6* genes and (less often) adaptation genes[35].

Class 2 CRISPR–Cas systems

Class 2 CRISPR–Cas systems are defined by the presence of a single subunit crRNA–effector module. This class includes type II CRISPR–Cas systems, as well as a putative new classification, type V.

Type II CRISPR–Cas systems. Type II CRISPR–Cas systems dramatically differ from types I and III, and are by far the simplest in terms of the number of genes. The signature gene for type II is *cas9*, which encodes a multidomain protein that combines the functions of the crRNA–effector complex with target DNA cleavage[25], and also contributes to adaptation[63,64]. In addition to *cas9*, all identified type II CRISPR–*cas* loci contain *cas1* and *cas2* (see REF. 65 for a detailed comparative analysis of type II systems) (FIG. 1) and most type II loci also encode a tracrRNA, which is partially complementary to the repeats within the respective CRISPR array[65–67].

The core of Cas9, which includes both nuclease domains and a characteristic Arg-rich cluster, most likely evolved from genes of transposable elements that are not associated with CRISPR[65]. Thus, owing to the significant sequence similarity between Cas9 and its homologues that are unrelated to CRISPR–Cas, Cas9 cannot be used as the only signature for identification of type II systems. Nevertheless, the presence of *cas9* in the vicinity of *cas1* and *cas2* genes is a hallmark of type II loci.

Type II CRISPR–Cas systems are currently classified into three subtypes, which were introduced in the previous classification (II-A and II-B)[13] or subsequently proposed on the basis of a distinct locus organization (II-C)[65,66,68] (FIG. 2; see Supplementary information S4 (table)). Subtype II-A systems include an additional gene, *csn2* (FIG. 2), which is considered a signature gene for this subtype. The long and short variants of Csn2 form compact clusters when superimposed over the Cas9 phylogeny and seem to correspond to two distinct variants of subtype II-A[65]. However, as with subtype I-B, we chose to keep these two variants within subtype II-A. It was recently shown that all four subtype II-A Cas proteins are involved in spacer acquisition[63].

Subtype II-B lacks *csn2* but includes *cas4*, which is otherwise typical of type I systems (FIG. 2). Moreover, subtype II-B *cas1* and *cas2* are more closely related to type I homologues than to subtype II-A, which is suggestive of a recombinant origin of subtype II-B[65]. Subtype II-C loci only have three protein-coding genes (*cas1*, *cas2* and *cas9*) and are the most common type II CRISPR–Cas system in bacteria[3,65,66]. A notable example of a subtype II-C system is the crRNA-processing-independent system found in *Neisseria meningitidis*[69] (BOX 1).

In the Cas9 phylogeny, subtypes II-A and II-B are monophyletic whereas subtype II-C is paraphyletic with respect to II-A (that is, subtype II-A originates from within II-C)[65]. Nevertheless, II-C was retained as a single subtype given the minimalist architecture of the effector modules shared by all II-C loci.

Putative type V CRISPR–Cas systems. A gene denoted *cpf1* (TIGRFAM reference TIGR04330) is present in several bacterial genomes and one archaeal genome, adjacent to *cas1*, *cas2* and a CRISPR array (for example, in the FNFX1_1431–FNFX1_1428 locus of *Francisella cf. novicida* Fx1)[70] (FIG. 2). These observations led us to putatively define a fifth type of CRISPR–Cas system, type V, which combines Cpf1 (the interference module) with an adaptor module (FIG. 1; see Supplementary information S4 (table)). Cpf1 is a large protein (about 1,300 amino acids) that contains a RuvC-like nuclease domain homologous to the respective domain of Cas9 and the TnpB protein of IS605 family transposons, along with putative counterparts to the characteristic Arg-rich region of Cas9 and the Zn finger of TnpB. However, Cpf1 lacks the HNH nuclease domain that is present in all Cas9 proteins[54,65]. Given the presence of a predicted single-subunit crRNA–effector complex, the putative type V systems are assigned to class 2 CRISPR–Cas. Some of the putative type V loci also encode Cas4 and accordingly resemble subtype II-B loci, whereas others lack Cas4 and are more similar in architecture to subtype II-C. Unlike Cas9, Cpf1 is encoded outside the CRISPR–Cas context in several genomes, and its high similarity with TnpB suggests that *cpf1* is a recent recruitment from transposable elements.

If future experiments were to show that these loci encode bona fide CRISPR–Cas systems and that Cpf1 is a functional analogue of Cas9, then these systems would arguably qualify as a novel type of CRISPR–Cas. Despite the overall similarity to type II CRISPR–Cas systems, the putative type V loci clearly differ from the established type II subtypes more than type II subtypes differ from each other, most notably in the distinct domain architectures of Cpf1 and Cas9. Furthermore, whereas type II systems are specific to bacteria, a putative type V system is present in at least one archaeon, *Candidatus* Methanomethylophilus alvus[35].

Rare, unclassifiable CRISPR–Cas systems

The classification of CRISPR–Cas systems outlined above covers nearly all of the CRISPR–*cas* loci identified in the currently sequenced archaeal and bacterial genomes

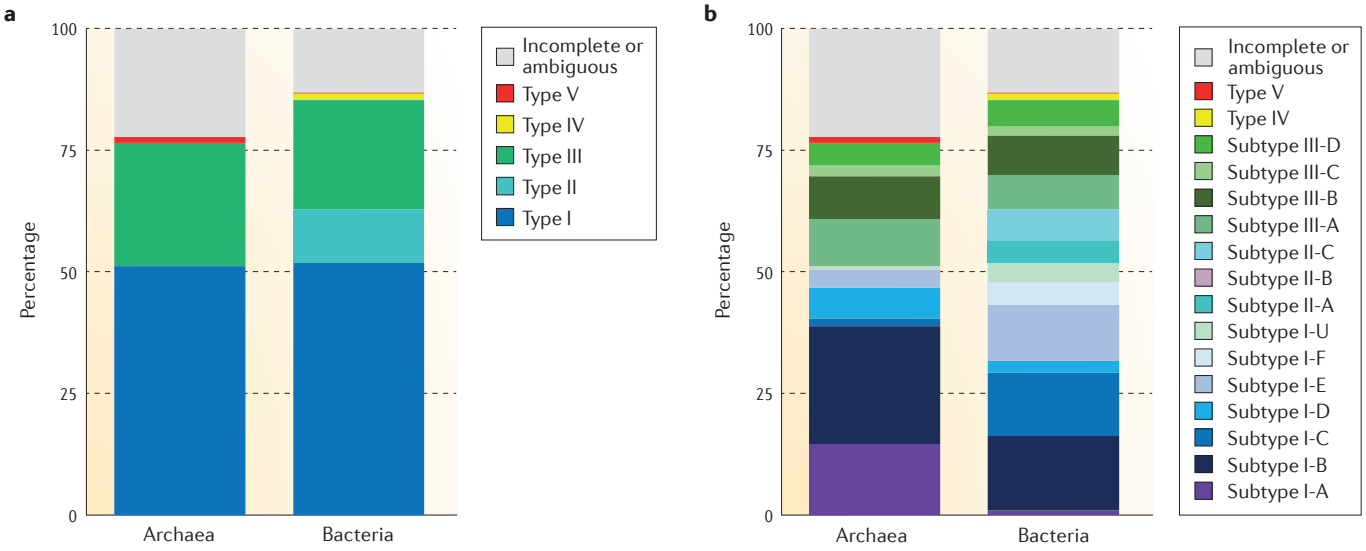


Figure 3 | **Distribution of CRISPR–Cas systems in sequenced archaeal and bacterial genomes.** **a** | Distribution by types. Chart showing the proportions of identified CRISPR–*cas* loci in bacterial or archaeal genomes that encode type I, type II, type III, type IV or type V CRISPR–Cas systems. The proportion of loci that encode incomplete systems or that we could not classify unambiguously is also shown. **b** | Distribution by subtypes. Chart showing the proportions of identified CRISPR–*cas* loci in bacterial or archaeal genomes that encode each of the subtypes of CRISPR–Cas systems included in the new classification described in this article. Note that type IV and V loci each encompass a single subtype. The proportion of loci that encode incomplete systems or that we could not classify unambiguously is also shown.

(FIG. 3). Nonetheless, owing to the rapid evolution of CRISPR–*cas* loci, which involves extensive recombination, it was not possible to account for all variants.

As a case in point, a putative CRISPR–Cas system was recently identified in *Thermococcus onnurineus*[71]. Based on some marginal similarities to protein components of crRNA–effector complexes, this locus was previously described as a Csf module[35], which here is classified as type IV. However, only the putative Cas7 protein from this locus (TON_0323) is most similar to the variant characteristic of type IV systems (Csf2), whereas Cas2 and Cas4 are uncharacteristic of type IV loci, and an uncharacterized large protein containing an HD domain is present instead of Csf1. These features suggest classification of the *T. onnurineus* as a derived type I system (notwithstanding the absence of the signature gene *cas3* or its variant *cas3′*), although it could not be assigned to any known subtype.

Several unusual variants of type III systems also posed a challenge for our classification. For example, the 15-gene locus in *Ignisphaera aggregans* has previously been classified as subtype III-D[35]. However, the III-D signature gene *csx10*, which encodes Cas5, is missing, and the other Cas proteins encoded by this locus show limited similarities to different type III subtypes[71]. Therefore, the *I. aggregans* locus seems to encode a type III system but cannot be unequivocally assigned to any subtype. Another distinct type III variant has been identified in several Crenarchaeota, primarily from the order Sulfolobales[35]. These loci lack detectable small subunits encoded by *csm2* or *cmr5* but contain a unique *cas* gene provisionally denoted *csx26*. Another variant is typified by the CRISPR–*cas* locus from *Thermotoga lettingae*[35], which is the only known type III system to encode a single Cas7 protein, a feature of type IV systems. These two type III variants share more similarity with subtype III-A than with other subtypes and are currently assigned to this subtype (see Supplementary information S9 (box)); however, subsequent analysis of new genomes along with experimental study might prompt their reclassification into separate subtypes.

This accumulation of unclassifiable variants suggests that the current approaches to CRISPR–Cas system classification will need to be further refined to cope with the challenge of ever increasing diversity.

Distribution in archaea and bacteria

Approximately 47% of analysed bacterial and archaeal genomes encode CRISPR–*cas* loci. As reported previously[13,72], CRISPR–Cas systems are much more prevalent in archaea (87% of genomes) than they are in bacteria (50% of genomes). For those genomes encoding CRISPR–*cas* loci, the rate of incomplete loci is similar for archaeal and bacterial genomes (17% and 12%, respectively). Complete single-unit loci are most commonly type I systems in both archaeal and bacterial genomes (64% and 60% of the loci, respectively), whereas putative type IV and type V systems are rare (<2% overall). Archaea possess significantly more type III systems than bacteria (34% versus 25% of the complete single-unit CRISPR–*cas* loci) but lack type II systems (13% in bacteria) (FIG. 3a). Thus, class 2 CRISPR–Cas systems are represented in archaea only by a single instance of the putative type V.

Overall, the most abundant CRISPR–Cas system is subtype I-B (20% of complete single-unit loci), followed by subtypes I-C and I-E (13% and 12%, respectively). In archaea, subtype I-A is the second most abundant after subtype I-B (18% and 30%, respectively), followed

by subtypes III-A and III-B; subtype I-F is missing[35] (FIG. 3b). Among the three type II subtypes, subtypes II-C and II-A are the most abundant, comprising 7% and 5% of bacterial single-unit *cas* loci, respectively; subtype II-B is a minority, with only six loci that are restricted to Proteobacteria (0.3%). Finally, archaea encompass a significantly greater fraction of multi-unit loci than bacteria (14% versus 6%). Of the 13% of all CRISPR–*cas* loci that are incomplete or unclassified, 48% are partial type I loci and 25% are partial type III loci.

Different archaeal and bacterial phyla show distinct trends in the distribution of CRISPR–Cas systems (see Supplementary information S8 (table)). Notably, the Crenarchaeota lack subtypes I-B and I-C systems, which are abundant in other archaea and bacteria, whereas the Euryarchaeota are enriched in subtype I-B loci[35]. The Actinobacteria show a strong preference for subtype I-E systems, and the Cyanobacteria for subtype III-B systems, whereas the Firmicutes account for most of the subtype II-A systems. Finally, the Proteobacteria lack subtype I-A systems but are strongly enriched in subtype I-F loci. Considering the extraordinary importance of type II CRISPR–Cas systems in biotechnology, it is worth emphasizing that these systems represent a minority of CRISPR–*cas* loci. They also seem to be specific to bacteria and are significantly over-represented in the Proteobacteria and the Firmicutes.

We expect that the bias of available sequence data towards cultivable microorganisms, especially those of medical or biotechnological importance, affects the currently observed distribution of CRISPR–Cas systems. Nevertheless, the remarkable stability of the overall fraction of CRISPR-possessing microorganisms over several years of observation seems to imply that at least the main trends are captured by the present analysis.

Modular organization and evolution

Similarly to other defence systems, CRISPR–*cas* loci evolve under strong selection pressure exerted by changing pathogens, resulting in rapid evolution that is largely uncoupled from the evolution of the rest of the respective genomes. Here we examine the evolutionary relationships between different components of the CRISPR–Cas systems and put forward the concept of modular organization, with semi-independent evolution of each module.

cas loci and CRISPR arrays. For the purpose of comparative analysis of CRISPR–Cas systems, CRISPR arrays were predicted in all genomes using CRISPRfinder[73,74] following the procedure described in CRISPRmap[75] and CRISPRstrand[76]. For each of the 1,949 *cas* loci, the nearest CRISPR array was identified, which showed a natural cut-off of 530 base pairs for the distance between *cas* loci and proximal CRISPR arrays (Supplementary information S8 (table)). Using this cut-off, 1,484 *cas* loci (75%) were classified as adjacent to a CRISPR array, 383 loci (22%) were present in CRISPR-positive genomes but far from any array, and 82 loci (54 complete and 28 incomplete, 3% total) were present in CRISPR-negative genomes. Although, as expected, the fraction of *cas* loci in CRISPR-negative genomes was significantly higher for incomplete (6.5%) than complete (2.3%) *cas* loci (χ^2 test P value of 7×10^{-5}), the existence of complete *cas* loci that were not accompanied by a recognizable CRISPR array anywhere in the genome was notable, as it defies the principle that crRNA–effector complexes are universally associated with CRISPR immunity. These CRISPR-less loci could be remnants of recently inactivated CRISPR–Cas systems or might function in a different way to the characterized CRISPR–Cas systems.

Conversely, of the 4,210 detected CRISPR arrays, 1,382 (33%) are adjacent (within 530 base pairs) to a *cas* locus, 2,365 arrays (56%) are located outside of *cas* loci in *cas*-positive genomes, and the remaining 463 arrays (11%) are orphans, present in genomes without detected *cas* loci. The orphan CRISPR arrays are probably remnants of formerly functional CRISPR–Cas systems.

CRISPR arrays are themselves classified into 18 structural families and 24 sequence families (only 23 were used here because one family could not be associated with any *cas* loci in our dataset), including unclassified repeats[75–77]. Both structural and sequence families of CRISPR show significant preferential association with particular types and subtypes of *cas* loci, although in most cases associations with other types or subtypes can also occur (FIG. 4; see Supplementary information S3,S10 (boxes)).

CRISPR–Cas systems and the species tree. Defence systems of bacteria and archaea evolve under extreme selection pressure from pathogens, particularly viruses, often using non-classic evolutionary processes, such as the seemingly Lamarckian adaptations represented by spacer integrations in CRISPR arrays[78], the partially selfish mode of reproduction in which toxin–antitoxin systems are maintained in the genome through their addictive properties[79], and pervasive horizontal gene transfer[72,80]. In line with these trends, evidence of extensive horizontal transfer of CRISPR–*cas* loci has been reported[8,13,34,81–83].

To quantify the propensity of CRISPR–Cas systems to evolve via horizontal — as opposed to vertical — transmission, we compared various system features with a provisional species tree of bacteria and archaea that was reconstructed from concatenated ribosomal protein alignments[84]. As expected, the classification of the *cas* loci showed only weak consistency with the species tree (FIG. 4). The association between the species tree and CRISPR repeat types was also weak for both structure-based and sequence-based repeat classification (FIG. 4; see Supplementary information S11 (table)). These observations quantitatively show that horizontal transfer dominates the evolution of CRISPR–*cas* loci.

Cas1 phylogeny, CRISPR–Cas classification and architecture of* cas *loci. We examined the key evolutionary trends of the CRISPR–Cas systems in connection with the classification outlined above. Cas1 is the most conserved Cas protein, in terms of both representation in CRISPR–*cas* loci and amino acid sequence conservation[85], and the Cas1 phylogeny generally correlates with the organization of CRISPR–*cas* loci[13]. Thus, until recently, Cas1 has been considered to be the signature of

the presence of CRISPR–Cas systems in a genome[13,32,34]. However, in this analysis we identified 86 genomes containing complete (and by inference, functional) effector modules but that lacked *cas1*. These include genomes encoding the putative type IV systems, most subtype III-B, III-C and III-D systems and rare variants of subtypes I-C and I-F; 14 of these genomes also lack readily identifiable CRISPR arrays (FIG. 2; see Supplementary information S7 (table)).

Conversely, in some archaea and bacteria *cas1* genes are located outside CRISPR–*cas* loci[4], often within predicted self-synthesizing transposable elements dubbed casposons[86]. Casposon-encoded Cas1 proteins probably function as integrases that mediate the mobility of these transposons. The discovery of casposons suggests that the CRISPR–Cas adaptive immunity system arose from the insertion of a casposon near an innate immunity locus that encoded an effector complex[87].

Of the 1,949 CRISPR–*cas* loci analysed, 1,404 encompass at least one *cas1* gene. We constructed a phylogenetic tree of all 1,418 Cas1 sequences (some composite loci contain at least two *cas1* genes) and rooted the tree using the modified midpoint procedure (FIG. 5; see Supplementary information S5 (box)). Mapping CRISPR–*cas* loci onto the Cas1 tree (FIG. 5) demonstrates a considerable agreement between the phylogeny of Cas1 and locus types and subtypes, consistent with previous observations. Thus, *cas1* genes of subtypes I-E, I-F, II-B and putative type V are strictly monophyletic, and *cas1* genes of subtypes I-C, I-U and II-A are largely monophyletic, with a few exceptions. In addition, *cas1* genes of subtypes II-A and II-C form a mostly homogeneous clade, in agreement with a previous analysis[65]. By contrast, *cas1* genes from the other type I subtypes and type III loci are scattered across the tree, suggestive of primarily horizontal evolution[13,34,35,88]. Thus, although substantial recombination occurs between the adaptation module and the other modules of the *cas* loci, the combination of the adaptation module with other modules is far from random.

As expected, the phylogeny of Cas1 is a poor match to the species tree of archaea and bacteria. The correlation of the distances between species with those between the corresponding *cas1* genes in the tree is much weaker than the correlation between the Cas1 phylogeny and CRISPR–*cas* locus classification (FIG. 4; see Supplementary information S11 (table)). These observations imply an extensive history of horizontal transfers, many of which involved complete CRISPR–*cas* loci, whereas a smaller number included the adaptation module alone.

Cas1 is crucial to the adaptation stage of the CRISPR-mediated immune response[17,89] and thus could be expected to co-evolve with CRISPR arrays[83,88]. We mapped structure-based and sequence-based repeat classification of CRISPR arrays adjacent to *cas* loci to the Cas1 tree. When only fully classified CRISPR repeats are considered, a high degree of consistency is observed between the Cas1 tree topology and repeat classification (FIG. 4; see Supplementary information S11 (table)), which probably reflects the direct recognition of repeats by Cas1 and its mechanistic involvement in the formation of the CRISPR arrays[89].

We also developed a quantitative measure to compare the architectures of the *cas* loci to one another and to generate a similarity dendrogram (see Supplementary information S12 (box)). Overall, the topology of the dendrogram is consistent with the subtype classification of CRISPR–Cas systems (FIG. 4; see Supplementary information S11 (table)). However, the clusters obtained by this method are much narrower than the respective subtypes, which is consistent with a frequent rearrangement of CRISPR–Cas loci. By contrast, clusters obtained from protein similarity searches, using proteins from the interference module, are broader and often directly correspond to individual subtypes (see Supplementary information S12,S13 (boxes)). As expected, the clustering of CRISPR–Cas systems by locus architecture is substantially more compatible with the Cas1 phylogeny than with the species tree (FIG. 4), in agreement with the considerable evolutionary coherence of the CRISPR–Cas systems despite frequent horizontal gene transfer of CRISPR–*cas* loci and of individual modules.

Figure 4 | **Comparison of different classifications of CRISPR–Cas systems.** This graph shows the strength of correlation between the new classification of CRISPR–Cas systems described here ('subtypes'; in the centre of the graph) and other classification measures. 'Interference genes tree' represents a phylogeny of interference module genes, which encode multisubunit CRISPR RNA (crRNA)–effector complexes or Cas9 proteins. This tree was created using a simple clustering approach based on aggregate protein sequence similarity. 'Adaptation genes tree' represents clustering produced by the same method but based on both components of the adaptation module, Cas1 and Cas2. 'Cas1 phylogeny' is the phylogenetic tree of Cas1 proteins shown in FIG. 5. 'Loci architecture tree' represents clustering based on a quantitative measure we developed to compare the architectures of CRISPR–*cas* loci. The measure is based on a weighted similarity index of the order of *cas* genes. 'Repeats (sequence)' denotes the classification of CRISPR sequences into 24 families on the basis of sequence similarity. 'Repeats (structure)' denotes the classification of CRISPR sequences into 18 families on the basis of structural similarity. The species tree represents the phylogeny of bacterial and archaeal translation systems. The distances depicted are inversely proportional to the degree of similarity. The full similarity matrix is shown in Supplementary information S11 (table).

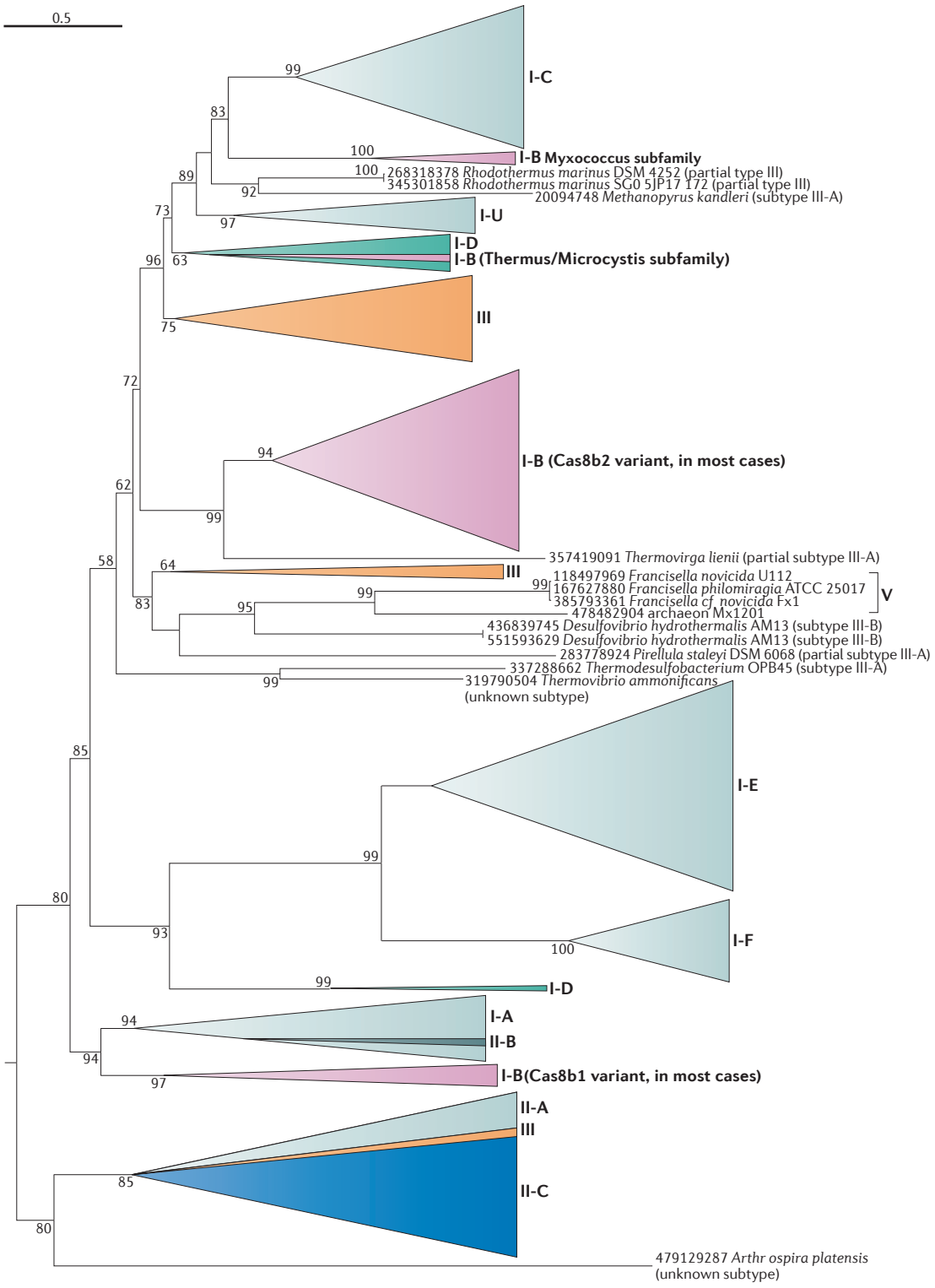


Figure 5 | **Mapping of the CRISPR–Cas classification onto the phylogenetic tree of Cas1.** Subtypes from the new classification of CRISPR–Cas systems described here were mapped onto a sequence-based phylogenetic reconstruction of 1,418 proteins from the Cas1 family, which is the most conserved Cas protein family. The phylogeny shows a close agreement with the subtype classification, as subtypes I-A, I-C, I-E, I-F, I-U, II-A, II-B, and putative type V are mostly or strictly monophyletic and are shown in gradients of light grey, except for II-B, which is shown in dark grey to indicate its origin from within I-A. The more discordant distribution of Cas1 for other subtypes probably results from horizontal transfer. None of the type III subtypes is monophyletic (in contrast to the Cas10 tree shown in Supplementary information S9 (box)), and so type III subtypes are not indicated. Note that Cas1 is absent in type IV loci and so these putative CRISPR–Cas systems are not shown. Triangles denote multiple collapsed branches. Individual genes are labelled with species names and gene identification numbers. Bootstrap values are indicated as percentage points; values below 50% are not shown.

Automated annotation of CRISPR–*cas* loci

Given the rapid pace of microbial genome sequencing, tools for the automated annotation of CRISPR–*cas* locus subtypes in newly sequenced genomes would be highly valuable. Although a careful inspection of combined features is required for accurate subtype annotation, we investigated whether an automated annotation method based on the similarity of the protein sequences of interference modules can faithfully reproduce the existing locus annotation.

To assess the value of the interference module as a proxy for the distribution of CRISPR–*cas* loci in our classification, we adopted a simple clustering approach based on aggregate sequence protein similarity[35]. This approach was chosen because of the lack of a universal marker suitable for phylogenetic analysis, as there is great variability in gene composition and module architecture between subtypes. The resulting cluster dendrogram (see Supplementary information S13 (box)) showed a high correlation with the subtype classification (FIG. 4; see Supplementary information S10 (box)). A similar cluster dendrogram constructed for Cas1 and Cas2 (see Supplementary information S14 (text)) showed a strong correlation with the Cas1 phylogeny but a considerably weaker correlation with the classification and architecture of CRISPR–*cas* loci than observed for the crRNA–effector complex dendrogram (FIG. 4; see Supplementary information S11 (table)). This difference supports our rationale in classifying CRISPR–*cas* loci on the basis of the interference module rather than Cas1 and demonstrates the ability of interference module protein clustering to closely reflect the new classification.

Having established the strong agreement between the clustering of interference module proteins and our classification, we constructed an automated classifier using prior information on the association between sequence PSSMs and CRISPR–*cas* loci and the corresponding classification of the effector modules. The classifier achieved 0.998 accuracy, which means that only 4 of 1,942 subtypes were incorrectly assigned (see Supplementary information S4,S15 (table, figure)). However, the accuracy of the method depends on the level of sequence similarity of the analysed Cas proteins to those available in the modelling phase, and predictably drops when the variants are only distantly related to the existing subtypes. Thus, the automated classifier described here has only limited applicability when annotating divergent variants of CRISPR–Cas subtypes.

Conclusions

The principal conclusion from the comparative analysis of the CRISPR–*cas* loci described here is the dynamic character and pronounced modularity of the evolution of this adaptive immunity system, which is conceivably driven by a perpetual arms race between the host genome and invading plasmids and viruses (dynamic evolution is a general theme in the evolution of defence systems[72,80]). In particular, the Cas1–Cas2 adaptation module evolved, to a large extent, independently of the operational modules (in particular, crRNA–effector complexes) of CRISPR–Cas systems, in agreement with the probable origin of the system as the result of the integration of a casposon-like mobile element next to an operon encoding a stand-alone effector complex[87]. The dynamic, modular evolution of CRISPR–Cas is also manifested at the level of the architecture of *cas* loci and the combination of different families of CRISPR arrays with different *cas* loci. However, a complementary trend is the frequent horizontal transfer of complete CRISPR–*cas* loci, which confers a degree of coherence to these systems and ensures that there is almost no congruence between the evolution of CRISPR–Cas and the species phylogeny as represented by the translation system[90].

The dynamic and modular character of CRISPR–Cas evolution hampers a straightforward classification based on evolutionary relationships. However, the classification approach we propose here, which combines signature genes with elements of the architecture of *cas* loci, assigned nearly all of the detected CRISPR–*cas* loci to specific subtypes. Furthermore, the resulting classification is largely compatible with the results of sequence-based clustering of crRNA–effector complexes, which can be adopted for automated classification of CRISPR–Cas systems from new genomes. The refinement of automated classification using more sophisticated machine learning and other computational techniques could lead to the development of fully automated classification of CRISPR–Cas systems.

In many respects, the new classification closely resembles the 2011 version[13], suggesting that the most common variants of CRISPR–Cas systems have already been discovered. However, we introduced a new top level, class, to account for the key differences between multisubunit and single-subunit crRNA–effector modules, as well as two new putative types (type IV and type V) and five new subtypes (II-C, III-C and III-D, together with the single subtypes of type IV and type V systems). Furthermore, the existence of currently unclassifiable variants implies that rare types and subtypes remain to be discovered and characterized, and the number of these is expected to substantially increase with the sequencing of new bacterial and archaeal genomes and metagenomes. In particular, the similarity between Cpf1 of the putative type V system and TnpB, which is usually found in transposons, suggests that multiple variants of single-subunit effector modules, and thus class 2 systems, might have evolved on independent occasions.

The classification of CRISPR–Cas systems and the principles of CRISPR–Cas evolution outlined here are expected to help the identification and focused discovery of new variants, some of which could become novel tools for genome engineering.

1. Deveau, H., Garneau, J. E. & Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **64**, 475–493 (2010).
2. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).
3. Koonin, E. V. & Makarova, K. S. CRISPR–Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **10**, 679–686 (2013).
4. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. The basic building blocks and evolution of CRISPR–Cas systems. *Biochem. Soc. Trans.* **41**, 1392–1400 (2013).
5. Barrangou, R. & Marraffini, L. A. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54**, 234–244 (2014).
6. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

7. Barrangou, R. CRISPR–Cas systems and RNA-guided interference. *Wiley Interdiscip. Rev. RNA* **4**, 267–278 (2013).
8. Westra, E. R. *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46**, 311–339 (2012).
9. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338 (2012).
10. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
11. Magadán, A. H., Dupuis, M. E., Villion, M. & Moineau, S. Cleavage of phage DNA by the *Streptococcus thermophilus* CRISPR3–Cas system. *PLoS ONE* **7**, e40913 (2012).
12. van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34**, 401–407 (2009).
13. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).
14. Westra, E. R., Buckling, A. & Fineran, P. C. CRISPR–Cas systems: beyond adaptive immunity. *Nat. Rev. Microbiol.* **12**, 317–326 (2014).
15. Sampson, T. R. & Weiss, D. S. CRISPR–Cas systems: new players in gene regulation and bacterial physiology. *Front. Cell. Infect. Microbiol.* **4**, 37 (2014).
16. Louwen, R., Staals, R. H., Endtz, H. P., van Baarlen, P. & van der Oost, J. The role of CRISPR–Cas systems in virulence of pathogenic bacteria. *Microbiol. Mol. Biol. Rev.* **78**, 74–88 (2014).
17. Nunez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).
18. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
19. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
20. Shah, S. A., Erdmann, S., Mojica, F. J. & Garrett, R. A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.* **10**, 891–899 (2013).
21. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
22. Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
23. Wang, R., Preamplume, G., Terns, M. P., Terns, R. M. & Li, H. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure* **19**, 257–264 (2011).
24. Deltcheva, E. *et al.* CRISPR RNA maturation by *trans*-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
25. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
26. Samai, P. *et al.* Co-transcriptional DNA and RNA cleavage during type III CRISPR–Cas immunity. *Cell* **161**, 1164–1174 (2015).
27. Hale, C. R. *et al.* Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell* **45**, 292–302 (2012).
28. Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol. Cell* **46**, 606–615 (2012).
29. van Duijn, E. *et al.* Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from *Escherichia coli* and *Pseudomonas aeruginosa*. *Mol. Cell Proteom.* **11**, 1430–1441 (2012).
30. Zhang, J. *et al.* Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol. Cell* **45**, 303–313 (2012).
31. Wiedenheft, B. *et al.* Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489 (2011).
32. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1**, e60 (2005).
33. Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR–Cas systems. *Biol. Direct* **6**, 38 (2011).
34. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7 (2006).
35. Vestergaard, G., Garrett, R. A. & Shah, S. A. CRISPR adaptive immune systems of Archaea. *RNA Biol.* **11**, 156–167 (2014).
36. Staals, R. H. *et al.* Structure and activity of the RNA-targeting Type III-B CRISPR–Cas complex of *Thermus thermophilus*. *Mol. Cell* **52**, 135–145 (2013).
37. Spilman, M. *et al.* Structure of an RNA silencing complex of the CRISPR–Cas immune system. *Mol. Cell* **52**, 146–152 (2013).
38. Staals, R. H. *et al.* RNA targeting by the type III-A CRISPR–Cas Csm complex of *Thermus thermophilus*. *Mol. Cell* **56**, 518–530 (2014).
39. Tamulaitis, G. *et al.* Programmable RNA shredding by the type III-A CRISPR–Cas system of *Streptococcus thermophilus*. *Mol. Cell* **56**, 506–517 (2014).
40. Benda, C. *et al.* Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. *Mol. Cell* **56**, 43–54 (2014).
41. Hale, C. R., Cocozaki, A., Li, H., Terns, R. M. & Terns, M. P. Target RNA capture and cleavage by the Cmr type III-B CRISPR–Cas effector complex. *Genes Dev.* **28**, 2432–2443 (2014).
42. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **12**, 479–492 (2014).
43. Jackson, R. N. Lavin, M., Carter, J., & Wiedenheft, B. Fitting CRISPR-associated Cas3 into the helicase family tree. *Curr Opin Struct Biol.* **24**, 106–114 (2014).
44. Mulepati, S., Heroux, A. & Bailey, S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).
45. Zhao, H. *et al.* Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature* **515**, 147–150 (2014).
46. Taylor, D. W. *et al.* Structures of the CRISPR–Cmr complex reveal mode of RNA target positioning. *Science* **348**, 581–585 (2015).
47. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).
48. Sander, J. D. & Joung, J. K. CRISPR–Cas systems for editing, regulating and targeting genomes. *Nat. Biotech.* **32**, 347–355 (2014).
49. Altschul, S. F. & Koonin, E. V. PSI-BLAST — a tool for making discoveries in sequence databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).
50. Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335–1342 (2011).
51. Gong, B. *et al.* Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc. Natl Acad. Sci. USA* **111**, 16359–16364 (2014).
52. Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21**, 771–777 (2014).
53. Mulepati, S. & Bailey, S. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J. Biol. Chem.* **286**, 31896–31903 (2011).
54. Makarova, K. S. & Koonin, E. V. Annotation and classification of CRISPR–Cas systems. *Methods Mol. Biol.* **1311**, 47–75 (2015).
55. Nam, K. H. *et al.* Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR–Cas system. *Structure* **20**, 1574–1584 (2012).
56. Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.* **30**, 482–496 (2002).
57. Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA–Cas protein complex. *Cell* **139**, 945–956 (2009).
58. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845 (2008).
59. Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. A. Conditional tolerance of temperate phages via transcription-dependent CRISPR–Cas targeting. *Nature* **514**, 633–637 (2014).
60. Deng, L., Garrett, R. A., Shah, S. A., Peng, X. & She, Q. A novel interference mechanism by a type IIIB CRISPR–Cmr module in *Sulfolobus*. *Mol. Microbiol.* **87**, 1088–1099 (2013).
61. Peng, W., Feng, M., Feng, X., Liang, Y. X. & She, Q. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Res.* **43**, 406–417 (2015).
62. White, M. F. Structure, function and evolution of the XPD family of iron-sulfur-containing 5′→3′ DNA helicases. *Biochem. Soc. Trans.* **37**, 547–551 (2009).
63. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR–Cas adaptation. **519**, 199–202 *Nature* (2015).
64. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. *Genes Dev.* **29**, 356–361 (2015).
65. Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR–Cas systems. *Nucleic Acids Res.* **42**, 6091–6105 (2014).
66. Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrRNA and Cas9 families of type II CRISPR–Cas immunity systems. *RNA Biol.* **10**, 726–737 (2013).
67. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* **56**, 333–339 (2014).
68. Fonfara, I. *et al.* Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR–Cas systems. *Nucleic Acids Res.* **42**, 2577–2590 (2014).
69. Zhang, Y. *et al.* Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Mol. Cell* **50**, 488–503 (2013).
70. Schunder, E., Rydzewski, K., Grunow, R. & Heuner, K. First indication for a functional CRISPR/Cas system in *Francisella tularensis*. *Int. J. Med. Microbiol.* **303**, 51–60 (2013).
71. Makarova, K. S. *et al.* Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. *Extremophiles* **18**, 877–893 (2014).
72. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**, 4360–4377 (2013).
73. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007).
74. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
75. Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S. & Backofen, R. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.* **41**, 8034–8044 (2013).
76. Alkhnbashi, O. S. *et al.* CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489–i496 (2014).
77. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* **8**, R61 (2007).
78. Koonin, E. V. & Wolf, Y. I. Is evolution Darwinian or/and Lamarckian? *Biol. Direct* **4**, 42 (2009).
79. Leplae, R. *et al.* Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families. *Nucleic Acids Res.* **39**, 5513–5525 (2011).
80. Koonin, E. V. & Wolf, Y. I. Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front. Cell Infect. Microbiol.* **2**, 119 (2012).
81. Godde, J. S. & Bickerton, A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* **62**, 718–729 (2006).

82. Almendros, C., Mojica, F. J., Díez-Villaseñor, C., Guzmán, N. M. & García-Martínez, J. CRISPR–Cas functional module exchange in *Escherichia coli*. *mBio* **5**, e00767–e00713 (2014).
83. Shah, S. A. & Garrett, R. A. CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol.* **162**, 27–38 (2011).
84. Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE* **7**, e36972 (2012).
85. Takeuchi, N., Wolf, Y. I., Makarova, K. S. & Koonin, E. V. Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol.* **194**, 1216–1225 (2012).
86. Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. & Koonin, E. V. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR–Cas immunity. *BMC Biol.* **12**, 36 (2014).
87. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* **16**, 184–192 (2015).
88. Garrett, R. A., Vestergaard, G. & Shah, S. A. Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol.* **19**, 549–556 (2011).
89. Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
90. Puigbo, P., Wolf, Y. I. & Koonin, E. V. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).
91. Hooton, S. P. & Connerton, I. F. *Campylobacter jejuni* acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. *Front. Microbiol.* **5**, 744 (2014).
92. Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904–912 (2009).
93. Kwon, A. R. *et al.* Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. *Nucleic Acids Res.* **40**, 4216–4228 (2012).
94. Makarova, K. S., Anantharaman, V., Aravind, L. & Koonin, E. V. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. *Biol. Direct* **7**, 40 (2012).
95. Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.* **283**, 20361–20371 (2008).
96. Nam, K. H. *et al.* Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287**, 35943–35952 (2012).
97. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
98. Rouillon, C. *et al.* Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell* **52**, 124–134 (2013). 99.
Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).
100. Beloglazova, N. *et al.* Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J.* **30**, 4616–4627 (2011).
101. Ramia, N. F. *et al.* Essential structural and functional roles of the Cmr4 subunit in RNA cleavage by the Cmr CRISPR–Cas complex. *Cell Rep.* **9**, 1610–1617 (2014).
102. Zhu, X. & Ye, K. Cmr4 is the slicer in the RNA-targeting Cmr CRISPR complex. *Nucleic Acids Res.* **43**, 1257–1267 (2015).
103. Brendel, J. *et al.* A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (crispr)-derived rnas (crrnas) in *Haloferax volcanii*. *J. Biol. Chem.* **289**, 7164–7177 (2014).
104. Osawa, T., Inanaga, H., Sato, C. & Numata, T. Crystal structure of the CRISPR–Cas RNA silencing Cmr complex bound to a target analog. *Mol. Cell* **58**, 418–430 (2015).
105. Jung, T. Y. *et al.* Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure* **23**, 782–790 (2015).
106. Sapranauskas, R. *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* **39**, 9275–9282 (2011).
107. Makarova, K. S., Anantharaman, V., Grishin, N. V., Koonin, E. V. & Aravind, L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front. Genet.* **5**, 102 (2014).
108. Nam, K. H., Kurinov, I. & Ke, A. Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca^{2+}-dependent double-stranded DNA binding activity. *J. Biol. Chem.* **286**, 30759–30768 (2011).
109. Koo, Y., Jung, D. K. & Bae, E. Crystal structure of *Streptococcus pyogenes* Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. *PLoS ONE* **7**, e33401 (2012).
110. Arslan, Z. *et al.* Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res.* **41**, 6347–6359 (2013).
111. Lee, K. H. *et al.* Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. *Proteins* **80**, 2573–2582 (2012).
112. Zhu, X. & Ye, K. Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR–Cas systems. *FEBS Lett.* **586**, 939–945 (2012).
113. Shao, Y. *et al.* Structure of the Cmr2–Cmr3 subcomplex of the Cmr RNA silencing complex. *Structure* **21**, 376–384 (2013).
114. Guy, C. P., Majernik, A. I., Chong, J. P. & Bolt, E. L. A novel nuclease-ATPase (Nar71) from archaea is part of a proposed thermophilic DNA repair system. *Nucleic Acids Res.* **32**, 6176–6186 (2004).
115. Cass, S. D. *et al.* The role of Cas8 in type I CRISPR interference. *Biosci. Rep.* **35**, e00197 (2015).
116. Reeks, J. *et al.* Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. *RNA Biol.* **10**, 762–769 (2013).
117. Jackson, R. N. & Wiedenheft, B. A conserved structural chassis for mounting versatile CRISPR RNA-guided immune responses. *Mol. Cell* **58**, 722–728 (2015).

Acknowledgements
K.S.M., Y.I.W., D.H. and E.V.K. are supported by the National Institutes of Health (NIH) Intramural Research Program at the National Library of Medicine, US Department of Health and Human Services. R.M.T. and M.P.T. are supported by NIH grants RO1 GM54682 and RO1 GM99876. J.v.d.O. was partly supported by SIAM Gravitation Grant 024.002.002 from the Netherlands Organization for Scientific Research (N.W.O.). S.J.J.B. was financially supported by an NWO Vidi grant (864.11.005) and European Research Council (ERC) Stg (639707). A.F.Y. is supported by the Natural Sciences and Engineering Research Council (NSERC) Strategic Network Grant IBN and NSERC Discovery grant. S.M. acknowledges funding from Natural Sciences and Engineering Research Council of Canada (Discovery program) and holds a Tier 1 Canada Research Chair in Bacteriophages. F.J.M.M. is supported by the Ministerio de Economía y Competitividad (BIO2014-53029). R.B. is supported by the Deutsche Forschungsgemeinschaft (DFG) grant (BA 2168/5-2). S.A.S. and R.A.G. were funded primarily by the Danish Natural Science Research Council. O.S.A., F.C., S.J.S., R.B., S.A.S and R.A.G. are grateful to all members of the FOR1680 for helpful discussions.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
TIGRFAM file directory: ftp://ftp.jcvi.org/pub/data/TIGRFAMs
TIGR02165 | TIGR04330

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table) | S3 (box) | S4 (table) | S5 (box) | S6 (box) | S7 (table) | S8 (table) | S9 (box) | S10 (box) | S11 (table) | S12 (box) | S13 (box) | S14 (box) | S15 (figure)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Characterizing leader sequences of CRISPR loci

Omer S. Alkhnbashi, Shiraz A. Shah, Roger A. Garrett, Sita J. Saunders, Fabrizio Costa and Rolf Backofen. **Bioinformatics Journal, 2016**, doi: 10.1093/bioinformatics/btw454.

Personal contribution

My contribution in this project is a major one. I conceived the method and its analysis. Furthermore, I implemented the software and was involved in planning and writing the article.

Omer S. Alkhnbashi

The following co-authors confirm the above stated contribution.

Shiraz A. Shah

Roger A. Garrett

Sita J. Saunders

Fabrizio Costa

Rolf Backofen

Bioinformatics, 32, 2016, i576–i585
doi: 10.1093/bioinformatics/btw454

OXFORD

Characterizing leader sequences of CRISPR loci

Omer S. Alkhnbashi[1], Shiraz A. Shah[2], Roger A. Garrett[2], Sita J. Saunders[1], Fabrizio Costa[1,*] and Rolf Backofen[1,3,*]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, 79110 Freiburg, Germany, [2]Archaea Centre, Department of Biology, University of Copenhagen N, DK2200 Copenhagen N, Denmark and [3]BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, University of Freiburg, Freiburg im Breisgau, Germany

*To whom correspondence should be addressed

Abstract

Motivation: The CRISPR-Cas system is an adaptive immune system in many archaea and bacteria, which provides resistance against invading genetic elements. The first phase of CRISPR-Cas immunity is called adaptation, in which small DNA fragments are excised from genetic elements and are inserted into a CRISPR array generally adjacent to its so called leader sequence at one end of the array. It has been shown that transcription initiation and adaptation signals of the CRISPR array are located within the leader. However, apart from promoters, there is very little knowledge of sequence or structural motifs or their possible functions. Leader properties have mainly been characterized through transcriptional initiation data from single organisms but large-scale characterization of leaders has remained challenging due to their low level of sequence conservation.

Results: We developed a method to successfully detect leader sequences by focusing on the consensus repeat of the adjacent CRISPR array and weak upstream conservation signals. We applied our tool to the analysis of a comprehensive genomic database and identified several characteristic properties of leader sequences specific to archaea and bacteria, ranging from distinctive sizes to preferential indel localization. *CRISPRleader* provides a full annotation of the CRISPR array, its strand orientation as well as conserved core leader boundaries that can be uploaded to any genome browser. In addition, it outputs reader-friendly HTML pages for conserved leader clusters from our database.

Availability and Implementation: *CRISPRleader* and multiple sequence alignments for all 195 leader clusters are available at http://www.bioinf.uni-freiburg.de/Software/CRISPRleader/.

Contact: costa@informatik.uni-freiburg.de or backofen@informatik.uni-freiburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

CRISPR-Cas is an adaptive immune system of archaea and bacteria that provides resistance against invading viruses and plasmids (Barrangou and van der Oost, 2013). 84 and 45% of sequenced archaeal and bacterial genomes, respectively, encode a CRISPR-Cas system (Barrangou and van der Oost, 2013). Each CRISPR-Cas locus comprises several regions. Central to the system is a small 19–48 bp sequence, the CRISPR repeat, which plays a key role in regulating all aspects of CRISPR-Cas function. The CRISPR repeat acts as a regulatory guide and the associated Cas proteins provide the main machinery required for the defence mechanism. The CRISPR array contains repetitions of a CRISPR *repeat* sequence interspaced by foreign DNA fragments (spacers) and can consist of hundreds of repeat-spacer units. Currently, CRISPR-Cas systems are classified into five types and at least 16 subtypes (Makarova *et al.*, 2015; Vestergaard *et al.*, 2014). CRISPR-Cas systems have had a monumental impact on biotechnology as a basis for developing cheap and effective genome-editing techniques for almost any organism (Hsu *et al.*, 2014; Li *et al.*, 2016).

The function of CRISPR-Cas systems can be divided into three major phases: (i) *adaptation*, where a short fragment of invading DNA is inserted into the CRISPR locus for future recognition of that invader; (ii) *expression*, which involves the biogenesis of guide RNA units (crRNA) and their integration into large RNA–protein effector complexes and (iii) *interference*, where these effector complexes vigilantly scan for and degrade invading genetic material previously identified by—and integrated into—the CRISPR-Cas system (Barrangou and van der Oost, 2013). The least understood phase in

CRISPR-Cas immunity is adaptation where a foreign DNA fragment from invading genetic material is integrated.

The integration usually occurs upstream of the first repeat, before a region denoted as the *leader*, which contains regulatory elements important for adaptation. The leaders vary in size, extending from 47 bp in some bacteria to a few hundred bp in some hyperthermophilic archaea, and they tend to exhibit longer regions of low complexity sequence, with limited sequence conservation (Shah and Garrett, 2011). Owing to their limited sequence conservation, even between very similar archaea and bacteria, very little information is available to date and no bioinformatic tool currently exists that can automatically annotate leaders and define their boundaries.

To improve our understanding of the adaptation phase, we studied leader sequences in more detail. Individual experimental studies have demonstrated that the leaders carry the main bacteria- or archaea-specific promoters for CRISPR transcription (Brouns *et al.*, 2008; Lillestol *et al.* 2006, 2009), and that they contain signals for CRISPR-Cas adaptation (Diez-Villasenor *et al.*, 2013; Erdmann and Garrett, 2012; Yosef *et al.*, 2012). The existence of adaptation signals in the leader region is also supported by the existence of leaderless CRISPR-arrays in some crenarchaea, which do not acquire new spacers (Gudbergsdottir *et al.*, 2011; Lillestol *et al.*, 2006, 2009). However, leaderless CRISPR are still functional in the remaining immunity steps because they yield processed CRISPR RNAs (crRNAs), presumably as a result of transcription from promoters taken up randomly in spacers (Deng *et al.*, 2012; Wurtzel *et al.*, 2010).

Concerning the typical length of a leader region, experimental studies of the type I–E CRISPR-Cas system of *Escherichia coli* provided evidence for 40–60 bp of the leader region, located immediately upstream from the first CRISPR repeat, being essential for spacer acquisition (Yosef *et al.*, 2012). Further experiments with the same type I–E system narrowed the critical region to positions −1 to −43 (in relation to the first CRISPR repeat) (Diez-Villasenor *et al.*, 2013). Moreover, for the type I–A system of *Sulfolobus*, a natural deletion of the leader region from positions −47 to −70 resulted in a low level of adaptation activity and also a decreased specificity of spacer acquisition whereby spacer insertions occurred all along the CRISPR array and not just at the first repeat (Erdmann and Garrett, 2012; Garrett *et al.*, 2015).

The existence of adaptation signals in the leader region is also supported by evolutionary studies. Despite their relatively low sequence conservation, sequence clustering studies for the *Sulfolobales* have shown that the leaders tend to coevolve with CRISPR repeat, the adaptation module (Cas1, 2 and 4) and the protospacer-adjacent motif (PAM) (Shah and Garrett, 2011). Experimental support for this coevolution was provided by studies on the *E.coli* type I–E system (Diez-Villasenor *et al.*, 2013). Leaders also carry conserved sequence motifs, currently of unknown function (Garrett *et al.*, 2015; Mojica and Garrett, 2013). The latter are possibly involved in aligning multiple RNA polymerase complexes for CRISPR transcription and/or in assembling Cas proteins adjacent to the CRISPR adaptation site (Lillestol *et al.*, 2009; Marraffini and Sontheimer, 2008; Mojica *et al.*, 2009; Rollie *et al.*, 2015; Shah *et al.*, 2009).

Existing CRISPR-prediction tools do not provide any information regarding CRISPR leaders. In this study, we developed *CRISPRleader*, an efficient approach to determining CRISPR leader boundaries by focusing on leader sequence conservation within groupings based on the similarity of the repeats in the adjacent CRISPR arrays. Our method utilizes a string-kernel technique that can capture more information than traditional sequence alignments and is especially capable of detecting a collection of local motifs. We built specialized HMM models for each of the 51 and the 144 CRISPR-leader clusters from archaea and bacteria, respectively. The method takes a complete genome or draft genome as input and first predicts all possible CRISPR arrays in the correct orientation, and then annotates the CRISPR-leader boundaries.

2 Materials and methods

2.1 CRISPR dataset

In this study, we use the comprehensive dataset of CRISPR arrays of archaeal and bacterial genomes which were downloaded from the CRISPRmap webserver (Alkhnbashi *et al.*, 2014; Lange *et al.*, 2013). The dataset contains 217 archaeal genomes that encode around 985 CRISPR arrays and 1409 bacterial genomes with 3515 CRISPR arrays (a total of 4500 CRISPR arrays). In achaeal CRISPR arrays, the average length of repeats is 29 nt and the average number of repeats per array is 18. In bacteria, in contrast, the average number of repeats per array is 13 and the average repeat length is 30 nt.

2.2 CRISPR leader sequence identification

Although the characteristic repeat-spacer architecture of CRISPR arrays can be easily detected, the orientation of the CRISPR array is inherently ambiguous and thus the determination of the strand from which crRNAs are generated is uncertain. Using the machine learning approach presented in Alkhnbashi *et al.* (2014), it is, however, possible to identify the most probable orientation. Given the array orientation, it is generally assumed that the 3′ boundary of the leader sequence is immediately adjacent to the first CRISPR repeat as both leader and CRISPR array is transcribed in a single transcript (Scholz *et al.*, 2013). Figure 1 depicts a schematic view of a CRISPR locus with the CRISPR array and its respective leader region.

2.2.1 Criteria to determine leaderless CRISPR arrays

In this work, we define a leaderless CRISPR array with the following criteria. First, the distance between the 3′ end of an annotated gene and the 5′ end of the CRISPR array should be less than 20 bp. In the literature, leader regions with experimentally verified function are definitely longer than 20 bp. Second, if the curve fitting procedure fails, it indicates a complete lack of detectable sequence similarity, and we, therefore, discard all the sequences in the leader cluster. Third, we check if the average pairwise similarity as computed by the Needleman–Wunsh algorithm is less than 50%. In this case, there could still be a functional leader present, however, since the sequence similarity between the associated CRISPR repeats is already high, we assume that it is unlikely for such a divergent leader to exist.

2.2.2 CRISPR-leader clusters

It has been shown that the leader sequence coevolves with CRISPR repeats, with the Cas1 protein and with the PAM motif (Shah and Garrett, 2011). To make use of this evolutionary information, we introduce the notion of a leader cluster, which consists of leaders grouped together according to their associated repeat families. By doing so, we overcome the problem of the limited sequence similarity of leaders. To group the repeat sequences, we follow the approach presented in CRISPRmap (Alkhnbashi *et al.*, 2014; Lange *et al.*, 2013). In detail, given a CRISPR array, we first compute the *consensus*-repeat sequence by aligning all repeat sequences without gaps and then take for each position the most frequent nucleotide. We define the similarity between two consensus repeat sequences as

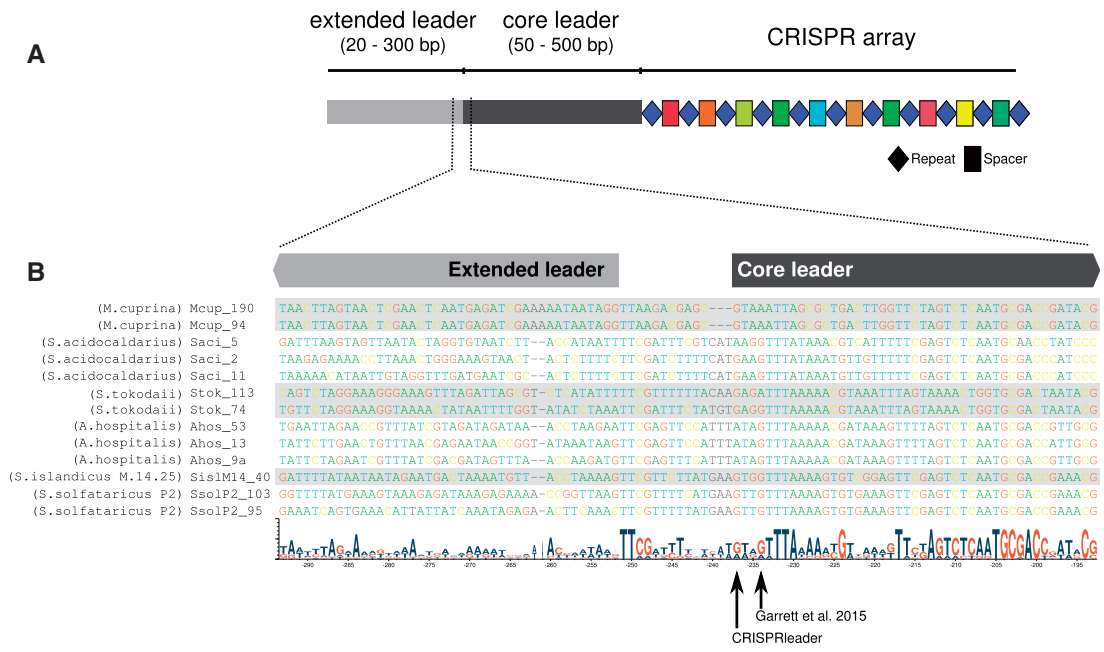


Fig. 1. (**A**) Schematic view of the elements of a CRISPR array showing the repeats (blue diamonds) and spacers (coloured rectangles) of a CRISPR array and the leader region, which we separate into a core and an extended leader. The *core* leader is generally conserved across different host species and is shorter than the *extended* leader which is normally only conserved between multiple leader copies in the same genome. (**B**) Sequences correspond to a cluster of related leaders shared between species of the genera Acidianus, Metalosphaera and Sulfolobus. Each leader is identified by the number of repeats in the adjacent CRISPR. *CRISPRleader* predicts the length of the core leader, since the extended leader is assumed to be functionally less important. In the bottom, we provide an example of a leader alignment to show a detailed view at the junction between the core and extended leader. Here it is possible to see how the extended part is only conserved between multiple copies in the same organism. In contrast, the core part is conserved across all of the different hosts, is underlined by the sequence logo below. The leader boundary predicted by *CRISPRleader* and the boundary determined by expert inspection are indicated by black arrows at the bottom

the global pairwise alignment score computed using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970). To obtain coherent sets, we then apply Markov Clustering (MCL) (Enright *et al.*, 2002). In CRISPRmap, it was found that better results can be obtained if the similarity matrix is thresholded, i.e. if we set to 0, the similarity value for pairs of repeat sequences that are not sufficiently similar. The only tunable parameter for the MCL algorithm is called 'inflation' and determines the scale of the clustering (i.e. if we prefer many small clusters or few large ones). We optimized these parameters to guarantee that archaea and bacteria are always placed in distinct clusters. This yielded a value of 86 for the similarity threshold and a value of 2.2 for inflation. In this setting, *CRISPRleader* identifies 52 clusters in our dataset of 770 archaeal CRISPR leaders (with a number of leader sequences per cluster that ranges from 3 to 69) and 144 clusters in our set of 2224 bacterial CRISPR leaders (with a number of leader sequences per cluster that ranges from 3 to 184). See Table 1 for details.

2.2.3 CRISPR-leader similarity profile

In the following, we describe how *CRISPRleader* estimates the 5′ leader boundary (CRISPR repeat distal) based loosely on the sequence conservation within a set of leader sequences that are clustered together. We exploit two key assumptions: (i) the 3′ end of the leader (CRISPR repeat proximal) is immediately upstream of the first repeat in the CRISPR array (Brouns *et al.*, 2008; Lillestol *et al.*, 2006, 2009) and (ii) due to evolution-related adaptation signals, the leader sequence will likely exhibit detectable signals of sequence conservation.

First, we trim all the leader sequences to the smallest of (i) an upper limit of 600 nt or (ii) the first occurrence of a predicted protein-coding gene using PRODIGAL (Hyatt *et al.*, 2010) version 2.6.2.

A traditional approach to finding the leader boundaries based on sequence conservation would be to perform a global or local multiple-sequence alignment. In practice, however, the resulting alignments are too noisy. This is likely due to the small size of the conserved regions relative to the sequences length (e.g. 40 nt within 600 nt) and to the small number of sequences (see Section 3.1 for a more detailed analysis). To overcome this issue, we developed a more robust approach based on string kernels (see the following Section for details on the notion of kernels). We start by exploiting the fact that the 3′ boundary of the leader is known to be adjacent to the CRISPR array. We then align all sequences at the CRISPR array's boundary. Subsequently, we apply a running windowing approach: we extract a subsequence from each leader that spans the same W positions and we consider a newly developed average pairwise similarity among these subsequences; we shift the window of a step of S nucleotides and repeat the procedure on the next set of subsequences. Our new pairwise sequence similarity is computed using the Neighbourhood Subgraph Pairwise Decomposition Kernel (NSPDK) (Costa and Grave, 2010). To normalize this similarity value, we consider the average pairwise similarity of the subsequences after a random di- or tri-nucleotide shuffle. The conservation signal is then the log ratio of these two average similarities: when the subsequences are not evolutionarily related, we expect the two similarity values to be comparable yielding log odds scores close to 0. To detect the end of the conserved region, we smoothed the log odds signal by subsequently fitting a parameterized sigmoid curve σ_θ with parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ under the constraints that it saturates to 0 at one of the extremes. The parameterization's semantics

Table 1 The leader clusters are summarized at the CRISPRmap repeat family level

Repeat	Repeat consensus sequence	Phylogenetic distribution	Clusters	# Leaders	Avg length
F3(402)	GXXXXXXXXXXXAXXGXATTGAAAG	Crenarchaeota	22	294	210 (±50)
F4(329)	GTTXXAATMAGACXXXWXXXGRATXGAAAX	Euryarchaeota	12	280	236 (±115)
F12(68)	GTTXCAGAXGXACCXTTGTGGGXTTGAA	Euryarchaeota	8	57	111 (±16)
F15(45)	GTTTCXGWAGACATGTXTGGAAA	Euryarchaeota	2	23	366 (±74)
F16(45)	CCAGAAATCAAAAGATAGTWGAAAC	Crenarchaeota	4	41	199 (±3)
F2(556)	GTTTXXAKXXTACCTATXXGGRATTGAAAC	Bacteroidetes/Crenarchaeota/Firmicutes/ Thermotogae	27	437	158 (±46)
F10(89)	TTXXARWXXXXTCCAXTAAAACAAGGATTGAAAC	Euryarchaeota/Firmicutes	6	45	254 (±121)
F1(671)	GTXXTCCCCGCGCXXGCGGGGATRXXCCX	Proteobacteria/Actinobacteria	21	540	103 (±53)
F5(296)	GTCGCXCCCYXXXXGXGXGCGTGGATTGAAAX	Actinobacteria/Planctomycetes/Firmicutes	7	208	83 (±66)
F6(264)	GTTCACTGCCGYAYAGGCAGCTTAGAAA	Proteobacteria	3	230	146 (±14)
F7(236)	XXTKXAMXXTAAXXXXXGWXGTATXTAAAT	Firmicutes/Fusobacteria	13	180	191 (±54)
F8(175)	TXXXXXXXXXCCCCGXXAGGGGAYKGAAAC	Actinobacteria/Deinococcus-Thermus	12	117	157 (±90)
F9(146)	TTXXAAXXXCCCTXTXAGGGATTGAAAC	Cyanobacteria/Firmicutes	10	112	149 (±63)
F11(76)	GTXXXAXXGXCCYGATKXXXARGGGATTRMGAC	Proteobacteria/Bacteroidetes	6	36	107 (±29)
F13(61)	GTTTTAGAGCXXTGTTRTTTXGAATGGTXCCAAAAC	Firmicutes	2	44	210 (±15)
F14(53)	GXXXCXXCGCXGXGGCCXCATTGAAGC	Proteobacteria/Planctomycetes/Firmicutes	3	29	137 (±74)
F17(38)	STGCXXTGATGCCGXWAGGCGTTGACAC	Cyanobacteria/Proteobacteria/Spirochaetes	1	4	213 (±7)
F18(36)	GTTTCYCCTGRRGGTTGAAA	Cyanobacteria/Firmicutes	5	16	176 (±70)
F19(29)	GTTKTAGYYCCYTTTYWMATTTCKYWRTGSTAAAT	Proteobacteria	3	18	116 (±14)
F20(26)	XXXXXGCGXXXCGGCGGXXGXGGX	Acidobacteria/Proteobacteria	1	4	101 (±15)
F21(25)	GTTGWYAAARTAAATTGAAAGCAAWTCACAAC	Bacteroidetes/Ignavibacteriae	1	15	100 (±0.0)
F22(19)	GTYTAGRTGATGTRATCAATAGKTYAAGAC	Firmicutes	2	10	599 (±0.46)
F23(14)	GTTTTGTACTCTARATTTAAGTAACGTAAAAC	Firmicutes	1	6	197 (±8)
F24(11)	WMRTAMCCCCXXAKXAXAGGGGACKARAAC	Firmicutes	1	3	382 (±2)

For each repeat family, the total number of members is given in parentheses, along with the consensus repeat sequence and the taxonomic distribution. The number of leader clusters within each repeat family is also given, along with the total number of leaders found, as well as their average length.

is: θ_1 represents the maximal conservation log odds value, θ_2 represents the length scale factor and θ_3 encodes the position of maximal slope, i.e. the point when the signal transitions from one of the saturated region to the other:

$$\sigma_\theta(x) = \theta_1 \cdot \frac{e^{-\frac{x-\theta_3}{\theta_2}}}{1 + e^{-\frac{(x-\theta_3)}{\theta_2}}}.$$

The estimated leader 5′ boundary is then directly read from θ_3. Figure 2 visualizes the complete process for detecting the boundary of the conservation signal within each leader cluster.

2.2.4 String kernels and explicit feature construction

A string kernel is a function that allows the computational manipulation of strings in a high-dimensional, implicit feature space without ever computing the actual coordinates of the string in that space, but rather by simply computing the inner products between the images of pairs of strings in the feature space. The inner product computed by the kernel function can be used to define a similarity notion. When normalized, the kernel maps pairs of strings s and s' into the interval $[0, 1]$, where 1 means that the two strings are indistinguishable (for the kernel) and 0 that they do not share any resemblance. Popular string kernels are based on the notion of *k-mers*, i.e. substrings of size k. The *k-mer* kernel (also called spectral kernel in Leslie *et al.*, 2002) between s and s', i.e. $K(s, s')$, is the number of the *k-mers* that are identical between s and s'. A normalized kernel computes the fraction of identical *k-mers* w.r.t. the total number of *k-mers* present in the two strings s and s', often as the quantity: $K(s, s')/\sqrt{K(s, s) \cdot K(s', s')}$. Since the occurrence of *k-mers* is exponentially less probable w.r.t. their size k, there is little to gain in considering large *k-mers* (e.g. $k > 10$) when comparing biological

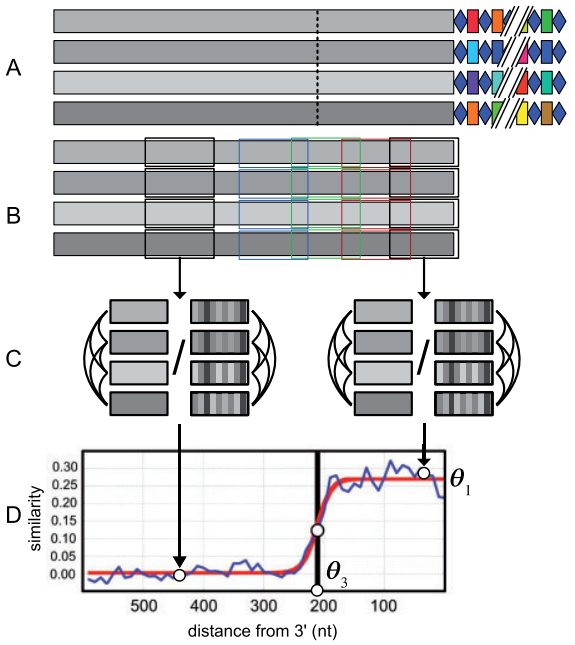


Fig. 2. Leader boundary identification: (**A**) leader sequences are clustered together according to the similarity between the associated repeat sequences; the 3′ end of the sequences in a cluster is aligned w.r.t. the first CRISPR repeat and (**B**) shifting windows spanning the same positions are extracted. (**C**) The average pairwise similarity between all subsequences in a window is computed using the proposed string kernel; the same procedure is applied to shuffled sequences to compute the log odds ratio and (**D**) a saturating function is fitted to distinguish the highly conserved region from the non-conserved one; the point of maximum slope θ_3 is returned as leader boundary

sequences from different species. Small *k-mers*, however, might not yield a sufficient discriminative power. To mitigate these problems, a notion of 'approximate match' was introduced in Leslie *et al.* (2004), where the insertion, deletion or mismatch of up to m components of the *k-mer* is tolerated when counting the correspondences. In practice however, these approximate techniques lead to an increase in run-times and are not always effective in significantly increasing the discriminative power.

The NSPDK approach tries to find a better compromise by restricting the type of mismatches. While the kernel introduced in Costa and Grave (2010) is primarily designed for graphs, here we develop a restricted version for sequences. In detail, the features considered here are pairs of *k-mers* at a fixed distance d, i.e. we assume that there exist a relation $\Phi(s,k,d)$ that is verified for pairs of substrings a, b of s that are of length k and such that their distance is d, we denote such a pair as ϕ_i. The distance between two substrings a, b of s is defined as the length of the substring between the first character of a and the first character of b. The kernel is defined as:

$$K^{k,d}(s,s') = \sum_{\substack{\phi_i \in \Phi^{-1}(s,k,d) \\ \phi_j \in \Phi^{-1}(s',k,d)}} \delta(\phi_i,\phi_j)$$

where $\Phi^{-1}(s,k,d)$ is the inverse of the relation $\Phi(s,k,d)$, i.e. it is the set of all ϕ_i, i.e. pairs of substrings of length k at distance d, and $\delta(x,y)$ is the Kronecker delta, i.e. the function that evaluates to 1 if $x=y$ and to 0 otherwise. We consider the normalized kernel: $\widehat{K}^{k,d}(s,s') = K^{k,d}(s,s')/\sqrt{K^{k,d}(s,s)\cdot K^{k,d}(s',s')}$. Given a maximal value for $k \leq k^*$ and $d \leq d^*$, we consider all the possible combination of values for k and d:

$$\kappa(s,s') = \sum_{\substack{k \leq k^* \\ d \leq d^*}} \widehat{K}^{k,d}(s,s')$$

and finally, we consider the normalized kernel: $\widehat{\kappa}(s,s') = \kappa(s,s')/\sqrt{\kappa(s,s)\cdot\kappa(s',s')}$.

Differently from the standard kernel approach, where the inner product is computed implicitly without having to compute the coordinates of a string in the high-dimensional space, a variant of NPDK, introduced in Frasconi *et al.* (2012), allows one to construct the explicit feature representation in an efficient way. The idea is to exploit a hashing encoding of the decomposed parts. Here, since each feature is a pair of *k-mers* at a given distance, we first hash the *k-mers* individually and then hash the integer triplet formed by the hash values for the two *k-mers* and the distance value into a single integer code. This integer code is then the feature indicator. For each such feature, we count how many times that specific pair of *k-mers* occurs in the sequence. The resulting data structure is a sparse vector representation of the string, which allows an efficient computation of $K^{k,d}(s,s')$ as a dot product.

2.2.5 Optimization of parameters

The method we have employed for the leader boundary determination exposes several parametric choices: the window and step size (W, S), the string kernel complexity (k^*, d^*), the shuffling order for the normalization. To optimize their values, we used supervised data from the work by Lillestol *et al.* (2009) which provides two sets of six and eight leaders for archaeal organisms with experimental evidence for their boundaries. We computed the discrepancy between the experimental and the predicted boundaries as the average squared difference expressed in number of nucleotides. In this setting, we obtained the best results when the window size was $W=40$ nt (selected from $\{10,20,30,40,60\}$), the step size $S=10$ nt (selected from $\{5,10,15,20,30\}$), the maximal *k-mer* size $k^*=3$ and the maximal gap size $d^*=3$ (selected from $\{1,2,3,4,5,6,7,8,9,10\}$), and the order of the shuffling 2 (selected from {1, 2, 3}), i.e. we used dinucleotide shuffling.

2.2.6 CRISPR-leader boundary adjustment via sequence alignment

CRISPR-leaders are inherently quite long, surpassing hundreds of nucleotides in many cases. Thus, there is a potential for indels to accumulate in regions within the leader which are relatively less important, functionally. This means that even closely related leaders can differ in size by tens of nucleotides. To deal with this problem, *CRISPRleader* implements a post-processing procedure to refine the boundary estimate for each individual leader within a cluster. After determining the leader cluster and an initial boundary estimate as previously detailed, we extend the length of each leader sequence (in the 5′ direction) by one third of the respective sequence alignment length of its leader cluster to accommodate undetected indels events. We then perform multiple sequence alignment using the MAFFT tool (Katoh *et al.*, 2002) on the extended sequences belonging to each cluster. The length of the conserved consensus sequence is then yielded as the adjusted boundary.

2.2.7 Automated annotation of core leaders

When given CRISPR arrays of a single organism, *CRISPRleader* automatically annotates the leader region according to our data and delivers a detailed report of all CRISPR arrays, including the boundary of the core leader and the consensus repeat. For the core leader annotation, we first identify the leader cluster from our dataset according to the best-matching consensus repeat. Second, we use the boundaries associated to the corresponding leader cluster to extract the candidate leader core region. Third, we determine whether the putative leader sequence shows sufficient sequence similarity to that leader cluster. For that purpose, we use Hidden Markov Models (HMMs) that we have computed for each leader cluster using HMMER (Eddy, 2011). The corresponding HMM is then used to compute the log-odds score to test whether the new candidate is similar enough to the sequences in the cluster. If the score lies within two standard deviations from the mean log-odds score of the group we accept the sequence, align it to the clustered leaders and compute the length of the conserved consensus sequence. Finally, we report the full alignment with the other clustered leader sequences to highlight insertion or deletion events.

3 Results and discussion

3.1 Conservation profiles could not be detected by alignment-based methods

The more traditional approach to finding the leader boundaries based on sequence conservation would be to perform a global multiple-sequence alignment. In practice, however, the resulting alignment is too noisy to be used to derive a reliable signal. We can hypothesize several reasons that contribute to this situation. First, the conserved region is generally small relative to the sequence lengths, e.g. values of 60 nt within the overall 600 nt sequence are not uncommon. Second, the number of leader sequences that are grouped together in a cluster can be small (less than five). Third, current alignment techniques cannot consistently accommodate transposition events. In practice, it is hard to globally align sequences

when relatively large insertions, deletions and transposition events are possible. For this reason, a more 'local' approach based on *k-mers* can be more effective. To experimentally determine the quality of the conservation signals that can be obtained via alignment strategies, we applied both a global multiple-sequence alignment and a local-alignment strategy to the sequences in the leader clusters. We proceed by incrementally extending the aligned sequence lengths by 10 nt, always starting from the CRISPR array boundary. As shown anecdotally in Supplementary Figure 2, a clear end of the conserved region cannot be reliably detected using global and local alignments, whereas our string-kernel approach shows a very clear conservation boundary on the same data. Owing to this described limitation of leader-boundary detection using the traditional alignment approaches, leaders have not been well characterized in the literature to date. Once detected with our method, however, we could produce well-conserved multiple sequence alignments of precomputed leader clusters for all published genomes that we publish on our website (see availability).

3.2 A well-conserved core leader controls adaptation and transcription

The region of sequence conservation in leaders tends to extend further upstream of the CRISPR locus when similar leaders are compared within the same genome or between closely related strains of the same species. In contrast, when comparing similar leaders across different species, the conserved regions end closer to the CRISPR locus. Here we define the former as the extended leader and the latter as the core leader (Fig. 1). The sequence conservation in the CRISPR-distal regions of the extended leader is likely to have resulted from relatively recent duplication events. The core leader, on the other hand, tends to be well conserved, even for divergent hosts, which implies that only the core region is of special functional significance. In the present study, we predict the boundary of the core leader on the assumption that the additional sequence in the extended leader carries less significant and unknown functions.

According to the literature, two types of regulatory signals fall into the core-leader region. First, both archaeal and bacterial promotors (for transcription) have been detected in the region directly upstream of the CRISPR locus in different type I systems (Brouns *et al.*, 2008; Lillestol *et al.*, 2009). Second, various lines of evidence have implicated this leader region in the adaptation mechanism. In a type I–A system in *Sulfolobus solfataricus*, a natural leader deletion (Fig. 5B) extending from positions −47 to −70 (from the first CRISPR repeat) led to relatively infrequent spacer insertions at different positions along the CRISPR locus (Erdmann and Garrett, 2012; Garrett *et al.*, 2015). In the type I–E system of *E.coli*, it was shown that exchanging leaders between similar CRISPR loci resulted in inverted spacer insertion and in altered sizes of incorporated spacers (Diez-Villasenor *et al.*, 2013). Moreover, attempts to localize the leader regions that are essential for adaptation in type I–E systems demonstrated that some sequences contained within the region −1 to −41 or −60 were essential for adaptation (Yosef *et al.*, 2012). Thus, it is likely that the sequence elements in the core leader normally regulate the frequency and specificity of spacer insertion at the first repeat (by Cas1, demonstrated experimentally to facilitate insertion (Rollie *et al.*, 2015)) as well as controlling the size and orientation of the new spacer.

3.3 The conservation of core leaders is more widespread than previously believed

Early studies characterizing the leader noted its lack of conservation beyond the species boundary (Jansen *et al.*, 2002; Mojica *et al.*, 2000), and this observation was reiterated in later studies (Horvath *et al.*, 2009; Lillestol *et al.*, 2006) when more genome data were available. The first report of related leaders spanning several species and genera was for the crenarchaeal order Sulfolobales (Lillestol *et al.*, 2009), but similar findings have not been made subsequently for other archaea or bacteria. This has probably been due to insufficient genomic data being available and to the difficulty in identifying the leaders using traditional alignment approaches. Nevertheless, the restriction of leaders within tight phylogenetic boundaries stands in contrast to what has otherwise been shown for CRISPR-Cas systems, where most subtypes (except those of type II systems) are shared between the bacterial and archaeal domains (Makarova *et al.*, 2015).

In our study, we find numerous archaeal leader clusters that are shared between several species and genera, but seldom cross the order boundary (Supplementary Fig. 7). For example, the largest archaeal leader cluster contains sequences from *Pyrococcus* and *Thermococcus* species only, both members of the order *Thermococcales*. Of the 10 largest archaeal leader clusters, only one is represented across more than one order (Table 1 and Supplementary Table *S*1). In contrast, bacterial leader clusters are much more diverse taxonomically (compare Fig. 3A and Supplementary Fig. 7). The two largest bacterial leader clusters are represented by several orders within the phylum Proteobacteria. The third-largest bacterial leader cluster contains members from multiple phyla, including, but not restricted to, Proteobacteria, Firmicutes, Actinobacteria, Chlorobi and Spirochaetes, all within a single leader cluster. The same staggering diversity is seen throughout the other major bacterial leader clusters (Fig. 3).

Conventional wisdom within the field has long been that CRISPR leaders are not conserved beyond the species boundary. Conservation across the order Sulfolobales was shown previously (Lillestol *et al.*, 2009) and our results show that order-wide conservation is normal for archaea. In contrast, for bacteria there seem to be no taxonomic boundaries for leader-cluster diversity. We found similar leaders in bacteria as diverse as *Pseudomonas* and *Clostridium* and the compatibility of leaders across diverse phyla is comparable to that of the CRISPR subtypes themselves. This kind of diversity within bacterial leader clusters seems to be the rule rather than the exception, but has so far gone undetected owing to the lack of reliable methods for leader identification. As for the stark difference between archaea and bacteria, in terms how widely conserved their leader clusters are, no straightforward explanation arose from the data. One factor may be that the currently sequenced archaeal genomes are strongly biased to extremophiles present in isolated environments which include salt lakes and acidophilic hot springs that exhibit more limited biodiversity. The few archaeal genomes that do originate from more complex microbial environments do not tend to carry CRISPRs. Thus, the lack of widespread conservation of currently sequenced archaeal leader clusters may simply result from the formidable barriers to horizontal gene transfer imposed by their habitats.

3.4 Core leaders display different patterns of conservation

Leader clusters that are more taxonomically restricted tend to show a relatively high and uniform sequence conservation throughout the

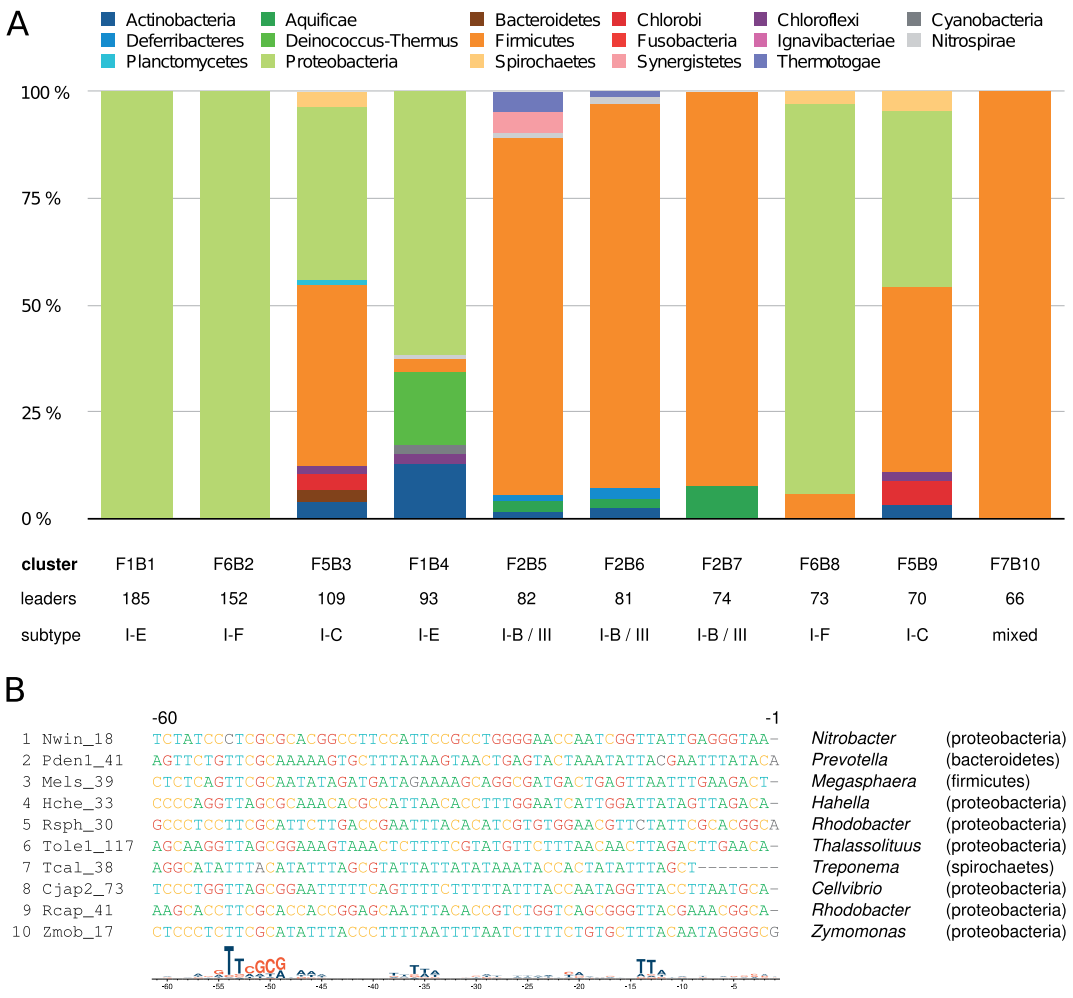


Fig. 3. (**A**) The taxonomic distribution is shown, on the phylum-level, for each of the ten largest bacterial leader clusters. Despite proteobacteria and firmicutes dominating the underlying genomic data, diversity is still evident with most clusters representing several additional phyla. The number of leaders in each family is also shown along with the principal CRISPR-Cas subtype associated with the leaders. (**B**) An alignment of the core leader from 10 randomly selected members of cluster 3 is shown, along with names of the genera and phyla they originated from. The logo plot at the bottom is based on all 109 members of bacterial cluster F5B3. The wide taxonomic distribution within the cluster is reflected in the individual leader sequences, which are evidently very diverse. Throughout much of the alignment, any sequence identity is undetectable. However, the alignment is anchored near either end by two prominent sequence motifs which are present in most sequences despite their divergence

entire core length (Fig. 1B). This uniform conservation may just reflect that the leaders have not yet had time to diverge sufficiently in order for functionally important regions to stand out from their background. In contrast, the taxonomically diverse bacterial leader clusters have diverged to such an extent that sequence identity is undetectable throughout most of the sequence length (Fig. 3B). Instead, small motifs exist that are conserved in both sequence and position across diverse members of the same cluster. These motifs not only confirm a common origin for the leaders within that cluster, but also may be crucial for their function. Prominent sequence motifs are featured towards repeat-distal ends of core leaders for the major bacterial leader clusters F5B3, F2B5 and F2B5 (Supplementary Table *S1*). In contrast, the repeat proximal end is more divergent with numerous indels (Fig. 5A), showing little to no overall sequence conservation. Low sequence conservation towards the repeat proximal end in bacterial leaders, although common, is not the rule, as some leader clusters (e.g. bacterial cluster F1B12) do show the opposite pattern with a conserved proximal end and a divergent distal end (Supplementary Table *S1*).

3.5 Predicted core leaders coincide with published results and are generally longer in archaea than in bacteria

Using our *CRISPRleader* approach, we determined the conservation boundaries and respective leader length distributions for archaea and bacteria separately. The frequency of leader lengths peaked at about 60 and 130 bp in bacteria with a smaller peak at 190 bp, while in archaea, lengths were larger with peaks at 100, 220 and 290 bp (Fig. 4A) suggestive of some diversity of function. The leader boundaries obtained coincided closely with previously described CRISPR leaders for a few organisms in the literature (summarized in Table 2).

3.6 CRISPR loci are frequently leaderless

Individual observations of leaderless CRISPR loci have been reported that are defective in transcription and inactive in adaptation but it remains unclear whether they have lost their leaders or whether they have simply been separated from the leader distal ends of other CRISPR loci, possibly as a result of transposition events There are no data available on the extent of leaderless loci and,

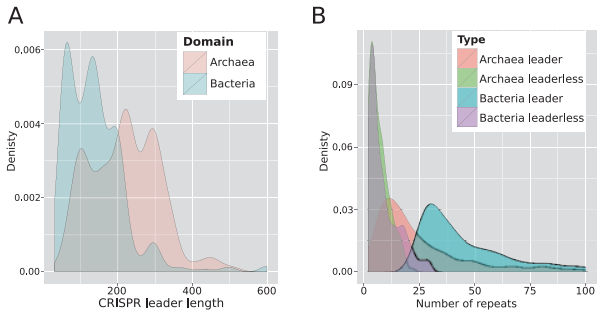


Fig. 4. (**A**) A comparison of the CRISPR leader length distributions between archaea and bacteria. It shows that archaea and bacteria are grouped into limited size ranges. The archaea peak leader sizes are larger with average values 100, 220 and 290 bp while the bacteria leader sizes are smaller with average values 60 and 160 bp. (**B**) The distribution of leader-containing and leaderless CRISPR loci in archaea and bacteria. The size distributions for leaderless CRISPR loci are similar for archaea and bacteria

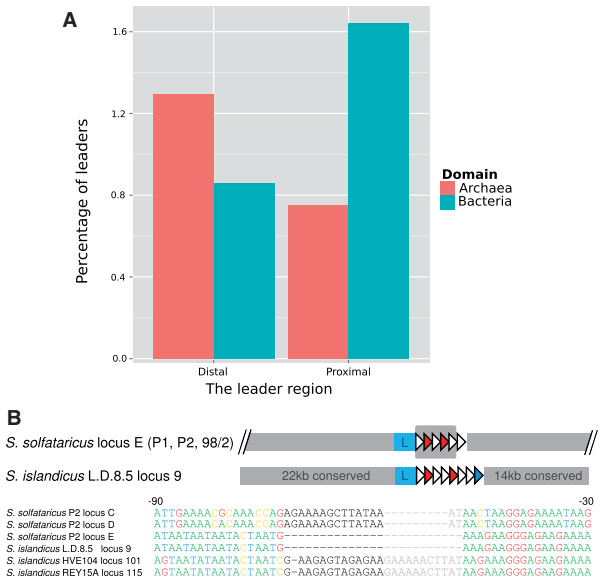


Fig. 5. (**A**) The distributions of insertions and deletions (indels) in the leader regions. Bacterial leaders more often carry indels towards the repeat proximal end, while archaeal leaders have them at the repeat distal end. (**B**) *Sulfolobus* leader deletions implicated in the adaptation phase. Part of an alignment between a series of Sulfolobus CRISPR leaders of cluster 2 is shown. *S. solfataricus* CRISPRs C and D acquire spacers during viral challenges, as does *S. islandicus* REY15A locus 115. *S.solfataricus* locus E is deficient in adaptation, acquiring spacers abnormally and at a very low rate, in turn making the CRISPR very small. A similar small locus is found in *S.islandicus* L.D.8.5. The leaders of both loci share a deletion around 50 bp from the first repeat, which is not found in the adaptation proficient leaders, consistent with a role in adaptation deficiency

therefore, we estimated the percentage of CRISPR loci that lack leaders and calculated the sizes of their arrays (number of spacer-repeat units) relative to those loci with conserved leaders. The results demonstrated that 13% of 980 archaeal CRISPR loci, and 24% of 2852 bacterial loci, were considered leaderless (Fig. 4B). Moreover, the sizes of the leaderless CRISPR arrays were much smaller on average (Fig. 4B). The smaller sizes are consistent with the leaderless loci being inactive in CRISPR adaptation and unable to increase in size but they are also consistent with them having separated from the ends of other CRISPR loci.

3.7 Leader clusters correlate more with Cas1 phylogeny than the subtype classification

Earlier studies have demonstrated that the sequences of leaders, repeats and Cas1 tend to coevolve for the type I–A CRISPR-Cas systems of the Sulfolobales (Shah and Garrett, 2011) and Thermoproteales (Garrett *et al.*, 2011). It was inferred that all these components were involved in spacer acquisition, whereas components of the interference effector complex evolved separately.

We quantified the degree of interdependence and coevolution of the leader clusters against Cas1 phylogeny and the cognate CRISPR subtype, respectively, by applying the Adjusted Rand Index (ARI) (Rand, 1971) that measures correlation between clusters. Leader clusters correlated with Cas1 clusters yielding an ARI value of 0.75, indicating a high degree of correlation. Conversely, the ARI between leader clusters and CRISPR subtypes was only 0.37. We infer that the lower correlation between the leader cluster and CRISPR subtype indicates that the same leader type can cofunction with CRISPR systems of different subtypes, and vice versa, as long as the correct adaptation module (i.e. Cas1, Cas2 and Cas4) is present to interact with the leader and maintain the CRISPR locus. This is consistent with the numerous reports of modular exchange where different adaptation and interference modules interchange to form new combinations of functional CRISPR-Cas systems (Garrett *et al.*, 2011; Makarova *et al.*, 2015; Vestergaard *et al.*, 2014). Since the latest CRISPR-subtype classification (Makarova *et al.*, 2015) primarily reflects the diversity of the interference modules, a lower correlation between CRISPR subtypes and leader clusters is to be expected.

Table 2. Comparison of predicted leader lengths against published leaders

Organism name	Published	Predicted	Difference
E.coli IYB5101 (Yosef *et al.*, 2012)	100	105	5
E.coli BL21-AI (Yosef *et al.*, 2012)	100	95	5
C.jejuni (Tasaki *et al.*, 2012)	146	144	2
Synechocystis pcc6803 (Scholz *et al.*, 2013)	125	116	9
S.pyogenes (Fonfara *et al.*, 2014)	109	108	1
S.solfataricus (Lillestol *et al.*, 2009)	238	237	1
M.marzei Göl (Nickel *et al.*, 2013)	108	108	0
M.marzei Göl (Nickel *et al.*, 2013)	108	111	3

3.8 Automated annotation of core leaders and CRISPR arrays using CRISPRleader

CRISPRleader accepts either a complete or partial genome sequence as input and provides a full annotation of the CRISPR array, their strand orientation as well as conserved core leader boundaries. In addition, it outputs reader-friendly HTML pages for conserved leader clusters from our database and it provides a standardized BED format that can be used to visualize CRISPR arrays and leader annotations in any genome browser.

4 Conclusion

Adaptation is currently the least understood of the main phases in the CRISPR-Cas immune system. Although it is known that

adaptation is affected by signals present in the region upstream of the CRISPR array, the so-called leader sequence, no bioinformatic tool exists that can automatically annotate these leader sequences to date. This is due to the fact that the known leader sequences exhibit only limited sequence conservation. To gain a deeper understanding, we developed a novel *k-mer*-based tool, *CRISPRleader*, that can reliably detect the CRISPR leader boundaries.

We analyzed 1426 archaeal and bacterial genomes using *CRISPRleader* and identified several characteristic properties of the leader sequences. Results show that although an extended region can be conserved between few very closely related species or CRISPR loci, generally a smaller core leader region, directly adjacent to the CRISPR locus, is conserved between more distantly related species.

We identified core leaders from 770 archaeal and 2224 bacterial CRISPR loci and observed significant differences between leader clusters. First, core leaders tend to be longer in archaea than in bacteria. Second, leader clusters in archaea are more homogeneous in terms of phyla than in bacteria. This may reflect the fact that archaea have survived primarily in low-energy environments which are often quite isolated (e.g. solfataric fields or hypersaline lakes) such that genetic exchange is much more limited than for most bacteria. Third, bacteria exhibit more indels in the CRISPR-proximal region of the core leaders than archaea. This core leader region has been shown to be important for CRISPR transcription and CRISPR-Cas adaptation and may be readily inactivated, or modulated, by indel activity, possibly triggered by an invader to circumvent targeting.

Regarding common characteristics, we showed that in both archaea and bacteria (i) leader sequences and repeats tend to coevolve with the Cas1 protein more broadly than previously believed, i.e. irrespectively of the system's subtype and (ii) leaderless CRISPR loci tend to be much smaller than loci with a leader present. This is possibly indicative of a displacement event from the leader-distal ends of other CRISPR loci. Leaderless CRISPR loci have been shown not to undergo adaptation but can still contribute to crRNA-directed interference.

Acknowledgements

The authors thank Anika Erxleben, Björn Grüning and Mummadi Chaithanya Kumar for their help.

Funding

This work was funded by the German Research Foundation (DFG) program FOR1680 'Unravelling the Prokaryotic Immune System' (grant BA 2168/5-1 to R.B.).

Conflict of Interest: none declared.

References

Alkhnbashi,O.S. *et al.* (2014) CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, **30**, i489–i496. In: *The proceedings of the 13th European Conference on Computational Biology (ECCB) 2014*.

Barrangou,R. and van der Oost, J., eds. (2013) *CRISPR-Cas Systems: RNA-Mediated Adaptive Immunity in Bacteria and Archaea*. Springer Press, Heidelburg, pp. 1–129.

Brouns,S.J.J. *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.

Costa,F. and Grave,K.D. (2010). Fast neighborhood subgraph pairwise distance kernel. In: *Proceedings of the 26 th International Conference on Machine Learning*. Omnipress, pp. 255–262.

Deng,L. *et al.* (2012) Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus. *Nucleic Acids Res.*, **40**, 2470–2480.

Diez-Villasenor,C. *et al.* (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I–E variants of *Escherichia coli*. *RNA Biol.*, **10**, 792–802.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Erdmann,S. and Garrett,R.A. (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol.*, **85**, 1044–1056.

Fonfara,I. *et al.* (2014) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.*, **42**, 2577–2590.

Frasconi,P. *et al.*, (2012). klog: a language for logical and relational learning with kernels. *arXiv preprint arXiv:1205.3981*.

Garrett,R.A. *et al.* (2011) Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol.*, **19**, 549–556.

Garrett,R.A. *et al.* (2015) CRISPR-Cas adaptive immune systems of the sulfolobales: unravelling their complexity and diversity. *Life (Basel)*, **5**, 783–817.

Gudbergsdottir,S. *et al.* (2011) Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol. Microbiol.*, **79**, 35–49.

Horvath,P. *et al.* (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int. J. Food Microbiol.*, **131**, 62–70.

Hsu,P.D. *et al.* (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Jansen,R. *et al.* (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Lange,S.J. *et al.* (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.*, **41**, 8034–8044. (SJL, OSA and DR contributed equally to this work.)

Leslie,C. *et al.* (2002) The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.*, **2002**, 564–575.

Leslie,C.S. *et al.* (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Li,Y. *et al.* (2016) Harnessing Type I and Type III CRISPR-Cas systems for genome editing. **44**, e34.

Lillestol,R.K. *et al.* (2006) A putative viral defence mechanism in archaeal cells. *Archaea*, **2**, 59–72.

Lillestol,R.K. *et al.* (2009) CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol. Microbiol.*, **72**, 259–272.

Makarova,K.S. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.

Marraffini,L.A. and Sontheimer,E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, **322**, 1843–1845.

Mojica,F.J. *et al.* (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.*, **36**, 244–246.

Mojica,F.J.M. and Garrett,R.A. (2013). Discovery and seminal developments in the CRISPR field. In: *CRISPR-Cas Systems*. Springer, Berlin, Heidelberg, pp. 1–31.

Mojica,F.J.M. *et al.* (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733–740.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. **48**, 443–453.

Nickel,L. *et al*. (2013) Two CRISPR-Cas systems in *Methanosarcina mazei* strain Go1 display common processing features despite belonging to different types I and III. *RNA Biol*., **10**, 779–791.

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc*., **66**, 846–850.

Rollie,C. *et al*. (2015) Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife*, **4**,

Scholz,I. *et al*. (2013) CRISPR-Cas systems in the *Cyanobacterium synechocystis* sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS One*, **8**, e56470. (IS and SJL contributed equally to this work.)

Shah,S.A. and Garrett,R.A. (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol*., **162**, 27–38.

Shah,S.A. *et al*. (2009) Distribution of CRISPR spacer matches in viruses and plasmids of *Crenarchaeal acidothermophiles* and implications for their inhibitory mechanism. *Biochem. Soc. Trans*., **37**, 23–28.

Tasaki,E. *et al*. (2012) Molecular identification and characterization of clustered regularly interspaced short palindromic repeats (CRISPRs) in a urease-positive thermophilic *Campylobacter* sp. (UPTC). *World J. Microbiol. Biotechnol*., **28**, 713–720.

Vestergaard,G. *et al*. (2014) CRISPR adaptive immune systems of Archaea. *RNA Biol*., **11**, 157–168.

Wurtzel,O. *et al*. (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res*., **20**, 133–141.

Yosef,I. *et al*. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res*., **40**, 5569–5576.

Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs

Viktoria Reimann, **Omer S. Alkhnbashi**, Sita J. Saunders, Ingeborg Scholz, Stephanie Hein, Rolf Backofen and Wolfgang R. Hess. **Nucleic Acids Research Journal, 2016**, doi: 10.1093/nar/gkw786.

Personal contribution

I have done a major contribution in this project, which led to share a first authorship with Viktoria Reimann. I performed the bioinformatics analysis of CRISPR-Cas system. Furthermore, I was involved in the interpretation of the results and writing the article.

Omer S. Alkhnbashi

The following co-authors confirm the above stated contribution.

Viktoria Reimann

Sita J. Saunders

Rolf Backofen

Wolfgang R. Hess

Nucleic Acids Research Advance Access published September 5, 2016

doi: 10.1093/nar/gkw786

Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs

Viktoria Reimann[1,†], Omer S. Alkhnbashi[2,†], Sita J. Saunders[2], Ingeborg Scholz[1], Stephanie Hein[1], Rolf Backofen[2,3,4,*] and Wolfgang R. Hess[1,3,*]

[1]Genetics and Experimental Bioinformatics group, Faculty of Biology, University of Freiburg, Schänzlestrasse 1, 79104 Freiburg, Germany, [2]Bioinformatics group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany, [3]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Habsburgerstrasse 49, D-79104 Freiburg, Germany and [4]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schänzlestrasse 18, D-79104 Freiburg, Germany

Received July 11, 2016; Revised August 24, 2016; Accepted August 26, 2016

ABSTRACT

A hallmark of defense mechanisms based on clustered regularly interspaced short palindromic repeats (CRISPR) and associated sequences (Cas) are the crRNAs that guide these complexes in the destruction of invading DNA or RNA. Three separate CRISPR-Cas systems exist in the cyanobacterium *Synechocystis* sp. PCC 6803. Based on genetic and transcriptomic evidence, two associated endoribonucleases, Cas6-1 and Cas6-2a, were postulated to be involved in crRNA maturation from CRISPR1 or CRISPR2, respectively. Here, we report a promiscuity of both enzymes to process *in vitro* not only their cognate transcripts, but also the respective non-cognate precursors, whereas they are specific *in vivo*. Moreover, while most of the repeats serving as substrates were cleaved *in vitro*, some were not. RNA structure predictions suggested that the context sequence surrounding a repeat can interfere with its stable folding. Indeed, structure accuracy calculations of the hairpin motifs within the repeat sequences explained the majority of analyzed cleavage reactions, making this a good measure for predicting successful cleavage events. We conclude that the cleavage of CRISPR1 and CRISPR2 repeat instances requires a stable formation of the characteristic hairpin motif, which is similar between the two types of repeats. The influence of surrounding sequences might partially explain variations in crRNA abundances and should be considered when designing artificial CRISPR arrays.

INTRODUCTION

Roughly 30% of bacterial and 70% of archaeal publicly available genomes encode clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins (1,2). These CRISPR-Cas systems provide an adaptable and inheritable immune system against viruses and other foreign genetic elements (3,4). Most CRISPR-Cas loci consist of an array of alternating identical repeat and varying spacer sequences and a set of genes encoding Cas proteins (1), which have been categorized into two major classes, five major types I–V and into at least 16 subtypes (5). Although there is a complex relationship between repeats and types, a strong correlation between Cas1 and structural motifs and sequence families was found and 4719 repeats were clustered into 18 structural motifs and 24 sequence families (2,6).

The repeat-spacer array gives rise to a long precursor transcript named pre-crRNA (7,8) that is, in Type I and III systems, processed by an endoribonuclease into intermediate crRNAs (70–80 nt) and, in some Type I and III systems, in a second ribonucleolytic step into the mature crRNAs (40–50 nt) (8–11). The known primary endoribonucleases in subtype I-A, I-B, I-E, I-F and Type III systems belong to the Cas6 family of proteins, whereas the enzymatic activity for the second ribonucleolytic step is unknown (8,12). In contrast, Cas5, which possesses a structural role in the interference complex of most other CRISPR subtypes, serves as the dedicated endoribonuclease in subtype I-C systems (13,14).

It is only poorly understood how different Cas6 endoribonucleases, present in organisms with multiple CRISPR systems, differentiate between their targets (15). Cai *et al*. reported that about 70% of available cyanobacterial genomes possess various types of CRISPR-Cas systems (16). Three separate CRISPR-Cas systems exist in the cyanobacterium

*To whom correspondence should be addressed. Tel: +49 761 203 2796; Fax: +49 761 203 2745; Email: wolfgang.hess@biologie.uni-freiburg.de
Correspondence may also be addressed to Rolf Backofen. Tel: +49 761 203 7461; Fax: +49 761 203 746; Email: backofen@informatik.uni-freiburg.de
†These authors contributed equally to the paper as first authors.

Synechocystis sp. PCC 6803, which were named CRISPR1, CRISPR2 and CRISPR3. From these, CRISPR2 and 3 are subtype III-D and III-B systems, based on the presence of a Cas10 protein (17) and the most recent classification system (5). CRISPR1 was classified as a subtype I-D CRISPR-Cas system, characterized by the presence of the type-specific protein Cas3 and the subtype-specific protein Cas10d (18); I-D systems are currently poorly described in the literature. All three CRISPR-Cas-systems (CRISPR1–3) are encoded on the ∼100 kb plasmid pSYSA (17), which in addition encodes at least nine distinct toxin–antitoxin systems, characterizing it as the major defense plasmid of this organism (19,20).

Three of the *Synechocystis* sp. PCC 6803 *cas* genes encode enzymes of the Cas6 endoribonuclease family: *slr7014, slr7068, sll7075* (17). These were named *cas6-1*, *cas6-2a* and *cas6-2b* according to their location upstream of the CRISPR1 and CRISPR2 arrays, respectively. Deleting *cas6-1* abolished the accumulation of CRISPR1 pre-crRNA, processing intermediates and mature crRNAs; whereas in the Δ*cas6-2a* mutant CRISPR2 transcripts >200 nt overaccumulated, but shorter intermediates and mature crRNAs were lacking (17). These phenotypes are consistent with a function of Cas6-1 and Cas6-2a as endoribonucleases. Cas6-1 and Cas6-2a are only 16.5% identical at the amino acid level (Supplementary Figure S1). Although closely related proteins exist in other cyanobacteria, the relation of Cas6-1 and Cas6-2a to biochemically characterized RNA endonucleases is vague and requires genetic and biochemical analysis.

The *Synechocystis* sp. PCC 6803 CRISPR1 and CRISPR2 hairpins are structurally similar and the last 11 nt of the repeat sequences are identical (Figure 1D). This is relevant because many Cas6 proteins cleave within CRISPR repeats that form hairpin structures: e.g. Cas6e of *Escherichia coli* (7,18), MmCas6b of *Methanococcus maripaludis* (21), Cas6f of *Pseudomonas aeruginosa* (22,23) and several enzymes from *Sulfolobus solfataricus* (24–26). In other cases, Cas6 proteins also bind an unstructured repeat sequence, e.g. PfCas6 of *Pyrococcus furiosus* (27).

Analysis of *Synechocystis* sp. PCC 6803 RNA-seq data led to the identification of a putative cleavage site within the repeat sequences of CRISPR1 and 2 (17). The cleavage at this site generates in both cases intermediate crRNAs with a length of 68–83 nt consisting of a single spacer sequence as well as an 8-nt-repeat handle at the 5′ end and a 29 nt repeat fragment at the 3′ end. However, *in vivo* data from northern hybridizations and RNA-seq showed the mature crRNAs to be shorter. They are 39 and 45 nt in case of CRISPR1 and 36 or 37 nt for CRISPR2 (17). Thus, in a second, so far uncharacterized step the crRNA intermediates are processed further into the mature crRNAs. Such a further processing by an unknown trimming nuclease that removes 3′ portions of the crRNA is also known from several Type III and at least one subtype I-A system (9,28–30).

Here, we demonstrate biochemically that Cas6-1, encoded within a cassette of subtype I-D *cas* genes in *Synechocystis* sp. PCC 6803, is the endoribonuclease that generates the 8-nt-repeat handle of CRISPR1 mature crRNAs and that Cas6-2a, encoded within a different cassette of *cas* gene (belonging to subtype III-D), is the endoribonuclease that processes the crRNAs of CRISPR2. We detected a promiscuity of both enzymes to process not only their cognate CRISPR1 or CRISPR2 transcripts, but to also cleave the transcripts from the other locus. This promiscuity is in striking contrast to the *in vivo* specificity of these enzymes found in the analysis of deletion mutants (17).

Moreover, cleavage of the non-cognate substrates was less efficient and not all possible cleavage sites were recognized. Bioinformatics analysis of a series of *in vitro* experiments suggested the successful cleavage to depend on the stable formation of a hairpin motif, which is similar between CRISPR1 and CRISPR2. However, the sequences of adjacent spacers can lead to alternative structures that inhibit stable folding of the hairpin motif and thus are incompatible with the cleavage reaction. The influence of surrounding sequences might partially explain variations in crRNA abundances *in vivo* and should be considered when designing artificial CRISPR arrays.

MATERIALS AND METHODS

Cloning, expression and purification of cyanobacterial Cas6 endonucleases

The genes *slr7014* and *slr7068* that encode Cas6-1 and Cas6-2a (17) were amplified by polymerase chain reaction (PCR) using primers containing BamHI and SphI or PstI and SacI restriction sites (Supplementary Table S1) and 10 ng of *Synechocystis* sp. PCC 6803 genomic DNA. PCR fragments were subcloned in *E. coli* DH5α after ligation into vector pJET1.2/blunt (CloneJET PCR Cloning Kit, Thermo Fisher Scientific). The *slr7014*-containing BamHI/SphI restriction fragment was isolated and ligated into the corresponding sites of vector pQE70 (QIAGEN) and transformed into *E. coli* M15[pREP4], whereas the *slr7068*–containing PstI/SacI fragment was recloned into vector pASK-IBA7plus (IBA-Solutions for Life Sciences) and transformed into *E. coli* Rosetta(DE3)pLysS. In this way, the reading frame of Cas6-1 was prolonged by six additional histidine residues at the C-terminus (His_6-tag), whereas Cas6-2a is featured with an N-terminal *Strep*-Tactin® affinity tag (*Strep*-tag® II). The cloned fragments were verified by DNA sequencing (GATC Biotech).

Escherichia coli M15[pREP4]/pQE70::*slr7014* was grown in LB medium (31) in a culture volume of 400 ml (100 μg/ml ampicillin, 50 μg/ml kanamycin) at 37°C to an OD_{600nm} of ∼0.8. Protein expression was induced by the addition of 1 mM IPTG at 30°C for 3 h. Cells were pelleted by centrifugation at 6500 *g* and 4°C for 15 min and frozen at −20°C, or immediately resuspended in 5 ml of lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole; pH 8) in the presence of protease inhibitor (cOmplete Protease Inhibitor Cocktail Tablets, Roche). Cells were disrupted by sonication (Sonifier 250, Branson) and debris was removed by centrifugation at 11000 *g* and 4°C for 30 min.

Expression of Cas6-2a was induced in *E. coli* Rosetta(DE3)pLysS/pASK-IBA7plus::*slr7068* in a culture volume of 400 ml (100 μg/ml ampicillin, 34 μg/ml chloramphenicol) at an OD_{550nm} of ∼0.6 by the addition of 200 ng/ml anhydrotetracycline. Cultures were grown at 22°C with 180 rpm overnight. Cells were pelleted as for Cas6-1 but then resuspended in 4 ml of buffer W (100 mM

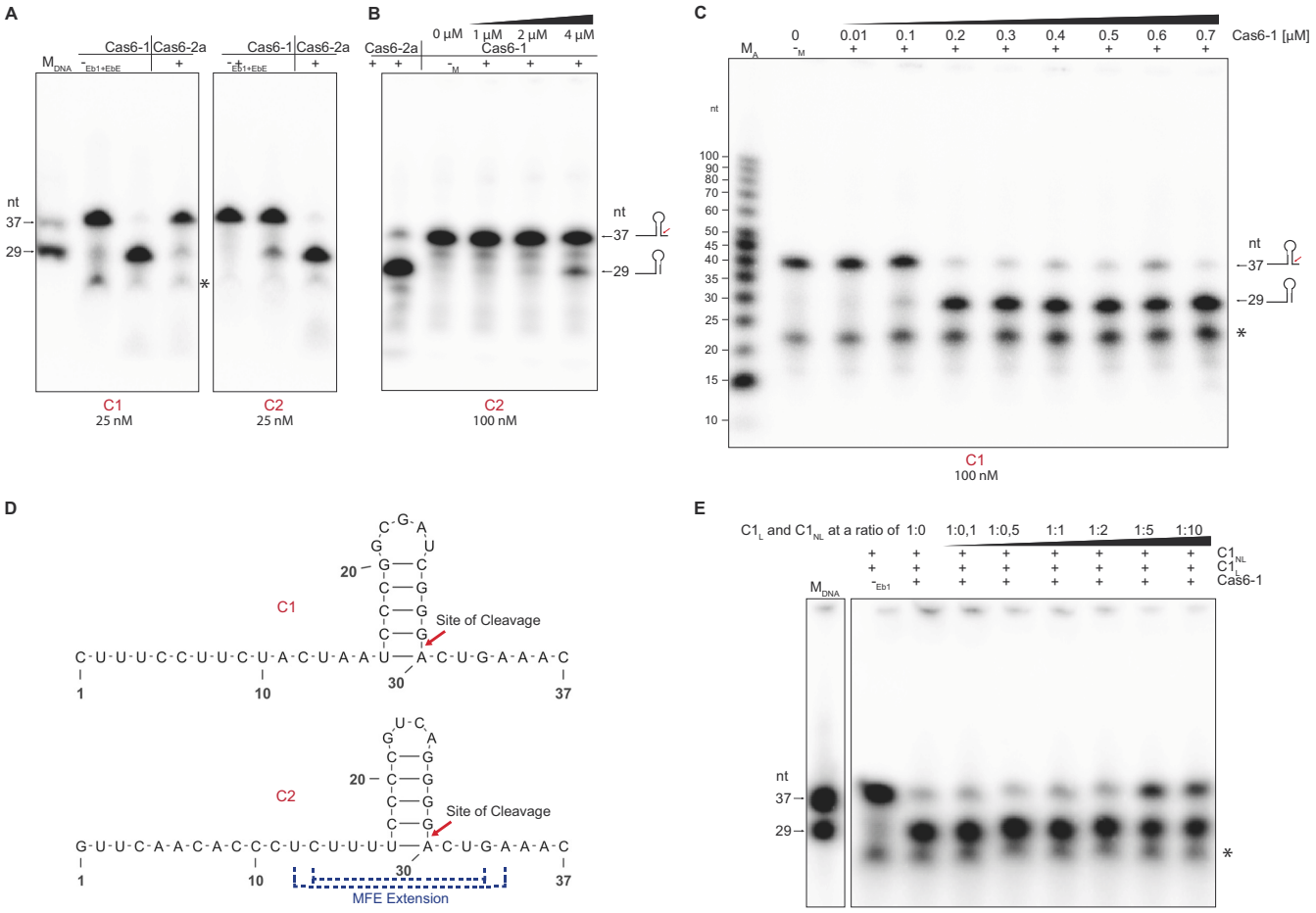


Figure 1. Cas6-1- and Cas6-2a-mediated cleavage of synthetic CRISPR repeat fragments. (**A**) A total of 25 nM of synthetic 5′ ^{32}P labeled CRISPR1 (C1) or CRISPR2 (C2) oligoribonucleotides were incubated in the presence (+) of enzyme (1 µM Cas6-1 or 2 µl of Cas6-2a elution fraction 3) in reaction buffer A for 1 h. (**B**) A total of 100 nM of synthetic 5′ ^{32}P labeled C2 RNA was incubated for 1 h with 1–4 µM Cas6-1. Incubation with Cas6-2a (2 µl of Cas6 2a elution fraction 3) served as a positive control using cleavage buffer B. (**C**) A total of 100 nM of synthetic 5′ ^{32}P labeled C1 RNA was incubated for 1 h with increasing concentrations of Cas6-1 in cleavage buffer B. (**D**) Predicted secondary structures of the *Synechocystis* sp. PCC 6803 CRISPR1 and CRISPR2 repeat RNAs (C1 and C2). The determined cleavage site (17) of the processing endoribonucleases within the CRISPR1 or CRISPR2 repeats is located 8 nt upstream of the 3′ repeat end, indicated by an arrow. The RNA secondary structures were predicted with the RNAfold web server (34) and drawn using VARNA (40). (**E**) Titration of 5′ labeled C1 RNA cleavage by Cas6-1 through the addition of unlabeled C1 RNA ($C1_{NL}$). To show that cleavage of 5′ labeled C1 RNA ($C1_L$) by Cas6-1 is inhibited by addition of unlabeled C1 RNA, 0.025 pmol of $C1_L$ and increasing amounts of $C1_{NL}$ (0.0025–0.25 pmol) were incubated with 500 nM Cas6-1 in reaction buffer A for 15 min. All reactions were separated by denaturing 8 M urea 15% polyacrylamide gel electrophoresis (PAGE) and bands were visualized by autoradiography. Three different negative controls were performed, by adding elution buffer 1 (Eb1), elution buffer E (EbE) or an aliquot from a mock purification from *Escherichia coli* cells containing the respective vector without an inserted gene (M). A byproduct of C1 oligonucleotide synthesis is labeled by the asterisk (*) in panels A, C and E. M_A: Low Molecular Weight DNA Marker (Affymetrix); M_{DNA}: radiolabeled oligonucleotides of the respective sizes serving as size markers.

Tris–HCL, pH 8, 150 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA), pH 8). Cells were disrupted in 7 ml tubes with 250 µl glass beads (Ø 0.5 mm) with a tissue homogenizer (Precellys24 with Cryolys-N_2-cooling, 6* 10 s, 6500 rpm with 5 s of break between each interval, Bertin Technologies). Debris was removed by centrifugation at 13000 *g* and 4°C for 15 min.

For the purification of Cas6-1, Ni^{2+}-NTA-agarose (QIAGEN) and chromatography columns (Poly-Prep®, BIO-RAD) were used. The protein purification was performed under native conditions as recommended by the the manufacturer (The QIAexpressionist, QIAGEN). A bed volume of 300 µl was used. The wash buffer contained 40 mM imidazole. To elute the bound proteins, elution buffers 1 and 2 (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM or 500 mM imidazole, pH 8) were used consecutively. The elution fractions 1 and 2 of purified Cas6-1 protein were 10-fold concentrated (Supplementary Figure S2A) using Amicon Ultra-0.5 or Ultra-2 Centrifugal Filter Units with Ultracel-10 membrane (Merck Millipore). Thereby the buffer was exchanged to PBS (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4; pH 7.3). The protein concentrations were determined using the Direct Detect® Spectrometer (Merck Millipore).

For the purification of Cas6-2a, *Strep*-Tactin® Sepharose (IBA-Solutions for Life Sciences) and chromatography columns (Poly-Prep®, BIO-RAD) were used. The purification was performed under native conditions according

to the manufacturer's recommendations (IBA-Solutions for Life Sciences) using a bed volume of 800 μl. The column was washed 5 times with 4 ml of buffer W and bound protein was eluted with 6* 400 μl of elution buffer E (100 mM Tris–HCl, pH 8, 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin). Purified proteins were analyzed via sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) (6% polyacrylamide (PAA) stacking gel, 15% PAA separating gel), visualized by GelCode Blue Safe Protein staining (Thermo Fisher Scientific) for 1 h, aliquoted and stored at −80°C until use. The purified Cas6-2a protein was further verified by western blot detection (Supplementary Figure S2B) using the *Strep*-Tactin® HRP conjugate according to the manufacturer's instructions (IBA-Solutions for Life Sciences) and detecting the chemiluminescence signal with the FUSION SL™ imaging system (peQlab). Additionally, the vectors and tags pQE30, pQE70 (His6-tag, from QIAGEN) and pET28a(+) (His6-tag, Merck Millipore), pGEX-6P-1 (GST-tag, GE Healthcare Life Sciences), pASK-IBA6 and pASK-IBA43plus (*Strep*-tag® II, IBA-Solutions for Life Sciences) were tested for the purification of Cas6-2a. As additional negative controls, all purification steps were repeated for both proteins with an empty vector *E. coli* control strain.

Generation of radiolabeled synthetic RNA oligonucleotides

Synthetic oligoribonucleotides C1 and C2 (SIGMA-ALDRICH®) correspond to the CRISPR1 and CRISPR2 repeat RNAs (Supplementary Table S1). A total of 12.5 pmol of each oligoribonucleotide were used for 5′ end-labeling with ^{32}P using 50 μCi of [γ^{32}P] ATP (3000 Ci/mmol, Hartmann Analytic) and 25 U of T4 polynucleotide kinase (Thermo Fisher Scientific) in a reaction volume of 50 μl. As size markers, DNA oligonucleotides of the respective sizes (M_{DNA}) or the Low Molecular Weight Marker (M_A, Affymetrix), that is also composed of DNA, were analogously 5′ end-labeled. Unbound nucleotides were removed with RNA Clean & Concentrator-5 (Zymo Research) and labeled RNA was eluted in 50 μl of nuclease free water.

RNase cleavage assays with synthetic RNA oligonucleotides

Cleavage reactions were performed in a volume of 10 μl in cleavage buffer A (20 mM HEPES-KOH, pH 8, 250 mM KCl, 1 mM DTT, 2 mM $MgCl_2$) or cleavage buffer B (20 mM Tris–HCL, pH 7.8, 400 mM KCl), with 25–100 nM 5′ labeled RNA and 0.01–4 μM Cas6-1 or, when indicated, with 2 μl of elution fraction 3 of the Cas6-2a purification. As negative controls served elution buffer 1 (Eb1, for Cas6-1 experiments), elution buffer E (EbE, Cas6-2a experiments) or analogous purifications with cells harboring the respective plasmid without an inserted gene (empty vector (mock) controls, MC). Reactions were incubated for 15–60 min at 37°C, stopped by the addition of 2× RNA loading dye (95% formamide, 0.025% SDS, 0.025% bromophenol blue, 0.025% xylene cyanol FF, 0.5 mM EDTA, pH 8) and stored on ice. Before loading onto denaturing 8 M urea 15% PAA gels, the reactions were incubated at 95°C for 5 min and cooled down on ice. Gels were exposed to a phosphor imaging screen (BIO-RAD) and radiolabeled RNA was visualized by phosphor imaging (Molecular Imager PharosFX™ Plus, BIO-RAD) and analyzed using Quantity One® software (BIO-RAD).

In vitro transcription and purification

In vitro transcription of CRISPR1, 2 and 3 was performed with the MEGAshortscript T7 transcription kit (Ambion®, Thermo Fisher Scientific). Suitable templates were PCR-amplified and thereby tagged with a T7 promoter as part of the primer sequences (Supplementary Table S1). The amplified and purified (NucleoSpin® Gel and PCR Clean-up, MACHEREY-NAGEL) fragments were used for *in vitro* transcription according to the manufacturer's specifications. All *in vitro* transcribed RNAs carry two additional guanidine nucleotides at their 5′ ends originating from the T7 promoter. The *in vitro* transcripts were used directly (products CRISPR1, CRISPR1*, CRISPR2 and CRISPR3; used for experiments shown in Figure 2) or after gel purification (products CRISPR1 I–IX and CRISPR2 I–IX; used for experiments shown in Figure 3). In the latter case, transcripts were size-fractionated by denaturing 8 M urea 10% PAGE, visualized with ethidium bromide under ultraviolet (UV) light and excised at the appropriate size. Transcripts were eluted for 18–24 h at 37°C by adding 300 μl of transcript elution buffer (20 mM Tris–HCL, pH 7.5, 250 mM sodium acetate, pH 5.2, 1 mM EDTA, pH 8). Afterward, the elution buffer containing the eluted *in vitro* transcript was transferred to a fresh reaction tube and the transcript was precipitated for 18–72 h at −20°C by adding two volumes of ethanol (99.8%). The RNA was pelleted at 11 000 *g* and 4°C for 30 min, washed once with 100 μl of ethanol (70%), pelleted again for 5 min and resuspended in 50 μl of nuclease free water.

RNase cleavage assays with *in vitro* transcripts

Cleavage assays of Cas6-1 were performed in cleavage buffer B (experiments shown in Figure 3A and B) or by adding Cas6-1 directly to the *in vitro* transcript (experiments shown in Figure 2) at 37°C for 0.5 h if not specified otherwise. Cleavage assays of Cas6-2a were performed in cleavage buffer C (10 mM HEPES-KOH, pH 8.0, 125 mM KCl, 0.5 mM DTT, 0.94 mM $MgCl_2$) with 1 or 2 μl of Cas6-2a elution fraction 3 at 37°C for 0.5 h. To stop the reactions, 2× RNA loading dye was added. Before loading onto denaturing 8 M urea 10% PAA gels, reactions were incubated at 95°C for 5 min. As size markers served the Low Range ssRNA Ladder from NEB (M_N), the RiboRuler Low Range RNA Ladder from Thermo Fisher Scientific (M_F) and the Low Molecular Weight Marker from Affymetrix (M_A). RNA was visualized after ethidium bromide staining under UV light (254 nm) in a gel documentation system (E-Box-3026, peQlab). For size determination of RNA fragments generated in cleavage assays with Cas6-1 or Cas6-2a separation of fragments was performed with a sequencing gel electrophoresis apparatus (Model S2, Biometra). The denaturing 8.3 M urea 10% PAA gel with a size of 31 × 38 cm was prerun at constant power (65 W) for 1 h and with a surface temperature of 42–46°C. After sample loading the

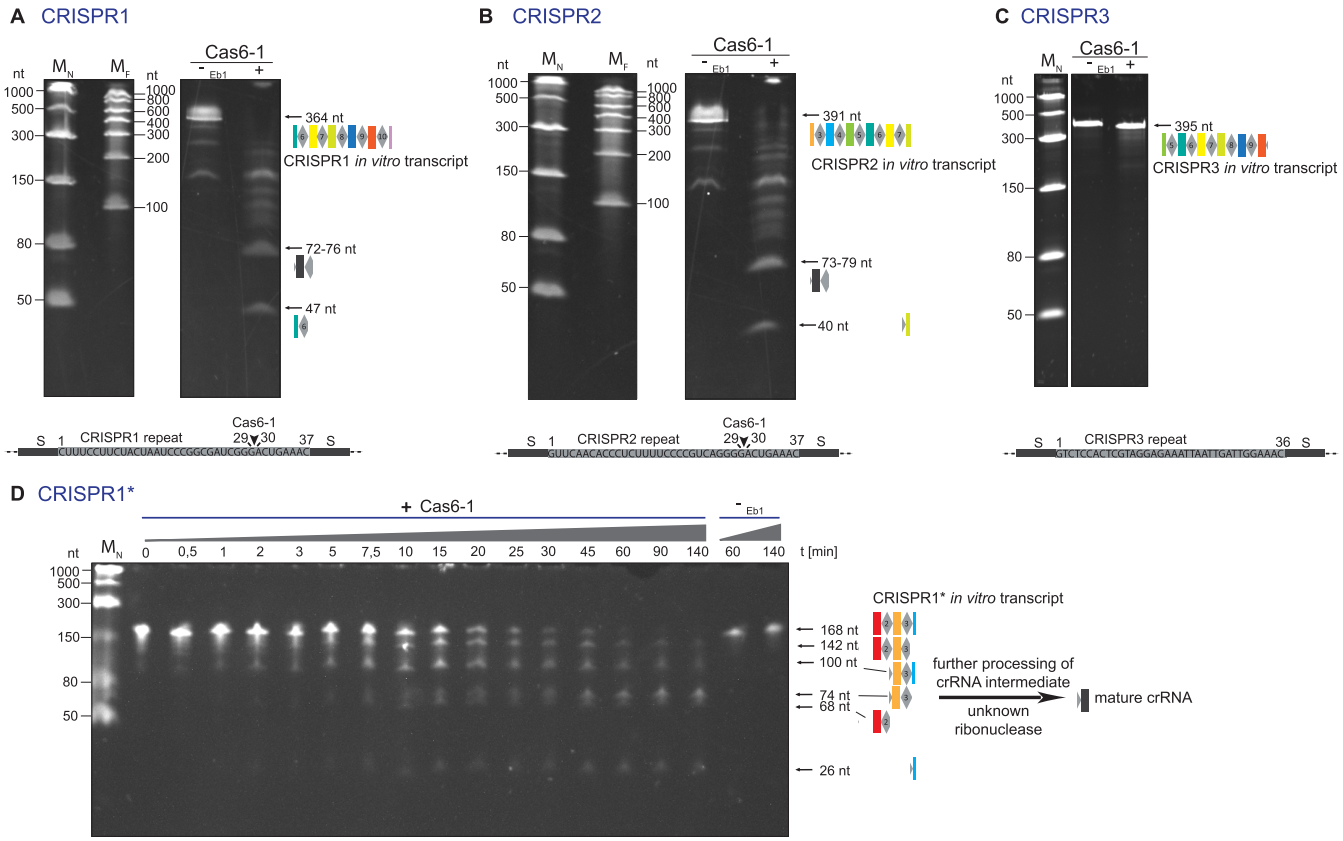


Figure 2. Incubation of *in vitro* transcripts of CRISPR1, 2 and 3 with Cas6-1. (**A**) A total of 2.3 μM of CRISPR1 transcript is cleaved by 6.2 μM Cas6-1 by incubation for 1 h. (**B**) A total of 2 μM of CRISPR2 transcript is cleaved by 6.2 μM Cas6-1 by incubation for 1 h. (**C**) A total of 3.2 μM of CRISPR3 transcript is not cleaved by 6.5 μM Cas6-1 by incubation for 2 h. (**D**) The cleavage of 6.6 μM of the shorter CRISPR1* *in vitro* transcript by 5.8 μM Cas6-1 monitored over a time of 2 h 20 min. Since no finally mature crRNA (17) was detected, the *in vivo* presence of an unknown ribonuclease is suggested. All reactions were performed at 37°C in a reaction volume of 5 μl. RNA was separated by 8 M urea 10% PAGE and bands were visualized by ethidium bromide staining. Diamonds represent repeats and rectangles spacer sequences. Transcripts were not gel purified, explaining the appearance of fragments smaller than the full length transcripts of CRISPR1 and 2 in absence of Cas6-1. Eb1, elution buffer 1 used as negative control. Molecular markers: MN, Low Range ssRNA Ladder (NEB); MF, RiboRuler Low Range RNA Ladder (Thermo Fisher Scientific).

gel was run for additional 4.5 h at 65 W. An alkaline hydrolysis ladder was produced by incubation of 20 pmol of a 358 nt *in vitro* transcript in a buffer containing 50 mM Tris–HCl, pH 8.5 and 20 mM $MgCl_2$ for 48 h at 30°C.

For staining of sequencing gel-separated RNA, SYBR® Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific) was used in a 1:10000 dilution with 0.5 × Tris-Borate-EDTA (TBE) buffer. The image was taken with the Laser Scanner Typhoon FLA 9500 (GE Healthcare Life Sciences) with the following settings: excitation: 473 nm, emission filter long pass blue ≥ 510 nm, photomultiplier value: 450 or 500.

Predicting stabilities of CRISPR hairpin motifs within their natural context

The functional consensus structure motifs for CRISPR1 and CRISPR2, as shown in Figure 1D, were taken from reference (17), where local sequence context was considered and thus the repeat structure that is most stable across the entire CRISPR array was determined. By this definition, the consensus motif is a *local* structure that consists of the base pairs defined in the consensus motif. We estimate the quality of formation of this local functional repeat structure in a specific fragment by determining the accuracy of this structure as previously defined (32). The *accuracy* of a local structure consisting of a set of base pairs $S^{loc} = \{(i_1, j_1), \ldots, (i_k, j_k)\}$ in an RNA-sequence R is defined as the expected overlap of the local structure S^{loc} with all possible global structures S of the sequence R:

$$Acc\left(S^{loc}, R\right) = \sum_{S \text{ structure } of \ R} \left| S^{loc} \cap^{S} \right| * \Pr(S|R)$$

where $\Pr(S|R)$ is the Boltzmann probability of the global structure S in the ensemble of all structures of R. Since this would require a summation over an exponential number of structures, this cannot be directly calculated this way. However, as shown in reference (32), this quantity is equivalent to:

$$Acc(S^{loc}, R) = \sum_{(i,j) \in S^{loc}} \Pr((i,j)|R)$$

which can easily be calculated. Here, $\Pr((i, j)|R)$ is the base pair probability of the base pair (i, j) in the sequence R as determined by the McCaskill approach, as e.g. implemented in the Vienna RNA package with RNAfold –p.

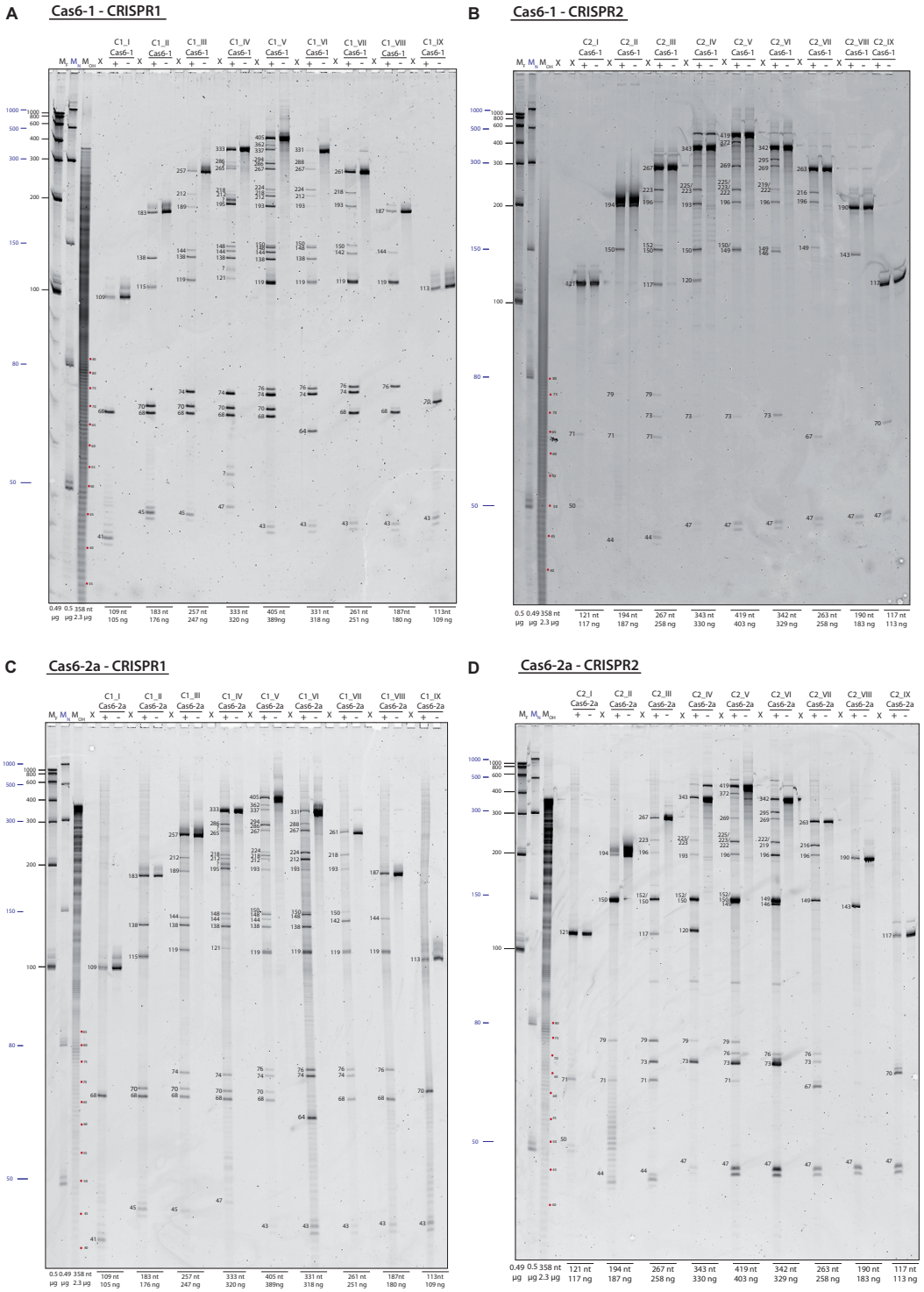


Figure 3. Cleavage of CRISPR1 or CRISPR2 *in vitro* transcripts by Cas6-1 or Cas6-2a. (**A** and **B**) Processing of CRISPR1 or CRISPR2 transcripts I–IX by Cas6-1. *In vitro* transcripts were incubated with 750 nM Cas6-1 in reaction buffer B. (**C** and **D**) Processing of CRISPR1 or CRISPR2 transcripts I-IX by Cas6-2a. *In vitro* transcripts of CRISPR1 were incubated in buffer C with 2 μl of Cas6-2a elution fraction 3. For the reactions shown in (**D**) only 1 μl of Cas6-2a elution fraction 3 was used. All *in vitro* transcripts were gel purified and used in a final concentration of 300 nM. The reactions were performed in a reaction volume of 10 μl in the presence or absence of enzyme at 37°C for 30 min. Reactions were separated on denaturing 8.3 M urea 10% PAA sequencing gels and bands were visualized with SYBR® Gold Nucleic Acid Gel Stain. The sizes of cleavage products are indicated in nucleotides next to the respective fragments. M_N, M_F, M_A: Molecular Weight Markers as in Figures 1 and 2. M_{OH}: alkaline hydrolysis ladder; X: empty lanes.

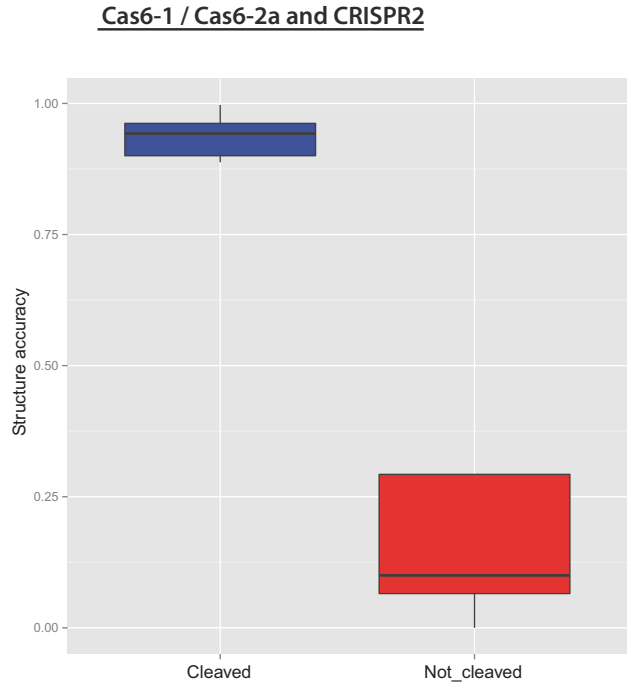


Figure 4. The structure stability of the CRISPR2 hairpin, measured as the base pair accuracy (y-axis), is compared between repeat instances that were cleaved (left) and not cleaved (right) by Cas6-1 and Cas6-2a in the *in vitro* experiments in Figure 3B and D. High base pair accuracies correspond to successful cleavage events, whereas low base pair accuracies explain repeats that were not cleaved. For both enzymes only 3 out of 25 experimental observations were not explained by the base pair accuracy (Supplementary Table S3).

In the case of longer fragments, the problem of long-range base pairs occurs. It is well known that predictions of these long-range base pairs are especially unreliable and noisy. To minimize this effect, and to also account for possible other effects like co-transcriptional folding or intermediate processing, we followed a local folding approach for determining base pair probabilities (see reference (32) for a discussion of various local folding approaches). The idea is to calculate base pair probabilities as usual as the sum of probabilities of structures that contain this base pair, but to restrict the set of possible structures to local structures by restricting the maximal span of a base pair. This implies, that in all possible structure considered in this calculation, the distance between the left and right end of any base pair is restricted. In our case, we used 80 nt as maximal span of a base pair. Technically, this is achieved by using RNAplfold (33) from Vienna package 1.8.4, where we set the window size (W) equal to the fragment size and the maximum base-pair span (L) equal to 80 nt. In addition, we used the option –noLP to disallow lonely base pairs, which usually improves the prediction quality. Dot plots were calculated for the repeat structure by taking the average of the sub-matrices for each repeat instance, where the base-pair probability matrix is computed for each window separately and then averaged over all windows using RNAfold (34), Vienna package version 1.8.4, with parameters ' -p -d2 –noLP'. The RNA secondary structures were drawn using VARNA (40).

RESULTS

Cas6-1 mediated cleavage of synthetic oligoribonucleotides

We cloned, expressed and purified recombinant *Synechocystis* sp. PCC 6803 Cas6-1 and Cas6-2a as soluble proteins (Supplementary Figure S2). Recombinant Cas6-1 and Cas6-2a cleaved their cognate synthetic repeat oligoribonucleotides C1 or C2 completely, whereas both enzymes cleaved their non-cognate repeats (C2 for Cas6-1 and C1 for Cas6-2a) only weakly (Figure 1A–C). Both enzymes cleave their respective targets at a single position, resulting in an 8 nt shorter 5′ ^{32}P labeled RNA product (29 nt). For both tested repeats, this product is consistent with the cleavage between repeat positions G29 and A30 (Figure 1D) that was determined by RNA-seq analysis (17). We verified 5′ ^{32}P labeled C1 RNA ($C1_L$) as a substrate of Cas6-1 by titrating the cleavage reaction through addition of increasing amounts of unlabeled substrate $C1_{NL}$. The addition of a 5- to 10-fold excess of $C1_{NL}$ over $C1_L$ caused a decrease of $C1_L$ cleavage by Cas6-1 since the protein likely reached a limit of saturation with substrate (Figure 1E).

Promiscuity in the cleavage of CRISPR precursor transcripts by Cas6-1 and Cas6-2a

Since a repeat sequence does not exist on its own but is part of a pre-crRNA transcript, the endoribonuclease activity of Cas6-1 on longer precursors was studied. We incubated precursors containing multiple repeat-spacer units of CRISPR1–3 with the purified protein *in vitro* and analyzed cleavage products by gel electrophoresis (Figure 2). Transcripts of CRISPR1 (364 nt and a shorter version CRISPR1* of 168 nt) were cleaved to products of the expected sizes (Figure 2A and D; Supplementary Table S2). Surprisingly, in this assay the *in vitro* transcript of CRISPR2 was cleaved by Cas6-1 with the similar efficiency as the CRISPR1 transcript (Figure 2A and B). In contrast, the CRISPR3 transcript was not cleaved by Cas6-1 (Figure 2C), consistent with the results of genetic analyses (17).

In the following, we characterized the ectopic Cas6-1-mediated processing of CRISPR2 transcripts by systematic substrate variation. In addition, we tested if Cas6-2a could possibly mediate processing of CRISPR1 transcripts as well. Each RNA fragment represented a subsequence of the original CRISPR1 or CRISPR2 array with a different number of repeats and spacers. The CRISPR1 and 2 fragments I–IX were incubated *in vitro* in the presence or absence of Cas6-1 or Cas6-2a and resulting cleavage fragments were analyzed by denaturing gel electrophoresis (Figure 3). Detected fragment sizes were consistent with the expected lengths when assuming a cleavage 8 nt upstream of the 3′ end of each repeat instance in the CRISPR1 or 2 fragments (Supplementary Table S2). Both enzymes delivered very similar patterns for the respective substrates, suggesting that the identical sites were recognized and cleaved. However, we noticed for both enzymes that for CRISPR1 all but for CRISPR2 not all theoretically possible fragments (Supplementary Table S2) were observed, consistent with the idea that they could generate some but not all of the theoretically possible products. The presence of potential contaminating RNase activities in the preparations is considered very low

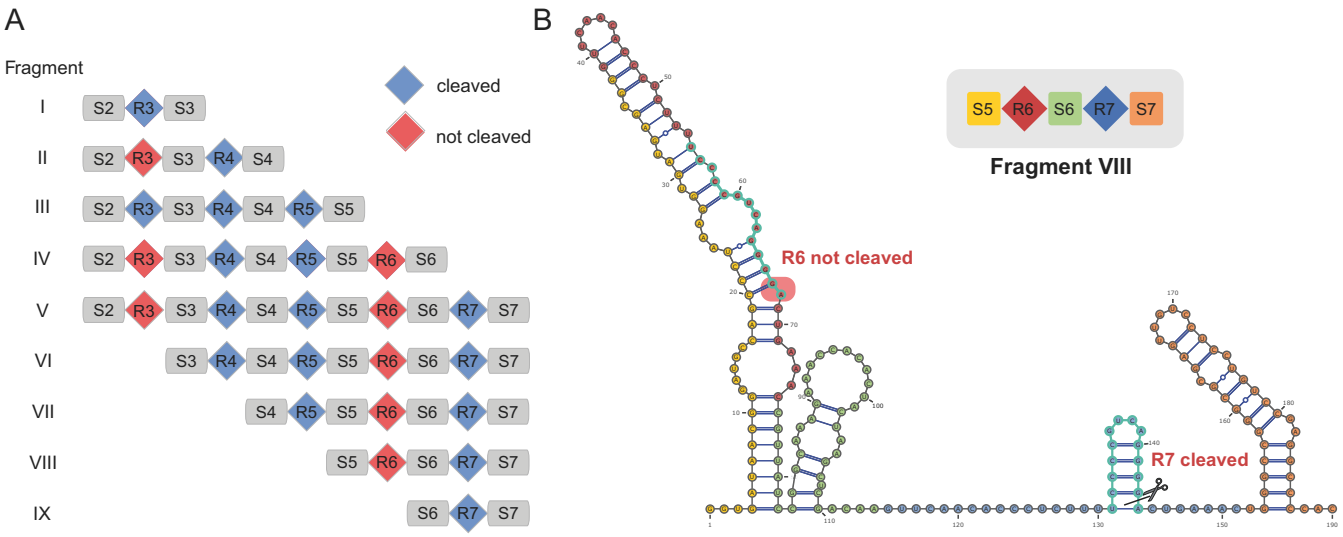


Figure 5. Systematic analysis of CRISPR2 cleavage by Cas6-1. (**A**) Schematic overview of full length CRISPR2 transcripts and positions of cleavage by Cas6-1 as determined by the experiment shown in Figure 3B. All data are also summarized in Supplementary Table S3. (**B**) Prediction of a global MFE structure to determine the most probable structure for the complete CRISPR2 fragment VIII. We have indicated the positions covered by the local functional repeat structure in turquoise and the remaining repeat sequence in red (R6) or blue (R7). Spacers are colored in yellow (S5), green (S6) or orange (S7). The local functional repeat structure is formed in the cleaved repeat R7, whereas the associated position is blocked by other stems in the non-cleaved repeat R6 of fragment VIII.

because there was no RNA processing or degradation in parallel incubations with empty-vector mock preparations.

Adjacent spacer sequences influence the formation of the substrate structure

Computational analysis of CRISPR structure suggested that adjacent spacer sequences can influence the formation of the repeat structure motif (17). To test whether surrounding sequence context influences Cas6-1 and Cas6-2a cleavage of CRISPR1 and 2 transcripts, we calculated the accuracies (see 'Materials and Methods' section) of the local functional repeat motifs for all the products obtained experimentally (Figure 3), each representing a subsequence of the original CRISPR1 or CRISPR2 array with a different number of repeats and spacers (Figure 4). As can be clearly seen, the accuracy of the functional local repeat structure is significantly lower for the non-cleaved compared to the cleaved fragments. To illustrate this further, we chose the CRISPR2 repeats R6 and R7 within fragment VIII as an example. We observed that repeat R7 was cleaved in this fragment, whereas repeat R6 was not cleaved as part of the same fragment VIII (Figure 5A), indicated by the lack of the 123, 76 and 67 nt fragments for CRISPR2-VIII in Figure 3B and D. Predictions of the secondary structure revealed that only the local functional repeat structure of R7 is formed in fragment VIII, whereas the associated positions are blocked by the alternative secondary structure in case of the non-cleaved repeat R6 (Figure 5B). The latter case is especially interesting: while all repeat instances are of identical sequence, the adjacent spacer sequences differ. Thus, this finding illustrated the possible relevance of local basepairing interactions between a repeat and its adjacent spacers. Therefore, we measured the predicted stability of the hairpin motif from Figure 1D for each repeat instance, using the base pair accuracy: a value close to 1 or 0 corresponds to a high or low predicted structure stability, respectively. We observed a very clear separation of base pair accuracies with respect to the presence or absence of cleavage events (Figures 3 and 6). In summary, the base pair accuracy could explain 43 out of 50 experimental cleavage outcomes for CRISPR2 (Supplementary Tables S3 and 4). These results justify using the base pair accuracy to predict cleavage events that depend on the stability of local structure motifs.

DISCUSSION

Mature crRNAs are integrated into large ribonucleoprotein complexes with their cognate Cas proteins and guide these complexes to invading foreign RNA (9) or DNA sequences (7,9,10,35). Therefore, the accurate processing of crRNA precursors is an essential step in the CRISPR-Cas antiviral defense mechanism. However, the variation in mechanisms and involved factors is amazing. RNases play also a key role in the control of mRNA stability and gene expression mediated by bacterial sRNAs (36) and host RNases are able to perform crucial functions in the maturation of CRISPR transcripts, too. For example, in Type II systems a trans-activating RNA (tracrRNA) together with the endogenous RNase III is the key enzyme for the maturation of crRNAs (37), while a CRISPR element in *Listeria monocytogenes* is processed by the endogenous polynucleotide phosphorylase (38). However, also the opposite situation exists, in which a native CRISPR-Cas system regulates the expression of an endogenous transcript encoding a bacterial lipoprotein requiring Cas9, together with tracrRNA and the scaRNA sRNA (39). These finding illustrate that it is worthwhile to study CRISPR-Cas systems of different subtypes and different organisms.

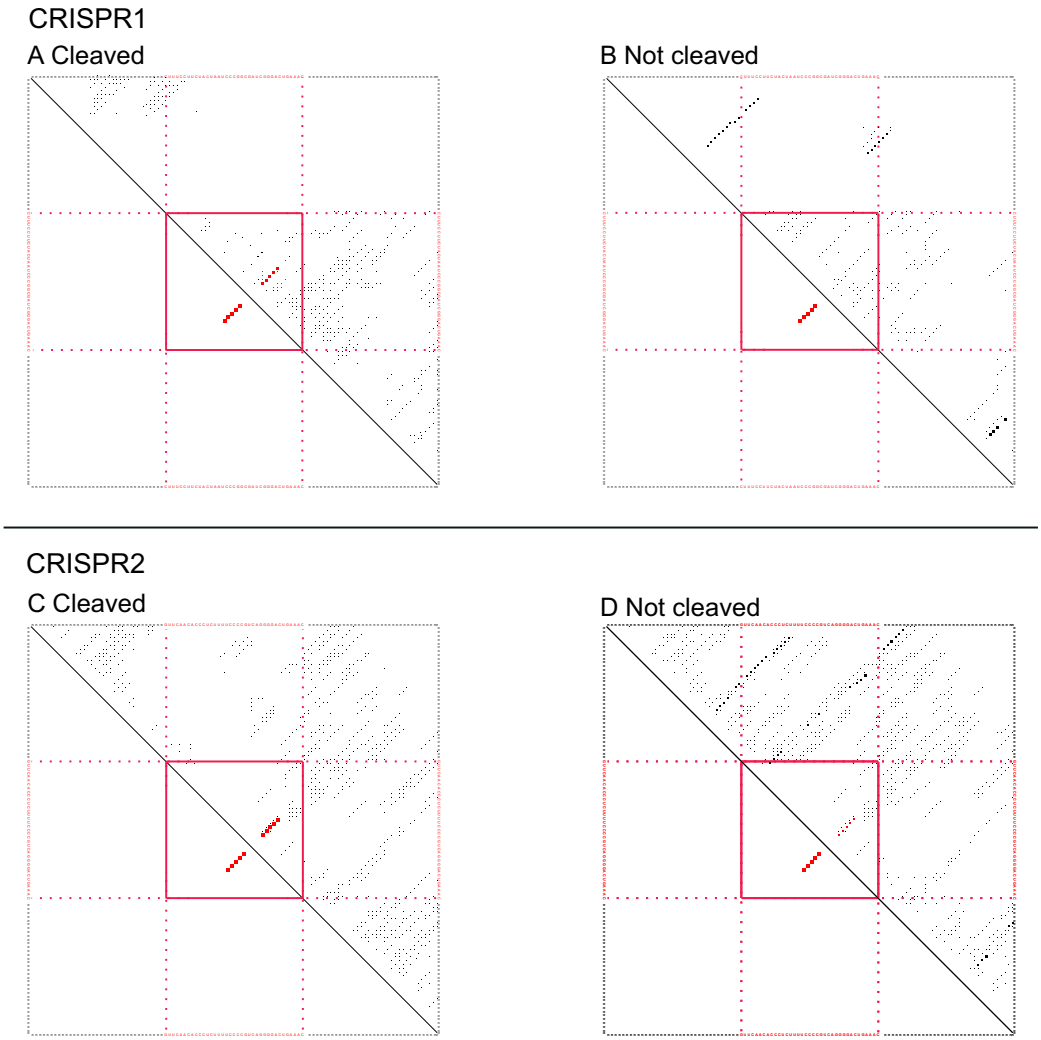


Figure 6. Average dot plot of (**A** and **B**) CRISPR1 or (**C** and **D**) CRISPR2 repeats that are cleaved or not cleaved in the artificial fragments. Each dot represents the base pairing potential of the nucleotides that can be read from the respective rows and columns. In the bottom-left triangle, we show the base pairs involved in the functional motif (highlighted in red). The top-right triangle represents the average base pair probability of each repeat instance in the respective fragments. Adjacent to the repeats are the average base pair probabilities with respect to the position in the adjacent spacer. (A and C) Repeat instances that were cleaved. (B and D) Repeat instances that were not cleaved. We observe that in (B and D) there are many more base pairs in the surrounding context than in (A and C) and that the base pairs of the functional motif are more probable on average in (A and C) than in (B and D). Note that we represent the spacers by their mode length.

Here, we provide the first biochemical analysis of pre-crRNA processing by Cas6 proteins in cyanobacteria and in a subtype I-D CRISPR-Cas system. The enzyme, Cas6-1, is able to specifically process synthetic repeat RNA of its corresponding CRISPR1 repeat-spacer array *in vitro,* but also CRISPR2 RNA. However, when the *cas6-2a* gene, which is located upstream of the CRISPR2 array, is knocked out, CRISPR2 RNA accumulates to lengths mainly >200 nt (17). An RNA-seq analysis of a *cas6-2a* mutant confirmed that no repeat-specific processing of the CRISPR2 array occurs (Supplementary Figure S3). The implication of these results is that Cas6-1 does not process CRISPR2 transcripts in the absence of Cas6-2a *in vivo,* despite its observed *in vitro* cleavage activity. Strikingly, we observed a similar promiscuity in the ability of Cas6-2a to correctly process CRISPR1-derived transcripts *in vitro*, whereas it did not substitute the missing Cas6-1 activity in deletion mutants of *cas6-1 in vivo* (17). Interestingly, when analyzing the cleavage of only a single artificial repeat sequence *in vitro*, Cas6-1 could cleave CRISPR1 (oligonucleotide C1 in Figure 1), as did Cas6-2a with the CRISPR2 repeat substrate (oligonucleotide C2 in Figure 1). Conversely, Cas6-1 and Cas6-2a cleaved the single non-cognate repeats only inefficiently (Figure 1A). These substrate specificities are consistent with the specificity for the two enzymes for either CRISPR1 or CRISPR2 *in vivo* (17) and, in view of the identical secondary structures, must be caused by the differences in the respective repeat sequences. Therefore, it is a possibility that these enzymes possess a higher affinity to their cognate CRISPR transcripts *in vivo*, which thus would outcompete the non-cognate substrates. However, that does not explain the results of the genetic analyses when one of

the respective endonuclease genes was deleted (17). Another explanation is that the cleavage of the respective precursor transcripts requires the specific binding of a further Cas protein that is assembled into a complex only with the correct endonuclease. The latter explanation fits well with the observed accumulation of longer CRISPR2 transcripts in absence of Cas6-2a (Supplementary Figure S3) that could be stabilized by the specific binding of a second Cas protein or a complex of proteins. Furthermore, the CRISPR1 and 2 *in vitro* transcript cleavage assays also confirmed that Cas6-1 and Cas6-2a are not sufficient to generate the CRISPR1 mature crRNA species detected *in vivo* (17): only fragments corresponding to intermediate crRNAs with a length of 72–76 nt were observed. Therefore, these intermediate crRNAs are expected to be processed in a second step into mature crRNAs by a so far unknown ribonuclease.

We noticed for both enzymes when incubating them with CRISPR repeat-spacer fragments of varying lengths that cleavage occurred in most repeat instances, but not in all (Figure 3). To shed light on the reasons why a repeat is cleaved or not cleaved depending on the position within a transcript, local secondary structure predictions were performed on the whole transcript VIII of CRISPR2 taking the influence of adjacent sequences into account. In Figure 5B we exemplify this for the repeat instances R6 and R7 and the cleavage behavior with Cas6-1: For the 5′ part of fragment VIII we see long helical regions that form between repeat R6 and the preceding spacer sequence that obstruct the formation of the characteristic hairpin structure motif and cover the cleavage site within the repeat R6, making it inaccessible for the enzyme. Indeed, repeat instance R6 was not cleaved. In contrast, the functional hairpin motif is clearly formed the 3′ part of fragment VIII and accessible for the enzyme, consistent with the cleavage of R7.

These findings can be generalized for all repeat instances. In Figure 4, the distribution for the accuracy of the local functional repeat motif from Figure 1D for cleaved and non-cleaved fragments, for CRISPR2, is shown. As can be clearly seen, the accuracy of the function local repeat structure is significantly lower for non-cleaved fragments. The reason for this low accuracy is a competition between the local functional repeat structure and competing stable stems, as shown for repeat R6 of fragment VIII in Figure 5B. To visualize this effect for all cleaved and non-cleaved artificial fragments, we averaged the dot plots (plus a context of 35 nt) of repeats that are cleaved or not cleaved in the artificial fragments of CRISPR1 and CRISPR2. In comparison, we observed that among uncleaved fragments (Figure 6B and D) there are many more base pairs in the surrounding context than among cleaved fragments (Figure 6A and C) and that the base pairing of the functional motif has a much higher average probability for the cleaved fragments (Figure 6B and D). We conclude that Cas6-1 and Cas6-2a are both enzymes that require the formation of the hairpin motif in repeats for substrate recognition.

We could successfully explain cleavage events by measuring the predicted structure stability of the hairpin motif in each repeat instance: high predicted stabilities led to a cleavage and low stabilities did not. Further computational investigation into these results showed that the sequence context (i.e. spacers) surrounding a repeat instance could form stable structures with the repeat sequence that sequester the formation of the functional hairpin motif. Despite each repeat instance always having the same adjacent spacers, some repeat instances were both cleaved and not cleaved, depending on the fragment length. This implies that long-range effects on the repeat structure exist that go beyond the directly adjacent spacer sequences.

In summary, we describe the dependency of Cas6-mediated cleavage on the RNA secondary structure by analyzing the cleavage patterns of nine CRISPR1 and nine CRISPR2 *in vitro* transcripts, varying in length and sequence, which revealed that a specific repeat is not necessarily always cleaved by Cas6-1 or Cas6-2a. A successful cleavage was furthermore influenced by the context of adjacent (and even more distantly located) sequences and was thereby dependent on secondary structure formation in the direct neighborhood of the repeat. The influence of surrounding sequences might lead to variations in crRNA abundances and should be taken into account when designing artificial CRISPR arrays.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank Thomas Wallner for his support in the preparation of alkaline hydrolysis ladders and sequencing gels.

FUNDING

German Research Foundation (DFG) program FOR1680 ‘Unravelling the Prokaryotic Immune System’ [HE 2544/8-2, BA 2168/5-2 to W.R.H., R.B.]. Funding for open access charge: Deutsche Forschungsgemeinschaft grants [HE 2544/8-2 and BA 2168/5-2].
Conflict of interest statement. None declared.

REFERENCES

1. Jansen,R., van Embden,J.D.A., Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.
2. Lange,S.J., Alkhnbashi,O.S., Rose,D., Will,S. and Backofen,R. (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.*, **41**, 8034–8044.
3. Bhaya,D., Davison,M. and Barrangou,R. (2011) CRISPR-Cas systems in Bacteria and Archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.*, **45**, 273–297.
4. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.
5. 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. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.
6. Alkhnbashi,O.S., Costa,F., Shah,S.A., Garrett,R.A., Saunders,S.J. and Backofen,R. (2014) CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformarmatics*, **30**, i489–496.
7. 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*, **321**, 960–964.

8. Hale,C., Kleppe,K., Terns,R.M. and Terns,M.P. (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcusfuriosus*. *RNA*, **14**, 2572–2579.
9. Hale,C.R., Zhao,P., Olson,S., Duff,M.O., Graveley,B.R., Wells,L., Terns,R.M. and Terns,M.P. (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, **139**, 945–956.
10. Karginov,F.V. and Hannon,G.J. (2010) The CRISPR system: small RNA-guided defense in Bacteria and Archaea. *Mol. Cell*, **37**, 7–19.
11. Przybilski,R., Richter,C., Gristwood,T., Clulow,J.S., Vercoe,R.B. and Fineran,P.C. (2011) Csy4 is responsible for CRISPR RNA processing in *Pectobacteriumatrosepticum*. *RNA Biol.*, **8**, 517–528.
12. 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 Dev.*, **22**, 3489–3496.
13. 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*, **20**, 1574–1584.
14. Punetha,A., Sivathanu,R. and Anand,B. (2014) Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR-Cas type I-C system. *Nucleic Acids Res.*, **42**, 3846–3856.
15. Hochstrasser,M.L. and Doudna,J.A. (2015) Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem. Sci.*, **40**, 58–66.
16. Cai,F., Axen,S.D. and Kerfeld,C.A. (2013) Evidence for the widespread distribution of CRISPR-Cas system in the phylum Cyanobacteria. *RNA Biol.*, **10**, 687–693.
17. Scholz,I., Lange,S.J., Hein,S., Hess,W.R. and Backofen,R. (2013) CRISPR-Cas systems in the cyanobacterium *Synechocystis* sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS One*, **8**, e56470.
18. 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. *et al.* (2011) Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
19. Kopfmann,S. and Hess,W.R. (2013) Toxin antitoxin systems on the large defense plasmid pSYSA of *Synechocystis* sp. PCC6803. *J. Biol. Chem.*, **288**, 7399–7409.
20. Kopfmann,S., Roesch,S.K. and Hess,W.R. (2016) Type II toxin-antitoxin systems in the unicellular cyanobacterium Synechocystis sp. PCC 6803. *Toxins*, **8**, E228.
21. Richter,H., Zoephel,J., Schermuly,J., Maticzka,D., Backofen,R. and Randau,L. (2012) Characterization of CRISPR RNA processing in *Clostridiumthermocellum* and *Methanococcusmaripaludis*. *Nucleic Acids Res.*, **40**, 9887–9896.
22. 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*, **329**, 1355–1358.
23. Haurwitz,R.E., Sternberg,S.H. and Doudna,J.A. (2012) Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *EMBO J.*, **31**, 2824–2832.
24. Lintner,N.G., Kerou,M., Brumfield,S.K., Graham,S., Liu,H., Naismith,J.H., Sdano,M., Peng,N., She,Q., Copie,V. *et al.* (2011) Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J. Biol. Chem.*, **286**, 21643–21656.
25. Shao,Y. and Li,H. (2013) Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6. *Structure*, **21**, 385–393.
26. Sokolowski,R.D., Graham,S. and White,M.F. (2014) Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system. *Nucleic Acids Res.*, **42**, 6532–6541.
27. Wang,R., Preamplume,G., Terns,M.P., Terns,R.M. and Li,H. (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure*, **19**, 257–264.
28. Plagens,A., Tjaden,B., Hagemann,A., Randau,L. and Hensel,R. (2012) Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. *J. Bacteriol.*, **194**, 2491–2500.
29. Plagens,A., Tripp,V., Daume,M., Sharma,K., Klingl,A., Hrle,A., Conti,E., Urlaub,H. and Randau,L. (2014) In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Res.*, **42**, 5125–5138.
30. Zhang,J., Rouillon,C., Kerou,M., Reeks,J., Brugger,K., Graham,S., Reimann,J., Cannone,G., Liu,H., Albers,S.-V. *et al.* (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol. Cell*, **45**, 303–313.
31. Bertani,G. (1951) Studies on lysogenesis. *J. Bacteriol.*, **62**, 293–300.
32. Lange,S.J., Maticzka,D., Mohl,M., Gagnon,J.N., Brown,C.M. and Backofen,R. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.*, **40**, 5215–5226.
33. Bernhart,S.H., Hofacker,I.L. and Stadler,P.F. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.
34. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
35. Marraffini,L.A. and Sontheimer,E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in Bacteria and Archaea. *Nat. Rev. Genet.*, **11**, 181–190.
36. Saramago,M., Bárria,C., Dos Santos,R.F., Silva,I.J., Pobre,V., Domingues,S., Andrade,J.M., Viegas,S.C. and Arraiano,C.M. (2014) The role of RNases in the regulation of small RNAs. *Curr. Opin. Microbiol.*, **18**, 105–115.
37. Deltcheva,E., Chylinski,K., Sharma,C.M., Gonzales,K., Chao,Y., Pirzada,Z.A., Eckert,M.R., Vogel,J. and Charpentier,E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, **471**, 602–607.
38. Sesto,N., Touchon,M., Andrade,J.M., Kondo,J., Rocha,E.P.C., Arraiano,C.M., Archambaud,C., Westhof,É., Romby,P. and Cossart,P. (2014) A PNPase dependent CRISPR System in Listeria. *PLoS Genet.*, **10**, e1004065.
39. Sampson,T.R., Saroj,S.D., Llewellyn,A.C., Tzeng,Y.-L. and Weiss,D.S. (2013) A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature*, **497**, 254–257.
40. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Supplementary material

This chapter contains the supplementary materials for all publications in a chronological order.

Supplementary information
CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems

Sita J. Lange[1,#] , Omer S. Alkhnbashi[1,#], Dominic Rose[1,#], Sebastian Will[1,5]and Rolf Backofen[1,2,3,4,*]

[1]Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-Universität Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany
[2]ZBSA Centre for Biological Systems Analysis, University of Freiburg, Habsburgerstr. 49, 79104 Freiburg, Germany
[3]BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, University of Freiburg, Germany
[4]Center for non-coding RNA in Technology and Health, University of Copenhagen, Gronnegardsvej 3, DK-1870 Frederiksberg C, Denmark
[5]Bioinformatics, Department of Computer Science, University of Leipzig,04107 Leipzig, Germany
[#]These authors contributed equally to this work

Email: Sita J. Lange - sita@informatik.uni-freiburg.de; Omer S. Alkhnbashi - alkhanbo@informatik.uni-freiburg.de; Dominic Rose - dominic@informatik.uni-freiburg.de; Sebastian Will - will@bioinf.uni-leipzig.de; Rolf Backofen*- backofen@informatik.uni-freiburg.de;

*Corresponding author

S1 Additional methods

S1.1 Cas subtype annotation from Haft et al. 2005.

To annotate the early Cas subtypes from Haft *et al.* [1], we followed the procedure given in Kunin *et al.* [2]. More specifically, we downloaded the single *cas* gene models created by Haft *et al.* from the TIGRFAM database. Using the HMMER program with the TIGRFAM models (same as for the single *cas* gene annotation), we searched the 20 kb of nucleotides up- and downstream of the array locus and annotated a *cas* gene if it was found with an E-value ≤ 0.001. We used a strict annotation of Cas subtypes, whereby all *cas* genes of a subtype were required.

S1.2 Webserver input: adding new repeat sequences to the existing CRISPR clustering

The user of our CRISPRmap webserver can enter any CRISPR sequences and they will be assigned to our sequence families and structure motifs, if possible, and integrated into the hierarchical CRISPRmap tree. Thus, information on conservation is available for not only sequences in our dataset, but also novel, yet unsequenced, CRISPRs. In the following, we describe the procedure for one input sequence, many sequences are done simultaneously in the same way:

1. *Is the repeat sequence in our database?* If the given repeat sequence is in our database, in either

orientation, we highlight this sequence (or one if many copies exist) in our CRISPRmap cluster tree, and automatically assign it to the corresponding structure motif and/or sequence family and stop here.

2. *What is the correct orientation?* If the user is not sure about the correct repeat orientation, i.e. the checkbox for repeat orientation has been activated, we first predict the orientation with our model described in the methods section of the main manuscript. The orientation should then be consistent with our data.

3. *Is it structured or unstructured?* The RNA structure prediction algorithm, RNAfold [3] is used to determine whether the repeat sequence is structured or unstructured. If the minimum free energy structure is the unstructured sequence, i.e. contains no base-pairs, it remains unassigned to a structure motif and we continue with Step 5.

4. *Does it belong to a structure motif?* Albeit a structure being predicted, the repeat does not necessarily belong to a conserved structure motif. We add the repeat sequence to all repeats assigned to one of our structure motifs and re-run RNAclust [4] with a modified UPGMA algorithm (see following section "Constrained Clustering"). In short, the modification allows the generation of the cluster tree by keeping the motifs intact, i.e. non-overlapping. If a repeat falls into or next to one of the existing structure motifs, we assign it to the motif by the following: (1) The repeat is folded by RNAfold [3] with the option -p to calculate a structure dotplot. (2) This dotplot is aligned with the consensus dotplot of the structure motif using LocARNA. (3) The repeat is assigned to be a member of the motif if it is able to fold into the consensus structure of that respective motif with at most one base-pair missing. We ensure that the new consensus structure contains at least four base-pairs and is at the same position as previously. A comparison of the new and old consensus structures and alignments is given on the web server results page.

5. *Does it belong to one of our conserved sequence families?* We assign the repeat to a conserved sequence family by comparing it to the previously calculated ClustalW sequence profiles [5], see Methods section "Clustering of repeat sequences into conserved sequence families". Let $sim(F, r)$ be the profile score of a repeat r compared with the profile of the family F, where $r \notin F$. For each family, the minimum F_{min} and maximum F_{max} profile similarity was determined by removing each sequence from the family, re-calculating the profile for the remaining sequences, and determining the similarity score of the respective repeat to the profile. A repeat r was then assigned to a sequence family F if (1) $sim(F, r)$ is greater or equal to F_{min} and (2) the distance between $sim(F, r)$ and F_{max} is the minimum for all families.

6. *Where is it located in the CRISPRmap cluster tree?* With a final run of RNAclust on all repeat sequences, we get the updated CRISPRmap cluster tree and we highlight the input sequence location in this tree. Any additional annotations (outer rings), such as Cas subtype, are not displayed for novel repeat sequences.

S1.3 Constrained Clustering

We consider the general problem to cluster a set of taxa hierarchically based on their distances. Additionally, we constrain the clustering such that certain, e.g. a priori known, clusters are prevented from mixing with each other.

Given is a set of taxa, indexed from 1 to n, together with all pairwise distances between the taxa; furthermore, a set $\mathcal{X}$ of disjoint clusters of these taxa, i.e. $\mathcal{X}$ is contained in the powerset of $\{1, \ldots, n\}$ and all non-identical clusters c and d in $\mathcal{X}$ do not intersect. Commonly, $\mathcal{X}$ covers only a subset of all taxa;

therefore, we distinguish *constrained taxa* (that are contained in some element of $\mathcal{X}$) and the remaining *unconstrained taxa*.

We aim to construct a cluster tree of the taxa, i.e. a rooted binary tree T with n leaves corresponding to the n taxa. First, this tree should reflect the given distances. Second it has to support the clustering given by $\mathcal{X}$ such that clusters in $\mathcal{X}$ are grouped together but unconstrained taxa can be interspersed freely. For this purpose, we require that no subtree of T contains leaves from two different clusters in $\mathcal{X}$ unless both clusters are completely contained in the subtree. We call this condition $\mathcal{X}$*-cluster constraint*. (Formally: for each subtree with leaves L and each pair of non-identical clusters c and d in $\mathcal{X}$, $c \cap L \subset c$ implies $d \cap L = \emptyset$.)

Our novel constrained clustering algorithm is based on the unweighted pair group method UPGMA. The original algorithm UPGMA starts from n singleton clusters corresponding to the n taxa. Until all clusters are combined, it iteratively merges the two nearest clusters. For the latter, the cluster distances are initially derived from the input distances and distances to new clusters are computed after each merge of clusters. The sequence of merges determines the cluster tree. The novel algorithm modifies UPGMA, such that, in each iteration, it merges the nearest pair of clusters that can be merged without violating the $\mathcal{X}$-cluster constraint. To check this condition efficiently, we keep track for each cluster whether it contains some elements of a cluster in $\mathcal{X}$ and whether it includes such a cluster completely. Merging two clusters does violate the constraint if and only if each cluster overlaps some cluster in $\mathcal{X}$ but does not cover it completely.

S1.4 Horizontal gene transfer between bacteria and archaea

Although archaeal CRISPRs are generally well-separated from bacterial ones in general, we observed a few instances where an archaeal CRISPR is located within a bacterial-dominated region and vice versa. To investigate whether these mixed regions could arise from potential horizontal transfer, we applied BLAST to search for homologous Cas1 (or Cas2) protein sequences (Cas1 and Cas2 are the most ubiquitous Cas proteins and exist in both bacteria and archaea). We identified 24 archaeal and 8 bacterial repeats that were assigned to sequence families or structure motifs dominated by the opposite domain. For 75% (18 out of 24) of the archaeal repeats, we identified Cas1 or Cas2 homologs in bacteria in the top five BLAST hits (E-value $\leq 2 \times 10^{-10}$); the same was true for only one of the four bacterial repeats.

S2 Supplementary tables

S2.1 Number of Cas subtype annotations

We annotated each CRISPR in our dataset according to the closest Cas subtypes as described in the methods of the manuscript. The two major Cas subtype annotation systems were considered [1, 6]; the number of CRISPRs we annotated with each subtype is given in Table S1.

S2.2 Summary tables of sequence families and structure motifs

Supplementary Tables S2–S19 summarise the sequence families and structure motifs, sorted according to the superclass they belong to. The numbering of the families is according to the number of repeats belonging to that family. The annotations in each column is done manually with respect to the majority of repeats in that family (see other supplementary file for the full list). For the Cas subtype, an annotation is

Subtype	Archaea	Bacteria	Total
10 subtypes from Makarova *et al.* 2011 [6]			
I-A	134	203	337
I-B	89	293	382
I-C	**14**	**322**	**336**
I-D	49	38	87
I-E	**8**	**447**	**455**
I-F	**1**	**155**	**156**
II-A	0	50	50
II-B	9	95	104
III-A	148	223	371
III-B	108	149	257
% CRISPR	87 %	68 %	72 %
8 subtypes from Haft *et al.* 2005 [1]			
Apern	65	0	65
Dvulg	**1**	**184**	**185**
Ecoli	**8**	**369**	**377**
Hmari	15	36	51
Mtube	8	9	17
Nmeni	0	27	27
Tneap	89	254	343
Ypest	**0**	**120**	**120**
% CRISPR	29 %	35 %	34 %

Table S1: The number of identified Cas subtype annotations for our REPEATS dataset. There were double as many annotations using the more recent classification from Makarova *et al.*, however, we did not require that all *cas* genes from the respective subtype to be present; whereas the annotations performed for Haft *et al.* were more strict, since we used full subtype models (see methods). In general, Dvulg, Ecoli, Hmari, Mtube, Nmeni, and Ypest correspond to I-C, I-E, I-B, III-A, both type II, and I-F, respectively. Structured repeats with very stable and conserved hairpin motifs, mainly found in bacteria, are written in bold. Note that the 9 subtype II-B CRISPRs in archaea are likely to be incorrect as we did not identify an RNase III in these organisms. Automated annotation of subtype II-B was especially difficult as it contains no subtype-specific Cas protein.

only given if this is more or less clear. If there is a complete mix of subtypes, no information is given. The Cas subtypes are summarised according to the *cas* genes that are found in the majority of chromosomes which contain the CRISPRs of each family or motif. More details of the majority *cas* genes is given on the web server. Archaeal families and motifs are highlighted in blue. If the CRISPRmap webserver is updated in future, then these tables supply a record for sequence families and structure motifs that are referred to in this work. The secondary structures of the motifs and sequence logos of the families are also provides in the tables.

Table S2: **Summary for the bacterial sequence families in Superclass A.**

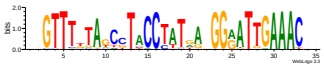
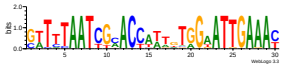
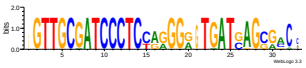
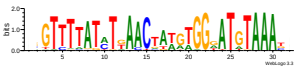
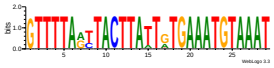
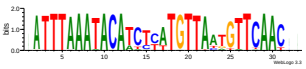
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F1		289	M10 un- structured	Firmicutes	I-B III-A III-B
F25		23	un- structured	mixed bacteria	I-A II-B III-A
F16		40	un- structured	Thermotogae	III-A
F30		19	M2	Actinobacteria	-
F6		124	M8 un- structured	Firmicutes	I-A
F28		20	un- structured	Firmicutes	I-A
F34		15	M21	Firmicutes	II-B
F9		76	M7	Firmicutes	III-B

Table S3: **Structure motif summary for bacterial motifs in Superclass A.**

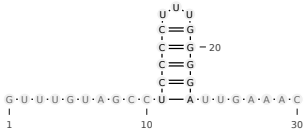
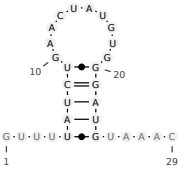
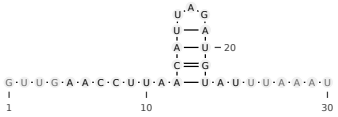
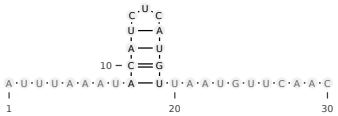
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M10		50	F1	Firmicutes	I-B II-B III-A
M8		55	F6	Firmicutes	I-A I-B III-A
M21		26	F34 unassigned	Firmicutes	-
M7		78	F9	Firmicutes	I-A III-B

Table S4: **Summary for the archaeal sequence families in Superclass A.**

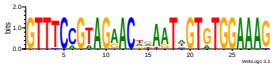
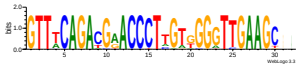
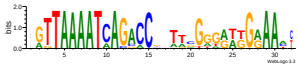
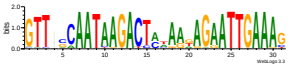
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F29		20	un-structured	Euryarchaeota Crenarchaeota	III-A
F19		32	un-structured	Euryarchaeota	-
F7		108	M15 M16 M27	Euryarchaeota	I-A
F10		70	un-structured	Euryarchaeota	I-B

Table S5: **Structure motif summary for archaeal motifs in Superclass A.**

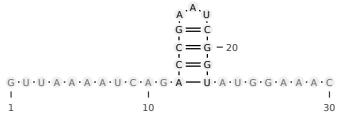
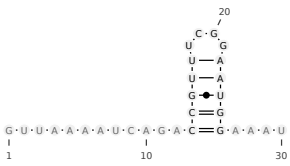
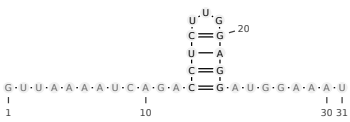
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M15		35	F7	Euryarchaeota	-
M27		17	F7	Euryarchaeota	-
M16		33	F7	Euryarchaeota	-

Table S6: **Sequence family summary for Superclass B.**

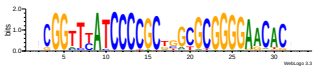
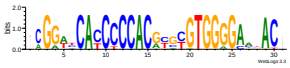
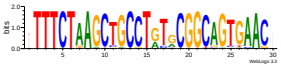
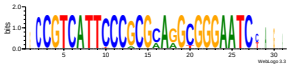
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F2		221	M1	Actinobacteria Proteobacteria	I-E
F18		35	M1	mixed bacteria	I-E II-B
F8		88	M6	Proteobacteria	I-F
F22		26	M18	Proteobacteria	I-E

Table S7: **Structure motif summary Superclass B.**

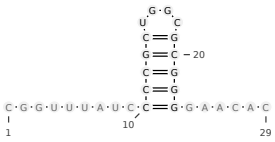
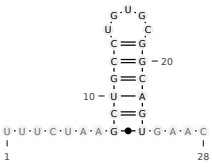
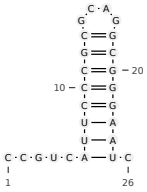
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M1		265	F2 F18	mixed bacteria	I-E
M6		89	F8	Proteobacteria	I-F
M18		28	F22	Proteobacteria	III-B

Table S8: **Sequence family summary for Superclass C.**

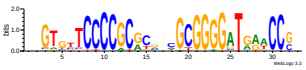
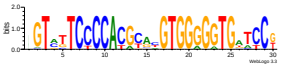
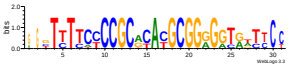
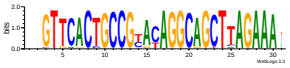
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F4		172	M2	Actinobacteria Proteobacteria	I-C I-E II-B
F21		27	M2	mixed bacteria	I-E
F33		16	M2	mixed bacteria	I-C I-E II-B
F5		135	M4	Proteobacteria	I-F

Table S9: **Structure motif summary for Superclass C.**

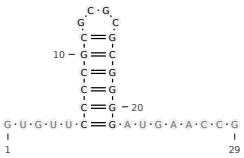
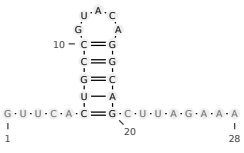
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M2		222	F4 F21 F30 F33 unassigned	mixed bacteria	I-E
M4		142	F5 unassigned	Proteobacteria	I-F

Table S10: **Sequence family summary for Superclass D.**

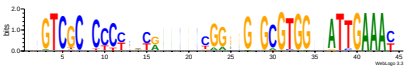
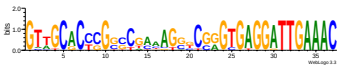
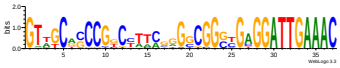
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F3		210	M3 M9	mixed bacteria	I-C
F37		14	M9	Deinococcus-Thermus	I-C III-B
F32		18	M9	Deinococcus-Thermus Proteobacteria	I-C

Table S11: **Summary for structure motifs in Superclass D with sequence conservation.**

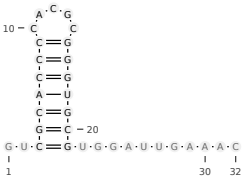
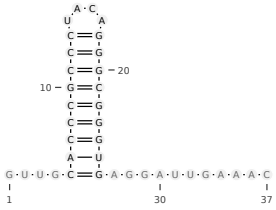
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M3		195	F3	mixed bacteria	I-C
M9		52	F3 F32 F37	mixed bacteria	I-C I-A

Table S12: **Summary for structure motifs in Superclass D without sequence conservation.**

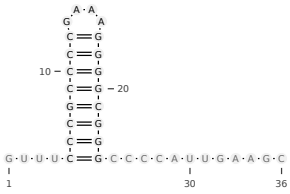
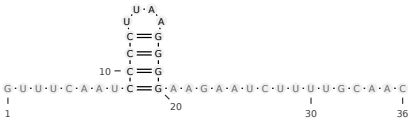
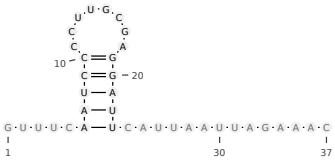
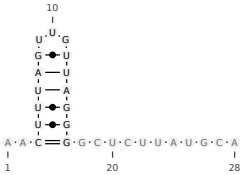
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M19		28	unassigned	mixed bacteria	I-A II-B III-B
M25		19	unassigned	mixed bacteria	III-A III-B
M30		13	unassigned	Cyanobacteria Chloroflexi	I-E II-B
M33		10	unassigned	mixed bacteria	II-B

Table S13: **Sequence family summary for Superclass E.**

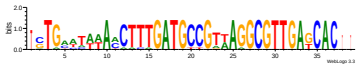
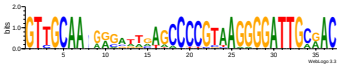
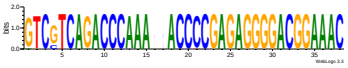
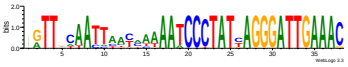
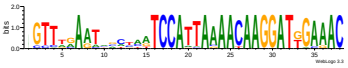
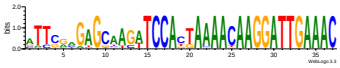
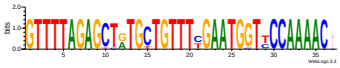
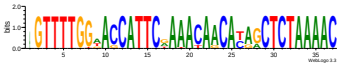
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F39		13	M5	mixed bacteria	I-A I-B II-B
F31		19	M5	Deinococcus-Thermus	III-A
F12		45	M5	Actinobacteria	II-B III-A
F23		24	M12	Cyanobacteria	I-D II-B
F20		28	M13 un-structured	Euryarchaeota mixed bacteria	I-B
F26		23	M13 un-structured	Euryarchaeota	-
F35		15	un-structured	Firmicutes	II-A
F27		22	M14 un-structured	Firmicutes	II-A II-B

Table S14: **Summary of bacterial structure motifs in Superclass E with sequence conservation.**

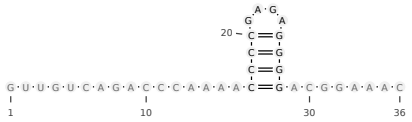
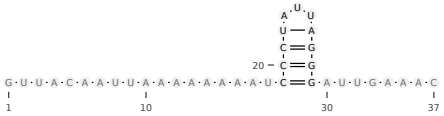
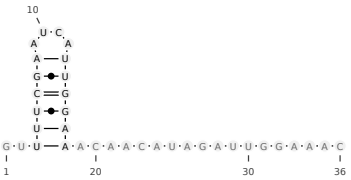
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M5		106	F12 F31 F39 unassigned	Cyanobacteria mixed bacteria	II-B III-A
M12		40	F23 unassigned	mixed bacteria	-
M14		35	F27 unassigned	Firmicutes Cyanobacteria	II-A II-B

Table S15: **Summary of bacterial structure motifs in Superclass E without sequence conservation.**

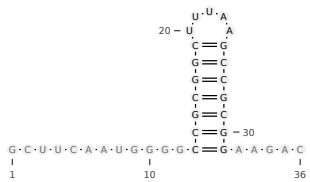
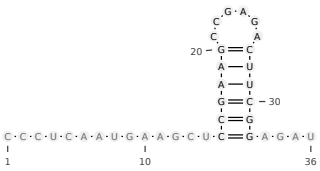
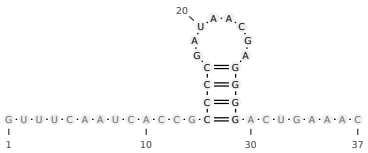
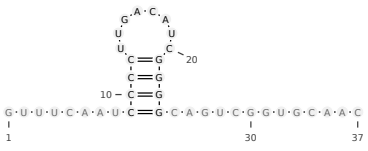
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M23		23	unassigned	mixed bacteria	-
M26		19	unassigned	Actinobacteria	-
M28		16	unassigned	mixed bacteria	I-C III-A
M24		21	unassigned	mixed bacteria	-

Table S16: **Summary of archaeal structure motifs in Superclass E.**

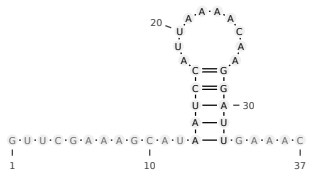
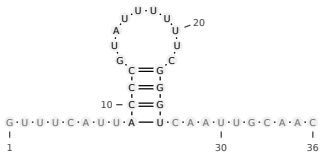
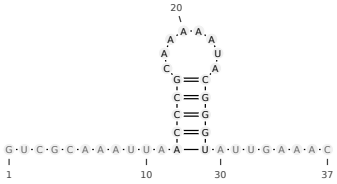
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M13		37	F20 F26	Euryarchaeota	I-A
M31		11	unassigned	Euryarchaeota mixed bacteria	-
M29		14	unassigned	Euryarchaeota mixed bacteria	II-B

Table S17: **Sequence family summary for Superclass F.**

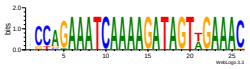
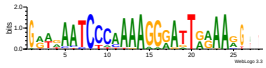
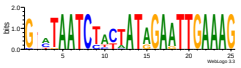
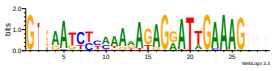
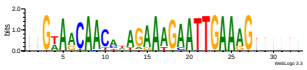
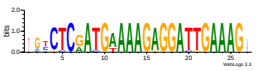
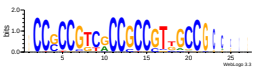
#	Sequence Logo	Size	Motifs	Taxonomy	Subtypes
F24		23	un-structured	Crenarchaeota	III-A III-B
F15		42	M22 un-structured	Crenarchaeota	I-A III-B
F13		44	M17 un-structured	Crenarchaeota	I-A III-B
F11		49	M11 un-structured	Crenarchaeota	III-B
F14		44	un-structured	Crenarchaeota	I-A I-D III-A
F38		13	un-structured	mixed archaea	I-A III-B
F36		15	M20	Firmicutes	-
F40		13	un-structured	Proteobacteria	I-B
F17		39	un-structured	Actinobacteria	-

Table S18: **Summary for archaeal structure motifs in Superclass F.**

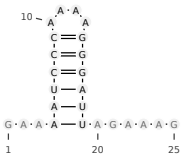
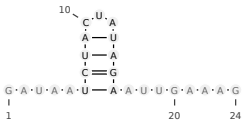
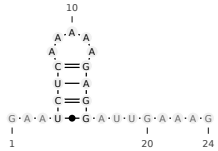
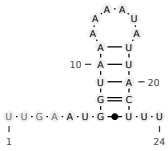
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M22		24	F15	Crenarchaeota	I-A III-B
M17		29	F13	Crenarchaeota	I-A III-B
M11		45	F11 unassigned	Crenarchaeota	III-A III-B
M20		27	F36 unassigned	Firmicutes Crenarchaeota	-

Table S19: **Final structure motif unassigned to a Superclass.**

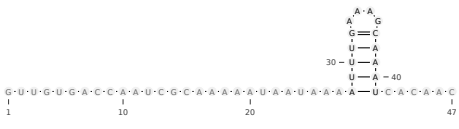
#	Structure Motif	Size	Families	Taxonomy	Subtypes
M32		10	unassigned	Becteroidetes	II-B

Table S20: Published CRISPR-Cas systems with experimental evidence of the processing mechanism. In particular, these are systems for which the Cas endoribonuclease is characterised and/or the repeat structure has been verified. Published results are consistent with our data. The IDs, a–o, are marked, in order, as red lines on the CRISPRmap tree in the manuscript in Figure 1.

ID	Organism	Family	Motif	Cas Subtype	Summary
Superclass A					
a	*Clostridium thermocellum* ATCC 27405	F1	-	I-B	Unstructured; 8-nt-5'-tag; biochemical evidence to show **Cas6b** activity [7]
b	*Pyrococcus furiosus* DSM 3638	F10	-	III-B	Unstructured; 8-nt-5'-tag; cleavage by **Cas6**; crystal structure of repeat wrapped around Cas6 [8]
Superclass C					
c	*Escherichia coli* K12 substr. W3110	F4	M2	I-E	Structure predicted, but stable; 8-nt-5'-tag; cleavage by **Cas6e**, biochemical experiments [9]
d	*Thermus thermophilus* HB8	F4	M2	I-E	Structured; 8-nt-5'-tag; cleavage by **Cas6e**; crystal structure of repeat hairpin in Cas6e (Cse3) [10, 11]
e	*Pseudomonas aeruginosa* UCBPP-PA14	F5	M4	I-F	Cleavage by Cas6f (Csy4); 8-nt-5'-tag; crystal structure and mutational analyses of repeat hairpin in **Cas6f** [12–14]
Superclass D					
f	*Bacillus halodurans* C-125	F3	M3	I-C	Cleavage by **Cas5d**; 11-nt-5'-tag mutational analysis of hairpin structure [15]
g	*Thermus thermophilus* HB27	F37	M9	I-C	Cleavage by **Cas5d**; 11-nt-5'-tag biochemical experiments [16]
h	*Nanoarchaeum equitans* Kin4-M	-	-	I-A	Biochemical evidence to show **Cas6b** activity; 8-nt-5'-tag [17]
Superclass E					
i	*Synechocystis* sp. PCC6803	-	M5	I-D & III-variant	Cleavage by **Cas6**; 8-nt-5'-tag; biochemical experiments, extended structure prediction of hairpin motif [18]
j	*Methanosarcina marzei* Gö1	F26	M13	I-B & III-B	Cleavage by **Cas6b**; 8-nt-5'-tag; structure probing experiment of hairpin [19]
k	*Clostridium thermocellum* ATCC 27405	F20	-	I-B	Biochemical evidence to show **Cas6b** activity; 8-nt-5'-tag [7]
l	*Staphylococcus epidermidis* RP62A	-	M28	III-A	Cleavage by **Cas6**; 8-nt-5'-tag; hairpin structure as in M28 verified by mutational analysis and sequence specificity around cleavage site [20]
m	*Methanococcus maripaludis* C5	-	M29	I-B	Cleavage by **Cas6b**; 8-nt-5'-tag; biochemical experiments [7]
n	*Synechocystis* sp. PCC6803	-	M14	III-variant	Biochemical analysis of **Cmr2** implicate its involvement in either cleavage, crRNA stabilisation, or array expression regulation; 13-nt-5'-tag [18]
o	*Streptococcus pyogenes* SF370 (M1 serotype)	F35	-	II-A	Cleavage with **tracrRNA**, host **RNase III** and **Cas9**, biochemical experiments; 22-nt-5'-tag [21]

S3 Supplementary figures

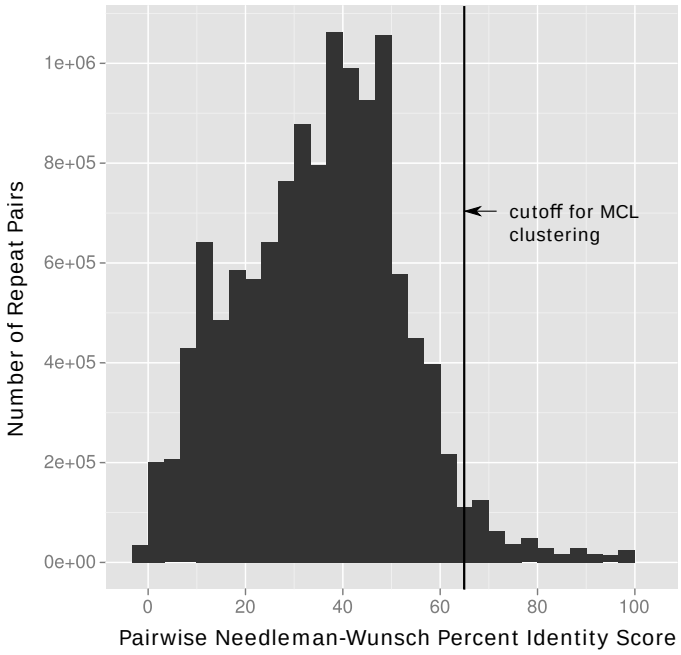


Figure S1: **Pairwise similarities for repeats.** We plotted the distribution of pairwise percent identities (x-axis) of Needleman-Wunsch [22] alignments for all repeats to determine a cutoff for the Markov clustering. Here we see that 65% is a reasonable cutoff in comparison to the background distribution. Repeats with a similarity below 65% are set to zero. Because of the short repeat length and conserved sequence motifs, it is necessary to choose such a high cutoff.

Figure S2: **Verifying repeat families with sequence profiles and re-assigning individual repeats.** All repeats were clustered into families using Markov clustering [23,24]. We verified these families using an independent method of sequence profiles, see Methods section "Clustering of repeat sequences into conserved sequence families". After the generation of one profile per family, we caculated the profile scores for each repeat in the REPEATS dataset. We plotted the profile scores (y-axis) for each repeat assigned to one of the families (x-axis) as red-coloured dots in Supplementary Figure S2. Subsequently, we used this range of profile scores to re-assign repeats to one of the existing families as stated in the main text of the manuscript. Profile scores for re-assigned dots are in blue (73 repeats). These profile scores are also used to assign new input repeat sequences from the webserver to one of our existing families.

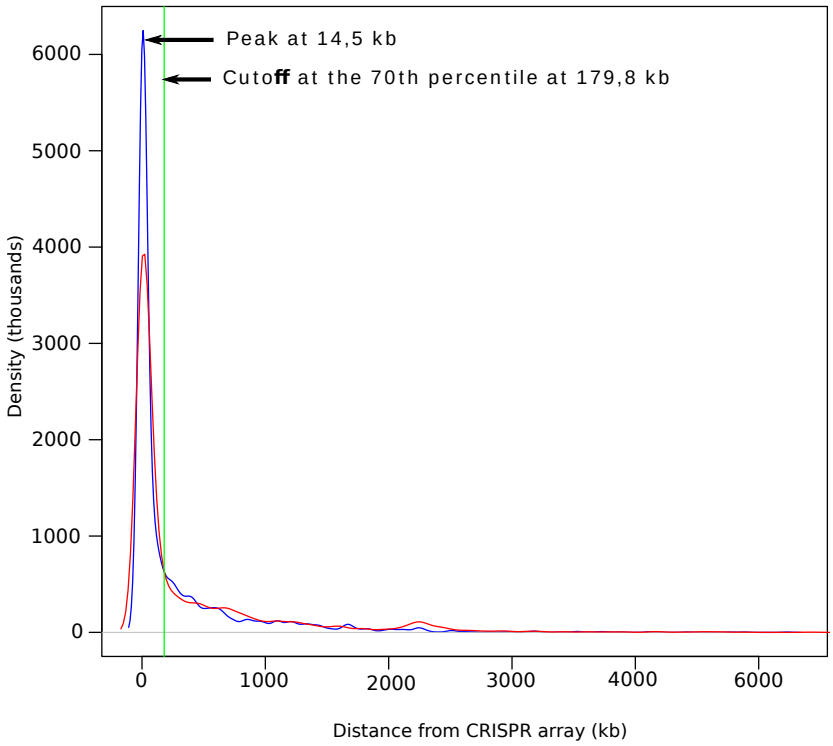


Figure S3: **Distance of *cas* genes in the annotation of subtypes from Makarova *et al.* 2011.** Distance of signature subtypes is in blue and the distance of signature types is in red; the cutoff is indicated with the green line. The plot shows the distribution of the closest signature genes to the CRISPR array. A signature gene is one that is unique to either the subtype or the type, respectively.

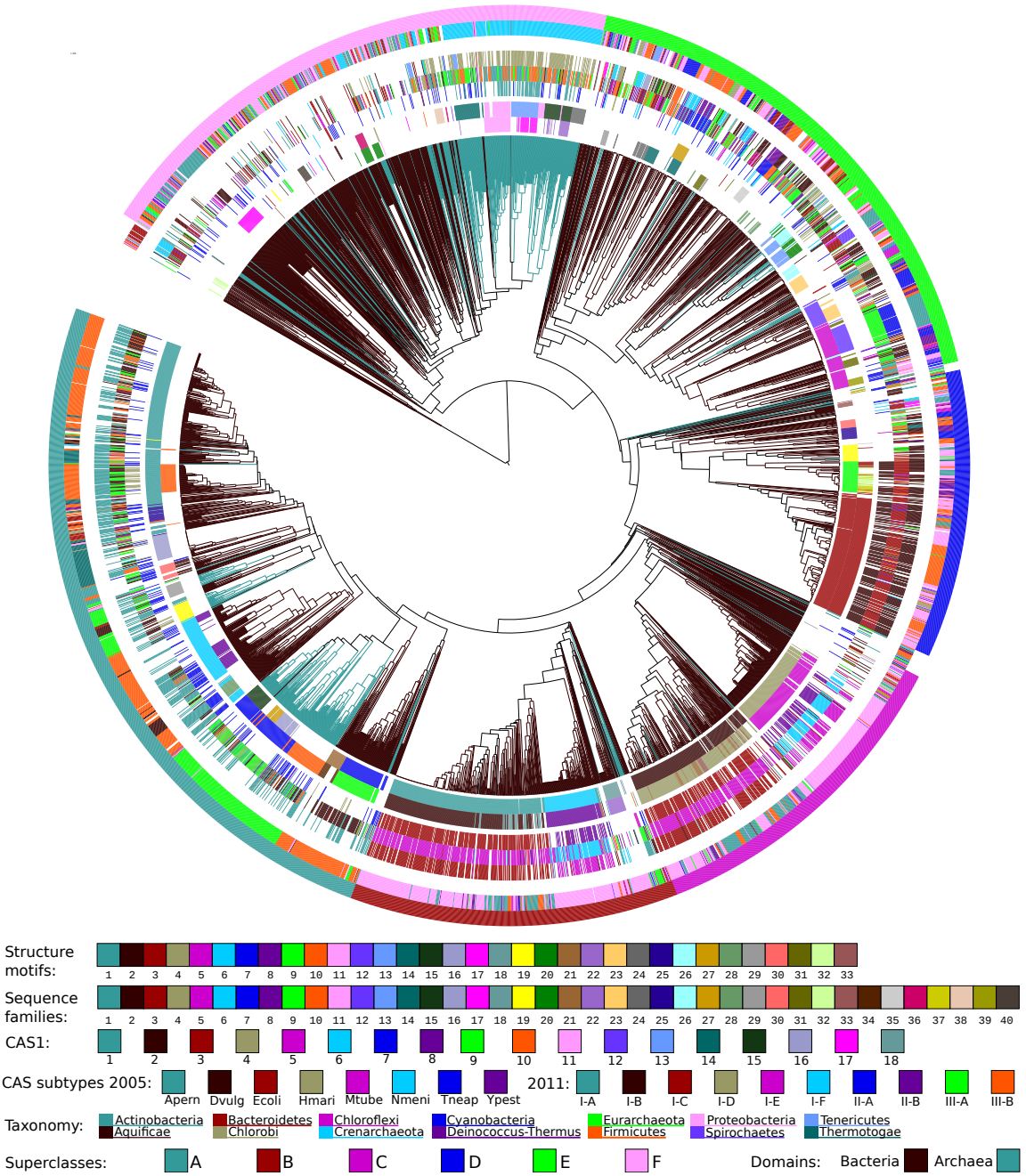


Figure S4: **CRISPR of repeat conservation including all annotations.** CRISPR repeats cluster into 33 structure motifs and 40 sequence families. Here we show the cluster tree with all annotation rings—the "alltogether" option in the webserver—colour coding starts from inside to outside, see the legend. The branches of the tree are labelled according to the origin of the repeat: blue-green for archaea and dark brown for bacteria. **Ring 1** (inner-most) 33 structure motifs, **ring 2** 40 sequence families, **ring 3** Haft 2005 subtype annotation, **ring 4** Makarova 2011 subtype annotation, **ring 5** 18 cas1 clusters, **ring 6** taxonomic phyla annotation and **ring 7** (outer-most) the six superclasses for general orientation.

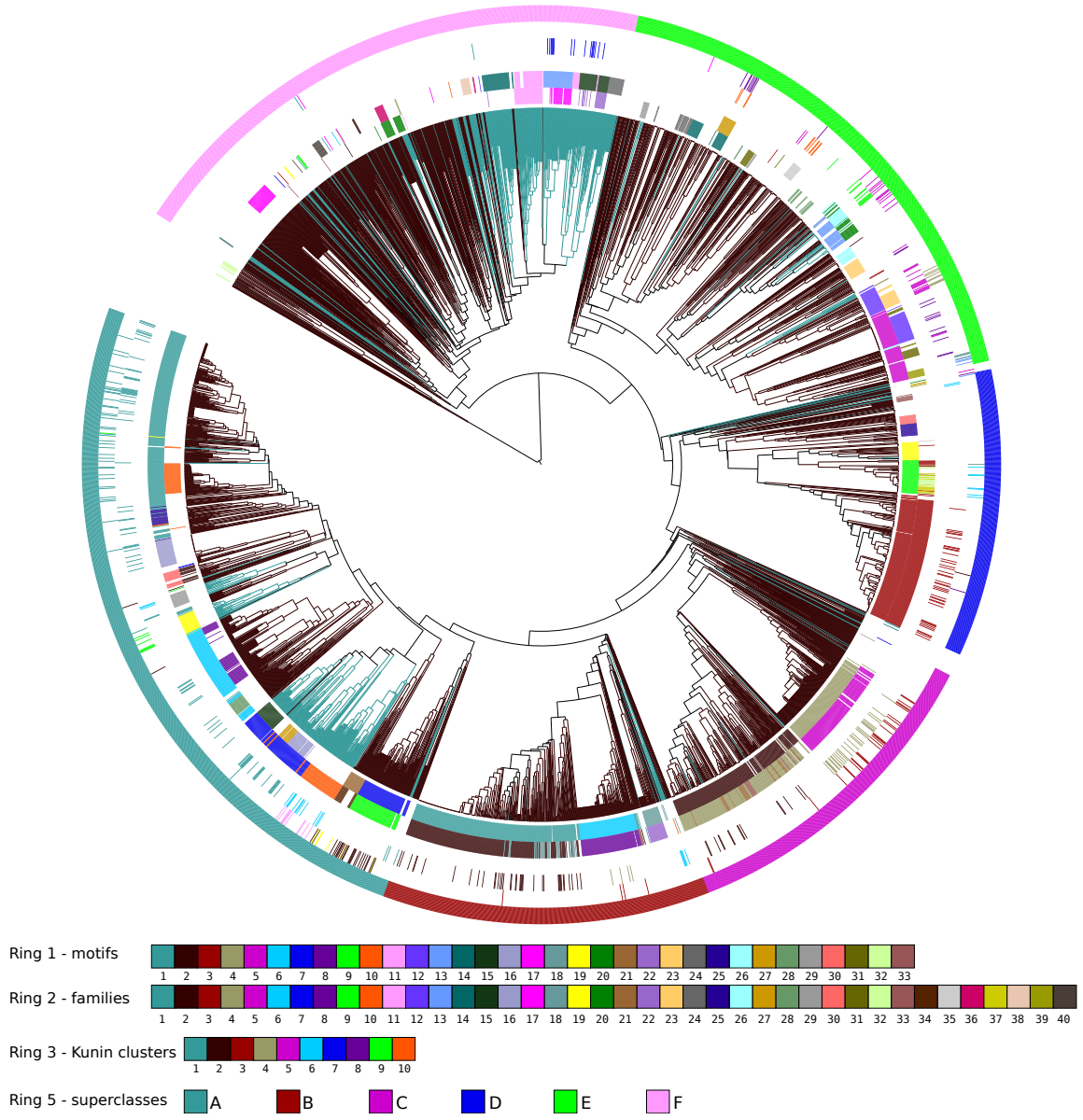


Figure S5: **Comparison of our clustering with previous domain-wide repeat clusters or families on our CRISPRmap tree.** The branches of the tree are labelled according to the origin of the repeat: blue-green for archaea and dark brown for bacteria. **Ring 1** (inner-most) shows our structure motifs, **ring 2** shows our sequence families. After the white ring, we show ten of the twelve clusters from Kunin *et al.* [2,25] in **Ring3**; clusters 11 and 12 contain fewer than ten repeats and to be consistent with our cluster minimum size, we have removed them here. **Ring 4** contains those sequences of the Rfam [26] database that are also contained in REPEATS (since we have all sequenced genomes to-date) and only families (16 out of 65) with at least ten sequences. We do not mark the family names here, but just want to show the relative locations of sequences in the CRISPRmap tree. **Ring 5** (outer-most) shows the six superclasses for general orientation. In summary, we clearly see that our data is significantly more comprehensive than previous work.

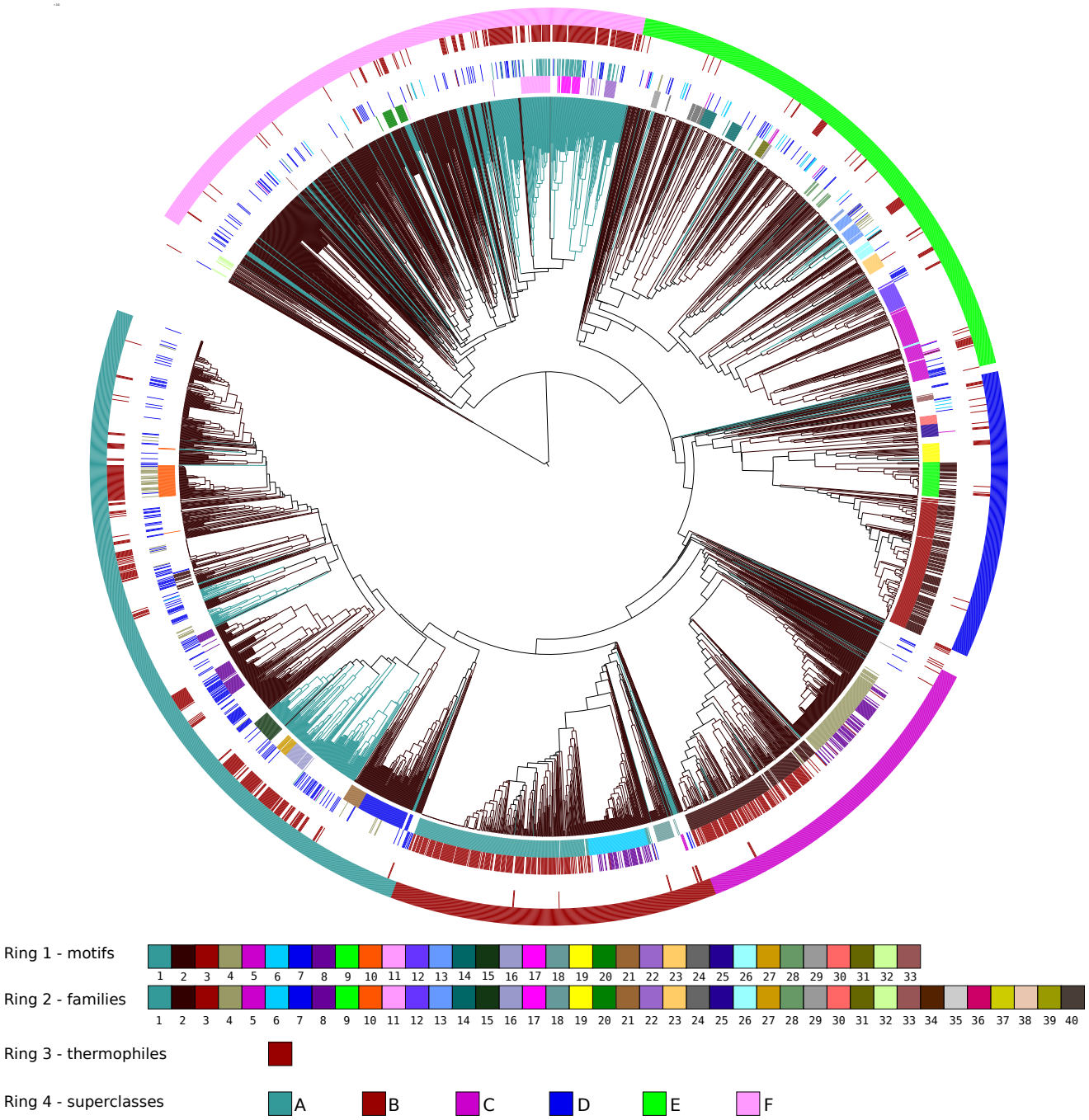


Figure S6: **CRISPRs found in thermophilic organisms. Ring 3** shows the number of CRISPRs that were found in thermophilic organisms (taken from ExtremeDB, `http://extrem.igib.res.in`, March 2013). At leat 17% of our CRISPRs stem from thermophiles. Of these CRISPRs, 81% are in superclasses A and F, which are associated with diverse types I-A, I-B, I-D, III-A and III-B. In contrast, only 7% of the bacterial CRISPRs in superclasses B, C, and D—with strong Cas subtype associations—stem from thermophiles. The same is true for bacteria only: 60% of the CRISPRs from bacterial thermophiles are in superclass A.

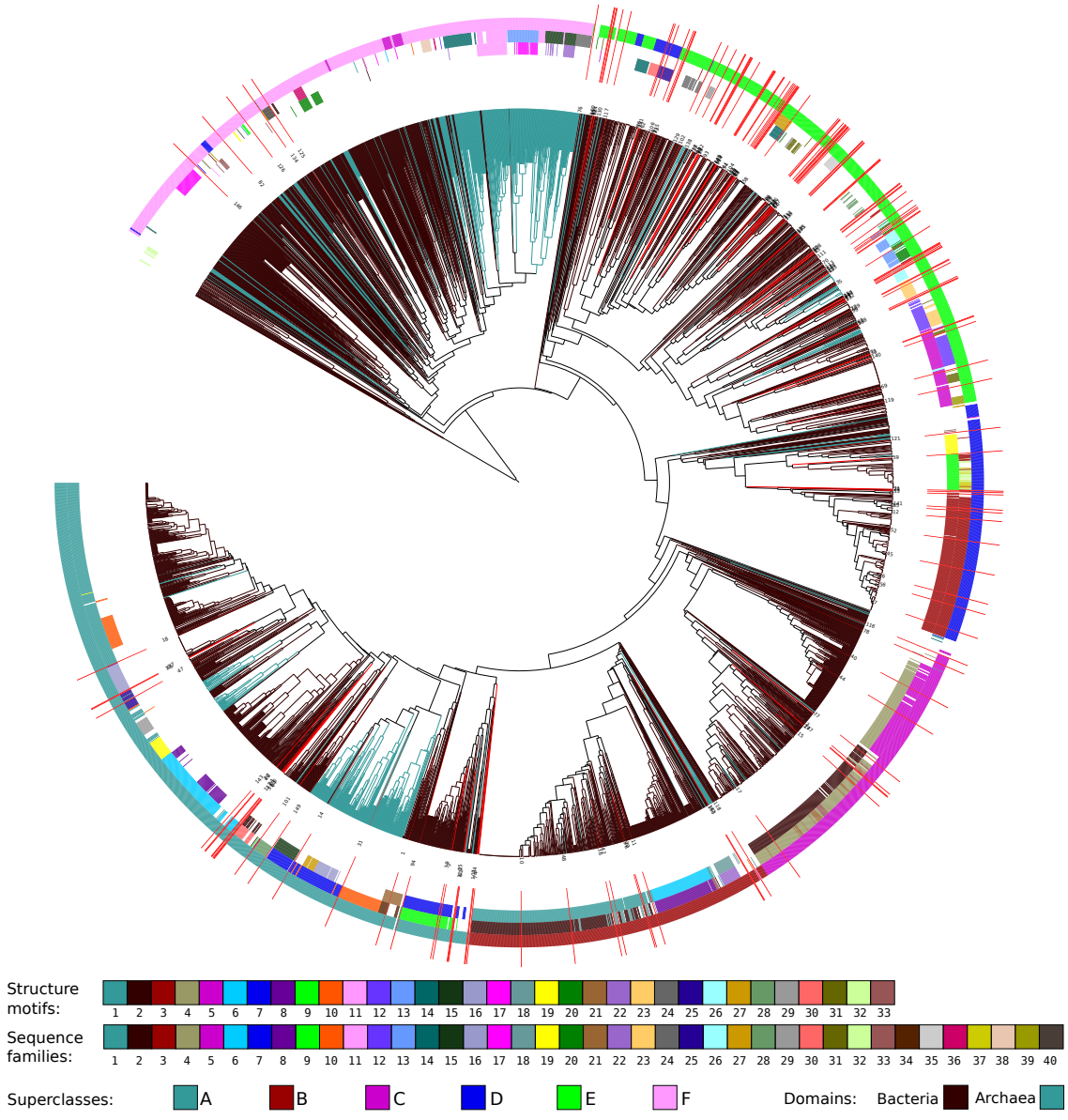


Figure S7: **CRISPRmap tree—a use-case study.** This is the CRISPRmap cluster tree after re-clustering 150 repeats from a human metagenomic studies [27] together with our REPEATS data. The new 150 repeats are maked with red lines. Interestingly, many repeats have been assigned to superclass E and cluster together to potentially form new classes of motifs or families.

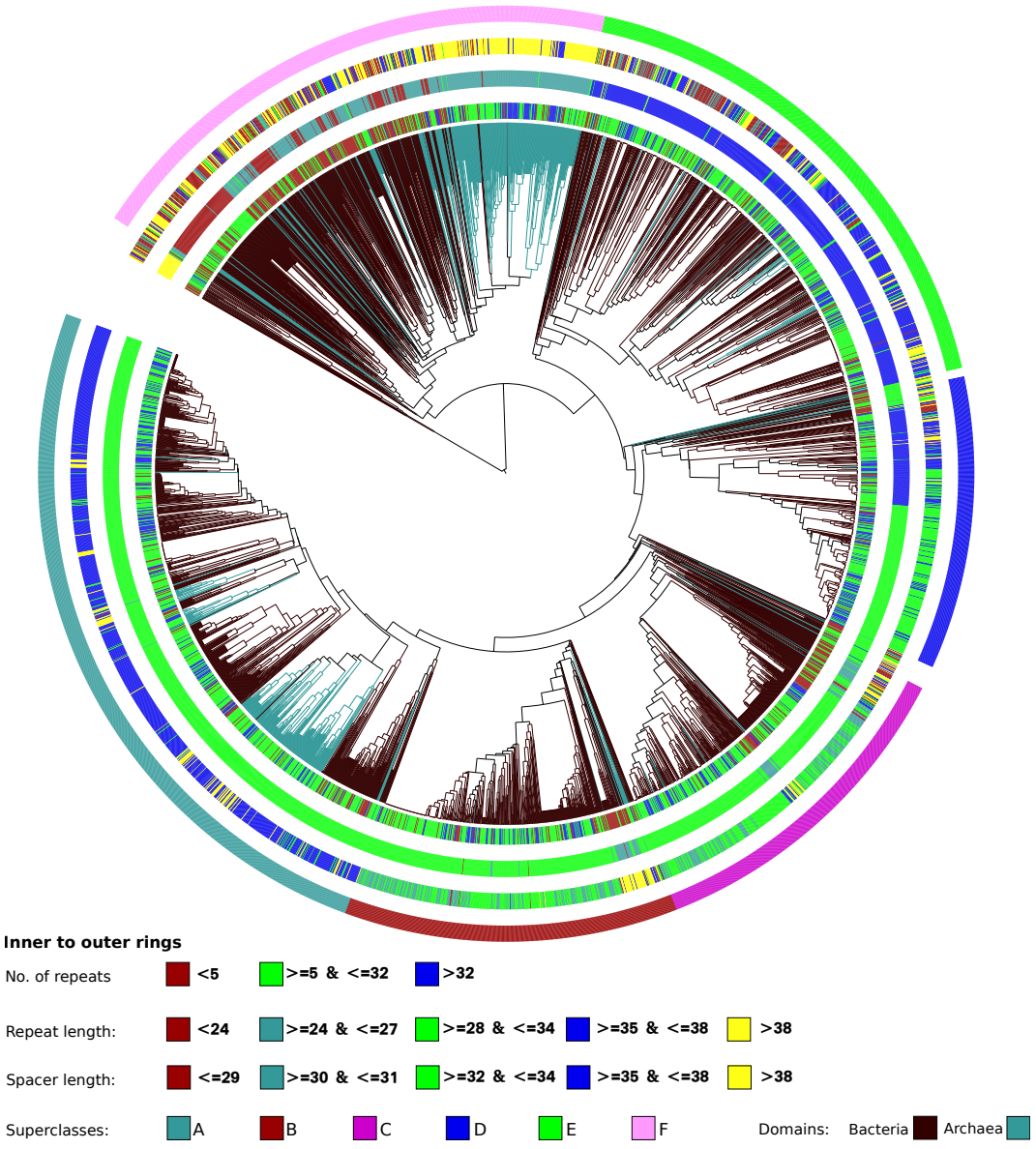


Figure S8: **Analysis of array, repeat and average spacer sizes.** First, we see the very small arrays containing less than 5 repeat instances (red-brown) are mostly located in the more divergent parts of the CRISPRmap tree; most are within the bacterial part of superclass F. Many of these arrays may not be functional CRISPR-Cas systems, but other repetitive elements instead. Second, superclass F contains both some unusually short and unusually long repeats, which also may not represent functional CRISPRs. In addition repeats in superclass F and half of D are longer than those in superclasses A to the first half of D. Third, repeats in superclasses A and F are longer than ones in B-D; this means the Cas subytpes I-C, I-E, and I-F associate with shorter spacers than the others. Spacers in Crenarchaeota are unusually long with most longer than 38 nt. Interestingly, shorter repeats seem to pair with longer spacers. Cutoffs were choses according to the distribution of each array characteristic.

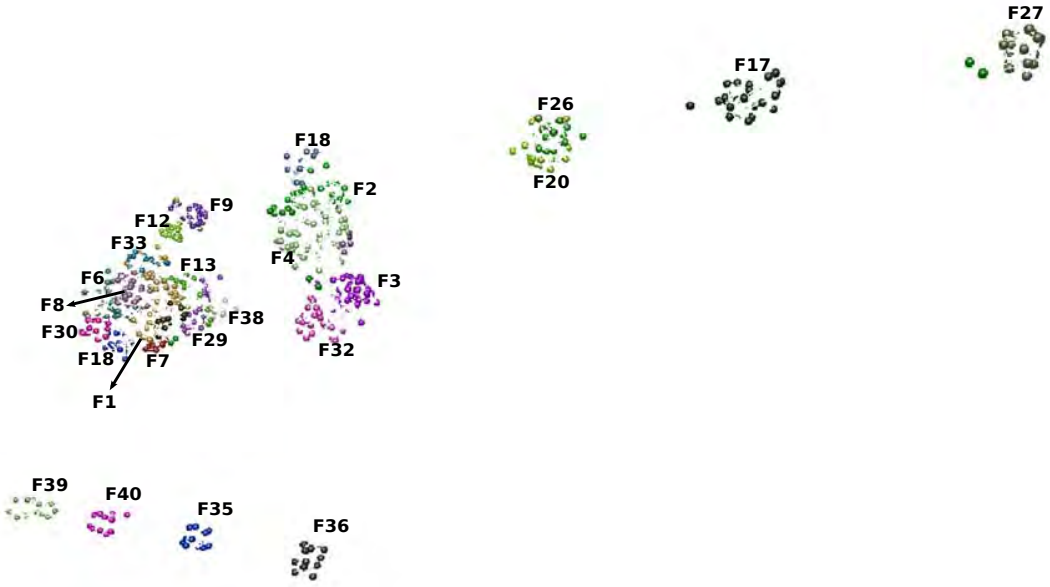


Figure S9: **Sequence families separated on a two-dimensional plane.** The 40 sequence famlies are mapped onto a two-dimensional plane by BioLayout [28] according to their percent identity scores. We have marked only those families that are clearly visible. The families are divided into two main groups with some that are more separated from the rest.

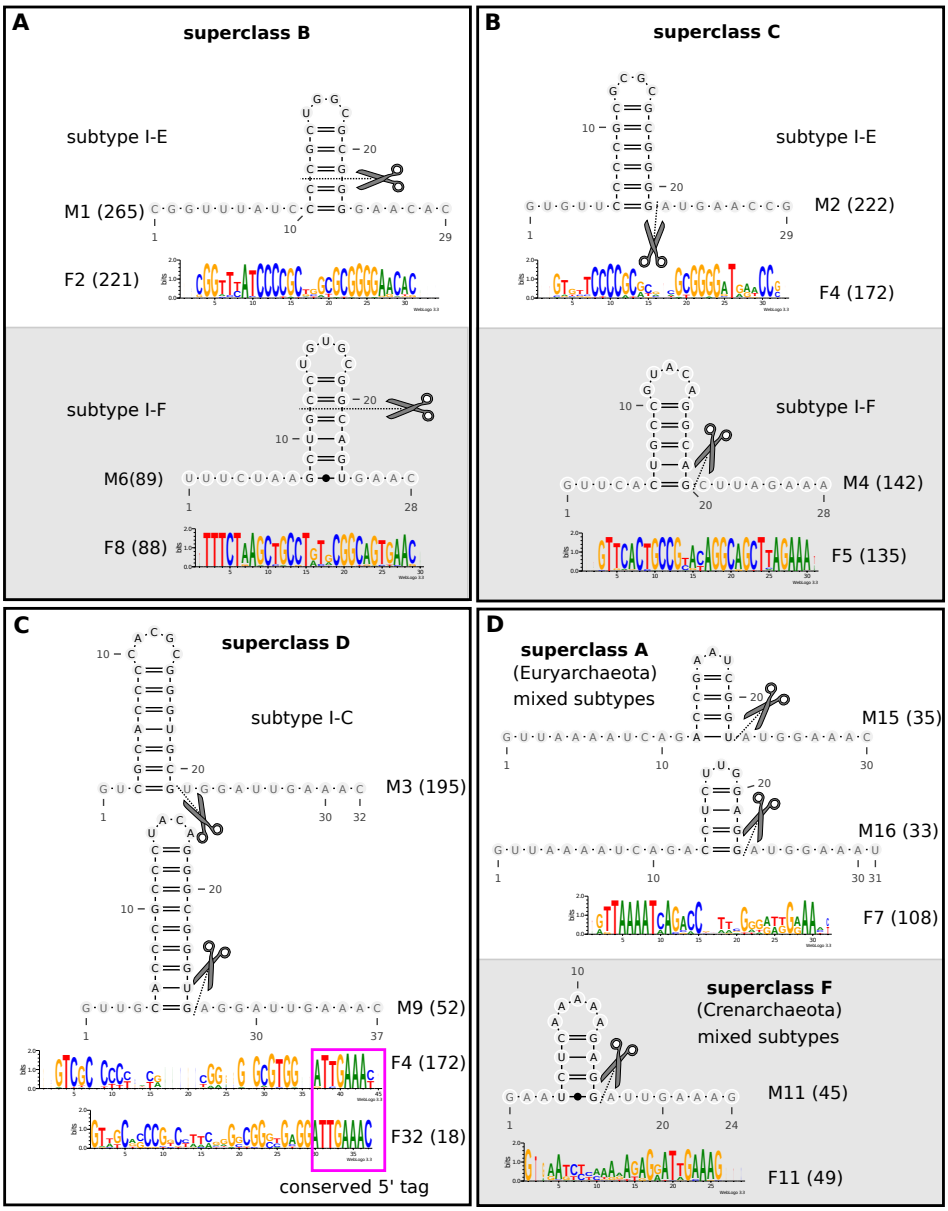


Figure S10: **Conserved structured CRISPRs fit well to published cleavage sites and display various patterns of sequence conservation.** The sequence family logos correspond to the depicted structure motifs. Potential cleavage sites are indicated as observed in the literature [**?**, 7–13, 15–18, 20]. **a.-b.** Superclasses B and C contain stable structure motifs of the subtypes I-E and I-F. The difference is that the structures in superclass B are closer to the 3' end of the repeat and that the potential cleavage site is in the double-stranded region of the stem instead of the 3' side of its base. **c.** Superclass D contains members of the I-C subtype with relatively long hairpin motifs. Note that the potential cleavage site leads to an 11 nt instead of an 8 nt tag in the mature crRNA and we also see the well-conserved 3' end of the repeat (*ATTGAAAC*); this 3' sequence is found in many CRISPRs, also in archaea. **d.** Examples of structure motifs found in archaeal repeats in superclasses A and F. These are smaller and less stable than the bacterial motifs.

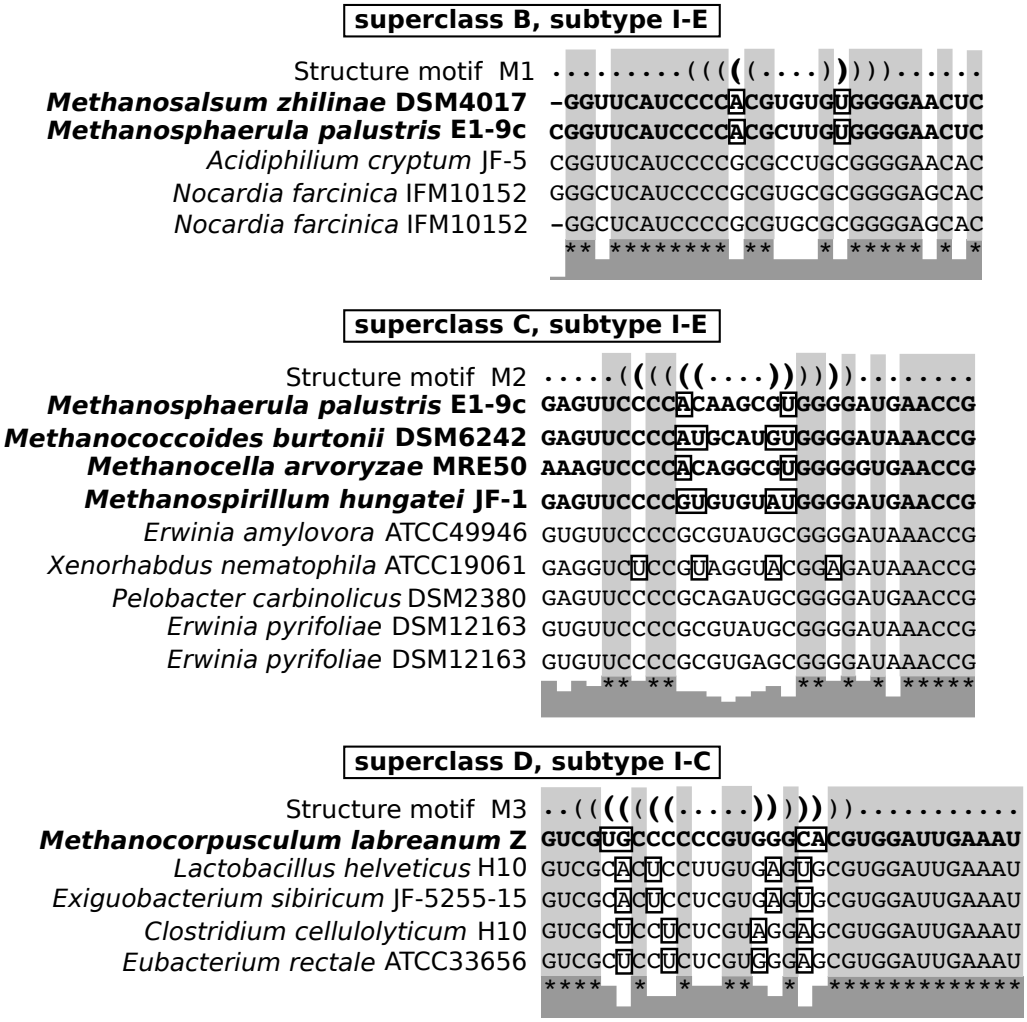


Figure S11: **Selected alignments showing evidence of horizontal transfer of structured CRISPRs from bacterial to archaeal genomes.** Archaeal CRISPRs are indicated in **bold** typeface. The secondary structure from the respective motif is written above in dot-bracket format: brackets and dots corresponds to base pairs and unpaired nucleotides, respectively. The highlighted brackets and squares show that the secondary RNA structure has been conserved by compensatory base pair mutations. These compensatory base pair mutations give excellent evidence for the conservation and importance of the respective structure motifs.

References

1. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes**. *PLoS Comput. Biol.* 2005, **1**(6):e60.
2. Kunin V, Sorek R, Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats**. *Genome Biol.* 2007, **8**(4):R61.
3. Hofacker IL, Stadler PF: **Memory efficient folding algorithms for circular RNA secondary structures**. *Bioinformatics* 2006, **22**(10):1172–6.
4. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R: **Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering**. *PLoS Comput. Biol.* 2007, **3**(4):e65.
5. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res.* 1994, **22**(22):4673–80.
6. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J, Koonin EV: **Evolution and classification of the CRISPR-Cas systems**. *Nat. Rev. Microbiol.* 2011, **9**(6):467–77.
7. Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, Randau L: **Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis**. *Nucleic Acids Res.* 2012.
8. Wang R, Preamplume G, Terns MP, Terns RM, Li H: **Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage**. *Structure* 2011, **19**(2):257–64.
9. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J: **Small CRISPR RNAs guide antiviral defense in prokaryotes**. *Science* 2008, **321**(5891):960–4.
10. Gesner EM, Schellenberg MJ, Garside EL, George MM, Macmillan AM: **Recognition and maturation of effector RNAs in a CRISPR interference pathway**. *Nat. Struct. Mol. Biol.* 2011, **18**(6):688–92.
11. Sashital DG, Jinek M, Doudna JA: **An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3**. *Nat. Struct. Mol. Biol.* 2011, **18**(6):680–7.
12. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA: **Sequence- and structure-specific RNA processing by a CRISPR endonuclease**. *Science* 2010, **329**(5997):1355–8.
13. Haurwitz RE, Sternberg SH, Doudna JA: **Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA**. *EMBO J.* 2012, **31**(12):2824–32.
14. Sternberg SH, Haurwitz RE, Doudna JA: **Mechanism of substrate selection by a highly specific CRISPR endoribonuclease**. *RNA* 2012, **18**(4):661–72.
15. Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A: **Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system**. *Structure* 2012, **20**(9):1574–84.
16. Garside EL, Schellenberg MJ, Gesner EM, Bonanno JB, Sauder JM, Burley SK, Almo SC, Mehta G, MacMillan AM: **Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases**. *RNA* 2012, **18**(11):2020–8.
17. Randau L: **RNA processing in the minimal organism Nanoarchaeum equitans**. *Genome Biol.* 2012, **13**(7):R63.
18. Scholz I, Lange SJ, Hein S, Hess WR, Backofen R: **CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 Exhibit Distinct Processing Pathways Involving at Least Two Cas6 and a Cmr2 Protein**. *PLoS One* 2013, **8**(2):e56470.
19. Nickel L, Weidenbach K, Jager D, Backofen R, Lange SJ, Heidrich N, Schmitz RA: **Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III**. *RNA Biol* 2013, **10**(5).

20. Hatoum-Aslan A, Maniv I, Marraffini LA: **Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site**. *Proc. Natl. Acad. Sci. USA* 2011, **108**(52):21218–22.

21. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E: **CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III**. *Nature* 2011, **471**(7340):602–7.

22. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J. Mol. Biol.* 1970, **48**(3):443–53.

23. van Dongen S: **Graph Clustering by Flow Simulation**. *PhD thesis*, University of Utrecht 2000.

24. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res.* 2002, **30**(7):1575–84.

25. Shah SA, Garrett RA: **CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems**. *Res. Microbiol.* 2011, **162**:27–38.

26. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A: **Rfam: Wikipedia, clans and the "decimal" release**. *Nucleic Acids Res.* 2011, **39**(Database issue):D141–5.

27. Rho M, Wu YW, Tang H, Doak TG, Ye Y: **Diverse CRISPRs evolving in human microbiomes**. *PLoS Genet.* 2012, **8**(6):e1002441.

28. Theocharidis A, van Dongen S, Enright AJ, Freeman TC: **Network visualization and analysis of gene expression data using BioLayout Express(3D)**. *Nat. Protoc.* 2009, **4**(10):1535–50.

Supplementary information S1 (table) | **The core proteins of CRISPR-Cas systems**

Family	Biochemical evidence/*in silico* prediction	Examples of available structures and structural features
Cas1	Metal-dependent deoxyribonuclease [56,115] that functions as the integrase during adaptation [17]; deletion of Cas1 in *E. coli* results in increased sensitivity to DNA damage and impaired chromosomal segregation [116].	PDB: 3GOD, 3LFX, 2YZS Unique fold with two domains: N-terminal β stranded domain and catalytic C-terminal α-helical domain.
Cas2	RNase specific to U-rich regions [60], double-stranded DNAse; forms a tight complex with Cas1 and appears to perform a structural role during adaptation.	PDB: 2IVY, 2I8E, 3EXC, 4P6I RRM (ferredoxin) fold.
Cas3 (helicase and HD domain)	Single-stranded DNA nuclease (HD domain) and ATP-dependent helicase [66]; required for interference [61].	PDB: 4QQW, 4QQX, 4QQZ, 4QQY
Cas3'' (stand alone HD nuclease)	Metal-dependent deoxyribonuclease specific for double-stranded oligonucleotides [117].	PDB: 3S4L, 3SKD
Cas4	PD-(DE)xK superfamily nuclease with four conserved cysteines coordinating one [4Fe-4S] or [2Fe-2S] cluster [15,118,119]; cleaves ssDNA in the 5' to 3' or both directions [118-120] .	PDB: 4IC1
Cas5	Subunit of effector complex interacting with large subunit and Cas7 subunit and binding the 5'-handle of crRNA [39,42,43,61,62,121,122]. in the subtype I-C system Cas5 is the ribonuclease that replaces the Cas6 function [94].	PDB: 3KG4; 3VZI; 3VZH Two domains of RRM (ferredoxin) fold, the C-terminal domain is deteriorated in many Cas5 protein of Type I; RAMP superfamily.
Cas6	Metal-independent endoribonuclease that generates crRNAs [61,74,83,121,123-125].	PDB: 2XLJ, 1WJ9,3I4H, 4C8Z, 4DZD Two domains of RRM (ferredoxin) fold, RAMP superfamily.
Cas7	Subunit of effector complexes binding crRNA [39,42,43,61,62]; often present in effector complexes in several copies.	PDB: 3PS0, 4N0L RRM (ferredoxin) fold with subdomains, RAMP superfamily.
Cas8abcef, (large subunit)	Subunit of effector complex, involved in PAM recognition [36-39,61].	PDB: 4AN8

Cas8abcef, (large subunit)	Subunit of effector complex, involved in PAM recognition [36-39,61].	PDB: 4AN8 belong to RRM (ferredoxin) fold; Zn finger containing domain and C-terminal alpha helical domain [79]; Fusion: HD nuclease domain.
Small subunit	Small, mostly alpha helical protein, subunit of effector complex [42,43,61,62,74,83,121,126].	PDB: 2ZCA (Cse2); 2ZOP, 2OEB (Cmr5); 3ZC4 (Csa5); Cse2 has two alpha helical bundle-like domains; Cmr5 has a domain matching N-terminal domain of Cse1 and Csa5 has a domain matching C-terminal domain of Cse2.
Cas9	In Type II CRISPR-Cas systems, Cas9 is sufficient both to generate crRNA and to cleave the target DNA [1,25], although it requires help of a house-keeping gene coding for RNase III and a special gene tracrRNA encoded in the respective CRISPR-*cas* locus [34]; Both the RuvC and HNH nuclease domains of Cas9 are involved in the cleavage of the target DNA [18,89]. Additionally, Cas9 contributes to adaptation, in particular by recognizing the PAM [54,55].	PDB: 4OGC, 4OO8, 4CMP Cas9 has several subdomains and adopt a two-lobed general structure. Beyond two catalytic nuclease domain its subdomains do not appear to be similar to other known protein structures [63,127].

Supplementary information S4 (table) | **The classes, types and subtypes of CRISPR-Cas systems, their signature proteins and key features**

Subtype	**Mono-phyletic in Cas1 tree**	**Signature proteins: Strong**/weak* (other name)	**Comment**
Class 1: multisubunit effector complexes			
Type I: Cascade effector complexes			
I-A	No	Cas8a, Csa5 (small subunit)	Cas3 is often split into the helicase Cas3' and HD nuclease Cas3'' and a separate gene for small subunit *csa5* is often present. Several distinct subfamilies of Cas8a exist.
I-B	No	Cas8b	I-B systems belong to several distinct clades on the Cas1 tree. Characterized only by gene composition: all loci have *cas5*, *cas7*, *cas8* and *cas6* genes. Usually the *cas3* gene is not split. Several distinct subfamilies of Cas8b exist.
I-C	No	Cas8c	These systems usually do not have a *cas6* gene. Cas5 is catalytically active and replaces Cas6 function.
I-D	No	**Cas10d (large subunit)**	The HD domain is associated with the large subunit rather than with Cas3 but lacks the circular permutation of the motifs like the HD domain fused with Cas10 in type III systems.
I-E	Yes	**Cse1 (Cas8e), Cse2 (small subunit)**	The *cas4* gene is not associated with this system.
I-F	Yes	**Csy1 (Cas8f), Csy2 (Cas5 group RAMP), Csy3 (Cas7 group RAMP), Cas6f**	The *cas4* gene is not associated with this system, *cas2* is fused to *cas3*. There is no separate gene for a small subunit, which is either missing or fused to the large subunit.
I-F variant 1	N/A	**Csy1/Csy2 (large subunit/Cas5f) fusion**	The *cas1-cas2-cas3* genes are not present. Usually three genes (*csy1/csy2* fusion, *csy3* and *cas6f*) are present in an operon, which is often found next to *tniQ*/*tnsD* family genes. These are potentially mobile effector complexes.
I-F variant 2	Yes	**PBPRB1993 (Cas5 group RAMP) PBPRB1992 (Cas7 group RAMP)**	A derived variant of I-F with two distinct genes of predicted group Cas5 (PBPRB1992) and group Cas7 (PBPRB1993) RAMPs.
I-U	No	**GSU0054**	These systems usually lack identifiable stand alone *cas6*. GSU0054 contains several specific insertions or fusions,

NEW		(Cas5 group RAMP)	including a region with limited similarity to Cas6. GSU0054 is likely to be catalytically active. Large subunits are highly variable and sometimes apparently missing. Cas3 contains a C-terminal HD domain.
Type III: Csm (III-A,D) and Cmr (III-B,C) effector complexes			
III-A	No	Csm2 (small subunit)	Also known as the Csm module and Cas10 usually has active catalytic motifs. The III-A loci typically contain several *cas7* group genes and is often linked to *csm6* which has CARF and C-terminal HEPN domain. Might be associated with *cas1-cas2* gene pairs of different origin.
III-B	No	Cmr5 (small subunit)	Also known as the Cmr (or RAMP) module. Cas10 often has active catalytic motifs. These systems are usually associated with several Cas7 group RAMPs and are rarely present in a genome as a stand-alone system They are usually not linked to *cas1-cas2* gene pair. Cmr1 has a duplication of RAMP domains both from the Cas7 group.
III-C **NEW**	No	MTH326 (Cas10 or Csx11)	Have small subunit and several Cas5 and Cas7 RAMP protein shared with Type III-B. The large subunit is often inactivated and some Cmr1 family proteins possess only one RAMP domain.
III-D **NEW**	No	**Csx10 (Cas5 group RAMP)** all1473 (uncharacterized component)	Have small subunit and several Cas5 and Cas7 RAMP protein shared with Type III-A. The signature gene is *csx10* which encodes a fusion of Cas5 and Cas7 proteins. Another specific gene *all1473* is likely to be a component of effector complex but is not similar to any known Cas proteins. The large subunit is often lacking the HD domain. Csx10 could be fused to the small subunit in some systems and Cas7 group RAMPs are often fused and have large insertions.
Putative type IV: Uncharacterized multisubunit effector complexes			
IV (putative) **NEW**	N/A	Csf1 (large subunit)	These systems possess a gene for a highly reduced large subunit Csf1. Some variants, in addition to the predicted large subunit, Cas7 and Cas5, encode a gene for a DinG-like helicase. Other variants often encode RHA1_ro10070-like proteins, which are putative small subunits of effector complexes. The latter systems are found mostly in Actinobacteria and often on plasmids.
Class 2: Single protein (multidomain) effector complexes			
Type II: Cas9 effector complexes			
II-A	Yes	Csn2 (helper protein)	Monophyletic group in Cas9 and Cas1 tree. There are four genes in these operons with *csn2* gene in addition to *cas1_2_9*. There are at least 5 distinct families of Csn2.
II-B	Yes	Cas9 (Csx12 subfamily)	Monophyletic group on Cas9 tree with four gene operons containing *cas4* in addition to *cas1_2_9*.

II-C	No	N/A	Only three genes are present in the II-C operon - *cas1_2_9*.
Putative type V: Cpf1 effector complexes			
V (putative) NEW	Yes	Cpf1	Cpf1 is a large protein which C-terminal region shares a significant similarity with TnpB of transposable element IS605. Contains RuvC-like nuclease. In several genomes *cpf1* is found as a stand-alone gene or in the context with other transposon-associated genes

Note: * **Strong**/weak – is the characteristic of the signature protein family with respect to subtype recognition/classification ability using the respective profile. Strong means that it has a relatively high specificity and high selectivity, i.e. is a reliable signature, whereas weak means that search for this family yields either a high level of false positives or false negatives, but nevertheless the family remains the best available signature for a particular subtype.

Supplementary information S8 (table) | **Distribution of different types and subtypes of CRISPR-Cas in archaeal and bacterial phyla**

	CAS-I-A	CAS-I-B	CAS-I-C	CAS-I-D	CAS-I-E	CAS-I-F	CAS-I-U	CAS-II-A	CAS-II-B	CAS-II-C	CAS-III-A	CAS-III-B	CAS-III-C	CAS-III-D	CAS-IV	CAS-V	part
A. Crenarchaeota	40.7	0	0	1	0	0	0	0	0	0	11	25.5	0	6.9	0	0	45.6
A. Euryarchaeota	9.2	72.1	5.4	18.5	12.4	0	2.7	0	0	0	22.2	4.6	7.7	3.2	0	4.3	21.8
A. Nanoarchaeota	0	6.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A. Thaumarchaeota	0	4.2	0	0	0	0	0	0	0	0	0	0	0	5.9	0	0	8.8
A. unclassified	0	0	0	2.3	0	0	0	0	0	0	0	0	0	0	0	0	0
B. Actinobacteria	0	1.4	7.8	0	44.8	0	3.5	8.6	0	2.7	1.2	0	0	9.2	7.2	0	22.2
B. Aquificae	0	9.1	0	0	0	0	0	0	0	0	4.2	0	2.8	0	0	0	3.3
B. Bacteroidetes/Chlorobi	0	22.4	12.9	0	5.8	0	0	0	0	25.1	12.9	11.7	2.3	4	0	0	26.4
B. Caldiserica	0	4.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B. Chlamydiae/Verrucomicrobia	0	0	0	0	0	0	0	0	0	4.8	0	0	0	5.3	0	0	5.3
B. Chloroflexi	4.4	12.5	6.6	6.7	8.8	0	0	0	0	0	15.1	12.8	5.5	11.3	0	0	9
B. Chrysiogenetes	0	0	0	0	0	3.7	0	0	0	0	3.7	0	0	0	0	0	0
B. Cyanobacteria	0.8	14.6	6.3	24	3.5	0	3.3	0	0	0	4	31.5	1	22.5	0	0	25.9
B. Deferribacteres	0	9.3	0	0	0	0	0	0	0	0	2.9	3.1	0	0	0	0	6
B. Deinococcus-Thermus	2	1	3.2	0	16	0	0.4	0	0	0	6.5	6.4	0	0	0	0	10.2
B. Dictyoglomi	0	3.5	0	0	0	0	0	0	0	0	1.7	0	0	0	0	0	0
B. Elusimicrobia	0	0	0	0	0	0	0	0	0	9.9	0	0	0	0	0	0	0
B. Fibrobacteres/Acidobacteria	0	4.4	3.8	0	0	0	13.3	0	0	5.7	3.4	8.9	0	0	0	0	5.7
B. Firmicutes	8.1	84.7	66.7	7.6	19.8	3.4	7.8	49.9	0	7.8	21.9	12.9	10.1	16.8	4.3	1.7	28.2
B. Fusobacteria	0	3.9	0	0	0	0	0	0.4	0	4.9	0	0	0	1.6	0	0	1.6
B. Nitrospirae	0	3.7	0	0	4.5	0	0	0	0	0	7.3	0	3.7	0	0	0	0
B. Planctomycetes	0	2.6	13	0	0	0	12.7	0	0	0	0	3.2	0	4.6	3.6	0	7
B. Proteobacteria	0	41.8	90.4	1	83.3	66.6	22.3	12.9	3.7	42.3	18.1	34	0	16.6	6.4	1.6	50.2
B. Spirochaetes	0	8.4	9.5	1.6	4.9	5.3	0	1.9	0	3.6	4.7	1.9	0	0	2.3	0	3
B. Synergistetes	0	6.7	0	0	0	0	0	0	0	0	0	2.2	0	0	0	0	10.1
B. Thermodesulfobacteria	0	0	0	0	0	0	0	0	0	0	2.7	2.4	0	0	0	0	2.7
B. Thermotogae	0	15.9	0	0	0	0	0	0	0	0	4.8	2.9	5.7	0	0	0	0
B. unclassified	0	7.9	0	0	3.1	0	4.2	0	0	0	4.8	3.1	0	0	0	0	4.8

Numbers and colors indicate the abundance of a system in a phylum (the sum of genome weights across single-unit loci; see Supplementary File 2 for details). "A" indicates archaeal phyla; "B" indicates bacterial phyla. The "part" column includes incomplete and ambiguously classified systems.

In format provided by Makarova *et al.* (NOVEMBER 2015)

Supplementary information S11 (table) | **Comparison of different classifications of CRISPR-Cas systems**

	Species tree	CRISPR-Cas subtypes	Cas1 tree	Repeats, structure (1)	Repeats, sequence (1)	Locus architecture	Interference genes similarity tree	Adaptation genes similarity tree
Species tree	x							
CRISPR-Cas subtypes	0.25	x						
Cas1 tree	0.12	0.65	x					
Repeats, structure (1)	0.30	0.73	0.64	x				
Repeats, sequence (1)	0.33	0.72	0.68	0.84	x			
Locus architecture	0.11	0.78	0.47	0.60	0.71	x		
Interference genes similarity tree	0.14	0.75	0.33	0.55	0.64	0.23	x	
Adaptation genes similarity tree	0.11	0.64	0.78	0.64	0.70	0.33	0.26	x

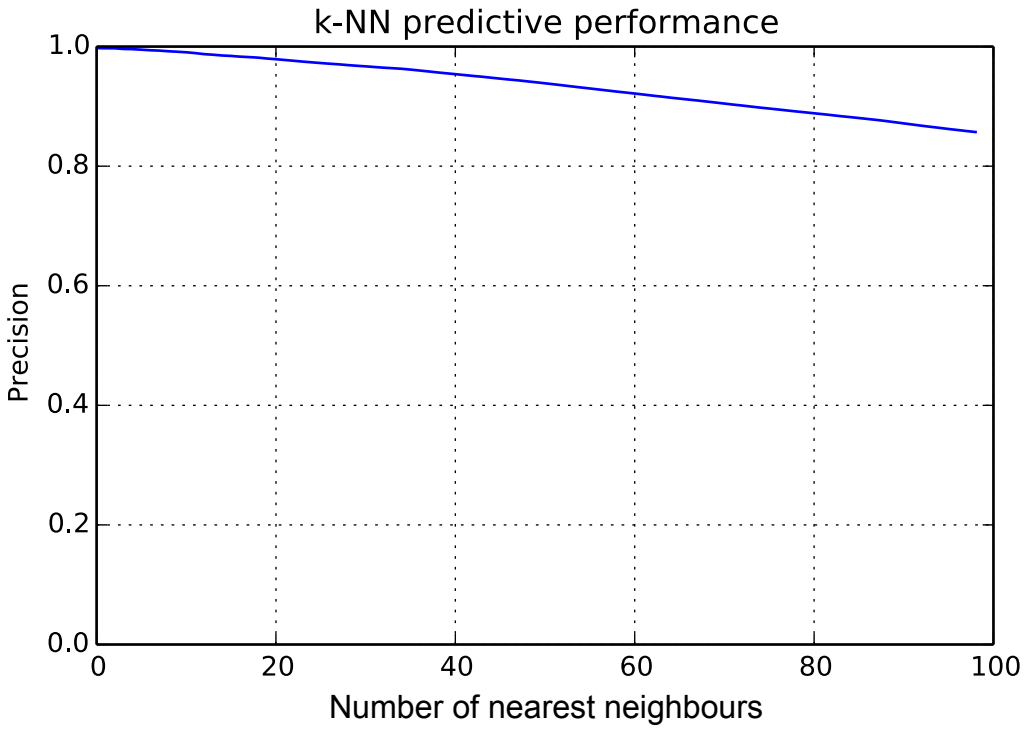
(1) - excluding unclassified repeats

Similarity between classifications of *cas* loci based on different features
Similarity was calculated using the following measures, depending on the nature of the classifications:

Category/Category	xxx	(Normalized Mutual Information)
Category/Tree	xxx	(adjusted Information Consistency Index)
Tree/Tree	xxx	(tree-induced distance Spearman correlation)

See the Supplementary File S2 for the details on the comparison methids

Supplementary information S15 (figure) | **Performance of an automated classifier for annotation of CRISPR–*cas* loci**



Methods

Genome weighting

The currently available collection of archaeal and bacterial genomes has a highly biased distribution of isolates across taxa. For example, it includes 46 strains of *Escherichia coli*, whereas entire phyla, such as Nanoarchaeota, Korarchaeota, Chrysiogenetes and others, are represented by a single genome. This extreme bias makes quantitative characterization of genomic features challenging and renders unusable most standard statistical methods that rely on random independent sampling as a null model. A relative genome weighting scheme that assigns low values to members of the densely sampled clades and high values to lone representatives of clades, can be used to mitigate the effects of the sampling bias.

Two notions are central to our model of relative genome weighting: first, closely related genomes should contribute individually less to the total clade weight than their more distant relatives; second, the relative contribution of the clades should reflect the number of independent evolutionary events that occurred in the history of the clade. Using the sum of branch lengths in a (sub)tree allows one to quantify both concepts.

Consider a node in a rooted phylogenetic tree that has several descendant clades, each with the sum of branch lengths (including the length of the branch connecting this subtree to the parent node) T_i. If the total weight assigned to this node is set to W, then it is distributed between the descendant subtrees as $W_i = WT_i/\Sigma T_i$. The sums of branch lengths for each internal tree node can be easily computed iteratively in the leaf-to-root direction and the total tree weight can be iteratively distributed between clades and leaves in the root-to-leaf direction.

To estimate the genome weights, we used an approximate phylogenetic tree reconstructed from concatenated alignments of ribosomal proteins [1] that was rooted between bacteria and archaea. The subtree encompassing the 1302 *cas*-positive genomes was extracted from the original tree. The weights calculated using this procedure are robust to minor perturbations of tree topology, especially those that involve deep clades and short internal branches.

CRISPR-*cas* loci identification

An exhaustive search for *cas* genes was performed within the set of protein sequences annotated in 2751 complete archaeal and bacterial genomes that were available at the NCBI as of February 1, 2014. The 185 multiple sequence alignments of Cas proteins that were not available through public databases were constructed and added to the ~29,500 CD, COG and PFAM profiles in the NCBI CDD database [2]. Altogether, 395 profiles represented 93 distinct Cas protein families. Searches were performed using PSI-BLAST [3], with the alignment consensus employed as the master query.

The 93 *cas* genes were classified by sequence similarity into 35 families that belong to 12 distinct functional classes according to the functions of the respective proteins in CRISPR-Cas systems (Supplementary File 1). Of the 35 families of cas genes, 11 constitute the *cas* core and the rest are classified as "ancillary".

The *cas* loci were identified using a two-step procedure. In the first step, PSI-BLAST search results with e-value threshold of 10^{-6} were used to annotate all proteins in the set of complete archaeal and bacterial genomes. The highest-scoring profile for all non-overlapping sequence segments were identified. In the second step, gene products from neighborhoods of ±20 genes around all identified *cas* genes were used as queries for the second round of PSI-BLAST search with e-value threshold of 0.01. Additional genes with moderately significant matches to Cas profiles and located in the vicinity of confidently predicted *cas* genes were identified.

Gene neighborhoods of ±5 genes around all identified *cas* genes were extracted; overlapping neighborhoods were merged and trimmed to the first and the last *cas* gene, to form the candidate loci. A locus that contains at least two *cas* genes, of which at least one gene belongs to the *cas* core, was identified as a valid *cas* locus.

Profile-based CRISPR-*cas* loci classification

A set of Cas sequence profiles was collected over the years since the previous publication on CRISPR-Cas classification [4]. Correspondence between the profiles, gene names and CRISPR-Cas system types and subtypes was reexamined in the course of this work. To assist the assembly of a non-redundant and self-consistent set of Cas protein profiles, the multiple profiles for Cas5, Cas7 and Large Subunit were aligned to each other using HMMER 3.0 [5] and cluster dendrograms were constructed from matrices of relative pairwise scores using UPGMA. The dendrograms were examined for inconsistent annotation of similar profiles; potential discrepancies were investigated on a case by case basis, and annotation was adjusted where required.

Loci were classified using the correspondence table between Cas sequence profiles and CRISPR-Cas (sub)types (Supplementary File 1). The classification procedure consisted of two steps. First, a gene group annotation was used to identify genes of the effector module (*cas5*-like, *cas7*-like and Large Subunit), *cas9* and *cpf1*. A genomic segment containing either each of the major effector module genes or one *cas9* gene or one *cpf1* gene was considered a complete CRISPR-Cas system unit of type I/III/IV, type II or type V, respectively. Loci that contained neither the full complement of effector module genes nor *cas9* or *cpf1* were classified as partial.

At the second step, each locus unit or single-unit locus was analyzed separately. Each Cas profile within the unit contributed a "vote" for the type and subtype that this profile corresponded to. Contributions from profiles with multiple affinities (such as, for example, *cas5* pfam09704 profile that does not discriminate between subtypes of type I) were equally divided between the corresponding (sub)types. The "votes" were tallied across the unit; if the dominant (sub)type accounted for at least 2/3 of the total, the locus (unit) was assigned to the respective

(sub)type. If no type or subtype received the qualified majority, the locus or unit was considered to be ambiguously classified.

Sequence-based phylogenetic reconstruction

Multiple sequence alignments were constructed using a combination of MUSCLE [6] to align closely related sequences and MAFFT [7] to merge these alignments. Sites with of gap character fraction >0.5 and homogeneity <0.1 [8] were removed from the alignment. Phylogenetic trees were reconstructed using the FastTree program [9] with the WAG evolutionary model and the discrete gamma model with 20 rate categories. The same program was used for bootstrap value calculation.

For the phylogenetic analysis of the Cas1 family, 1418 Cas1 protein sequences were used. A filtered Cas1 alignment (306 positions) was used for tree reconstruction. For the phylogenetic analysis of the Cas3 family, 1093 protein sequences were used, and the filtered Cas3 alignment for tree reconstruction included 283 positions. For the Cas10 family, three alignments for three distinct families were constructed and used for phylogenetic analysis. These families consisted of 443, 36 and 10 protein sequences, and the respective filtered alignments included 427, 910 and 652 positions.

Classification comparison and information consistency index

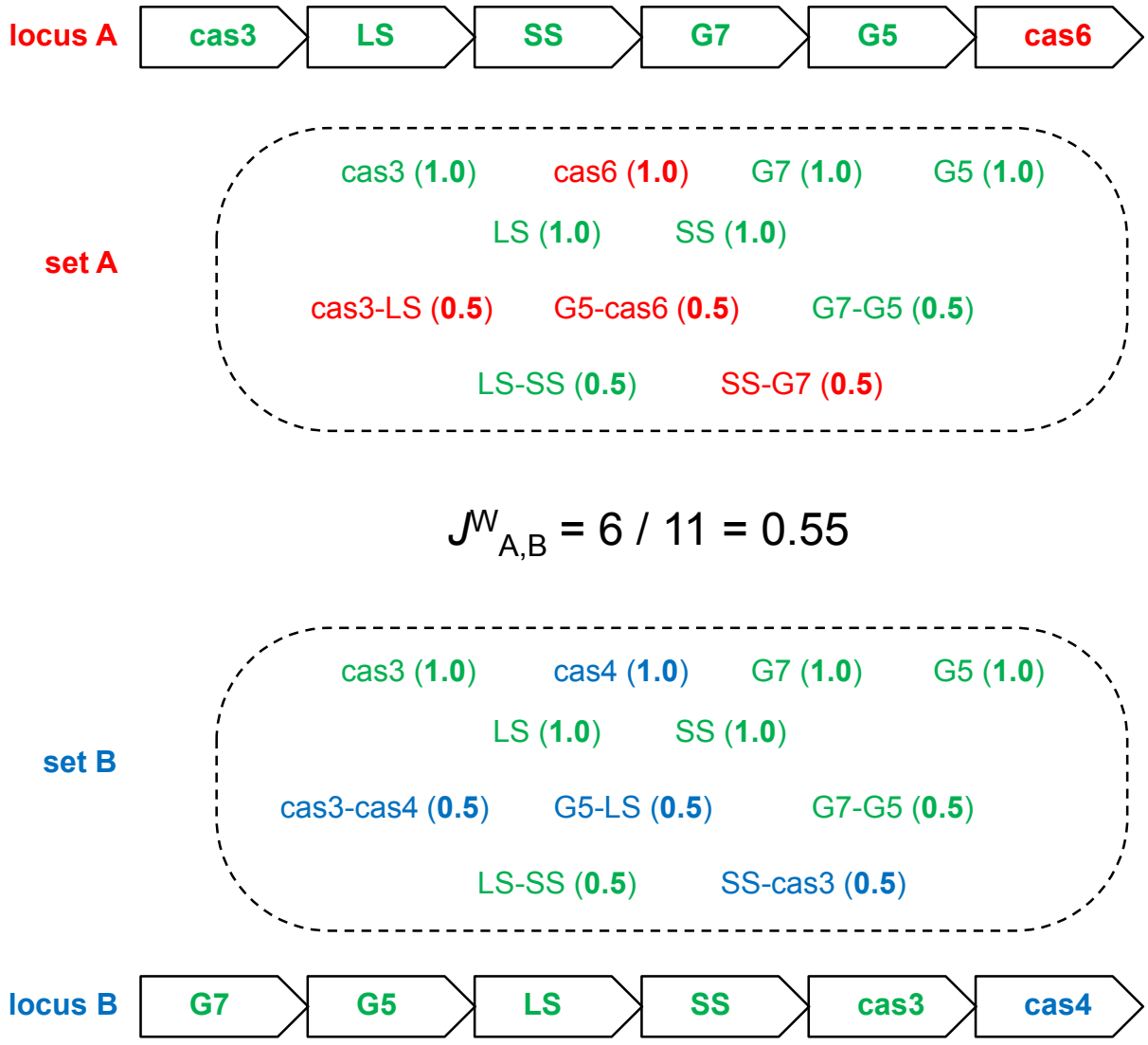
To compare classifications of CRISPR-*cas* loci based on different criteria (e.g. according to the CRISPR-Cas subtypes or according to the sequence classification of repeats in the adjacent CRISPR cassette), we used the Normalized Mutual Information index (mutual information divided by the geometric mean of the entropies of both classifications) [10].

To compare different trees reconstructed for the same set of leaves, the distances between the leaves along the tree branches were computed by summing the branch lengths along the path, connecting the leaves; then, the Spearman rank correlation coefficients between the distances induced by the two trees were calculated.

To quantify the fit between the tree structure and classification of the leaves, the following procedure was used. Within each clade of a rooted tree, the clade entropy was calculated from the distribution of its descendant leaves across the classes. Weighted average of clade entropies was calculated across the tree using clade weights, producing the tree-wide estimate of the classification entropy E_T. Then, the tree labels were scrambled, and the procedure was repeated 10 times to estimate the expectation of the tree entropy for the random labeling, E_R. The information consistency index then is calculated as $1\text{-}\min(E_T,E_R)/E_R$. A perfect tree that segregates at the root into clades corresponding to pure classes has E_T equal to zero, and therefore, has an information consistency index of 1. The tree with the entropy as high as that of a tree with random leaf labeling (or higher) has the information consistency index of 0.

Locus architecture dendrogram

To compare the architectures of the *cas* loci, the following procedure was developed. First, the gene order in the loci encoded in the negative strand was inverted. All non-*cas* genes were removed; *cas* genes were classified according to the family classification (Supplementary File 1). Each locus was encoded as a set of weighted components in the following way: all individual genes were included in the set with weights of 1; all ordered pairs of adjacent genes were included in the set with weights of 1/2 (see figure below). The weighted Jaccard similarity index $J^{W}_{A,B}$ for the component sets of loci ***A*** and ***B*** was computed as the sum of weights of the intersection of sets ***A*** and ***B*** divided by the sum of weights of the union of sets ***A*** and ***B***. The distance between the loci ***A*** and ***B*** was computed as $-\ln(J^{W}_{A,B})$. The figure below shows an example of the weighted Jaccard similarity index calculation.



The loci architecture similarity dendrogram was constructed from the pairwise loci distance matrix using the UPGMA method.

Locus sequence similarity dendrogram

In order to automatically group *cas* loci, we introduce a clustering approach based on protein similarity. Given a *cas* locus, as defined in the Section *CRISPR-cas loci identification*, we select the interference proteins for Type I and Type III and Cas9 for Type II. Becuase some *cas* loci contain multiple effector modules of different types, the effector proteins of each locus were separated according to their types. For each pair of proteins p_i and p_j (belonging to different *cas* loci), the FASTA [11] protein sequence similarity score $S(p_i, p_j)$ was computed. To guarantee appropriate metric properties, the similarity was symmetrized to $S^*(p_i, p_j) = \left(S(p_i, p_j) + S(p_j, p_i)\right)/2$, and the score was normalized to

$$\hat{S}(p_i, p_j) = S^*(p_i, p_j) / \sqrt{S^*(p_i, p_i) S^*(p_j, p_j)}.$$

The similarity between two *cas* loci L_m and L_n is then defined as the average pairwise similarity between all possible protein pairings:

$$S_L(L_m, L_n) = \frac{1}{|L_m||L_n|} \sum_{p_i \in L_m} \sum_{p_j \in L_n} \hat{S}(p_i, p_j).$$

Finally, the dendrogram was generated by manually rooting the unrooted tree obtained by Rapid Neighbor-Joining [12] on the derived pairwise distance, $D_L = 1 - S_L$. The software and instructions for clustering CRISPR-cas loci by protein sequence similarity and automatic subtype assignment are available from http://www.bioinf.uni-freiburg.de/Supplements/NRMmicro_Koonin_2015/.

References

1 Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS One* **7**, e36972, (2012).
2 Marchler-Bauer, A. *et al.* CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**, D205-210 (2009).
3 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
4 Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477, (2011).
5 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, (2011).
6 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
7 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, (2013).
8 Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I. & Koonin, E. V. The deep archaeal roots of eukaryotes. *Mol Biol Evol* **25**, 1619-1630 (2008).
9 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, (2010).
10 Strehl, A. & Ghosh, J. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *The Journal of Machine Learning Research* **3**, 583-617 (2002).
11 Pearson, W. Finding protein and nucleotide similarities with FASTA. Ch. 3, Unit 3.9 (2004).
12 Simonsen. M & Pedersen. Rapid computation of distance estimators from nucleotide and amino acid alignment. C.N.S. in *SAC2011* (2011).

Supplemental Material for:

Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs

Viktoria Reimann[1*], Omer S. Alkhnbashi[2*], Sita J. Saunders[2], Ingeborg Scholz[1], Stephanie Hein[1], Rolf Backofen[2,3,4,§] and Wolfgang R. Hess[1,3,§]

[1]Genetics and Experimental Bioinformatics group, Faculty of Biology, University of Freiburg, Schänzlestr. 1, 79104 Freiburg, Germany;

[2]Bioinformatics group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany;

[3]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Habsburgerstr. 49, D-79104 Freiburg, Germany;

[4]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schänzlestr. 18, D-79104 Freiburg, Germany;

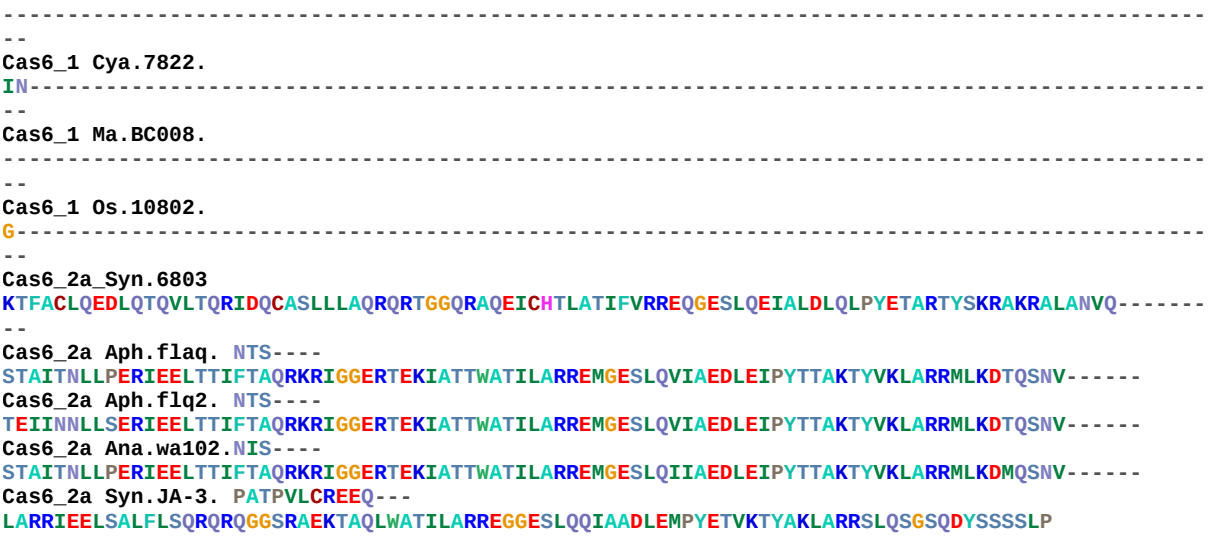
*Corresponding author: University of Freiburg, Faculty of Biology, Schänzlestr. 1, D-79104 Freiburg, Germany; Tel. +49-(0)761-203-2796; FAX: +49-(0)761-203-2745

Email: wolfgang.hess@biologie.uni-freiburg.de

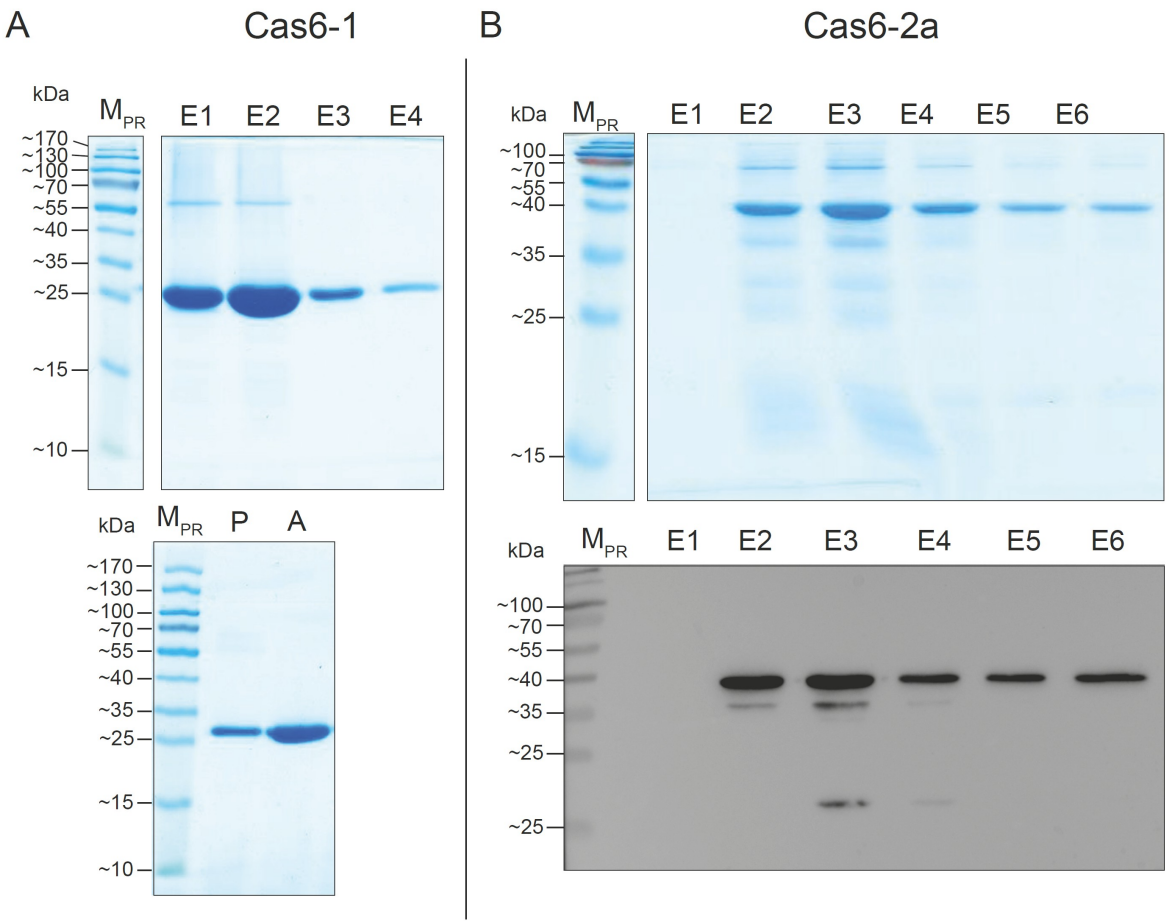
*Co-sharing first authors ; §co-sharing authors for correspondence

```
                         10        20        30        40        50        60        70
80        90        100
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
...|....|....|....|....|
Cas6_1_Syn.6803  --------------------MFDDRYSLYSVVIELGAAKKGFPTGILGRALHSQVLEWLKIGEPSLAEELHQSQ-
ISPFSISPLIGKRRSKLTEEGDRFF
Cas6_1 Syn.7002. --------------------MPNDPYSLYSIVIELGAAEKGFPTGILGRSLHSQVLQWFKQDNPFLATELHQSQ-
ISPFSISPLMGKRHAKLTKAGDRLF
Cas6_1 Cya.7822. -------------MVQDIWQMSNHSYTLQSMVIELGAAAKGKPPATLGRAIHAQVLEWINLGNSELANQVHESQ-
IAPLSLSGLLGNRRPKGTEAGDNFY
Cas6_1 Ma.BC008. -------------MEQKISSQPDKSSNLHSLVVELAAAERGKFSPTLSRALHALVLRWLHLGSPEVAEFVHNTQ-
```

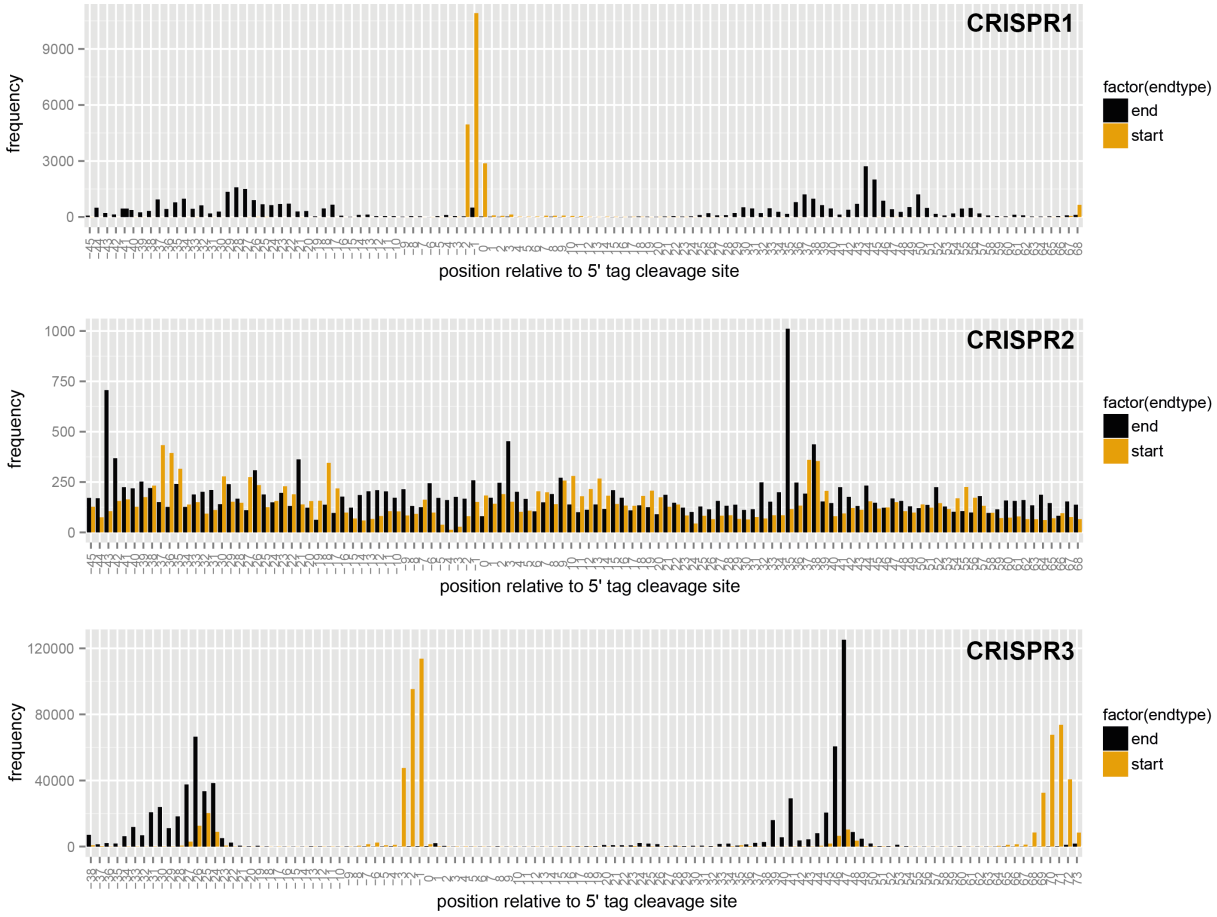
```
TSPIGISRLLGNRRPKGTQPGDIFY
Cas6_1 Os.10802.  ------------MEPAMSLQYPRYS-LQSLVVELGAANQGMPPSTLGRALHALVLNWLQRGNPELAEAVHDSQ-
ESPLSISGLVGERRPKGTQAGDEFF
Cas6_2a_Syn.6803  --------------MVD--
LKSLAGAEMVGLRWQLRFDRPCRLESHYVKGLHAWFLHQVQAIDPDVSAWLHDGQGEKPFTISRLIGPTLWQEGHWHWQIN
Cas6_2a Aph.flaq.
-MSELATSSKHKNPSKNTPLIWADDTELVSLVFELEVTDSTALYSQYTIGLHAWFLDQVRQINPTLSAYLHDSESEKPFSISALEGQLLPTGKQ
LQLQAN
Cas6_2a Aph.flq2.
-MSELATSSKHKNPSKNTPLIWADDTELVSLVFELEVTDSTALYSQYTIGLHAWFLDQVRQINPTLSAYLHDSESEKPFSISALEGQLLPTGKQ
LQLQAN
Cas6_2a Ana.wa102.-
MSELATSSNRKNPSKNTPLIWADDTELVSLVFDLEVTDSTALYSQYTIGLHAWFLDQVRQINPTLSAYLHDGESEKPFSISALEGQLLPTGKQL
QLQSN
Cas6_2a Syn.JA-3. MLTKLSFSEPPVAGAEN--
SKWPAGSELVGIALEVQAPQPYLLDPHYAKGLHAWFLSQVQETDPQLSAYLHDGESEKPFTLSRLMGPFREQGGRLLIPPQ

                          110       120       130       140       150       160       170
180       190       200
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
...|....|....|....|....|
Cas6_1_Syn.6803   FRISLLNGSLLQPLLKGLEQQDKQIVMLDKFAFRLCHIHILPGSHSLARASSYALTTQA-
PTSSKITLKFHSATSFKIDRNTIQPFPLGDSVFNSLLRRW
Cas6_1 Syn.7002.  FRICLLRGDLLQPLLNGIEQTVNQSVCLDKFRFRLCQTHILPGSHPLAGASHYSLISQT-
PVSSKITLDFKSSTSFKVDRKIIQVFPLGEHVFNSLLRRW
Cas6_1 Cya.7822.
FRIGLLNGNLIEPLLKGIEIWGTQPLIFANFPFVLRNYYTLPGTHRLAGATNYYVLTQTVQEQRKITLKFLSPTSFK-
QQKSIQLFPLSDLVFGSLQRRW
Cas6_1 Ma.BC008.  FRISLLDGNLIAPLLKGLEQQERQPVILAKFSFIIRNIYTLPGTHTLVEATNYGAISKN-
SNSTDIQLTFFSPTSFK-QLQNIQTFPLPELVFNSLHRRW
Cas6_1 Os.10802.
FRIGVLDGRLISPLLSGLEHWGASPISLGKFPFAIRSVYAMPGTHPWAGSSDYYALANAPALSNDVTLSFLSPTSFK-
QKQGVQPFPLTELVFDGLRRRW
Cas6_2a_Syn.6803  KTYHWQLNLLSGALIEALQPWLAR---
LPNKIVLARQTLWVEAVDCYLAPHNYQQLWPQGALPRRQEFTFTSPTSFR-RQGNHYPLPEPRNVLQSYLRRW
Cas6_2a Aph.flaq. YRYRWQINAISQPVVQFLSQWLTQ---
PPTTLKLRDACLQVKQISIVNPPTTYNQLLQSSINHRNINLSFISPTSFR-RKGHHFPLPVPFNLFHSYLRRW
Cas6_2a Aph.flq2. YRYRWQINAISQPVVQFLSQWLTQ---
PPTTLKLRDAYLQIKQISIVNPPTTYNQLLQSFINHRNINLSFISPTSFR-RKGHHFPLPVPFNLFHSYLRRW
Cas6_2a Ana.wa102.QIYRWQINAISQPVVQFLSQWLIQ---
PPTTLKLRDACLQVKQISIVNPPTTYNKLLQSSINHKNINLSFISPTSFR-RKGHHFPLPVPFNLFHSYLRRW
Cas6_2a Syn.JA-3. IPFRWSITALNPQVVEWLREWCRR---
LPPWLELRGSPLQILGWKVSAPPRTYRQLLEQPLSPRSWSLSFVSPTSFR-HRGHHLPLPIPRNLFHSYLRRW

                          210       220       230       240       250       260       270
280       290       300
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
...|....|....|....|....|
Cas6_1_Syn.6803   NHFA-----PEELYFPSVSWQIPVAAFELKTYSVQLKK--SEIGSEGWVTYLFPDQEQ------
AKIASVLSQFAFFAGVGRKTSMGMGQVSVNNHG---
Cas6_1 Syn.7002.  NNFA-----PEDLHFSQVDWSIPIAAFDVKTIPIHLKK--VEIGAQGWVTYIFPNTEQ------
AKIASVLSEFAFFSGVGRKTTMGMGQVQVRS-----
Cas6_1 Cya.7822.  NEFA-----PENLQIPALTWDALVTAYELKTYALKLESG-AEIGSQGWISYEFNSSEQ------
AKYASVLSQFAFFAGVGRKTTMGMGQVQLTVNNDIT
Cas6_1 Ma.BC008.  NHFA-----PNEYKFPSVEWMGLVSAYELKTHALRMEGG-AEIGAQGWVRYRFRDEEQ------
AKIASVLAHFAFFSGVGRKTSMGMGQTRIGG-----
Cas6_1 Os.10802.  NAFA-----PEELLFPAVEWNGLVSAFELQTHALKMKGG-AEIGAQGWVRYRFPDAEQ------
AKVAAVLAHFAFFSGVGRKTAMGMGQVKLLAESSSG
Cas6_2a_Syn.6803
NDFSGLAFEPEPFLDYWVPQNVVIDRHWLESVKTTAGKQGSVVGFVGAVSLVLTPQARNDGDDYGRLFHALCRYGPYCGTGHKTTFGLGQTMAG
WATPDL
Cas6_2a Aph.flaq. NDFSGMPIEQEAFLE-WIDENVIIHKHCLESVKVAAGKRGSVTGFTGAISLGLTKTALDN-
PEFTQLFDALVQLAPYCGTGHKTTFGLGQTCLNWVNPES
Cas6_2a Aph.flq2. NDFSGMPIEQEAFLE-WIDENVIIHKHCLESVKVAAGKRGSVTGFTGAISLGLTKTALDN-
PEFTQLFDALVQLAPYCGTGHKTTFGLGQTCLNWVNPES
Cas6_2a Ana.wa102.NDFSGMPIEQDGFLE-WIDENVIIHKHRLESVKVAAGKRGSVTGFTGAISLGLTKTALNN-
IEFTQLFYALVQLAPYCGTGHKTTFGLGQTCLDWVNPES
Cas6_2a Syn.JA-3. NDFSGLPIEAEPFLD-WVDGEVIIQRHRLESVKTTAGRQGSVTGFIGCVQLAVSSRAP---
ELLQQQLQALIHLAPYCGTGHKTPFGLGQTRLGWLAEEL

                          310       320       330       340       350       360       370
380       390
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
...|....|....|....|.
Cas6_1_Syn.6803
--------------------------------------------------------------------------------
--
Cas6_1 Syn.7002.
```



Supplemental Figure S1. Multiple sequence alignment Cas6-1 and Cas6-2a of *Synechocystis* sp. PCC 6803 with the 4 best matches found for each protein in Genbank using blastP. These top-matching proteins are all from the phylum cyanobacteria. For Cas6-1 these homologs are from: *Synechococcus* sp. PCC 7002 (Syn.7002), *Cyanothece* sp. PCC 7822 (Cya.7822), *Mastigocoleus testarum* BC008 (Ma.BC008), *Oscillatoria* sp. PCC 10802 (Os.10802). The homologs to Cas6-2a are from: *Aphanizomenon flos-aquae* 2012/KM1/D3 (Aph.flaq.), *Aphanizomenon flos-aquae* (Aph.flq2), *Anabaena* sp. wa102 (Ana.wa102) and *Synechococcus* sp. JA-3-3Ab (Syn. JA-3). The Cas6 motif (consensus GhGxxxxGhG, where h is hydrophobic and xxxxx has at least one lysine or arginine (1, 2)), conserved among all 10 compared sequences, is boxed.



Supplemental Figure S2. SDS-PAGE of purified recombinant proteins Cas6-1 and Cas6-2a of *Synechocystis* sp. PCC 6803 expressed in *E. coli*. **(A)** Cas6-1 (gene *slr7014*) was purified after cloning in pQE70 and expression in *E. coli* M15[pREP4] containing a C-terminal His_6-tag. 16 µl of each elution fraction were applied on the gel. The bottom part shows 1 µl of purified Cas6-1 before (P) and after (A) concentration of elution fractions 1 and 2 with Amicon Ultra Centrifugal Filter Units. **(B)** Cas6-2a (gene *slr7068*) was expressed in *E. coli* Rosetta(DE3)pLysS containing an N-terminal *Strep*-tag® II after cloning in pASK-IBA7plus. 16 µl of each elution fraction were separated on the gel. The lower part shows detection of the N-terminal *Strep*-tag® II with *Strep*-Tactin®-HRP conjugate after Western Blotting. Two of the lower molecular weight bands were recognized by the antiserum and consequently represent a small amount of C-terminally truncated Cas6-2a. The molecular weights of the recombinant proteins are 30.5 kDa (Cas6-1) or 45 kDa (Cas6-2a) and were calculated with the ExPASy Compute pI/MW tool. M_{PR} shows a prestained protein molecular weight marker (Thermo Scientific™ PageRuler™).



Supplemental Figure S3. The effects of a knockout of gene *slr7068* encoding Cas6-2a on the accumulation of reads from the CRISPR loci 1, 2 and 3. Whereas the accumulation of CRISPR1 and CRISPR3-derived crRNAs is unaffected (compare to the analysis of wildtype RNA-seq data in Figure 3 in Scholz *et al*. (3), no specific 5' or 3' ends can be detected for CRISPR2-derived transcripts.

Supplemental Table S1. Oligonucleotides used in this study. Restriction sites used for cloning are underlined. T7-Promotor sequences are highlighted in bold. All oligonucleotides were ordered from SIGMA-ALDRICH®. RNA oligonucleotides C1 and C2 were ordered HPLC purified.

DNA oligonucleotides		**product**
Heterologous overexpression		
1	CATGCATGCTTGATGATCGCTACAGTTTGTATTCC	Cas6-1
2	CATGGATCCTCCATGATTGTTAACTGAAACCTGTCC	Cas6-1
3	GATGCCGAGCTCGTGGTGGATCTAAAATCCTTAGCTGG	Cas6-2a
4	GATGCCCTGCAGTTATTGAACATTGGCTAAGGCCCGCTTAGC	Cas6-2a
Template synthesis for *in vitro* transcription		
5	**TAATACGACTCACTATAGG**AGAAAGAACCTCAGAACTTTCCTTC	CRISPR1
6	TTGCCTGGGATGGCAAGTTTCAG	CRISPR1
7	**TAATACGACTCACTATAGG**ACGATTGTTGTGCCCCTGGCGGTCG	CRISPR1*
8	AGCTGAAGCAAAATCTGTGTTTCAG	CRISPR1*
9	**TAATACGACTCACTATAGG**CCCCGCCCCCGTGGTGGGAG	CRISPR2
10	CTCGGACAGGAGGACAACTCG	CRISPR2
11	**TAATACGACTCACTATAGG**AGCGCACTGCCTGTCATTACTATTAG	CRISPR3
12	AGACATCACCACGAAATCCACTTGG	CRISPR3
13	**TAATACGACTCACTATAGG**TAATCCTAATTAGGTTTGAGTTAG	CRISPR1 I, II, III, IV, V
14	CTAGCCTTTGGTACTAGCTG	CRISPR1 I
15	CCACATTCATCGCTACAGAC	CRISPR1 II
16	TTCTGAGGTTCTTTCTAAAATTCTTC	CRISPR1 III
17	CTGATAGATCGTAGCGGAATG	CRISPR1 IV
18	**TAATACGACTCACTATAGG**ACAGATTTTGCTTCAGCTAGTAC	CRISPR1 VI
19	CCACCACCACCACCACGATTAATC	CRISPR1 V, VI, VII, VIII, IX
20	**TAATACGACTCACTATAGG**TCGTAACTACCTTCTAAGTCTG	CRISPR1 VII
21	**TAATACGACTCACTATAGG**CAAAATATAGGGAAGAATTTTAGAAAG	CRISPR1 VIII
22	**TAATACGACTCACTATAGG**AACTAAGGCATCCCATAGCATTC	CRISPR1 IX
23	**TAATACGACTCACTATAGG**CGGGGCTTGGGGGGTTGGAGTCC	CRISPR2 I, II, III, IV, V
24	TTATGAGTCCAACAGCCATTAGGAGG	CRISPR2 I
25	TGCCTAAAACACTCATCGAGAACTAC	CRISPR2 II
26	CCGCTCATCACCTTTAGGGCTGG	CRISPR2 III
27	TTGTCGAGCTTAGTAGTGTGG	CRISPR2 IV
28	**TAATACGACTCACTATAGG**TGTGAGTTGCATAATGCCTCCTAATGG	CRISPR2 VI
29	GTGGGGCCTCGGACAGGAGG	CRISPR2 V, VI, VII, VIII, IX
30	**TAATACGACTCACTATAGG**CTTGGTATTTGTAGTTCTCGATGAGTG	CRISPR2 VII
31	**TAATACGACTCACTATAGG**TGATAACGGGATGCCAGCCCTAAAGG	CRISPR2 VIII
32	**TAATACGACTCACTATAGG**CGTTATCCGGCAAAGAAACCACAC	CRISPR2 IX
RNA oligonucleotides		**ID**
33	CUUUCCUUCUACUAAUCCCGGCGAUCGGGACUGAAAC	C1
34	GUUCAACACCCUCUUUUCCCCGUCAGGGGACUGAAAC	C2

Supplemental Table S2. Sequences of *in vitro* transcripts used in this study. The two additional G nucleotides at the 5' ends of the transcripts originate from the sequence of the T7 promoter. The repeat sequences are italicized and determined cleavage sites **(3)** are indicated by diagonal slashes. Calculated lengths of fragments and intermediates upon cleavage at the indicated cleavage sites are given in the last column in nucleotides. The length of the full length *in vitro* transcripts is represented in nucleotides (nt) as part of the ID.

CRISPR1		
ID	**Sequence**	**Possible fragment lengths [nt]**
CRISPR1_364nt	GGAGAAAGAACCUCAGAA*CUUUCCUUCUACUAAU CCCGGCGAUCGGG/ACUGAAAC*AACUAAGGCAUC CCAUAGCAUUCCGCUACGAUCUAUCAG*CUUUCC UUCUACUAAUCCCGGCGAUCGGG/ACUGAAAC*UA AUAAUUGCUGAUUAAUCGUGGUGGUGGUGGUGG *CUUUCCUUCUACUAAUCCCGGCGAUCGGG/ACUG AAAC*GAUAUAUGGCUAAAUAUUGCUCAAAAGAUU UUAAUA*CUUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*AUUAGAUUGGUCGUGUUUUGAUU AACGGUGCUAGC*CUUUCCUUCUACUAAUCCCGG CGAUCGGG/ACUGAAAC*UUGCCAUCCCAGGCAA	364, 340, 317, 293, 268, 241, 221, 217, 195, 169, 148, 145, 123, 96, 76, 73, 72, 47, 24
CRISPR1*_168nt	GGACGAUUGUUGUGCCCCUGGCGGUCGCUUUCA AUGCCUC*UUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*UAAUCCUAAUUAGGUUUGAGUUA GUAUCUAGUGCCAU*CUUUCCUUCUACUAAUCCC GGCGAUCGGG/ACUGAAAC*ACAGAUUUUGCUUCA GCU	168, 142, 100, 74, 68, 26
CRISPR1_I_109nt	GGUAAUCCUAAUUAGGUUUGAGUUAGUAUCUAG UGCCAU*CUUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*ACAGAUUUUGCUUCAGCUAGUAC CAAAGGCUAG	109, 68, 41
CRISPR1_II_183nt	GGUAAUCCUAAUUAGGUUUGAGUUAGUAUCUAG UGCCAU*CUUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*ACAGAUUUUGCUUCAGCUAGUAC CAAAGGCUAG*CUUUCCUUCUACUAAUCCCGGCG AUCGGG/ACUGAAAC*UCGUAACUACCUUCUAAGU CUGUAGCGAUGAAUGUGG	183, 138, 115, 70, 68, 45
CRISPR1_III_257nt	GGUAAUCCUAAUUAGGUUUGAGUUAGUAUCUAG UGCCAU*CUUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*ACAGAUUUUGCUUCAGCUAGUAC CAAAGGCUAG*CUUUCCUUCUACUAAUCCCGGCG AUCGGG/ACUGAAAC*UCGUAACUACCUUCUAAGU CUGUAGCGAUGAAUGUGG*CUUUCCUUCUACUAA UCCCGGCGAUCGGG/ACUGAAAC*CAAAAUAUAGG GAAGAAUUUUAGAAAGAACCUCAGAA	257, 212, 189, 144, 138, 119, 74, 70, 68, 45
CRISPR1_IV_333nt	GGUAAUCCUAAUUAGGUUUGAGUUAGUAUCUAG UGCCAU*CUUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*ACAGAUUUUGCUUCAGCUAGUAC CAAAGGCUAG*CUUUCCUUCUACUAAUCCCGGCG AUCGGG/ACUGAAAC*UCGUAACUACCUUCUAAGU CUGUAGCGAUGAAUGUGG*CUUUCCUUCUACUAA UCCCGGCGAUCGGG/ACUGAAAC*CAAAAUAUAGG GAAGAAUUUUAGAAAGAACCUCAGAA*CUUUCCUU CUACUAAUCCCGGCGAUCGGG/ACUGAAAC*AACU AAGGCAUCCCAUAGCAUUCCGCUACGAUCUAUCA G	333, 286, 265, 218, 212, 195, 148, 144, 138, 121, 74, 74, 70, 68, 47
CRISPR1_V_405nt	GGUAAUCCUAAUUAGGUUUGAGUUAGUAUCUAG UGCCAU*CUUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*ACAGAUUUUGCUUCAGCUAGUAC CAAAGGCUAG*CUUUCCUUCUACUAAUCCCGGCG AUCGGG/ACUGAAAC*UCGUAACUACCUUCUAAGU CUGUAGCGAUGAAUGUGG*CUUUCCUUCUACUAA UCCCGGCGAUCGGG/ACUGAAAC*CAAAAUAUAGG GAAGAAUUUUAGAAAGAACCUCAGAA*CUUUCCUU CUACUAAUCCCGGCGAUCGGG/ACUGAAAC*AACU AAGGCAUCCCAUAGCAUUCCGCUACGAUCUAUCA G*CUUUCCUUCUACUAAUCCCGGCGAUCGGG/ACU*	405, 362, 337, 294, 286, 267, 224, 218, 212, 193, 150, 148, 144, 138, 119, 76, 74, 74, 70, 68, 43

	*GAAAC*UAAUAAUUGCUGAUUAAUCGUGGUGGUG GUGGUGG	

CRISPR1_VI_331nt	GGACAGAUUUUGCUUCAGCUAGUACCAAAGGCU AG*CUUUCCUUCUACUAAUCCCGGCGAUCGGG/AC UGAAAC*UCGUAACUACCUUCUAAGUCUGUAGCGA UGAAUGUGG*CUUUCCUUCUACUAAUCCCGGCGA UCGGG/ACUGAAAC*CAAAAUAUAGGGAAGAAUUU UAGAAAGAACCUCAGAA*CUUUCCUUCUACUAAUC CCGGCGAUCGGG/ACUGAAAC*AACUAAGGCAUCC CAUAGCAUUCCGCUACGAUCUAUCAG*CUUUCCU UCUACUAAUCCCGGCGAUCGGG/ACUGAAAC*UAA UAAUUGCUGAUUAAUCGUGGUGGUGGUGGUGG	331, 288, 267, 224, 212, 193, 150, 148, 138, 119, 76, 74, 74, 64, 43
CRISPR1_VII_261nt	GGUCGUAACUACCUUCUAAGUCUGUAGCGAUGA AUGUGG*CUUUCCUUCUACUAAUCCCGGCGAUCG GG/ACUGAAAC*CAAAAUAUAGGGAAGAAUUUUAGA AAGAACCUCAGAA*CUUUCCUUCUACUAAUCCCGG CGAUCGGG/ACUGAAAC*AACUAAGGCAUCCCAUA GCAUUCCGCUACGAUCUAUCAG*CUUUCCUUCUA CUAAUCCCGGCGAUCGGG/ACUGAAAC*UAAUAAU UGCUGAUUAAUCGUGGUGGUGGUGGUGG	261, 218, 193, 150, 142, 119, 76, 74, 68, 43
CRISPR1_VIII_187nt	GGCAAAAUAUAGGGAAGAAUUUUAGAAAGAACCU CAGAA*CUUUCCUUCUACUAAUCCCGGCGAUCGG G/ACUGAAAC*AACUAAGGCAUCCCAUAGCAUUCC GCUACGAUCUAUCAG*CUUUCCUUCUACUAAUCC CGGCGAUCGGG/ACUGAAAC*UAAUAAUUGCUGAU UAAUCGUGGUGGUGGUGGUGG	187, 144, 119, 76, 68, 43
CRISPR1_IX_113nt	GGAACUAAGGCAUCCCAUAGCAUUCCGCUACGAU CUAUCAG*CUUUCCUUCUACUAAUCCCGGCGAUC GGG/ACUGAAAC*UAAUAAUUGCUGAUUAAUCGUG GUGGUGGUGGUGG	113, 70, 43
CRISPR2		
ID	**Sequence**	**Possible fragment sizes [nt]**
CRISPR2_391nt	GGCCCCGCCCCCGUGGUGGGA*GUUCAACACCCU CUUUUCCCCGUCAGGGG/ACUGAAAC*UGUGAGUU GCAUAAUGCCUCCUAAUGGCUGUUGGACUCAUA A*GUUCAACACCCUCUUUUCCCCGUCAGGGG/ACU GAAAC*CUUGGUAUUUGUAGUUCUCGAUGAGUGU UUUAGGCA*GUUCAACACCCUCUUUUCCCCGUCA GGGG/ACUGAAAC*UGAUAACGGGAUGCCAGCCCU AAAGGUGAUGAGCGG*GUUCAACACCCUCUUUUC CCCGUCAGGGG/ACUGAAAC*CGUUAUCCGGCAAA GAAACCACACUACUAAGCUCGACAA*GUUCAACAC CCUCUUUUCCCCGUCAGGGG/ACUGAAAC*UGGGC CGGGCGCGAGUUGUCCUCCUGUCCGAG	391, 351, 341, 301, 275, 262, 225, 222, 202, 189, 152, 149, 146, 129, 116, 79, 76, 73, 50, 40
CRISPR2_I_121nt	GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCC CGUGGUGGGA*GUUCAACACCCUCUUUUCCCCGU CAGGGG/ACUGAAAC*UGUGAGUUGCAUAAUGCCU CCUAAUGGCUGUUGGACUCAUAA	121, 71, 50
CRISPR2_II_194nt	GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCC CGUGGUGGGA*GUUCAACACCCUCUUUUCCCCGU CAGGGG/ACUGAAAC*UGUGAGUUGCAUAAUGCCU CCUAAUGGCUGUUGGACUCAUAA*GUUCAACACC CUCUUUUCCCCGUCAGGGG/ACUGAAAC*CUUGGU AUUUGUAGUUCUCGAUGAGUGUUUUAGGCA	194, 150, 123, 79, 71, 44
CRISPR2_III_267nt	GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCC CGUGGUGGGA*GUUCAACACCCUCUUUUCCCCGU CAGGGG/ACUGAAAC*UGUGAGUUGCAUAAUGCCU CCUAAUGGCUGUUGGACUCAUAA*GUUCAACACC CUCUUUUCCCCGUCAGGGG/ACUGAAAC*CUUGGU	267, 223, 196, 152 150, 117, 79, 73, 71, 44

	AUUUGUAGUUCUCGAUGAGUGUUUUAGGCA*GUU CAACACCCUCUUUUCCCCGUCAGGGG/ACUGAAA C*UGAUAACGGGAUGCCAGCCCUAAAGGUGAUGA GCGG	

CRISPR2_IV_343nt	GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCC CGUGGUGGGA*GUUCAACACCCUCUUUUCCCCGU CAGGGG/ACUGAAAC*UGUGAGUUGCAUAAUGCCU CCUAAUGGCUGUUGGACUCAUAA*GUUCAACACC CUCUUUUCCCCGUCAGGGG/ACUGAAAC*CUUGGU AUUUGUAGUUCUCGAUGAGUGUUUUAGGCA*GUU CAACACCCUCUUUUCCCCGUCAGGGG/ACUGAAA* CUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGA GCGG*GUUCAACACCCUCUUUUCCCCGUCAGGGG /ACUGAAAC*CGUUAUCCGGCAAAGAAACCACACUA CUAAGCUCGACAA	343, 296, 272, 225 223, 193, 152, 150, 146, 120, 79, 73, 71, 47
CRISPR2_V_419nt	GGCGGGGCUUGGGGGGUUGGAGUCCCCGCCCC CGUGGUGGGA*GUUCAACACCCUCUUUUCCCCGU CAGGGG/ACUGAAAC*UGUGAGUUGCAUAAUGCCU CCUAAUGGCUGUUGGACUCAUAA*GUUCAACACC CUCUUUUCCCCGUCAGGGG/ACUGAAAC*CUUGGU AUUUGUAGUUCUCGAUGAGUGUUUUAGGCA*GUU CAACACCCUCUUUUCCCCGUCAGGGG/ACUGAAA* CUGAUAACGGGAUGCCAGCCCUAAAGGUGAUGA GCGG*GUUCAACACCCUCUUUUCCCCGUCAGGGG /ACUGAAAC*CGUUAUCCGGCAAAGAAACCACACUA CUAAGCUCGACAA*GUUCAACACCCUCUUUUCCCC GUCAGGGG/ACUGAAAC*UGGGCCGGGCGCGAGU UGUCCUCCUGUCCGAGGCCCCAC	419, 372, 348, 301, 296, 269, 225, 223, 222, 196, 152, 150, 149, 146, 123, 79, 76, 73, 73, 71, 47
CRISPR2_VI_342nt	GGUGUGAGUUGCAUAAUGCCUCCUAAUGGCUGU UGGACUCAUAA*GUUCAACACCCUCUUUUCCCCG UCAGGGG/ACUGAAAC*CUUGGUAUUUGUAGUUCU CGAUGAGUGUUUUAGGCA*GUUCAACACCCUCUU UUCCCCGUCAGGGG/ACUGAAAC*UGAUAACGGGA UGCCAGCCCUAAAGGUGAUGAGCGG*GUUCAACA CCCUCUUUUCCCCGUCAGGGG/ACUGAAAC*CGUU AUCCGGCAAAGAAACCACACUACUAAGCUCGACA A*GUUCAACACCCUCUUUUCCCCGUCAGGGG/ACU GAAAC*UGGGCCGGGCGCGAGUUGUCCUCCUGUC CGAGGCCCCAC	342, 295, 269, 222, 219, 196, 149, 146, 146, 123, 76, 73, 73, 73, 47
CRISPR2_VII_263nt	GGCUUGGUAUUUGUAGUUCUCGAUGAGUGUUUU AGGCA*GUUCAACACCCUCUUUUCCCCGUCAGGG G/ACUGAAAC*UGAUAACGGGAUGCCAGCCCUAAA GGUGAUGAGCGG*GUUCAACACCCUCUUUUCCCC GUCAGGGG/ACUGAAAC*CGUUAUCCGGCAAAGAA ACCACACUACUAAGCUCGACAA*GUUCAACACCCU CUUUUCCCCGUCAGGGG/ACUGAAAC*UGGGCCG GGCGCGAGUUGUCCUCCUGUCCGAGGCCCCAC	263, 216, 196, 149, 140, 123, 76, 73, 67, 47
CRISPR2_VIII_190nt	GGUGAUAACGGGAUGCCAGCCCUAAAGGUGAUG AGCGG*GUUCAACACCCUCUUUUCCCCGUCAGGG G/ACUGAAAC*CGUUAUCCGGCAAAGAAACCACAC UACUAAGCUCGACAA*GUUCAACACCCUCUUUUCC CCGUCAGGGG/ACUGAAAC*UGGGCCGGGCGCGA GUUGUCCUCCUGUCCGAGGCCCCAC	190, 143, 123, 76, 67, 47
CRISPR2_IX_117nt	GGCGUUAUCCGGCAAAGAAACCACACUACUAAGC UCGACAA*GUUCAACACCCUCUUUUCCCCGUCAG GGG/ACUGAAAC*UGGGCCGGGCGCGAGUUGUCC UCCUGUCCGAGGCCCCAC	117, 70, 47
CRISPR3		
ID	**Sequence**	**Possible fragment sizes [nt]**
CRISPR3_394nt	GGAGCGCACUGCCUGUCAUUACUAUUA*GUCUCC ACUCGUAGGAGAAAUUA/AUUGAUUGGAAAC*UUA GAUUGCGGGGGCUAGUGACGCCAUAGUUUAACG ACA*GUCUCCACUCGUAGGAGAAAUUA/AUUGAUU*	395, 345, 341, 291, 270, 268, 218, 216, 198, 197, 147, 144, 143, 127, 125, 75,

	*GGAAAC*AAAGAUAUUAGGCUAUCCUUCGGGGUA GUCUUUCUU*GUCUCCACUCGUAGGAGAAAUUA/A* *UUGAUUGGAAAC*AAGUGUUGUUGCCUAGUGUUA UACCAGAAUAUCCC*GUCUCCACUCGUAGGAGAAA* *UUA/AUUGAUUGGAAAC*CUCAUUAGUGCUAUCUU CUUGUUGAUGGAUUAGAACA*GUCUCCACUCGUA* *GGAGAAAUUA/AUUGAUUGGAAAC*UCUACUCUGG UGAAGACCAAGUGGAUUUCGUGGUGAUGUCU	73, 72, 71, 54, 50

Supplemental Table S3. Structure accuracy values of CRISPR2 repeats. Predictions were performed on the whole transcript. In case of cleavage a high accuracy value (→ 1) and in case of non-cleavage a low accuracy value (→ 0) was expected. Cleavage occurrence of CRISPR2 repeats by Cas6-1 or Cas6-2a was determined experimentally (**Figure 3B, D**). Red colored background highlights cases where the experimental result and the definition of the accuracy value are not in agreement.

CRISPR2 Fragment	Repeat	Cleavage by Cas6-1	Cleavage by Cas6-2a	Accuracy of structural repeats
I	R3	Yes	Yes	0.0001
II	R3	No	No	0
II	R4	Yes	Yes	0.9971
III	R3	Yes	Yes	0.8019
III	R4	Yes	Yes	0.9886
III	R5	Yes	Yes	0.9311
IV	R3	No	No	0.8019
IV	R4	Yes	Yes	0.9427
IV	R5	Yes	Yes	0.8878
IV	R6	No	No	0.1229
V	R3	No	No	0.8019
V	R4	Yes	Yes	0.9427
V	R5	Yes	Yes	0.9003
V	R6	No	No	0.1
V	R7	Yes	Yes	0.962
VI	R4	Yes	Yes	0.8906
VI	R5	Yes	Yes	0.929
VI	R6	No	No	0.1
VI	R7	Yes	Yes	0.962
VII	R5	Yes	Yes	0.9128
VII	R6	No	No	0.0857
VII	R7	Yes	Yes	0.962
VIII	R6	No	No	0.004
VIII	R7	Yes	Yes	0.9773
IX	R7	Yes	Yes	0.9832

Supplemental Table S4. Average accuracy values of CRISPR2 repeats. Predictions were performed on the whole CRISPR array.

CRISPR 2 Repeat	Cleaved	Not cleaved	Conflict	% Cleaved	Average accuracy in fragments	Accuracy in CRISPR2 array
R3	2	3	Yes	40	0.55246	0.5453
R4	5	0	No	100	0.9272	0.9060
R5	5	0	No	100	0.9040	0.8509
R6	0	5	No	0	0.2953	0.2425
R7	5	0	No	100	0.9459	0.8877

References

1. 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 Comput. Biol.*, **1**, e60.

2. 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 Res.*, **30**, 482–496.

3. Scholz,I., Lange,S.J., Hein,S., Hess,W.R. and Backofen,R. (2013) CRISPR-Cas systems in the cyanobacterium *Synechocystis* sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS ONE*, **8**, e56470.

Bibliography

[1] R. Barrangou and J. van der Oost, eds., *CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea*. Heidelburg: Springer Press, pp. 1-129, 2013.

[2] J. E. Garneau, M.-E. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadan, and S. Moineau, "The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA," *Nature*, vol. 468, no. 7320, pp. 67–71, 2010.

[3] C. R. Hale, S. Majumdar, J. Elmore, N. Pfister, M. Compton, S. Olson, A. M. Resch, C. V. C. r. Glover, B. R. Graveley, R. M. Terns, and M. P. Terns, "Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs," *Mol Cell*, vol. 45, no. 3, pp. 292–302, 2012.

[4] J. Zhang, C. Rouillon, M. Kerou, J. Reeks, K. Brugger, S. Graham, J. Reimann, G. Cannone, H. Liu, S.-V. Albers, J. H. Naismith, L. Spagnolo, and M. F. White, "Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity," *Mol Cell*, vol. 45, no. 3, pp. 303–13, 2012.

[5] M. M. Jore, M. Lundgren, E. van Duijn, J. B. Bultema, E. R. Westra, S. P. Waghmare, B. Wiedenheft, U. Pul, R. Wurm, R. Wagner, M. R. Beijer, A. Barendregt, K. Zhou, A. P. L. Snijders, M. J. Dickman, J. A. Doudna, E. J. Boekema, A. J. R. Heck, J. van der Oost, and S. J. J. Brouns, "Structural basis for CRISPR RNA-guided DNA recognition by Cascade," *Nat Struct Mol Biol*, vol. 18, no. 5, pp. 529–36, 2011.

[6] K. H. Nam, C. Haitjema, X. Liu, F. Ding, H. Wang, M. P. DeLisa, and A. Ke, "Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system," *Structure*, vol. 20, no. 9, pp. 1574–84, 2012.

[7] S. J. Lange, O. S. Alkhnbashi, D. Rose, S. Will, and R. Backofen, "CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems," vol. 41, no. 17, pp. 8034–44, 2013. SJL, OSA and DR contributed equally to this work.

[8] O. S. Alkhnbashi, F. Costa, S. A. Shah, R. A. Garrett, S. J. Saunders, and R. Backofen, "CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci," *Bioinformatics*, vol. 30, no. 17, pp. i489–i496, 2014. In the proceedings of the 13th European Conference on Computational Biology (ECCB) 2014.

[9] O. S. Alkhnbashi, S. A. Shah, R. A. Garrett, S. J. Saunders, F. Costa, and R. Backofen, "Characterizing leader sequences of CRISPR loci," *Bioinformatics*, vol. 32, no. 17, pp. i576–i585, 2016.

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

[11] V. Reimann, O. S. Alkhnbashi, S. J. Saunders, I. Scholz, S. Hein, R. Backofen, and W. R. Hess, "Structural constraints and enzymatic promiscuity in the Cas6-dependent generation of crRNAs," 2016.

[12] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–3, 1970.

[13] C. A. Suttle, "Viruses in the sea," *Nature*, vol. 437, no. 7057, pp. 356–61, 2005.

[14] J. Ortin and F. Parra, "Structure and function of RNA replication," vol. 60, pp. 305–26, 2006.

[15] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nat Rev Genet*, vol. 5, no. 7, pp. 522–31, 2004.

[16] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals," *Nat Rev Mol Cell Biol*, vol. 10, no. 2, pp. 126–39, 2009.

[17] S. Erdmann and R. A. Garrett, "Archaeal Viruses of the Sulfolobales: Isolation, Infection, and CRISPR Spacer Acquisition," *Methods Mol Biol*, vol. 1311, pp. 223–32, 2015.

[18] J. John P. Donahue, Richard M. Peek, ed., *Restriction and Modification Systems, chapter 24*. In Mobley H, Mendz G, Hazell S (ed), Helicobacter pylori. ASM Press, Washington, DC, 2001.

[19] G. G. Wilson and N. E. Murray, "Restriction and modification systems," vol. 25, pp. 585–627, 1991.

[20] G. G. Wilson, "Organization of restriction-modification systems," vol. 19, no. 10, pp. 2539–66, 1991.

[21] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, "Comparative genomics of defense systems in archaea and bacteria," vol. 41, no. 8, pp. 4360–77, 2013.

[22] M. P. Terns and R. M. Terns, "CRISPR-based adaptive immune systems," *Curr Opin Microbiol*, vol. 14, no. 3, pp. 321–7, 2011.

[23] D. Bhaya, M. Davison, and R. Barrangou, "CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation," vol. 45, pp. 273–97, 2011.

[24] R. Sorek, V. Kunin, and P. Hugenholtz, "CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea," *Nat Rev Microbiol*, vol. 6, no. 3, pp. 181–6, 2008.

[25] R. Jansen, J. D. A. v. Embden, W. Gaastra, and L. M. Schouls, "Identification of genes that are associated with DNA repeats in prokaryotes," *Mol Microbiol*, vol. 43, no. 6, pp. 1565–75, 2002.

[26] R. K. Lillestol, S. A. Shah, K. Brugger, P. Redder, H. Phan, J. Christiansen, and R. A. Garrett, "CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties," *Mol Microbiol*, vol. 72, no. 1, pp. 259–72, 2009.

[27] I. Grissa, G. Vergnaud, and C. Pourcel, "The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats," *BMC Bioinformatics*, vol. 8, p. 172, 2007.

[28] F. J. M. Mojica and R. A. Garrett, "Discovery and seminal developments in the CRISPR field," in *CRISPR-Cas Systems*, pp. 1–31, Springer Berlin Heidelberg, 2013.

[29] S. Chakraborty, A. P. Snijders, R. Chakravorty, M. Ahmed, A. M. Tarek, and M. A. Hossain, "Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria," *Mol Phylogenet Evol*, vol. 56, no. 3, pp. 878–87, 2010.

[30] S. Minot, R. Sinha, J. Chen, H. Li, S. A. Keilbaugh, G. D. Wu, J. D. Lewis, and F. D. Bushman, "The human gut virome: inter-individual variation and dynamic response to diet," *Genome Res*, vol. 21, no. 10, pp. 1616–25, 2011.

[31] I. Garcia-Heredia, A.-B. Martin-Cuadrado, F. J. M. Mojica, F. Santos, A. Mira, J. Anton, and F. Rodriguez-Valera, "Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses," *PLoS One*, vol. 7, no. 3, p. e33802, 2012.

[32] B. Greve, S. Jensen, K. Brugger, W. Zillig, and R. A. Garrett, "Genomic comparison of archaeal conjugative plasmids from Sulfolobus," *Archaea*, vol. 1, no. 4, pp. 231–9, 2004.

[33] I. Grissa, G. Vergnaud, and C. Pourcel, "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats," *NAR*, vol. 35, no. Web Server issue, pp. W52–7, 2007.

[34] C. Bland, T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, and P. Hugenholtz, "CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats," *BMC Bioinformatics*, vol. 8, p. 209, 2007.

[35] R. C. Edgar, "PILER-CR: fast and accurate identification of CRISPR repeats," *BMC Bioinformatics*, vol. 8, p. 18, 2007.

[36] Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, and A. Nakata, "Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product," *J Bacteriol*, vol. 169, no. 12, pp. 5429–33, 1987.

[37] E. R. Westra, D. C. Swarts, R. H. J. Staals, M. M. Jore, S. J. J. Brouns, and J. van der Oost, "The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity," vol. 46, pp. 311–39, 2012.

[38] J. D. van Embden, T. van Gorkom, K. Kremer, R. Jansen, B. A. van Der Zeijst, and L. M. Schouls, "Genetic variation and evolutionary origin of the direct repeat locus of Mycobacterium tuberculosis complex bacteria," *J Bacteriol*, vol. 182, no. 9, pp. 2393–401, 2000.

[39] N. Hoe, K. Nakashima, D. Grigsby, X. Pan, S. J. Dou, S. Naidich, M. Garcia, E. Kahn, D. Bergmire-Sweat, and J. M. Musser, "Rapid molecular genetic subtyping of serotype M1 group A Streptococcus strains," *Emerg Infect Dis*, vol. 5, no. 2, pp. 254–63, 1999.

[40] B. Masepohl, K. Gorlitz, and H. Bohme, "Long tandemly repeated repetitive (LTRR) sequences in the filamentous cyanobacterium Anabaena sp. PCC 7120," *Biochim Biophys Acta*, vol. 1307, no. 1, pp. 26–30, 1996.

[41] F. J. Mojica, C. Ferrer, G. Juez, and F. Rodriguez-Valera, "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," *Mol Microbiol*, vol. 17, no. 1, pp. 85–93, 1995.

[42] P. M. Groenen, A. E. Bunschoten, D. van Soolingen, and J. D. van Embden, "Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method," *Mol Microbiol*, vol. 10, no. 5, pp. 1057–65, 1993.

[43] V. Kunin, R. Sorek, and P. Hugenholtz, "Evolutionary conservation of sequence and secondary structures in CRISPR repeats," *Genome Biol*, vol. 8, no. 4, p. R61, 2007.

[44] F. J. M. Mojica, C. Diez-Villasenor, J. Garcia-Martinez, and E. Soria, "Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements," *J Mol Evol*, vol. 60, no. 2, pp. 174–82, 2005.

[45] C. Pourcel, G. Salvignol, and G. Vergnaud, "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies," *Microbiology*, vol. 151, no. Pt 3, pp. 653–63, 2005.

[46] A. Stern, L. Keren, O. Wurtzel, G. Amitai, and R. Sorek, "Self-targeting by CRISPR: gene regulation or autoimmunity?," vol. 26, no. 8, pp. 335–40, 2010.

[47] A. Bolotin, B. Quinquis, A. Sorokin, and S. D. Ehrlich, "Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin," *Microbiology*, vol. 151, no. Pt 8, pp. 2551–61, 2005.

[48] R. K. Lillestol, P. Redder, R. A. Garrett, and K. Brugger, "A putative viral defence mechanism in archaeal cells," *Archaea*, vol. 2, no. 1, pp. 59–72, 2006.

[49] M. Touchon and E. P. C. Rocha, "The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella," *PLoS One*, vol. 5, no. 6, p. e11126, 2010.

[50] C. Diez-Villasenor, C. Almendros, J. Garcia-Martinez, and F. J. M. Mojica, "Diversity of CRISPR loci in Escherichia coli," *Microbiology*, vol. 156, no. Pt 5, pp. 1351–61, 2010.

[51] S. Gudbergsdottir, L. Deng, Z. Chen, J. V. K. Jensen, L. R. Jensen, Q. She, and R. A. Garrett, "Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers," *Mol Microbiol*, vol. 79, no. 1, pp. 35–49, 2011.

[52] S. A. Shah and R. A. Garrett, "CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems," *Res Microbiol*, vol. 162, no. 1, pp. 27–38, 2011.

[53] S. J. J. Brouns, M. M. Jore, M. Lundgren, E. R. Westra, R. J. H. Slijkhuis, A. P. L. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, and J. van der Oost, "Small CRISPR RNAs guide antiviral defense in prokaryotes," *Science*, vol. 321, no. 5891, pp. 960–4, 2008.

[54] I. Yosef, M. G. Goren, and U. Qimron, "Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli," vol. 40, no. 12, pp. 5569–76, 2012.

[55] S. Erdmann and R. A. Garrett, "Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms," *Mol Microbiol*, vol. 85, no. 6, pp. 1044–56, 2012.

[56] C. Diez-Villasenor, N. M. Guzman, C. Almendros, J. Garcia-Martinez, and F. J. M. Mojica, "CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli," *RNA Biol*, vol. 10, no. 5, pp. 792–802, 2013.

[57] O. Wurtzel, R. Sapra, F. Chen, Y. Zhu, B. A. Simmons, and R. Sorek, "A single-base resolution map of an archaeal transcriptome," *Genome Res*, vol. 20, no. 1, pp. 133–41, 2010.

[58] L. Deng, C. S. Kenchappa, X. Peng, Q. She, and R. A. Garrett, "Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus," vol. 40, no. 6, pp. 2470–80, 2012.

[59] R. A. Garrett, G. Vestergaard, and S. A. Shah, "Archaeal CRISPR-based immune systems: exchangeable functional modules," *Trends Microbiol*, vol. 19, no. 11, pp. 549–56, 2011.

[60] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath, "CRISPR provides acquired resistance against viruses in prokaryotes," *Science*, vol. 315, no. 5819, pp. 1709–12, 2007.

[61] D. C. Swarts, C. Mosterd, M. W. J. van Passel, and S. J. J. Brouns, "CRISPR interference directs strand specific spacer acquisition," *PLoS One*, vol. 7, no. 4, p. e35888, 2012.

[62] E. Semenova, M. M. Jore, K. A. Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, and K. Severinov, "Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence," vol. 108, no. 25, pp. 10098–103, 2011.

[63] S. A. Shah, N. R. Hansen, and R. A. Garrett, "Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism," *Biochem Soc Trans*, vol. 37, no. Pt 1, pp. 23–8, 2009.

[64] S. A. Shah, S. Erdmann, F. J. M. Mojica, and R. A. Garrett, "Protospacer recognition motifs: Mixed identities and functional diversity," *RNA Biol*, vol. 10, no. 5, 2013.

[65] D. H. Haft, J. Selengut, E. F. Mongodin, and K. E. Nelson, "A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes," *PLoS Comput Biol*, vol. 1, no. 6, p. e60, 2005.

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

[67] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, "The basic building blocks and evolution of CRISPR-CAS systems," *Biochem Soc Trans*, vol. 41, no. 6, pp. 1392–400, 2013.

[68] K. S. Makarova, D. H. Haft, R. Barrangou, S. J. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. M. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost, and E. V. Koonin, "Evolution and classification of the CRISPR-Cas systems," *Nat Rev Microbiol*, vol. 9, no. 6, pp. 467–77, 2011.

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

[70] N. Takeuchi, Y. I. Wolf, K. S. Makarova, and E. V. Koonin, "Nature and intensity of selection pressure on CRISPR-associated genes," *J Bacteriol*, vol. 194, no. 5, pp. 1216–25, 2012.

[71] M. Krupovic, K. S. Makarova, P. Forterre, D. Prangishvili, and E. V. Koonin, "Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity," *BMC Biol*, vol. 12, p. 36, 2014.

[72] N. Beloglazova, G. Brown, M. D. Zimmerman, M. Proudfoot, K. S. Makarova, M. Kudritska, S. Kochinyan, S. Wang, M. Chruszcz, W. Minor, E. V. Koonin, A. M. Edwards, A. Savchenko, and A. F. Yakunin, "A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats," vol. 283, no. 29, pp. 20361–71, 2008.

[73] A.-R. Kwon, J.-H. Kim, S. J. Park, K.-Y. Lee, Y.-H. Min, H. Im, I. Lee, K.-Y. Lee, and B.-J. Lee, "Structural and biochemical characterization of HP0315 from Helicobacter pylori as a VapD protein with an endoribonuclease activity," vol. 40, no. 9, pp. 4216–28, 2012.

[74] D. A. Daines, J. Jarisch, and A. L. Smith, "Identification and characterization of a nontypeable Haemophilus influenzae putative toxin-antitoxin locus," *BMC Microbiol*, vol. 4, p. 30, 2004.

[75] K. S. Makarova, L. Aravind, N. V. Grishin, I. B. Rogozin, and E. V. Koonin, "A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis," vol. 30, no. 2, pp. 482–96, 2002.

[76] G. Vestergaard, R. A. Garrett, and S. A. Shah, "CRISPR adaptive immune systems of Archaea," *RNA Biol*, vol. 11, no. 2, pp. 157–168, 2014.

[77] T. Liu, Y. Li, X. Wang, Q. Ye, H. Li, Y. Liang, Q. She, and N. Peng, "Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition," vol. 43, no. 2, pp. 1044–55, 2015.

[78] M. Li, R. Wang, D. Zhao, and H. Xiang, "Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process," vol. 42, no. 4, pp. 2483–92, 2014.

[79] C. Richter, R. L. Dy, R. E. McKenzie, B. N. J. Watson, C. Taylor, J. T. Chang, M. B. McNeil, R. H. J. Staals, and P. C. Fineran, "Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer," vol. 42, no. 13, pp. 8516–26, 2014.

[80] K. C. Cady, J. Bondy-Denomy, G. E. Heussler, A. R. Davidson, and G. A. O'Toole, "The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages," *J Bacteriol*, vol. 194, no. 21, pp. 5728–38, 2012.

[81] Y. Wei, M. T. Chesne, R. M. Terns, and M. P. Terns, "Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus," vol. 43, no. 3, pp. 1749–58, 2015.

[82] B. Wiedenheft, K. Zhou, M. Jinek, S. M. Coyle, W. Ma, and J. A. Doudna, "Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense," *Structure*, vol. 17, no. 6, pp. 904–12, 2009.

[83] N. L. Held, A. Herrera, H. Cadillo-Quiroz, and R. J. Whitaker, "CRISPR associated diversity within a population of Sulfolobus islandicus," *PLoS One*, vol. 5, no. 9, 2010.

[84] F. J. M. Mojica, C. Diez-Villasenor, J. Garcia-Martinez, and C. Almendros, "Short motif sequences determine the targets of the prokaryotic CRISPR defence system," *Microbiology*, vol. 155, no. Pt 3, pp. 733–40, 2009.

[85] P. C. Fineran and E. Charpentier, "Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information," *Virology*, vol. 434, no. 2, pp. 202–9, 2012.

[86] E. Charpentier, H. Richter, J. van der Oost, and M. F. White, "Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity," *FEMS Microbiol Rev*, vol. 39, no. 3, pp. 428–41, 2015.

[87] T.-H. Tang, J.-P. Bachellerie, T. Rozhdestvensky, M.-L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius, and A. Huttenhofer, "Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus," vol. 99, no. 11, pp. 7536–41, 2002.

[88] T.-H. Tang, N. Polacek, M. Zywicki, H. Huber, K. Brugger, R. Garrett, J. P. Bachellerie, and A. Huttenhofer, "Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus," *Mol Microbiol*, vol. 55, no. 2, pp. 469–81, 2005.

[89] N. G. Lintner, M. Kerou, S. K. Brumfield, S. Graham, H. Liu, J. H. Naismith, M. Sdano, N. Peng, Q. She, V. Copie, M. J. Young, M. F. White, and C. M. Lawrence, "Structural and

functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE)," vol. 286, no. 24, pp. 21643–56, 2011.

[90] E. M. Gesner, M. J. Schellenberg, E. L. Garside, M. M. George, and A. M. Macmillan, "Recognition and maturation of effector RNAs in a CRISPR interference pathway," *Nat Struct Mol Biol*, vol. 18, no. 6, pp. 688–92, 2011.

[91] A. Hatoum-Aslan, I. Maniv, and L. A. Marraffini, "Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site," vol. 108, no. 52, pp. 21218–22, 2011.

[92] R. E. Haurwitz, S. H. Sternberg, and J. A. Doudna, "Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA," *EMBO J*, vol. 31, no. 12, pp. 2824–32, 2012.

[93] R. Wang, H. Zheng, G. Preamplume, Y. Shao, and H. Li, "The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex," *Protein Sci*, vol. 21, no. 3, pp. 405–17, 2012.

[94] S. Juranek, T. Eban, Y. Altuvia, M. Brown, P. Morozov, T. Tuschl, and H. Margalit, "A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs," *RNA*, vol. 18, no. 4, pp. 783–94, 2012.

[95] J. Brendel, B. Stoll, S. J. Lange, K. Sharma, C. Lenz, A.-E. Stachler, L.-K. Maier, H. Richter, L. Nickel, R. A. Schmitz, L. Randau, T. Allers, H. Urlaub, R. Backofen, and A. Marchfelder, "A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of crRNAs in Haloferax volcanii," vol. 289, no. 10, pp. 7164–77, 2014.

[96] B. Stoll, L.-K. Maier, S. J. Lange, J. Brendel, S. Fischer, R. Backofen, and A. Marchfelder, "Requirements for a successful defence reaction by the CRISPR-Cas subtype I-B system," *Biochem Soc Trans*, vol. 41, no. 6, pp. 1444–8, 2013.

[97] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity," *Science*, vol. 337, no. 6096, pp. 816–21, 2012.

[98] E. Deltcheva, K. Chylinski, C. M. Sharma, K. Gonzales, Y. Chao, Z. A. Pirzada, M. R. Eckert, J. Vogel, and E. Charpentier, "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III," *Nature*, vol. 471, no. 7340, pp. 602–7, 2011.

[99] Y. Zhang, N. Heidrich, B. J. Ampattu, C. W. Gunderson, H. S. Seifert, C. Schoen, J. Vogel, and E. J. Sontheimer, "Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis," *Mol Cell*, vol. 50, no. 4, pp. 488–503, 2013.

[100] C. R. Hale, P. Zhao, S. Olson, M. O. Duff, B. R. Graveley, L. Wells, R. M. Terns, and M. P. Terns, "RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex," *Cell*, vol. 139, no. 5, pp. 945–56, 2009.

[101] H. Deveau, R. Barrangou, J. E. Garneau, J. Labonte, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, and S. Moineau, "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus," *J Bacteriol*, vol. 190, no. 4, pp. 1390–400, 2008.

[102] S. Fischer, L.-K. Maier, B. Stoll, J. Brendel, E. Fischer, F. Pfeiffer, M. Dyall-Smith, and A. Marchfelder, "An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA," vol. 287, no. 40, pp. 33351–63, 2012.

[103] S. Erdmann, S. Le Moine Bauer, and R. A. Garrett, "Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus," *Mol Microbiol*, vol. 91, no. 5, pp. 900–17, 2014.

[104] P. C. Fineran, M. J. H. Gerritzen, M. Suarez-Diez, T. Kunne, J. Boekhorst, S. A. F. T. van Hijum, R. H. J. Staals, and S. J. J. Brouns, "Degenerate target sites mediate rapid primed CRISPR adaptation," vol. 111, no. 16, pp. E1629–38, 2014.

[105] L. A. Marraffini and E. J. Sontheimer, "Self versus non-self discrimination during CRISPR RNA-directed immunity," *Nature*, vol. 463, no. 7280, pp. 568–71, 2010.

[106] B. Wiedenheft, G. C. Lander, K. Zhou, M. M. Jore, S. J. J. Brouns, J. van der Oost, J. A. Doudna, and E. Nogales, "Structures of the RNA-guided surveillance complex from a bacterial immune system," *Nature*, vol. 477, no. 7365, pp. 486–9, 2011.

[107] A. Plagens, V. Tripp, M. Daume, K. Sharma, A. Klingl, A. Hrle, E. Conti, H. Urlaub, and L. Randau, "In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex," vol. 42, no. 8, pp. 5125–38, 2014.

[108] L. Deng, R. A. Garrett, S. A. Shah, X. Peng, and Q. She, "A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus," *Mol Microbiol*, vol. 87, no. 5, pp. 1088–99, 2013.

[109] W. Peng, M. Feng, X. Feng, Y. X. Liang, and Q. She, "An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference," vol. 43, no. 1, pp. 406–17, 2015.

[110] H. Wang, M. La Russa, and L. S. Qi, "CRISPR/Cas9 in Genome Editing and Beyond," vol. 85, pp. 227–64, 2016.

[111] B. L. Stoddard, "Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification," *Structure*, vol. 19, no. 1, pp. 7–15, 2011.

[112] F. D. Urnov, E. J. Rebar, M. C. Holmes, H. S. Zhang, and P. D. Gregory, "Genome editing with engineered zinc finger nucleases," *Nat Rev Genet*, vol. 11, no. 9, pp. 636–46, 2010.

[113] A. M. Scharenberg, P. Duchateau, and J. Smith, "Genome engineering with TAL-effector nucleases and alternative modular nuclease technologies," *Curr Gene Ther*, vol. 13, no. 4, pp. 291–303, 2013.

[114] A. J. Bogdanove and D. F. Voytas, "TAL effectors: customizable proteins for DNA targeting," *Science*, vol. 333, no. 6051, pp. 1843–6, 2011.

[115] J. D. Sander and J. K. Joung, "CRISPR-Cas systems for editing, regulating and targeting genomes," *Nat Biotechnol*, vol. 32, no. 4, pp. 347–55, 2014.

[116] P. D. Hsu, E. S. Lander, and F. Zhang, "Development and applications of CRISPR-Cas9 for genome engineering," *Cell*, vol. 157, no. 6, pp. 1262–78, 2014.

[117] A. Smola and V. S.V.N., eds., *Introduction to Machine Learning*. Cambridge University Press, pp. 1-234, 2008.

[118] S. Prompramote, Y. Chen, and Y.-P. Chen, "Machine learning in bioinformatics: in bioinformatics technologies," *Springer Heidelberg,Germany*, no. 36, pp. 117–153, 2005.

[119] A. Hojjat and H. Shih-Lin, eds., *Machine Learning: Neural Networks, Genetic Algorithms, and Fuzzy Systems*. Cambridge University Press, pp. 1-320, 1994.

[120] T. J. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning: Chapter in the book, machine learning: An artificial intelligence approach," *TIOGA Publishing Co., Palo Alto*, no. 36, pp. 3–23, 1983.

[121] L. Breiman, "Heuristics of instability and stabilization in model selection," *Ann. Statist.*, vol. 24, pp. 2350–2383, 12 1996.

[122] G. M. James, "Variance and bias for general loss functions," *Machine Learning*, vol. 51, no. 2, pp. 115–135, 2003.

[123] P. Dayan, "Unsupervised learning," *In R. A. Wilson and F. Keil (Eds). The MIT encyclopedia of the cognitive sciences*, no. 7, 1999.

[124] K. Kundu and R. Backofen, "Cluster based prediction of PDZ-peptide interactions," *BMC Genomics*, vol. 15, no. Suppl 1, p. S5, 2014.

[125] J. Brownlee, "Supervised and unsupervised machine learning algorithms," *In Machine Learning Algorithms*, 2016.

[126] S. Thirumuruganathan, "A detailed introduction to k-nearest neighbor (knn) algorithm," *https://saravananthirumuruganathan.wordpress.com/*, 2010.

[127] T. Srivastava, "Introduction to k-nearest neighbors : Simplified," *Big data: Analytics Vidhya, learn everything about analytics*, 2014.

[128] R. Sebastian, ed., *Python Machine Learning: Unlock deeper insights into Machine Leaning with this vital guide to cutting-edge predictive analytics*. Packt Publishing Ltd. (September 24th, 2015), 2015.

[129] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, ACM Press, 1992.

[130] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," in *Automat. Remote Contr.*, vol. 25, pp. 917 – 936, 1964.

[131] F. Costa and K. D. Grave, "Fast neighborhood subgraph pairwise distance kernel," in *Proceedings of the 26 th International Conference on Machine Learning*, pp. 255–262, Omnipress, 2010.

[132] D. Haussler, "Convolution kernels on discrete structures," Technical Report UCS-CRL-99-10, University of California at Santa Cruz, Santa Cruz, CA, USA, 1999.

[133] F. Costa, "Learning an efficient constructive sampler for graphs," *Artificial Intelligence*, 2016.

[134] E. M. Luks, "Isomorphism of graphs of bounded valence can be tested in polynomial time," *J. Comput. Syst. Sci*, vol. 25, pp. 42–65, 1982.

[135] B. D. McKay, "Practical graph isomorphism," *Congressus Numerantium*, vol. 30, pp. 45–87, 1981.

[136] X. Yan and J. Han., "gSpan: Graph-based substructure pattern mining," in *Proc. 2002 Int. Conf. Data Mining (ICDM ´02)*, pp. 721–724, 2002.

[137] S. Sorlin and C. Solnon, "A parametric filtering algorithm for the graph isomorphism problem," *Constraints*, vol. 13, pp. 518–537, 2008.

[138] I. Damgård, "A design principle for hash functions," in *Advances in Cryptology-CRYPTO´89 Proceedings*, pp. 416–427, Springer, 1990.

[139] C. Leslie, E. Eskin, and W. S. Noble, "The spectrum kernel: a string kernel for SVM protein classification," *Pac. Symp. Biocomput.*, pp. 564–575, 2002.

[140] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20, no. 4, pp. 467–76, 2004.

[141] F. J. Mojica, C. Diez-Villasenor, E. Soria, and G. Juez, "Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria," *Mol Microbiol*, vol. 36, no. 1, pp. 244–6, 2000.

[142] D. G. Sashital, M. Jinek, and J. A. Doudna, "An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3," *Nat Struct Mol Biol*, vol. 18, no. 6, pp. 680–7, 2011.

[143] S. H. Sternberg, R. E. Haurwitz, and J. A. Doudna, "Mechanism of substrate selection by a highly specific CRISPR endoribonuclease," *RNA*, vol. 18, no. 4, pp. 661–72, 2012.

[144] I. Scholz, S. J. Lange, S. Hein, W. R. Hess, and R. Backofen, "CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 Exhibit Distinct Processing Pathways Involving at Least Two Cas6 and a Cmr2 Protein," *PLoS One*, vol. 8, no. 2, p. e56470, 2013. IS and SJL contributed equally to this work.

[145] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler, "Identification and Classification of Conserved RNA Secondary Structures in the Human Genome," *PLoS Comput Biol*, vol. 2, no. 4, p. e33, 2006.

[146] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen, "Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering," *PLoS Comput Biol*, vol. 3, no. 4, p. e65, 2007.

[147] J. H. Havgaard, E. Torarinsson, and J. Gorodkin, "Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix," *PLoS Comput Biol*, vol. 3, no. 10, pp. 1896–908, 2007.

[148] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen, "LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs," *RNA*, vol. 18, no. 5, pp. 900–14, 2012.

[149] A. R. Gruber, S. H. Bernhart, I. L. Hofacker, and S. Washietl, "Strategies for measuring evolutionary conservation of RNA secondary structures," *BMC Bioinformatics*, vol. 9, p. 122, 2008.

[150] P. P. Gardner, J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, R. D. Finn, E. P. Nawrocki, D. L. Kolbe, S. R. Eddy, and A. Bateman, "Rfam: Wikipedia, clans and the "decimal" release," vol. 39, no. Database issue, pp. D141–5, 2011.

[151] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," vol. 30, no. 7, pp. 1575–84, 2002.

[152] L. Nickel, K. Weidenbach, D. Jager, R. Backofen, S. J. Lange, N. Heidrich, and R. A. Schmitz, "Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III," *RNA Biol*, vol. 10, no. 5, pp. 779–791, 2013.

[153] E. L. Garside, M. J. Schellenberg, E. M. Gesner, J. B. Bonanno, J. M. Sauder, S. K. Burley, S. C. Almo, G. Mehta, and A. M. MacMillan, "Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases," *RNA*, vol. 18, no. 11, pp. 2020–8, 2012.

[154] A. Biswas, P. Fineran, and C. Brown, "Accurate computational prediction of the transcribed strand of CRISPR noncoding RNAs," *Bioinformatics*, 2014.

[155] R. A. Garrett, S. A. Shah, S. Erdmann, G. Liu, M. Mousaei, C. Leon-Sobrino, W. Peng, S. Gudbergsdottir, L. Deng, G. Vestergaard, X. Peng, and Q. She, "CRISPR-Cas Adaptive Immune Systems of the Sulfolobales: Unravelling Their Complexity and Diversity," *Life (Basel)*, vol. 5, no. 1, pp. 783–817, 2015.

[156] L. A. Marraffini and E. J. Sontheimer, "CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA," *Science*, vol. 322, no. 5909, pp. 1843–5, 2008.

[157] C. Rollie, S. Schneider, A. S. Brinkmann, E. L. Bolt, and M. F. White, "Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition," *Elife*, vol. 4, 2015.

[158] W. Pearson, "Finding protein and nucleotide similarities with FASTA," *Curr Protoc Bioinformatics*, vol. Chapter 3, p. Unit3.9, 2004.

[159] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases," vol. 23, no. 11, pp. 444–7, 1998.

[160] H. Richter, J. Zoephel, J. Schermuly, D. Maticzka, R. Backofen, and L. Randau, "Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis," vol. 40, no. 19, pp. 9887–96, 2012.